# On the Computational and Statistical Efficiency of the Empirical Maximum Entropy on the Mean Method

**Matthew King-Roskamp**[†] MATTHEW.KING-ROSKAMP@MAIL.MCGILL.CA
*Department of Mathematics and Statistics*
*McGill University*
*Montreal, Quebec H3A 0G4, Canada*

**Gabriel Rioux**[†] G.RIOUX@IMPERIAL.AC.UK
*Department of Mathematics*
*Imperial College London*
*South Kensington, London SW7 2AZ, United Kingdom*

**Rustum Choksi** RUSTUM.CHOKSI@MCGILL.CA
*Department of Mathematics and Statistics*
*McGill University*
*Montreal, Quebec H3A 0G4, Canada*

**Tim Hoheisel** TIM.HOHEISEL@MCGILL.CA
*Department of Mathematics and Statistics*
*McGill University*
*Montreal, Quebec H3A 0G4, Canada*

## Abstract

The *Maximum Entropy on the Mean (MEM)* method provides a flexible computational framework for solving inverse problems by combining data fidelity with entropy-based regularization. In practice, however, the prior distribution is typically unknown but can be estimated from data, giving rise to the empirical MEM method. We establish a parametric convergence rate of $O(n^{-1/2})$ in expectation for empirical MEM, improving upon the previously established $O(n^{-1/4})$ guarantee by King-Roskamp et al. (2026). Our proof is based on a novel stability analysis of the primal and dual optimization problems under perturbations of the underlying probability measure, relying only on foundational tools from convex analysis and probability. We further show that the MEM dual problem admits a reformulation as an expected risk minimization problem, thereby placing MEM within the modern framework of stochastic optimization and enabling scalable stochastic gradient algorithms for large-scale inverse problems. Together, these results place empirical MEM as a statistically and computationally efficient methodology for data-driven inverse problems.

**Keywords:** Maximum entropy on the mean method, inverse problems, Fenchel-Rockafellar duality, empirical measures, statistical stability, stochastic optimization, stochastic subgradient method

## 1 Introduction

Linear inverse problems of the form

$$Ax \approx b$$

[†] Denotes equal contribution to this work.

are fundamental in statistics, signal processing, machine learning (in particular imaging). Such problems are frequently ill-conditioned or underdetermined, and the observation may additionally be corrupted by noise. Stable recovery therefore requires regularization, often via the incorporation of prior information about the unknown signal.

The *Maximum Entropy on the Mean (MEM) method* provides a convex and information-theoretic framework for incorporating such prior information; see, e.g., (Le Besnerais et al., 1999; Rioux et al., 2021; Vaisbourd et al., 2025). Given a prior probability measure $\mu \in \mathcal{P}(\mathcal{X})$ supported on a compact set $\mathcal{X} \subset \mathbb{R}^m$, MEM associates with $\mu$ the (convex) function

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

which we call the *MEM function*, see, e.g., (Vaisbourd et al., 2025) for an extensive study of this object, and Section 2.3 for more details. The corresponding reconstruction is obtained by solving

$$x_\mu \in \operatorname*{argmin}_{x\in\mathbb{R}^m} \{\alpha g(Ax - b) + \kappa_\mu(x)\}, \tag{$P_\mu$}$$

where $g$ is a proper, lower semicontinuous, and convex data-fidelity function and $\alpha > 0$ balances data fidelity against prior information/regularization. This is the *MEM method* in a nutshell, see Sections 2.2 and 2.3 for more details.

Emerging from ideas of E.T. Jaynes in 1957 (Jaynes, 1957a,b), various forms and interpretations of MEM (see Brown, 1986; Gamboa, 1989; Maréchal and Lannes, 1997; Dacunha-Castelle and Gamboa, 1990; Le Besnerais et al., 1999) have appeared in the literature. Applications have occurred in different disciplines such as earth sciences (Fermin et al., 2006; Rietsch, 1976; Urban, 1996), crystallography (Navaza, 1985, 1986), and medical imaging (Amblard et al., 2004; Cai et al., 2022; Chowdhury et al., 2013; Grova et al., 2006; Heers et al., 2016). 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) (Rioux et al., 2021, 2020). 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.

The key to approaching the MEM problem theoretically and computationally is Fenchel-Rockafellar duality (Rockafellar, 1970). The (Fenchel) dual of the (primal) MEM problem $(P_\mu)$ reads:

$$z_\mu \in \operatorname*{argmin}_{z\in\mathbb{R}^d} \Big\{\alpha g^*(-z/\alpha) - \langle b, z\rangle + L_\mu(A^\top z)\Big\}, \tag{$D_\mu$}$$

where $L_\mu(y) := \log \int_{\mathcal{X}} \exp\langle y, \cdot\rangle d\mu$ is the log-moment generating function of the prior $\mu$. The advantages of appealing to the dual problem $(D_\mu)$ instead of the primal MEM problem $(P_\mu)$ are as follows: The log-moment generating function has closed form for many priors $\mu$ and is essentially smooth (in the sense of Rockafellar (1970)) thus enabling expedient primal-dual recovery

$$x_\mu = \nabla L_\mu(A^\top z_\mu). \tag{1}$$

In many applications, however, the prior distribution $\mu$ is unknown. What is available instead is a collection of independent training samples from $\mu$

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mu.$$

A natural data-driven construction replaces $\mu$ by an empirical measure

$$\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}. \tag{2}$$

The MEM method based on this measure, which we call the *empirical MEM method*, was analyzed in King-Roskamp et al. (2026). That work established almost sure convergence of empirical MEM solutions $x_{\hat{\mu}_n}$ to $x_\mu$, including suitable approximate solutions, and obtained finite-sample convergence guarantees in expectation. To wit, an $n^{-1/4}$ sample complexity with the quadratic data-fidelity term, $g = \frac{1}{2}\|\cdot\|^2$, was derived, but it was conjectured that this rate was not sharp and that the parametric rate $n^{-1/2}$ should hold.

The first objective of the present paper is to resolve this question: We prove that the empirical MEM estimator converges in expectation at the parametric rate

$$\mathbb{E}\left[\|x_{\hat{\mu}_n} - x_\mu\|\right] = O\left(n^{-1/2}\right) \tag{3}$$

for all convex and $L$-smooth fidelity terms (see Assumption (A1) below). Similar to the analysis in King-Roskamp et al. (2026), we heavily rely on Fenchel-Rockafellar duality: First, we establish stability results for solutions of the (Fenchel-Rockafellar) dual MEM problem $(D_\mu)$ under perturbations of the underlying probability measure (Theorem 12). Using this, we then prove the $n^{-1/2}$-rate of convergence of empirical dual solutions in expectation. Moreover, by establishing that the primal-dual recovery (1) is Lipschitz-stable (Lemma 14), we obtain stability of primal solutions with respect to the underlying measure (Proposition 16) and, ultimately, the desired convergence from (3) (Theorem 17). Moreover, we show that, for an $\varepsilon$-optimal empirical dual solution $z_{\hat{\mu}_n,\varepsilon}$ and its associated primal reconstruction $x_{\hat{\mu}_n,\varepsilon} := \nabla L_{\hat{\mu}_n}(A^\top z_{\hat{\mu}_n,\varepsilon})$, the bound

$$\mathbb{E}\left[\|x_{\hat{\mu}_n,\varepsilon} - x_\mu\|\right] \le C_1 n^{-1/2} + C_2\sqrt{\varepsilon}, \tag{4}$$

holds for constants independent of $n$ and $\varepsilon$ (Theorem 18).

The second objective of the paper is computational: The empirical prior in (2) is attractive precisely when a large training set is available, but the standard empirical MEM dual involves the full sample at every evaluation. Consequently, deterministic full-batch methods become increasingly costly as $n$ grows. To address this limitation, we show that the empirical dual can be reformulated as a risk minimization problem. Concretely, we show that with $\ell(x,z) = \exp\left(\alpha g^*(-z/\alpha) - \langle b, z\rangle + \langle A^\top z, x\rangle\right)$, the MEM dual problem $(D_\mu)$ has the same solution(s) as

$$\min_{z\in\mathbb{R}^d} \mathbb{E}_{X\sim\mu}[\ell(X,z)] \tag{5}$$

Consequently, the empirical MEM dual can be cast as

$$\min_{z\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell(X_i,z). \tag{6}$$

This reformulation places empirical MEM within the standard framework of stochastic optimization. It permits stochastic-gradient updates based on individual samples or mini-batches and removes the need to process the entire prior dataset at every iteration.

The exponential terms inherited from the moment generating function create a difficulty that is specific to this formulation: stochastic gradients need not be uniformly bounded and may become numerically unstable when the iterates move into unfavorable regions. Motivated by stochastic gradient clipping methods, including the analysis of Mai and Johansson (2021), we study a projected clipped stochastic-gradient scheme adapted to the empirical MEM objective. In this extended real-valued setting, we refine the previous treatment to explicitly account for domain considerations on $\ell$. We establish almost sure convergence under explicit assumptions and investigate the method numerically on inverse problems of increasing scale. The experiments demonstrate that the stochastic formulation can use substantially larger empirical priors than the original full-batch formulation while retaining competitive reconstruction quality.

The statistical and computational contributions are complementary. The statistical analysis controls the error introduced by replacing the population prior $\mu$ with the empirical prior $\hat{\mu}_n$. The optimization analysis controls the additional error introduced by solving the empirical problem only approximately, cf. (4). This decomposition separates the statistical error from the optimization error and gives a natural stopping principle: solving the empirical problem far beyond the accuracy permitted by the sample size provides little statistical benefit.

To summarize, our main contributions are the following:

(i) We prove nonasymptotic stability bounds for the dual and primal MEM solutions under perturbations of the prior measure.

(ii) We establish the parametric convergence rate $O(n^{-1/2})$ in expectation for empirical dual and primal solutions, improving the previously known $O(n^{-1/4})$ rate from King-Roskamp et al. (2026).

(iii) We extend the statistical analysis to approximate empirical solutions and obtain an explicit decomposition of statistical and optimization errors.

(iv) We reformulate the empirical MEM dual as an empirical risk minimization problem, thereby enabling stochastic optimization methods whose per-iteration cost does not scale with the full sample size.

(v) We develop and analyze a projected clipped stochastic-(sub)gradient method for the reformulated problem and demonstrate its performance on large-scale image reconstruction examples.

The paper is organized as follows: In Section 2 we collect the required material from convex analysis and probability theory, and provide some more details on the MEM method. Section 3 develops the stability theory and proves the parametric convergence rates for exact and approximate empirical solutions. Section 4 derives the expected-risk reformulation and studies stochastic-subgradient methods for the empirical dual. Section 5 presents the numerical experiments.

*Notation:* Throughout we equip $\mathbb{R}^m$ with the Euclidean norm $\|\cdot\| := \|\cdot\|_2$. The (Euclidean) distance of a point $x \in \mathbb{R}^m$ to a set $C \in \mathbb{R}^m$ is denoted by $\operatorname{dist}(x, C)$. The closed ball of radius

$r$ (in the Euclidean norm) centered at 0 is denoted by $B_r$. For a matrix $A \in \mathbb{R}^{d\times m}$ we denote $\|A\|$ as its operator 2-norm (i.e., its largest singular value), and $\sigma_{\min}(A) = \sqrt{\lambda_{\min}(A^\top A)}$ as the smallest singular value of $A$. For any $C \subset \mathbb{R}^d$ set $|C|_\infty := \sup_{c\in C} \|c\|$.

## 2 Preliminaries

In this section we review some concepts from convex analysis and probability theory that are essential for our study. We then recall the MEM problem briefly introduced above with some more technical details.

