Towards optimal sensor placement for inverse problems in spaces of measures

OPEN ACCESS
IOP Publishing
Inverse Problems

Inverse Problems **40** (2024) 055007 (43pp) https://doi.org/10.1088/1361-6420/ad2cf8

Towards optimal sensor placement for inverse problems in spaces of measures

Phuoc-Truong Huynh[1], **Konstantin Pieper**[2] and **Daniel Walter**[3,*]

[1] Institut für Mathematik, Alpen-Adria-Universität Klagenfurt, 9020 Klagenfurt, Austria
[2] Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, United States of America
[3] Institut für Mathematik, Humboldt-Universität zu Berlin, 10117 Berlin, Germany

E-mail: daniel.walter@hu-berlin.de, phuoc.huynh@aau.at and pieperk@ornl.gov

Received 7 August 2023; revised 24 January 2024
Accepted for publication 26 February 2024
Published 25 March 2024

Abstract

The objective of this work is to quantify the reconstruction error in sparse inverse problems with measures and stochastic noise, motivated by optimal sensor placement. To be useful in this context, the error quantities must be explicit in the sensor configuration and robust with respect to the source, yet relatively easy to compute in practice, compared to a direct evaluation of the error by a large number of samples. In particular, we consider the identification of a measure consisting of an unknown linear combination of point sources from a finite number of measurements contaminated by Gaussian noise. The statistical framework for recovery relies on two main ingredients: first, a convex but non-smooth variational Tikhonov point estimator over the space of Radon measures and, second, a suitable mean-squared error based on its Hellinger–Kantorovich distance to the ground truth. To quantify the error, we employ a non-degenerate source condition as well as careful linearization arguments to derive a computable upper bound. This leads to asymptotically sharp error estimates in expectation that are explicit in the sensor configuration. Thus they can be used to estimate the expected reconstruction error for a given sensor configuration and guide the placement of sensors in sparse inverse problems.

* Author to whom any correspondence should be addressed.

Keywords: inverse problems, optimal sensor placement, Radon measures, off-the-grid sparse recovery, frequentistic-inference

1. Introduction

The identification of an unknown signal $\mu^\dagger$ comprising finitely many point sources lies at the heart of challenging applications such as acoustic inversion [20, 30], microscopy [10, 25], astronomy [35], low-rank tensor decomposition [23], linear system identification [3], as well as initial value identification [7, 8, 22]. Moreover, the recovery of an unknown function by one-hidden-layer neural networks [2, 9, 29] is intrinsically linked to this task. In all of these contexts the problem is to identify an unknown linear combination (superposition) of functions indexed by a nonlinear parameter from a finite number of measurements. Motivated by inverse point source location tasks we will refer to the linear parameters as *amplitudes* and nonlinear parameters as *locations*. Moreover, we will assume that measurements are associated to certain spatial locations, motivated by point-wise measurements of physical quantities. Denoting by $\Omega_s \subset \mathbb{R}^d$ and $\Omega_o \subset \mathbb{R}^{d_o}$, $d, d_o \geqslant 1$, compact sets of possible source locations and measurement points, a common mathematical framework for the recovery of the locations $y_n^\dagger \in \Omega_s$ and amplitudes $q_n^\dagger$ of its $N_s^\dagger$ individual point sources can be given by equations of the form

$$z_j^d(\varepsilon) = \sum_{n=1}^{N_s^\dagger} q_n^\dagger k\left(x_j, y_n^\dagger\right) + \varepsilon_j \quad \text{for } j = 1, \ldots, N_o; \tag{1.1}$$

Here, $k \in \mathcal{C}(\Omega_o \times \Omega_s)$ denotes a sufficiently smooth given integral kernel (resulting from the modeling of the physical process and the properties of the sensors), and $x_j \in \Omega_o$ denote measurement locations. Moreover, ε_j is a measurement error for each sensor that, for the purposes of this paper is thought of as a random perturbation stemming from measurement noise. This type of *ill-posed inverse problem* is challenging for a variety of reasons. First and foremost, we neither assume knowledge of the amplitudes and positions of the sources nor of their number. This adds an additional combinatorial component to the generally nonlinear nonconvex problem. Second, inference on $\mu^\dagger$ is only possible through a finite number of indirect measurements z^d. Additional challenges are given by the appearance of unobservable measurement noise ε in the problem.

To alleviate some of these difficulties we identify $\mu^\dagger$ with a finite linear combination of Dirac measures

$$\mu^\dagger = \sum_{n=1}^{N_s^\dagger} q_n^\dagger \delta_{y_n^\dagger} \quad \text{and } z_j^d(\varepsilon) = \int_{\Omega_s} k(x_j, y)\, \mathrm{d}\mu^\dagger(y) + \varepsilon_j. \tag{1.2}$$

Subsequently, we try to recover $\mu^\dagger$ by the stable solution of the linear, ill-posed, operator equation:

$$\text{Find } \mu \in \mathcal{M}(\Omega_s): \quad K\mu \approx z^d(\varepsilon) \quad \text{where } K\mu = \left(\int_\Omega k(x_1, y)\ \mathrm{d}\mu(y); \ldots; \int_\Omega k(x_{N_o}, y)\ \mathrm{d}\mu(y) \right).$$

Here, $\mathcal{M}(\Omega_s)$ is the space of Radon measures defined on the location set Ω_s. At first glance, this might seem counter-intuitive: The space $\mathcal{M}(\Omega_s)$ is much larger than the set of 'sparse'

signals of the form (1.2). Thus, this lifting should only contribute to the ill-posedness of the problem. However, it also bypasses the nonlinear dependency of $k(x_j, \cdot)$ onto the location of the sources and enables the use of powerful tools from variational regularization theory for the reconstruction of $\mu^\dagger$. In this work, stable recovery of μ is facilitated by a variational Tikhonov estimator in the space of Radon measures [4, 19], which amounts to solving a nonsmooth minimization problem over this space.

However, measurements stemming from experiments are always affected by errors, either due to external influences, imperfectness of the measurement devices or human failure. These have to be taken into account in order to guarantee a stable recovery of $\mu^\dagger$. In particular, it is evident that the choice of the measurement locations and the quality of the employed sensors is a key factor for the successful and robust reconstruction of the signal. This directly leads to the problem of sensor design, which is to identify a measurement configuration leading to recovery guarantees with minimal error for the given effort, in a suitable way. Since the sensor design must usually be chosen before the exact source is known and the practical measurement has been performed (thus yielding a realization of the noise), this usually calls for a stochastic framework for the noise. Although much is know about the error caused by deterministic noise [1, 11, 13, 33, 34], we are not aware of any works pertaining to the case of stochastic noise in this context. Moreover, existing deterministic bounds on the error of the recovery $\mu(\varepsilon)$ to the ground truth $\mu^\dagger$ are not explicit in terms of the measurement locations x_j the statistical properties of the error ε_j and ground truth $\mu^\dagger$, and thus can not directly be used to quantify the influence of the measurement locations on the error. The explicit dependency on the measurement setup is needed to guide the choice of an optimal design that minimizes the expected recovery error for a given cost (often measured in terms of number and quality of sensors), while robustness with respect to the ground truth is desirable if only an approximate guess of the exact source is available (which is the realistic case, in practice).

In addition, to quantify the error, often estimates are given separately in terms of positions and coefficients, which can then be translated into an an upper bound of the error of the measure, which may be an overestimate by a large factor. To provide a useful bound for sensor placement, we start from the error in the recently developed Hellinger–Kantorovich metric [24], which we then link to the parameters and the quantitative bound that is asymptotically sharp.

1.1. Sparse inverse problems with deterministic noise

Despite the popularity of sparse inverse problems, most of the existing work, to the best of our knowledge, focuses on deterministic noise ε. Central objects in this context, are the (noiseless) *minimum norm problem*

$$\min_{\mu \in \mathcal{M}(\Omega_s)} ||\mu||_{\mathcal{M}(\Omega_s)} \quad \text{subject to } K\mu = K\mu^\dagger \tag{$\mathcal{P}_0$}$$

as well as the question whether $\mu^\dagger$ is *identifiable*, i.e. its unique solution. A sufficient condition for the latter, is, e.g. the injectivity of the restricted operator $K_{|\operatorname{supp}\mu^\dagger}$ as well as the existence of a so-called *dual certificate* $\eta^\dagger \in \mathcal{C}^2(\Omega_s)$, [11], i.e. a subgradient $\eta^\dagger \in \partial\|\mu^\dagger\|_{\mathcal{M}(\Omega_s)}$, which is in some sense minimal, satisfying a *strengthened source condition*

$$|\eta^\dagger(y)| \leqslant 1 \quad \text{for all } y \in \Omega_s,\ \eta^\dagger\left(y_n^\dagger\right) = \operatorname{sign}\left(q_n^\dagger\right), \quad |\eta^\dagger(y)| < 1 \quad \text{for all } y \in \Omega_s \setminus \left\{y^\dagger\right\}_{n=1}^{N_s}.$$

For example, in particular settings, the groundbreaking paper [6] shows that $\mu^\dagger$ is identifiable if the source locations $y_n^\dagger$ are sufficiently well separated. In this context, several manuscripts, see e.g. [1, 11, 13, 34] for a non-exhaustive list, study the approximation of an identifiable $\mu^\dagger$ by solutions to the Tikhonov-regularized problem

$$\bar{\mu}(\varepsilon) \in \mathfrak{M}(\varepsilon) := \operatorname*{arg\,min}_{\mu \in \mathcal{M}(\Omega_s)} \left[\frac{1}{2} ||K\mu - z^d(\varepsilon)||^2_{\Sigma_0^{-1}} + \beta ||\mu||_{\mathcal{M}(\Omega_s)} \right], \tag{$\mathcal{P}_{\beta,\varepsilon}$}$$

where Σ_0 is a positive definite diagonal matrix and the regularization parameter $\beta = \beta(\|\varepsilon\|) > 0$ is adapted to the strength of the noise. This represents a challenging *nonsmooth* minimization problem over the infinite-dimensional and non-reflexive space of Radon measures. Moreover, due to its lack of strict convexity, its solutions are typically not unique. Under mild conditions on the choice of β, arbitrary solutions $\bar{\mu}(\varepsilon)$ approximate $\mu^\dagger$ in the weak*-sense as ε goes to zero. Moreover, it was shown in [11] that if the minimal dual certificate $\eta^\dagger$ associated to problem ($\mathcal{P}_0$) satisfies the strengthened source condition and its curvature does not degenerate around $y_n^\dagger$, $\bar{\mu}(\varepsilon)$ is unique and of the form

$$\bar{\mu}(\varepsilon) = \sum_{n=1}^{N_s^\dagger} \bar{q}_n(\varepsilon)\, \delta_{\bar{y}_n(\varepsilon)} \quad \text{with} \quad |\bar{q}_n(\varepsilon) - q_n^\dagger| + ||\bar{y}_n(\varepsilon) - y_n^\dagger|| = \mathcal{O}(\|\varepsilon\|)$$

provided that $\|\varepsilon\|$ and β are small enough.

1.2. Sparse inverse problems with random noise

From a practical perspective, assuming knowledge on the norm of the error is very restrictive or even unrealistic and a statistical model for the measurement error is more appropriate. While the literature on deterministic sparse inversion is very rich, there are only few works dealing with randomness in the problem. We point out, e.g. [5] in which the authors consider additive i.i.d. noise stemming from a low-pass filtering of the signal. A reconstruction $\bar{\mu}(\varepsilon)$ is obtained by solving a constrained version of ($\mathcal{P}_{\beta,\varepsilon}$) and the authors show that, with high probability, there holds $Q_{\text{hi}}(\bar{\mu}(\varepsilon)) \approx Q_{\text{hi}}(\mu^\dagger)$ where Q_{hi} is a convolution with a high-resolution kernel. Moreover, in [34] the authors consider deterministic noise but allow for randomness in the forward operator K. Their main result provides an estimate on an optimal transport energy between two positive measures derived from source and reconstruction. These again hold with high probability. Finally, we also mention [12] in which the authors propose a first step towards *Bayesian inversion* for sparse problems, i.e. both measurement noise as well as the unknown $\mu^\dagger$ are considered to be random variables. A suitable prior is constructed and well-posedness of the associated Bayesian inverse problem is shown.

In this paper, similar to [5], we adopt a frequentist viewpoint on sparse inverse problems and assume that the measurement errors follow a known probability distribution. In contrast, the unknown signal $\mu^\dagger$ is treated as a deterministic object. More in detail, we assume unbiased independent Gaussian noise with diagonal covariance matrix $\Sigma = \operatorname{diag}(\sigma_j)$, corresponding to independent measurements with variable quality sensors at different locations. We consider the Tikhonov-type estimator ($\mathcal{P}_{\beta,\varepsilon}$) with

$$\Sigma_0^{-1} = \Sigma^{-1}/p, \quad \text{where } p = \operatorname{tr}\left(\Sigma^{-1}\right)$$

and investigate its error to the ground truth, where we have to account for the randomness of the noise. In statistical terms, Σ^{-1} is the precision matrix of the sensor array, and p can be interpreted as an overall precision of the combined measurement, roughly representing an analogue to $1/||\varepsilon||$ in the stochastic setting. First and foremost, the uncertainty of the noise propagates to the estimator and thus $\bar{\mu}$ has to be interpreted as a random variable. Second, unlike the deterministic setting of [11], the asymptotic analysis cannot exclusively rely on smallness assumptions on the Euclidean norm of the noise: some realizations of ε might be very large, albeit with small probability. Consequently, reconstructions can exhibit undesirable features such as clustering phenomena around $y_n^\dagger$ or spurious sources far away from the true support. In particular, the reconstructed signal may comprise more or less than $N_s^\dagger$ sources. Thus, we require a suitable distance between signed measures that is compatible with weak* convergence on bounded subsets of $\mathcal{M}(\Omega_s)$. We find a suitable candidate in generalizations of optimal transport energies [24]; see also [9, 34].

Despite its various difficulties, stochastic noise also provides new opportunities. For example, unlike the deterministic case, we are given a whole distribution of the measurement data and not only one particular realization. Clearly, the uncertainty in the estimate critically depends on the appropriate choice of measurement locations $\boldsymbol{x} = (x_j)_{j=1,\dots,N_o}$, the overall precision p, and relative precision of each sensor Σ_0^{-1}. Formalizing this connection enables the mathematical program of *optimal sensor placement* or optimal design, i.e. an optimization of the measurement setup to mitigate the influence of the noise before any data is collected in a real experiment. This requires a cheap-to-evaluate *design criterion* which allows to compare the quality of different sensor setups. For linear inverse problems in Hilbert spaces, a popular performance indicator is the mean-squared error of the associated least-squares estimator, which admits a closed form representation through its decomposition into variance and bias; see, e.g. [18]. For nonlinear problems, *locally optimal* sensor placement approaches rely on a linearization of the forward model around a best guess for the unknown parameters; see, e.g. [36]. To the best of our knowledge, optimal sensor placement for nonsmooth estimation problems and for infinite dimensional parameter spaces beyond the Hilbert space setting is uncharted territory.

1.3. Contribution

Taking the mentioned difficulties in the stochastic setting into consideration, we are led to the analysis of the *worst-case mean-squared-error* of the estimator

$$\mathrm{MSE}\left[\bar{\mu}\right] := \mathbb{E}\left[\sup_{\bar{\mu}\in\mathfrak{M}} d_{\mathrm{HK}}\left(\bar{\mu},\mu^\dagger\right)^2\right] = \int_{\mathbb{R}^{N_o}} \sup_{\bar{\mu}\in\mathfrak{M}(\varepsilon)} d_{\mathrm{HK}}\left(\bar{\mu},\mu^\dagger\right)^2 \mathrm{d}\gamma_p(\varepsilon)\,, \quad (1.3)$$

where d_{HK} denotes an extension of the Hellinger–Kantorovich distance introduced in [24] to signed measures (see section 4) and γ_p is the noise distribution $\mathcal{N}(0,\Sigma)$. We point out that, in comparison to linear inverse problems in Hilbert space, $\mathrm{MSE}[\bar{\mu}]$ does not admit a closed form expression and its computation requires both, a sampling of the expected value, as well as an efficient way to calculate the Hellinger–Kantorovich distance. This prevents its direct use in the context of optimal sensor placement for sparse inverse problems.

To enable efficient sensor design, we first need to select an appropriate regularization parameter, depending on the noise level. Here, we focus on the *a priori* choice rule of $\beta(p) = \beta_0/\sqrt{p}$ for some tunable $\beta_0 > 0$, that only takes into account the overall precision of the sensor.

For this choice, we provide the following upper bound:

$$\mathrm{MSE}\left[\bar{\mu}\right] \leqslant \frac{8}{p}\,\psi_{\beta_0}\left(\boldsymbol{x}, \Sigma_0\right) + \bar{c}\exp\left(-\bar{\lambda}\beta_0^2\right), \tag{1.4}$$

where the constant $\psi_{\beta_0}(\boldsymbol{x}, \Sigma_0)$ (further detailed below) explicitly depends on the locations and relative precisions while the constants $\bar{c}$ and $\bar{\lambda}$ depend on the problem setup (the kernel and domain, some basic bounds on the ground truth), a non-degeneracy parameter of the dual certificate $\eta^\dagger$ (further detailed below), and quantities that can be bounded by $\psi_{\beta_0}(\boldsymbol{x}, \Sigma_0)$, but do not depend on p or β_0 for $p \geqslant \bar{p} > 0$ and $\beta_0 \geqslant \bar{\beta}_0 > 0$; see theorem 6.1. Thus, under these basic assumptions and by choosing β_0 large enough, the second term in (1.4) becomes negligible and the first term dominates and closely predicts the mean-squared error. This behavior is confirmed by numerical examples; see section 7.

To further illustrate the meaning of the constant $\psi_{\beta_0}(\boldsymbol{x}, \Sigma_0)$, let us denote by $\boldsymbol{q} = (q_1; \ldots; q_{N_s})$ and $\boldsymbol{y} = (y_1; \ldots; y_{N_s})$ the vectors of coefficients and positions of sources, respectively. Additionally, we collect all the parameters of a given finite source μ in the vector $\boldsymbol{m} = (\boldsymbol{q}; \boldsymbol{y})$, and introduce the *parameter-to-observation* map $G(\boldsymbol{m}) = K\mu$, as well as its Jacobian $G'(\boldsymbol{m}^\dagger)$ evaluated at the parameters of the ground truth. Associated to this, we denote the Fisher information matrix $\mathcal{I}_0$ by

$$\mathcal{I}_0 := G'\left(\boldsymbol{m}^\dagger\right)^\top \Sigma_0^{-1} G'\left(\boldsymbol{m}^\dagger\right). \tag{1.5}$$

Then the constant in the estimate above is computed as

$$\psi_{\beta_0}\left(\boldsymbol{x}, \Sigma_0\right) = \mathrm{tr}\left(W_\dagger \mathcal{I}_0^{-1}\right) + \beta_0^2 \left\|\mathcal{I}_0^{-1}\left(\boldsymbol{\rho}; \boldsymbol{0}\right)\right\|_{W_\dagger}^2,$$

with the sign vector $\boldsymbol{\rho} = \mathrm{sign}\,\boldsymbol{q}^\dagger$ and a weighted Euclidean norm $\|\cdot\|_{W_\dagger}$ which is induced by a positive definite matrix $W_\dagger$ connected to the ground truth $\boldsymbol{m}^\dagger$. This clarifies how the multiplicative constant in the estimate explicitly depends on the measurement setup and we note that it closely resembles the 'classical' A-optimal design criterion; see [18]. Together with the estimate (1.4), and the smallness of the second term, this suggests that $\psi_{\beta_0}(\boldsymbol{x}, \Sigma_0)$ is a suitable criterion to quantify the quality of a given design in terms of the MSE (1.3).

Concerning the smallness of the second term, we note that the constant $\bar{\lambda}$ also depends on a non-degeneracy constant $\theta > 0$, which is a further tightening of the assumption on the dual certificate. This non-degenerate source condition on $\mu^\dagger$ requires the associated minimal norm dual certificate $\eta^\dagger$ to fulfill

$$\left|\eta^\dagger(y)\right| \leqslant 1 - \theta \min\left\{\theta, \min_{n=1,\ldots,N_s} \left\|\sqrt{|q_n^\dagger|}\left(y - y_n^\dagger\right)\right\|_2^2\right\} \quad \text{for all } y \in \Omega_s \tag{1.6}$$

for some $\theta > 0$. This condition has been employed in many previous works, and is known to uniformly hold for several settings under general assumptions on the measurement and a separation of the condition of the sources; see, e.g. [33] and the references therein.

The proof of the main result relies on a splitting of the set of measurement errors $\mathbb{R}^{N_o}$ into a set of 'nice' events $\mathcal{A}_{\mathrm{nice}}$ as well as an estimate of the probability of its complement $\mathbb{R}^{N_o} \setminus \mathcal{A}_{\mathrm{nice}}$, related to the second term in (1.4). On $\mathcal{A}_{\mathrm{nice}}$, there is a unique optimal parameter $\bar{\boldsymbol{m}} = (\bar{\boldsymbol{q}}, \bar{\boldsymbol{y}})$ with the correct number of sources that parametrizes $\bar{\mu}$. Then, the distance between the reconstruction and the ground truth in the Hellinger–Kantorovich distance can

be estimated by a weighted Euclidean distance of the parameters. Those can be further estimated with a linearization of G, which leads to (1.4) after explicitly computing the expectation. This estimate is specific to the choice of d_{HK} and relies on its interpretation as an unbalanced Wasserstein-2 distance. While similar estimates can be derived for other popular metrics such as the Kantorovich-Rubinstein distance (related to the Wasserstein-1 distance; see appendix C) this would introduce additional constants in the first term of (1.4) stemming from an inverse inequality of discrete ℓ_1 and weighted ℓ_2 norms. Thus, the first term in the modified estimate would overestimate the true error by a potentially substantial factor. In contrast, the first term in (1.4) is sharp in the sense that the convenient factor of 8 can, *mutatis mutandis*, be replaced by any $c > 1$, at the cost of increasing the constant in the second term.

1.4. Further related work

1.4.1. Sparse minimization problems beyond inverse problems. Minimization problems over spaces of measures represent a sensible extension of ℓ_1-regularization towards decision variables on continuous domains. Consequently, problems of the form ($\mathcal{P}_{\beta,\varepsilon}$) naturally appear in a variety of different applications, detached from inverse problems. We point out, e.g. optimal actuator placement, optimal sensor placement [26], as well as the training of shallow neural networks [2]. Non-degeneracy conditions similar to (1.6) play a crucial role in this context and form the basis for an in-depth (numerical) analysis of the problem, e.g. concerning the derivation of fast converging solution methods, [9, 14, 31], or finite element error estimates [22].

1.4.2. Inverse problems with random noise. Frequentist approaches to inverse problems have been studied previously in, e.g. [16, 37]. These works focus on the 'lifting' of deterministic regularization methods as well as of their consistency properties and convergence rates to the random noise setting. This only relies on minimal assumptions on the inverse problem, e.g. classical source conditions, and thus covers a wide class of settings. Similar to the present work, an important role is played by a splitting of the possible events into a set on which the deterministic theory holds and its small complement. However, we want to stress that the proof of the main estimate in (1.4) is problem-taylored and relies on exploiting specific structural properties of inverse problems in spaces of measures. Moreover, our main goal is *not* the consistency analysis of an estimator but the derivation of a useful and mathematically sound design criterion for sparse inverse problems.

1.5. Organization of the paper

The paper is organized as follows: In section 3, we recall some properties of the minimum norm problem ($\mathcal{P}_{\beta,\varepsilon}$) and the Tikhonov regularized problem ($\mathcal{P}_{\beta,\varepsilon}$) as well as its solutions. In section 4, we define the Hellinger–Kantorovich distance and investigate its properties. Section 5 is devoted to study the linearized estimate $\delta\widehat{\boldsymbol{m}}$. Using these results, we then investigate sparse inverse problems with random noise in section 6 and provide a sharp upper bound for $\mathrm{MSE}[\bar{\mu}]$ in section 6.2. Finally, in section 7 we present some numerical examples to verify our theory.

