Data-Driven Priors in the Maximum Entropy on the Mean Method for Linear Inverse Problems.

Matthew King-Roskamp, Rustum Choksi, and Tim Hoheisel

Abstract. We establish the theoretical framework for implementing the maximum entropy on the mean (MEM) method for linear inverse problems in the setting of approximate (data-driven) priors. We prove a.s. convergence for empirical means and further develop general estimates for the difference between the MEM solutions with different priors μ and ν based upon the epigraphical distance between their respective log-moment generating functions. These estimates allow us to establish a rate of convergence in expectation for empirical means. We illustrate our results with denoising on MNIST and Fashion-MNIST data sets.

1. Introduction. Linear inverse problems are pervasive in data science. A canonical example (and our motivation here) is denoising and deblurring in image processing. Machine learning algorithms, particularly neural networks trained on large data sets, have proven to be a game changer in solving these problems. However, most machine learning algorithms suffer from the lack of a foundational framework upon which to rigorously assess their performance. Thus, there is a need for mathematical models which are on one end, data driven, and on the other end, open to rigorous evaluation. In this article, we devise and analyze one such model based upon what is known as *Maximum Entropy on the Mean* (MEM) (described in some detail in subsection 1.2).

1.1. History and state of the art of the MEM method. Emerging from ideas of E.T. Jaynes in 1957 [23, 24], various forms and interpretations of MEM (see [7, 20, 30, 13, 29]) have appeared in the literature. Applications have occurred in different disciplines such as earth sciences [17, 33, 43], crystallography [31, 32], and medical imaging [1, 10, 11, 21, 22]. Recently, the MEM method has been shown to be a powerful tool for blind deblurring of images that possess some form of symbology (for example, UPC and QR barcodes) [35, 34]. However, MEM methods are not widely used and have yet to become a modern tool for solving contemporary data-driven inverse problems in image processing and machine learning.

MEM methods require problem-specific knowledge in the form of a *statistical prior*. For a prior based upon a known distribution, the theory is well understood, with dedicated algorithms for its implementation [44]. This work is the first to incorporate data in a systematic way into the MEM framework for linear inverse problems. While the incorporation is natural, we provide theoretical guarantees of convergence and upper bounds on rates of convergence without requiring model assumptions. In addition to providing the theoretical framework, we present several numerical examples for denoising images from *MNIST* [15] and *Fashion-MNIST* [47] data sets; showcasing that our work results in a data-driven model with numerical implementation via standard optimization routines.

1.2. Brief overview of the MEM method. Let us now provide some details, summarizing the MEM method for linear inverse problems. Full details will be provided in the next section. Our canonical inverse problem takes the following form

$$b = C\overline{x} + \eta. \tag{1.1}$$

The unknown solution $\overline{x}$ is a vector in $\mathbb{R}^d$; the observed data is $b \in \mathbb{R}^m$; $C \in \mathbb{R}^{m\times d}$, and $\eta \sim \mathcal{Z}$ is an random noise vector in $\mathbb{R}^m$ drawn from noise distribution $\mathcal{Z}$. In the setting of image processing, $\overline{x}$ denotes the ground truth image with d pixels, C is a blurring matrix with typically $d = m$, and the observed noisy (and blurred image) is b. For known C, we seek to recover the ground truth $\overline{x}$ from b. In certain classes of images, the case where C is also unknown (blind deblurring) can also be solved with the MEM framework (cf. [35, 34]) but we will not focus on this here. In fact, our numerical experiments will later focus purely on denoising, i.e., $C = I$. The power of MEM is to exploit the fact that there exists a prior distribution μ for the space of admissible ground truths.

The basis of the method is the *MEM function* $\kappa_\mu : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ defined as

$$\kappa_\mu(x) := \inf \{\mathrm{KL}(Q\|\mu) \ : \ Q \in \mathcal{P}(\mathcal{X}), \mathbb{E}_Q = x\},$$

where $\mathrm{KL}(Q\|\mu)$ denotes the Kullback-Leibler (KL) divergence between the probability distributions μ and Q (see Subsection 2.2 for the definition). With κ_μ in hand, our proposed solution to (1.1) is

$$\overline{x}_\mu = \operatorname*{argmin}_{x \in \mathbb{R}^d} \{\alpha g_b(Cx) \ + \ \kappa_\mu(x)\}, \tag{1.2}$$

where g_b is any (closed, proper) convex function that measures *fidelity* of Cx to b. The function g_b depends on b and can in principle be adapted to the noise distribution $\mathcal{Z}$. For example, as was highlighted in [44], one can take the *MEM estimator* (an alternative to the well-known *maximum likelihood estimator*) based upon a family of distributions (for instance, if the noise is Gaussian, then the MEM estimator is the familiar $g_b(\cdot) = \frac{1}{2}\|(\cdot) - b\|_2^2$). Finally $\alpha > 0$ is a fidelity parameter.

The variational problem (1.2) is solved via its Fenchel dual. As we explain in Subsection 2.2, we exploit the well-known connection in the large deviations literature that, under appropriate assumptions, the MEM function κ_μ is simply the Cramér rate function defined as the Fenchel conjugate of the log-moment generating function (LMGF)

$$L_\mu(y) := \log \int_{\mathcal{X}} \exp\langle y, \cdot \rangle d\mu.$$

Under certain assumptions on g_b (cf. Subsection 2.2) we obtain strong duality

$$\min_{x \in \mathbb{R}^d} \alpha g_b(Cx) + \kappa_\mu(x) = - \min_{z \in \mathbb{R}^m} \alpha g^*(-z/\alpha) + L_\mu(C^T z), \tag{1.3}$$

and, more importantly, a primal-dual recovery is readily available: If $\overline{z}_\mu$ is a solution to the dual problem (the argmin of the right-hand-side of (1.3)) then

$$\overline{x}_\mu := \nabla L_\mu(C^T \overline{z})$$

is the unique solution of the primal problem. This is the MEM method in a nutshell.

1.3. Contributions and outline of the article. In this article, we address the following questions: Suppose we do not have full access to the underlying prior distribution μ; rather we have access to an approximation sequence μ_n which in a suitable sense (e.g. weak convergence of measures) converges to μ. Does the approximate MEM solution $\overline{x}_{\mu_n}$ converge to the solution $\overline{x}_\mu$, and if so, at which rate? A key feature of the MEM approach is that one does not need to quantify the convergence of μ_n to μ, but rather only approximate the LMGF L_μ from data. Hence our analysis is based on the *closeness* of L_{μ_n} to L_μ. This results in the *closeness* of the dual solutions $\overline{z}_n$ and in turn the primal solutions $\overline{x}_{\mu_n}$. Here, we leverage the fundamental work of Wets et al. on *epigraphical distances, epigraphical convergence, and epi-consistency* ([37],[40],[26]).

Our results are presented in four sections. In Section 3, we work with a general g_b satisfying standard assumptions. We consider the simplest way of approximating μ via empirical means of n i.i.d. samples from μ. In Theorem 3.9, we prove that the associated MEM solutions $\overline{x}_{\mu_n}$ converge almost surely to the solution $\overline{x}_\mu$ with full prior. In fact, we prove a slightly stronger result pertaining to ε_n-solutions as $\varepsilon_n \searrow 0$. This result opens the door to two natural questions: (i) At which rate do the solutions converge? (ii) Empirical means is perhaps the simplest way of approximating μ and what is the corresponding rate? Given that the MEM method rests entirely on the LMGF of the prior, it is natural to ask how the rate depends on an approximation to the LMGF. So, if we used a different way of approximating μ, what would the rate look like? We address these questions for the case $g_b = \frac{1}{2}\|(\cdot) - b\|_2^2$. In Section 4 we provide insight into the second question first via a deterministic estimate which controls the difference in the respective solutions associated with

two priors ν and μ based upon the epigraphical distance between their respective LMGFs. We again prove a general result for ε-solutions associated with prior μ (cf. Theorem 4.7). In Section 5, we apply this bound to the particular case of the empirical means approximation, proving a $\frac{1}{n^{1/4}}$ convergence rate (cf. Theorem 5.5) in expectation.

Finally, in Section 6, we present several numerical experiments for denoising based upon a finite MNIST data set. These serve not to compete with any of the state-of-the-art machine learning-based denoising algorithm, but rather to highlight the effectiveness of our data-driven mathematical model which is fully supported by theory.

Remark 1.1 (Working at the higher level of the probability distribution of the solution). As in [35, 34], an equivalent formulation of the MEM problem is to work not at the level of the x, but rather at the level of the probability distribution of the ground truth, i.e., we seek to solve

$$\overline{Q} \;=\; \mathrm{argmin}_{Q\in\mathcal{P}(\mathcal{X})}\;\left\{\alpha g_b(C\mathbb{E}_Q) \;+\; \mathrm{KL}(Q\|\mu)\right\},$$

where one can recover the previous image-level solution as $\overline{x}_\mu = \mathbb{E}_{\overline{Q}}$. As shown in [34], under appropriate assumptions this reformulated problem has exactly the same dual formulation as in the right-hand-side of (1.3). Because of this one has full access to the entire probability distribution of the solution, not just its expectation. This proves useful in our MNIST experiments where the optimal ν is simply a weighted sum of images uniformly sampled from the MNIST set. For example, one can do thresholding (or masking) at the level of the optimal ν (cf. the examples in Section 6).

Notation: $\overline{\mathbb{R}} := \mathbb{R}\cup\{\pm\infty\}$ is the extended real line. The standard inner product on $\mathbb{R}^n$ is $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ is the Euclidean norm. For $C\in\mathbb{R}^{m\times d}$, $\|C\| = \sqrt{\lambda_{\max}(C^TC)}$ is its spectral norm, and analogously $\sigma_{\min}(C) = \sqrt{\lambda_{\min}(C^TC)}$ is the smallest singular value of C. The trace of C is denoted $\mathrm{Tr}(C)$. For smooth $f:\mathbb{R}^d\to\mathbb{R}$, we denote its gradient and Hessian by ∇f and $\nabla^2 f$, respectively.

2. Tools from convex analysis and the MEM method for solving the problem (1.1) .

2.1. Convex analysis. We present here the tools from convex analysis essential to our study. We refer the reader to the standard texts by Bauschke and Combettes [5] or Chapters 2 and 11 of Rockafellar and Wets [37] for further details. Let $f:\mathbb{R}^d\to\overline{\mathbb{R}}$. The domain of f is $\mathrm{dom}(f) := \{x\in\mathbb{R}^d \mid f(x) < +\infty\}$. We call f proper if $\mathrm{dom}(f)$ is nonempty and $f(x) > -\infty$ for all x. We say that f is lower semicontinuous (lsc) if $f^{-1}([-\infty,a])$ is closed (possibly empty) for all $a\in\mathbb{R}$. We define the (Fenchel) conjugate $f^*:\mathbb{R}^d\to\overline{\mathbb{R}}$ of f as $f^*(x^*) := \sup_{x\in\mathbb{R}^d}\{\langle x,x^*\rangle - f(x)\}$. A proper f is said to be convex, if $f(\lambda x+(1-\lambda)y)\le\lambda f(x)+(1-\lambda)f(y)$ for every $x,y\in\mathrm{dom}(f)$ and all $\lambda\in(0,1)$. If the former inequality is strict for all $x\neq y$, then f is said to be strictly convex. Finally, if f is proper and there is a $c>0$ such that $f-\frac{c}{2}\|\cdot\|^2$ is convex we say f is c-strongly convex. In the case where f is (continuously) differentiable on $\mathbb{R}^d$, then f is c-strongly convex if and only if

$$f(y)-f(x)\ge\nabla f(x)^T(y-x)+\frac{c}{2}\|y-x\|_2^2\quad\forall x,y\in\mathbb{R}^d. \tag{2.1}$$

The subdifferential of a convex function $f:\mathbb{R}^d\to\overline{\mathbb{R}}$ at $\overline{x}\in\mathrm{dom}(f)$ is $\partial f(\overline{x}) = \{x^*\in\mathbb{R}^d \,|\langle x-\overline{x},x^*\rangle\le f(x)-f(\overline{x}),\,\forall x\in\mathbb{R}^d\}$. A function $f:\mathbb{R}^d\to\overline{\mathbb{R}}$ is said to be level-bounded if for every $\alpha\in\mathbb{R}$, the set $f^{-1}([-\infty,\alpha])$ is bounded (possibly empty). A function f is (level) coercive if it is bounded below on bounded sets and satisfies

$$\liminf_{\|x\|\to+\infty}\frac{f(x)}{\|x\|}>0.$$

In the case f is proper, lsc, and convex, level-boundedness is equivalent to level-coerciveness [37, Corollary 3.27]. A function f is said to be supercoercive if $\liminf_{\|x\|\to+\infty}\frac{f(x)}{\|x\|}=+\infty$.
A point $\overline{x}$ is said to be an ε-minimizer of a proper function f if $f(\overline{x})\le\inf_{x\in\mathbb{R}^d}f(x)+\varepsilon$ for some $\varepsilon>0$. We denote the set of all such points as $S_\varepsilon(f)$. Correspondingly, the solution set of proper function f is denoted as $\mathrm{argmin}(f) = S_0(f) =: S(f)$.

The epigraph of a function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is the set $\text{epi}(f) := \{(x, \alpha) \in \mathbb{R}^d \times \overline{\mathbb{R}} \mid \alpha \geq f(x)\}$. A sequence of functions $f_n : \mathbb{R}^d \to \overline{\mathbb{R}}$ epigraphically converges (epi-converges)[1] to f, written $f_n \xrightarrow[n\to+\infty]{e} f$, if and only if

$$(i)\ \forall z, \forall z_n \to z :\ \liminf f_n(z_n) \geq f(z), \quad (ii)\ \forall z\ \exists z_n \to z : \limsup f_n(z_n) \leq f(z).$$

2.2. Maximum Entropy on the Mean Problem. For basic concepts of measure and probability, we follow most closely the standard text of Billingsley [6, Chapter 2]. Globally in this work, μ will be a Borel probability measure defined on compact $\mathcal{X} \subset \mathbb{R}^d$. Precisely, we work on the probability space $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \mu)$, where $\mathcal{X} \subset \mathbb{R}^d$ is compact and $\mathcal{B}_\mathcal{X} = \{B \cap \mathcal{X} \ :\ B \in \mathbb{B}_d\}$ where $\mathbb{B}_d$ is the σ-algebra induced by the open sets in $\mathbb{R}^d$. [2] We will denote the set of all probability measures on the measurable space $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ as $\mathcal{P}(\mathcal{X})$, and refer to elements of $\mathcal{P}(\mathcal{X})$ as probability measures on $\mathcal{X}$, with the implicit understanding that these are always Borel measures. For $Q, \mu \in \mathcal{P}(\mathcal{X})$, we say Q is absolutely continuous with respect to μ (and write $Q \ll \mu$) if for all $A \in \mathcal{B}_\mathcal{X}$ with $\mu(A) = 0$, then $Q(A) = 0$. [6, p. 422]. For $Q \ll \mu$, the Radon-Nikodym derivative of Q with respect to μ is defined as the (a.e.) unique function $\frac{dQ}{d\mu}$ with the property $Q(A) = \int_A \frac{dQ}{d\mu} d\mu$ for $A \in \mathcal{B}_\mathcal{X}$ [6, Theorem 32.3].

The Kullback-Leibler (KL) divergence [28] of $Q \in \mathcal{P}(\mathcal{X})$ with respect to $\mu \in \mathcal{P}(\mathcal{X})$ is defined as

$$\text{KL}(Q\|\mu) := \begin{cases} \int_\mathcal{X} \log(\frac{dQ}{d\mu}) d\mu, & Q \ll \mu, \\ +\infty, & \text{otherwise.} \end{cases} \tag{2.2}$$

For $\mu \in \mathcal{P}(\mathcal{X})$, the expected value $\mathbb{E}_\mu \in \mathbb{R}^d$ and moment generating function $M_\mu : \mathbb{R}^d \to \mathbb{R}$ function of μ are defined as [6, Ch.21]

$$\mathbb{E}_\mu := \int_\mathcal{X} x d\mu(x), \qquad M_\mu(y) := \int_\mathcal{X} \exp\langle y, x\rangle d\mu(x),$$

respectively. The log-moment generating function of μ is defined as

$$L_\mu(y) := \log M_\mu(y) = \log \int_\mathcal{X} \exp\langle y, x\rangle d\mu(x).$$

As $\mathcal{X}$ is bounded, M_μ is finite-valued everywhere. By standard properties of moment generating functions (see e.g. [41, Theorem 4.8]) it is then analytic everywhere, and in turn so is L_μ.