### 2.1 Fundamentals from convex analysis

Throughout, we work with extended real-valued functions $f : \mathbb{R}^m \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$. The *domain* of a function $f$ is the set $\mathrm{dom}(f) = \{x \in \mathbb{R}^m \ : f(x) < +\infty\}$. If $\mathrm{dom}(f) \neq \emptyset$ and $f$ is never $-\infty$, $f$ is called *proper*. A function $f$ is said to be *lower semicontinuous (lsc)* if the preimage $f^{-1}([-\infty, a])$ is a closed (possibly empty) set for all $a \in \mathbb{R}$. We say that a proper function $f$ is *convex* if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$

for every $x, y \in \mathrm{dom}(f)$ and all $\lambda \in (0, 1)$.

For $f : \mathbb{R}^m \to \overline{\mathbb{R}}$, its *subdifferential* at $\overline{x}$ is the collection of all *subgradients* of $f$ at $\overline{x}$, given by

$$\partial f(\overline{x}) := \{v \in \mathbb{R}^m \ : \ f(x) \geq f(\overline{x}) + \langle v, x - \overline{x}\rangle, \ \forall x \in \mathbb{R}^m\}, \tag{7}$$

which reduces to singleton $\{\nabla f(\bar{x})\}$ when $f$ is convex and differentiable at $\overline{x}$. We defer to Rockafellar (1970) for more details on subdifferential calculus.

The *(Fenchel) conjugate* of $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ is the function $f^* : \mathbb{R}^m \to \overline{\mathbb{R}}$ defined as

$$f^*(x^*) := \sup_{x\in\mathbb{R}^m} \{\langle x, x^*\rangle - f(x)\}.$$

An important notion for our study is *strong convexity*. The following result records useful properties of strongly convex functions from standard reference texts, see, e.g., (Rockafellar and Wets, 1998).

**Proposition 1** (Strong convexity)**.** *Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be lsc proper convex and let $\sigma > 0$. The following hold:*

*i) (Characterizations) The following are equivalent:*

1. *$f$ is $\sigma$-strongly convex.*
2. *$f^*$ is $1/\sigma$-*smooth*, i.e., $\nabla f^*$ is (globally) $1/\sigma$-Lipschitz.*
3. *$\partial f$ is* strongly monotone*, i.e., for all $x_1, x_2 \in \mathbb{R}^m$:*

$$\langle v_1 - v_2, x_1 - x_2\rangle \geq \sigma\|x_1 - x_2\|^2, \qquad \forall v_1 \in \partial f(x_1), v_2 \in \partial f(x_2). \tag{8}$$

4. *For all $x_1, x_2 \in \mathbb{R}^m$:*

$$f(x_2) - f(x_1) \geq \langle v, x_2 - x_1\rangle + \frac{\sigma}{2}\|x_2 - x_1\|^2, \quad \forall v \in \partial f(x_1). \tag{9}$$

*ii) (Unique minimizers) If $f$ is $\sigma$-strongly convex it has a unique minimizer $\bar{x}$ that satisfies*

$$f(x) - f(\bar{x}) \geq \frac{\sigma}{2}\|x - \bar{x}\|^2, \quad \forall x \in \mathbb{R}^m .$$

*iii) When $f$ is differentiable and $\sigma$-strongly convex, we have*

$$\sigma\|x_1 - x_2\| \leq \|\nabla f(x_1) - \nabla f(x_2)\| \quad \forall x_1, x_2 \in \mathbb{R}^m .$$

The central convex-analytic tool for our study is Fenchel-Rockafellar duality which we briefly recall here, see, e.g., Rockafellar and Wets (1998) for more details.

**Theorem 2** (Fenchel-Rockafellar duality)**.** *Let $\kappa : \mathbb{R}^\ell \to \overline{\mathbb{R}}$, $\phi : \mathbb{R}^k \to \overline{\mathbb{R}}$ be proper, lsc, and convex and $L \in \mathbb{R}^{k\times\ell}$. Define*

$$\min_{x\in\mathbb{R}^\ell} \phi(Lx) + \kappa(x) \tag{P}$$

*and*

$$\min_{y\in\mathbb{R}^k} \kappa^*(L^\top y) + \phi^*(-y). \tag{D}$$

*Set*

$$p := \inf_{x\in\mathbb{R}^\ell}\{\phi(Lx) + \kappa(x)\} \quad \text{and} \quad d := -\inf_{y\in\mathbb{R}^k}\{\kappa^*(L^\top y) + \phi^*(-y)\},$$

*and assume the qualification condition*

$$0 \in \operatorname{int}\left(L(\operatorname{dom}(\kappa)) - \operatorname{dom}(\phi)\right).$$

*Then the following hold:*

*i) (Strong duality) $p = d$, and the dual problem has a solution.*

*ii) (Primal-dual recovery) If $\bar{y}$ solves the dual problem* (D) *and $\kappa^*$ is differentiable (at $L^\top\bar{y}$), then $\bar{x} = \nabla\kappa^*(L^\top\bar{y})$ (uniquely) solves the primal problem* (P)*.*

### 2.2 Basic notions from probability theory

Throughout this paper, we fix a compact set $\mathcal{X} \subset \mathbb{R}^m$ and let $\mathcal{P}(\mathcal{X})$ denote the set of Borel probability measures on $\mathcal{X}$. For any two probability measures $\nu, \mu \in \mathcal{P}(\mathcal{X})$, $\nu + \mu$ is a finite Borel measure. Hence, for any $f$ which is integrable with respect to $\mu$ and $\nu$ we often write $\int f d\nu + \int f d\mu$ as $\int f d(\nu + \mu)$ and similarly for differences of Borel measures.

The *expectation* of a random variable $X$ with distribution $\mu \in \mathcal{P}(\mathcal{X})$ is

$$\mathbb{E}[X] = \int_{\mathcal{X}} x d\mu(x) \in \mathbb{R}^m .^*$$

We sometimes write $\mathbb{E}_\mu[X]$ to stress that the expectation is a function of the underlying distribution. The *covariance matrix* of $X$ is

$$\operatorname{cov}(X) = \mathbb{E}_\mu\left[(X - \mathbb{E}_\mu[X])(X - \mathbb{E}_\mu[X])^\top\right] \in \mathbb{R}^{m\times m}, \tag{10}$$

*This vector-valued integral is to be understood component-wise which is consistent with the usual definition $\mathbb{E}[X] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_m])^\top$ of the random vector $X$. This idea also extends to random matrices.

which we sometimes write as $\mathrm{cov}_\mu(X)$. For scalar random variables (i.e., $m = 1$), we instead write var for the *variance*.

The *Kullback-Leibler (KL) divergence* (Kullback and Leibler, 1951) of $Q \in \mathcal{P}(\mathcal{X})$ with respect to $\mu \in \mathcal{P}(\mathcal{X})$ is defined as

$$\mathrm{KL}(Q\|\mu) := \begin{cases} \int_{\mathcal{X}} \log(\frac{dQ}{d\mu}) dQ, & \text{if } Q \ll \mu, \\ +\infty, & \text{otherwise.} \end{cases} \tag{11}$$

### 2.3 The MEM problem

As mentioned in the introduction, the *Maximum Entropy on the Mean (MEM) function* is given by

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

where $\mu \in \mathcal{P}(\mathcal{X})$ is a given prior distribution. Given a signal $b \in \mathbb{R}^d$ and an operator $A \in \mathbb{R}^{d\times m}$, the MEM problem, introduced in (Le Besnerais et al., 1999), consists of solving the convex optimization problem

$$x_\mu = \operatorname*{argmin}_{x\in\mathbb{R}^m} \{\alpha g(Ax - b) \; + \; \kappa_\mu(x)\}, \tag{$P_\mu$}$$

where $g$ is a lsc, proper, and convex function which measures the discrepancy between $Ax$ and $b$ and $\alpha > 0$ is a hyperparameter that enables balancing the influence of both terms in the objective. Therefore, the MEM problem aims to recover a signal based on a transformed and perhaps noisy observation thereof. The function $g$ is generally used to enforce the fidelity of the signal, $Ax \approx b$, whereas the MEM function enables incorporating prior beliefs about the distribution of the signal. As shown in, e.g., King-Roskamp et al. (2026, Lemma 2.1), this problem has a unique solution under the qualification condition $0 \in \mathrm{int}(\mathrm{dom}(g) - A\,\mathrm{dom}(\kappa_\mu))$ and solutions can be obtained by first solving the Fenchel-Rockafellar dual problem

$$z_\mu \in \operatorname*{argmin}_{z\in\mathbb{R}^d} \phi_\mu(z) := \alpha g^*(-z/\alpha) - \langle b, z\rangle + L_\mu(A^\top z), \tag{$D_\mu$}$$

Here $L_\mu$ is the log-moment generating function (also known as the cumulant generating function) of $\mu$ given by

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

Under our standing assumption that $\mathcal{X}$ is compact, we record the following key properties of $L_\mu$ and $\kappa_\mu$ for ease of reference.

**Lemma 3.** *For compact $\mathcal{X}$, let $\mu \in \mathcal{P}(\mathcal{X})$. The following hold:*

*(i) $L_\mu$ is convex, finite-valued, and analytic. In particular, $L_\mu$ is continuously differentiable with gradient*

$$\nabla L_\mu(y) = \frac{\int (\cdot) \exp\langle y, \cdot\rangle d\mu}{\int \exp\langle y, \cdot\rangle d\mu}. \tag{12}$$

*(ii)* $\kappa_\mu = L_\mu^*$ *and* $\kappa_\mu^* = L_\mu$.

*(iii)* $\lim_{\|x\|\to\infty} \frac{\kappa_\mu(x)}{\|x\|} = \infty$, *and* $\kappa_\mu$ *is strictly convex on every convex subset of* $\{x \,|\, \partial\kappa_\mu(x) \neq \emptyset\}$. *Namely,* $\kappa_\mu$ *is supercoercive and essentially strictly convex.*

These properties are collected in, e.g., King-Roskamp et al. (2026, Section 2.2).

Given a solution $z_\mu$ to $(D_\mu)$, the solution, $x_\mu$, to $(P_\mu)$ can be obtained via the primal-dual recovery formula given by Theorem 2,

$$x_\mu = \nabla L_\mu(A^\top z_\mu). \tag{13}$$

It is noteworthy that the dual solution $z_\mu$ can also be used to recover

$$\bar{Q} = \operatorname{argmin}\{\mathrm{KL}(Q \,\|\, \mu) : \mathbb{E}_Q[X] = x_\mu, Q \in \mathcal{P}(\mathcal{X})\},$$

i.e., the optimal distribution in the definition of $\kappa_\mu(x_\mu)$, see (Rioux et al., 2020, Lemma 3). Namely, it is given by $\bar{Q} = \gamma_{\mu,z_\mu}$ where, for each $z \in \mathbb{R}^d$, $\gamma_{\mu,z} \in \mathcal{P}(\mathcal{X})$ is the unique probability distribution which admits the Radon-Nikodym derivative

$$\frac{d\gamma_{\mu,z}}{d\mu} : u \in \mathcal{X} \mapsto \frac{\exp\langle A^\top z, u\rangle}{\int \exp\langle A^\top z, \cdot\rangle d\mu}. \tag{14}$$

In effect, in light of the primal-dual recovery condition, it is easily seen that the expectation of $\gamma_{\mu,z_\mu}$ is indeed $x_\mu$. Precisely, we observe that

$$\mathbb{E}_{\gamma_{\mu,z_\mu}}[X] = \frac{\int u \exp\langle A^\top z_\mu, u\rangle d\mu(u)}{\int \exp\langle A^\top z_\mu, u\rangle d\mu(u)} = \nabla L_\mu(A^\top z_\mu) = x_\mu. \tag{15}$$

Here the left equality is found by applying $\mathbb{E}_{\gamma_{\mu,z_\mu}}[X] = \int x d\gamma_{\mu,z_\mu}(x) = \int x \frac{d\gamma_{\mu,z_\mu}}{d\mu}(x) d\mu(x)$ as follows from the Radon-Nikodym theorem, see e.g., (Dudley, 2018, Theorem 5.5.4), and the right equality follows from evaluating (12) at $A^\top z_\mu$.