2. Notation and preliminaries

Before going into the main part of the paper, we introduce the basic notation used throughout the paper and gather preliminary assumptions concerning the considered integral kernels as well as pertinent facts on Radon measures.

2.1. Notation

Throughout the paper, c_i, C_i, $i = 1,2,\ldots$ denote generic constants that may vary from line to line. By $C = C(a,b,\ldots)$, we indicate that C depends on $a,b,\ldots$. We denote by $\Omega_s \subset \mathbb{R}^d$ and $\Omega_o \subset \mathbb{R}^{d_o}$ the compact location and observation set, where $d_o, d \geqslant 1$ and Ω_s has a nonempty interior. A vector in X^m for a set X and $m > 1$, will be written in bold face, for instance $\boldsymbol{y} = (y_1;\ldots;y_{N_s}) \in \Omega_s^{N_s}$, $\boldsymbol{q} = (q_1;\ldots;q_{N_s}) \in \mathbb{R}^{N_s}$ and $\boldsymbol{x} = (x_1;\ldots;x_{N_o}) \in \Omega_o^{N_o}$ are vectors of coefficients, positions of sources and positions of observations, respectively, where the formal definitions are introduced in the sequel. We write $(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n)$ and $(\boldsymbol{a}_1;\ldots;\boldsymbol{a}_n)$ to stack vectors $\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n$ horizontally and vertically, respectively. We write $\|\cdot\|_p$ for the usual ℓ^p-norm on $\mathbb{R}^m$. For a vector $x \in \mathbb{R}^m$ and a positively defined matrix $W \in \mathbb{R}^{m\times m}$, we define the weighted W-norm of x as $\|x\|_W := \left\|W^{1/2}x\right\|_2$. The closed ball in this weighted norm is denoted by $B_W(x,r) := \{x' \in \mathbb{R}^m : \|x'-x\|_W \leqslant r\}$. For a linear map $A : X \to Y$, the operator norm of A is given by $\|A\|_{X\to Y} = \sup_{\|x\|_X \leqslant 1} \|Ax\|_Y$. Similarly, any bilinear map $A : X_1 \times X_2 \to Y$ has a natural operator norm $\|A\|_{X_1\times X_2\to Y} := \sup_{\|x_1\|_{X_1}\leqslant 1, \|x_2\|_{X_2}\leqslant 1} \|A(x_1,x_2)\|_Y$.

Furthermore, let $k : \Omega_o \times \Omega_s \to \mathbb{R}$ be a real-valued kernel. We introduce the following notations which turn k into vector-valued kernels: $k[\boldsymbol{x}](y) = k[\boldsymbol{x},y]$ is a column vector with

$$k[\boldsymbol{x},y] := (k(x_1,y);\ldots;k(x_{N_o},y)), \quad \boldsymbol{x} = (x_1;\ldots;x_{N_o}) \in \Omega_o^{N_o}, \quad y \in \Omega_s, \tag{2.1}$$

while $k[x,\boldsymbol{y}]$ is a row vector with

$$k[x,\boldsymbol{y}] := (k(x,y_1),\ldots,k(x,y_{N_s})), \quad x \in \Omega_o, \quad \boldsymbol{y} = (y_1;\ldots;y_{N_s}) \in \Omega_s^{N_s}. \tag{2.2}$$

Similarly, we also have the matrix $k[\boldsymbol{x},\boldsymbol{y}]$ defined as

$$k[\boldsymbol{x},\boldsymbol{y}] := (k(x_1,\boldsymbol{y});\ldots;k(x_{N_o},\boldsymbol{y})). \tag{2.3}$$

When $k = k(x,\cdot)$ is a smooth function in variable y, we consider the r^{th}-derivative of k the tensor of partial derivatives is y by $\nabla^r_{y\ldots y}k(x,y)$. In particular, $\nabla_y k(x,y)$ and $\nabla^2_{yy}k(x,y)$ are the gradient and Hessian of k (with respect to variable y,) respectively. We note that $\nabla_y k\colon \Omega_o \times \Omega_s \to \mathbb{R}^{N_s}$ is a vector valued kernel and thus we define $\nabla_y^\top k[x,\boldsymbol{y}]$ as a matrix defined by

$$\nabla_y^\top k[x,\boldsymbol{y}] = \left(\nabla_y k(x,y_1)^\top, \nabla_y k(x,y_2)^\top, \ldots, \nabla_y k(x,y_{N_s})^\top\right). \tag{2.4}$$

Similarly, $\nabla_y^\top k[\boldsymbol{x},\boldsymbol{y}]$ is a block matrix defined by

$$\nabla_y^\top k[\boldsymbol{x},\boldsymbol{y}] = \left(\nabla_y^\top k[x_1,\boldsymbol{y}],\ldots,\nabla_y^\top k[x_{N_o},\boldsymbol{y}]\right). \tag{2.5}$$

Throughout the paper, by a slight abuse of notation, we denote by ε a variable deterministic noise, a random variable, or its realization, which will be clear from the context. By γ_p we denote the density of a multivariate Gaussian random variable with expectation zero and covariance Σ. Further notation, specific to the present manuscript, will be introduced at first appearance. For quicker reference, a notation table can be found in appendix D.

2.2. Preliminaries

We also recall some basic facts and assumptions for inverse source location.

2.2.1. Integral kernels. Throughout the paper, we assume that the kernel is sufficiently regular:

(A1) The kernel $k \in \mathcal{C}(\Omega_o \times \Omega_s)$ is three-times differentiable in the variable y. For abbreviation, we further set

$$C_k := \sup_{x\in\Omega_o, y\in\Omega_s} |k(x,y)|, \qquad C_k' := \sup_{x\in\Omega_o, y\in\Omega_s} \|\nabla_y k(x,y)\|_2,$$
$$C_k'' := \sup_{x\in\Omega_o, y\in\Omega_s} \left\|\nabla_{yy}^2 k(x,y)\right\|_{2\to 2}, \quad C_k''' := \sup_{x\in\Omega_o, y\in\Omega_s} \left\|\nabla_{yyy}^3 k(x,y)\right\|_{2\times 2\to 2}.$$

By means of the kernel k, we introduce the weak* continuous *source-to-measurements* operator $K\colon \mathcal{M}(\Omega_s) \to \mathbb{R}^{N_o}$ with

$$K\mu = \left(\int_{\Omega_s} k(x_1,y)\,\mathrm{d}\mu(y); \ldots; \int_{\Omega_s} k(x_{N_o},y)\,\mathrm{d}\mu(y) \right). \tag{2.6}$$

Moreover, consider the operator $K^*\colon \mathbb{R}^{N_o} \to \mathcal{C}^2(\Omega_s)$ given by

$$[K^* z](y) = \sum_{j=1}^{N_o} z_j k(x_j, y) \quad \text{for all} \quad z \in \mathbb{R}^{N_o}. \tag{2.7}$$

Then K^* is linear and continuous and there holds

$$\int_\Omega [K^* z](y)\ \mathrm{d}\mu(y) = z^\top [K\mu] \quad \text{for all} \quad \mu \in \mathcal{M}(\Omega_s),\ z \in \mathbb{R}^{N_o}.$$

2.2.2. Space of Radon measures. We recall some properties of Radon measures. Let $\Omega \subset \mathbb{R}^d$, $d \geqslant 1$ be a compact set. We define the space of Radon measures $\mathcal{M}(\Omega)$ as the topological dual of the space $\mathcal{C}(\Omega)$ of continuous functions on Ω endowed with the supremum norm. It is then a Banach space equipped with the dual norm

$$\|\mu\|_{\mathcal{M}(\Omega)} := \sup\left\{ \int_\Omega f \mathrm{d}\mu : f \in \mathcal{C}(\Omega), \|f\|_{\mathcal{C}(\Omega)} \leqslant 1 \right\}.$$

Weak* convergence of a sequence in $\mathcal{M}(\Omega)$ will be denoted by '$\rightharpoonup^*$'. More specifically, we have

$$\mu_n \rightharpoonup^* \mu \quad \text{if and only if} \quad \int_\Omega f \mathrm{d}\mu_n \to \int_\Omega f \mathrm{d}\mu \quad \text{for all} \quad f \in \mathcal{C}(\Omega).$$

Next, by the definition of the total variation norm, its subdifferential is defined by

$$\partial \|\mu\|_{\mathcal{M}(\Omega_s)} := \left\{ \eta \in \mathcal{C}(\Omega_s) : |\eta(y)| \leqslant 1, \forall y \in \Omega_s \text{ and } \int_{\Omega_s} \eta \mathrm{d}\mu = \|\mu\|_{\mathcal{M}(\Omega_s)} \right\},$$

see for instance [11]. In particular, for a discrete measure $\mu = \sum_{n=1}^{N} q_n \delta_{y_n}$ one has

$$\partial \|\mu\|_{\mathcal{M}(\Omega_s)} = \{\eta \in \mathcal{C}(\Omega_s) : |\eta(y)| \leqslant 1, \forall y \in \Omega_s \text{ and } \eta(y_n) = \operatorname{sign}(q_n), \forall n = 1, \dots, N\}.$$

Finally, by $\mathcal{M}^+(\Omega)$ we refer to the set of positive Radon measures on Ω.

3. Sparse inverse problems with deterministic noise

Our interest lies in the stable recovery of a sparse ground truth measure

$$\mu^\dagger = \sum_{n=1}^{N^\dagger} q_n^\dagger \delta_{y_n^\dagger} \quad \text{for some} \quad q_n^\dagger \in \mathbb{R},$$

by solving the Tikhonov regularization ($\mathcal{P}_{\beta,\varepsilon}$) associated to the inverse problem $z^d = K\mu$ given noisy data z^d. In this preliminary section, we give some meaningful examples of this abstract setting and briefly recap the key concepts and results in the case of additive deterministic noise

$$z^d(\varepsilon) = K\mu^\dagger + \varepsilon \quad \text{for some} \quad \varepsilon \in \mathbb{R}^{N_o}. \tag{3.1}$$

In particular, we clarify the connection between ($\mathcal{P}_{\beta,\varepsilon}$) and ($\mathcal{P}_0$) and recall a first qualitative statement on the asymptotic behavior of solutions to ($\mathcal{P}_{\beta,\varepsilon}$) for a suitable *a priori* regularization parameter choice $\beta = \beta(\varepsilon)$.

3.1. Examples

Sparse inverse problems appear in a variety of interesting applications. In the following, we give some examples which fit into our setting.

Example 3.1. Consider the advection-diffusion equation

$$\partial_t u - \nabla(\boldsymbol{D} \cdot \nabla u) + \nabla \cdot (\kappa u) = 0 \text{ in } (0,T) \times \mathbb{R}^d, \tag{3.2}$$

together with the initial value $u(0,\cdot) = \mu$. The boundary condition is given by $u \to 0$ as $x \to \infty$. This equation describes the rate of change of the concentration of the contaminant $u(t,x)$. For simplicity, we consider a two-dimensional medium, and both $\kappa = (\kappa_1, \kappa_2)$ and $\boldsymbol{D} = \operatorname{diag}(D_1, D_2)$ are independent of x. Here the solution to (3.2) is given by

$$u(t,x) = \int_{\mathbb{R}^2} G(x-y,t)\, \mathrm{d}\mu(y)$$

where $G(x,t)$ is the Green's function of the advection-diffusion equation, which is given by

$$G(x,t) = \frac{1}{4\pi\sqrt{D_1 D_2 t}} \exp\left(-\|x - \kappa t\|^2_{\boldsymbol{D}^{-1}} / (4t)\right).$$

Here, if one seeks to identify the initial value μ from finite number of measurements at time $T_o > 0$ in the observation set $\Omega_o \subset \mathbb{R}^2$, the kernel is given by $k(x,y) = G(x-y, T_o)$.

Example 3.2. Consider the advection-diffusion equation on a bounded smooth domain Ω, together with the Dirichlet boundary conditions $u|_{(0,T)\times\partial\Omega} = 0$, then there exists a kernel $G(x,y,t)$ such that

$$u(t,x) = \int_\Omega G(x,y,t)\,\mathrm{d}\mu(y),$$

see, e.g. [15]. In this case, for observations at time T_o we choose $k = G(\cdot,\cdot,T_o)$. For $\Omega_o \subset \Omega$ (i.e. no observation near the boundary), the regularity requirements on $\partial\Omega$ are not necessary since one can employ interior regularity arguments; see, e.g. [17].

3.2. Tihkonov regularization of sparse inverse problems

In this section, we briefly summarize some preliminary results concerning the regularized problem ($\mathcal{P}_{\beta,\varepsilon}$) as well as its solution set. We start by discussing its well-posedness.

Proposition 3.3. *Problem ($\mathcal{P}_{\beta,\varepsilon}$) admits a solution $\bar\mu$. Furthermore, any solution $\bar\mu$ to ($\mathcal{P}_{\beta,\varepsilon}$) satisfies $\|\bar\mu\|_{\mathcal{M}(\Omega_s)} \leqslant \|\varepsilon\|^2_{\Sigma_0^{-1}}/(2\beta) + \|\mu^\dagger\|_{\mathcal{M}(\Omega_s)}$ and the solution set*

$$\mathfrak{M}(\varepsilon) = \arg\min(\mathcal{P}_{\beta,\varepsilon})$$

is weak* compact.

Proof. Existence of a minimizer of ($\mathcal{P}_{\beta,\varepsilon}$) is guaranteed by [4, proposition 3.1] noticing that the forward operator $K : \mathcal{M}(\Omega_s) \to \mathbb{R}^{N_o}$ of ($\mathcal{P}_{\beta,\varepsilon}$) is weak*-to-strong continuous. For the upper bound we use the optimality of $\bar\mu$ compared to $\mu^\dagger$ as well as the definition of $z^d(\varepsilon)$ to get

$$\beta\|\bar\mu\|_{\mathcal{M}(\Omega_s)} \leqslant \frac{1}{2}\left\|K\bar\mu - z^d\right\|^2_{\Sigma_0^{-1}} + \beta\|\bar\mu\|_{\mathcal{M}(\Omega_s)} \leqslant \frac{1}{2}\|\varepsilon\|^2_{\Sigma_0^{-1}} + \beta\left\|\mu^\dagger\right\|_{\mathcal{M}(\Omega_s)}.$$

Moreover, $\mathfrak{M}(\varepsilon)$ is weak* closed since the objective functional in ($\mathcal{P}_{\beta,\varepsilon}$) is weak* lower semi-continuous. Combining both observations, we conclude the weak* compactness of $\mathfrak{M}(\varepsilon)$. □

In particular, note that $\mathfrak{M}(\varepsilon)$ is, in general, not a singleton due to the lack of strict convexity in ($\mathcal{P}_{\beta,\varepsilon}$). Moreover, we recall that the inverse problem was introduced as a lifting of the nonconvex and combinatorial integral equation (1.1). From the same perspective, ($\mathcal{P}_{\beta,\varepsilon}$) can be interpreted as a convex relaxation of the parametrized problem

$$\inf_{\substack{\boldsymbol{y}\in\Omega_s^N,\ \boldsymbol{q}\in\mathbb{R}^N,\\ N\in\mathbb{N}}} \left[\frac{1}{2}\left\|k[\boldsymbol{x},\boldsymbol{y}]\boldsymbol{q} - z^d\right\|^2_{\Sigma_0^{-1}} + \beta\|\boldsymbol{q}\|_1\right], \tag{3.3}$$

In the following proposition, we show that this relaxation is exact, i.e. there exists at least one solution to (3.3) and its minimizers parametrize sparse solutions to ($\mathcal{P}_{\beta,\varepsilon}$).

Proposition 3.4. *There holds* min *($\mathcal{P}_{\beta,\varepsilon}$)* $=$ inf *(3.3). For a triple $(\bar N,\bar{\boldsymbol{y}},\bar{\boldsymbol{q}})$ with $\bar y_i \neq \bar y_j$, $i \neq j$, the following statements are equivalent:*

- *The triple $(\bar N,\bar{\boldsymbol{y}},\bar{\boldsymbol{q}})$ is a solution of (3.3).*
- *The parametrized measure $\bar\mu = \sum_{n=1}^{\bar N}\bar q_n\delta_{\bar y_n}$ is a solution of ($\mathcal{P}_{\beta,\varepsilon}$).*

Moreover, ($\mathcal{P}_{\beta,\varepsilon}$) admits at least one solution of this form with $\bar N \leqslant N_o$.

Proof. Given $(N,\boldsymbol{y},\boldsymbol{q})$ with $y_i \neq y_j$, $i \neq j$, note that the sparse measure

$$\mu(\boldsymbol{y},\boldsymbol{q}) = \sum_{n=1}^{N} q_n \delta_{y_n} \quad \text{satisfies} \quad K\mu(\boldsymbol{y},\boldsymbol{q}) = k[\boldsymbol{x},\boldsymbol{y}]\boldsymbol{q},\ \|\mu(\boldsymbol{y},\boldsymbol{q})\|_{\mathcal{M}(\Omega_s)} = \|\boldsymbol{q}\|_1 .$$

Hence, one readily verifies $\min(\mathcal{P}_{\beta,\varepsilon}) = \inf(3.3)$ as well as the claimed equivalence due to the weak* density of the set of sparse measures in $\mathcal{M}(\Omega_s)$ and since the objective functional in $(\mathcal{P}_{\beta,\varepsilon})$ is weakly* lower semicontinuous. The existence of a sparse solution to $(\mathcal{P}_{\beta,\varepsilon})$ follows similarly to [30, theorem 3.7]. □

The equivalence between both of these problems will play a significant role in our subsequent analysis. Additional insight on the structure of solutions to $(\mathcal{P}_{\beta,\varepsilon})$ can be gained through the study of its first order necessary and sufficient optimality conditions. Since our interest lies in sparse solutions, we restrict the following proposition to this particular case.

Proposition 3.5. *A measure $\bar{\mu} = \sum_{n=1}^{\bar{N}} \bar{q}_n \delta_{\bar{y}_n}$ is a solution of $(\mathcal{P}_{\beta,\varepsilon})$ if and only if*

$$|\bar{\eta}(y)| \leqslant 1 \text{ for all } y \in \Omega_s, \quad \bar{\eta}(\bar{y}_n) = \operatorname{sign}(\bar{q}_n), \quad \forall n = 1,\dots,\bar{N},$$

where

$$\bar{\eta} = -K^*\Sigma_0^{-1}\left(K\bar{\mu} - z^d\right)/\beta = K^*\Sigma_0^{-1}\left(z^d - k[\boldsymbol{x},\bar{\boldsymbol{y}}]\bar{\boldsymbol{q}}\right)/\beta. \tag{3.4}$$

Note that $\bar{\eta}$ is independent of the particular choice of the solution to $(\mathcal{P}_{\beta,\varepsilon})$. We will refer to it as the dual certificate associated to $(\mathcal{P}_{\beta,\varepsilon})$ in the following. Finally, we give a connection between $(\mathcal{P}_{\beta,\varepsilon})$ and the minimum norm problem $(\mathcal{P}_0)$ in the vanishing noise limit. The following general convergence property follows directly from [19].

Proposition 3.6. *Assume that $\beta = \beta(\varepsilon)$ is chosen such that*

$$\beta \to 0 \text{ and } \frac{\|\varepsilon\|^2_{\Sigma_0^{-1}}}{\beta} \to 0 \text{ as } \|\varepsilon\|_{\Sigma_0^{-1}} \to 0.$$

Then solutions to $(\mathcal{P}_{\beta,\varepsilon})$ subsequentially converge weakly-* towards solutions of $(\mathcal{P}_0)$.

3.3. Radon minimum norm problems

Following proposition 3.6, guaranteed recovery of the ground truth measure requires that $\mu^\dagger$ is identifiable, i.e. the unique solution of $(\mathcal{P}_0)$. In this section, we briefly summarize some key concepts regarding $(\mathcal{P}_0)$ and state sufficient assumptions for the latter. For this purpose, introduce the associated Fenchel dual problem

$$\min_{\zeta \in \mathbb{R}^{N_o}} \left[-\langle \mu^\dagger, K^*\Sigma_0^{-1}\zeta \rangle + \mathbb{I}_{\|K^*\Sigma_0^{-1}\zeta\|_{C(\Omega_s)} \leqslant 1} \right]. \tag{3.5}$$

as well as the minimal-norm dual certificate

$$\eta^\dagger := K^*\Sigma_0^{-1}\zeta^\dagger \in \mathcal{C}^2(\Omega_s) \quad \text{where} \quad \zeta^\dagger = \operatorname*{arg\,min}_{\zeta \in \mathbb{R}^{N_o}} \{ \|\zeta\|_2 \,:\, \zeta \in \arg\min(3.5) \}. \tag{3.6}$$

Note that the existence of $\zeta^\dagger$, and therefore the minimum-norm dual certificate $\eta^\dagger$, is guaranteed in this setting following [30, proposition A.2] as well as due to $K^*\colon \mathbb{R}^{N_o} \to \mathcal{C}^2(\Omega_s)$. Moreover, by standard results from convex analysis, a given $\mu \in \mathcal{M}(\Omega_s)$ is a solution to $(\mathcal{P}_0)$ if and only if $\eta^\dagger \in \partial\|\mu\|_{\mathcal{M}(\Omega_s)}$. The following assumptions on $\mu^\dagger$ and $\eta^\dagger$ are made throughout the paper:

(A2) *Structure of* $\mu^\dagger$: We assume that there holds

$$\mu^\dagger = \sum_{n=1}^{N_s^\dagger} q_n^\dagger \delta_{y_n^\dagger} \quad \text{where} \quad q_n^\dagger \neq 0,\ y_n^\dagger \in \operatorname{int}(\Omega_s) \quad \text{for all} \quad n = 1,\dots,N_s^\dagger.$$

(A3) *Source condition*: We assume that the minimum-norm dual certificate $\eta^\dagger$ satisfies

$$|\eta^\dagger(y)| \leqslant 1 \quad \text{for all} \quad y \in \Omega_s \quad \text{and} \quad \eta^\dagger\left(y_n^\dagger\right) = \operatorname{sign}\left(q_n^\dagger\right) \quad \text{for all} \quad n = 1,\dots,N_s.$$

(A4) *Strengthened source condition*: We assume that

$$|\eta^\dagger(y)| < 1 \quad \text{for all} \quad y \in \Omega_s \setminus \left\{y_n^\dagger\right\}_{n=1}^{N_s^\dagger}$$

and the operator $K_{|\operatorname{supp}\mu^\dagger} := k[\boldsymbol{x}, \boldsymbol{y}^\dagger]$ is injective.

Here, assumption A3 is equivalent to $\eta^\dagger \in \partial\left\|\mu^\dagger\right\|_{\mathcal{M}(\Omega_s)}$, i.e. $\mu^\dagger$ is indeed a solution to $(\mathcal{P}_0)$, whereas assumptions A2 and A3 imply its uniqueness. While assumption A4 seems very strong at first glance, it can be explicitly verified in some settings (see, e.g. [6]) and is often numerically observed in practice. According to [11, proposition 5] we have the following:

Proposition 3.7. *Let assumptions A2–A4 hold. Then* $\mu^\dagger$ *is the unique solution of* $(\mathcal{P}_0)$.