Given $\mu \in \mathcal{P}(\mathcal{X})$, the Maximum Entropy on the Mean (MEM) function [44] $\kappa_\mu : \mathbb{R}^d \to \overline{\mathbb{R}}$ is

$$\kappa_\mu(y) := \inf\{\text{KL}(Q \parallel \mu) : \mathbb{E}_Q = y, Q \in \mathcal{P}(\mathcal{X})\}.$$

The functions κ_μ and L_μ are paired in duality in a way that is fundamental to this work. We will flesh out this connection, as well as give additional properties of κ_μ for our setting; a Borel probability measure μ on compact $\mathcal{X}$. A detailed discussion of this connection under more general assumptions is the subject of [44].

For any $\mu \in \mathcal{P}(\mathcal{X})$ we have a vacuous tail-decay condition of the following form: for any $\sigma > 0$,

$$\int_\mathcal{X} e^{\sigma\|x\|} d\mu(x) \leq \max_{x\in\mathcal{X}} \|x\| e^{\sigma\|x\|} < +\infty.$$

Consequently, by [16, Theorem 5.2 (iv)][3] we have that

$$\kappa_\mu(x) = \sup_{y\in\mathbb{R}^d} \left[\langle y, x\rangle - \log \int_\mathcal{X} e^{\langle y,x\rangle} d\mu(x)\right] (= L^*_\mu(x)).$$

[1] This is one of many equivalent conditions that characterize epi-convergence, see e.g. [37, Proposition 7.2].
[2] Equivalently, we could work with a Borel measure μ on $\mathbb{R}^d$ with support contained in $\mathcal{X}$.
[3] Applied to μ considered as a measure over $\mathbb{R}^d$ with support in $\mathcal{X}$.

Note that the conjugate L_μ^* is known in the large deviations literature as the (Cramér) rate function. For a more full development and alternative derivations of this conjugacy we refer to [29, 31].

Returning to the setting of interest with our standing assumption that $\mathcal{X}$ is compact, $\kappa_\mu = L_\mu^*$. This directly implies the following properties of κ_μ: (i) As L_μ is proper, lsc, and convex, so is its conjugate $L_\mu^* = \kappa_\mu$. (ii) Reiterating that L_μ is proper, lsc, convex, we may assert $(L_\mu^*)^* = L_\mu$ via Fenchel-Moreau ([40, Theorem 5.23]), and hence $\kappa_\mu^* = L_\mu$. (iii) As $\operatorname{dom}(L_\mu) = \mathbb{R}^d$ we have that κ_μ is supercoercive [37, Theorem 11.8 (d)]. (iv) Recalling that L_μ is everywhere differentiable, κ_μ is strictly convex on every convex subset of $\operatorname{dom}(\partial\kappa_\mu)$, which is also referred to as essentially strictly convex [36, p. 253].

With these preliminary notions, we can (re-)state the problem of interest in full detail. We work with images represented as vectors in $\mathbb{R}^d$, where d is the number of pixels. Given observed image $b \in \mathbb{R}^m$ which may be blurred and noisy, and known matrix $C \in \mathbb{R}^{m\times d}$, we wish to recover the ground truth $\hat{x}$ from the linear inverse problem $b = C\hat{x} + \eta$, where $\eta \sim \mathcal{Z}$ is an unknown noise vector in $\mathbb{R}^m$ drawn from noise distribution $\mathcal{Z}$. We remark that, in practice, it is usually the case that $m = d$ and C is invertible, but this is not necessary from a theoretical perspective. We assume the ground truth $\hat{x}$ is the expectation of an underlying image distribution - a Borel probability measure - μ on compact set $\mathcal{X} \subset \mathbb{R}^d$. Our best guess of $\hat{x}$ is then obtained by solving

$$\text{(P)} \qquad \overline{x}_\mu = \operatorname*{argmin}_{x\in\mathbb{R}^d} \alpha g(Cx) + \kappa_\mu(x).$$

where $g = g_b$ is a proper, lsc, convex function which may depend on b and serves as a fidelity term, and $\alpha > 0$ a parameter. For example, if $g = \frac{1}{2}\|b - (\cdot)\|_2^2$ one recovers the so-called reformulated MEM problem, first seen in [29].

Lemma 2.1. *For any lsc, proper, convex g, the primal problem (P) always has a solution.*

Proof. By the global assumption of compactness of $\mathcal{X}$, we have κ_μ is proper, lsc, convex and supercoercive, following the discussion above. As $g \circ C$ and κ_μ are convex, so is $\alpha g \circ C + \kappa_\mu$ for $\alpha > 0$. Further as both $\alpha g \circ C$ and κ_μ are proper and lsc, and κ_μ is supercoercive, the summation $\alpha g \circ C + \kappa_\mu$ is supercoercive, [37, Exercise 3.29, Lemma 3.27]. A supercoercive function is, in particular, level-bounded, so by [37, Theorem 1.9] the solution set $\operatorname{argmin}(\alpha g \circ C + \kappa_\mu)$ is nonempty. ∎

We make one restriction on the choice of g, which will hold globally in this work:

Assumption 2.2. $0 \in \operatorname{int}(\operatorname{dom}(g) - C\operatorname{dom}(\kappa_\mu))$.

We remark that this property holds vacuously whenever g is finite-valued, e.g., $g = \frac{1}{2}\|b - (\cdot)\|_2^2$.

Instead of solving (P) directly, we use a dual approach. As $\kappa_\mu^* = L_\mu$ (by compactness of $\mathcal{X}$), the primal problem (P) has Fenchel dual (e.g., [5, Definition 15.19]) given by

$$\text{(D)} \qquad (\arg)\min_{z\in\mathbb{R}^m} \alpha g^*(-z/\alpha) + L_\mu(C^T z).$$

We will hereafter denote the dual objective associated with $\mu \in \mathcal{P}(\mathcal{X})$ as

$$\phi_\mu(z) := \alpha g^*(-z/\alpha) + L_\mu(C^T z). \tag{2.3}$$

We remark that our sign convention and use of minimization in the dual agrees with [5], but the dual problem appears elsewhere in the literature as $\max -\phi_\mu(z)$, see e.g. [48, Corollary 2.8.5]. We record the following result which highlights the significance of Assumption 2.2 to our study.

Theorem 2.3. *The following are equivalent:*
(i) Assumption 2.2 holds; (ii) $\operatorname{argmin}\phi_\mu$ is nonempty and compact; (iii) ϕ_μ is level-coercive. In particular, under Assumption 2.2, the primal problem (P) has a unique solution given by

$$\overline{x}_\mu = \nabla L_\mu(C^T\overline{z}), \tag{2.4}$$

where $\overline{z} \in \operatorname{argmin}\phi_\mu$ is any solution of the dual problem (D).

Proof. As ϕ_μ is proper, convex and lsc, the equivalence of (i)-(iii) is exactly [4, Proposition 3.1.3 (a),(c),(d)] as $\operatorname{int}(\operatorname{dom}(\phi_\mu^*)) = \operatorname{int}(C\operatorname{dom}(\kappa_\mu) - \operatorname{dom}(g))$. Furthermore [4, Theorem 5.2.1][4], yields the primal-dual recovery formula (2.4) using the differentiability of L_μ. ∎

2.3. Approximate and Empirical Priors, Random Functions, and Epi-consistency. If one has access to the true underlying image distribution μ, then the solution recipe is complete: solve (D) and use the primal-dual recovery formula (2.4) to find a solution to (P). But in practical situations, such as the imaging problems of interest here, it is unreasonable to assume full knowledge of μ and instead one models μ from domain specific knowledge or data set e.g. the discussions of [14, 42]. That is, one specifies a prior $\nu \in \mathcal{P}(\mathcal{X})$ with $\nu \approx \mu$, and solves the approximate dual problem

$$\min_{z\in\mathbb{R}^m} \phi_\nu(z). \tag{2.5}$$

Given $\varepsilon > 0$ and any ε-solution to (2.5), i.e. given any $z_{\nu,\varepsilon} \in S_\varepsilon(\nu)$, we define

$$\overline{x}_{\nu,\varepsilon} := \nabla L_\nu(C^T z_{\nu,\varepsilon}), \tag{2.6}$$

with the hope, inspired by the recovery formula (2.4), that with a "reasonable" choice of $\nu \approx \mu$, and small ε, then also $\overline{x}_{\nu,\varepsilon} \approx \overline{x}_\mu$. The remainder of this work is dedicated to formalizing how well $\overline{x}_{\nu,\varepsilon}$ approximates $\overline{x}_\mu$ under various assumptions on g and ν.

A natural first approach is to construct ν from sample data. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We model image samples as i.i.d. $\mathcal{X}$-valued random variables $\{X_1, \dots, X_n, \dots\}$ with shared law $\mu := \mathbb{P}\, X_1^{-1}$. That is, each $X_i : \Omega \to \mathcal{X}$ is an $(\Omega, \mathcal{F}) \to (\mathcal{X}, \mathcal{B}_\mathcal{X})$ measurable function with the property that $\mu(B) = \mathbb{P}(\omega \in \Omega \,:\, X_1(\omega) \in B)$, for any $B \in \mathcal{B}_\mathcal{X}$. In particular, the law μ is by construction a Borel probability measure on $\mathcal{X}$. Intuitively, a random sample of n images is a given sequence of realizations $\{X_1(\omega), \dots, X_n(\omega), \dots\}$, from which we take only the first n vectors. In practice, such a sequence could arise from sampling a fixed dataset uniformly at random, in which case n is dictated by the amount of data that is available and is feasible to compute with. We then approximate μ via the empirical measure

$$\mu_n^{(\omega)} := \frac{1}{n}\sum_{i=1}^n \delta_{X_i(\omega)}.$$

With this choice of $\nu = \mu_n^{(\omega)}$, we have the approximate dual problem

$$\min_{z\in\mathbb{R}^m} \phi_{\mu_n^{(\omega)}}(z) \quad \text{with} \quad \phi_{\mu_n^{(\omega)}}(z) = \alpha g^*\left(\frac{-z}{\alpha}\right) + \log \frac{1}{n}\sum_{i=1}^n e^{\langle C^T z, X_i(\omega)\rangle}. \tag{2.7}$$

And exactly analogous to (2.6), given an ε-solution $\overline{z}_{n,\varepsilon}(\omega)$ of (2.7), we define

$$\overline{x}_{n,\varepsilon}(\omega) := \nabla L_{\mu_n^{(\omega)}}(C^T \overline{z}_{n,\varepsilon}(\omega)) = \frac{\sum_{i=1}^n C X_i(\omega) e^{\langle C^T \overline{z}_{n,\epsilon}(\omega), X_i(\omega)\rangle}}{\sum_{i=1}^n e^{\langle C^T \overline{z}_{n,\epsilon}(\omega), X_i(\omega)\rangle}}. \tag{2.8}$$

Clearly, while the measure $\mu_n^{(\omega)}$ is well-defined and Borel for any given ω, the convergence properties of $\overline{z}_{n,\varepsilon}(\omega)$ and $\overline{x}_{n,\epsilon}(\omega)$ should be studied in a stochastic sense over Ω. To this end, we leverage a probabilistic version of epi-convergence for random functions known as epi-consistency [26].

Let $(T, \mathcal{A})$ be a measurable space. A function $f : \mathbb{R}^m \times T \to \overline{\mathbb{R}}$ is called a random[5] lsc function (with respect to $(T, \mathcal{A})$) [40, Definition 8.50] if the (set-valued) map $S_f : T \rightrightarrows \mathbb{R}^{m+1}$, $S_f(t) = \operatorname{epi} f(\cdot, t)$ is closed-valued and measurable in the sense $S_f^{-1}(O) = \{t \in T \,:\, S_f(x) \cap O \neq \emptyset\} \in \mathcal{A}$.

[4] Note that there is a sign error in equation (5.3) in the reference.

[5] The inclusion of the word 'random' in this definition need not imply a priori any relation to a random process; we simply require measurability properties of f. Random lsc functions are also known as normal integrands in the literature, see [37, Chapter 14].

Our study is fundamentally interested in random lsc functions on $(\Omega, \mathcal{F})$, in service of proving convergence results for $\overline{x}_{n,\epsilon}(\omega)$. But we emphasize that random lsc functions with respect to $(\Omega, \mathcal{F})$ are tightly linked with random lsc functions on $(X, \mathcal{B}_{\mathcal{X}})$. Specifically, if $X : \Omega \to \mathcal{X}$ is a random variable and $f : \mathbb{R}^m \times \mathcal{X} \to \overline{\mathbb{R}}$ is a random lsc function with respect to $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$, then the composition $f(\cdot, X(\cdot)) : \mathbb{R}^m \times \Omega \to \mathbb{R}$ is a random lsc function with respect to the measurable space $(\Omega, \mathcal{F})$, see e.g. [37, Proposition 14.45 (c)] or the discussion of [39, Section 5]. This link will prove computationally convenient in the next section.

While the definition of a random lsc function is unwieldy to work with directly, it is implied by a host of easy to verify conditions [40, Example 8.51]. We will foremost use the following one: Let $(T, \mathcal{A})$ be a measurable space. If a function $f : \mathbb{R}^m \times T \to \overline{\mathbb{R}}$ is finite valued, with $f(\cdot, t)$ continuous for all t, and $f(z, \cdot)$ measurable for all z, we say f is a Carathéodory function. Any function which is Carathéodory is random lsc [38, Example 14.26].

Immediately, we can assert $\phi_{\mu_n^{(\cdot)}}$ is a random lsc function from $\mathbb{R}^d \times \Omega \to \overline{\mathbb{R}}$, as it is Carathéodory. In particular, by [37, Theorem 14.37] or [39, Section 5], the ε-solution mappings

$$\omega \mapsto \left\{ z \ : \ \phi_{\mu_n^{(\omega)}}(z) \leq \inf \phi_{\mu_n^{(\omega)}} + \varepsilon \right\}$$

are measurable (in the set valued sense defined above), and for all $\varepsilon \geq 0$ there always exists a $\mathbb{P}$-measurable selection $z_{n,\varepsilon}(\omega) \in S_\varepsilon(\mu_n^{(\omega)})$ [37, Theorem 14.37, Theorem 14.33].

We conclude with the definition of epi-consistency as seen in [26, p. 86]; a sequence of random lsc functions $h_n : \mathbb{R}^m \times \Omega \to \overline{\mathbb{R}}$ is said to be epi-consistent with limit function $h : \mathbb{R}^m \to \overline{\mathbb{R}}$ if

$$\mathbb{P}\left(\left\{ \omega \in \Omega \mid h_n(\cdot, \omega) \xrightarrow[\mathrm{n} \to +\infty]{\mathrm{e}} h \right\} \right) = 1. \tag{2.9}$$

3. Epigraphical convergence and convergence of minimizers.

The goal of this section is to prove convergence of minimizers in the empirical case, i.e., that $\overline{x}_{n,\varepsilon}(\omega)$ as defined in (2.8) converges to $\overline{x}_\mu$, the solution of (P), for $\mathbb{P}$-almost every $\omega \in \Omega$ as $\varepsilon \searrow 0$. To do so, we prove empirical approximations of the moment generating function are epi-consistent with M_μ, and leverage this to prove epi-consistency of $\phi_{\mu_n^{(\omega)}}$ with limit ϕ_μ. Via classic convex analysis techniques, this guarantees the desired convergence of minimizers with probability one.

3.1. Epi-consistency of the empirical moment generating functions.

Given $\{X_1, \ldots, X_n, \ldots\}$ i.i.d. with shared law $\mu = \mathbb{P} X_1^{-1} \in \mathcal{P}(\mathcal{X})$, we denote the moment generating function of $\mu_n^{(\omega)}$ as $M_n(y, \omega) := \frac{1}{n} \sum_{i=1}^n e^{\langle y, X_i(\omega) \rangle}$. Define $f : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}$ as $f(z, x) = e^{\langle C^T z, x \rangle}$. Then

$$M_\mu(C^T z) = \int_{\mathcal{X}} e^{\langle C^T z, \cdot \rangle} d\mu = \int_{\mathcal{X}} f(z, \cdot) d\mu,$$

$$M_n(C^T z, \omega) = \frac{1}{n} \sum_{i=1}^n e^{\langle C^T z, X_i(\omega) \rangle} = \frac{1}{n} \sum_{i=1}^n f(z, X_i(\omega)).$$

This explicit decomposition is useful to apply a specialized version of the main theorem of King and Wets [26, Theorem 2], which we restate without proof.