Throughout this work we make the following mild standing assumption:

$$g : \mathbb{R}^d \to \mathbb{R} \text{ is convex and } \frac{1}{\beta}\text{-smooth, i.e., } \nabla g \text{ is (globally) } \frac{1}{\beta}\text{-Lipschitz.} \tag{A1}$$

Moreover, (A1) implies that $g^*$ is $\beta$-strongly convex which, in particular, yields that

The dual objective function $\phi_\mu$ is $\frac{\beta}{\alpha}$-strongly convex for all $\mu \in \mathcal{P}(\mathcal{X})$.

The prototypical example $g = \frac{1}{2}\|\cdot\|^2$ satisfies (A1), but there are other examples of interest beyond this. For example, any $g$ which is the Moreau envelope of another proper lsc convex function satisfies (A1) (Rockafellar and Wets, 1998, Example 10.32). From the machine learning literature, the softplus $g(x) = \log(1+e^{\langle a,x\rangle})$ with $a \in \mathbb{R}^d$ and the radial loss $g(x) = \sqrt{1+\|x\|^2}$ also satisfy this condition. Finally, we may also consider any separable loss functions which satisfies the assumptions component-wise, e.g., $g = \sum_{i=1}^d g_i(x_i)$ for

$g_i$ convex and $\frac{1}{\beta_i}$-smooth. This last category includes, for example, the log-cosh loss $g_i = \log\cosh(\cdot)$ with $\beta_i = 1$.

As noted in Section 1, in the case where $\mu$ is unknown – such as in imaging problems – a natural approach is to approximate $\mu$ via observed data samples. Explicitly, given samples

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mu,$$

the *empirical distribution*,

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i},$$

serves as a natural surrogate for a population-level prior distribution. This choice of prior leads to an *empirical dual problem*

$$z_{\hat{\mu}_n} \in \operatorname*{argmin}_{z\in\mathbb{R}^d} \Big\{ \alpha g^*\Big(-\frac{z}{\alpha}\Big) - \langle b, z\rangle + L_{\hat{\mu}_n}(A^\top z)\Big\}$$

for the log-moment generating $L_{\hat{\mu}_n}$ of $\hat{\mu}_n$.

## 3 Parametric rates for convergence in expectation

In the sequel, we establish the parametric mean rate of convergence of the empirical primal and dual solutions towards their population-level counterparts. The main result of this section is as follows.

**Theorem 4** (Primal and dual rates)**.** *Fix $\alpha > 0$, a fidelity function $g : \mathbb{R}^d \to \mathbb{R}$ satisfying* (A1)*, a prior $\mu \in \mathcal{P}(\mathcal{X})$, and let $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mu$. Further, let $z_{\hat{\mu}_n,\varepsilon}$ be a $\varepsilon$-approximate solution to the (empirical) dual MEM problem for $\hat{\mu}_n$ for some $\varepsilon \geq 0$ , i.e.,*

$$\phi_{\hat{\mu}_n}(z_{\hat{\mu}_n,\varepsilon}) \leq \inf_z \phi_{\hat{\mu}_n}(z) + \varepsilon,$$

*which is obtained by a measurable selection from the set of all $\varepsilon$-approximate solutions. Then, letting $z_\mu$ denote the solution of the (population) dual MEM problem for $\mu$,*

$$\mathbb{E}\left[\|z_{\hat{\mu}_n,\varepsilon} - z_\mu\|\right] \leq \frac{C}{\sqrt{n}} + \sqrt{\frac{2\alpha}{\beta}\varepsilon}, \tag{16}$$

*where $C = \frac{\alpha}{\beta}\|A\|C'$ for $C' = 2|\operatorname{supp}(\mu)|_\infty \exp\left(2\|A\|\|z_\mu\||\operatorname{supp}(\mu)|_\infty\right)$. Moreover, setting $K' = |\operatorname{supp}(\mu)|_\infty^2\|A\|$, if $x_{\hat{\mu}_n,\varepsilon} := \nabla L_{\hat{\mu}_n}(A^\top z_{\hat{\mu}_n,\varepsilon})$ is the recovered (approximate) primal solution for $\hat{\mu}_n$ and $x_\mu$ is the primal solution for $\mu$,*

$$\mathbb{E}\left[\|x_{\hat{\mu}_n,\varepsilon} - x_\mu\|\right] \leq \frac{K'C + C'}{\sqrt{n}} + K'\sqrt{\frac{2\alpha}{\beta}\varepsilon}. \tag{17}$$

*If, moreover, $A$ is injective, and $g$ is $\frac{1}{\gamma}$-strongly convex, then*

$$\mathbb{E}\left[\|x_{\hat{\mu}_n,\varepsilon} - x_\mu\|\right] \leq \frac{\gamma C}{\alpha\sigma_{\min}(A)}\frac{1}{\sqrt{n}} + K'\sqrt{\frac{2\alpha}{\beta}\varepsilon}.$$

We first comment on the condition that solutions are obtained via a measurable selection.

**Remark 5** (Measurable selection)**.** *Fix $n \in \mathbb{N}$ and consider the function*

$$F_n : (x_1, \dots, x_n) \times z \in \mathcal{X}^n \times \mathbb{R}^d \mapsto \alpha g^*\Big(-\frac{z}{\alpha}\Big) - \langle b, z\rangle + \log\Big(\frac{1}{n}\sum_{i=1}^n \exp\langle z, Ax_i\rangle\Big),$$

*which is evidently (jointly) lsc. By Rockafellar and Wets (1998, Example 14.31), $F_n$ is a normal integrand so that*

$$(x_1, \dots, x_n) \in \mathcal{X}^n \mapsto \{z \in \mathbb{R}^d : F_n(x_1, \dots, x_n, z) \leq \inf_{z\in\mathbb{R}^d} F_n(x_1, \dots, x_n, z) + \varepsilon\}$$

*is closed-valued and measurable, noting that the infimum is always attained by strong convexity of the objective, see Proposition 14.33 and Theorem 14.37 from the same reference. By Rockafellar and Wets (1998, Corollary 14.6), there exists a measurable selection of $\varepsilon$-minimizer for $F_n(x_1, \dots, x_n, \cdot)$ and, since the data generating process $\omega \in \Omega \mapsto (X_1(\omega), \dots, X_n(\omega)) \in \mathcal{X}^n$ is measurable (by definition), there exists a measurable selection of the dual $\varepsilon$-minimizers $z_{\hat{\mu}_n,\varepsilon}$ as the composition of Borel measurable maps is measurable. It follows that the corresponding selection of primal solutions is measurable. If $\varepsilon = 0$, no selection is required as each problem admits a unique solution.*

*Practically, any algorithm which takes the samples and returns a $\varepsilon$-minimizer based on an iterative application of measurable operations is a measurable selection rule provided that the initialization and stopping rules are also measurable.*

The proof of Theorem 4 follows the following main steps.

1. We show that for any $\rho, \eta \in \mathcal{P}(\mathcal{X})$, the unique minimizers $z_\rho, z_\eta$, of $\phi_\rho, \phi_\eta$, satisfy

$$\|z_\eta - z_\rho\| \leq C_1 \left| \int \exp\langle A^\top z_\eta, \cdot\rangle d(\eta - \rho)\right| + C_2 \left\| \int (\cdot) \exp\langle A^\top z_\eta, \cdot\rangle d(\rho - \eta)\right\|, \quad (18)$$

   for constants $C_1, C_2 \geq 0$ which are independent of the choice of $\rho$ and $\eta$, see Corollary 10 and Lemma 11. This is accomplished by leveraging the strong monotonicity of the subgradients of $\phi_\rho$ and $\phi_\eta$ and the fact that the subgradients depend on $\rho, \eta$ only through the terms $\nabla L_\rho$ and $\nabla L_\eta$ whose difference can be controlled in terms of the integral terms in (18), see Lemma 6.

2. Next, for a prior $\mu \in \mathcal{P}(\mathcal{X})$ and the empirical measure $\hat{\mu}_n$ from $n$ i.i.d. samples $X_1, \dots, X_n$ from $\mu$, it follows from (18) that $\mathbb{E}\big[\|z_{\hat{\mu}_n} - z_\mu\|\big]$ is bounded above by

$$C_1\sqrt{\operatorname{var}\Big[\frac{1}{n}\sum_{i=1}^n \exp\langle A^\top z_\mu, X_i\rangle\Big]} + C_2\sqrt{\operatorname{tr}\left(\operatorname{cov}\Big[\frac{1}{n}\sum_{i=1}^n X_i \exp\langle A^\top z_\mu, X_i\rangle\Big]\right)},$$

   so that both terms scale as $n^{-1/2}$ upon showing that that the variance/covariance of the summands are uniformly bounded, see Theorem 12. To this effect, we show in Theorem 8 that $\|z_\mu\|$ can be bounded *a priori*. Applying the triangle inequality,

$$\mathbb{E}\big[\|z_{\hat{\mu}_n,\varepsilon} - z_\mu\|\big] \leq \mathbb{E}\big[\|z_{\hat{\mu}_n} - z_\mu\|\big] + \mathbb{E}\big[\|z_{\hat{\mu}_n,\varepsilon} - z_{\hat{\mu}_n}\|\big],$$

   where the first term on the right-hand side is bounded as described above whereas the second is bounded as $\sqrt{\frac{2\alpha}{\beta}(\phi_{\hat{\mu}_n}(z_{\hat{\mu}_n,\varepsilon}) - \phi_{\hat{\mu}_n}(z_{\hat{\mu}_n}))} \leq \sqrt{\frac{2\alpha}{\beta}\varepsilon}$ as follows from (9).

3. Finally, we transfer the rates for dual solutions to the primal solutions by using the (approximate) primal-dual recovery formula $x_{\hat{\mu}_n,\varepsilon} = \nabla L_{\hat{\mu}_n}(A^\top z_{\hat{\mu}_n,\varepsilon})$. This is accomplished by leveraging the stability of the gradient with respect to perturbations in the measures mentioned previously and by establishing that $\nabla L_\mu$ is Lipschitz continuous. The error due to the approximate nature of solutions is handled exactly as in item 2.

The remainder of this section is dedicated to proving Theorem 4 starting with two technical lemmas. The first will enable us to characterize the stability of $\nabla L_\rho$ at a fixed point $z$ under perturbations of the underlying measure as outlined in item 1 above.