As a consequence, proposition 3.6 implies $\bar{\mu} \rightharpoonup^* \mu^\dagger$. Moreover, according to [11, Proposition 1], the dual certificates $\bar{\eta}$ associated to $(\mathcal{P}_{\beta,\varepsilon})$ approximate the minimal norm dual certificate $\eta^\dagger$ in a suitable sense. Taking into account assumption A3 as well as proposition 3.5, we thus conclude that the reconstruction of $\mu^\dagger$ from (3.3) is governed by the convergence of the global extrema of $\bar{\eta}$ towards those of $\eta^\dagger$. However, in order to capitalize on this observation in our analysis, we need to compute a closed form expression for $\eta^\dagger$. In general, this is intractable due to the global constraint $|\eta^\dagger(z)| \leqslant 1$, $z \in \Omega_s$. As a remedy, the authors of [11] introduce a simpler proxy replacing this constraint by finitely many linear ones noting that

$$\nabla\eta^\dagger\left(y_n^\dagger\right) = 0, \quad \eta^\dagger\left(y_n^\dagger\right) = \operatorname{sign}\left(q_n^\dagger\right) \quad \text{for all} \quad n = 1,\dots,N_s^\dagger.$$

The computation of the associated vanishing derivative pre-certificate $\eta_{\mathrm{PC}} := K^*\Sigma_0^{-1}\zeta_{\mathrm{PC}} \in \mathcal{C}^2(\Omega_s)$ where

$$\zeta_{\mathrm{PC}} = \operatorname*{arg\,min}_{\zeta \in \mathbb{R}^{N_o}} \left\{ \|\zeta\|_2 : \nabla\eta_{\mathrm{PC}}\left(y_i^\dagger\right) = 0, \quad \eta_{\mathrm{PC}}\left(y_n^\dagger\right) = \operatorname{sign}\left(q_n^\dagger\right) \quad \text{for all} \quad n = 1,\dots,N_s^\dagger \right\} \tag{3.7}$$

only requires the solution of a linear systems of equations and coincides with $\eta^\dagger$ under appropriate conditions, see [11, proposition 7]. Finally, in order to derive quantitative statements on the reconstruction error between $\bar{\mu}$ and $\mu^\dagger$, we require the non-degeneracy of the minimal

norm dual certificate of $\mu^\dagger$ in the sense of [11]. Since we aim to use (1.4) in the context of optimal sensor placement, that is, we need to track the dependence of the involved constants on the measurement setting, we utilize the following quantitative definition; see [33].

Definition 3.8. We say that $\eta \in \mathcal{C}^2(\Omega_s)$ is $\theta-$non-degenerate or $\theta-$admissible for the sparse measure $\mu = \sum_{n=1}^{N_s} q_n \delta_{y_n}$ and $\theta \in (0,1]$ if there holds

$$|\eta(y)| \leqslant 1 - \theta \min\left\{\theta, \min_{n=1,\dots,N_s} \left\|w_n^\dagger (y - y_n)\right\|_2^2\right\}, \quad \eta(y_n) = \operatorname{sign}(q_n) \qquad \text{for all} \quad y \in \Omega_s \tag{3.8}$$

and weights $w_n^\dagger = \sqrt{|q_n^\dagger|}$.

Due to the regularity of η one readily verifies that (3.8) is equivalent to

$$-\operatorname{sign}\eta(y_n)\, \nabla^2 \eta(y_n) \geqslant 2\theta |w_n^\dagger|^2 \operatorname{Id} \quad \text{for every} \quad n = 1, 2, \dots, N_s, \tag{3.9}$$

as well as

$$|\eta(y)| \leqslant 1 - \theta^2, \quad \text{for all} \quad y \in \Omega_s \setminus \bigcup_{n=1,\dots,N_s} B_{w_n^\dagger}\left(y_n, \sqrt{\theta}\right). \tag{3.10}$$

4. Distances on spaces of measures

In order to quantitatively study the reconstruction error of estimators of the source $\mu^\dagger$, we introduce a distance function on $\mathcal{M}(\Omega_s)$ which measures the error between the estimated source measure $\widehat{\mu}$ and the reference measure $\mu^\dagger$. An obvious choice of distance would be the total variation norm on $\mathcal{M}(\Omega_s)$, however it is not suitable for quantifying the reconstruction error. In fact, evaluating $d_{\mathrm{TV}}(\mu_1, \mu_2) = \|\mu_1 - \mu_2\|_{\mathcal{M}(\Omega_s)}$ for sparse measures $\mu_1, \mu_2 \in \mathcal{M}(\Omega_s)$ is simple by noting that

$$d_{\mathrm{TV}}(q_1 \delta_{y_1}, q_2 \delta_{y_1}) = |q_1 - q_2|,$$

but for $y_1 \neq y_2$, one has

$$d_{\mathrm{TV}}(q_1 \delta_{y_1}, q_2 \delta_{y_2}) = |q_1| + |q_2|,$$

that is, d_{TV} does not quantify the reconstruction error of the source positions, and small perturbations of the source points lead to a constant error in the metric. Hence, in general one cannot rely on TV distance to evaluate the quality of the reconstruction. In the following, we consider an extension of the Hellinger–Kantorovich (H-K) metric [24] to signed measures, which possesses certain properties that will be discussed below. The construction of the H-K distance is more involved than another often used candidate, namely the Kantorovich–Rubinstein (K-R) distance (see, e.g. [21, 28]) or flat metric, which is directly obtained as a dual norm of a space of Lipschitz functions (see appendix C). It induces the same topology of weak* convergence, and is bounded by the H-K metric [24]. Since our estimates are going to be asymptotically sharp in H-K, but only an upper bound in K-R, we focus on H-K in the following.

The Hellinger–Kantorovich metric [24] is a generalization of the Wasserstein-2 distance (see, e.g. [27]) for measures which are not necessarily of the same norm. We first assume the

case of positive measures $\mu_1, \mu_2 \geqslant 0$ and define the H-K metric in terms of the Wasserstein-2 metric as:

$$d_{\mathrm{HK}}(\mu_1,\mu_2)^2 := \inf\left\{ W_2(\widetilde{\mu}_1,\widetilde{\mu}_2) \;\middle|\; \widetilde{\mu}_1,\widetilde{\mu}_2 \in \mathcal{P}_2\left(\mathbb{R}^+ \times \Omega_s\right) : h_2(\widetilde{\mu}_1) = \mu_1, h_2(\widetilde{\mu}_2) = \mu_2 \right\}.$$

Here, $\mathcal{P}_2(\mathbb{R}^+ \times \Omega_s)$ are the probability measures of with finite second moment on $\mathbb{R}^+ \times \Omega_s$, the two-homogeneous marginal is

$$h_2(\widetilde{\mu}) = \int_{\mathbb{R}^+} r^2 \mathrm{d}\widetilde{\mu}(r,\cdot) \in \mathcal{M}(\Omega_s),$$

and $\mathbb{R}^+ \times \Omega_s$ is endowed with a conic metric

$$d_{\mathrm{cone}}\left((r_1,y_1),(r_2,y_2)\right)^2 := \left(\sqrt{r_1} - \sqrt{r_2}\right)^2 + 4\sqrt{r_1 r_2}\sin_+^2\left(\|y_1 - y_2\|_2/2\right), \quad (4.1)$$

where $\sin_+(z) := \sin(\min\{z, \pi/2\})$. For a detailed study of this metric and its properties as well as equivalent formulations in terms of Entropy-Transport problems we refer to [24].

For signed measures, we note that for any distance based on a norm (such as the TV or K-R distance) one observes that

$$d(\mu_1,\mu_2) = \left\|\left(\mu_1^+ + \mu_2^-\right) - \left(\mu_2^+ + \mu_1^-\right)\right\| = d\left(\mu_1^+ + \mu_2^-, \mu_2^+ + \mu_1^-\right), \quad (4.2)$$

by using the Jordan decomposition $\mu_i = \mu_i^+ - \mu_i^-$. Motivated by (4.2), we define

$$d_{\mathrm{HK}}(\mu_1,\mu_2) := d_{\mathrm{HK}}\left(\mu_1^+ + \mu_2^-, \mu_2^+ + \mu_1^-\right), \quad (4.3)$$

which is indeed a metric on $\mathcal{M}(\Omega_s)$ and fulfills $d_{\mathrm{HK}}(\mu_1,\mu_2) \leqslant d_{\mathrm{HK}}(\mu_1^+,\mu_2^+) + d_{\mathrm{HK}}(\mu_1^-,\mu_2^-)$.

In contrast to the total variation distance, the Hellinger–Kantorovich distance between two Dirac measures $q_1\delta_{y_1}$ and $q_2\delta_{y_2}$ can be computed by

$$d_{\mathrm{HK}}(q_1\delta_{y_1}, q_2\delta_{y_2})^2 = \left(\sqrt{|q_1|} - \sqrt{|q_2|}\right)^2 + 4\sqrt{|q_1||q_2|}\sin_+^2\left(\|y_1 - y_2\|_2/2\right),$$

which is exactly the conic metric given in (4.1). Clearly, it is evidence that for small perturbations of both the source positions and coefficients, the resulting change of the H-K distance remains small. Hence, it is reasonable to employ this type of distance to measure the reconstruction error.

One next advantage of the H-K distance is that it is compatible with the weak* topology on $\mathcal{M}(\Omega_s)$, namely it induced weak* convergence on bounded set in $\mathcal{M}(\Omega_s)$.

Proposition 4.1. *The Hellinger–Kantorovich distance of signed measures defined in (4.3) metrizes weak* convergence of signed measures on bounded set in $\mathcal{M}(\Omega_s)$. More precisely, a bounded sequence $\{\mu_n\}_{n\in\mathbb{N}} \subset \mathcal{M}(\Omega_s)$ converges weakly* to a measure μ if only if $d_{\mathrm{HK}}(\mu_n,\mu) \to 0$ as $n \to \infty$.*

Proof. Assume that $d_{\mathrm{HK}}(\mu_n,\mu) \to 0$ as $n \to \infty$. One can write

$$\mu_n - \mu = \left(\mu_n^+ + \mu^-\right) - \left(\mu^+ + \mu_n^-\right) =: \mu_n^1 - \mu_n^2, \quad (4.4)$$

which implies $d_{\mathrm{HK}}(\mu_n^1,\mu_n^2) = d_{\mathrm{HK}}(\mu_n,\mu) \to 0$. Since $\left\|\mu_n^i\right\|_{\mathcal{M}} \leqslant \|\mu_n^{\pm}\|_{\mathcal{M}} + \|\mu^{\mp}\|_{\mathcal{M}} \leqslant 2M$ and the HK-distance metrizes weak$*$ convergence on bounded sequences of non-negative measures (see [24, theorem 7.15]), we have $\mu_n^1 - \mu_n^2 \rightharpoonup^* 0$, which means that $\mu_n \rightharpoonup^* \mu$.

Conversely, assume that $\mu_n \rightharpoonup^* \mu$. Consider the decomposition (4.4) and suppose that the distance $d_{\mathrm{HK}}(\mu_n, \mu)$ does not converges to zero. Then there exists a subsequence, denoted by the same symbol, such that

$$d_{\mathrm{HK}}\left(\mu_n^1, \mu_n^2\right) = d_{\mathrm{HK}}\left(\mu_n, \mu\right) \geqslant \delta > 0. \tag{4.5}$$

We now use the fact that $\left\|\mu_n^i\right\|_{\mathcal{M}} \leqslant 2M$ to extract a further subsequence (again with the same symbol) such that $\mu_n^i \rightharpoonup^* \widehat{\mu}^i$, which implies $\mu_n - \mu = \mu_n^1 - \mu_n^2 \rightharpoonup^* \widehat{\mu}^1 - \widehat{\mu}^2$. Due to (4.5) and the fact that the HK-distance metrizes weak* convergence on bounded sequences of non-negative measures we have that $\widehat{\mu}^1 \neq \widehat{\mu}^2$ and thus $\mu_n - \mu \rightharpoonup^* \widehat{\mu}^1 - \widehat{\mu}^2 \neq 0$. Thus the subsequence $\{\mu_n\}_{n\in\mathbb{N}}$ does not converge weak* to μ and the original sequence $\{\mu_n\}_{n\in\mathbb{N}}$ can not converge to μ.

□

To evaluate the reconstruction error, the distance between finitely supported measures is needed since the reference measure as well as the reconstructed measure are known to be sparse. In fact, we only need a (sharp) upper bound for the H-K distance, which will be provided for the finitely supported case below in term of a (weighted) ℓ^2-type distance. This is yet another advantage of the H-K distance in comparison to other distances.

Proposition 4.2. *Let μ and $\mu^\dagger$ be finitely supported with the same number N of support points and* $\operatorname{sign} q_n = \operatorname{sign} q_n^\dagger$, *for all $n = 1, \ldots, N$. Then we have*

$$d_{\mathrm{HK}}\left(\mu, \mu^\dagger\right)^2 \leqslant R\left(\boldsymbol{q}, \boldsymbol{q}^\dagger\right) \sum_{n=1}^{N} \left(\frac{|q_n - q_n^\dagger|^2}{4|q_n^\dagger|} + |q_n^\dagger| \left\|y_n - y_n^\dagger\right\|_2^2 \right),$$

where $R(\boldsymbol{q}, \boldsymbol{q}^\dagger) := \max\{ \sqrt{|q_n|/|q_n^\dagger|}, \sqrt{|q_n^\dagger|/|q_n|} \ : \ n = 1, \ldots, N\}$.

Loosely speaking, the H-K distance between two discrete measures μ and $\mu^\dagger$ with the same number of support points could be upper bounded by a weighted ℓ^2-type distance of their corresponding coefficients and positions.

Proof. We use that any finitely supported positive measure with N support points μ can be extended with $h_2(\widetilde{\mu}) = \mu$ according to

$$\widetilde{\mu} = \frac{1}{N} \sum_{n=1}^{N} \delta_{(r_n, y_n)}, \quad \text{where } r_n = \sqrt{N|q_n|}.$$

In addition, notice that $d_{\mathrm{HK}}(\mu, \mu^\dagger) = d_{\mathrm{HK}}(\mu^1, \mu^2)$ where $\mu^1 := \mu^+ + \mu^{\dagger,-}$ and $\mu^2 := \mu^{\dagger,+} + \mu^-$ are positive measures with N support of points. Thus, combining this with the fact that $(1/N)\, d_{\mathrm{cone}}((r_1, y_1), (r_2, y_2)) = d_{\mathrm{cone}}((r_1/\sqrt{N}, y_1), (r_2/\sqrt{N}, y_2))$ it follows:

$$\begin{aligned} d_{\mathrm{HK}}\left(\mu, \mu^\dagger\right)^2 &\leqslant \sum_{n=1}^{N_\dagger} \left[\left(\sqrt{|q_n|} - \sqrt{|q_n^\dagger|} \right)^2 + 4\sqrt{|q_n||q_n^\dagger|} \cdot \sin_+^2\left(\left\|y_n - y_n^\dagger\right\|_2 / 2 \right) \right] \\ &\leqslant \sum_{n=1}^{N_\dagger} \left(\frac{\left(q_n - q_n^\dagger\right)^2}{4\sqrt{|q_n||q_n^\dagger|}} + \sqrt{|q_n||q_n^\dagger|} \cdot \left\|y_n - y_n^\dagger\right\|_2^2 \right). \end{aligned}$$

Here, we have used $\sin_+^2(\cdot) \leqslant (\cdot)^2$ and $(\sqrt{a}-\sqrt{b})^2 = (a-b)^2/(\sqrt{a}+\sqrt{b})^2 \leqslant (a-b)^2/(4\sqrt{ab})$. This immediately implies the estimate. □

The previous result motivates to define a weighted ℓ^2-norm for the given parameters $(\boldsymbol{q};\boldsymbol{y}) \in (\mathbb{R}\setminus\{0\})^N \times \Omega_s^N$. More precisely, we define the weight $w = \sqrt{|\boldsymbol{q}|} := (\sqrt{|q_1|},\cdots,\sqrt{|q_N|}) \in (\mathbb{R}\setminus\{0\})^N$ and the associated weighted norm for a perturbation $(\delta\boldsymbol{q};\delta\boldsymbol{y}) \in \mathbb{R}^N \times \mathbb{R}^{dN}$ as

$$\|(\delta\boldsymbol{q};\delta\boldsymbol{y})\|_W^2 := \frac{1}{4}\left\|w^{-1}\delta\boldsymbol{q}\right\|_2^2 + \|w\,\delta\boldsymbol{y}\|_2^2 = \sum_{n=1}^{N}\left(\frac{|\delta q_n|^2}{4|q_n|} + |q_n|\,\|\delta y_n\|_2^2\right), \tag{4.6}$$

where $(w\delta\boldsymbol{y})_n = w_n\delta y_n$ denotes the entry-wise (Hadamard) product. Here, the diagonal matrix $W = \mathrm{diag}((w^{-2}/4; w^2; \ldots; w^2))$ induces the norm in (4.6). Then by proposition 4.2, we have

$$d_{\mathrm{HK}}\left(\mu,\mu^\dagger\right)^2 \leqslant R\left(\boldsymbol{q},\boldsymbol{q}^\dagger\right)\left\|\left(\boldsymbol{q}-\boldsymbol{q}^\dagger;\boldsymbol{y}-\boldsymbol{y}^\dagger\right)\right\|_{W_\dagger}^2,$$

where $W_\dagger$ is the diagonal weight matrix defined above for the weight $w^\dagger = \sqrt{|\boldsymbol{q}^\dagger|}$. Moreover, two different weighted norms are equivalent up to the same factor

$$R\left(\boldsymbol{q},\boldsymbol{q}^\dagger\right)^{-1}\|(\delta\boldsymbol{q};\delta\boldsymbol{y})\|_{W_\dagger}^2 \leqslant \|(\delta\boldsymbol{q};\delta\boldsymbol{y})\|_W^2 \leqslant R\left(\boldsymbol{q},\boldsymbol{q}^\dagger\right)\|(\delta\boldsymbol{q};\delta\boldsymbol{y})\|_{W_\dagger}^2 \tag{4.7}$$

because $R(\boldsymbol{q},\boldsymbol{q}^\dagger) = \max\{\left\|w/w^\dagger\right\|_\infty, \left\|w^\dagger/w\right\|_\infty\}$. For $\mu \approx \mu^\dagger$ the factor $R(\boldsymbol{q},\boldsymbol{q}^\dagger)$ is arbitrarily close to one. In other words, asymptotically for $\mu \approx \mu^\dagger$ the upper bound from proposition 4.2 is sharp:

$$d_{\mathrm{HK}}\left(\mu,\mu^\dagger\right)^2 \approx \left\|\left(\boldsymbol{q}-\boldsymbol{q}^\dagger;\boldsymbol{y}-\boldsymbol{y}^\dagger\right)\right\|_{W_\dagger}^2.$$

5. Fully explicit estimates for the deterministic reconstruction error

The Hellinger–Kantorovich distance allows us to quantify the reconstruction error between the unknown source $\mu^\dagger$ and measures obtained by solving ($\mathcal{P}_{\beta,\varepsilon}$). This will be done in two steps. First, we study the approximation of $\boldsymbol{m}^\dagger = (\boldsymbol{q}^\dagger;\boldsymbol{y}^\dagger)$, i.e. the support points and coefficients of the ground truth, by stationary points $\widehat{\boldsymbol{m}} = \widehat{\boldsymbol{m}}(\varepsilon)$ of the nonconvex parametrized problem

$$\min_{\boldsymbol{m}=(\boldsymbol{q};\boldsymbol{y})\in(\mathbb{R}\times\Omega_s)^{N_s}}\left[\frac{1}{2}\left\|G(\boldsymbol{m}) - G\left(\boldsymbol{m}^\dagger\right) - \varepsilon\right\|_{\Sigma_0^{-1}}^2 + \beta\|\boldsymbol{q}\|_1\right], \tag{5.1}$$

where the source-to-observable map G satisfies

$$G(\boldsymbol{m}) = G(\boldsymbol{q};\boldsymbol{y}) = k[\boldsymbol{x},\boldsymbol{y}]\boldsymbol{q} = \sum_{n=1}^{N} q_n k[\boldsymbol{x},y_n]. \tag{5.2}$$

By assumption A1, the latter is three times differentiable. Notice that (5.1) is obtained from (3.3) by fixing $N_s = N_s^\dagger$ points of sources in the formulation. Hence, solutions, let alone stationary points, of problem (5.1) do not parametrize minimizers of ($\mathcal{P}_{\beta,\varepsilon}$) in general. Moreover, it is clear that problem (5.1) is primarily of theoretical interest since its practical realization requires knowledge of $N_s^\dagger$. Thus, in a second step, we investigate for which noises ε, $\widehat{\boldsymbol{m}}$ parametrizes the unique solution of ($\mathcal{P}_{\beta,\varepsilon}$). While these results build upon similar techniques as [11], we give a precise, quantitative characterization of this asymptotic regime and clarify

the dependence of the involved constants on the problem parameters, e.g. the measurement points $\boldsymbol{x}$. This is necessary, for both, lifting these deterministic results to the stochastic setting in section 3 as well utilizing the derived error estimates in the context of optimal sensor placement. However, since these are merely intermediate steps in the derivation of our main result, we omit a detailed exposition at this point and direct the interested reader to appendix B. In the following, a central role will be played by the linearized problem

$$\min_{\delta\boldsymbol{m}=(\delta\boldsymbol{q};\delta\boldsymbol{y})\in\mathbb{R}^{(1+d)N_s}} \left[\frac{1}{2}\left\|G'\left(\boldsymbol{m}^\dagger\right)\delta\boldsymbol{m}-\varepsilon\right\|^2_{\Sigma_0^{-1}}+\beta\,\mathrm{sign}\left(\boldsymbol{q}^\dagger\right)^\top\delta\boldsymbol{q}\right]. \tag{5.3}$$

Note that here we have linearized both, the mapping G as

$$\begin{aligned} G\left(\boldsymbol{q}^\dagger+\delta\boldsymbol{q},\boldsymbol{y}^\dagger+\delta\boldsymbol{y}\right) &\approx G\left(\boldsymbol{q}^\dagger,\boldsymbol{y}^\dagger\right)+G'\left(\boldsymbol{q}^\dagger,\boldsymbol{y}^\dagger\right)(\delta\boldsymbol{q},\delta\boldsymbol{y}) \\ &= k\left[\boldsymbol{x},\boldsymbol{y}^\dagger\right]\boldsymbol{q}^\dagger+k\left[\boldsymbol{x},\boldsymbol{y}^\dagger\right]\delta\boldsymbol{q}+\left(\nabla_y^\top k\left[\boldsymbol{x},\boldsymbol{y}^\dagger\right]\circ\boldsymbol{q}^\dagger\right)\delta\boldsymbol{y}, \end{aligned}$$

using that

$$G'(\boldsymbol{m})=\begin{pmatrix}k[\boldsymbol{x},\boldsymbol{y}] & \nabla_y^\top k[\boldsymbol{x},\boldsymbol{y}]\circ\boldsymbol{q}\end{pmatrix} \quad \text{where} \quad \left(\nabla_y^\top k\left[\boldsymbol{x},\boldsymbol{y}^\dagger\right]\circ\boldsymbol{q}^\dagger\right)_{i,j}:=\nabla_y k\left(x_i,y_j^\dagger\right)^\top q_j^\dagger,$$

as well as the $\|\cdot\|_1$ −norm with

$$\left\|\boldsymbol{q}^\dagger+\delta\boldsymbol{q}\right\|_1\approx\left\|\boldsymbol{q}^\dagger\right\|_1+\mathrm{sign}\left(\boldsymbol{q}^\dagger\right)^\top\delta\boldsymbol{q}.$$

The following proposition characterizes the solutions of (5.1) and (5.3). Since its proof relies on standard computations, we omit it for the sake of brevity.

Proposition 5.1. *The solutions $\bar{\boldsymbol{m}}$ to (5.1) fulfill the stationarity condition*

$$S(\bar{\boldsymbol{m}}):=G'(\bar{\boldsymbol{m}})^\top\Sigma_0^{-1}\left(G(\bar{\boldsymbol{m}})-G\left(\boldsymbol{m}^\dagger\right)-\varepsilon\right)+\beta\left(\bar{\boldsymbol{\rho}};\boldsymbol{0}\right)=0, \tag{5.4}$$

for some $\bar{\boldsymbol{\rho}}\in\partial\|\bar{\boldsymbol{q}}\|_1$. The solutions of (5.3) satisfy

$$G'\left(\boldsymbol{m}^\dagger\right)^\top\Sigma_0^{-1}\left(G'\left(\boldsymbol{m}^\dagger\right)\delta\widehat{\boldsymbol{m}}-\varepsilon\right)+\beta\left(\boldsymbol{\rho};\boldsymbol{0}\right)=0,$$

where $\boldsymbol{\rho}=\mathrm{sign}\boldsymbol{q}^\dagger$. If $G'(\boldsymbol{m}^\dagger)$ has full column rank then the Fisher information matrix

$$\mathcal{I}_0:=G'\left(\boldsymbol{m}^\dagger\right)^\top\Sigma_0^{-1}G'\left(\boldsymbol{m}^\dagger\right) \tag{5.5}$$

is invertible and the unique solution of (5.3) is given by

$$\begin{aligned} \delta\widehat{\boldsymbol{m}}(\varepsilon) &:= \mathcal{I}_0^{-1}\left(G'\left(\boldsymbol{m}^\dagger\right)^\top\Sigma_0^{-1}\varepsilon-\beta\left(\boldsymbol{\rho};\boldsymbol{0}\right)\right) \\ &= \left(\Sigma_0^{-1/2}G'\left(\boldsymbol{m}^\dagger\right)\right)^+\Sigma_0^{-1/2}\varepsilon-\beta\mathcal{I}_0^{-1}\left(\boldsymbol{\rho};\boldsymbol{0}\right) \end{aligned} \tag{5.6}$$

where $(\Sigma_0^{-1/2}G'(\boldsymbol{m}^\dagger))^+=\mathcal{I}_0^{-1}G'(\boldsymbol{m}^\dagger)^\top\Sigma_0^{-1/2}$ is the pseudo-inverse of $\Sigma_0^{-1/2}G'(\boldsymbol{m}^\dagger)$.