Proposition 3.1. *Let $f : \mathbb{R}^m \times \mathcal{X} \to \overline{\mathbb{R}}$ be a random lsc function such that $f(\cdot, x)$ is convex and differentiable for all x. Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{X}$-valued random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with shared law $\mu \in \mathcal{P}(\mathcal{X})$. If there exists $\overline{z} \in \mathbb{R}^m$ such that*

$$\int_{\mathcal{X}} f(\overline{z}, \cdot) d\mu < +\infty, \qquad \text{and} \qquad \int_{\mathcal{X}} \|\nabla_z f(\overline{z}, \cdot)\| d\mu < +\infty,$$

then the sequence of (random lsc) functions $S_n : \mathbb{R}^m \times \Omega \to \overline{\mathbb{R}}$ given by

$$S_n(z, \omega) := \frac{1}{n} \sum_{i=1}^n f(z, X_i(\omega))$$

is epi-consistent with limit $S_\mu : z \mapsto \int_{\mathcal{X}} f(z,\cdot)d\mu$*, which is proper, convex, and lsc.*

Via a direct application of the above we have the following.

Corollary 3.2. *The sequence* $M_n(C^T(\cdot),\cdot)$ *is epi-consistent with limit* $M_\mu \circ C^T$*.*

Proof. Define $f(z,x) = e^{\langle C^T z, x\rangle}$. For any x, $\langle C^T(\cdot), x\rangle$ is a linear function, and $e^{(\cdot)}$ is convex - giving that the composition $f(\cdot,x)$ is convex. As f is differentiable (hence continuous) in z for fixed x and vice-versa, it is Carathéodory and thus a random lsc function (with respect to $(\mathcal{X},\mathcal{B}_\mathcal{X})$).

Next we claim $\overline{z} = 0$ satisfies the conditions of the proposition. First, by direct computation

$$\int_{\mathcal{X}} e^{\langle 0,x\rangle} d\mu(x) = \int_{\mathcal{X}} d\mu(x) = 1 < +\infty$$

as μ is a probability measure on $\mathcal{X}$. As $f(\cdot,x)$ is differentiable, we can compute $\nabla_z f(\overline{z},x) = Cxe^{\langle C^T 0,x\rangle} = Cx$. Hence

$$\int_{\mathcal{X}} \|\nabla_z f(\overline{z},x)\| d\mu(x) = \int_{\mathcal{X}} \|Cx\| d\mu(x) \leq \|C\| \max_{x\in\mathcal{X}} \|x\| < +\infty,$$

where we have used the boundedness of $\mathcal{X}$, and once again that μ is a probability measure. Thus we satisfy the assumptions of Proposition 3.1, and can conclude that the sequence of random lsc functions S_n given by $S_n(z,\omega) = \frac{1}{n}\sum_{i=1}^n f(z, X_i(\omega))$ are epi-consistent with limit $S_\mu : z \mapsto \int_{\mathcal{X}} f(z,\cdot)d\mu$. But,

$$S_n(z,\omega) = \frac{1}{n}\sum_{i=1}^n e^{\langle C^T z, X_i(\omega)\rangle} = M_n(C^T z,\omega) \qquad \text{and} \qquad S_\mu(z) = \int_{\mathcal{X}} e^{\langle C^T z,\cdot\rangle} d\mu = M_\mu(C^T z),$$

and so we have shown the sequence $M_n(C^T(\cdot),\cdot)$ is epi-consistent with limit $M_\mu \circ C^T$. ∎

Corollary 3.3. *The sequence* $L_{\mu_n^{(\omega)}} \circ C^T$ *is epi-consistent with limit* $L_\mu \circ C^T$*.*

Proof. Let

$$\Omega_e = \left\{ \omega \in \Omega \mid M_n(C^T(\cdot),\omega) \xrightarrow[n\to+\infty]{e} M_\mu \circ C^T(\cdot) \right\},$$

which has $\mathbb{P}(\Omega_e) = 1$ by Corollary 3.2, and let $\omega \in \Omega_e$. Both M_n and M_μ are finite valued and strictly positive, and furthermore the function $\log : \mathbb{R}_{++} \to \mathbb{R}$ is continuous and increasing. Hence, by a simple extension of [36, Exercise 7.8(c)], it follows, for all $\omega \in \Omega_e$, that

$$L_{\mu_n^{(\omega)}} \circ C^T = \log M_n(C^T(\cdot),\omega) \xrightarrow[n\to+\infty]{e} \log M_\mu \circ C^T = L_\mu \circ C^T.$$

∎

3.2. Epi-consistency of the dual objective functions. We now use the previous lemma to obtain an epi-consistency result for the entire empirical dual objective function. This is not an immediately clear, as epi-convergence is not generally preserved by even simple operations such as addition, see, e.g., the discussion in [37, p. 276] and the note [8] that eludes to subtle difficulties when dealing with extended real-valued arithmetic in this context.

We recall the following pointwise convergence result for compact $\mathcal{X}$, which is classical in the statistics literature.

Lemma 3.4. *If* $\mu \in \mathcal{P}(\mathcal{X})$*, for almost every* $\omega \in \Omega$*, and all* $z \in \mathbb{R}^m$

$$M_n(C^T z,\omega) \to M_\mu \circ C^T(z),$$

namely pointwise convergence in z*.*

We remark that the literature contains stronger uniform convergence results, observed first in Csörgö [**?**] without proof, and later proven in [18] and [12, Proposition 1]. Noting that both $M_n(z,\omega), M_\mu(z) > 0$ are strictly positive for all $z \in \mathbb{R}^m$, and that the logarithm is continuous on the strictly positive real line, we have an immediate corollary:

Corollary 3.5. *For almost every $\omega \in \Omega$, for all $z \in \mathbb{R}^m$*

$$L_{\mu_n^{(\omega)}}(C^T z) = \log M_n(C^T z, \omega) \to \log M_\mu(C^T z) = L_\mu(C^T z).$$

Using this we prove the first main result:

Theorem 3.6. *For any lsc, proper, convex function g, the empirical dual objective function $\phi_{\mu_n^{(\omega)}}$ is epi-consistent with limit ϕ_μ*

Proof. Define

$$\Omega_e = \left\{ \omega \in \Omega \mid L_{\mu_n^{(\omega)}} \circ C^T(\cdot) \xrightarrow[\mathrm{n}\to+\infty]{\mathrm{e}} L_\mu \circ C^T(\cdot) \right\}.$$

By Corollary 3.3, $\mathbb{P}(\Omega_e) = 1$. Similarly denote

$$\Omega_p = \left\{ \omega \in \Omega \mid L_{\mu_n^{(\omega)}} \circ C^T(\cdot) \to L_\mu \circ C^T(\cdot) \text{ pointwise} \right\}.$$

By Corollary 3.5, we also have $\mathbb{P}(\Omega_p) = 1$. In particular we observe that $\mathbb{P}(\Omega_e \cap \Omega_p) = 1$.

On the other hand we have vacuously that the constant sequence of convex, proper, lsc functions $\alpha g^* \circ (-\mathrm{Id}/\alpha)$ converges to $\alpha g^* \circ (-\mathrm{Id}/\alpha)$ both epigraphically and pointwise.

Thus for any fixed $\omega \in \Omega_p \cap \Omega_e$ we have constructed two sequences, namely $g_n \equiv \alpha g^* \circ (-\mathrm{Id}/\alpha)$ and $L_n = L_{\mu_n^{(\omega)}} \circ C^T$, which both converge epigraphically and pointwise for all $\omega \in \Omega_e \cap \Omega_p$. Therefore, by [37, Theorem 7.46(a)], for all $\omega \in \Omega_e \cap \Omega_p$

$$\alpha g^* \circ (-\mathrm{Id}/\alpha) + L_{\mu_n^{(\omega)}} \circ C^T \xrightarrow[\mathrm{n}\to+\infty]{\mathrm{e}} \alpha g^* \circ (-\mathrm{Id}/\alpha) + L_\mu \circ C^T.$$

As $\mathbb{P}(\Omega_e \cap \Omega_p) = 1$, this proves the result. ∎

3.3. Convergence of minimizers. We now use epi-consistency to prove convergence of minimizers. At the dual level this can be summarized in the following lemma, essentially [26, Proposition 2.2]; which was stated therein without proof.[6]

Lemma 3.7. *There exists a subset $\Xi \subset \Omega$ of measure one, such that for any $\omega \in \Xi$ we have: Let $\{\varepsilon_n\} \searrow 0$ and $z_n(\omega)$ such that*

$$\phi_{\mu_n^{(\omega)}}(z_n(\omega)) \le \inf_z \phi_{\mu_n^{(\omega)}}(z) + \varepsilon_n.$$

Let $\{z_{n_k}(\omega)\}$ be any convergent subsequence of $\{z_n(\omega)\}$. Then $\lim_{k\to+\infty} z_{n_k}(\omega)$ is a minimizer of ϕ_μ. If ϕ_μ admits a unique minimizer $\overline{z}_\mu$, then $z_n \to \overline{z}_\mu$.

Proof. Denote

$$\Xi = \left\{ \omega \in \Omega \mid \phi_{\mu_n^{(\omega)}} \xrightarrow[\mathrm{n}\to+\infty]{\mathrm{e}} \phi_\mu \right\}.$$

By Theorem 3.6, $\mathbb{P}(\Xi) = 1$. Fix any $\omega \in \Xi$.

[6] We remark that (as observed in [26]) epigraphical convergence of a (multi-)function depending on a parameter (such as ω) guarantees convergence of minimizers in much broader contexts, see e.g. [3, Theorem 1.10] or [38, Theorem 3.22]. Here we include a first principles proof.

As we have fixed $\omega \in \Xi$, we have that the sequence $\phi_{\mu_n^{(\omega)}} \xrightarrow[n\to+\infty]{e} \phi_\mu$ epi-converges. Also, by Theorem 2.3, our global Assumption 2.2 holds if and only if ϕ_μ is level-bounded. These two observations together imply by [37, Theorem 7.32 (c)] that the sequence $\phi_{\mu_n}^{(\omega)}$ is eventually level-bounded.[7] Altogether, this means the sequence of lsc, proper, eventually level-bounded functions $\phi_{\mu_n^{(\omega)}}$ epi-converge to ϕ_μ - which is also lsc and proper. This set of properties is precisely the necessary assumptions of [37, Theorem 7.33], which then asserts any sequence of approximate minimizers $\{z_n(\omega)\}$ is bounded with all cluster points belonging to $\operatorname{argmin} \phi_\mu$. Namely, any convergent subsequence $\{z_{n_k}(\omega)\}$ has the property that its limit $\lim_{k\to+\infty} z_{n_k} \in \operatorname{argmin} \phi_\mu$. Lastly, if we also have $\operatorname{argmin} \phi_\mu = \{\overline{z}_\mu\}$, then from the same result [37, Theorem 7.33], then necessarily $z_n(\omega) \to \overline{z}_\mu$. ∎

We now push this convergence to the primal level by using, in essence, Attouch's Theorem [2], [3, Theorem 3.66], in the form of a corollary of Rockafellar and Wets [37, Theorem 12.40].

Lemma 3.8. *Let $\hat{z} \in \mathbb{R}^m$, and let $z_n \to \hat{z}$ be any sequence converging to $\hat{z}$. Then for almost every ω,*

$$\lim_{n\to+\infty} \nabla L_{\mu_n^{(\omega)}}(C^T z_n) = \nabla L_\mu(C^T \hat{z}).$$

Proof. We first observe that $\operatorname{dom}(L_\mu \circ C^T) = \mathbb{R}^m$ so that $\hat{z} \in \operatorname{int}(\operatorname{dom}(L_\mu \circ C^T))$. Also as M_μ is everywhere finite-valued, $L_\mu(C^T\hat{z}) = \log M_\mu(C^T\hat{z}) < +\infty$. Furthermore for all n, the function $L_{\mu_n^{(\omega)}} \circ C^T$ is proper, convex, and differentiable. Finally, we have shown in Corollary 3.3, that for almost every $\omega \in \Omega$, we have $L_{\mu_n^{(\omega)}} \circ C^T \xrightarrow[n\to+\infty]{e} L_\mu \circ C^T$.
These conditions together are the necessary assumptions of [37, Theorem 12.40 (b)]. Hence we have convergence $\lim_{n\to+\infty} \nabla L_{\mu_n^{(\omega)}}(C^T z_n) = \nabla L_\mu(C^T\hat{z})$ for almost every $\omega \in \Omega$. ∎

We now prove the main result.

Theorem 3.9. *There exists a set $\Xi \subseteq \Omega$ of probability one such that for each $\omega \in \Xi$ the following holds: Given $\varepsilon_n \searrow 0$, and $z_n(\omega)$ such that $\phi_{\mu_n^{(\omega)}}(z_n(\omega)) \le \inf_z \phi_{\mu_n^{(\omega)}}(z) + \varepsilon_n$, define*

$$x_n(\omega) := \nabla L_{\mu_n^{(\omega)}}(C^T z_n).$$

If $z_{n_k}(\omega)$ is any convergent subsequence of $z_n(\omega)$ then $\lim_{k\to+\infty} x_{n_k}(\omega) = \overline{x}_\mu$, where $\overline{x}_\mu$ is the unique solution of (P). If (2.7) admits a unique solution $\overline{z}_\mu$, then in fact $x_n(\omega) \to \overline{x}_\mu$.

Proof. Let

$$\Xi = \{\omega \in \Omega \mid \phi_{\mu_n^{(\omega)}} \xrightarrow[n\to+\infty]{e} \phi_\mu\},$$

recalling that by Proposition 3.1, $\mathbb{P}(\Xi) = 1$. Fix $\omega \in \Xi$. By Lemma 3.7, for any convergent subsequence $z_{n_k}(\omega)$ with limit $\overline{z}(\omega)$, we have that $\overline{z}(\omega) \in \operatorname{argmin} \phi_\mu$. Furthermore, by Lemma 3.8

$$\lim_{k\to+\infty} x_{n_k}(\omega) = \lim_{k\to+\infty} \nabla L_{\mu_{n_k}^{(\omega)}}(C^T z_{n_k}) = \nabla L_\mu(C^T\overline{z}(\omega))$$

Using the primal-dual optimality conditions (2.4) we have that $\nabla L_\mu(C^T\overline{z}(\omega))$ solves the primal problem (P). As (P) admits a unique solution $\overline{x}_\mu$, necessarily $\lim_{k\to+\infty} x_{n_k}(\omega) = \overline{x}_\mu$. If additionally $\operatorname{argmin} \phi_\mu = \{\overline{z}_\mu\}$, then necessarily $z_n \to \overline{z}_\mu$ via Lemma 3.7, and the result follows from an identical application of Lemma 3.8 and (2.4). ∎

The key novelty of this result is the almost sure convergence of minimizers, particularity as $\varepsilon \searrow 0$. The compact support of the measure μ is key for this result. In particular, this is stronger than the convergence in probability guaranteed by common statistical techniques such as m-estimation [46, Section 3.2].

[7] A sequence of functions $f_n : \mathbb{R}^d \to \overline{\mathbb{R}}$ is eventually level-bounded if for each α, the sequence of sets $\{f_n^{-1}([-\infty, \alpha])\}$ is eventually bounded, see [37, p. 266].

4. Convergence rates for quadratic fidelity. When proving rates of convergence, we restrict ourselves to the case where $g = \frac{1}{2}\|b - (\cdot)\|_2^2$. Thus, the dual objective function reads

$$\phi_\mu(z) = \frac{1}{2\alpha}\|z\|^2 - \langle b, z\rangle + L_\mu(C^T z). \tag{4.1}$$

Clearly, ϕ_μ is finite valued and $(1/\alpha$-$)$strongly convex, hence admits a unique minimizer $\overline{z}_\mu$. Recalling what was laid out in Subsection 2.2, as global Assumption 2.2 holds vacuously with $g = \frac{1}{2}\|b - (\cdot)\|^2$, the unique solution to the MEM primal problem (P) is given by $\overline{x}_\mu = \nabla L_\mu(C^T\overline{z}_\mu)$. Further by our global compactness assumption on $\mathcal{X}$, ϕ_μ is (infinitely many times) differentiable.