**Lemma 6** (Gradient stability). *Fix $B \in \mathbb{R}^{d\times m}$, $z \in \mathbb{R}^d$, and $\rho, \eta \in \mathcal{P}(\mathcal{X})$. Then:*

*(i)*
$$\left\| \frac{\int B(\cdot)\exp\langle A^\top z, \cdot\rangle d\rho}{\int \exp\langle A^\top z, \cdot\rangle d\rho} - \frac{\int B(\cdot)\exp\langle A^\top z, \cdot\rangle d\eta}{\int \exp\langle A^\top z, \cdot\rangle d\eta} \right\|$$
$$\leq \kappa(B,\rho,\eta,z)\left| \int \exp\langle A^\top z, \cdot\rangle d(\eta-\rho)\right| + \zeta(B,\eta,z)\left\| \int (\cdot)\exp\langle A^\top z, \cdot\rangle d(\rho-\eta)\right\|$$
*where, recalling the measure $\gamma_{\rho,z}$ as defined in* (14),

$$\zeta(B,\eta,z) := \|B\|\left(\int \exp\langle A^\top z, \cdot\rangle d\eta\right)^{-1}, \quad \kappa(B,\rho,\eta,z) := \|\mathbb{E}_{\gamma_{\rho,z}}[X]\|\zeta(B,\eta,z).$$

*(ii) For $\gamma_{\rho,z}$ as defined by* (14), $\|\mathbb{E}_{\gamma_{\rho,z}}[X]\| \leq \mathbb{E}_{\gamma_{\rho,z}}[\|X\|] \leq |\operatorname{supp}(\rho)|_\infty$.

*(iii) $\kappa(B,\rho,\eta,z)$ and $\zeta(B,\eta,z)$ satisfy the bounds*

$$\kappa(B,\rho,\eta,z) \leq \|B\||\operatorname{supp}(\rho)|_\infty \exp\left(\|A\|\|z\||\operatorname{supp}(\eta)|_\infty\right)$$
$$\zeta(B,\eta,z) \leq \|B\|\exp\left(\|A\|\|z\||\operatorname{supp}(\eta)|_\infty\right).$$

**Proof** For the first assertion, we directly compute

$$\frac{\int B(\cdot)\exp\langle A^\top z, \cdot\rangle d\rho}{\int \exp\langle A^\top z, \cdot\rangle d\rho} - \frac{\int B(\cdot)\exp\langle A^\top z, \cdot\rangle d\eta}{\int \exp\langle A^\top z, \cdot\rangle d\eta}$$
$$= \frac{\int B(\cdot)\exp\langle A^\top z, \cdot\rangle d\rho \int \exp\langle A^\top z, \cdot\rangle d\eta - \int B(\cdot)\exp\langle A^\top z, \cdot\rangle d\eta \int \exp\langle A^\top z, \cdot\rangle d\rho}{\int \exp\langle A^\top z, \cdot\rangle d\rho \int \exp\langle A^\top z, \cdot\rangle d\eta}$$
$$= \frac{\int B(\cdot)\exp\langle A^\top z, \cdot\rangle d\rho \int \exp\langle A^\top z, \cdot\rangle d(\eta-\rho) + \int \exp\langle A^\top z, \cdot\rangle d\rho \int B(\cdot)\exp\langle A^\top z, \cdot\rangle d(\rho-\eta)}{\int \exp\langle A^\top z, \cdot\rangle d\rho \int \exp\langle A^\top z, \cdot\rangle d\eta}$$

where in the last line we have added and subtracted $\int B(\cdot)\exp\langle A^\top z, \cdot\rangle d\rho \int \exp\langle A^\top z, \cdot\rangle d\rho$ from the numerator. Taking the norm on both sides, using the linearity of the integral, and

using the fact that $\|Bw\| \leq \|B\|\|w\|$ for any $w \in \mathbb{R}^m$,

$$\begin{aligned}
&\left\| \frac{\int B(\cdot)\exp\langle A^\top z, \cdot\rangle d\rho}{\int \exp\langle A^\top z, \cdot\rangle d\rho} - \frac{\int B(\cdot)\exp\langle A^\top z, \cdot\rangle d\eta}{\int \exp\langle A^\top z, \cdot\rangle d\eta} \right\| \\
&= \left\| \frac{\int B(\cdot)\exp\langle A^\top z, \cdot\rangle d\rho \int \exp\langle A^\top z, \cdot\rangle d(\eta-\rho) + \int \exp\langle A^\top z, \cdot\rangle d\rho \int B(\cdot)\exp\langle A^\top z, \cdot\rangle d(\rho-\eta)}{\int \exp\langle A^\top z, \cdot\rangle d\rho \int \exp\langle A^\top z, \cdot\rangle d\eta} \right\| \\
&\leq \frac{\|B\|}{\int \exp\langle A^\top z, \cdot\rangle d\eta} \left\| \frac{\int (\cdot)\exp\langle A^\top z, \cdot\rangle d\rho}{\int \exp\langle A^\top z, \cdot\rangle d\rho} \right\| \left| \int \exp\langle A^\top z, \cdot\rangle d(\eta-\rho) \right| \\
&\qquad + \frac{\|B\|}{\int \exp\langle A^\top z, \cdot\rangle d\eta} \left\| \int (\cdot)\exp\langle A^\top z, \cdot\rangle d(\rho-\eta) \right\|,
\end{aligned}$$

which corresponds to item (i).

On the other hand, item (ii) follows from a direct application of Jensen's inequality, whereby $\|\mathbb{E}_{\gamma_{\rho,z}}[X]\| \leq \mathbb{E}_{\gamma_{\rho,z}}[\|X\|] \leq |\operatorname{supp}(\rho)|_\infty$, where the latter inequality is a consequence of the definition of $\gamma_{\rho,z}$ in (14) which asserts that $\operatorname{supp}(\gamma_{\rho,z}) \subset \operatorname{supp}(\rho)$.

For the estimates in item (iii), we first observe that for any $u \in \operatorname{supp}(\eta)$, we have that $\exp\left(-\|A\|\|z\||\operatorname{supp}(\eta)|_\infty\right) \leq \exp\langle A^\top z, u\rangle \leq \exp\left(\|A\|\|z\||\operatorname{supp}(\eta)|_\infty\right)$. Integrating this inequality gives

$$\exp\left(-\|A\|\|z\||\operatorname{supp}(\eta)|_\infty\right) \leq \int \exp\langle A^\top z, \cdot\rangle d\eta \leq \exp\left(\|A\|\|z\||\operatorname{supp}(\eta)|_\infty\right).$$

We conclude that $\zeta(B,\eta,z) = \|B\|\left(\int \exp\langle A^\top z, \cdot\rangle d\eta\right)^{-1} \leq \|B\|\exp\left(\|A\|\|z\||\operatorname{supp}(\eta)|_\infty\right)$. Together with item (ii), this also shows that

$$\kappa(B,\rho,\eta,z) = \|\mathbb{E}_{\gamma_{\rho,z}}[X]\|\zeta(B,\eta,z) \leq \|B\||\operatorname{supp}(\rho)|_\infty \exp\left(\|A\|\|z\||\operatorname{supp}(\eta)|_\infty\right).$$

■

The second technical lemma establishes the existence of subgradients at certain points, which is used when carrying out item 1 in the proof outline.

**Lemma 7** (Nonempty subdifferential)**.** *Fix $\rho, \eta \in \mathcal{P}(\mathcal{X})$ and let $z_\rho, z_\eta$ be solutions to the dual MEM problem with prior $\rho$ and $\eta$. Then, $\partial\phi_\rho(z_\eta)$ is nonempty and, in particular,*

$$A\nabla L_\rho(A^\top z_\eta) - A\nabla L_\eta(A^\top z_\eta) \in \partial\phi_\rho(z_\eta). \tag{19}$$

**Proof** First, observe that $\operatorname{rint}(\operatorname{dom}(L_\eta \circ A^\top - \langle b, \cdot\rangle)) = \mathbb{R}^d$, and as both $g^*$ and $L_\rho \circ A^\top - \langle b, \cdot\rangle$ are proper convex functions, the sum rule (Rockafellar, 1970, Theorem 23.8)

$$\partial\phi_\eta(z) = \partial(\alpha g^*(-z/\alpha)) + \partial\left(-\langle b, z\rangle + L_\eta(A^\top z)\right) = -\partial g^*(-z/\alpha) - b + A\nabla L_\eta(A^\top z),$$

holds for each $z \in \mathbb{R}^d$, where the final equality uses the differentiability of $L_\eta$ from Lemma 3. At the minimizer $z_\eta$, the first-order optimality condition reads

$$0 \in \partial\phi_\eta(z_\eta) = -\partial g^*(-z_\eta/\alpha) - b + A\nabla L_\eta(A^\top z_\eta),$$

and hence

$$b - A\nabla L_\eta(A^\top z_\eta) \in -\partial g^*(-z_\eta/\alpha).$$

For any other measure $\rho$, the sum rule then gives $\partial\phi_\rho(z_\eta) = -\partial g^*(-z_\eta/\alpha) - b + A\nabla L_\rho(A^\top z_\eta)$, which is nonempty as it contains, e.g., $A\nabla L_\rho(A^\top z_\eta) - A\nabla L_\eta(A^\top z_\eta)$ as follows from the previous display. ■

With the technical results of Lemma 6 and Lemma 7 in hand, we now move to analyzing the stability properties of dual solutions of the MEM problem under perturbations of the prior measure which will naturally lead to the claimed statistical rates.

### 3.1 Stability and rates for dual solutions

As noted in the proof outline, our approach relies on the $\beta/\alpha$-strong convexity of the dual objective, which follows from (A1). In effect, it enables us to bound the difference between solutions for different priors in terms of the norm of a subgradient of one of the objectives and to obtain *a priori* estimates on the norm of dual solutions as stated next.

**Theorem 8** (Estimates for dual solutions)**.** *Fix $\rho, \eta \in \mathcal{P}(\mathcal{X})$, and let $z_\rho, z_\eta$ be solutions to the dual MEM problem with prior $\rho$ and $\eta$, respectively. Then,*

*(i) Letting $G = \partial\phi_\eta(z_\rho) \cup \partial\phi_\rho(z_\eta)$, we have that*

$$\|z_\rho - z_\eta\| \leq \frac{\alpha}{\beta} \min_{v \in G} \|v\|.$$

*(ii) For any $z \in \mathbb{R}^d$, it holds that*

$$\|z_\eta - z\| \leq \frac{\alpha}{\beta} \left[ \min_{v_g \in \partial g^*(-z/\alpha)} \|v_g\| + \|b\| + \|A\| |\operatorname{supp}(\eta)|_\infty \right],$$

*whereby $\|z_\eta\| \leq \frac{\alpha}{\beta} \left[\min_{v_g \in \partial g^*(-z/\alpha)} \|v_g\| + \|b\| + \|A\| |\operatorname{supp}(\eta)|_\infty\right] + \|z\|$.*

**Proof** (i): First by Lemma 7, there exists $v \in \partial\phi_\eta(z_\rho)$, and $w \in \partial\phi_\eta(z_\eta)$. Then, as $\phi_\eta$ is $\beta/\alpha$-strongly convex, by Equation (8)

$$\|z_\rho - z_\eta\|^2 \leq \frac{\alpha}{\beta} \langle v - w, z_\rho - z_\eta \rangle \leq \frac{\alpha}{\beta} \|v - w\| \|z_\rho - z_\eta\|,$$

so that $\|z_\rho - z_\eta\| \leq \frac{\alpha}{\beta}\|v - w\|$. In particular, as $z_\eta$ is a minimizer of $\phi_\eta$, we may choose $w = 0 \in \partial\phi_\eta(z_\eta)$. Hence

$$\|z_\rho - z_\eta\| \leq \frac{\alpha}{\beta}\|v\| \quad \forall v \in \partial\phi_\eta(z_\rho).$$

As $\phi_\rho$ is also $\beta/\alpha$-strongly convex, we may repeat this argument with reversed roles of $\eta$ and $\rho$ to find

$$\|z_\rho - z_\eta\| \leq \frac{\alpha}{\beta}\|v\| \quad \forall v \in \partial\phi_\rho(z_\eta)$$

Combining the last two inequalities, we find

$$\|z_\rho - z_\eta\| \leq \frac{\alpha}{\beta} \inf_{v \in G} \|v\|,$$

for $G$ as defined above. Finally, as $G$ is closed and $\|\cdot\|$ has compact level sets, the infimum is attained.