Since (5.1) is nonconvex, the stationarity condition (5.4) is only necessary but not sufficient for optimality. In the following, we call any solution to (5.4) a stationary point.

5.1. Error estimates for stationary points

In this section, we show that for sufficiently small noise ε, problem (5.1) admits a unique stationary point $\widehat{\boldsymbol{m}}(\varepsilon)$ in the vicinity of $\boldsymbol{m}^\dagger$. Moreover, loosely speaking, $\boldsymbol{m}^\dagger$ and $\boldsymbol{m}^\dagger + \delta\widehat{\boldsymbol{m}}(\varepsilon)$ provide Taylor expansions of zeroth and first order, respectively, for $\widehat{\boldsymbol{m}}(\varepsilon)$.

Proposition 5.2. *Suppose that $G'(\boldsymbol{m}^\dagger)$ has full column rank. Then, for some constant $C_1 = C_1(k, \mu^\dagger, \|\mathcal{I}_0^{-1}\|_{W_\dagger^{-1} \to W_\dagger})$ and radius $\hat{r} > 0$ and all ε with $C_1(\|\varepsilon\|_{\Sigma_0^{-1}} + \beta) \leqslant 1$, the stationarity condition (5.1) admits a unique solution $\widehat{\boldsymbol{m}} = \widehat{\boldsymbol{m}}(\varepsilon)$ on $B_{W^\dagger}(\boldsymbol{m}^\dagger, (3/2)\hat{r})$. Moreover, the stationary point satisfies $\widehat{\boldsymbol{m}} \in B_{W^\dagger}(\boldsymbol{m}^\dagger, \hat{r})$ as well as*

$$\begin{aligned} \left\|\widehat{\boldsymbol{m}} - \boldsymbol{m}^\dagger\right\|_{W_\dagger} &\leqslant 2\left\|\delta\widehat{\boldsymbol{m}}\right\|_{W_\dagger} \leqslant C_1\left(\|\varepsilon\|_{\Sigma_0^{-1}} + \beta\right), \\ \left\|\widehat{\boldsymbol{m}} - \boldsymbol{m}^\dagger - \delta\widehat{\boldsymbol{m}}\right\|_{W_\dagger} &\leqslant C_1^2\left(\|\varepsilon\|_{\Sigma_0^{-1}} + \beta\right)^2. \end{aligned}$$

For the sake of brevity, we omit by now the proof of proposition 5.2, which is then presented in appendix B.

Remark 5.3. We note that C_1 depends monotonically on the norm of the inverse Fisher information matrix; see Remark B.3. Moreover, the dependency on the ground truth $\mu^\dagger$ is only in terms of the norm $\left\|\boldsymbol{q}^\dagger\right\|_1$, and distances of $y_n^\dagger$ to the boundary and $q_n^\dagger$ to zero.

5.2. Error estimates for reconstructions of the ground truth

As mentioned in the preceding section, solving the stationarity equation (5.4) for $\widehat{\boldsymbol{m}} = (\widehat{\boldsymbol{y}}, \widehat{\boldsymbol{q}})$ is not feasible in practice since it presupposes knowledge of $N_s^\dagger$. Moreover, recalling that $\widehat{\boldsymbol{m}}$ is merely a stationary point, the parametrized measure

$$\widehat{\mu} = \sum_{n=1}^{N_s^\dagger} \widehat{q}_n \delta_{\widehat{y}_n} \tag{5.7}$$

is not necessarily a minimizer of ($\mathcal{P}_{\beta,\varepsilon}$). In this section, our primary goal is to show that $\widehat{\boldsymbol{m}}$ indeed parametrizes the unique solution of problem ($\mathcal{P}_{\beta,\varepsilon}$) if the minimum norm dual certificate $\eta^\dagger$ associated to ($\mathcal{P}_0$) is θ-admissible and if the set of admissible noises ε is further restricted. A fully-explicit estimate for the reconstruction error between $\widehat{\mu}$ and the ground truth $\mu^\dagger$ in the Hellinger–Kantorovich distance then follows immediately. For this purpose, recall from [11, proposition 7] that the non-degeneracy of $\eta^\dagger$ implies

$$\eta^\dagger = \eta_{\mathrm{PC}} = K^*\Sigma_0^{-1/2}\left(G'\left(\boldsymbol{m}^\dagger\right)\Sigma_0^{-1/2}\right)^+(\boldsymbol{\rho};\mathbf{0}) = K^*\Sigma_0^{-1}G'\left(\boldsymbol{m}^\dagger\right)\mathcal{I}_0^{-1}(\boldsymbol{\rho};\mathbf{0}). \tag{5.8}$$

where η_{PC} denotes the vanishing derivative pre-certificate from section 3.3.

We first prove that

$$\widehat{\eta} = \beta^{-1}K^*\Sigma_0^{-1}\left(z^d(\varepsilon) - K\widehat{\mu}\right) = \beta^{-1}K^*\Sigma_0^{-1}\left(G\left(\boldsymbol{m}^\dagger\right) + \varepsilon - G(\widehat{\boldsymbol{m}})\right)$$

is $\theta/2$-admissible for certain ε and β.

Proposition 5.4. *Let the assumptions in proposition 5.2 be satisfied and $\eta^\dagger$ be $\theta-$admissible for $\mu^\dagger$, $\theta \in (0,1]$. Then there exists a constant $C_2 = C_2(k, \mu^\dagger, \|\mathcal{I}_0^{-1}\|_{W_\dagger^{-1} \to W_\dagger})$ such that if*

$$C_1\left(\|\varepsilon\|_{\Sigma_0^{-1}} + \beta\right) \leqslant \sqrt{\theta/32}, \tag{5.9}$$

$$C_2\beta^{-1}\left(\left(\|\varepsilon\|_{\Sigma_0^{-1}}+\beta\right)^2+\|\varepsilon\|_{\Sigma_0^{-1}}\right)\leqslant\theta^2/32, \tag{5.10}$$

then the function $\widehat{\eta}$ is $\theta/2-$admissible for $\widehat{\mu}$.

The proof of proposition 5.4 is then provided in appendix B.

Remark 5.5. We note that C_2 depends monotonically on the norm of the inverse Fisher information matrix; see remark B.5. Moreover, the dependency on the ground truth $\mu^\dagger$ is only in terms of the norm $\left\|\boldsymbol{q}^\dagger\right\|_1$, and distances of $y_n^\dagger$ to the boundary and $q_n^\dagger$ to zero.

As a consequence, we conclude that the solution to ($\mathcal{P}_{\beta,\varepsilon}$) is unique and parametrized by $\widehat{\boldsymbol{m}}$. Moreover, its H-K distance to $\mu^\dagger$ can be bounded in terms of the linearization $\delta\widehat{\boldsymbol{m}}$.

Theorem 5.6. *Let the assumptions of proposition 5.4 hold. Then the solution of ($\mathcal{P}_{\beta,\varepsilon}$) is unique and given by $\widehat{\mu}$ from (5.7). There holds*

$$d_{\mathrm{HK}}\left(\widehat{\mu},\mu^\dagger\right)^2\leqslant 8\left\|\delta\widehat{\boldsymbol{m}}\right\|_{W_\dagger}^2. \tag{5.11}$$

Proof. From proposition 5.4, we conclude that $\widehat{\eta}$ is $\theta/2$-admissible for $\widehat{\mu}$. Consequently, we have $\widehat{\eta}\in\partial\|\widehat{\mu}\|_{\mathcal{M}(\Omega_s)}$, i.e. $\widehat{\mu}$ is a solution of ($\mathcal{P}_{\beta,\varepsilon}$). It remains to show its uniqueness. For this purpose, it suffices to argue that

$$K_{|\operatorname{supp}\widehat{\mu}}=k[\boldsymbol{x},\widehat{\boldsymbol{y}}]\in\mathbb{R}^{N_o\times N_s^\dagger}$$

is injective, see, e.g. the proof of [31, proposition 3.6]. Assume that this is not the case. Then, following [30, theorem B.4], there is $\boldsymbol{v}\neq 0$ with $k[\boldsymbol{x},\boldsymbol{y}]\boldsymbol{v}=0$ and $\tau\neq 0$ such that the measure $\tilde{\mu}$ parametrized by $\tilde{\boldsymbol{m}}=(\tilde{\boldsymbol{q}};\widehat{\boldsymbol{y}})$ with $\tilde{\boldsymbol{q}}=\widehat{\boldsymbol{q}}+\tau\boldsymbol{v}$ is also a solution of ($\mathcal{P}_{\beta,\varepsilon}$) (choose the sign of τ to not increase the ℓ_1-regularization, and the magnitude small not to change the sign of $\tilde{\boldsymbol{q}}$) and $\tilde{\boldsymbol{q}}\neq\widehat{\boldsymbol{q}}$. For $s\in(0,1)$, set $\boldsymbol{q}_s=(1-s)\widehat{\boldsymbol{q}}+s\tilde{\boldsymbol{q}}$. By convexity of ($\mathcal{P}_{\beta,\varepsilon}$), the measure parametrized by $\boldsymbol{m}_s=(\boldsymbol{q}_s;\widehat{\boldsymbol{y}})$ is also a minimizer of ($\mathcal{P}_{\beta,\varepsilon}$). Consequently, $\boldsymbol{m}_s$ is a solution of (5.1) and thus also a stationary point. Finally, noting that $\boldsymbol{m}_s\neq\widehat{\boldsymbol{m}}$, $s\in(0,1)$, and $\lim_{s\to 0}\boldsymbol{m}_s=\widehat{\boldsymbol{m}}$, we arrive at a contradiction to the uniqueness of stationary points in the vicinity of $\boldsymbol{m}^\dagger$. The estimate in (5.11) immediately follows from

$$d_{\mathrm{HK}}\left(\widehat{\mu},\mu^\dagger\right)^2\leqslant R\left(\widehat{\boldsymbol{q}},\boldsymbol{q}^\dagger\right)\left\|\widehat{\boldsymbol{m}}-\boldsymbol{m}^\dagger\right\|_{W_\dagger}^2\leqslant 2\left\|\widehat{\boldsymbol{m}}-\boldsymbol{m}^\dagger\right\|_{W_\dagger}^2\leqslant 8\left\|\delta\widehat{\boldsymbol{m}}\right\|_{W_\dagger}^2$$

□

6. Inverse problems with random noise

Finally, let $(\mathcal{D},\mathcal{F},\mathbb{P})$ denote a probability space and consider the stochastic measurement model

$$z^d(\varepsilon)=K\mu^\dagger+\varepsilon,$$

where the noise is distributed according to $\varepsilon\sim\gamma_p=\mathcal{N}(0,p^{-1}\Sigma_0)$ for some $p>0$ representing the overall precision of the measurements. Mimicking the deterministic setting, we are interested in the reconstruction of the ground truth $\mu^\dagger$ by solutions obtained from ($\mathcal{P}_{\beta,\varepsilon}$) for realizations of the random variable ε. By utilizing the quantitative analysis presented in the preceding section, we provide an upper bound on the worst-case mean-squared error

$$\mathbb{E}_{\gamma_p}\left[\sup_{\mu\in\mathfrak{M}(\cdot)} d_{\mathrm{HK}}\left(\mu,\mu^\dagger\right)^2\right] = \int_{\mathbb{R}^{N_o}} \sup_{\mu\in\mathfrak{M}(\varepsilon)} d_{\mathrm{HK}}\left(\mu,\mu^\dagger\right)^2 \,\mathrm{d}\gamma_p\left(\varepsilon\right)$$

for a suitable *a priori* parameter choice rule $\beta = \beta(p)$. Note that the expectation is well-defined according to appendix A.2.

6.1. A priori *parameter choice rule*

Before stating the main result of the manuscript, let us briefly motivate the particular choice of the misfit term in ($\mathcal{P}_{\beta,\varepsilon}$) as well as the employed parameter choice rule from the perspective of the stochastic noise model. Since we consider independent measurements, their covariance matrix $\Sigma = p^{-1}\Sigma_0$ is diagonal with $\Sigma_{jj} = \sigma_j^2$ for variances $\sigma_j^2 > 0$, $j = 1,\dots,N_o$. This corresponds to performing the individual measurements with independent sensors of variable precision $p_j = 1/\sigma_j^2$. We call

$$p = \sum_{j=1}^{N_o} p_j = \sum_{n=1}^{N_o} \sigma_j^{-2} = \mathrm{tr}\left(\Sigma^{-1}\right).$$

the total precision of the sensor array. It can be seen that its reciprocal $\sigma_{\mathrm{tot}}^2 = 1/p$ corresponds to the harmonic average of the variances divided by the number of sensors N_o. Therefore, the misfit in ($\mathcal{P}_{\beta,\varepsilon}$) satisfies

$$\left\|K\mu - z^d\left(\varepsilon\right)\right\|_{\Sigma_0^{-1}}^2 = \frac{1}{p}\sum_{j=1}^{N_o} \sigma_j^{-2} |[K\mu]_j - z^d\left(\varepsilon\right)_j|^2.$$

For identical sensors and measurements $\varepsilon \sim \mathcal{N}(0, \mathrm{Id}_{N_o})$ this simply leads to the scaled Euclidean norm $(1/N_o)\left\|K\mu - z^d(\varepsilon)\right\|_2^2$. In general, by increasing the total precision of the sensor setup p, we improve the measurements by proportionally decreasing the variances by σ_{tot}^2. While this will decrease the expected level of noise through its distribution, it will not affect the misfit functional, which is just influenced by Σ_0, or the normalized variances $\sigma_{0,j}^2 = \sigma_j^2/\sigma_{\mathrm{tot}}^2$.

Moreover, since $\varepsilon \sim \mathcal{N}(0,\Sigma)$, we have $\Sigma^{-1/2}\varepsilon \sim \mathcal{N}(0,_{N_o})$ and by direct calculations, the following estimate holds

$$\frac{N_o}{\sqrt{N_o+1}} \leqslant \mathbb{E}_{\gamma_p}\left[\|\varepsilon\|_{\Sigma^{-1}}\right] \leqslant \sqrt{N_o}.$$

Hence, with high probability, realizations of the error fulfill the estimate

$$\sqrt{\sum_{j=1}^{N_o} \varepsilon_j^2/\sigma_j^2} = \|\varepsilon\|_{\Sigma^{-1}} = \sqrt{p}\,\|\varepsilon\|_{\Sigma_0^{-1}} \leqslant C\sqrt{N_o}$$

and thus $\|\varepsilon\|_{\Sigma_0^{-1}} \lesssim 1/\sqrt{p}$. Thus, we consider the expected noise $\sigma_{\mathrm{tot}} = 1/\sqrt{p}$ as an (expected) upper bound for the noise. This motivates the parameter choice rule

$$\beta\left(p\right) = \beta_0/\sqrt{p} = \beta_0\,\mathrm{tr}\left(\Sigma^{-1}\right)^{-1/2}$$

for some $\beta_0 > 0$ large enough.

6.2. Quantitative error estimates in the stochastic setting

We are now prepared to prove a quantitative estimate on the worst-case mean-squared error by lifting the deterministic result of theorem 5.6 to the stochastic setting.

Theorem 6.1. *Assume that $\eta^\dagger$ is θ-admissible for $\theta \in (0,1)$ and set $\beta(p) = \beta_0/\sqrt{p}$. Then there exists*

$$\overline{p} = \beta_0^2 c_p\left(\theta, k, \mu^\dagger, \left\|\mathcal{I}_0^{-1}\right\|_{W_\dagger^{-1}\to W_\dagger}\right)$$

such that for $p \geqslant \overline{p}$, there holds

$$\mathbb{E}_{\gamma_p}\left[\sup_{\mu\in\mathfrak{M}(\cdot)} d_{\mathrm{HK}}\left(\mu,\mu^\dagger\right)^2\right] \leqslant 8\mathbb{E}_{\gamma_p}\left[\|\delta\widehat{\boldsymbol{m}}\|_{W_\dagger}^2\right] + C_3\exp\left[-\left(\frac{\theta^2\beta_0}{64C_4}\right)^2/(2N_o)\right], \tag{6.1}$$

where $C_3 = 2\|\mu^\dagger\|_{\mathcal{M}(\Omega_s)} + \sqrt{2N_o}/(2\beta_0\sqrt{p})$ and $C_4 = \max\{C_1, C_2\}$. In addition, the expectation $\mathbb{E}_{\gamma_p}[\|\delta\widehat{\boldsymbol{m}}\|_{W_\dagger}^2]$ has the closed form

$$\mathbb{E}_{\gamma_p}\left[\|\delta\widehat{\boldsymbol{m}}\|_{W_\dagger}^2\right] = \frac{1}{p}\left(\mathrm{tr}\left(W_\dagger\mathcal{I}_0^{-1}\right) + \beta_0^2\left\|\mathcal{I}_0^{-1}(\boldsymbol{\rho};\mathbf{0})\right\|_{W_\dagger}^2\right). \tag{6.2}$$

Proof. Define the sets

$$A_1 = \left\{\varepsilon : C_4\beta(p)^{-1}\|\varepsilon\|_{\Sigma_0^{-1}} \leqslant \frac{\theta^2}{64}\right\}, \quad A_2 = \left\{\varepsilon : C_4\beta(p)^{-1}\left(\|\varepsilon\|_{\Sigma_0^{-1}} + \beta(p)\right)^2 \leqslant \frac{\theta^2}{64}\right\}.$$

By a case distinction, we readily verify

$$\mathbb{R}^{N_o}\setminus(A_1\cap A_2) \subset \left(\mathbb{R}^{N_o}\setminus A_1\right)\cup\left(\left(\mathbb{R}^{N_o}\setminus A_2\right)\cap A_1\right)$$

and thus

$$\begin{aligned}\mathbb{E}_{\gamma_p}\left[\sup_{\mu\in\mathfrak{M}(\cdot)} d_{\mathrm{HK}}\left(\mu,\mu^\dagger\right)^2\right] \leqslant &\int_{A_1\cap A_2}\sup_{\mu\in\mathfrak{M}(\varepsilon)} d_{\mathrm{HK}}\left(\mu,\mu^\dagger\right)^2 \mathrm{d}\gamma_p(\varepsilon)\\ &+ \underbrace{\int_{\mathbb{R}^{N_o}\setminus A_1}\sup_{\mu\in\mathfrak{M}(\varepsilon)} d_{\mathrm{HK}}\left(\mu,\mu^\dagger\right)^2 \mathrm{d}\gamma_p(\varepsilon)}_{I_1}\\ &+ \underbrace{\int_{(\mathbb{R}^{N_o}\setminus A_2)\cap A_1}\sup_{\mu\in\mathfrak{M}(\varepsilon)} d_{\mathrm{HK}}\left(\mu,\mu^\dagger\right)^2 \mathrm{d}\gamma_p(\varepsilon)}_{I_2}.\end{aligned} \tag{6.3}$$

For $\varepsilon \in A_1\cap A_2$, we have

$$\begin{aligned}C_2\beta(p)^{-1}\left(\left(\|\varepsilon\|_{\Sigma_0^{-1}} + \beta(p)\right)^2 + \|\varepsilon\|_{\Sigma_0^{-1}}\right) &\leqslant C_4\beta^{-1}(p)\left(\|\varepsilon\|_{\Sigma_0^{-1}} + \beta(p)\right)^2 + C_4\beta^{-1}(p)\|\varepsilon\|_{\Sigma_0^{-1}}\\ &\leqslant \frac{\theta^2}{64} + \frac{\theta^2}{64} = \frac{\theta^2}{32},\end{aligned}$$

i.e. ε satisfies (5.10). Moreover, expanding the square in the definition of A_2, we conclude that (5.9) also holds due to

$$2C_4\left(\|\varepsilon\|_{\Sigma_0^{-1}}+\beta(p)\right)\leqslant\frac{\theta^2}{32}\leqslant\frac{\sqrt{\theta}}{2\sqrt{32}}.$$

Hence, for $\varepsilon\in A_1\cap A_2$, there holds $\mathfrak{M}(\varepsilon)=\{\widehat{\mu}\}$ and

$$\sup_{\mu\in\mathfrak{M}(\varepsilon)} d_{\mathrm{HK}}\left(\mu,\mu^\dagger\right)^2=d_{\mathrm{HK}}\left(\widehat{\mu},\mu^\dagger\right)^2\leqslant 8\left\|\delta\widehat{\boldsymbol{m}}\right\|_{W_\dagger}^2 \tag{6.4}$$

by proposition 5.6. Next, we estimate I_1 by

$$d_{\mathrm{HK}}\left(\widehat{\mu},\mu^\dagger\right)^2\leqslant\left\|\mu^\dagger\right\|_{\mathcal{M}(\Omega_s)}+\|\widehat{\mu}\|_{\mathcal{M}(\Omega_s)}\leqslant 2\left\|\mu^\dagger\right\|_{\mathcal{M}(\Omega_s)}+\|\varepsilon\|_{\Sigma_0^{-1}}^2/\left(2\beta_0/\sqrt{p}\right)$$

applying proposition 3.3 and [24, proposition 7.8]. Together with lemma A.1 this yields

$$\begin{aligned}I_1=\int_{\mathbb{R}^{N_o}\setminus A_1} d_{\mathrm{HK}}\left(\widehat{\mu},\mu^\dagger\right)^2\mathrm{d}\gamma_p(\varepsilon)&\leqslant\int_{\mathbb{R}^{N_o}\setminus A_1}\left(2\left\|\mu^\dagger\right\|_{\mathcal{M}(\Omega_s)}+\|\varepsilon\|_{\Sigma^{-1}}^2/\left(2\beta_0\sqrt{p}\right)\right)\mathrm{d}\gamma_p(\varepsilon)\\&\leqslant\left(2\left\|\mu^\dagger\right\|_{\mathcal{M}(\Omega_s)}+\frac{\sqrt{2N_o}}{2\beta_0\sqrt{p}}\right)\exp\left[-\left(\frac{\theta^2\beta_0}{64C_4}\right)^2\Big/\left(2N_o\right)\right].\end{aligned} \tag{6.5}$$

Finally, for $\varepsilon\in(\mathbb{R}^{N_o}\setminus A_2)\cap A_1$, one has

$$p^{1/4}\left(\frac{\theta^2\beta_0}{64C_4}\right)^{1/2}-\beta_0<\|\varepsilon\|_{\Sigma^{-1}}=p^{-1/2}\|\varepsilon\|_{\Sigma_0^{-1}}\leqslant\frac{\theta^2\beta_0}{64C_4}$$

where the first inequality follows from $\varepsilon\notin A_2$ and the second follows from $\varepsilon\in A_1$. Hence, if we choose

$$p\geqslant\beta_0^2\left(\frac{\theta^2}{64C_4}+1\right)^4\Big/\left(\frac{\theta^2}{64C_4}\right)^2:=\beta_0^2\bar{c}_p:=\bar{p}$$

then $(\mathbb{R}^{N_o}\setminus A_2)\cap A_1$ is empty and $I_2=0$. Together with (6.3)–(6.5), we obtain (6.1) for every $p\geqslant\bar{p}$. The equality in (6.2) follows immediately from the closed form expression (5.6) for $\delta\widehat{\boldsymbol{m}}$ and $\varepsilon\sim\mathcal{N}(0,p^{-1}\Sigma_0)$.