4.1. Epigraphical distances. Our main tool to prove convergence rates are epigraphical distances. We mainly follow the presentation in Royset and Wets [40, Chapter 6.J], but one may find similar treatment in Rockafellar and Wets [37, Chapter 7]. For any norm $\|\cdot\|_*$ on $\mathbb{R}^d$, the distance (in said norm) between a point c and a set D is defined as $d_D(c) = \inf_{d\in D}\|c - d\|_*$. For C, D subsets of $\mathbb{R}^d$ we define the excess of C over D [40, p. 399] as

$$\operatorname{exc}(C, D) := \begin{cases} \sup_{c\in C} d_D(c) & \text{if } C, D \neq \emptyset, \\ +\infty & \text{if } C \neq \emptyset, D = \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

We note that this excess explicitly depends on the choice of norm used to define d_D. For the specific case of the 2-norm, we denote the projection of a point $a \in \mathbb{R}^d$ onto a closed, convex set $B \subset \mathbb{R}^d$ as the unique point $\operatorname{proj}_B(a) \in B$ which achieves the minimum $\|\operatorname{proj}_B(a) - a\| = \min_{b\in B}\|b - a\|$. The truncated ρ-Hausdorff distance [40, p. 399] between two sets $C, D \subset \mathbb{R}^d$ is defined as

$$\hat{d}_\rho(C, D) := \max\left\{\operatorname{exc}(C \cap B_\rho, D), \operatorname{exc}(D \cap B_\rho, C)\right\},$$

where $B_\rho = \{x \in \mathbb{R}^d : \|x\|_* \leq \rho\}$ is the closed ball of radius ρ in $\mathbb{R}^d$. When discussing distances on $\mathbb{R}^d$, we will consistently make the choice of $\|\cdot\|_* = \|\cdot\|$ the 2-norm. We note we can recover the usual Pompeiu-Hausdorff distance by taking $\rho \to +\infty$.

However, to extend the truncated ρ-distance to epigraphs of functions - which here are subsets of $\mathbb{R}^{d+1}$ - we equip $\mathbb{R}^{d+1}$ with a very particular norm. For any $z \in \mathbb{R}^{d+1}$, write $z = (x, a)$ for $x \in \mathbb{R}^d, a \in \mathbb{R}$. Then for any $z_1, z_2 \in \mathbb{R}^{d+1}$ we define the norm

$$\|z_1 - z_2\|_{*,d+1} = \|(x_1, a_1) - (x_2, a_2)\|_{*,d+1} := \max\{\|x_1 - x_2\|_2, |a_1 - a_2|\}.$$

With this norm, we can define an epi-distance as in [40, Equation 6.36]: for $f, h : \mathbb{R}^d \to \overline{\mathbb{R}}$ not identically $+\infty$, and $\rho > 0$ we define

$$\hat{d}_\rho(f, h) := \hat{d}_\rho(\operatorname{epi} f, \operatorname{epi} h), \tag{4.2}$$

where $\mathbb{R}^{d+1}$ has been equipped with the norm $\|\cdot\|_{*,d+1}$. This epi-distance quantifies epigraphical convergence in the following sense: [37, Theorem 7.58][8] if f is a proper function and f_n a sequence of proper functions, then for any constant $\rho_0 > 0$:

$$f_n \xrightarrow[n\to+\infty]{e} f \text{ if and only if } \hat{d}_\rho(f_n, f) \to 0 \text{ for all } \rho > \rho_0.$$

4.2. Convergence Rates. We begin with a technical lemma which will prove expedient for future results.

[8] We remark that while at first glance the definition of epi-distance seen in [37, Theorem 7.58] differs from ours (which agrees with [40]), it is equivalent up to multiplication by a constant and rescaling in ρ. See [40, Proposition 6.58] and [37, Proposition 7.61] - the Kenmochi conditions - for details.

Lemma 4.1. *Let $\rho > 0$ and $\nu \in \mathcal{P}(\mathcal{X})$. Then, for all $z \in B_\rho$ we have*

$$M_\nu(C^T z) = \int_{\mathcal{X}} e^{\langle C^T z, \cdot \rangle} d\nu \in \left[\exp\left(-\rho\|C\||\mathcal{X}|\right), \exp\left(\rho\|C\||\mathcal{X}|\right)\right].$$

Proof. For all $x \in \mathcal{X}$, $z \in B_\rho$, we have, via Cauchy-Schwarz, that

$$\exp\left(-\rho\|C\||\mathcal{X}|\right) \leq \exp\left(-\|z\|\|Cx\|\right) \leq \exp\langle C^T z, x\rangle.$$

In particular, $\exp\left(-\rho\|C\||\mathcal{X}|\right) \leq \min_{x\in\mathcal{X}} \exp\langle C^T z, x\rangle$. On the other hand, we find that

$$\exp\langle C^T z, x\rangle \leq \exp\left(\|z\|\|Cx\|\right) \leq \exp\left(\rho\|C\||\mathcal{X}|\right).$$

Thus, $\max_{x\in\mathcal{X}} \exp\langle C^T z, x\rangle \leq \exp\left(\rho\|C\||\mathcal{X}|\right)$. Hence for any $\nu \in \mathcal{P}(\mathcal{X})$ we find

$$1 \cdot \exp\left(-\rho\|C\||\mathcal{X}|\right) \leq \nu(\mathcal{X}) \min_{x\in\mathcal{X}} e^{\langle C^T z, x\rangle} \leq \int_{\mathcal{X}} e^{\langle C^T z, \cdot\rangle} d\nu \leq \nu(\mathcal{X}) \max_{x\in\mathcal{X}} e^{\langle C^T z, x\rangle} \leq 1 \cdot \exp\left(\rho\|C\||\mathcal{X}|\right).$$

∎

We now prove rates of convergence for arbitrary prior ν, and later specialize to the empirical case. To this end, we construct a key global constant ρ_0 induced by C, b, α in (4.1): We define

$$\rho_0 := \max\left\{\hat{\rho}, \frac{\hat{\rho}^2}{2\alpha} + \|b\|\hat{\rho} + \hat{\rho}\|C\||\mathcal{X}|\right\}, \tag{4.3}$$

where $\hat{\rho} = 2\alpha(\|b\| + \|C\||\mathcal{X}|)$ and $|\mathcal{X}| := \max_{x\in\mathcal{X}} \|x\|$. We emphasize that our running compactness assumption on $\mathcal{X}$ is essential for finiteness of ρ_0. The main feature of this constant is the following.

Lemma 4.2. *For any $\nu \in \mathcal{P}(\mathcal{X})$, let ϕ_ν be the corresponding dual objective function as defined in (4.1), which has a unique minimizer $\overline{z}_\nu$. Then ρ_0 has the following two properties:*

$$(a) \quad \phi_\nu(\overline{z}_\nu) \in [-\rho_0, \rho_0], \qquad (b) \quad \|\overline{z}_\nu\| \leq \rho_0.$$

Proof. We first claim that $\|\overline{z}_\nu\| \leq \hat{\rho}$. Let $z \in \mathbb{R}^d$ be such that $\|z\| > \hat{\rho}$. Then,

$$\phi_\nu(z) \geq \frac{\|z\|^2}{2\alpha} - \|b\|\|z\| + \log\exp\left(-\|C\|\|x\|\|z\|\right) = \|z\|\left(\frac{\|z\|}{2\alpha} - \|b\| - \|C\||\mathcal{X}|\right). \tag{4.4}$$

Where in the first inequality we have used Cauchy-Schwarz on $\langle b, z\rangle$, and Lemma 4.1 with $\rho = \|z\|$ to bound $M_\mu(C^T z) \geq \exp\left(-\|C\|\|x\|\|z\|\right)$. From (4.4) it is clear that $\|z\| > \hat{\rho}$ implies $\phi_\nu(z) > 0$. But observing that $\phi_\nu(0) = 0$, such z cannot be a minimizer. Hence necessarily $\|\overline{z}_\nu\| \leq \hat{\rho} \leq \rho_0$. Once more, via Cauchy-Schwarz and Lemma 4.1 we compute

$$\begin{aligned} |\phi_\nu(\overline{z}_\nu)| = \left|\frac{\|\overline{z}_\nu\|^2}{2\alpha} - \langle b, \overline{z}_\nu\rangle + L_\nu(\overline{z}_\nu)\right| &\leq \frac{\hat{\rho}^2}{2\alpha} + \hat{\rho}\|b\| + \log\exp(\|C\||\mathcal{X}|\hat{\rho}) \\ &= \frac{\hat{\rho}^2}{2\alpha} + \hat{\rho}\|b\| + \hat{\rho}\|C\||\mathcal{X}|. \end{aligned}$$

∎

Lemma 4.3. *Let ρ_0 be given by (4.3). Then for all $\rho > \rho_0$ and all $\mu, \nu \in \mathcal{P}(\mathcal{X})$, we have*

$$\hat{d}_\rho(\phi_\mu, \phi_\nu) \leq \max_{z\in B_\rho} |L_\nu(C^T z) - L_\mu(C^T z)|.$$

Proof. Lemma 4.2 guarantees that for both measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$, we have

$$\phi_\nu(\overline{z}_\nu), \phi_\mu(\overline{z}_\mu) \in [-\rho_0, \rho_0] \qquad \text{and} \qquad \|\overline{z}_\nu\|, \|\overline{z}_\mu\| \leq \rho_0. \tag{4.5}$$

These conditions imply, for any $\rho > \rho_0$, that the set $C_\rho := (\{z : \phi_\mu(z) \le \rho\} \cup \{z : \phi_\nu(z) \le \rho\}) \cap B_\rho$ is nonempty. This follows from (4.5) as for any $\rho > \rho_0$ the nonempty set $\{\overline{z}_\mu, \overline{z}_\nu\} \cap B_{\rho_0}$ is contained in $C_{\rho_0} \subset C_\rho$. As C_ρ is nonempty we may apply [40, Theorem 6.59] with $f = \phi_\mu$ and $g = \phi_\nu$ to obtain $\hat{d}_\rho(\phi_\mu, \phi_\nu) \le \sup_{z \in C_\rho} |\phi_\nu(z) - \phi_\mu(z)|$. Then from the definition of ϕ_μ and ϕ_ν, we have

$$\begin{aligned} \sup_{z \in C_\rho} |\phi_\nu(z) - \phi_\mu(z)| &= \sup_{z \in C_\rho} |L_\nu(C^T z) - L_\mu(C^T z)| \\ &\le \sup_{z \in B_\rho} |L_\nu(C^T z) - L_\mu(C^T z)| \\ &= \max_{z \in B_\rho} |L_\nu(C^T z) - L_\mu(C^T z)|, \end{aligned}$$

where in the penultimate line uses that $C_\rho \subseteq B_\rho$, and the final equality follows as the continuous function $L_\mu \circ C^T - L_\nu \circ C^T$ achieves a maximum over the compact set B_ρ. ∎

For notational convenience, we will hereafter denote

$$D_\rho(\nu, \mu) := \max_{z \in B_\rho} |L_\mu(C^T z) - L_\nu(C^T z)|.$$

We also recall from Subsection 2.1 that $S_\varepsilon(\nu)$ denotes the set of ε-minimizers of ϕ_ν.

Lemma 4.4. *Let ρ_0 be given by (4.3). Then, for all $\mu, \nu \in \mathcal{P}(\mathcal{X})$, all $\rho > \rho_0$, and all $\varepsilon \in [0, \rho - \rho_0]$, the following holds: If*

$$\delta > \varepsilon + 2D_\rho(\nu, \mu),$$

then

$$|\phi_\nu(\overline{z}_\nu) - \phi_\mu(\overline{z}_\mu)| \le D_\rho(\nu, \mu) \qquad \text{and} \qquad \operatorname{exc}(S_\varepsilon(\nu) \cap B_\rho, S_\delta(\mu)) \le D_\rho(\nu, \mu).$$

Proof. Let $\rho > \rho_0$ and $\varepsilon \in [0, \rho - \rho_0]$. By choice, we have $\varepsilon < 2\rho$. By Lemma 4.2(a) we have $\phi_\nu(\overline{z}_\nu), \phi_\mu(\overline{z}_\mu) \in [-\rho_0, \rho_0]$, and in turn by the choice of ρ, ε we have $[-\rho_0, \rho_0] \subseteq [-\rho, \rho_0] \subseteq [-\rho, \rho - \varepsilon]$. Also, as $\rho > \rho_0$, by Lemma 4.2(b) we have $\{z_\nu\} = \operatorname{argmin} \phi_\nu \cap B_\rho$ and $\{z_\mu\} = \operatorname{argmin} \phi_\mu \cap B_\rho$. These properties of ρ, ε are exactly the assumptions of [40, Theorem 6.56] for $f = \phi_\mu$ and $g = \phi_\nu$. This result yields that, if $\delta > \varepsilon + 2\hat{d}_\rho(\phi_\mu, \phi_\nu)$, then $|\phi_\nu(\overline{z}_\nu) - \phi_\mu(\overline{z}_\mu)| \le \hat{d}_\rho(\phi_\nu, \phi_\mu)$ and $\operatorname{exc}(S_\varepsilon(\nu) \cap B_\rho, S_\delta(\mu)) \le \hat{d}_\rho(\phi_\nu, \phi_\mu)$.
However as $\rho > \rho_0$, we may apply Lemma 4.3 to assert $\hat{d}_\rho(\phi_\mu, \phi_\nu) \le D_\rho(\nu, \mu)$. Hence, for any $\delta > \varepsilon + 2D_\rho(\nu, \mu) \ge \varepsilon + 2\hat{d}_\rho(\phi_\mu, \phi_\nu)$ we obtain

$$\begin{aligned} |\phi_\nu(\overline{z}_\nu) - \phi_\mu(\overline{z}_\mu)| &\le D_\rho(\nu, \mu), \\ \operatorname{exc}(S_\varepsilon(\nu) \cap B_\rho, S_\delta(\mu)) &\le D_\rho(\nu, \mu). \end{aligned}$$

∎

For the main results, Theorem 4.7 and Theorem 5.5, we require additional auxiliary results. With some additional computation, we can infer the following Lipschitz bound on ∇L_ν.

Corollary 4.5. *Let $\hat{\rho} > 0$ and $\nu \in \mathcal{P}(\mathcal{X})$. Then for all $x, y \in B_{\hat{\rho}} \subset \mathbb{R}^d$, we have that*

$$\|\nabla L_\nu(x) - \nabla L_\nu(y)\| \le K\|x - y\| \tag{4.6}$$

for an explicit constant $K > 0$ which depends on $\hat{\rho}, d, |\mathcal{X}|$, but not on ν.

Proof. As discussed in Subsection 2.2, L_ν is twice continuously differentiable. Hence, using the fundamental theorem of calculus, we have $\nabla L_\nu(x) - \nabla L_\nu(y) = \int_0^1 \nabla^2 L_\mu(x + t(y - x)) \cdot (y - x) dt$.

Thus, as $x + t(y - x) \in B_{\hat{\rho}}$ for all $t \in [0, 1]$, we have

$$\begin{aligned}\|\nabla L_\nu(z) - \nabla L_\nu(y)\| &\leq \int_0^1 \|\nabla^2 L_\mu(x + t(y - x))\|\|y - x\| dt \\ &\leq \int_0^1 \max_{z \in B_{\hat{\rho}}} \|\nabla^2 L_\nu(z)\|\|y - x\| dt \\ &= \max_{z \in B_{\hat{\rho}}} \|\nabla^2 L_\nu(z)\|\|y - x\|.\end{aligned}$$

By convexity of L_ν, we observe that $\nabla^2 L_\nu(z)$ is (symmetric) positive semidefinite (for any z). Hence, $\max_{z \in B_{\hat{\rho}}} \|\nabla^2 L_\nu(z)\| \leq \max_{z \in B_{\hat{\rho}}} \operatorname{Tr}(\nabla^2 L_\nu(z))$. Now, observe that

$$\frac{\partial}{\partial z_i} L_\nu(z) = \frac{1}{M_\nu(z)} \left[\int_{\mathcal{X}} x_i \exp\langle z, x\rangle d\nu(x)\right] = \frac{1}{\int_{\mathcal{X}} \exp\langle z, x\rangle d\nu(x)} \left[\int_{\mathcal{X}} x_i \exp\langle z, x\rangle d\nu(x)\right],$$

where the interchange of the derivative and integral is permitted by the Leibniz rule for finite measures, see e.g. [19, Theorem 2.27] or [27, Theorem 6.28]. Hence,

$$\frac{\partial^2}{\partial z_i^2} L_\nu(z) = \frac{-1}{(M_\nu(z))^2} \left[\int_{\mathcal{X}} x_i \exp\langle z, x\rangle d\nu(x)\right]^2 + \frac{1}{M_\nu(z)} \left[\int_{\mathcal{X}} x_i^2 \exp\langle z, x\rangle d\nu(x)\right].$$

Taking the absolute value in the last identity, we may bound $|x_i| \leq \|x\| \leq |\mathcal{X}|$, $\|z\| \leq \hat{\rho}$, and apply Lemma 4.1 to bound $M_\nu(z)$. This eventually yields

$$\left|\frac{\partial^2}{\partial z_i^2} L_\nu(z)\right| \leq \frac{|\mathcal{X}|}{\exp(-\hat{\rho}|\mathcal{X}|)^2} \exp(\hat{\rho}|\mathcal{X}|)^2 + \frac{|\mathcal{X}|^2}{\exp(-\hat{\rho}|\mathcal{X}|)} \exp(\hat{\rho}|\mathcal{X}|) =: \hat{K},$$

with $\hat{K} > 0$ which depends on $\hat{\rho}$ and $|\mathcal{X}|$. As this uniformly bounds every term in the trace, $K := d \cdot \hat{K}$ is the desired constant. ∎

The key feature of the constant K is that it does not depend on the choice of measure ν. Hence we can uniformly apply this bound over a family of measures, the most pertinent example being $\left\{\mu_n^{(\omega)}\right\}$. We remark that our upper bound on K is a vast overestimate for practical examples, which can be observed numerically. Finally we state a useful property of the excess.