(ii): By the subdifferential sum rule (Rockafellar, 1970, Theorem 23.8),

$$\partial \phi_\eta(z) = -\partial g^*(-z/\alpha) - b + A \nabla L_\eta(A^\top z), \forall z \in \mathbb{R}^d.$$

We split the proof into two cases: first, if $\partial \phi_\eta(z) \neq \emptyset$, for each $v \in \partial \phi_\eta(z)$, and for the particular point $0 \in \partial \phi_\eta(z_\eta)$, we have (by the same arguments as above) that

$$\|z_\eta - z\| \leq \frac{\alpha}{\beta} \|v\|.$$

Since $v \in \partial \phi_\eta(z)$, it is of the form $v = -v_g - b + A \nabla L_\eta(A^\top z)$ for some $v_g \in \partial g^*(-z/\alpha)$, so

$$\begin{aligned}
\|z_\eta - z\| &\leq \frac{\alpha}{\beta} \min_{v_g \in \partial g^*(-z/\alpha)} \left\| -v_g - b + A \nabla L_\eta(A^\top z) \right\| \\
&= \frac{\alpha}{\beta} \min_{v_g \in \partial g^*(-z/\alpha)} \left\| -v_g - b + A\, \mathbb{E}_{\gamma_{\eta,z}}[X] \right\| \\
&\leq \frac{\alpha}{\beta} \left[ \min_{v_g \in \partial g^*(-z/\alpha)} \|v_g\| + \|b\| + \|A\| \cdot |\operatorname{supp}(\eta)|_\infty \right],
\end{aligned}$$

where the penultimate line uses the property $\nabla L_\eta(A^\top z) = \mathbb{E}_{\gamma_{\eta,z}}[X]$ from (15) and where the last line follows from Lemma 6 (ii) and the triangle inequality. Once again, we emphasize that the minimum over subgradients is attained.

In the case that $\partial \phi_\eta(z) = \emptyset$, then also $-\partial g^*(-z/\alpha) = \partial \phi_\eta(z) + b - A \nabla L_\eta(A^\top z) = \emptyset$. Therefore the upper bound (ii) is a minimization over the empty set and takes the value $+\infty$ so that the result holds vacuously. ■

**Remark 9** (On solution estimates)**.** *As stated, the bound on the norm of solutions in Theorem 8 (ii) may be vacuous depending on the choice of evaluation point $z \in \mathbb{R}^d$. Indeed, as stated in the proof of that result, $\partial g^*(-z/\alpha)$ may be empty.*

*Nevertheless, we point out that* $\operatorname{dom} \partial g^* = \{u : \partial g^*(u) \neq \emptyset\}$ *is non-empty as $g^*$ is proper, see, e.g., Rockafellar (1970, Corollary 23.4) (and use the fact that $g^*$ is proper).*

We now show that the upper bound on $\|z_\rho - z_\eta\|$ furnished in Theorem 8 can be rewritten in terms of an expression involving the differences between the measures $\eta$ and $\rho$. In the sequel, it will be expedient to establish the following shorthand notation. Given a matrix $B$, and measures $\rho, \eta \in \mathcal{P}(\mathcal{X})$ we define

$$\begin{aligned}
\kappa'(B, \rho, \eta) &:= \max\left\{\kappa(B, \rho, \eta, z_\eta), \kappa(B, \eta, \rho, z_\eta)\right\}, \\
\zeta'(B, \rho, \eta) &:= \max\left\{\zeta(B, \rho, z_\eta), \zeta(B, \eta, z_\eta)\right\},
\end{aligned}$$

where $\zeta(B, \eta, z) = \|B\| \left(\int \exp\langle A^\top z, \cdot \rangle d\eta\right)^{-1}$ and $\kappa(B, \rho, \eta, z) = \|\mathbb{E}_{\gamma_{\rho,z}}[X]\| \zeta(B, \eta, z)$, as defined in Lemma 6 (i).

**Corollary 10** (Stability of dual solutions)**.** *Fix $\rho, \eta \in \mathcal{P}(\mathcal{X})$ and let $z_\rho, z_\eta$ solve the dual MEM problem with prior $\rho$ and $\eta$, respectively. Then*

$$\|z_\eta - z_\rho\| \leq \frac{\alpha}{\beta}\kappa'(A,\rho,\eta)\left|\int \exp\langle A^\top z_\eta, \cdot\rangle d(\eta-\rho)\right| + \frac{\alpha}{\beta}\zeta'(A,\rho,\eta)\left\|\int (\cdot)\exp\langle A^\top z_\eta, \cdot\rangle d(\rho-\eta)\right\|.$$

**Proof** We first apply Theorem 8, which asserts that $\|z_\eta - z_\rho\| \leq \frac{\alpha}{\beta}\min_{v\in\partial\phi_\rho(z_\eta)\cup\partial\phi_\eta(z_\rho)}\|v\|$. Taking $v = A\nabla L_\rho(A^\top z_\eta) - A\nabla L_\eta(A^\top z_\eta) \in \partial\phi_\rho(z_\eta)$ from Lemma 7 (19),

$$\|z_\eta - z_\rho\| \leq \frac{\alpha}{\beta}\|A\nabla L_\rho(A^\top z_\eta) - A\nabla L_\eta(A^\top z_\eta)\|, \tag{20}$$

and it remains to bound the right-hand side. Recalling the expressions for $\nabla L_\rho$ and $\nabla L_\eta$ from (12), we then apply Lemma 6 (i) with $B = A$ and $z = z_\eta$ to obtain that

$$\begin{aligned}
\|v\| &= \|A\nabla L_\rho(A^\top z_\eta) - A\nabla L_\eta(A^\top z_\eta)\| \\
&= \left\|\frac{\int A(\cdot)\exp\langle A^\top z_\eta, \cdot\rangle d\rho}{\int \exp\langle A^\top z_\eta, \cdot\rangle d\rho} - \frac{\int A(\cdot)\exp\langle A^\top z_\eta, \cdot\rangle d\eta}{\int \exp\langle A^\top z_\eta, \cdot\rangle d\eta}\right\| \\
&\leq \kappa(A,\rho,\eta,z_\eta)\left|\int \exp\langle A^\top z_\eta, \cdot\rangle d(\eta-\rho)\right| + \zeta(A,\eta,z_\eta)\left\|\int (\cdot)\exp\langle A^\top z_\eta, \cdot\rangle d(\rho-\eta)\right\|.
\end{aligned}$$

Repeating this computation with $-v$ in place of $v$, we apply Lemma 6 (i) once more $B = A$ and $z = z_\eta$ with the roles of $L_\rho$ and $L_\eta$ interchanged. This gives

$$\begin{aligned}
\|v\| &= \|-v\| \\
&= \|A\nabla L_\eta(A^\top z_\eta) - A\nabla L_\rho(A^\top z_\eta)\| \\
&\leq \left\|\frac{\int A(\cdot)\exp\langle A^\top z_\eta, \cdot\rangle d\eta}{\int \exp\langle A^\top z_\eta, \cdot\rangle d\eta} - \frac{\int A(\cdot)\exp\langle A^\top z_\eta, \cdot\rangle d\rho}{\int \exp\langle A^\top z_\eta, \cdot\rangle d\rho}\right\| \\
&\leq \kappa(A,\eta,\rho,z_\eta)\left|\int \exp\langle A^\top z_\eta, \cdot\rangle d(\rho-\eta)\right| + \zeta(A,\rho,z_\eta)\left\|\int (\cdot)\exp\langle A^\top z_\eta, \cdot\rangle d(\eta-\rho)\right\|.
\end{aligned}$$

Finally, as $|\int \exp\langle A^\top z_\eta, \cdot\rangle d(\eta-\rho)| = |\int \exp\langle A^\top z_\eta, \cdot\rangle d(\rho-\eta)|$, and $\|\int(\cdot)\exp\langle A^\top z_\eta, \cdot\rangle d(\eta-\rho)\| = \|\int(\cdot)\exp\langle A^\top z_\eta, \cdot\rangle d(\rho-\eta)\|$, we take the maximum of these bounds and substitute into (20) which yields

$$\|z_\eta - z_\rho\| \leq \frac{\alpha}{\beta}\kappa'(A,\rho,\eta)\left|\int \exp\langle A^\top z_\eta, \cdot\rangle d(\eta-\rho)\right| + \frac{\alpha}{\beta}\zeta'(A,\rho,\eta)\left\|\int (\cdot)\exp\langle A^\top z_\eta, \cdot\rangle d(\rho-\eta)\right\|$$

with the constants $\kappa'(A,\rho,\eta), \zeta'(A,\rho,\eta)$ defined previously. ■

In order to establish the desired parametric rate of convergence for the empirical dual MEM solutions, it remains to provide uniform bounds on the constants from Corollary 10.

**Lemma 11** (Constant bounds)**.** *Let $\mu \in \mathcal{P}(\mathcal{X})$ and set $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ to be the empirical measure from $n$ i.i.d. samples $X_1, \ldots, X_n$ from $\mu$. Then, with probability 1,*

$$\begin{aligned}
\kappa'(B,\hat{\mu}_n,\mu) &\leq \|B\|\,|\operatorname{supp}(\mu)|_\infty \exp(\|A\|\|z_\mu\|\,|\operatorname{supp}(\mu)|_\infty), \\
\zeta'(B,\hat{\mu}_n,\mu) &\leq \|B\|\exp(\|A\|\|z_\mu\|\,|\operatorname{supp}(\mu)|_\infty).
\end{aligned}$$

**Proof** (i) First, from Lemma 6 (iii),

$$\begin{aligned}\kappa(B,\hat{\mu}_n,\mu,z_\mu) &\leq \|B\|\,|\operatorname{supp}(\hat{\mu}_n)|_\infty \exp(\|A\|\|z_\mu\|\,|\operatorname{supp}(\mu)|_\infty)\\ &\leq \|B\|\,|\operatorname{supp}(\mu)|_\infty \exp(\|A\|\|z_\mu\|\,|\operatorname{supp}(\mu)|_\infty),\end{aligned}$$

where the second inequality holds since $\operatorname{supp}(\hat{\mu}_n) \subset \operatorname{supp}(\mu)$ with probability 1. Similarly, we have:

$$\begin{aligned}\kappa(B,\mu,\hat{\mu}_n,z_\mu) &\leq \|B\|\,|\operatorname{supp}(\mu)|_\infty \exp(\|A\|\|z_\mu\|\,|\operatorname{supp}(\hat{\mu}_n)|_\infty)\\ &\leq \|B\|\,|\operatorname{supp}(\mu)|_\infty \exp(\|A\|\|z_\mu\|\,|\operatorname{supp}(\mu)|_\infty).\end{aligned}$$

It readily follows that

$$\begin{aligned}\kappa'(B,\hat{\mu}_n,\mu) &= \max\{\kappa(B,\mu,\hat{\mu}_n,z_\mu),\kappa(B,\hat{\mu}_n,\mu,z_\mu)\}\\ &\leq \|B\|\,|\operatorname{supp}(\mu)|_\infty \exp(\|A\|\|z_\mu\|\,|\operatorname{supp}(\mu)|_\infty).\end{aligned}$$

(ii) As for item (i), applying Lemma 6 (iii) we have

$$\zeta(B,\hat{\mu}_n,z_\mu) \leq \|B\| \exp(\|A\|\|z_\mu\|\,|\operatorname{supp}(\hat{\mu}_n)|_\infty) \leq \|B\| \exp(\|A\|\|z_\mu\|\,|\operatorname{supp}(\mu)|_\infty),$$

where we have again used that $\operatorname{supp}(\hat{\mu}_n) \subset \operatorname{supp}(\mu)$. On the other hand, we have directly from Lemma 6 (iii) that

$$\zeta(B,\mu,z_\mu) \leq \|B\| \exp(\|A\|\|z_\mu\|\,|\operatorname{supp}(\mu)|_\infty).$$

Hence, $\zeta'(B,\hat{\mu}_n,\mu) = \max\{\zeta(B,\mu,z_\mu),\zeta(B,\hat{\mu}_n,z_\mu)\} \leq \|B\| \exp(\|A\|\|z_\mu\|\,|\operatorname{supp}(\mu)|_\infty)$. ■

With these preliminary results in hand, we now establish the parametric rate of convergence of empirical dual solutions as stated in Theorem 4.