□

Let us interpret this result: By choosing β_0 large enough, the second term on the right hand side of (6.6) becomes negligible, i.e.

$$\mathbb{E}_{\gamma_p}\left[\sup_{\mu\in\mathfrak{M}(\cdot)} d_{\mathrm{HK}}\left(\mu,\mu^\dagger\right)^2\right]\leqslant 8\mathbb{E}_{\gamma_p}\left[\left\|\delta\widehat{\boldsymbol{m}}\right\|_{W_\dagger}^2\right]+\delta \tag{6.6}$$

for some $0<\delta\ll 1$. As a consequence, due to its closed form representation (6.2), $\mathbb{E}_{\gamma_p}[\|\delta\widehat{\boldsymbol{m}}\|_{W_\dagger}^2]$ provides a computationally inexpensive, approximate upper surrogate for the worst-case mean-squared error which vanishes as $p\to\infty$. Moreover, due to its explicit dependence on the measurement setup, it represents a suitable candidate for an optimal design criterion in the context of optimal sensor placement for the class of sparse inverse problems under consideration. This potential will be further investigated in a follow-up paper.

Remark 6.2. It is worth mentioning that the constant 8 appearing on the right hand side of (6.6) is not optimal and is primarily a result of the proof technique. In fact, by appropriately selecting constants in propositions B.2 and 5.2, it is possible to replace 8 by $1+\delta$, where $0<\delta\ll 1$ at the cost of increasing $\bar{p}$. We will illustrate the sharpness of the estimate of the worst-case mean-squared error by $\mathbb{E}_{\gamma_p}[\|\delta\widehat{\boldsymbol{m}}\|^2_{W_\dagger}]$ in the subsequent numerical results.

Remark 6.3. Relying on similar arguments as in the proof of theorem 6.1, we are also able to derive pointwise estimates on the Hellinger–Kantorovich distance which hold with high probability. Indeed, noticing that (6.4) holds in the set $A_1\cap A_2$, we derive a lower probability bound for $\mathbb{P}(\varepsilon\in A_1\cap A_2)$ by noticing that

$$\begin{aligned}\mathbb{P}\left(\varepsilon\in A_1\cap A_2\right)&\geqslant \mathbb{P}\left(\varepsilon\in A_1\right)+\mathbb{P}\left(\varepsilon\in A_2\right)-1\\&\geqslant 1-\mathbb{P}\left(\varepsilon\in\mathbb{R}^{N_o}\backslash A_1\right)-\mathbb{P}\left(\varepsilon\in\mathbb{R}^{N_o}\backslash A_2\right).\end{aligned}$$

By invoking lemma A.1, one has

$$\mathbb{P}\left(\varepsilon\in\mathbb{R}^{N_o}\backslash A_1\right)=\mathbb{P}\left(\|\varepsilon\|_{\Sigma^{-1}}>\frac{\theta^2\beta_0}{64C_4}\right)\leqslant 2\exp\left[-\left(\frac{\theta^2\beta_0}{64C_4}\right)^2\Big/\left(2N_o\right)\right],$$

$$\mathbb{P}\left(\varepsilon\in\mathbb{R}^{N_o}\backslash A_2\right)=\mathbb{P}\left(\|\varepsilon\|_{\Sigma^{-1}}>\frac{p^{1/2}\theta\beta_0^{1/2}}{8C_4^{1/2}}-\beta_0\right)\leqslant 2\exp\left[-\left(\frac{p^{1/2}\theta\beta_0^{1/2}}{8C_4^{1/2}}-\beta_0\right)^2\Big/\left(2N_o\right)\right].$$

Hence, since $\exp(-x^2)\to 0$ as $x\to\infty$, we can see that for every $\delta\in(0,1)$, one can choose β_0 and p large enough such that

$$\exp\left[-\left(\frac{\theta^2\beta_0}{64C_4}\right)^2\Big/\left(2N_o\right)\right]<\delta/4,\quad \exp\left[-\left(\frac{p^{1/2}\theta\beta_0^{1/2}}{8C_4^{1/2}}-\beta_0\right)^2\Big/\left(2N_o\right)\right]<\delta/4,$$

which implies $\mathbb{P}(\varepsilon\in A_1\cap A_2)\geqslant 1-\delta$. Therefore, with probability at least $1-\delta$, we have

$$\sup_{\mu\in\mathfrak{M}(\varepsilon)} d_{\mathrm{HK}}\left(\mu,\mu^\dagger\right)^2\leqslant 8\,\|\delta\widehat{\boldsymbol{m}}\|^2_{W_\dagger}$$

for realization ε of the noise. Furthermore, employing lemma A.1 again, we know that with probability at least $1-\delta$, and independently from p, one has $\|\varepsilon\|_{\Sigma^{-1}}\leqslant\sqrt{-2N_o\ln(\delta/2)}$. Hence, by proposition 5.2 together with $\varepsilon\in A_1\cap A_2$, we have

$$\begin{aligned}\sup_{\mu\in\mathfrak{M}(\varepsilon)} d_{\mathrm{HK}}\left(\mu,\mu^\dagger\right)^2&\leqslant 8C_1p^{-1/2}\left(\|\varepsilon\|_{\Sigma^{-1}}+\beta_0\right)\\&\leqslant 8C_1p^{-1/2}\left(\sqrt{-2N_o\ln\left(\delta/2\right)}+\beta_0\right)\end{aligned}$$

with probability at least $1-2\delta$.

7. Numerical results

We end this paper with the study of some numerical examples to illustrate our theory. We consider a simplified version of example 3.1:

- The source domain Ω_s and observation domain Ω_o are the interval $[-1,1]$.
- The reference measure is given by $\mu^\dagger = 0.4\delta_{-0.7} + 0.3\delta_{-0.3} - 0.2\delta_{0.3} \in \mathcal{M}(\Omega_s)$.
- The kernel $k : [-1,1] \times [-1,1] \to \mathbb{R}$ is defined as

$$k(x,y) = \exp\left(-\frac{(x-y)^2}{2\sigma^2}\right), \sigma = 0.2, \quad x,y \in [-1,1].$$

- The measurement points $\{x_1,\dots,x_{N_o}\} \subset \Omega_o$ vary between the individual examples and are marked by grey points in the respective plots. The associated noise model is given by $\varepsilon \sim \mathcal{N}(0,\Sigma)$ with $\Sigma^{-1} = p\Sigma_0^{-1}$, where $\Sigma_0^{-1} = (1/N_o)\,\mathrm{Id}_{N_o}$.

Following our theory, we attempt to recover $\mu^\dagger$ by solving ($\mathcal{P}_{\beta,\varepsilon}$) using the *a priori* parameter choice rule $\beta(p) = \beta_0/\sqrt{p}$. The regularized problems are solved by the Primal-Dual-Active-Points method, [26, 31], yielding a solution $\bar{\mu}$. Since the action of the forward operator K on sparse measures can be computed analytically, the algorithm is implemented on a grid free level. In addition, we compute a stationary point $\widehat{\boldsymbol{m}}$ of the nonconvex problem (5.1) inducing the measure $\widehat{\mu}$ from (5.7). This is done by a similar iteration to the Gauss-Newton sequence (B.10) with a nonsmooth adaptation to handle the ℓ_1-norm and an added globalization procedure to make it converge without restrictions on the data. We note that this solution depends on the initialization of the algorithm at $\boldsymbol{m}^\dagger$, which is usually unavailable in practice. To evaluate the reconstruction results in a qualitative way, we follow [11] by considering the dual certificates and pre-certificates; see section 3. Our Matlab implementation is available at https://github.com/hphuoctruong/OED_SparseInverseProblems.

7.1. Example 1

In the first example, we illustrate the reconstruction capabilities of the proposed ansatz for different measurement setups and with and without noise in the observations. To this end, we attempt to recover the reference measure $\mu^\dagger$ using a variable number N_o of uniformly distributed sensors. For noisy data, the regularization parameter is selected as $\beta = \beta_0/\sqrt{p}$ where $\beta_0 = 2$ and $p = 10^4$. We first consider the exact measurement data with $N_o \in \{6,9,11\}$ and try to obtain $\mu^\dagger$ by solving ($\mathcal{P}_0$). The results are shown in figure 1. We observe that with 6 sensors, the pre-certificate η_{PC} is not admissible. Recalling [11, proposition 7], this implies that $\mu^\dagger$ is not a minimum norm solution. In contrast, the experiments with 9 and 11 uniform sensors provide admissible pre-certificates. In these situations, the pre-certificates coincide with the minimum norm dual certificates and the ground truth $\mu^\dagger$ is indeed an identifiable minimum norm solution.

Next, we consider noisy data and solve ($\mathcal{P}_{\beta,\varepsilon}$) for the aforementioned choice of $\beta(p)$. Following the observation in the first example, we only evaluate the reconstruction results obtained by 9 and 11 uniform sensors. In the absence of the measurement data obtained from experiments, we generate synthetic noisy measurements where the noise vector ε is a realization of the Gaussian random noise $\varepsilon \sim \mathcal{N}(0,\Sigma)$. The results are shown in figure 2. Since $\mu^\dagger$ is identifiable in these cases, $\widehat{\mu}$ and $\bar{\mu}$ coincide and closely approximate $\mu^\dagger$ with high probability for an appropriate choice of β_0 and p large enough. Both properties can be clearly observed in the plots, where $\beta_0 = 2$.

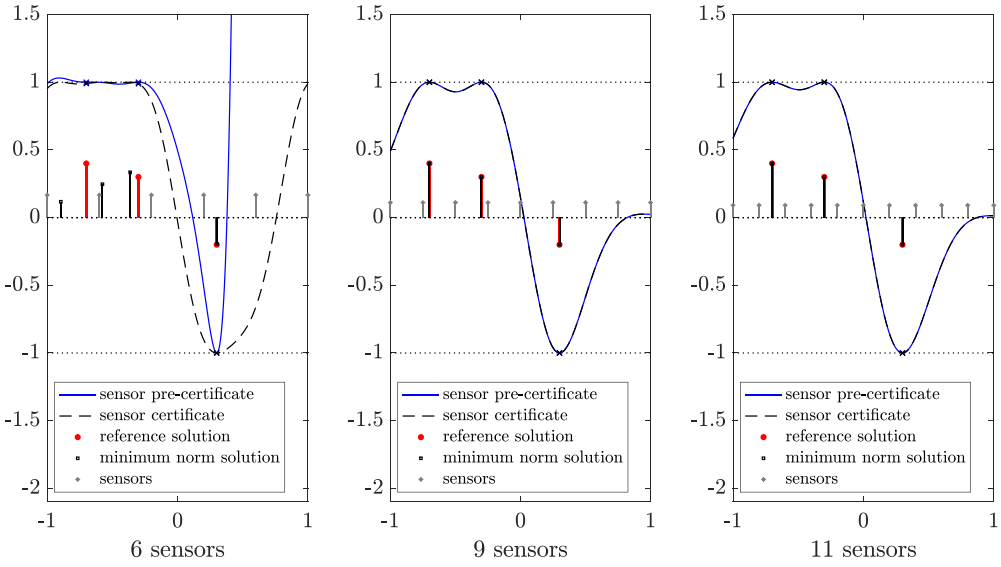

Figure 1. Reconstruction results with exact data using 6 sensors (left), 9 sensors (middle) and 11 sensors (right).

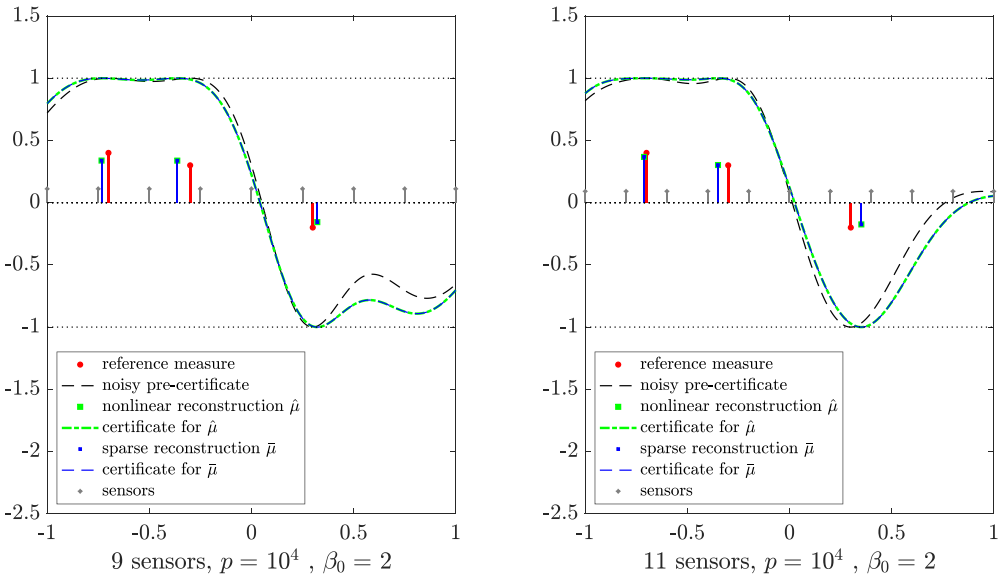

Figure 2. Reconstruction results with noisy data using 9 sensors (left) and 11 sensors (right).

7.2. *Example 2*

In the second example we study the influence of the parameter choice rule on the reconstruction result. To this end, we fix the measurement setup to 9 uniformly distributed sensors. We recall that the *a priori* parameter choice rule is given by $\beta(p) = \beta_0/\sqrt{p}$. According to section 6.2,

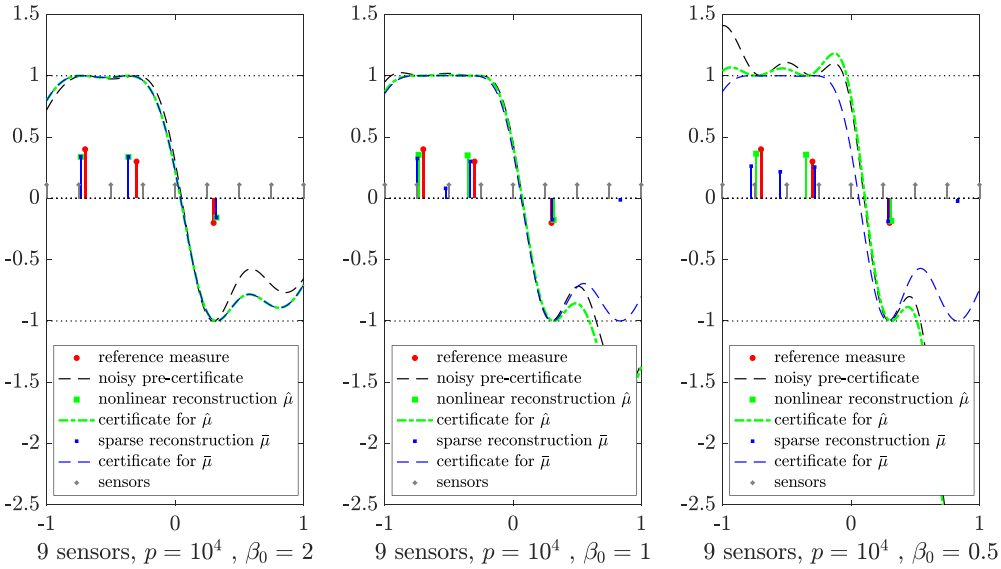

Figure 3. Reconstruction results with $\beta_0 = 2$ (left), $\beta_0 = 1$ (middle) and $\beta_0 = 0.5$ (right).

selecting a sufficiently large value for β_0 is recommended to achieve a high quality reconstruction. To determine a useful range of regularization parameters, we solve problem ($\mathcal{P}_{\beta,\varepsilon}$) for a sequence of regularization parameters using PDAP. Here, we choose $\beta_0 \in \{0.5, 1, 2\}$ and $p \in \{10^4, 10^5, 10^6\}$.

In figure 3, different reconstruction results are shown for the same realization of noise, $\beta_0 \in \{0.5, 1, 2\}$ and $p = 10^4$. As one can see, for this particular realization of the noise, the number of spikes is recovered exactly in the case $\beta_0 = 2$ and we again observe that $\widehat{\mu} = \bar{\mu}$. In contrast, for smaller β_0, the noisy pre-certificate is not admissible. Hence, while $\widehat{\mu}$ still provides a good approximation of $\mu^\dagger$, $\bar{\mu}$ admits two additional spikes away from the support of $\mu^\dagger$. These observations can be explained by looking at theorem 6.1 the second term on the right hand side of the inequality becomes negligible for increasing β_0 and large enough p. Thus, roughly speaking, the parameter β_0 controls the probability of the 'good events' in which $\widehat{\mu}$ is the unique solution of ($\mathcal{P}_{\beta,\varepsilon}$).

Finally, we address the reconstruction error from a quantitative perspective. For this purpose, we simplify the evaluation of the maximum mean-squared error (MSE) by inserting the solution $\bar{\mu}$ computed algorithmically. We note that this could only lead to an under-estimation of the maximum error in the case of non-unique solutions of ($\mathcal{P}_{\beta,\varepsilon}$); a degenerate case that is unlikely to occur in practice. Moreover, the expectation is approximated using 10^3 Monte-Carlo samples. Additionally, we use the closed form expression (6.2) for evaluating the linearized estimate $\mathbb{E}_{\gamma_p}[\|\delta\widehat{\boldsymbol{m}}\|^2_{W_\dagger}]$ exactly. Here, the expectations are computed for $\beta_0 \in \{2, 0.5\}$. The results are collected in table 1. We make several observations: Clearly, the MSE decreases for increasing p, i.e. lower noise level. For increased β_0, the behavior differs: For the theoretical quantities $\widehat{\boldsymbol{m}}$ and $\delta\widehat{\boldsymbol{m}}$ increased β_0 only introduces additional bias and thus increases error. For the estimator $\bar{\mu}$, the increased regularization however leads to generally improved results, since the probability of $\widehat{\mu} \neq \bar{\mu}$ is decreased. We highlight in bold the estimator which performed best for each β_0. Here, the results conform to theorem 6.1: For larger β_0, the second term on the right-hand side of (6.1) is negligible and the linearized estimate provides an excellent bound

Table 1. Reconstruction results with $\beta_0 = 2$ and $\beta_0 = 0.5$.

		$p = 10^4$	$p = 10^5$	$p = 10^6$	$p = 10^7$
$\beta_0 = 2$	$\mathbb{E}_{\gamma_p}[\|\|\delta\widehat{\boldsymbol{m}}\|\|^2_{W_\dagger}]$	$7.97 \cdot 10^{-3}$	$7.97 \cdot 10^{-4}$	$7.97 \cdot 10^{-5}$	$7.97 \cdot 10^{-6}$
	$\mathbb{E}_{\gamma_p}[d_{\mathrm{HK}}(\mu^\dagger, \widehat{\mu})^2]$	$5.43 \cdot 10^{-3}$	$6.49 \cdot 10^{-4}$	$7.35 \cdot 10^{-5}$	$7.76 \cdot 10^{-6}$
	$\mathbb{E}_{\gamma_p}[d_{\mathrm{HK}}(\mu^\dagger, \bar{\mu})^2]$	$5.44 \cdot 10^{-3}$	$\mathbf{6.51 \cdot 10^{-4}}$	$\mathbf{7.42 \cdot 10^{-5}}$	$\mathbf{7.99 \cdot 10^{-6}}$
$\beta_0 = 0.5$	$\mathbb{E}_{\gamma_p}[\|\|\delta\widehat{\boldsymbol{m}}\|\|^2_{W_\dagger}]$	$2.07 \cdot 10^{-3}$	$2.07 \cdot 10^{-4}$	$2.07 \cdot 10^{-5}$	$2.07 \cdot 10^{-6}$
	$\mathbb{E}_{\gamma_p}[d_{\mathrm{HK}}(\mu^\dagger, \widehat{\mu})^2]$	$1.71 \cdot 10^{-3}$	$1.94 \cdot 10^{-4}$	$2.03 \cdot 10^{-5}$	$2.06 \cdot 10^{-6}$
	$\mathbb{E}_{\gamma_p}[d_{\mathrm{HK}}(\mu^\dagger, \bar{\mu})^2]$	$\mathbf{4.12 \cdot 10^{-3}}$	$9.83 \cdot 10^{-4}$	$2.65 \cdot 10^{-4}$	$7.94 \cdot 10^{-5}$

Table 2. Reconstruction results with different sensor setups.

		11 sensors	'selected' 6 sensors	6 sensors
$\mathbb{E}_{\gamma_p}[d_{\mathrm{HK}}(\mu^\dagger, \bar{\mu})^2]$	$p = 10^4$	$5.03 \cdot 10^{-3}$	$\mathbf{4.25 \cdot 10^{-3}}$	$1.54 \cdot 10^{-2}$
	$p = 10^5$	$5.61 \cdot 10^{-4}$	$\mathbf{4.58 \cdot 10^{-4}}$	$1.19 \cdot 10^{-2}$
	$p = 10^6$	$6.31 \cdot 10^{-5}$	$\mathbf{4.65 \cdot 10^{-5}}$	$1.18 \cdot 10^{-2}$
$\mathbb{E}_{\gamma_p}[\|\|\delta\widehat{\boldsymbol{m}}\|\|^2_{W_\dagger}]$	$p = 10^4$	$6.09 \cdot 10^{-3}$	$4.77 \cdot 10^{-3}$	
	$p = 10^5$	$6.09 \cdot 10^{-4}$	$4.77 \cdot 10^{-4}$	Inf
	$p = 10^6$	$6.09 \cdot 10^{-5}$	$4.77 \cdot 10^{-5}$	

on the MSE for both $\widehat{\mu}$ and $\bar{\mu}$. We also note that the estimate is closer to the MSE in the limiting case for larger p. In contrast, for $\beta = 0.5$, the linearized estimate and the MSE of $\widehat{\mu}$ are much smaller than the MSE of the estimator $\bar{\mu}$. This underlines the observation that theorem 5.6 requires further restrictions on the admissible noises in comparison to proposition 5.2.

7.3. Example 3

The final example is devoted to compare the reconstruction results obtained by uniform designs and an improved design chosen by heuristics. To this end, we consider three measurement setups: uniformly distributed setups with 6 and 11 sensors, respectively, and one with 6 sensors selected on purpose. More precisely, in the later case, we place the sensors at $\Omega_o = \{-0.8, -0.6, -0.4, -0.1, 0.1, 0.4\}$. The different error measures are computed as in the previous example and the results are gathered in table 2. We observe that the measurement setup with 6 selected sensors performs better than the uniform ones. Moreover, the linearized estimate again provides a sharp upper bound on the error for both ten uniform and six selected sensors but yields numerically singular Fisher information matrices for six uniform sensors (denoted as Inf in the table), i.e. $\mu^\dagger$ is not stably identifiable in this case. Note that the estimator $\bar{\mu}$ still yields somewhat useful results, which are however affected by a constant error due to the difference in minimum norm solution and exact source as depicted in figure 1 and do not improve with lower noise level. These results suggest that the reconstruction quality does not only rely on the amount of measurements taken but also on their specific setup. In this case, we point out that the selected sensors are chosen to be adapted to the sources as every two sensors are placed on the two sides of every source. Thus the obtained results imply that if we have some reasonable prior information on the source positions and amplitudes, one may obtain a better sensor placement setup by incorporating it in the design of the measurement setup. This

leads to the concept of optimal sensor placement problems for sparse inversion which we will consider in a future work.