Lemma 4.6. *Let $A, B \subset \mathbb{R}^d$ be nonempty and let B be closed and convex. Then for $\overline{a} \in A$ and $\overline{b} = proj_B(\overline{a})$ we have*

$$\|\overline{a} - \overline{b}\| \leq \operatorname{exc}(A; B).$$

We now have developed all the necessary tools to state and prove the main result for the case of $g = \frac{1}{2}\|(\cdot) - b\|$.

Theorem 4.7. *Let ρ_0 be given by (4.3), and suppose $\operatorname{rank}(C) = d$. Then for all $\mu, \nu \in \mathcal{P}(\mathcal{X})$, all $\rho > \rho_0$ and all $\varepsilon \in [0, \rho - \rho_0]$, we have the following: If $\overline{z}_{\nu,\varepsilon}$ is an ε-minimizer of ϕ_ν as defined in (4.1), then*

$$\overline{x}_{\nu,\varepsilon} := \nabla L_\nu(C^T \overline{z}_{\nu,\varepsilon})$$

satisfies the error bound

$$\|\overline{x}_{\nu,\varepsilon} - \overline{x}_\mu\| \leq \frac{1}{\alpha \sigma_{\min}(C)} D_\rho(\nu, \mu) + \frac{2\sqrt{2}}{\sqrt{\alpha}\sigma_{\min}(C)} \sqrt{D_\rho(\nu, \mu)} + \left(K\|C\|\sqrt{2\alpha} + \frac{2}{\sqrt{\alpha}\sigma_{\min}(C)}\right)\sqrt{\varepsilon},$$

where $\overline{x}_\mu$ is the unique solution to the MEM primal problem (P) for μ and $K > 0$ is a constant which does not depend on μ, ν.

Proof. Let $\rho > \rho_0$, $\nu, \mu \in \mathcal{P}(\mathcal{X})$ and $\varepsilon \in [0, \rho - \rho_0]$. Let $\overline{z}_{\nu,\varepsilon}$ be a ε-minimizer of ϕ_ν, and denote the unique minimizers of ϕ_μ and ϕ_μ as $\overline{z}_\nu$ and $\overline{z}_\mu$, respectively. Then

$$\begin{aligned}\|\overline{x}_{\nu,\varepsilon} - \overline{x}_\mu\| &= \left\|\nabla L_\nu(C^T\overline{z}_{\nu,\varepsilon}) - \nabla L_\mu(C^T\overline{z}_\mu)\right\| \\ &= \left\|\nabla L_\nu(C^T\overline{z}_{\nu,\varepsilon}) - \nabla L_\nu(C^T\overline{z}_\nu) + \nabla L_\nu(C^T\overline{z}_\nu) - \nabla L_\mu(C^T\overline{z}_\mu)\right\|,\end{aligned}$$

and so

$$\|\overline{x}_{\nu,\varepsilon} - \overline{x}_\mu\| \leq \left\|\nabla L_\nu(C^T\overline{z}_{\nu,\varepsilon}) - \nabla L_\nu(C^T\overline{z}_\nu)\right\| + \left\|\nabla L_\nu(C^T\overline{z}_\nu) - \nabla L_\mu(C^T\overline{z}_\mu)\right\|. \tag{4.7}$$

To estimate the first term on the right hand side of (4.7), we require an auxiliary bound. Observe that, as ϕ_ν is strongly $1/\alpha$-convex with $\nabla\phi_\nu(\overline{z}_\nu) = 0$, we have

$$\|\overline{z}_\nu - \overline{z}_{\nu,\varepsilon}\| \leq \sqrt{2\alpha}|\phi_\nu(\overline{z}_\nu) - \phi_\nu(\overline{z}_{\nu,\varepsilon})|^{1/2} \leq \sqrt{2\alpha\varepsilon}. \tag{4.8}$$

Here the first inequality uses (2.1), while the second follows from the definition of $\overline{z}_\nu$ and $\overline{z}_{\nu,\varepsilon}$, as $|\phi_\nu(z_\nu) - \phi_\nu(z_{\nu,\varepsilon})| = \phi_\nu(z_{\nu,\varepsilon}) - \phi_\nu(z_\nu) \leq \varepsilon$.

From Lemma 4.2(b), we find that $\|\overline{z}_\nu\| \leq \rho_0$. Thus, (4.8) yields $\|\overline{z}_{\nu,\varepsilon}\| \leq \rho_0 + \sqrt{2\alpha\varepsilon}$. This implies $\|C^T\overline{z}_\nu\|, \|C^T\overline{z}_{\nu,\varepsilon}\| \leq \|C\|(\rho_0 + \sqrt{2\alpha\varepsilon})$. Hence, Corollary 4.5 with $\hat{\rho} = \|C\|(\rho_0 + \sqrt{2\alpha\varepsilon})$ yields

$$\left\|\nabla L_\nu(C^T\overline{z}_{\nu,\varepsilon}) - \nabla L_\nu(C^T\overline{z}_\nu)\right\| \leq K\|C^T\overline{z}_\nu - C^T\overline{z}_{\nu,\varepsilon}\|,$$

where K depends on $\hat{\rho}, |\mathcal{X}|, d$ and therefore on $|\mathcal{X}|, \|C\|, b, \varepsilon, \alpha, d$. The right-hand side in the last inequality can be further estimated with (4.8) to find

$$\|C^T\overline{z}_\nu - C^T\overline{z}_{\nu,\varepsilon}\| \leq \|C\|\|\overline{z}_\nu - \overline{z}_{\nu,\varepsilon}\| \leq \|C\|\sqrt{2\alpha\varepsilon}. \tag{4.9}$$

We now turn to the second term on the right-hand side of (4.7). First order optimality conditions give

$$0 = -\frac{\overline{z}_\nu}{\alpha} + b + C\nabla L_\nu(C^T\overline{z}_\nu), \qquad 0 = -\frac{\overline{z}_\mu}{\alpha} + b + C\nabla L_\mu(C^T\overline{z}_\mu),$$

and therefore $\left\|C(\nabla L_\nu(C^T\overline{z}_\nu) - \nabla L_\mu(C^T\overline{z}_\mu))\right\| = \frac{1}{\alpha}\|\overline{z}_\nu - \overline{z}_\mu\|$. Furthermore, as $\operatorname{rank}(C) = d$ we have $\sigma_{\min}(C) > 0$. We also have for, any $x \in \mathbb{R}^d$, that $\|Cx\| \geq \sigma_{\min}(C)\|x\|$, and hence

$$\left\|\nabla L_\nu(C^T\overline{z}_\nu) - \nabla L_\mu(C^T\overline{z}_\mu)\right\| \leq \frac{1}{\sigma_{\min}(C)}\left\|C(\nabla L_\nu(C^T\overline{z}_\nu) - \nabla L_\mu(C^T\overline{z}_\mu))\right\| = \frac{1}{\alpha\sigma_{\min}(C)}\|\overline{z}_\nu - \overline{z}_\mu\|.$$

In order to bound $\|\overline{z}_\nu - \overline{z}_\mu\|$ from above, we define $\delta := 2(\varepsilon + 2D_\rho(\nu, \mu))$. Denoting as usual $S_\delta(\mu)$ as the set of δ-minimizers of ϕ_μ, which is a closed, convex set by the continuity and convexity of ϕ_μ respectively, define $y = \operatorname{proj}_{S_\delta(\mu)}(\overline{z}_\nu)$. The triangle inequality gives

$$\|\overline{z}_\nu - \overline{z}_\mu\| \leq \|\overline{z}_\nu - y\| + \|y - \overline{z}_\mu\|. \tag{4.10}$$

By the choice of $\rho > \rho_0$, we have by Lemma 4.2(b) that $\overline{z}_\nu \in S_\varepsilon(\nu) \cap B_\rho$. Therefore applying Lemma 4.6 with $A = S_\varepsilon(\nu) \cap B_\rho$, $B = S_\delta(\mu)$ we can bound the first term on the right hand side of (4.10) as

$$\|\overline{z}_\nu - y\| \leq \operatorname{exc}(S_\varepsilon(\nu) \cap B_\rho; S_\delta(\mu)). \tag{4.11}$$

For the remaining term of the right hand side of (4.10), we use the characterization (2.1) of the $\frac{1}{\alpha}$-strong convexity in the differentiable case for ϕ_μ, noting $\nabla\phi_\mu(\overline{z}_\mu) = 0$. Hence

$$\|y - \overline{z}_\mu\| \leq \sqrt{2\alpha}|\phi_\mu(y) - \phi_\mu(\overline{z}_\mu)|^{1/2} \leq \sqrt{2\alpha\delta}, \tag{4.12}$$

where $y \in S_\delta(\mu)$ for the second inequality. Combining (4.9)–(4.12) with (4.7), we find that

$$\|\overline{x}_{\nu,\varepsilon} - \overline{x}_\mu\| \leq K\|C\|\sqrt{2\alpha\varepsilon} + \frac{1}{\alpha\sigma_{\min}(C)}\mathrm{exc}(S_\varepsilon(\nu) \cap B_\rho; S_\delta(\mu)) + \frac{1}{\alpha\sigma_{\min}(C)}\sqrt{2\alpha\delta}.$$

By the choice of $\delta = 2(\varepsilon + 2D_\rho(\nu,\mu))$, Lemma 4.4 asserts $\mathrm{exc}(S_\varepsilon(\nu) \cap B_\rho; S_\delta(\mu)) \leq D_\rho(\nu,\mu)$. Therefore

$$\begin{aligned}
\|\overline{x}_{\nu,\varepsilon} - \overline{x}_\mu\| &\leq \frac{1}{\alpha\sigma_{\min}(C)}\mathrm{exc}(S_\varepsilon(\nu) \cap B_\rho; S_\delta(\mu)) + \frac{1}{\alpha\sigma_{\min}(C)}\sqrt{2\alpha\delta} + K\|C\|\sqrt{2\alpha\varepsilon} \\
&\leq \frac{1}{\alpha\sigma_{\min}(C)}D_\rho(\nu,\mu) + \frac{1}{\sigma_{\min}(C)}\sqrt{\frac{4\varepsilon}{\alpha}} + \frac{1}{\sigma_{\min}(C)}\sqrt{\frac{8D_\rho(\nu,\mu)}{\alpha}} + K\|C\|\sqrt{2\alpha\varepsilon} \\
&= \frac{1}{\alpha\sigma_{\min}(C)}D_\rho(\nu,\mu) + \frac{2\sqrt{2}}{\sqrt{\alpha}\sigma_{\min}(C)}\sqrt{D_\rho(\nu,\mu)} + \left(K\|C\|\sqrt{2\alpha} + \frac{2}{\sqrt{\alpha}\sigma_{\min}(C)}\right)\sqrt{\varepsilon}
\end{aligned}$$

where in the second line we have used the definition of δ and the concavity of $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$. ∎

Note that we may set $\varepsilon = 0$ for a corollary on exact minimizers. However, the error bound still has the same scaling in terms of $D_\rho(\nu,\mu)$. This theorem is the first of its type to link MEM solutions with epigraphical distances. This result has the benefit applying uniformly in the choice η, so this theorem can be applied for *any* choice of η which provides a bound on $D_\rho(\mu,\eta)$ - and we demonstrate the particular case of $\nu = \mu_n^{(\omega)}$ in the next section. While this result has the stringent assumption that $\mathrm{rank}(C) = d$, we emphasize that this does not appear in the qualitative Theorem 3.9 which guarantees the almost sure convergence of solutions. We conjecture that weakening this assumption may be possible, but will require alternative techniques. Finally we remark that the scaling $\sqrt{D_\rho(\mu,\eta)}$ arises as a direct consequence of the smoothness and strong convexity of $g = \frac{1}{2}\|b - (\cdot)\|^2$, see e.g. [40, Theorem 4.2] and the following discussion - and hence it may be feasible to prove improved rates for other specialized choices of g.

5. A statistical dependence on n. This section is devoted to making the dependence on n explicit in Theorem 4.7 for the special case $\nu = \mu_n^{(\omega)}$. We briefly recall the empirical setting developed in Subsection 2.3. Given i.i.d. random vectors $\{X_1, X_2, \ldots, X_n, \ldots\}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ with shared law $\mu = \mathbb{P}X_1^{-1}$, we define $\mu_n^{(\omega)} = \sum_{i=1}^n \delta_{X_i(\omega)}$. For this measure, the dual objective reads

$$\phi_{\mu_n^{(\omega)}}(z) = \frac{1}{2\alpha}\|z\|^2 - \langle b, z\rangle + \log\frac{1}{n}\sum_{i=1}^n e^{\langle C^T z, X_i(\omega)\rangle}.$$

Given $\overline{z}_{n,\varepsilon}(\omega)$, an ε-minimizer of $\phi_{\mu_n^{(\omega)}}(z)$, define

$$\overline{x}_{n,\varepsilon}(\omega) := \nabla L_{\mu_n^{(\omega)}}(C^T\overline{z}_{n,\epsilon}(\omega)) = \frac{\sum_{i=1}^n CX_i(\omega)e^{\langle C^T\overline{z}_{n,\epsilon}(\omega), X_i(\omega)\rangle}}{\sum_{i=1}^n e^{\langle C^T\overline{z}_{n,\epsilon}(\omega), X_i(\omega)\rangle}}.$$

We begin with a simplifying lemma, recalling the notation developed in Section 4 of the moment generating function M_μ of μ and empirical moment generating function $M_n(\cdot,\omega)$ of $\mu_n^{(\omega)}$.

Lemma 5.1. *Let $\rho > 0$, $n \in \mathbb{N}$ and $\omega \in \Omega$ and set $K := \exp(\rho\|C\||\mathcal{X}|)$. Then*

$$D_\rho(\mu, \mu_n^{(\omega)}) \leq K \max_{z \in B_\rho} \left|M_\mu(C^T z) - M_n(C^T z, \omega)\right|.$$

Proof. Applying Lemma 4.1 to the particular probability measures μ and $\mu_n^{(\omega)}$ gives

$$M_\mu(C^T z), M_n(C^T z, \omega) \in [\exp(-\rho\|C\||\mathcal{X}|), \exp(\rho\|C\||\mathcal{X}|)] =: [c, d] \tag{5.1}$$

where $0 < c < d$. Furthermore, for any $s, t \in [c,d]$ we have $|\log(s) - \log(t)| \leq \frac{1}{c}|s - t|$, and hence

$$D_\rho(\mu, \mu_n^{(\omega)}) = \max_{z \in B_\rho} |L_{\mu_n^{(\omega)}}(C^T z) - L_\mu(C^T z)| \leq \exp(\rho\|C\||\mathcal{X}|) \max_{z \in B_\rho} \left|M_\mu(C^T z) - M_n(C^T z, \omega)\right|.$$

∎

Lemma 5.2. *Let ρ_0 be as defined in (4.3) and suppose* $\operatorname{rank}(C) = d$. *Then, for all $\rho > \rho_0$, all $\varepsilon \in [0, \rho - \rho_0]$, and for all $n \in \mathbb{N}$ we have: if $\overline{z}_{n,\varepsilon}(\omega) \in S_\varepsilon(\mu_n^{(\omega)})$, then $\overline{x}_{n,\varepsilon}(\omega) = \nabla L_{\mu_n^{(\omega)}}(C^T \overline{z}_{n,\varepsilon})$ satisfies*

$$\begin{aligned}\|\overline{x}_{n,\varepsilon}(\omega) - \overline{x}_\mu\| \leq{}& \frac{K_1}{\alpha\sigma_{\min}(C)} \max_{z\in B_\rho} \left|M_\mu(C^T z) - M_n(C^T z, \omega)\right| \\ &+ \frac{2\sqrt{2K_1}}{\sqrt{\alpha}\sigma_{\min}(C)} \sqrt{\max_{z\in B_\rho} |M_\mu(C^T z) - M_n(C^T z, \omega)|} + \left(K_2\|C\|\sqrt{2\alpha} + \frac{2}{\sqrt{\alpha}\sigma_{\min}(C)}\right)\sqrt{\varepsilon}\end{aligned}$$

where K_1 is a constant which depends on $\rho, |\mathcal{X}|, \|C\|$, and K_2 on $|\mathcal{X}|, \|C\|, b, \varepsilon, d, \alpha$.