**Theorem 12** (Parametric rates for dual solutions)**.** *Fix $\mu \in \mathcal{P}(\mathcal{X})$ and let $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ be the empirical measure from $n$ i.i.d. samples $X_1,\ldots,X_n$ from $\mu$. Denoting $z_\mu, z_{\hat{\mu}_n}$ as the solutions to the dual MEM problem with priors $\mu$ and $\hat{\mu}_n$, respectively, we have*

$$\mathbb{E}[\|z_{\hat{\mu}_n} - z_\mu\|] \leq \frac{C}{\sqrt{n}},$$

*where $C = \frac{2\alpha}{\beta}\|A\|\,|\operatorname{supp}(\mu)|_\infty \exp\left(2\|A\|\|z_\mu\|\,|\operatorname{supp}(\mu)|_\infty\right)$.*

**Proof** We first apply Corollary 10 with $\rho = \hat{\mu}_n$ and $\eta = \mu$ to obtain that

$$\begin{aligned}\|z_{\hat{\mu}_n} - z_\mu\| \leq{}& \frac{\alpha}{\beta}\kappa'(A,\hat{\mu}_n,\mu)\left|\int \exp\langle A^\top z_\mu,\cdot\rangle d(\mu-\hat{\mu}_n)\right|\\ &+ \frac{\alpha}{\beta}\zeta'(A,\hat{\mu}_n,\mu)\left\|\int (\cdot)\exp\langle A^\top z_\mu,\cdot\rangle d(\hat{\mu}_n-\mu)\right\|.\end{aligned} \tag{21}$$

Given that $\kappa'(A,\hat{\mu}_n,\mu), \zeta'(A,\hat{\mu}_n,\mu)$ admit upper bounds which do not depend on the choice of samples by Lemma 11 we proceed by bounding the expected value of the two integral terms.

First, let $f = \exp\langle A^\top z_\mu, \cdot\rangle$ and observe that

$$\begin{aligned}\mathbb{E}\left[\left|\int f d\mu - \frac{1}{n}\sum_{i=1}^n f(X_i)\right|\right] &\leq \sqrt{\mathbb{E}\left[\left|\int f d\mu - \frac{1}{n}\sum_{i=1}^n f(X_i)\right|^2\right]} \\ &= \sqrt{\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \int f d\mu - \frac{1}{n}\sum_{i=1}^n f(X_i)\right|^2\right]} \\ &= \sqrt{\frac{1}{n^2}\operatorname{var}\left[\sum_{i=1}^n f(X_i)\right]},\end{aligned}$$

where the first line follows from Jensen's inequality, and we have used the fact that $\int f d\mu = \frac{1}{n}\sum_{i=1}^n \int f d\mu$ in the next equality. The final equality follows from setting $Y = \sum_{i=1}^n f(X_i)$ and applying the definition of the variance of $Y$, (10). As the samples $X_1, \ldots, X_n$ are independent,

$$\operatorname{var}\left[\sum_{i=1}^n f(X_i)\right] = n\operatorname{var}[f(X_1)] \leq n\mathbb{E}_\mu[f^2(X_1)] \leq n\exp\left(2\|A\|\|z_\mu\|\,|\operatorname{supp}(\mu)|_\infty\right),$$

where we have used that $f(X_1)^2 = (\exp\langle A^\top z_\mu, X_1\rangle)^2 \leq \exp\left(2\|A\|\|z_\mu\|\,|\operatorname{supp}(\mu)|_\infty\right)$ in the final inequality. Hence, we have shown

$$\mathbb{E}\left[\left|\int \exp\langle A^\top z_\mu, \cdot\rangle d(\mu - \hat{\mu}_n)\right|\right] \leq \frac{1}{\sqrt{n}}\exp\left(\|A\|\|z_\mu\|\,|\operatorname{supp}(\mu)|_\infty\right) \tag{22}$$

For the other term of (21), define $h = (\cdot)\exp\langle A^\top z_\mu, \cdot\rangle$. Then once more via Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\int h d\mu - \frac{1}{n}\sum_{i=1}^n h(X_i)\right\|\right] &\leq \sqrt{\mathbb{E}\left[\left\|\int h d\mu - \frac{1}{n}\sum_{i=1}^n h(X_i)\right\|^2\right]} \\ &= \sqrt{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(\int h d\mu - h(X_i)\right)\right\|^2\right]} \\ &= \sqrt{\frac{1}{n^2}\operatorname{Tr}\left(\operatorname{cov}\left[\sum_{i=1}^n h(X_i)\right]\right)}.\end{aligned}$$

Here, the final equality can be observed by computing the covariance matrix of the random vector $Y = \sum_{i=1}^n h(X_i)$. Moreover, $\operatorname{cov}\left[\sum_{i=1}^n h(X_i)\right] = n\operatorname{cov}\left[h(X_1)\right]$ due to independence of the samples. Now, expanding the diagonal entries of the covariance matrix,

$$\operatorname{Tr}\left(\operatorname{cov}\left[h(X_1)\right]\right) = \mathbb{E}\left[\|h(X_1) - \mathbb{E}[h(X_1)]\|^2\right] \leq |\operatorname{supp}(\mu)|_\infty^2 \exp(2\|A\|\|z_\mu\|\,|\operatorname{supp}(\mu)|_\infty),$$

where the inequality follows from observing that

$$\mathbb{E}\left[\|h(X_1) - \mathbb{E}[h(X_1)]\|^2\right] = \mathbb{E}[\|h(X_1)\|^2] - \|\mathbb{E}[h(X_1)]\|^2 \leq \mathbb{E}\left[\|h(X_1)\|^2\right].$$

Therefore, we have

$$\mathbb{E}\left[\left\|\int (\cdot)\exp\langle A^\top z_\mu, \cdot\rangle d(\hat{\mu}_n - \mu)\right\|\right] \leq \frac{1}{\sqrt{n}}|\operatorname{supp}(\mu)|_\infty \exp\left(\|A\|\|z_\mu\||\operatorname{supp}(\mu)|_\infty\right). \tag{23}$$

Substituting (22) and (23) into (21) along with the bounds on $\kappa'(A, \hat{\mu}_n, \mu)$ and $\zeta'(A, \hat{\mu}_n, \mu)$ from Lemma 11 yields that

$$\mathbb{E}\left[\|z_{\hat{\mu}_n} - z_\mu\|\right] \leq \frac{1}{\sqrt{n}}\frac{2\alpha}{\beta}\|A\||\operatorname{supp}(\mu)|_\infty \exp\left(2\|A\|\|z_\mu\||\operatorname{supp}(\mu)|_\infty\right).$$

■

The above result naturally generalizes to approximate dual solutions as stated next.

**Corollary 13** (Parametric rates for approximate dual solutions)**.** *In the setting of Theorem 12, let $z_{\hat{\mu}_n,\varepsilon}$ be a measurable selection of an $\varepsilon$-approximate minimizer of $\phi_{\hat{\mu}_n}$. Then,*

$$\mathbb{E}[\|z_{\hat{\mu}_n,\varepsilon} - z_\mu\|] \leq \frac{C}{\sqrt{n}} + \sqrt{\frac{2\alpha}{\beta}\varepsilon}.$$

**Proof** By the triangle inequality,

$$\mathbb{E}[\|z_{\hat{\mu}_n,\varepsilon} - z_\mu\|] \leq \mathbb{E}[\|z_{\hat{\mu}_n} - z_\mu\|] + \mathbb{E}[\|z_{\hat{\mu}_n,\varepsilon} - z_{\hat{\mu}_n}\|].$$

The first term on the right-hand side is bounded by $Cn^{-1/2}$ by Theorem 12. Since $0 \in \partial\phi_{\hat{\mu}_n}(z_{\hat{\mu}_n})$, (9) implies that

$$\|z_{\hat{\mu}_n,\varepsilon} - z_{\hat{\mu}_n}\|^2 \leq \frac{2\alpha}{\beta}\left(\phi_{\hat{\mu}_n}(z_{\hat{\mu}_n,\varepsilon}) - \phi_{\hat{\mu}_n}(z_{\hat{\mu}_n})\right) \leq \frac{2\alpha}{\beta}\varepsilon,$$

proving the claimed result. ■

With this result in hand, we proceed to proving the corresponding convergence rate for the primal solutions.

### 3.2 Stability and rates for primal solutions

To derive stability and convergence rate results for primal solutions, we now first establish that the gradient of $L_\rho$ is Lipschitz continuous for any choice of $\rho \in \mathcal{P}(\mathcal{X})$.

**Lemma 14** (Lipschitz continuity)**.** *For any $\rho \in \mathcal{P}(\mathcal{X})$, $L_\rho \circ A^\top$ is $K$-smooth, where $K = |\operatorname{supp}(\rho)|_\infty^2\|A\|^2$.*

**Proof** We compute the first and second partial derivatives of $L_\rho \circ A^\top$ below:

$$\partial_i(L_\rho \circ A^\top) : y \in \mathbb{R}^d \mapsto \frac{\int \sum_{k=1}^m A_{ik}u_k \exp\langle y, Au\rangle d\rho(u)}{\int \exp\langle y, Au\rangle d\rho(u)},$$

$$\begin{aligned}\partial_j\partial_i(L_\rho \circ A^\top) : y \in \mathbb{R}^d \mapsto &\frac{\int \sum_{k=1}^m A_{ik}u_k \sum_{l=1}^m A_{jl}u_l \exp\langle y, Au\rangle d\rho(u)}{\int \exp\langle y, Au\rangle d\rho(u)},\\ &- \frac{\int \sum_{k=1}^m A_{ik}u_k \exp\langle y, Au\rangle d\rho(u) \int \sum_{l=1}^m A_{jl}u_l \exp\langle y, Au\rangle d\rho(u)}{\left(\int \exp\langle y, Au\rangle d\rho(u)\right)^2}.\end{aligned}$$

Consequently, the Hessian of $L_\rho \circ A^\top$ can be expressed as

$$\nabla^2(L_\rho \circ A^\top)(y) = \mathbb{E}_{\gamma_{\rho,y}}\left[AUU^\top A^\top\right] - \mathbb{E}_{\gamma_{\rho,y}}\left[AU\right]\mathbb{E}_{\gamma_{\rho,y}}\left[AU\right]^\top$$

where $\gamma_{\rho,y}$ is as defined in (14) and $U \sim \gamma_{\rho,y}$. We can rewrite this expression as

$$\nabla^2(L_\rho \circ A^\top)(y) = A\mathbb{E}_{\gamma_{\rho,y}}\left[\left(U - \mathbb{E}_{\gamma_{\rho,y}}\left[U\right]\right)\left(U - \mathbb{E}_{\gamma_{\rho,y}}\left[U\right]\right)^\top\right]A^\top,$$

so that, for any $w \in \mathbb{R}^d$ with $\|w\| \leq 1$,

$$w^\top(\nabla^2(L_\rho \circ A^\top)(y))w = \mathbb{E}_{\gamma_{\rho,y}}\left[\left\|\left(U - \mathbb{E}_{\gamma_{\rho,y}}\left[U\right]\right)^\top A^\top w\right\|^2\right] \leq \mathbb{E}_{\gamma_{\rho,y}}\left[\left\|U - \mathbb{E}_{\gamma_{\rho,y}}\left[U\right]\right\|^2\right]\|A\|^2,$$

where we have applied the Cauchy-Schwarz inequality. We may further bound the expectation as

$$\mathbb{E}_{\gamma_{\rho,y}}\left[\left\|U - \mathbb{E}_{\gamma_{\rho,y}}[U]\right\|^2\right] = \mathbb{E}_{\gamma_{\rho,y}}\left[\|U\|^2\right] - \left\|\mathbb{E}_{\gamma_{\rho,y}}[U]\right\|^2 \leq |\operatorname{supp}(\rho)|_\infty^2.$$

It follows from the previous two lines that all eigenvalues of $\nabla^2(L_\rho \circ A^\top)(y)$ are bounded above by $|\operatorname{supp}(\rho)|_\infty^2\|A\|^2 =: K$ so that $\nabla(L_\rho \circ A^\top)$ is $K$-Lipschitz continuous. ■

We briefly comment on some implications of Lemma 14.