8. Conclusion

In the present work, we have considered the inverse problem of estimating an unknown sparse signal $\mu^\dagger$ from finitely many measurements perturbed by Gaussian random noise which was formulated as a linear, ill-posed operator equation in the space of Radon measures. The main result of the paper is an asymptotical sharp upper bound on the mean-squared error defined in terms of the Hellinger–Kantorovich distance of a nonsmooth Tikhonov-type estimator which is confirmed by extensive numerical experiments. Its proof relies on three key concepts: A suitable *a priori* regularization parameter choice rule $\beta = \beta(p)$ which is adapted to the overall precision of the measurements p, the non-degeneracy of the minimal-norm dual certificate as well as a careful linearization argument for the H-K distance on a quantifiable set of random events. In comparison to the intractable mean-squared error, the new bound is easily computable and explicitly depends on the locations of the measurement sensors as well as their relative precision. In perspective, these observations suggest the application of this new-found upper estimate in the context of optimal sensor design for sparse inverse problems. However, we also point out that a practical realization of such an approach is not straightforward since the derived upper bound, i.e. the prospective design criterion, depends on the unknown source $\mu^\dagger$ and the non-degeneracy of the minimal-norm certificate, a property that also inherently depends on the measurement setup. Addressing these problems goes beyond the scope of the current paper and will be addressed in future work. Moreover, the extension of the presented result towards vector measures as, e.g. encountered in acoustic inversion is of great interest.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/hphuoctruong/OED_SparseInverseProblems.

Acknowledgments

The work of P-T Huynh was supported by the Austrian Science Fund FWF under the Grant DOC 78. The material in this manuscript is based on work supported by the Laboratory Directed Research and Development Program at Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC, under Contract No. DE–AC05–00OR22725. The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

Appendix A. Auxiliary results

A.1. Gaussian tail bounds

The following lemma gives an estimate on the tail probabilities of a Gaussian random variables as well as on its moments.

Lemma A.1. *Assume that* $\varepsilon_0 \sim \gamma_E = \mathcal{N}(0, \mathrm{Id}_{N_o})$. *Then for every* $\alpha > 0$ *there holds*

$$\mathbb{P}\left(\|\varepsilon_0\|_2 > \alpha\right) \leqslant 2\exp\left(-\frac{\alpha^2}{2N_o}\right).$$

Moreover for $l \geqslant 1$, *with* $C_l = (2l-1)!! = (2l-1)(2l-3)\cdots 1$, *we have*

$$\int_{\|\varepsilon_0\|_2 > \alpha} \|\varepsilon_0\|_2^l \, \mathrm{d}\gamma_E(\varepsilon_0) \leqslant \sqrt{2N_o C_l}\exp\left(-\frac{\alpha^2}{4N_o}\right) = \exp\left(-\frac{\alpha^2}{4N_o} + \log\left(2N_o C_l\right)/2\right).$$

Proof. According to remark 4 in [32], we get for any $\lambda > 0$ that

$$\mathbb{P}\left(\|\varepsilon_0\|_2 > \alpha\right) = \int_{\|\varepsilon_0\| > \alpha} \mathrm{d}\gamma_E(\varepsilon_0) \leqslant 2\exp\left(-\lambda\alpha\right)\exp\left(\lambda^2 N_o/2\right).$$

Minimizing the right-hand side with respect to λ yields $\lambda = \alpha/N_o$ and the first estimate. The second inequality is due to

$$\int_{\|\varepsilon_0\| > \alpha} \|\varepsilon_0\|_2^l \, \mathrm{d}\gamma_E(\varepsilon_0) \leqslant \sqrt{\int \|\varepsilon_0\|_2^{2l} \, \mathrm{d}\gamma_E(\varepsilon_0)}\sqrt{2\exp\left(-\frac{\alpha^2}{2N_o}\right)}$$

with Cauchy-Schwarz. The proof is finished noting that $\mathbb{E}[\|\varepsilon_0\|_2^{2l}] = N_o C_l$ where $C_l = (2l-1)!!$ denotes the $2l$-the moment of the univariate standard normal distribution. □

A.2. Some results on measurability

In this section we address the measurability of the worst-case distance function

$$\varepsilon \mapsto \sup_{\mu \in \mathfrak{M}(\varepsilon)} d_{\mathrm{HK}}\left(\mu, \mu^\dagger\right) \tag{A.1}$$

as well as the boundedness of its second moment. For this purpose, recall the definition of the solution set

$$\mathfrak{M}(\varepsilon) = \operatorname*{arg\,min}_{\mu \in \mathcal{M}(\Omega_s)} \left[\frac{1}{2}\left\|K\mu - z^d(\varepsilon)\right\|_{\Sigma_0^{-1}}^2 + \beta\|\mu\|_{\mathcal{M}(\Omega_s)}\right]$$

and note that $\mathfrak{M}(\varepsilon)$ is weak* compact. We first show that the supremum in (A.1) is attained and give a useful upper bound on it.

Lemma A.2. *For every* $\varepsilon \in \mathbb{R}^{N_o}$, *there is* $\bar{\mu}(\varepsilon) \in \mathfrak{M}(\varepsilon)$ *with*

$$d_{\mathrm{HK}}\left(\bar{\mu}(\varepsilon), \mu^\dagger\right) = \sup_{\mu \in \mathfrak{M}(\varepsilon)} d_{\mathrm{HK}}\left(\mu, \mu^\dagger\right).$$

Moreover, we have

$$\sup_{\mu\in\mathfrak{M}(\varepsilon)} d_{\mathrm{HK}}\left(\mu,\mu^{\dagger}\right)^2 \leqslant 2\left\|\mu^{\dagger}\right\|_{\mathcal{M}(\Omega_s)} + \frac{1}{2\beta}\|\varepsilon\|^2_{\Sigma_0^{-1}}. \tag{A.2}$$

Proof. The inequality (A.2) follows from proposition 3.3 and [24, proposition 7.18]. Hence, there exists a supremizing sequence $\{\mu_k\}_k \subset \mathfrak{M}(\varepsilon)$, i.e,

$$\lim_{k\to\infty} d_{\mathrm{HK}}\left(\mu_k,\mu^{\dagger}\right) = \sup_{\mu\in\mathfrak{M}(\varepsilon)} d_{\mathrm{HK}}\left(\mu,\mu^{\dagger}\right).$$

Due to weak* compactness of $\mathfrak{M}(\varepsilon)$, it admits a subsequence, denoted by the same subscript, as well as $\bar{\mu}(\varepsilon)\in\mathfrak{M}(\varepsilon)$ with $\mu_k \rightharpoonup^* \bar{\mu}(\varepsilon)$. Since d_{HK} metrizes weak* convergence, the inverse triangle inequality yields

$$\left|d_{\mathrm{HK}}\left(\mu_k,\mu^{\dagger}\right) - d_{\mathrm{HK}}\left(\bar{\mu}\left(\varepsilon\right),\mu^{\dagger}\right)\right| \leqslant d_{\mathrm{HK}}\left(\bar{\mu}\left(\varepsilon\right),\mu_k\right) \to 0$$

and thus

$$d_{\mathrm{HK}}\left(\bar{\mu}\left(\varepsilon\right),\mu^{\dagger}\right) = \lim_{k\to\infty} d_{\mathrm{HK}}\left(\mu_k,\mu^{\dagger}\right) = \sup_{\mu\in\mathfrak{M}(\varepsilon)} d_{\mathrm{HK}}\left(\mu,\mu^{\dagger}\right).$$

□

Using lemma A.2, we conclude the measurability of the worst-case distance.

Proposition A.3. *The function defined in (A.1) is γ_p-measurable. Moreover, there holds*

$$\mathbb{E}_p\left[\sup_{\mu\in\mathfrak{M}(\cdot)} d_{\mathrm{HK}}\left(\mu,\mu^{\dagger}\right)^2\right] \leqslant 2\left\|\mu^{\dagger}\right\|_{\mathcal{M}(\Omega_s)} + \frac{N_0}{2\beta p} < \infty$$

Proof. For abbreviation, define

$$\mathrm{Err}\left[\bar{\mu}\right](\varepsilon) = \sup_{\mu\in\mathfrak{M}(\varepsilon)} d_{\mathrm{HK}}\left(\mu,\mu^{\dagger}\right)$$

and let $\{\varepsilon_k\}_k$ denote a convergent sequence with limit ε. Note that the set $\bigcup_k \mathfrak{M}(\varepsilon_k)$ as well as the sequence $\mathrm{Err}[\bar{\mu}](\varepsilon_k)$ are bounded. Now, choose a subsequence $\{\varepsilon_{k,i}\}_i$ such that

$$\lim_{i\to\infty} \mathrm{Err}\left[\bar{\mu}\right](\varepsilon_{k,i}) = \limsup_{k\to\infty} \mathrm{Err}\left[\bar{\mu}\right](\varepsilon_k)$$

and let $\{\bar{\mu}(\varepsilon_{k,i})\}_i$ be a corresponding sequence of maximizers from lemma A.2. By possibly extracting another subsequence, there is $\tilde{\mu}$ with $\bar{\mu}(\varepsilon_{k,i}) \rightharpoonup^* \tilde{\mu}$. By standard arguments, we verify that $\tilde{\mu}\in\mathfrak{M}(\varepsilon)$. Consequently, we have

$$\limsup_{k\to\infty} \mathrm{Err}\left[\bar{\mu}\right](\varepsilon_k) = \lim_{i\to\infty} \mathrm{Err}\left[\bar{\mu}\right](\varepsilon_{k,i}) = \lim_{i\to\infty} d_{\mathrm{HK}}\left(\mu\left(\varepsilon_{k,i}\right),\mu^{\dagger}\right)^2 = d_{\mathrm{HK}}\left(\tilde{\mu},\mu^{\dagger}\right)^2 \leqslant \mathrm{Err}\left[\bar{\mu}\right](\varepsilon).$$

Hence $\mathrm{Err}[\bar{\mu}]$ is upper semicontinuous and thus measurable w.r.t γ_p. Finally, we apply (A.2) to conclude

$$\mathbb{E}_p\left[\sup_{\mu\in\mathfrak{M}(\cdot)} d_{\mathrm{HK}}\left(\mu,\mu^{\dagger}\right)^2\right] \leqslant 2\left\|\mu^{\dagger}\right\|_{\mathcal{M}(\Omega_s)} + \frac{1}{2\beta}\mathbb{E}_p\left[\|\varepsilon\|^2_{\Sigma_0^{-1}}\right] \leqslant 2\left\|\mu^{\dagger}\right\|_{\mathcal{M}(\Omega_s)} + \frac{N_0}{2\beta p}.$$

□

Appendix B. Proofs of proposition

In this section we provide the omitted proofs of proposition 5.2 and 5.4, respectively, as well as all the auxiliary results needed in their derivation.

Proposition B.1. *The following estimates hold:*

$$\left\|\Sigma_0^{-1/2} G'(\boldsymbol{m})\,\delta\boldsymbol{m}\right\|_2 \leqslant C_k \left\|\delta\boldsymbol{q}/w\right\|_2 + C_k' \sqrt{\|\boldsymbol{q}\|_1}\,\|w\delta\boldsymbol{y}\|_2,$$
$$\left\|\Sigma_0^{-1/2} G''(\boldsymbol{m})\,(\delta\boldsymbol{m},\tau\boldsymbol{m})\right\|_2 \leqslant C_k' \left(\|\delta\boldsymbol{q}/w\|_2\,\|w\tau\boldsymbol{y}\|_2 + \|\tau\boldsymbol{q}/w\|_2\,\|w\delta\boldsymbol{y}\|_2\right) + C_k''\,\|w\delta\boldsymbol{y}\|_2\,\|w\tau\boldsymbol{y}\|_2,$$

where $w_n = \sqrt{|q_n|}$. *In particular, with the W-norm* $\|\delta\boldsymbol{m}\|_W^2 := \|\delta\boldsymbol{q}/w\|_2^2/4 + \|w\delta\boldsymbol{y}\|_2^2$ *with* $W = W(\boldsymbol{m})$, *we have*

$$\left\|\Sigma_0^{-1/2} G'(\boldsymbol{m})\right\|_{W\to 2} \leqslant (2C_k + C_k')\sqrt{\|\boldsymbol{q}\|_1}, \tag{B.1}$$
$$\left\|\Sigma_0^{-1/2} G''(\boldsymbol{m})\right\|_{W\times W\to 2} \leqslant 4C_k' + C_k''. \tag{B.2}$$

Proof. We first notice that $\|v\|_{\Sigma_0^{-1}} \leqslant \|v\|_\infty$ due to $\operatorname{tr}\Sigma_0^{-1} = 1$. In addition, one can write

$$\begin{aligned}
[G'(\boldsymbol{m})\,\delta\boldsymbol{m}]_k &= \sum_{n=1}^{N_s} k(x_k,y_n)\,\delta q_n + (\nabla_y k(x_k,y_n))^\top \delta y_n\, q_n \\
&= \sum_{n=1}^{N_s} k(x_k,y_n)\,w_n\,\delta q_n/w_n + (\nabla_y k(x_k,y_n))^\top w_n \delta y_n\, q_n/w_n \\
[G''(\boldsymbol{m})\,(\delta\boldsymbol{m},\tau\boldsymbol{m})]_k &= \sum_{n=1}^{N_s} (\nabla_y k(x_k,y_n))^\top \delta y_n\,\tau q_n + (\nabla_y k(x_k,y_n))^\top \tau y_n\,\delta q_n \\
&\quad + \delta y_n^\top \nabla_{yy}^2 k(x_k,y_n)\,\tau y_n q_n.
\end{aligned}$$

Here, we choose $w_n = \sqrt{|q_n|}$. Hence, by estimating term by term, we have

$$\|G'(\boldsymbol{m})\,\delta\boldsymbol{m}\|_{\Sigma_0^{-1}} \leqslant C_k\,\|w\|_2\,\|\delta\boldsymbol{q}/w\|_2 + C_k'\,\|\boldsymbol{q}/w\|_2\,\|w\delta\boldsymbol{y}\|_2 \leqslant (2C_k + C_k')\sqrt{\|\boldsymbol{q}\|_1}\,\|\delta\boldsymbol{m}\|_W,$$

which implies (B.1). A similar argument gives (B.2). □

Proposition B.2. *Define the constant*

$$r^\dagger = r(\mu^\dagger) := \min\left\{\min\left\{w_n^\dagger/8,\, d_{w_n^\dagger}\left(y_n^\dagger,\partial\Omega\right)/2\right\} \mid n = 1,\dots,N_s\right\}.$$

Then for every $\boldsymbol{m} \in B_{W_\dagger}(\boldsymbol{m}^\dagger, r^\dagger)$, *there holds* $\operatorname{sign}\boldsymbol{q} = \operatorname{sign}\boldsymbol{q}^\dagger$ *and*

$$R(\boldsymbol{q},\boldsymbol{q}^\dagger) \leqslant 2 \quad \textit{and}\ 1/2\left\|\boldsymbol{q}^\dagger\right\|_1 \leqslant \|\boldsymbol{q}\|_1 \leqslant 2\left\|\boldsymbol{q}^\dagger\right\|_1, \tag{B.3}$$

where $R(\boldsymbol{q},\boldsymbol{q}^\dagger)$ *is the maximal ratio of the weights* w_n *and* $w_n^\dagger$ *from proposition 4.2.*

In addition, for all $\boldsymbol{m},\boldsymbol{m}' \in B_{W_\dagger}(\boldsymbol{m}^\dagger, r^\dagger)$ *and* $\delta\boldsymbol{m}$, *there holds*

$$\left\|\Sigma_0^{-1/2}\left(G(\boldsymbol{m}) - G(\boldsymbol{m}')\right)\right\|_2 \leqslant L_G\,\|\boldsymbol{m} - \boldsymbol{m}'\|_{W_\dagger}, \tag{B.4}$$

$$\left\|\Sigma_0^{-1/2} G'(\boldsymbol{m})\,\delta\boldsymbol{m}\right\|_2 \leqslant L_G \|\delta\boldsymbol{m}\|_{W_\dagger}, \tag{B.5}$$

$$\left\|\Sigma_0^{-1/2} \left(G'(\boldsymbol{m}) - G'(\boldsymbol{m}')\right)\delta\boldsymbol{m}\right\|_2 \leqslant L_{G'} \|\boldsymbol{m}-\boldsymbol{m}'\|_{W_\dagger} \|\delta\boldsymbol{m}\|_{W_\dagger}, \tag{B.6}$$

where $L_G := 4(2C_k + C_k')\sqrt{\|\boldsymbol{q}^\dagger\|_1}$ *and* $L_{G'} := 2(4C_k' + C_k'')$.

Proof. For $\boldsymbol{m} \in B_{W_\dagger}(\boldsymbol{m}^\dagger, r)$, one has

$$|q_n - q_n^\dagger|/w_n^\dagger \leqslant \left\|\left(\boldsymbol{q}-\boldsymbol{q}^\dagger\right)/w^\dagger\right\|_2 \leqslant 4\left\|\boldsymbol{m}-\boldsymbol{m}^\dagger\right\|_{W_\dagger} \leqslant 1/2w_n^\dagger, \quad \forall n = 1,\ldots,N_s.$$

This implies $1/2 \leqslant q_n/q_n^\dagger \leqslant 3/2$ for all $n = 1,2,\ldots,N_s$. Hence, $\operatorname{sign}\boldsymbol{q} = \operatorname{sign}\boldsymbol{q}^\dagger$ and (B.3) follows. Also, the condition $r^\dagger \leqslant d_{w_n^\dagger}(y_n^\dagger, \partial\Omega)/2$ guarantees that $y_n \in \Omega_s$ for all $n = 1,\ldots,N_s$. By proposition (B.1) and (B.3), it can now be seen that

$$\begin{aligned}\left\|\Sigma_0^{-1/2}\left(G(\boldsymbol{m}) - G(\boldsymbol{m}')\right)\right\|_2 &= \left\|\Sigma_0^{-1/2}\int_0^1 G'\left(\boldsymbol{m}' + t(\boldsymbol{m}-\boldsymbol{m}')\right)\mathrm{d}t\,(\boldsymbol{m}-\boldsymbol{m}')\right\|_2 \\ &\leqslant \int_0^1 \left\|\Sigma_0^{-1/2} G'\left(\boldsymbol{m}' + t(\boldsymbol{m}-\boldsymbol{m}')\right)\right\|_{W_\dagger \to 2} \mathrm{d}t\,\|\boldsymbol{m}-\boldsymbol{m}'\|_{W_\dagger},\end{aligned} \tag{B.7}$$

for every $\boldsymbol{m}, \boldsymbol{m}' \in B_{W_\dagger}(\boldsymbol{m}^\dagger, r^\dagger)$. Next, since $\boldsymbol{m}' + t(\boldsymbol{m}-\boldsymbol{m}') \in B_{W_\dagger}(\boldsymbol{m}^\dagger, r^\dagger)$, for $W = W(\boldsymbol{m}' + t(\boldsymbol{m}-\boldsymbol{m}'))$ and $W_\dagger = W_\dagger(\boldsymbol{m}^\dagger)$, we use (4.7), (B.3) and (B.1) to deduce that

$$\begin{aligned}\left\|\Sigma_0^{-1/2} G'\left(\boldsymbol{m}' + t(\boldsymbol{m}-\boldsymbol{m}')\right)\right\|_{W_\dagger \to 2} &\leqslant 2\left\|\Sigma_0^{-1/2} G'\left(\boldsymbol{m}' + t(\boldsymbol{m}-\boldsymbol{m}')\right)\right\|_{W \to 2} \\ &\leqslant 2\left(2C_k + C_k'\right)\sqrt{\|\boldsymbol{q}' + t(\boldsymbol{q}-\boldsymbol{q}')\|_1} \\ &\leqslant 4\left(2C_k + C_k'\right)\sqrt{\|\boldsymbol{q}^\dagger\|_1}.\end{aligned} \tag{B.8}$$

Combining (B.7) and (B.8), we deduce now (B.4). Similarly, (B.5) follows from (B.1) and (B.3) as well. Moreover, (B.6) can be proved using the estimate (B.2) with a similar argument. □

Proof of proposition 5.2. Since $G'(\boldsymbol{m}^\dagger)$ has full column rank, the Fisher information matrix $\mathcal{I}_0$ defined in (5.5) is invertible. Hence, the map $T(\boldsymbol{m}) := \boldsymbol{m} - \mathcal{I}_0^{-1} S(\boldsymbol{m})$ is well-defined, where $S(\boldsymbol{m})$ is the residual of the stationarity equation given in (5.4) with $\bar{\boldsymbol{\rho}} = \boldsymbol{\rho} = \operatorname{sign}\boldsymbol{q}^\dagger$. In order to obtain the claimed results, we aim to show that T is a contraction and argue similarly to the proof of the Banach fixed point theorem. However, since the correct domain of definition for the map T is difficult to determine beforehand, we provide a direct proof.

We start by showing that T is Lipschitz continuous on the ball $B_{W^\dagger}(\boldsymbol{m}^\dagger, \hat{r})$ for some as of yet undetermined $0 < \hat{r} \leqslant r$ with Lipschitz constant $\kappa(\hat{r}) \leqslant 1/2$ if ε is chosen suitably. For this purpose, consider two points $\boldsymbol{m}$ and $\boldsymbol{m}'$ in $B_{W^\dagger}(\boldsymbol{m}^\dagger, \hat{r})$, their difference $\delta\boldsymbol{m} = \boldsymbol{m} - \boldsymbol{m}'$ and the difference of their images $\delta\boldsymbol{m}_T = T(\boldsymbol{m}) - T(\boldsymbol{m}')$. Note that

$$\begin{aligned}\mathcal{I}_0\delta\boldsymbol{m}_T &= \mathcal{I}_0\delta\boldsymbol{m} - \left(S(\boldsymbol{m}) - S(\boldsymbol{m}')\right) \\ &= \left(G'\left(\boldsymbol{m}^\dagger\right) - G'(\boldsymbol{m})\right)^\top \Sigma_0^{-1} G'\left(\boldsymbol{m}^\dagger\right)\delta\boldsymbol{m} \\ &\quad + G'(\boldsymbol{m})^\top \Sigma_0^{-1}\left(G'\left(\boldsymbol{m}^\dagger\right)\delta\boldsymbol{m} - \left(G(\boldsymbol{m}) - G(\boldsymbol{m}')\right)\right) \\ &\quad - \left(G'(\boldsymbol{m}) - G'(\boldsymbol{m}')\right)^\top \Sigma_0^{-1}\left(G(\boldsymbol{m}') - G\left(\boldsymbol{m}^\dagger\right) - \varepsilon\right).\end{aligned}$$

We multiply this equation from the left with $(\delta \boldsymbol{m}_T)^\top$ and consider each term on the right hand side separately. Using proposition B.2, we have for the first term

$$\begin{aligned}
&(\delta \boldsymbol{m}_T)^\top \left(G'\left(\boldsymbol{m}^\dagger\right) - G'(\boldsymbol{m})\right)^\top \Sigma_0^{-1} G'\left(\boldsymbol{m}^\dagger\right) \delta \boldsymbol{m} \\
&\quad = \left(\Sigma_0^{-1/2}\left(G'\left(\boldsymbol{m}^\dagger\right) - G'(\boldsymbol{m})\right)\delta \boldsymbol{m}_T\right)^\top \Sigma_0^{-1/2} G'\left(\boldsymbol{m}^\dagger\right)\delta \boldsymbol{m} \\
&\quad \leqslant \left\|\Sigma_0^{-1/2}\left(G'\left(\boldsymbol{m}^\dagger\right) - G'(\boldsymbol{m})\right)\delta \boldsymbol{m}_T\right\|_2 \left\|\Sigma_0^{-1/2} G'\left(\boldsymbol{m}^\dagger\right)\delta \boldsymbol{m}\right\|_2 \\
&\quad \leqslant L_G L_{G'} \left\|\boldsymbol{m}^\dagger - \boldsymbol{m}\right\|_{W_\dagger} \|\delta \boldsymbol{m}\|_{W_\dagger} \|\delta \boldsymbol{m}_T\|_{W_\dagger}.
\end{aligned}$$