Proof. As $\mu_n^{(\omega)} \in \mathcal{P}(\mathcal{X})$, Theorem 4.7 yields for all n, for ρ_0 as defined in (4.3), for all $\rho > \rho_0$, and all $\varepsilon \in [0, \rho - \rho_0]$, if $\overline{z}_{n,\varepsilon}(\omega) \in S_\varepsilon(\mu_n^{(\omega)})$, then $\overline{x}_{n,\varepsilon}(\omega) = \nabla L_\nu(C^T z_{n,\varepsilon})$ satisfies

$$\begin{aligned}\|\overline{x}_{n,\varepsilon}(\omega) - \overline{x}_\mu\| \leq{}& \frac{1}{\alpha\sigma_{\min}(C)} D_\rho(\mu, \mu_n^{(\omega)}) + \frac{2\sqrt{2}}{\sqrt{\alpha}\sigma_{\min}(C)} \sqrt{D_\rho(\mu, \mu_n^{(\omega)})} \\ &+ \left(K_2\|C\|\sqrt{2\alpha} + \frac{2}{\sqrt{\alpha}\sigma_{\min}(C)}\right)\sqrt{\varepsilon},\end{aligned}$$

where we stress that the constant K_2 depends on $|\mathcal{X}|, \|C\|, b, \varepsilon, \alpha, d$, but does not depend on n. Applying Lemma 5.1 to bound $D_\rho(\mu, \mu_n^{(\omega)}) \leq K_1 \sup_{z\in B_\rho} \left|M_\mu(C^T z) - M_n(C^T z, \omega)\right|$ gives the result. ∎

In order to construct a final bound which depends explicitly on n it remains to estimate the term $\max_{z\in B_\rho} \left|M_\mu(C^T z) - M_n(C^T z, \omega)\right|$. This fits into the language of empirical process theory where this type of convergence is well studied. The main reference of interest here is van der Vaart [45].

For compact $\mathcal{X} \subset \mathbb{R}^d$, let $f : \mathcal{X} \to \mathbb{R}$ be a function and $\beta = (\beta_1, \ldots, \beta_d)$ be a multi-index, i.e. a vector of d nonnegative integers. We call $|\beta| = \sum_i \beta_i$ the order of β, and define the differential operator $D^\beta = \frac{\partial^{\beta_1}}{\partial x_1^{\beta_1}} \cdots \frac{\partial^{\beta_d}}{\partial x_d^{\beta_d}}$. For integer k, we denote by $\mathcal{C}^k(\mathcal{X})$ as the space of k-smooth (also known as k-Hölder continuous) functions on $\mathcal{X}$, namely, those f which satisfy [45, p. 2131]

$$\|f\|_{\mathcal{C}^k(\mathcal{X})} := \max_{|\beta|\leq k} \sup_{x\in\operatorname{int}(\mathcal{X})} \|D^\beta f(x)\| + \max_{|\beta|=k} \sup_{\substack{x,y\in\operatorname{int}(\mathcal{X}) \\ x\neq y}} \left|\frac{D^\beta f(x) - D^\beta f(y)}{\|x - y\|}\right| < +\infty.$$

Moreover, let $\mathcal{C}_R^k(\mathcal{X})$ denote the ball of radius R in $\mathcal{C}^k(\mathcal{X})$. With this notation developed we can state the classical results of van der Vaart [45]. In the notation therein, we apply the machinery of Sections 1 and 2 to the measure space $(\mathcal{X}_1, \mathcal{A}_1) = (\mathcal{X}, \mathcal{B}_\mathcal{X})$, equipped with probability measure μ. Taking $\mathbb{G}_n = \sqrt{n}(\mu_n^{(\omega)} - \mu)$ and $\mathcal{F}_1 = \mathcal{F} = C_R^k(\mathcal{X})$, this induces the norm $\|\mathbb{G}_n\|_\mathcal{F} = \sup_{f\in C_R^k(\mathcal{X})}\{\left|\int_\mathcal{X} f d\mathbb{G}_n\right|\}$, and hence the results of [45, p. 2131] give

Theorem 5.3. *Let $\mu \in \mathcal{P}(\mathcal{X})$. If $k > d/2$, then for any $R > 0$,*

$$\mathbb{E}_\mathbb{P}^* \left[\sup_{f\in C_R^k(\mathcal{X})} \sqrt{n} \left|\int_\mathcal{X} f d\mu_n^{(\cdot)} - \int_\mathcal{X} f d\mu\right|\right] \leq D$$

where D is a constant depending (polynomially) on $k, d, |\mathcal{X}|, R$, and $\mathbb{E}_\mathbb{P}^*$ is the outer expectation to avoid concerns of measurablity (see e.g. [46, Section 1.2]).

Here, the outer expectation is defined for $f : \Omega \to \overline{\mathbb{R}}$ as

$$\mathbb{E}_\mathbb{P}^*(f) := \inf_{\substack{h\geq f \text{ pointwise} \\ h \text{ measurable}}} \mathbb{E}_\mathbb{P}(h)$$

which coincides with the usual expectation for measurable functions. We remark that outer expectation is also known as the outer integral [37, Chapter 14.F]. A self-contained proof of Theorem 5.3

is non-trivial, requiring the development of entropy and bracketing numbers of function spaces which is beyond the scope of this article.[9] We simply take this result as given. However, we show the following corollary.

Corollary 5.4. *For all $n \in \mathbb{N}$, we have*

$$\mathbb{E}_{\mathbb{P}}\left[\max_{z\in B_\rho}\left|M_\mu(C^Tz) - M_n(C^Tz,\cdot)\right|\right] \le \frac{D}{\sqrt{n}},$$

where D is a constant depending on $d, |\mathcal{X}|, \|C\|, \rho$.

Proof. Observe that for each z, the function $f_z(x) = \exp\langle C^Tz, \cdot\rangle$ is an infinitely differentiable function on the compact set $\mathcal{X}$, and thus has bounded derivatives of all orders, in particular, of order $k = d > d/2$. Hence, for $\rho > 0$, the set of functions f_z parameterized by $z \in B_\rho$ satisfies

$$\left\{f_z(x) = \exp\langle C^\top z, x\rangle : z \in B_\rho\right\} \subset \mathcal{C}^d_{R_d}(\mathcal{X}),$$

where R_d is a constant which depends on $d, \rho, \|C\|$, and $|\mathcal{X}|$. Furthermore, as $\mathbb{Q}^m \cap B_\rho \subset B_\rho$ is a countable dense subset, $\max_{z\in B_\rho}\left|M_\mu(C^Tz) - M_n(C^Tz,\cdot)\right| = \sup_{z\in\mathbb{Q}^m\cap B_\rho}\left|M_\mu(C^Tz) - M_n(C^Tz,\cdot)\right|$ is a supremem of countably many $\mathbb{P}$-measurable functions and is hence $\mathbb{P}$-measurable. In particular the usual expectation agrees with the outer expectation. Hence applying Theorem 5.3 we may assert

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}\left[\max_{z\in B_\rho}\left|M_\mu(C^Tz) - M_n(C^Tz,\cdot)\right|\right] &= \mathbb{E}^*_{\mathbb{P}}\left[\sup_{z\in B_\rho}\left|\int_{\mathcal{X}} f_z d\mu_n^{(\cdot)} - \int_{\mathcal{X}} f_z\mu\right|\right] \\
&\le \mathbb{E}^*_{\mathbb{P}}\left[\sup_{f\in\mathcal{C}^d_{R_d}(\mathcal{X})}\left|\int_{\mathcal{X}} f d\mu_n^{(\cdot)} - \int_{\mathcal{X}} f\mu\right|\right] \\
&\le \frac{D}{\sqrt{n}},
\end{aligned}$$

for a constant D which depends on $d, |\mathcal{X}|$ and R_d. We remark that the choice of $k = d$ in the above was aesthetic, to remove the dependence of D on k. ∎

The final result now follows as a simple consequence of Corollary 5.4 and Lemma 5.2:

Theorem 5.5. *Suppose* $\operatorname{rank}(C) = d$. *For all $n \in \mathbb{N}$, and all $\overline{z}_{n,\varepsilon}(\omega) \in S_\varepsilon(\mu_n^{(\omega)})$, the associated $\overline{x}_{n,\varepsilon}(\omega) = \nabla L_{\mu_n^{(\omega)}}(C^T\overline{z}_{n,\varepsilon}(\omega))$ satisfies*

$$\begin{aligned}
\mathbb{E}^*_{\mathbb{P}}\left\|\overline{x}_{n,\varepsilon}(\cdot) - \overline{x}_\mu\right\| &\le \frac{DK_1}{\alpha\sigma_{\min}(C)}\frac{1}{\sqrt{n}} + \frac{2D\sqrt{2K_1}}{\sqrt{\alpha}\sigma_{\min}(C)}\sqrt{\frac{1}{\sqrt{n}}} + \left(K_2\|C\|\sqrt{2\alpha} + \frac{2}{\sqrt{\alpha}\sigma_{\min}(C)}\right)\sqrt{\varepsilon} \\
&= O\left(\frac{1}{n^{1/4}} + \sqrt{\varepsilon}\right),
\end{aligned}$$

where the leading constants K_1, K_2, D depend on $|\mathcal{X}|, \|C\|, b, \varepsilon, \alpha, d$. In particular, for the case $\varepsilon = 0$, the unique minimizer $\overline{x}_n(\omega)$ is always measurable and hence the outer expectation is the usual expectation.

[9] Bounds of this "Donsker" type have previously been applied to empirical approximations of stochastic optimization problems, to derive large deviation-style results for specific problems, see [39, Section 5.5] for a detailed exposition and discussion, in particular [39, Theorem 5.2]. In principle this machinery could be used here to derive similar large deviation results.

Proof. Take $\rho = 2\rho_0$ in Lemma 5.2. Then for any n, if $\overline{z}_{n,\varepsilon}(\omega) \in S_\varepsilon(\mu_n^{(\omega)})$, we have for all ω

$$\|\overline{x}_{n,\varepsilon}(\omega) - \overline{x}_\mu\| \leq \frac{K_1}{\alpha\sigma_{\min}(C)} \sup_{z\in B_\rho} \left|M_\mu(C^T z) - M_n(C^T z, \omega)\right| + \frac{2\sqrt{2}\sqrt{K_1}}{\sqrt{\alpha}\sigma_{\min}(C)} \sqrt{\sup_{z\in B_\rho} |M_\mu(C^T z) - M_n(C^T z, \omega)|} + \left(K_2\|C\|\sqrt{2\alpha} + \frac{2}{\sqrt{\alpha}\sigma_{\min}(C)}\right)\sqrt{\varepsilon},$$

holds with constants K_1, K_2 that depend only on $|\mathcal{X}|, \|C\|, b, \varepsilon, \alpha, d$. Taking the outer expectation on both sides, and applying Corollary 5.4 to the measurable right hand side gives the result.
For the latter assertion, by [37, Theorem 14.37] there always exists a measurable selection $\omega \to \operatorname{argmin} \phi_{\mu_n^{(\omega)}} = \{\overline{z}_n(\omega)\}$. In particular, $\overline{z}_n(\omega)$ is unique by the $\frac{1}{\alpha}$-strong convexity of $\phi_{\mu_n^{(\omega)}}$, and thus there is only one possible selection which is immediately measurable. Hence the function $\omega \to \overline{x}_n(\omega)$ is the composition of the continuous function $\nabla L_{\mu_n^{(\omega)}}$ and the measurable function $C^T\overline{z}_n(\omega)$ - and is hence measurable. Remarking also that $\overline{x}_n(\omega)$ is unique, the left hand side is always $\mathbb{P}$-measurable and the outer expectation agrees with the usual expectation. ∎

More generally we note that for all $\varepsilon > 0$, there always exists a measurable selection $z_{n,\varepsilon}(\omega) \in S_\varepsilon(\mu_n^{(\omega)})$ via [37, Theorem 14.6, Proposition 14.33]. Furthermore any algorithm $\mathcal{A}$ which performs a composition of operations which preserve Borel measurability - such summations, products, differentiations, and evaluations of measurable functions (see e.g. [19, Section 2.1]) - defines a selection $\omega \to \mathcal{A}(\phi_{\mu_n^{(\omega)}}) = \overline{z}_{n,\varepsilon}(\omega)$ which is a composition of Borel-measurable functions and hence measurable. With this in mind, issues of measurability are a technical note rather than of practical concern.

Finally, we remark that this result is the first to give a parametric rate for convergence of approximate minimizers of the empirical MEM problem. We conjecture however that this result is not sharp, and that a sharper convergence rate of $n^{1/2}$ may be proven using different analytical techniques. This conjecture will be examined numerically in the next section.

6. Numerical experiments.

We now shift to a numerical examination of the convergence $\overline{x}_{n,\varepsilon} \to \overline{x}_\mu$. We focus entirely on the most recent setting of Section 5, the MEM problem with an empirical prior $\mu_n^{(\omega)}$ and the fidelity term $g = \frac{1}{2}\|b - (\cdot)\|^2$. As discussed throughout, the empirical dual objective function $\phi_{\mu_n}^{(\omega)}$ is smooth and strongly convex, with easily computable derivatives:[10]

$$\nabla\phi_{\mu_n}^{(\omega)}(z) = \frac{1}{2\alpha}z - b + \frac{\sum_{i=1}^n CX_i(\omega)e^{\langle C^T z, X_i(\omega)\rangle}}{\sum_{i=1}^n e^{\langle C^T z, X_i(\omega)\rangle}}.$$

Hence we may solve this problem with any standard optimization package, and here we choose to use limited memory BFGS [9]. As a companion to this paper, we include a jupyter notebook, https://github.com/mattkingros/MEM-Denoising-and-Deblurring.git, which can be used to reproduce and extend all computations performed hereafter. We emphasize that, as this work is the first to propose using empirical data in the MEM framework, there are many numerical considerations that we will not address. More sophisticated and higher order optimization routines would naturally be of interest for moderate n, as are questions of how to efficiently solve this problem when n grows prohibitively large or when C is exceedingly close to singular.

For our numerical experiments we will focus purely on denoising, namely the case $C = I$. We will use two datasets and two distributions of noise as proof of concept. For noise, we use additive Gaussian noise and "salt-and-pepper" corruption noise where each pixel has an independent probability of being set to purely black or white. For datasets, the first is the MNIST digits dataset [15] which contains 60000 28x28 grayscale pixel images of hand-drawn digits 0-9, and the second is

[10] The derivatives are easily available analytically, but a naïve implementation may run into stability issues stemming from overflow when computing large sums of exponential terms. This can be easily addressed, see e.g. [25].

the more expressive dataset of Fashion-MNIST [47], which is once again 60000 28x28 grayscale pixel images but of various garments such as shoes, sneakers, bags, pants, and so on. We remark these choices of noise, dataset, and matrix are far from real world problems, and serve as a lightweight implementation of data into the MEM framework.

In all experiments we *always* include enough noise so that the nearest neighbour to b in the dataset belongs to a different class than the ground truth, ensuring that recovery is non-trivial. We include this in all figures, captioned "closest point". Furthermore, we take the ground truth to be a new data-point hand drawn by the authors; in particular, it is *not present* in either of the original MNIST or MNIST fashion datasets.

We begin with a baseline examination of the error in n for the MNIST digit dataset, for various choices of b. To generate $\mu_n^{(\omega)}$ practically, we sample n datapoints uniformly at random without replacement. As a remark on this methodology, the target best possible approximation is the one which uses all possible data, i.e., $\mu_D := \mu_{60{,}000}^{(\omega)}$. Hence, for error plots we compare to $\overline{x}_{\mu_D}$, as we do not have access to the full image distribution to construct $\overline{x}_\mu$.

Given a noisy image b, the experimental setup is as follows: for 20 values of n, spaced linearly between 10000 and 60000, we perform 15 random samples of size n. For each random sample, we compute an approximation $\overline{x}_{n,\epsilon}$ and the relative approximation error $\frac{\|\overline{x}_{n,\epsilon} - \overline{x}_{\mu_D}\|}{\|\overline{x}_{\mu_D}\|}$, which is then averaged over the trials. Superimposed is the upper bound $K_1 n^{1/4}$ convergence rate, as well as the conjectured $K_2 n^{1/2}$ rate, for moderate constants K_1, K_2 which changes between figures. We also visually exhibit several reconstructions, the nearest neighbour, and postprocessed images for comparison. This methodology is used to create figures Figures 1, 3, 6, 8, 10 and 12

A remark on postprocessing: As alluded to in the introduction (see Remark 1.1), an advantage of the dual approach is the ability to reconstruct the optimal measure Q_n which solves the measure-valued primal problem as seen in [34]. In particular as $Q_n \ll \mu_n^{(\omega)}$, we have $\overline{x}_n = \mathbb{E}_{Q_n}$ is a particular weighted linear combination of the input data. This allows for two types of natural postprocessing: at the level of the linear combination, i.e., setting all weights below some given threshold to zero, or at the level of the pixels: i.e. setting all pixels above 1-γ to 1 and below γ to 0. These postprocessing steps are motivated by the observation that solutions are compressible both in the linear combination and the pixel-intensity level. Hence our figures also include the final measure Q_n with all entries below 0.01 set to zero, the corresponding linear combination of the remaining datapoints (bottom right image), and a further masking at the pixel level (top right image).