**Remark 15.** *We make two observations as immediate consequences of Lemma 14:*

*(i) As* $\operatorname{supp}(\rho) \subset \mathcal{X}$, $\nabla(L_\rho \circ A^\top)$ *is* $|\mathcal{X}|_\infty^2\|A\|^2$*-Lipschitz for all* $\rho \in \mathcal{P}(\mathcal{X})$.

*(ii) For the choice* $A = I$*, this also shows that* $L_\rho$ *is* $|\operatorname{supp}(\rho)|_\infty^2$*-smooth.*

With this technical lemma in place, we can leverage the primal-dual recovery formula to establish a stability property for primal solutions under perturbations of the prior.

**Proposition 16** (Primal solution stability)**.** *Fix* $\rho, \eta \in \mathcal{P}(\mathcal{X})$*. Let* $x_\rho, x_\eta$ *solve the primal MEM problem with priors* $\rho, \eta$*, and* $z_\rho, z_\eta$ *solve the respective dual MEM problems. Then:*

*(i) Setting* $K' = \|A\||\operatorname{supp}(\rho)|_\infty^2$,

$$\|x_\rho - x_\eta\| \leq K'\|z_\rho - z_\eta\| + \kappa'(I,\rho,\eta)\left|\int \exp\langle A^\top z_\eta, \cdot\rangle d(\eta-\rho)\right| + \zeta'(I,\rho,\eta)\left\|\int (\cdot)\exp\langle A^\top z_\eta, \cdot\rangle d(\rho-\eta)\right\|.$$

*(ii) If* $A$ *is injective and* $g$ *is* $\frac{1}{\gamma}$*-strongly convex, then:*

$$\|x_\rho - x_\eta\| \leq \frac{\gamma}{\alpha\sigma_{\min}(A)}\|z_\rho - z_\eta\|.$$

**Proof** (i) As a consequence of the primal-dual recovery formula (13), we have

$$\begin{aligned}\|x_\rho - x_\eta\| &= \|\nabla L_\rho(A^\top z_\rho) - \nabla L_\eta(A^\top z_\eta)\| \\ &\leq \|\nabla L_\rho(A^\top z_\rho) - \nabla L_\rho(A^\top z_\eta)\| + \|\nabla L_\rho(A^\top z_\eta) - \nabla L_\eta(A^\top z_\eta)\|.\end{aligned}$$

For the former term, we use Remark 15 (ii) to obtain that

$$\begin{aligned}\|\nabla L_\rho(A^\top z_\rho) - \nabla L_\rho(A^\top z_\eta)\| &\leq |\operatorname{supp}(\rho)|_\infty^2\|A^\top(z_\rho - z_\eta)\| \\ &\leq |\operatorname{supp}(\rho)|_\infty^2\|A\|\|z_\rho - z_\eta\|.\end{aligned}$$

For the latter term, in view of (12), $\|\nabla L_\rho(A^\top z_\eta) - \nabla L_\eta(A^\top z_\eta)\|$ is exactly in the form of Lemma 6 (i) for the choices of $B = I$ and $z = z_\eta$. Hence

$$\begin{aligned}&\|\nabla L_\rho(A^\top z_\eta) - \nabla L_\eta(A^\top z_\eta)\| \\ &\leq \kappa(I,\rho,\eta,z_\eta)\left|\int \exp\langle A^\top z_\eta, \cdot\rangle d(\eta-\rho)\right| + \zeta(I,\eta,z_\eta)\left\|\int (\cdot)\exp\langle A^\top z_\eta, \cdot\rangle d(\rho-\eta)\right\|.\end{aligned}$$

Observing that $\|\nabla L_\rho(A^\top z_\eta) - \nabla L_\eta(A^\top z_\eta)\| = \|\nabla L_\eta(A^\top z_\eta) - \nabla L_\rho(A^\top z_\eta)\|$, we also apply Lemma 6 (i) for the choices of $B = I$ and $z = z_\eta$, with the order of $L_\eta$ and $L_\rho$ interchanged. This gives

$$\begin{aligned}&\|\nabla L_\eta(A^\top z_\eta) - \nabla L_\rho(A^\top z_\eta)\| \\ &\leq \kappa(I,\eta,\rho,z_\eta)\left|\int \exp\langle A^\top z_\eta, \cdot\rangle d(\eta-\rho)\right| + \zeta(I,\rho,z_\eta)\left\|\int (\cdot)\exp\langle A^\top z_\eta, \cdot\rangle d(\rho-\eta)\right\|.\end{aligned}$$

Taking the maximum over these two bounds gives the result with the corresponding constants $\kappa'(I,\rho,\eta)$ and $\zeta'(I,\rho,\eta)$ as was stated in the theorem.

(ii) Assuming that $A$ is injective and $g$ is $\frac{1}{\gamma}$-strongly convex, $g^*$ is differentiable and $\gamma$-smooth. Rearranging the first order optimality conditions $\nabla\phi_\rho(z_\rho) = \nabla\phi_\eta(z_\eta) = 0$ gives

$$A\nabla L_\eta(A^\top z_\eta) - A\nabla L_\rho(A^\top z_\rho) = \nabla g^*(-z_\eta/\alpha) - \nabla g^*(-z_\rho/\alpha).$$

As $\|Ax\| \geq \sigma_{\min}(A)\|x\|$ for any $x \in \mathbb{R}^m$, and we also have $\sigma_{\min}(A) > 0$ as $A$ is injective,

$$\sigma_{\min}(A)\|\nabla L_\eta(A^\top z_\eta) - \nabla L_\rho(A^\top z_\rho)\| \leq \|\nabla g^*(-z_\eta/\alpha) - \nabla g^*(-z_\rho/\alpha)\|. \tag{24}$$

Returning to the primal-dual recovery formula Equation (13), we have $x_\eta - x_\rho = \nabla L_\eta(A^\top z_\eta) - \nabla L_\rho(A^\top z_\rho)$. Hence, we find

$$\begin{aligned}\|x_\eta - x_\rho\| &= \|\nabla L_\eta(A^\top z_\eta) - \nabla L_\rho(A^\top z_\rho)\| \\ &\leq \frac{1}{\sigma_{\min}(A)}\|\nabla g^*(-z_\eta/\alpha) - \nabla g^*(-z_\rho/\alpha)\| \\ &\leq \frac{\gamma}{\alpha\sigma_{\min}(A)}\|z_\eta - z_\rho\|,\end{aligned}$$

where we have used (24) for the first inequality, and the $\gamma$-smoothness of $g^*$ in the latter. ■

As with the dual solutions, the stability properties provided above enable us to obtain the parametric rate for empirical primal solutions.

**Theorem 17** (Parametric rates for primal solutions)**.** *Fix $\mu \in \mathcal{P}(\mathcal{X})$ and let $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ be the empirical measure from $n$ i.i.d. samples $X_1, \ldots, X_n$ from $\mu$. Let $x_\mu, x_{\hat{\mu}_n}$ denote the solutions to the primal MEM problem with priors $\mu$ and $\hat{\mu}_n$ respectively. Then:*

*(i) Letting $C' = 2|\operatorname{supp}(\mu)|_\infty \exp\left(2\|A\|\|z_\mu\||\operatorname{supp}(\mu)|_\infty\right)$, $C = \frac{\alpha}{\beta}\|A\|C'$ be the constants from Theorem 12 and $K' = |\operatorname{supp}(\mu)|_\infty^2\|A\|$, we have*

$$\mathbb{E}[\|x_{\hat{\mu}_n} - x_\mu\|] \leq \frac{K'C + C'}{\sqrt{n}}.$$

*(ii) If $A$ is injective and $g$ is $\frac{1}{\gamma}$-strongly convex, then*

$$\mathbb{E}\left[\|x_{\hat{\mu}_n} - x_\mu\|\right] \leq \frac{\gamma C}{\alpha\sigma_{\min}(A)}\frac{1}{\sqrt{n}}.$$

**Proof** (i) Applying Proposition 16 (i) with $\eta = \mu, \rho = \hat{\mu}_n$ gives

$$\begin{aligned}\|x_{\hat{\mu}_n} - x_\mu\| \leq K\|z_{\hat{\mu}_n} - z_\mu\| &+ \kappa'(I, \hat{\mu}_n, \mu)\left|\int \exp\langle A^\top z_\mu, \cdot\rangle d(\mu - \hat{\mu}_n)\right| \\ &+ \zeta'(I, \hat{\mu}_n, \mu)\left\|\int (\cdot)\exp\langle A^\top z_\mu, \cdot\rangle d(\hat{\mu}_n - \mu)\right\|,\end{aligned}$$

where $K = \|A\||\operatorname{supp}(\hat{\mu}_n)|_\infty^2 \leq K'$, noting that $\operatorname{supp}(\hat{\mu}_n) \subset \operatorname{supp}(\mu)$. Taking the expectation of the above inequality and applying Theorem 12 to the first term yields

$$\begin{aligned}\mathbb{E}[\|x_{\hat{\mu}_n} - x_\mu\|] \leq \frac{K'C}{\sqrt{n}} &+ \mathbb{E}\left[\kappa'(I, \hat{\mu}_n, \mu)\left|\int \exp\langle A^\top z_\mu, \cdot\rangle d(\mu - \hat{\mu}_n)\right|\right] \\ &+ \mathbb{E}\left[\zeta'(I, \hat{\mu}_n, \mu)\left\|\int (\cdot)\exp\langle A^\top z_\mu, \cdot\rangle d(\hat{\mu}_n - \mu)\right\|\right].\end{aligned}$$

Hence, it remains to bound the expectation of the latter terms. From (22) and (23),

$$\begin{aligned}\mathbb{E}\left[\left|\int \exp\langle A^\top z_\mu, \cdot\rangle d(\mu - \hat{\mu}_n)\right|\right] &\leq \frac{1}{\sqrt{n}}\exp\left(\|A\|\|z_\mu\||\operatorname{supp}(\mu)|_\infty\right), \\ \mathbb{E}\left[\left\|\int (\cdot)\exp\langle A^\top z_\mu, \cdot\rangle d(\mu - \hat{\mu}_n)\right\|\right] &\leq \frac{1}{\sqrt{n}}|\operatorname{supp}(\mu)|_\infty\exp\left(\|A\|\|z_\mu\||\operatorname{supp}(\mu)|_\infty\right).\end{aligned}$$

As for $\kappa'(I, \hat{\mu}_n, \mu), \zeta'(I, \hat{\mu}_n, \mu)$, we apply Lemma 11 with $B = I$ to obtain that $\kappa'(I, \hat{\mu}_n, \mu) \leq |\operatorname{supp}(\mu)|_\infty \exp\left(\|A\|\|z_\mu\||\operatorname{supp}(\mu)|_\infty\right)$ and $\zeta'(I, \hat{\mu}_n, \mu) \leq \exp(\|A\|\|z_\mu\||\operatorname{supp}(\mu)|_\infty)$. This enables us to control the terms involving the integrals as

$$\frac{1}{\sqrt{n}}\left(|\operatorname{supp}(\mu)|_\infty \exp\left(2\|A\|\|z_\mu\||\operatorname{supp}(\mu)|_\infty\right) + |\operatorname{supp}(\mu)|_\infty \exp\left(2\|A\|\|z_\mu\||\operatorname{supp}(\mu)|_\infty\right)\right),$$

which coincides with $\frac{C'}{\sqrt{n}}$, thus completing the proof of item (i).