For the second term we estimate

$$\begin{aligned}
&(\delta \boldsymbol{m}_T)^\top G'(\boldsymbol{m})^\top \Sigma_0^{-1}\left(G'\left(\boldsymbol{m}^\dagger\right)\delta \boldsymbol{m} - (G(\boldsymbol{m}) - G(\boldsymbol{m}'))\right) \\
&\quad = \left(\Sigma_0^{-1/2} G'(\boldsymbol{m})\,\delta \boldsymbol{m}_T\right)^\top \Sigma_0^{-1/2} \int_0^1 \left(G'\left(\boldsymbol{m}^\dagger\right) - G'(\tau \boldsymbol{m} + (1-\tau)\boldsymbol{m}')\right)\delta \boldsymbol{m} \mathrm{d}\tau \\
&\quad \leqslant L_G L_{G'} \int_0^1 \left\|\boldsymbol{m}^\dagger - (\tau \boldsymbol{m} + (1-\tau)\boldsymbol{m}')\right\|_{W_\dagger} \mathrm{d}\tau \, \|\delta \boldsymbol{m}\|_{W_\dagger} \|\delta \boldsymbol{m}_T\|_{W_\dagger}
\end{aligned}$$

and for the third term we have

$$\begin{aligned}
&(\delta \boldsymbol{m}_T)^\top \left(G'(\boldsymbol{m}) - G'(\boldsymbol{m}')\right)^\top \Sigma_0^{-1}\left(G(\boldsymbol{m}') - G\left(\boldsymbol{m}^\dagger\right) - \varepsilon\right) \\
&\quad = \left(\Sigma_0^{-1/2}\left(G'(\boldsymbol{m}) - G'(\boldsymbol{m}')\right)\delta \boldsymbol{m}_T\right)^\top \Sigma_0^{-1/2}\left(G(\boldsymbol{m}') - G\left(\boldsymbol{m}^\dagger\right) - \varepsilon\right) \\
&\quad \leqslant L_{G'}\left(L_G \left\|\boldsymbol{m}^\dagger - \boldsymbol{m}'\right\|_{W_\dagger} + \|\varepsilon\|_{\Sigma_0^{-1}}\right)\|\delta \boldsymbol{m}\|_{W_\dagger} \|\delta \boldsymbol{m}_T\|_{W_\dagger}.
\end{aligned}$$

Since $\boldsymbol{m}, \boldsymbol{m}'$ are contained in the ball $B_{W^\dagger}(\boldsymbol{m}^\dagger, \hat{r})$ it follows that

$$\left\|\mathcal{I}_0^{-1}\right\|_{W_\dagger^{-1} \to W_\dagger}^{-1} \|\boldsymbol{m}_T\|_{W_\dagger}^2 \leqslant (\delta \boldsymbol{m}_T)^\top \mathcal{I}_0 \delta \boldsymbol{m}_T \leqslant L_{G'}\left(3 L_G \hat{r} + \|\varepsilon\|_{\Sigma_0^{-1}}\right)\|\delta \boldsymbol{m}_T\|_{W_\dagger} \|\delta \boldsymbol{m}\|_{W_\dagger},$$

using the fact that one has

$$\begin{aligned}
\boldsymbol{m}^\top \mathcal{I}_0 \boldsymbol{m} &= \left(W_\dagger^{1/2} \boldsymbol{m}\right)^\top \left[W_\dagger^{-1/2} \mathcal{I}_0 W_\dagger^{-1/2}\right]\left(W_\dagger^{1/2} \boldsymbol{m}\right) \\
&\geqslant \|\boldsymbol{m}\|_{W_\dagger}^2 \left\|W_\dagger^{1/2} \mathcal{I}_0^{-1} W_\dagger^{1/2}\right\|_{2\to 2}^{-1} = \|\boldsymbol{m}\|_{W_\dagger}^2 \left\|\mathcal{I}_0^{-1}\right\|_{W_\dagger^{-1} \to W_\dagger}^{-1}.
\end{aligned}$$

Dividing by $\|\boldsymbol{m}_T\|_{W_\dagger}$, the estimate

$$\|T(\boldsymbol{m}) - T(\boldsymbol{m}')\|_{W_\dagger} = \|\delta \boldsymbol{m}_T\|_{W_\dagger} \leqslant \kappa(\hat{r}) \|\delta \boldsymbol{m}\|_{W_\dagger} = \kappa(\hat{r}) \|\boldsymbol{m} - \boldsymbol{m}'\|_{W_\dagger}$$

follows with

$$\kappa(\hat{r}) := L_{G'} \left\|\mathcal{I}_0^{-1}\right\|_{W_\dagger^{-1} \to W_\dagger}\left(3 L_G \hat{r} + \|\varepsilon\|_{\Sigma_0^{-1}}\right).$$

The contraction estimate above holds for any $\hat{r} \leqslant r$ under the assumption that the points under consideration lie in the appropriate ball. In order to ensure contraction, we need to establish an appropriate bound and assumptions on the data. For this, we consider the linearized estimate

$$\delta \widehat{\boldsymbol{m}} = -\mathcal{I}_0^{-1} S\left(\boldsymbol{m}^\dagger\right) = \mathcal{I}_0^{-1}\left[G'\left(\boldsymbol{m}^\dagger\right)^\top \Sigma_0^{-1} \varepsilon - \beta(\boldsymbol{\rho}; \mathbf{0})\right],$$

from (5.6). Using the weighted $W_\dagger$-norm defined in proposition B.1, one has

$$\begin{aligned}\|\delta\widehat{\boldsymbol{m}}\|_{W_\dagger} &\leqslant \left\|\mathcal{I}_0^{-1}\right\|_{W_\dagger^{-1}\to W_\dagger}\left(\left\|\left(\Sigma_0^{-1/2}G'\left(\boldsymbol{m}^\dagger\right)\right)^\top\Sigma_0^{-1/2}\varepsilon\right\|_{W_\dagger^{-1}}+\beta\left\|(\boldsymbol{\rho};\mathbf{0})\right\|_{W_\dagger^{-1}}\right)\\ &\leqslant \left\|\mathcal{I}_0^{-1}\right\|_{W_\dagger^{-1}\to W_\dagger}\left(\left\|\left(\Sigma_0^{-1/2}G'\left(\boldsymbol{m}^\dagger\right)\right)^\top\right\|_{2\to W_\dagger^{-1}}\left\|\Sigma_0^{-1/2}\varepsilon\right\|_2+\beta\sqrt{\|\boldsymbol{q}^\dagger\|_1}\right)\\ &\leqslant \left\|\mathcal{I}_0^{-1}\right\|_{W_\dagger^{-1}\to W_\dagger}\left(L_G\left\|\varepsilon\right\|_{\Sigma_0^{-1}}+\beta\sqrt{\|\boldsymbol{q}^\dagger\|_1}\right),\end{aligned}$$

where we have used (B.1) together with $\left\|A^\top\right\|_{2\to W_\dagger^{-1}}=\|A\|_{W_\dagger\to 2}$ and $\|(\boldsymbol{\rho};\mathbf{0})\|_{W_\dagger^{-1}}=\sqrt{\|\boldsymbol{q}^\dagger\|_1}$. In the following, we denote

$$c_1:=\left\|\mathcal{I}_0^{-1}\right\|_{W_\dagger^{-1}\to W_\dagger}\left(L_G+\sqrt{\|\boldsymbol{q}^\dagger\|_1}\right),\quad c_2:=L_{G'}\left\|\mathcal{I}_0^{-1}\right\|_{W_\dagger^{-1}\to W_\dagger}(6L_Gc_1+1).$$

If we now choose $\hat{r}=\min\left\{c_1/c_2,\,r^\dagger/2\right\}$ and assume that

$$\left\|\varepsilon\right\|_{\Sigma_0^{-1}}+\beta\leqslant\frac{\hat{r}}{2c_1}=\min\left\{\frac{1}{2c_2},\,\frac{r^\dagger}{4c_1}\right\},\tag{B.9}$$

then it follows immediately with the previous estimates that

$$\|\delta\widehat{\boldsymbol{m}}\|_{W_\dagger}\leqslant c_1\left(\left\|\varepsilon\right\|_{\Sigma_0^{-1}}+\beta\right)\leqslant\frac{\hat{r}}{2}\quad\text{and }\kappa(\hat{r})\leqslant 1/2.$$

We are now ready to show the existence of a fixed point in $B_{W_\dagger}(\boldsymbol{m}^\dagger,\hat{r})$ as well as the claimed estimates. For this purpose, consider the simplified Gauss-Newton iterative sequence

$$\boldsymbol{m}_0=\boldsymbol{m}^\dagger,\quad \boldsymbol{m}^{k+1}=T\left(\boldsymbol{m}^k\right)=\boldsymbol{m}^k-\mathcal{I}_0^{-1}S\left(\boldsymbol{m}^k\right),\quad k\geqslant 1.\tag{B.10}$$

Put $\delta\boldsymbol{m}^k:=\boldsymbol{m}^k-\boldsymbol{m}^{k-1}$, $k\geqslant 1$. It can be seen that the first Gauss-Newton step is given by $\delta\boldsymbol{m}^1=\delta\widehat{\boldsymbol{m}}$. We use induction to prove that $\boldsymbol{m}^k\in B_{W_\dagger}(\boldsymbol{m}^\dagger,\hat{r})$ for all $k\geqslant 0$. Indeed, if ε satisfies (B.9), we have $\left\|\boldsymbol{m}^1-\boldsymbol{m}^\dagger\right\|_{W_\dagger}=\|\delta\widehat{\boldsymbol{m}}\|_{W_\dagger}\leqslant\hat{r}/2$, which implies $\boldsymbol{m}^1\in B_{W_\dagger}(\boldsymbol{m}^\dagger,\hat{r})$. Assume that $\boldsymbol{m}^k\in B_{W_\dagger}(\boldsymbol{m}^\dagger,\hat{r})$. Notice that it holds $\left\|\delta\boldsymbol{m}^{k+1}\right\|_{W_\dagger}=\left\|T(\boldsymbol{m}^k)-T(\boldsymbol{m}^{k-1})\right\|_{W_\dagger}\leqslant\kappa\left\|\delta\boldsymbol{m}^k\right\|_{W_\dagger}$. Then, with $d^k:=\left\|\delta\boldsymbol{m}^k\right\|_{W_\dagger}$ and $e^k:=\sum_{i=1}^k d^i$ we have

$$d^{k+1}\leqslant\kappa d^k\quad\text{and }e^k\leqslant\frac{1-\kappa^k}{1-\kappa}d^1\leqslant\frac{1}{1-\kappa}d^1.$$

Hence,

$$\left\|\boldsymbol{m}^{k+1}-\boldsymbol{m}^\dagger\right\|_{W_\dagger}\leqslant\sum_{i=1}^{k+1}\left\|\boldsymbol{m}^i-\boldsymbol{m}^{i-1}\right\|_{W_\dagger}=e^k\leqslant\frac{1}{1-\kappa}d^1\leqslant 2\left\|\delta\widehat{\boldsymbol{m}}\right\|_{W_\dagger}\leqslant\hat{r},\tag{B.11}$$

and thus $\boldsymbol{m}^{k+1}\in B_{W_\dagger}(\boldsymbol{m}^\dagger,\hat{r})$. Going to the limit, by standard arguments, we obtain that $\boldsymbol{m}^k\to\widehat{\boldsymbol{m}}\in B_{W_\dagger}(\boldsymbol{m}^\dagger,\hat{r})$ with $T(\widehat{\boldsymbol{m}})=\widehat{\boldsymbol{m}}$ and thus $S(\widehat{\boldsymbol{m}})=0$. Furthermore, by letting $k\to\infty$ in (B.11), we obtain $\left\|\widehat{\boldsymbol{m}}-\boldsymbol{m}^\dagger\right\|_{W_\dagger}\leqslant 2\left\|\delta\widehat{\boldsymbol{m}}\right\|_{W_\dagger}$.

For the second estimate, we rewrite the difference between the error and the perturbation in terms of all updates

$$\widehat{\boldsymbol{m}} - \boldsymbol{m}^{\dagger} - \delta\widehat{\boldsymbol{m}} = \widehat{\boldsymbol{m}} - \boldsymbol{m}^{0} - \delta\boldsymbol{m}^{1} = \sum_{k=2}^{\infty} \delta\boldsymbol{m}^{k}.$$

Now, choosing $\boldsymbol{m} := \boldsymbol{m}^{k+1}$ and $\boldsymbol{m}' := \boldsymbol{m}^{k}$ we have the contraction estimate

$$\left\|\delta\boldsymbol{m}^{k}\right\|_{W_{\dagger}} \leqslant \kappa\left(\tilde{r}\right)\left\|\delta\boldsymbol{m}^{k-1}\right\|_{W_{\dagger}},$$

where $\hat{r}$ is now replaced by $\tilde{r} = \max\{\left\|\boldsymbol{m}^{k+1} - \boldsymbol{m}^{\dagger}\right\|, \left\|\boldsymbol{m}^{k} - \boldsymbol{m}^{\dagger}\right\|\} \leqslant 2d^{1} \leqslant 2c_{1}(\|\varepsilon\|_{\Sigma_{0}^{-1}} + \beta)$ and thus $\kappa(\tilde{r}) \leqslant c_{2}(\|\varepsilon\|_{\Sigma_{0}^{-1}} + \beta)$. Hence, bounding the updates by $\left\|\delta\boldsymbol{m}^{k}\right\|_{W_{\dagger}} = d^{k} \leqslant \kappa(\tilde{r})^{k-1}d^{1} \leqslant (1/2)^{k-2}\kappa(\tilde{r})d^{1}$, we conclude

$$\left\|\widehat{\boldsymbol{m}} - \boldsymbol{m}^{\dagger} - \delta\widehat{\boldsymbol{m}}\right\|_{W_{\dagger}} \leqslant \sum_{k=2}^{\infty} (1/2)^{k-2}\,\kappa\left(\tilde{r}\right) d^{1} \leqslant 2c_{1}c_{2}\left(\|\varepsilon\|_{\Sigma_{0}^{-1}} + \beta\right)^{2}. \tag{B.12}$$

It remains to argue that $\widehat{\boldsymbol{m}}$ is the *unique* stationary point of (5.1) on $B_{W_{\dagger}}(\boldsymbol{m}^{\dagger}, (3/2)\hat{r})$. Replacing $\hat{r}$ with $\tilde{r} = (3/2)\hat{r})$ we still obtain the Lipschitz constant $\kappa((3/2)\hat{r}) \leqslant 3/4$ on the slightly larger ball. Now, assume that $\widetilde{\boldsymbol{m}}$ is any stationary point in the larger ball, thus also fixed point of T and

$$\|\widetilde{\boldsymbol{m}} - \widehat{\boldsymbol{m}}\|_{W_{\dagger}} = \|T(\widetilde{\boldsymbol{m}}) - T(\widehat{\boldsymbol{m}})\|_{W_{\dagger}} \leqslant (3/4)\,\|\widetilde{\boldsymbol{m}} - \widehat{\boldsymbol{m}}\|_{W_{\dagger}},$$

yielding $\widetilde{\boldsymbol{m}} = \widehat{\boldsymbol{m}}$.

□

Remark B.3. Following (B.9) and (B.12), the constant C_1 in the statement of proposition 5.2 can be chosen explicitly as

$$C_{1} = \max\left\{c_{1}, \frac{4c_{1}}{r^{\dagger}}, 2c_{2}\right\}.$$

Next, in order to prove proposition 5.4, we require the following estimates on K^{*}.

Lemma B.4. *Suppose that* $\eta(y) = [K^{*}\Sigma_{0}^{-1}\zeta](y)$ *for* $y \in \Omega_{s}$, $\zeta \in \mathbb{R}^{N_{o}}$. *Then*

$$\sup_{y\in\Omega_{s}} |D\eta\left(y\right)| \leqslant C_{D}\left\|\Sigma_{0}^{-1/2}\zeta\right\|_{2}.$$

for $D \in \{\mathrm{Id}, \nabla, \nabla^{2}, \nabla^{3}\}$ *and* $C_{D} \in \{C_{k}, C_{k}', C_{k}'', C_{k}'''\}$, *respectively.*

Proof. One can see that η can be written as

$$\eta\left(y\right) = \left[K^{*}\Sigma_{0}^{-1}\zeta\right]\left(y\right) = \left(\Sigma_{0}^{-1}\zeta, k\left[\boldsymbol{x}, y\right]\right)_{2} = \sum_{n=1}^{N_{o}} \left(\Sigma_{0}^{-1}\zeta\right)_{n} k\left(x_{n}, y\right). \tag{B.13}$$

Hence, for every $y \in \Omega_{s}$ there holds

$$|\eta\left(y\right)| \leqslant \sum_{n=1}^{N_{o}} |\left(\Sigma_{0}^{-1}\zeta\right)_{n}||k\left(x_{n}, y\right)| \leqslant \sup_{x\in\Omega_{o}, y\in\Omega_{s}} |k\left(x, y\right)| \sum_{n=1}^{N_{o}} |\left(\Sigma_{0}^{-1}\zeta\right)_{n}| = C_{k}\left\|\Sigma_{0}^{-1}\zeta\right\|_{1}.$$

Since $\sum_{n=1}^{N_o}\sigma_{0,n}^{-2}=1$, we have

$$\left\|\Sigma_0^{-1/2}v\right\|_1=\sum_{n=1}^{N_o}\sigma_{0,n}^{-1}v_n\leqslant\sqrt{\sum_{n=1}^{N_o}\sigma_{0,n}^{-2}}\sqrt{\sum_{n=1}^{N_o}|v_n|^2}=\|v\|_2\,,\quad\forall v\in\mathbb{R}^{N_o}$$

Hence, $\left\|\Sigma_0^{-1}\zeta\right\|_1\leqslant\left\|\Sigma_0^{-1/2}\zeta\right\|_2$. From (B.13) the other estimates follow similarly by taking derivatives. □

Proof of proposition 5.4. By the definition of $\widehat{\eta}$, one has

$$\widehat{\eta}\left(\widehat{y}_n\right)=\operatorname{sign}\left(\widehat{q}_n\right)=\operatorname{sign}\left(q_n^\dagger\right)\ \text{and}\ \nabla\widehat{\eta}\left(\widehat{y}_n\right)=0,\quad n=1,\ldots,N_s.$$

We now prove the $\theta/2$-admissibility of $\widehat{\eta}$ if (5.9)–(5.10) hold, namely

$$-\operatorname{sign}\widehat{\eta}\left(\widehat{y}_n\right)\nabla^2\widehat{\eta}\left(\widehat{y}_n\right)\geqslant\theta|w_n^\dagger|^2\,\mathrm{Id},\quad\forall n=1,\ldots,N_s,\tag{B.14}$$

$$|\widehat{\eta}(y)|\leqslant1-(\theta/2)^2,\quad\forall y\in\Omega_s\setminus\bigcup_{n=1,\ldots,N_s}B_{w_n^\dagger}\left(\widehat{y}_n,\sqrt{\theta/2}\right)\tag{B.15}$$

Compare this to (3.9)–(3.10) for η_{PC}. To this end, consider the noisy pre-certificate

$$\begin{aligned}\eta_{\mathrm{PC},\varepsilon}&:=-\beta^{-1}K^*\Sigma_0^{-1}\left(G'\left(\boldsymbol{m}^\dagger\right)\delta\widehat{\boldsymbol{m}}-\varepsilon\right)\\&=\beta^{-1}K^*\Sigma_0^{-1}\left[G'\left(\boldsymbol{m}^\dagger\right)\mathcal{I}_0^{-1}\left(\beta(\boldsymbol{\rho};\mathbf{0})-G'\left(\boldsymbol{m}^\dagger\right)^\top\Sigma_0^{-1}\varepsilon\right)+\varepsilon\right]\\&=\eta_{\mathrm{PC}}-\beta^{-1}K^*\Sigma_0^{-1/2}\left[\Sigma_0^{-1/2}G'\left(\boldsymbol{m}^\dagger\right)\mathcal{I}_0^{-1}G'\left(\boldsymbol{m}^\dagger\right)^\top\Sigma_0^{-1/2}-\mathrm{Id}\right]\left(\Sigma_0^{-1/2}\varepsilon\right)\\&=\eta_{\mathrm{PC}}-\beta^{-1}K^*\Sigma_0^{-1/2}\left[P-\mathrm{Id}\right]\left(\Sigma_0^{-1/2}\varepsilon\right),\end{aligned}\tag{B.16}$$

where η_{PC} is given in (5.8) and P is an orthogonal projection to the $N_s-N_o(1+d)$ dimensional range of $\Sigma_0^{-1/2}G'(\boldsymbol{m}^\dagger)$. This implies

$$\begin{aligned}\widehat{\eta}&=-\beta^{-1}K^*\Sigma_0^{-1}\left(G(\widehat{\boldsymbol{m}})-G\left(\boldsymbol{m}^\dagger\right)-\varepsilon\right)\\&=\eta_{\mathrm{PC},\varepsilon}-\beta^{-1}K^*\Sigma_0^{-1}\left(G(\widehat{\boldsymbol{m}})-G\left(\boldsymbol{m}^\dagger\right)-G'\left(\boldsymbol{m}^\dagger\right)\delta\widehat{\boldsymbol{m}}\right)\\&=\eta_{\mathrm{PC}}-\beta^{-1}\left[\underbrace{K^*\Sigma_0^{-1/2}\left[P-\mathrm{Id}\right]\left(\Sigma_0^{-1/2}\varepsilon\right)}_{e_1}-\underbrace{K^*\Sigma_0^{-1}\left(G(\widehat{\boldsymbol{m}})-G\left(\boldsymbol{m}^\dagger\right)-G'\left(\boldsymbol{m}^\dagger\right)\delta\widehat{\boldsymbol{m}}\right)}_{e_2}\right].\end{aligned}$$

Applying lemma B.4, we have

$$\begin{aligned}\|e_1\|_{\mathcal{C}(\Omega_s)}&\leqslant C_k\left\|\Sigma_0^{-1/2}\left[P-\mathrm{Id}\right]\Sigma_0^{-1/2}\varepsilon\right\|_1\\&\leqslant C_k\left\|\left[P-\mathrm{Id}\right]\Sigma_0^{-1/2}\varepsilon\right\|_2\leqslant C_k\|\varepsilon\|_{\Sigma_0^{-1}}.\end{aligned}$$

In order to estimate e_2, we apply lemma B.4 and proposition B.1 to have

$$\begin{aligned}\|e_2\|_{\mathcal{C}(\Omega_s)}&\leqslant C_k\left\|\Sigma_0^{-1}\left(G(\widehat{\boldsymbol{m}})-G\left(\boldsymbol{m}^\dagger\right)-G'\left(\boldsymbol{m}^\dagger\right)\delta\widehat{\boldsymbol{m}}\right)\right\|_1\\&\leqslant C_k\left\|\Sigma_0^{-1/2}\left(G(\widehat{\boldsymbol{m}})-G\left(\boldsymbol{m}^\dagger\right)-G'\left(\boldsymbol{m}^\dagger\right)\delta\widehat{\boldsymbol{m}}\right)\right\|_2.\end{aligned}\tag{B.17}$$

Notice that

$$G(\widehat{\boldsymbol{m}}) - G(\boldsymbol{m}^\dagger) - G'(\boldsymbol{m}^\dagger)\,\delta\widehat{\boldsymbol{m}} = G'(\boldsymbol{m}^\dagger)(\widehat{\boldsymbol{m}} - \boldsymbol{m}^\dagger - \delta\widehat{\boldsymbol{m}}) + \int_0^1 \left(G'(\boldsymbol{m}_\tau) - G'(\boldsymbol{m}^\dagger)\right)(\widehat{\boldsymbol{m}} - \boldsymbol{m}^\dagger)\,\mathrm{d}\tau$$

where $\boldsymbol{m}_\tau = \boldsymbol{m}^\dagger + \tau(\widehat{\boldsymbol{m}} - \boldsymbol{m}^\dagger)$. Using this together with propositions B.2 and 5.2, we have

$$\begin{aligned}\|e_2\|_{\mathcal{C}(\Omega_s)} &\leqslant C_k(\|\Sigma_0^{-1/2}G'(\boldsymbol{m}^\dagger)(\widehat{\boldsymbol{m}} - \boldsymbol{m}^\dagger - \delta\widehat{\boldsymbol{m}})\|_2 \\ &\quad + \int_0^1 \|\Sigma_0^{-1/2}(G'(\boldsymbol{m}_\tau) - G'(\boldsymbol{m}^\dagger))(\widehat{\boldsymbol{m}} - \boldsymbol{m}^\dagger)\|_2 \mathrm{d}\tau) \\ &\leqslant C_k(L_G\|\widehat{\boldsymbol{m}} - \boldsymbol{m}^\dagger - \delta\widehat{\boldsymbol{m}}\|_{W_\dagger} + L_{G'}\|\widehat{\boldsymbol{m}} - \boldsymbol{m}^\dagger\|_{W_\dagger}^2) \\ &\leqslant C_k(L_G + L_{G'})C_1^2(\|\varepsilon\|_{\Sigma_0^{-1}} + \beta)^2.\end{aligned} \tag{B.18}$$

Combining (B.16)–(B.18), we have

$$\|\widehat{\eta} - \eta_{\mathrm{PC}}\|_{\mathcal{C}(\Omega_s)} \leqslant c_3\beta^{-1}\left[\left(\|\varepsilon\|_{\Sigma_0^{-1}} + \beta\right)^2 + \|\varepsilon\|_{\Sigma_0^{-1}}\right],$$

where $c_3 := C_k((L_G + L_{G'})C_1^2 + 1)$. This yields

$$\begin{aligned}|\widehat{\eta}(y)| &\leqslant |\widehat{\eta}(y) - \eta_{\mathrm{PC}}(y)| + |\eta_{\mathrm{PC}}(y)| \\ &\leqslant \|\widehat{\eta} - \eta_{\mathrm{PC}}\|_{\mathcal{C}(\Omega_s)} + |\eta_{\mathrm{PC}}(y)| \\ &\leqslant c_3\beta^{-1}\left[\left(\|\varepsilon\|_{\Sigma_0^{-1}} + \beta\right)^2 + \|\varepsilon\|_{\Sigma_0^{-1}}\right] + |\eta_{\mathrm{PC}}(y)|.\end{aligned} \tag{B.19}$$

We first prove (B.15). Assume that (5.9) holds. Using proposition 5.2, we know that for $y \in \Omega_s \setminus \bigcup_{n=1,\ldots,N_s} B_{w_n^\dagger}(\widehat{y}_n, \sqrt{\theta/2})$, there holds

$$\left\|w_n^\dagger(y - y_n^\dagger)\right\|_2 \geqslant \left\|w_n^\dagger(y - \widehat{y}_n)\right\|_2 - \left\|\widehat{\boldsymbol{m}} - \boldsymbol{m}^\dagger\right\|_{W_\dagger} \geqslant \sqrt{\theta/2} - \sqrt{\theta/32} = \sqrt{9\theta/32}.$$

Hence, since η_{PC} is non-degenerate, we have by (3.8) that

$$|\eta_{\mathrm{PC}}(y)| \leqslant 1 - \theta\min\left\{\theta, \left\|w_n^\dagger(y - y_n^\dagger)\right\|_2^2\right\} \leqslant 1 - \theta\min\{\theta, 9\theta/32\} = 1 - 9\theta^2/32.$$

This, (B.19) and condition (5.10) with $C_2 = c_3$ imply that

$$|\widehat{\eta}(y)| \leqslant \theta^2/32 + \left(1 - 9\theta^2/32\right) = 1 - (\theta/2)^2, \quad \text{for every } y \in \Omega_s \setminus \bigcup_{n=1,\ldots,N_s} B_{w_n^\dagger}\left(\widehat{y}_n, \sqrt{\theta/2}\right),$$

which is indeed (B.15).