With this methodology developed, we present the results. For the first figures, the ground truth is a hand-drawn 8, which is approximated by sampling the MNIST digit dataset. For Figure 1 we use additive Gaussian noise of variance $\sigma = 0.10\|x\|$. For Figure 3 we use salt-and-pepper corruptions with an equal probability of 0.2 of any given pixel being set to 1 or 0.

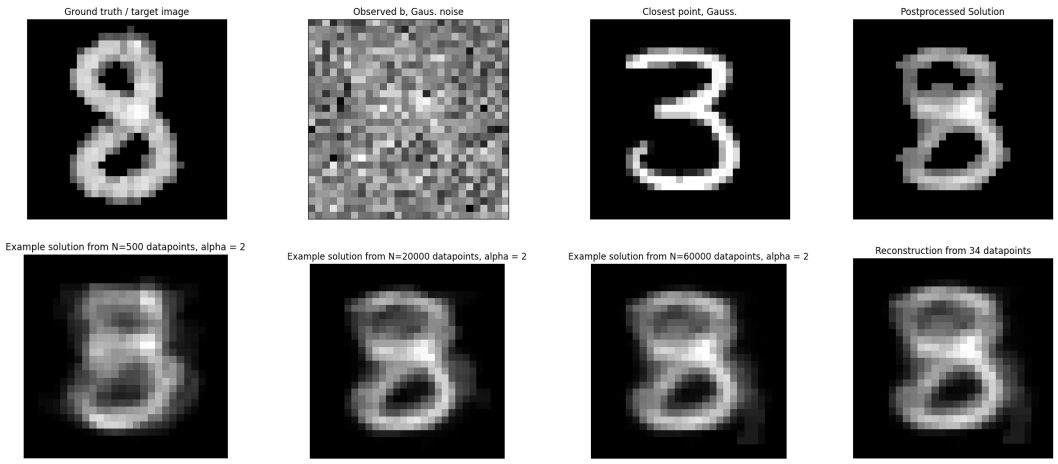

Figure 1: Recovery of an 8 with additive Gaussian noise

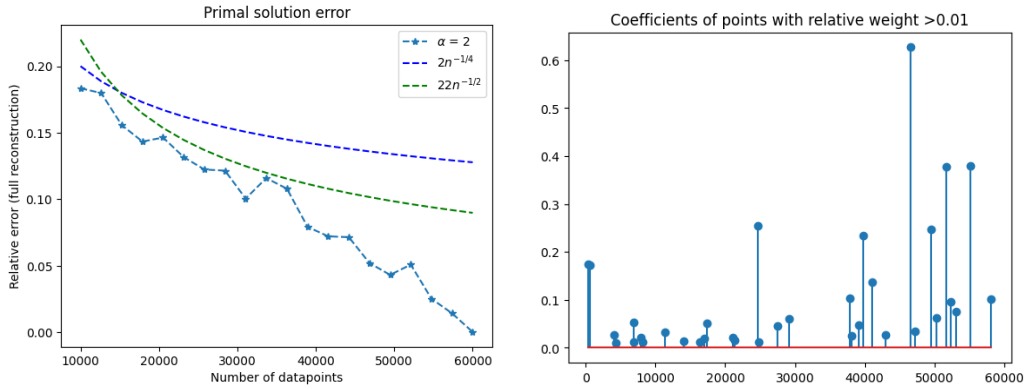

Figure 2: Rates and thresholding of optimal measure Q_n, for Figure 1

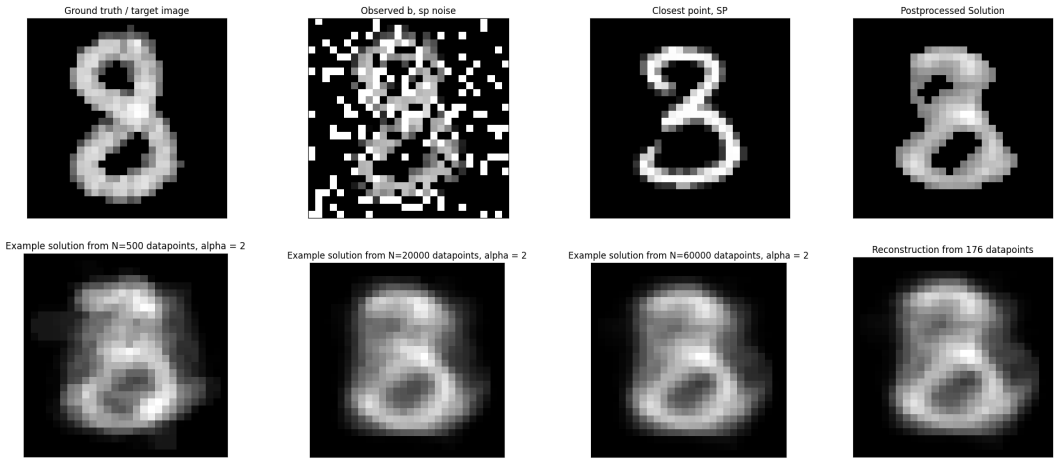

Figure 3: Recovery of an 8 with salt-and-pepper corruption noise

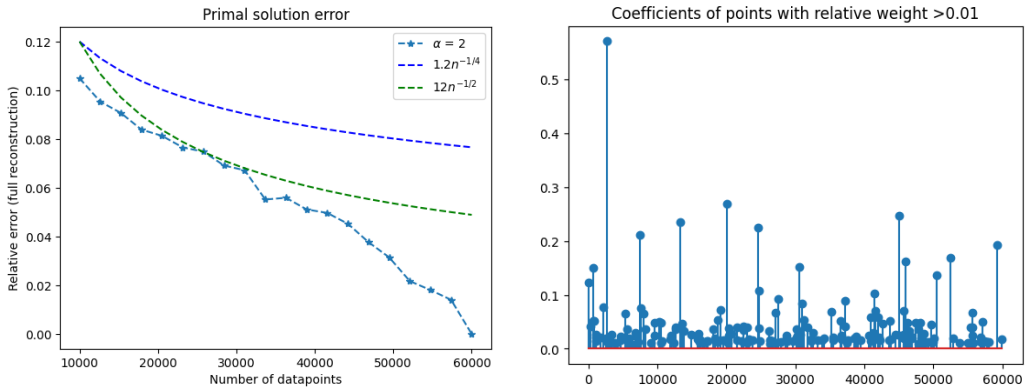

Figure 4: Rates and thresholding of optimal measure Q_n, for Figure 3

Examining Figures 1 and 3, we can make several observations which will persist globally in all numerics. As seen in Figures 2 and 4, the theoretical rate $n^{1/4}$ agrees with practical results, and appears to be tight for moderate n on the order of $n < 30000$. While the leading constants given by theory are quite large, growing polynomially with d, ρ e.g., in practice these are much more moderate - here $O(1)$. We also note the linear combination forming the final solution is quite compressible, here having only 157 datapoints having dual measure above 0.01, and after thresholding being visually indistinguishable from the full solution. Furthermore the *visual* convergence is fast, in the sense that there is not an appreciable visual difference between the solution given 20,000 and 60,000 datapoints. As they are formed as linear combinations of observed data, all solutions suffer from artefacting in the form of blurred edges, which can be solved by a final mask at the pixel level.

Before shifting away from the MNIST dataset, we also include a cautionary experiment where, for small random samples, the method is visually "confidently incorrect", which can be seen in Figure 5. In the previous examples Figures 1 and 3 for small n we simply have blurry images. This type of failure is generally representative of how the method performs when n is too small. However there is another failure case which distinctly highlights the risk of taking n too small, which can lead to a biased sample which does not well approximate μ - leading to correspondingly biased solutions. In Figure 5 we see the recovered image after masking is clearly a 3 for one random samples of size $n = 1000$, but a 5 for other random samples of size $n = 800, 2000$.

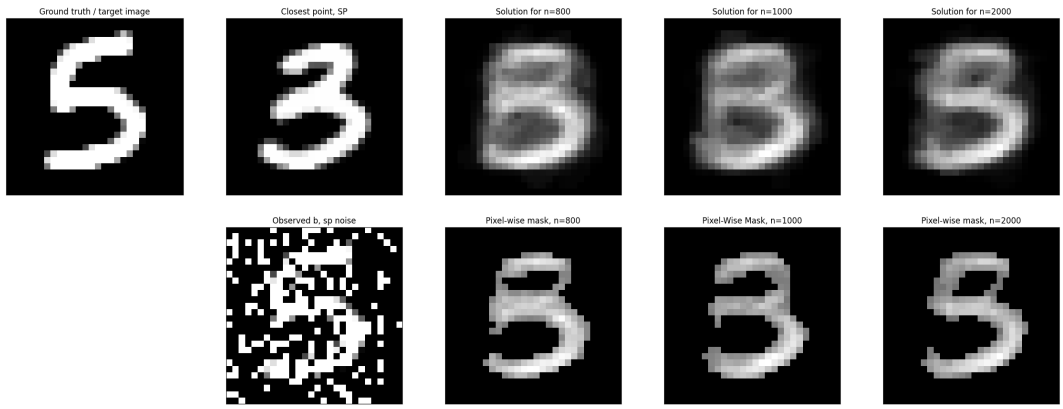

Figure 5: A specific type of failure case for small n.

We now move to the more expressive MNIST fashion dataset. We once again use a hand drawn target, which is not originally found in the dataset. To emphasize differences compared to handwritten digits, the fashion dataset is immediately more challenging, especially in regards to

fine details. To illustrate, there are few examples of garments with spots, stripes, or other detailed patterns - and as such are much more difficult to learn from uniform random samples. Similarly, if the target ground truth image contains details or patterns which are not present in the fashion dataset, there is little hope of constructing a reasonable linear combination to approximate the ground truth. On the other hand, some classes - such as heels or sandals - are extremely easy to learn, as there are many near-identical examples in the dataset, and are visually quite distinct from many other classes. Disregarding the change in dataset, our methodology for remains the same as Figure 1.

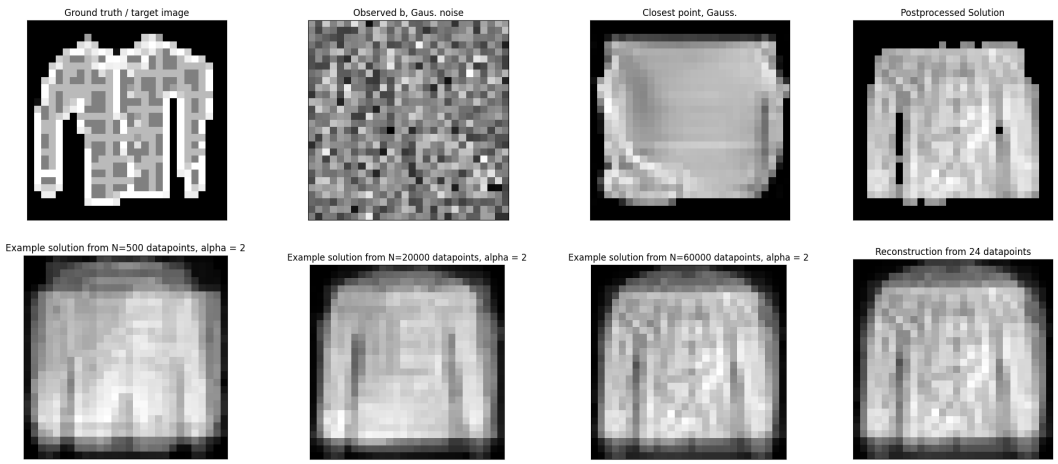

Figure 6: Recovery of a hand drawn shirt with additive Gaussian noise

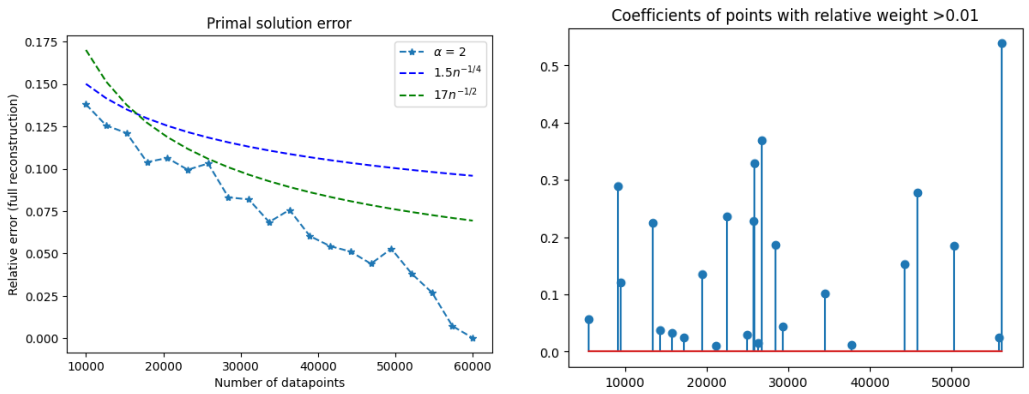

Figure 7: Rates and thresholding of optimal measure Q_n, for Figure 6

For the experiment with salt-and-pepper noise Figure 8, while MEM denoising clearly recovers a shirt, many of the finer details are lost or washed out. In contrast, with Gaussian noise Figure 6, there is some remnant of the shirts' pattern visible in the final reconstruction. In both cases the nearest neighbor is a bag or purse, and not visually close to the ground truth. While the constant leading constant is larger, once again we are firmly below the theoretically expected convergence rate of $n^{1/4}$, and solutions are compressible with respect to the optimal measure Q_n.

Finally, we conclude with an experiment on MNIST fashion with a hand drawn target of a heel, where we observe once again recovery well within the expected convergence rate, and visually recovers (after postprocessing) a reasonable approximation to the ground truth.

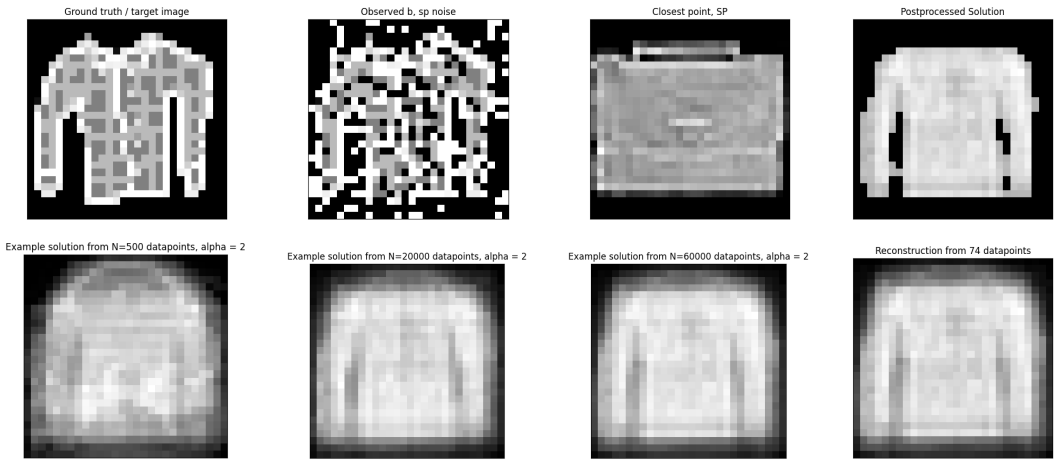

Figure 8: Recovery of a hand drawn shirt with salt and pepper corruption noise.

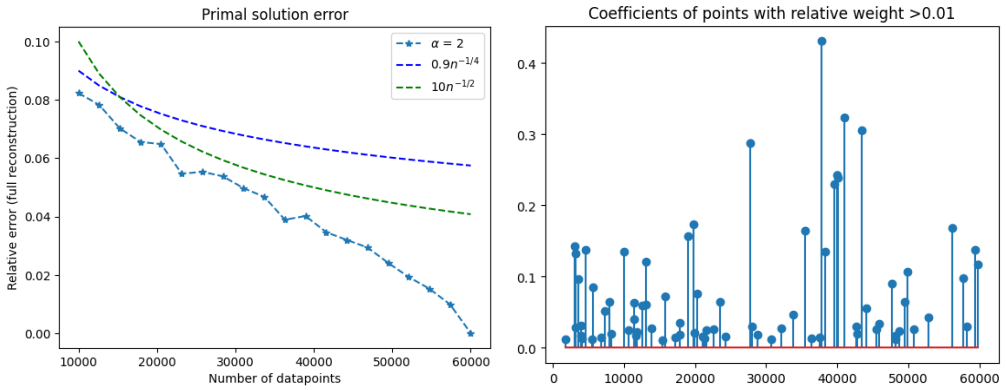

Figure 9: Rates and thresholding of optimal measure Q_n, for Figure 8

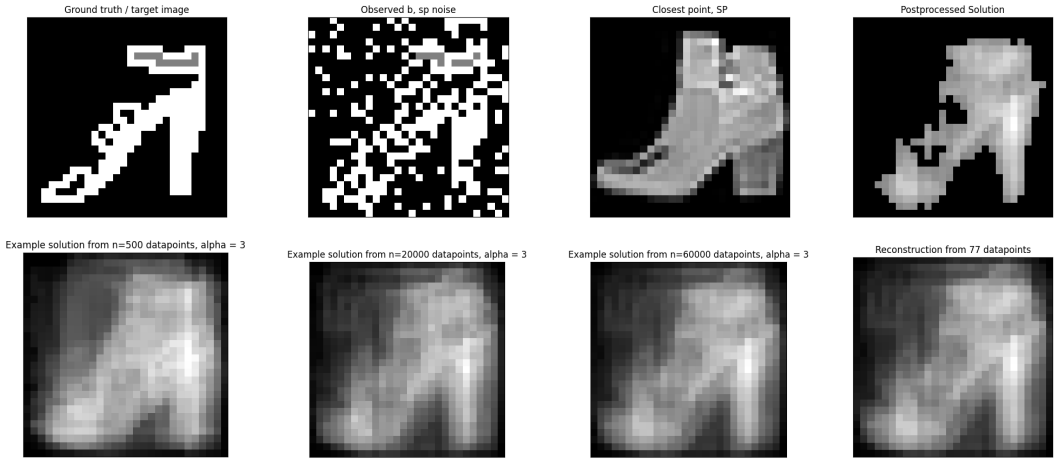

Figure 12: Recovery of hand drawn heel with Salt and Pepper corruption.

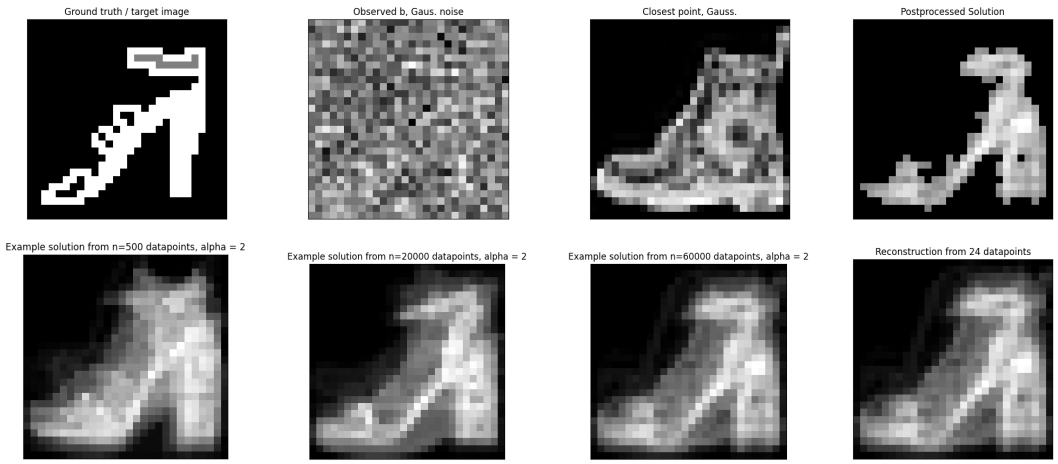

Figure 10: Recovery hand drawn heel with additive Gaussian noise.

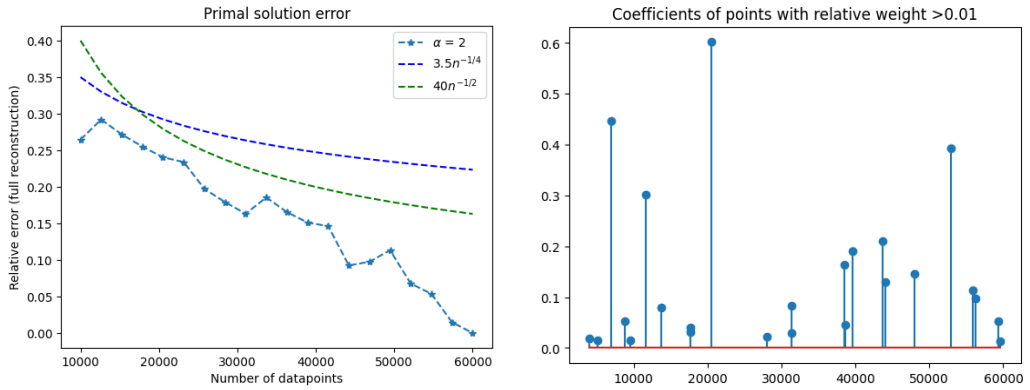

Figure 11: Rates and thresholding of optimal measure Q_n, for Figure 10

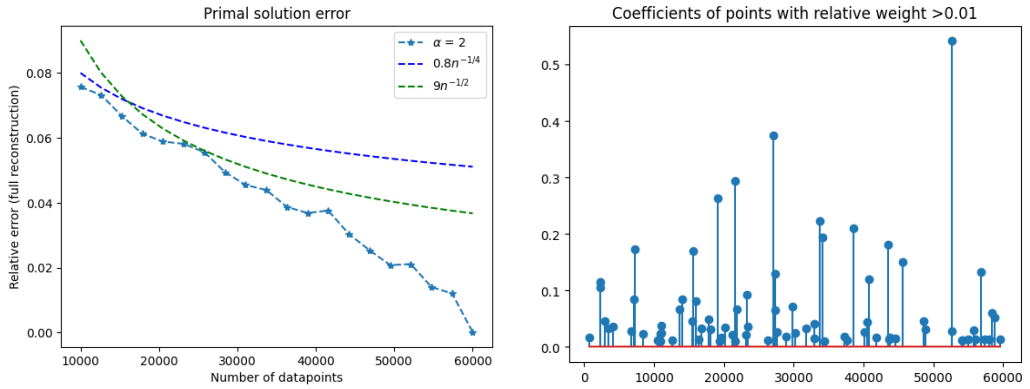

Figure 13: Rates and thresholding of optimal measure Q_n, for Figure 12

7. Conclusion and Future Work. Using the MEM framework, we have proposed using empirical priors $\mu_n^{(\omega)}$ derived from data to approximate the unknown solution $\overline{x}_\mu$. We have shown that this method has desirable theoretical properties, with Theorem 3.9 proving an almost sure convergence of $\overline{x}_{n,\varepsilon} \to \overline{x}_\mu$, while Theorem 4.7 and Theorem 5.5 give upper bounds on the error $\|x_{\nu,\varepsilon} - x_\mu\|$ for an

arbitrary prior ν in terms of epigraphical distances and a parametric rate $\|x_{n,\varepsilon} - x_\mu\| = O\left(\frac{1}{n^{1/4}}\right)$. One advantage of our approach here is that most of our results are valid not only for minimizers but also for ϵ-minimizers. We conjecture, however, that at least for exact minizers ($\epsilon = 0$) the rate result is not sharp and can be improved to $O(\frac{1}{\sqrt{n}})$ - as one might suspect from a function-valued central limit theorem.

As proof of concept, we numerically demonstrated the success of our method for denoising ($C = I$) based upon certain standard test data sets. These experiments were performed with modest computing resources and off-the-shelf optimization routines. In future work, our aim is to design specialized optimization routines that would give us access to moderate and large regimes in n and hence open the door to larger data sets. To this end, it would also be useful to reformulate our dual problem so that one can take advantage of stochastic gradient descent. In addition to denoising, it would be natural to present experiments for tasks such as deconvolution and inpainting.

Although we focused on the empirical approximation $\mu_n^{(\omega)}$, it would be natural to investigate more sophisticated approximations of μ. This is particularly true given that the approach taken here has been via epigraphical distances between the respective LMGFs.

Finally, in a certain sense this paper aims to learn a *regularizer* based upon a data set. It would also be interesting to approach the *fidelity* term, which can be understood as the MEM estimator based on a noise distribution (see [44]), in a similar vein. That is, can one learn the fidelity norm based upon samples of noise?

REFERENCES

[1] C. Amblard, E. Lapalme, and J.-M. Lina, *Biomagnetic source detection by maximum entropy and graphical models*, IEEE Trans. Biomed. Eng., 51 (2004), pp. 427–442.

[2] H. Attouch, *Convergence de fonctions convexes, des sous-différentiels et semi-groupes associés*, C.R. Acad. Sci. Paris, 284 (1977), pp. 539–542.

[3] H. Attouch, *Variational Convergence for Functions and Operators*, Pitman Advanced Pub. Program, Boston, 1984.

[4] A. Auslender and M. Teboulle, *Interior gradient and proximal methods for convex and conic optimization*, SIAM J. Optim., 16 (2006), pp. 697–725.

[5] H. Bauschke and P. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces.*, Springer, New York, 2019.

[6] P. Billingsley, *Probability and Measure*, John Wiley & Sons, Hoboken, NJ, 2017.

[7] L. D. Brown, *Fundamentals of Statistical Exponential Families: with Applications in Statistical Decision Theory*, Institute of Mathematical Statistics, Hayward, California, 1986.

[8] J. Burke and T. Hoheisel, *A note on epi-convergence of sums under the inf-addition rule*, Optimization Online, (2015).

[9] R. H. Byrd, J. Nocedal, and R. B. Schnabel, *Representations of quasi-newton matrices and their use in limited memory methods*, Math. Program., 63 (1994), pp. 129–156.

[10] Z. Cai, A. Machado, R. A. Chowdhury, A. Spilkin, T. Vincent, Ü. Aydin, G. Pellegrino, J.-M. Lina, and C. Grova, *Diffuse optical reconstructions of functional near infrared spectroscopy data using maximum entropy on the mean*, Scientific reports, 12 (2022). 2316.

[11] R. A. Chowdhury, J. M. Lina, E. Kobayashi, and C. Grova, *MEG source localization of spatially extended generators of epileptic activity: comparing entropic and hierarchical bayesian approaches*, PloS one, 8 (2013). E55969.

[12] S. Csörgő, *Kernel-transformed empirical processes*, J. Multivariate Anal., 13 (1983), pp. 517–533.

[13] D. Dacunha-Castelle and F. Gamboa, *Maximum d'entropie et problème des moments*, Ann. Inst. Henri Poincaré, Probab. Stat., 26 (1990), pp. 567–596.

[14] A. den Dekker and J. Sijbers, *Data distributions in magnetic resonance images: A review*, Phys. Med., 30 (2014), pp. 725–741.

[15] L. Deng, *The MNIST database of handwritten digit images for machine learning research [best of the web]*, IEEE Signal Process. Mag., 29 (2012), pp. 141–142.

[16] M. D. Donsker and S. S. Varadhan, *Asymptotic evaluation of certain Markov process expectations for large time. III*, Comm. Pure. Appl. Math., 29 (1976), pp. 389–461.

[17] A. Fermin, J.-M. Loubes, and C. Ludena, *Bayesian methods for a particular inverse problem seismic tomography*, Int. J. Tomogr. Stat, 4 (2006), pp. 1–19.

[18] A. Feuerverger, *On the empirical saddlepoint approximation*, Biometrika, 76 (1989), pp. 457–464.

[19] G. Folland, *Real Analysis: Modern Techniques and their Applications*, vol. 40, John Wiley & Sons, Hoboken, NJ, 1999.

[20] F. Gamboa, *Méthode du Maximum d'Entropie sur la Moyenne et Applications*, PhD thesis, Université Paris-Sud, 1989.

[21] C. Grova, J. Daunizeau, J.-M. Lina, C. G. Bénar, H. Benali, and J. Gotman, *Evaluation of EEG localization methods using realistic simulations of interictal spikes*, Neuroimage, 29 (2006), pp. 734–753.

[22] M. Heers, R. A. Chowdhury, T. Hedrich, F. Dubeau, J. A. Hall, J.-M. Lina, C. Grova, and E. Kobayashi, *Localization accuracy of distributed inverse solutions for electric and magnetic source imaging of interictal epileptic discharges in patients with focal epilepsy*, Brain Topography, 29 (2016), pp. 162–181.

[23] E. T. Jaynes, *Information theory and statistical mechanics*, Phys. Rev., 106 (1957), pp. 620–630.

[24] E. T. Jaynes, *Information theory and statistical mechanics. II*, Phys. Rev., 108 (1957), pp. 171–190.

[25] N. Kantas et al., *On Particle Methods for Parameter Estimation in State-Space Models*, Statist. Sci., 30 (2015), pp. 328 – 351.

[26] A. J. King and R. Wets, *Epi-consistency of convex stochastic programs*, Stoch. and Stoch. Rep., 34 (1991), pp. 83–92.

[27] A. Klenke, *Probability Theory: a Comprehensive Course*, Springer, Cham, Switzerland, 2013.

[28] S. Kullback and R. Leibler, *On information and sufficiency*, Ann. Math. Stat., 22 (1951), pp. 79–86.

[29] G. Le Besnerais, J.-F. Bercher, and G. Demoment, *A new look at entropy for solving linear inverse problems*, IEEE Trans. Inform. Theory, 45 (1999), pp. 1565–1578.

[30] P. Maréchal and A. Lannes, *Unification of some deterministic and probabilistic methods for the solution of linear inverse problems via the principle of maximum entropy on the mean*, Inverse Problems, 13 (1997).

[31] J. Navaza, *On the maximum-entropy estimate of the electron density function*, Acta Crystallogr. Sect. A, 41 (1985), pp. 232–244.

[32] J. Navaza, *The use of non-local constraints in maximum-entropy electron density reconstruction*, Acta Crystallogr. Sect. A, 42 (1986), pp. 212–223.

[33] E. Rietsch et al., *The maximum entropy approach to inverse problems-spectral analysis of short data records*

and density structure of the earth, J. Geophys., 42 (1976), pp. 489–506.
[34] G. Rioux, R. Choksi, T. Hoheisel, P. Maréchal, and C. Scarvelis, *The maximum entropy on the mean method for image deblurring*, Inverse Problems, 37 (2020). 015011.
[35] G. Rioux, C. Scarvelis, R. Choksi, T. Hoheisel, and P. Maréchal, *Blind deblurring of barcodes via Kullback-Leibler divergence*, IEEE Trans. on Pattern Anal. Mach. Intell., 43 (2021), pp. 77–88.
[36] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1997.
[37] R. T. Rockafellar and R. Wets, *Variational Analysis*, Springer, Berlin, 1998.
[38] R. T. Rockafellar and R. Wets, *Variational systems, an introduction*, in Multifunctions and Integrands: Stochastic Analysis, Approximation and Optimization Proceedings of a Conference held in Catania, Italy, June 7–16, 1983, Springer, 2006, pp. 1–54.
[39] W. Römisch and R. Wets, *Stability of ε-approximate solutions to convex stochastic programs*, SIAM J. Optim, 18 (2007), pp. 961–979.
[40] J. Royset and R. Wets, *An Optimization Primer*, Springer, Cham, Switzerland, 2021.
[41] T. Severini, *Elements of Distribution Theory*, Cambridge University Press, Cambridge, UK, 2005.
[42] A. Torralba and A. Oliva, *Statistics of natural image categories*, Network: Computation in Neural Systems, 14 (2003). 391.
[43] B. Urban, *Retrieval of atmospheric thermodynamical parameters using satellite measurements with a maximum entropy method*, Inverse Problems, 12 (1996). 779.
[44] Y. Vaisbourd, R. Choksi, A. Goodwin, T. Hoheisel, and C.-B. Schönlieb, *Maximum entropy on the mean and the cramér rate function in statistical estimation and inverse problems: properties, models, and algorithms*, Math. Program, in press, (2025).
[45] A. van der Vaart, *New donsker classes*, Ann. Probab., 24 (1996), pp. 2128–2140.
[46] A. van der Vaart and J. Wellner, *Weak Convergence*, Springer, 1996.
[47] H. Xiao, K. Rasul, and R. Vollgraf, *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. 2017, https://arxiv.org/abs/cs.LG/1708.07747.
[48] C. Zălinescu, *Convex Analysis in General Vector Spaces*, World Scientific, Singapore, 2002.