We next prove (B.14). Following the same arguments as for (B.17)–(B.18) together with lemma B.4, we have for $c_3'' := C_k''((L_G + L_{G'})C_1^2 + 1)$ that

$$\sup_{y\in\Omega_s}\left\|\nabla^2\widehat{\eta} - \nabla^2\eta_{\mathrm{PC}}\right\|_{2\to 2} \leqslant c_3''\beta^{-1}\left[\left(\|\varepsilon\|_{\Sigma_0^{-1}} + \beta\right)^2 + \|\varepsilon\|_{\Sigma_0^{-1}}\right]. \tag{B.20}$$

In addition, by invoking assumption A1 on the boundedness of the third derivative of k, we obtain

$$\begin{aligned}\left\|\nabla^2\eta_{\mathrm{PC}}\left(\widehat{y}_n\right)-\nabla^2\eta_{\mathrm{PC}}\left(y_n^\dagger\right)\right\|_{2\to 2} &\leqslant \left\|\widehat{y}_n-y_n^\dagger\right\|_2 \sup_{y\in\Omega_s}\left\|\nabla^3\eta_{\mathrm{PC}}\left(y\right)\right\|_{2\times 2\to 2}\\ &\leqslant |w_n^\dagger|^{-1}\left\|w_n^\dagger\left(\widehat{y}_n-y_n^\dagger\right)\right\|_2 C_k'''\left\|\Sigma_0^{-1/2}G'\left(\boldsymbol{m}^\dagger\right)\mathcal{I}_0^{-1}\left(\boldsymbol{\rho};\mathbf{0}\right)\right\|_2\\ &\leqslant |w_n^\dagger|^{-1}\left\|\widehat{\boldsymbol{m}}-\boldsymbol{m}^\dagger\right\|_{W_\dagger} C_k'''L_G\left\|\mathcal{I}_0^{-1}\right\|_{W_\dagger^{-1}\to W_\dagger}\left\|\left(\boldsymbol{\rho};\mathbf{0}\right)\right\|_{W_\dagger}\\ &\leqslant |w_n^\dagger|^{-1}C_k'''L_G\sqrt{\left\|\boldsymbol{q}^\dagger\right\|_1}\left\|\mathcal{I}_0^{-1}\right\|_{W_\dagger^{-1}\to W_\dagger} C_1\left(\|\varepsilon\|_{\Sigma_0^{-1}}+\beta\right)\\ &\leqslant c_4\beta^{-1}\left[\left(\|\varepsilon\|_{\Sigma_0^{-1}}+\beta\right)^2+\|\varepsilon\|_{\Sigma_0^{-1}}\right],\end{aligned} \tag{B.21}$$

with $c_4 := \max_n |w_n^\dagger|^{-1}C_k'''L_G\sqrt{\|\boldsymbol{q}^\dagger\|_1}\left\|\mathcal{I}_0^{-1}\right\|_{W_\dagger^{-1}\to W_\dagger} C_1$.

From (B.20)–(B.21), we have

$$\begin{aligned}\left\|\nabla^2\widehat{\eta}\left(\widehat{y}_n\right)-\nabla^2\eta_{\mathrm{PC}}\left(y_n^\dagger\right)\right\|_{2\to 2} &= \left\|\nabla^2\widehat{\eta}\left(\widehat{y}_n\right)-\nabla^2\eta_{\mathrm{PC}}\left(\widehat{y}_n\right)\right\|_{2\to 2}+\left\|\nabla^2\eta_{\mathrm{PC}}\left(\widehat{y}_n\right)-\nabla^2\eta_{\mathrm{PC}}\left(y_n^\dagger\right)\right\|_{2\to 2}\\ &\leqslant \left(c_3''+c_4\right)\beta^{-1}\left[\left(\|\varepsilon\|_{\Sigma_0^{-1}}+\beta\right)^2+\|\varepsilon\|_{\Sigma_0^{-1}}\right]\end{aligned}$$

for every $n=1,\ldots,N_s$. If we set $C_2=(c_3''+c_4)\max_n|w_n^\dagger|^{-2}$ and require

$$C_2\beta^{-1}\left[\left(\|\varepsilon\|_{\Sigma_0^{-1}}+\beta\right)^2+\|\varepsilon\|_{\Sigma_0^{-1}}\right]\leqslant\theta^2/32$$

and that η_{PC} is $\theta-$admissible with $0<\theta\leqslant 1$, we have with (3.9) for η_{PC} for any vector ξ that

$$\begin{aligned}-\mathrm{sign}\widehat{\eta}\left(\widehat{y}_n\right)\xi^\top\nabla^2\widehat{\eta}\left(\widehat{y}_n\right)\xi &\geqslant -\mathrm{sign}\eta_{\mathrm{PC}}\left(y_n^\dagger\right)\xi^\top\nabla^2\eta_{\mathrm{PC}}\left(y_n^\dagger\right)\xi-\|\xi\|_2^2\left\|\nabla^2\widehat{\eta}\left(\widehat{y}_n\right)-\nabla^2\eta_{\mathrm{PC}}\left(y_n^\dagger\right)\right\|_{2\to 2}\\ &\geqslant 2\theta|w_n^\dagger|^2\|\xi\|_2^2-\theta^2/32|w_n^\dagger|^2\|\xi\|_2^2\geqslant\theta|w_n^\dagger|^2\|\xi\|_2^2\end{aligned}$$

and $\widehat{\eta}$ satisfies (B.14). Hence, we conclude that $\widehat{\eta}$ is $\theta/2$-admissible for $\widehat{\mu}$.

□

Remark B.5. In fact, the constant C_2 in the proof of proposition 5.4 can be chosen as

$$C_2=\max\left\{c_3,(c_3''+c_4)\max_n|w_n^\dagger|^{-2}\right\}.$$

Since these constants depend monotonically on $\left\|\mathcal{I}_0^{-1}\right\|_{W_\dagger^{-1}\to W_\dagger}$, we also have the monotone dependence of C_2 on $\left\|\mathcal{I}_0^{-1}\right\|_{W_\dagger^{-1}\to W_\dagger}$.

Appendix C. Discussion on possible distance candidates

The Hellinger–Kantorovich distance introduced in this paper turns out to be a suitable distance to quantify the reconstruction. Nevertheless, it is worth mentioning that other choices of distances are possible, for instance the Kantorovich–Rubinstein distance.

C.1. Kantorovich-Rubinstein distance

The distance induced by the Kantorovich–Rubinstein (KR) norm (equivalent to the 'Bounded-Lipschitz' norm) is also referred to as the 'flat' metric, and metricises weak* convergence on bounded sets in $\mathcal{M}(\Omega_s)$. The norm can be defined by

$$\|\mu\|_{\mathrm{KR}} := \sup\left\{\int_{\Omega_s} f \mathrm{d}\mu : f \in C^{0,1}(\Omega_s), \|f\|_{C(\Omega_s)} \leqslant 1 \text{ and } \mathrm{Lip}(f) \leqslant 1\right\}.$$

Here, $C^{0,1}(\Omega_s)$ is the space of Lipschitz functions on Ω_s and

$$\mathrm{Lip}(f) := \sup_{x \neq y} \frac{\|f(x) - f(y)\|}{\|x - y\|}, \quad x, y \in \Omega_s.$$

The KR distance is then set to

$$d_{\mathrm{KR}}(\mu_1, \mu_2) = \|\mu_1 - \mu_2\|_{\mathrm{KR}}.$$

Although its evaluation for general sparse measures requires the solution of a minimization problem (see [21]) it has many useful properties. For positive measures $\mu_1, \mu_2 \geqslant 0$ it is equivalent to a generalized Wasserstein-1 distance for measures not necessarily of the same TV-norm as in [28];

$$d_{\mathrm{KR}}(\mu_1, \mu_2) = \inf_{\left\{\tilde{\mu}_1, \tilde{\mu}_2 : \ \|\tilde{\mu}_1\|_{\mathcal{M}} = \|\tilde{\mu}_2\|_{\mathcal{M}}\right\}} \left[\|\tilde{\mu}_1 - \mu_1\|_{\mathcal{M}} + \|\tilde{\mu}_2 - \mu_2\|_{\mathcal{M}} + W_1(\tilde{\mu}_1, \tilde{\mu}_2)\right]. \tag{C.1}$$

For signed measures, we use the Jordan decomposition $\mu_i = \mu_i^+ - \mu_i^-$ and observe that

$$d_{\mathrm{KR}}(\mu_1, \mu_2) = \left\|\left(\mu_1^+ + \mu_2^-\right) - \left(\mu_2^+ + \mu_1^-\right)\right\|_{\mathrm{KR}} = d_{\mathrm{KR}}\left(\mu_1^+ + \mu_2^-, \mu_2^+ + \mu_1^-\right),$$

which then allows to apply the characterization (C.1). This representation, together with the help of [28, proposition 2] allows to characterize the KR distance of single point sources with weight of equal sign:

$$\begin{aligned} d_{\mathrm{KR}}(q_1\delta_{y_1}, q_2\delta_{y_2}) &= \min_{0 \leqslant \theta \operatorname{sign} q_1 \leqslant \min\{|q_1|, |q_2|\}} \left[|\theta - q_1| + |\theta - q_2| + \theta \|y_1 - y_2\|_2\right] \\ &= |q_1 - q_2| + \min\{|q_1|, |q_2|\} \min\{\|y_1 - y_2\|_2, 2\}. \end{aligned} \tag{C.2}$$

For the case of $\operatorname{sign} q_1 \neq \operatorname{sign} q_2$, we instead have $d_{\mathrm{KR}}(q_1\delta_{y_1}, q_2\delta_{y_2}) = |q_1| + |q_2|$. The above formula can be used for all finitely supported measures with the same number of support points by applying the triangle inequality. Motivated by property (C.2), the Kantorovich-Rubinstein distance d_{KR} is also appropriate to quantify the distance between two discrete measures, and a similar upper bound as in proposition 4.2 can be obtained, i.e.

$$d_{\mathrm{KR}}\left(\mu, \mu^\dagger\right) \leqslant \sum_{n=1}^{N} \left(|q_n - q_n^\dagger| + |q_n^\dagger| \left\|y_n - y_n^\dagger\right\|_2\right) \leqslant 2\sqrt{2\|\mu^\dagger\|} \left\|\boldsymbol{m} - \boldsymbol{m}^\dagger\right\|_{W_\dagger},$$

However, this bound either uses an ℓ_1 like sum over the weighted errors, which is not suitable for the rest of our analysis, or the equivalence between a weighted ℓ_1 and ℓ_2 norm (by Hölder's inequality), and is not an asymptotically sharp bound.

Appendix D. Notation table

Ω_s, Ω_o	location and observation set, both compact
$N_s^\dagger$, $y_n^\dagger$, $q_n^\dagger$	Unknown number, positions and coefficients of ground truth sources
$\boldsymbol{y}^\dagger$, $\boldsymbol{q}^\dagger$, $\boldsymbol{m}^\dagger$	Concatenated source/measurement locations, coefficients, $\boldsymbol{m}^\dagger = (\boldsymbol{y}^\dagger; \boldsymbol{q}^\dagger)$
$w^\dagger$, $W^\dagger$	weight vector and weight matrix induced by $\boldsymbol{q}^\dagger$, (4.6) et sqq.
N_o, x_j	Given number and locations of measurements
k	Integral kernel, see section 2.2.1
$\nabla_y k$, $\nabla_{yy}^2 k$, $\nabla_{yyy}^3 k$	(Higher-order) partial derivatives of k w.r.t y
C_k, C_k', C_k'', C_k'''	Bounds on k, $\nabla_y k$, $\nabla_{yy}^2 k$, $\nabla_{yyy}^3 k$, see assumption A1
$k[\boldsymbol{x},\boldsymbol{y}]$, $k[x,\boldsymbol{y}]$	Evaluation of $k(\cdot,y)$, $k(x,\cdot)$ along $\boldsymbol{x}$, $\boldsymbol{y}$, respectively, (2.1) and (2.2)
$\nabla_y^\top k[x,\boldsymbol{y}]$	Evaluation of $\nabla_y k(x,\cdot)^\top$ along $\boldsymbol{y}$, (2.4)
$k[\boldsymbol{x},\boldsymbol{y}]$, $\nabla_y^\top k[\boldsymbol{x},\boldsymbol{y}]$	Evaluation of $k[\cdot,\boldsymbol{y}]$, $\nabla_y^\top k[\cdot,\boldsymbol{y}]$ along $\boldsymbol{x}$, (2.3) and (2.5)
K, K^*	Source-to-measurements operator, (2.6). (pre)-adjoint $K = (K^*)^*$, (2.7)
ε	Measurement noise, deterministic or random
$z^d(\varepsilon)$	Observed measurements given noise ε, (3.1)
$\mathcal{M}(\Omega_s)$, $\|\cdot\|_{\mathcal{M}(\Omega_s)}$	Space of Radon measures on Ω_s and associated norm, section 2.2.2
$\mu^\dagger$	Sparse ground truth measure, (1.2)
$(\mathcal{P}_0)$, $(\mathcal{P}_{\beta,\varepsilon})$	Minimum norm problem, regularized problem
$\mathfrak{M}(\varepsilon)$	Solution set of problem $(\mathcal{P}_{\beta,\varepsilon})$
$\beta(\varepsilon)$, $\beta(p)$, β_0	Parameter choice rules, $\beta(p) = \beta_0/\sqrt{p}$
$\eta^\dagger$, η_{PC}, $\bar{\eta}$	Minimum norm dual certificate, (3.6), vanishing derivative pre-certificate, (3.7) dual certificate, dual certificate of problem $(\mathcal{P}_{\beta,\varepsilon})$
θ	Non-degeneracy parameter, definition 3.8
d_{TV}, d_{KR}, d_{HK}	Total variation, Kantorovich–Rubinstein, Hellinger Kantorovich metrics
$\mathcal{I}_0$, ρ	Fisher information and sign vector, (1.5)
p, Σ_0	Overall precision of measurements, normalized covariance matrix
Σ, γ_p	Parametrized covariance matrix $\Sigma = p^{-1}\Sigma_0$ and Gaussian $\gamma_p = \mathcal{N}(0,\Sigma)$
$G(\boldsymbol{m})$	Measurements $k[\boldsymbol{x},\boldsymbol{y}]\boldsymbol{q}$ given $\boldsymbol{m} = (\boldsymbol{y};\boldsymbol{q})$, (5.2)
$\widehat{\boldsymbol{m}}(\varepsilon)$, $\delta\widehat{\boldsymbol{m}}(\varepsilon)$	Stationary point of (5.1), linear approximation of $\widehat{\boldsymbol{m}}(\varepsilon)$, (5.6).

ORCID iDs

Phuoc-Truong Huynh https://orcid.org/0000-0002-8072-8530
Konstantin Pieper https://orcid.org/0000-0002-4982-2588
Daniel Walter https://orcid.org/0000-0003-3647-0728

References

[1] Azaïs J-M, de Castro Y and Gamboa F 2015 Spike detection from inaccurate samplings *Appl. Comput. Harmon. Anal.* **38** 177–95
[2] Bach F 2017 Breaking the curse of dimensionality with convex neural networks *J. Mach. Learn. Res.* **18** 53
[3] Boyd N, Schiebinger G and Recht B 2017 The alternating descent conditional gradient method for sparse inverse problems *SIAM J. Optim.* **27** 616–39
[4] Bredies K and Pikkarainen H K 2012 Inverse problems in spaces of measures *ESAIM Control Optim. Calc. Var.* **19** 190–218
[5] Candès E J and Fernandez-Granda C 2013 Super-resolution from noisy data *J. Fourier Anal. Appl.* **19** 1229–54
[6] Candès E J and Fernandez-Granda C 2014 Towards a mathematical theory of super-resolution *Commun. Pure Appl. Math.* **67** 906–56

[7] Casas E and Kunisch K 2019 Using sparse control methods to identify sources in linear diffusion-convection equations *Inverse Problems* **35** 114002
[8] Casas E, Vexler B and Zuazua E 2015 Sparse initial data identification for parabolic PDE and its finite element approximations *Math. Control Relat. Fields* **5** 377–99
[9] Chizat L 2022 Sparse optimization on measures with over-parameterized gradient descent *Math. Program.* **194** 487–532
[10] Denoyelle Q, Duval V, Peyré G and Soubies E 2020 The sliding Frank-Wolfe algorithm and its application to super-resolution microscopy *Inverse Problems* **36** 014001
[11] Duval V and Peyré G 2014 Exact support recovery for sparse spikes deconvolution *Found. Comput. Math.* **15** 1315–55
[12] Engel S, Hafemeyer D, Münch C and Schaden D 2019 An application of sparse measure valued Bayesian inversion to acoustic sound source identification *Inverse Problems* **35** 33
[13] Fernandez-Granda C 2013 Support detection in super-resolution (arXiv:1302.3921)
[14] Flinth A, de Gournay F and Weiss P 2021 On the linear convergence rates of exchange and continuous methods for total variation minimization *Math. Program.* **190** 221–57
[15] Fukushima M, Oshima Y and Takeda M 2010 *Dirichlet Forms and Symmetric Markov Processes* (De Gruyter) (https://doi.org/10.1515/9783110218091)
[16] Gerth D, Hofinger A and Ramlau R 2017 On the lifting of deterministic convergence rates for inverse problems with stochastic noise *Inverse Problems Imaging* **11** 663–87
[17] Grisvard P 2011 *Elliptic Problems in Nonsmooth Domains* (Society for Industrial and Applied Mathematics) (https://doi.org/10.1137/1.9781611972030)
[18] Haber E, Magnant Z, Lucero C and Tenorio L 2012 Numerical methods for *A*-optimal designs with a sparsity constraint for ill-posed inverse problems *Comput. Optim. Appl.* **52** 293–314
[19] Hofmann B, Kaltenbacher B, Pöschl C and Scherzer O 2007 A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators *Inverse Problems* **23** 987–1010
[20] Kunisch K, Trautmann P and Vexler B 2016 Optimal control of the undamped linear wave equation with measure valued controls *SIAM J. Control Optim.* **54** 1212–44
[21] Lellmann J, Lorenz D A, Schönlieb C and Valkonen T 2014 Imaging with Kantorovich–Rubinstein discrepancy *SIAM J. Imaging Sci.* **7** 2833–59
[22] Leykekhman D, Vexler B and Walter D 2020 Numerical analysis of sparse initial data identification for parabolic problems *ESAIM, Math. Model. Numer. Anal.* **54** 1139–80
[23] Li Q, Prater A, Shen L and Tang G 2015 Overcomplete tensor decomposition via convex optimization *2015 IEEE 6th Int. Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)* (IEEE) pp 53–56
[24] Liero M, Mielke A and Savaré G 2018 Optimal entropy-transport problems and a new Hellinger-Kantorovich distance between positive measures *Invent. Math.* **211** 969–1117
[25] McCutchen C W 1967 Superresolution in microscopy and the abbe resolution limit *J. Opt. Soc. Am.* **57** 1190–2
[26] Neitzel I, Pieper K, Vexler B and Walter D 2019 A sparse control approach to optimal sensor placement in PDE-constrained parameter estimation problems *Numer. Math.* **143** 943–84
[27] Peyré G and Cuturi M 2019 Computational optimal transport: with applications to data science *Found. Trends Mach. Learn.* **11** 355–607
[28] Piccoli B and Rossi F 2016 On properties of the generalized Wasserstein distance *Arch. Ration. Mech. Anal.* **222** 1339–65
[29] Pieper K and Petrosyan A 2022 Nonconvex regularization for sparse neural networks *Appl. Comput. Harmon. Anal.* **61** 25–56
[30] Pieper K, Tang B Q, Trautmann P and Walter D 2020 Inverse point source location with the Helmholtz equation on a bounded domain *Comput. Optim. Appl.* **77** 213–49
[31] Pieper K and Walter D 2021 Linear convergence of accelerated conditional gradient algorithms in spaces of measures *ESAIM Control Optim. Calc. Var.* **27** 37
[32] Pinelis I F and Sakhanenko A I 1986 Remarks on inequalities for large deviation probabilities *Theory Probab. Appl.* **30** 143–8
[33] Poon C, Keriven N and Peyré G 2019 Support localization and the Fisher metric for off-the-grid sparse regularization *Proc. 22nd Int. Conf. on Artificial Intelligence and Statistics* (*Proc. Machine Learning Research* vol 89) (*16–18 April 2019*) ed K Chaudhuri and M Sugiyama (PMLR) pp 1341–50

[34] Poon C, Keriven N and Peyré G 2023 The geometry of off-the-grid compressed sensing *Found. Comput. Math.* **23** 241–327
[35] Puschmann K G and Kneer F 2005 On super-resolution in astronomical imaging *Astron. Astrophys.* **436** 373–8
[36] Uciński D 2005 *Optimal Measurement Methods for Distributed Parameter System Identification* (CRC Press) (https://doi.org/10.1201/9780203026786)
[37] Werner F and Hohage T 2012 Convergence rates in expectation for Tikhonov-type regularization of inverse problems with Poisson data *Inverse Problems* **28** 104004