(ii) Supposing that $A$ is injective and $g$ is $\frac{1}{\gamma}$-strongly convex, we apply Proposition 16 (ii) with $\rho = \hat{\mu}_n, \eta = \mu$, giving

$$\|x_{\hat{\mu}_n} - x_\mu\| \leq \frac{\gamma}{\alpha\sigma_{\min}(A)}\|z_{\hat{\mu}_n} - z_\mu\|.$$

Taking the expectation and applying Theorem 12 yields

$$\mathbb{E}\left[\|x_{\hat{\mu}_n} - x_\mu\|\right] \leq \frac{\gamma}{\alpha\sigma_{\min}(A)}\,\mathbb{E}\left[\|z_{\hat{\mu}_n} - z_\mu\|\right] \leq \frac{\gamma C}{\alpha\sigma_{\min}(A)}\frac{1}{\sqrt{n}}.$$

■

As before, we extend the sample complexity to the case of approximate minimizers.

**Theorem 18** (Parametric rates for approximate solutions)**.** *Let $z_{\hat{\mu}_n,\varepsilon}$ be a measurable selection of an $\varepsilon$-approximate solution to the dual MEM problem for $\hat{\mu}_n$and set $x_{\hat{\mu}_n,\varepsilon} := \nabla L_{\hat{\mu}_n}(A^\top z_{\hat{\mu}_n,\varepsilon})$. Then,*

$$\mathbb{E}\left[\|x_{\hat{\mu}_n,\varepsilon} - x_\mu\|\right] \leq \frac{K'C + C'}{\sqrt{n}} + K'\sqrt{\frac{2\alpha}{\beta}\varepsilon},$$

*where $K', C, C'$ are the constants of Theorem 17. If, additionally, $A$ is injective and $g$ is $\frac{1}{\gamma}$-strongly convex, then*

$$\mathbb{E}\left[\|x_{\hat{\mu}_n,\varepsilon} - x_\mu\|\right] \leq \frac{\gamma C}{\alpha\sigma_{\min}(A)}\frac{1}{\sqrt{n}} + K'\sqrt{\frac{2\alpha}{\beta}\varepsilon}.$$

**Proof** The triangle inequality yields

$$\begin{aligned}\|x_{\hat{\mu}_n,\varepsilon} - x_\mu\| &\leq \|x_{\hat{\mu}_n,\varepsilon} - x_{\hat{\mu}_n}\| + \|x_{\hat{\mu}_n} - x_\mu\| \\ &= \|\nabla L_{\hat{\mu}_n}(A^\top z_{\hat{\mu}_n,\varepsilon}) - \nabla L_{\hat{\mu}_n}(A^\top z_{\hat{\mu}_n})\| + \|x_{\hat{\mu}_n} - x_\mu\|. \end{aligned} \tag{25}$$

The expectation of the latter term has already been bounded in these different cases in Theorem 17, so it remains to bound the former term. However, Lemma 14 yields that $\nabla L_{\hat{\mu}_n}$ is $|\operatorname{supp}(\hat{\mu}_n)|_\infty^2$-Lipschitz continuous and hence

$$\begin{aligned}\|\nabla L_{\hat{\mu}_n}(A^\top z_{\hat{\mu}_n,\varepsilon}) - \nabla L_{\hat{\mu}_n}(A^\top z_{\hat{\mu}_n})\| &\leq |\operatorname{supp}(\hat{\mu}_n)|_\infty^2\|A^\top(z_{\hat{\mu}_n,\varepsilon} - z_{\hat{\mu}_n})\| \\ &\leq |\operatorname{supp}(\mu)|_\infty^2\|A\|\|z_{\hat{\mu}_n,\varepsilon} - z_{\hat{\mu}_n}\|.\end{aligned}$$

In turn, using the $\beta/\alpha$-strong convexity of $\phi_{\hat{\mu}_n}$, we use (9) and the fact that $z_{\hat{\mu}_n}$ is a minimizer of $\phi_{\hat{\mu}_n}$ (hence $0 \in \partial\phi_{\hat{\mu}_n}(z_{\hat{\mu}_n})$) to obtain

$$\frac{\beta}{2\alpha}\|z_{\hat{\mu}_n,\varepsilon} - z_{\hat{\mu}_n}\|^2 \leq \phi_{\hat{\mu}_n}(z_{\hat{\mu}_n,\varepsilon}) - \phi_{\hat{\mu}_n}(z_{\hat{\mu}_n}) \leq \varepsilon,$$

where the second inequality follows by the definition of $z_{\hat{\mu}_n,\varepsilon}$. Hence,

$$\|\nabla L_{\hat{\mu}_n}(A^\top z_{\hat{\mu}_n,\varepsilon}) - \nabla L_{\hat{\mu}_n}(A^\top z_{\hat{\mu}_n})\| \leq |\operatorname{supp}(\mu)|_\infty^2 \|A\| \sqrt{\frac{2\alpha}{\beta}\varepsilon}.$$

Taking the expectation in (25) then gives the result. ■

## 4 Maximum Entropy on the Mean as Empirical Risk Minimization

In the sequel, we show that the dual MEM problem can be construed as a risk minimization problem. This enables us to solve the dual empirical MEM problem using stochastic gradient-based methods and to leverage standard results on the statistical properties of empirical risk minimization problems.

### 4.1 An Empirical Risk Minimization Perspective

The classical study of empirical risk minimization consists of analyzing the effects of estimating $\theta_\mu \in \operatorname{argmin}_{\theta\in\Theta} \mathbb{E}_\mu[\ell(X,\theta)]$ by $\theta_{\hat{\mu}_n} \in \operatorname{argmin}_{\theta\in\Theta} \mathbb{E}_{\hat{\mu}_n}[\ell(X,\theta)]$ where $\ell$ is a loss function. In words, we wish to perform inference on the minimizer of a function of interest, but only observe surrogates of this function which are obtained via sample-based approximations. Standard results regarding consistency, rates of convergence, and limit distributions are available in the reference textbooks (Kosorok, 2008; van der Vaart and Wellner, 1996; van der Vaart, 2000) under the broader study of $M$-estimation.

Remarkably, for a given prior $\mu \in \mathcal{P}(\mathcal{X})$, the dual problem $(D_\mu)$ can be written in the standard form described above. Precisely, for each $z \in \mathbb{R}^d$,

$$\alpha g^*(-z/\alpha) - \langle b, z\rangle + \log \int \exp\langle A^\top z, \cdot\rangle d\mu = \log \int \exp\Big(\alpha g^*(-z/\alpha) - \langle b, z\rangle + \langle A^\top z, \cdot\rangle\Big)\, d\mu.$$

As the logarithm is strictly monotonically increasing, the optimization problem

$$\min_{z\in\mathbb{R}^d} \int \exp\Big(\alpha g^*(-z/\alpha) - \langle b, z\rangle + \langle A^\top z, \cdot\rangle\Big)\, d\mu$$

has the exact same solution(s) as the MEM dual problem $(D_\mu)$. It follows that the population-level dual problem can be written equivalently as

$$z_\mu \in \operatorname*{argmin}_{z\in\mathbb{R}^d} \mathbb{E}_\mu\left[\ell(X,z)\right] \text{ where } \ell(x,z) = \exp\Big(\alpha g^*(-z/\alpha) - \langle b, z\rangle + \langle A^\top z, x\rangle\Big),$$

whereas the empirical dual problem is equivalent to

$$z_{\hat{\mu}_n} \in \operatorname*{argmin}_{z\in\mathbb{R}^d} \mathbb{E}_{\hat{\mu}_n}\left[\ell(X,z)\right] = \operatorname*{argmin}_{z\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \exp\Big(\alpha g^*(-z/\alpha) - \langle b, z\rangle + \langle A^\top z, X_i\rangle\Big),$$

from which one can recover $x_\mu$ and $x_{\hat{\mu}_n}$ from their respective primal-dual recovery formulae, see (13). This formulation coincides with the standard empirical risk minimization setting described above and, importantly, preserves strong convexity of the original objective.

**Proposition 19** (Strong convexity under composition with exp)**.** *Suppose that $h : \mathbb{R}^d \to \overline{\mathbb{R}}$ is proper, lsc, and $\gamma$-strongly convex. Then, $z \in \mathbb{R}^d \mapsto e^{h(z)}$ is $\gamma'$-strongly convex for $\gamma' = \gamma e^{\inf_{\mathbb{R}^d} h} > 0$ and $\partial(e^h)(z) = \{ve^{h(z)} : v \in \partial h(z)\}$.*

**Proof** We record that exp is proper, lsc, convex, increasing, and finite valued. Hence, with the convention $e^{+\infty} = +\infty$, for any $z \in \mathrm{dom}(h) = \mathrm{dom}(e^h)$, we may apply the chain rule of Burke et al. (2021, Corollary 4) to obtain that

$$\partial(e^h)(z) = e^{h(z)}\partial h(z).$$

The same reference also asserts that $e^h$ is convex so that this subdifferential is well-defined. Now, let $z_1, z_2 \in \mathrm{dom}(e^h)$, and take $v_2 \in \partial h(z_2)$. Then,

$$e^{h(z_1)} - e^{h(z_2)} \geq e^{h(z_2)}\left(h(z_1) - h(z_2)\right) \geq e^{h(z_2)}\left(\langle v_2, z_1 - z_2\rangle + \frac{\gamma}{2}\|z_1 - z_2\|^2\right),$$

where the first inequality uses the (sub)gradient inequality (7) for the convex function $e^x$, and the latter uses the strong convexity of $h$, namely (9). We conclude by using the chain rule at $v_2$, observing $\{v_2 e^{h(z_2)} : v_2 \in \partial h(z_2)\} = \partial(e^h)(z_2)$ and $\gamma e^{h(z_2)} \geq \gamma e^{\inf_{\mathbb{R}^d} h}$, whereby

$$e^{h(z_1)} - e^{h(z_2)} \geq \langle u_2, z_1 - z_2\rangle + \frac{\gamma'}{2}\|z_1 - z_2\|^2, \qquad \forall u_2 \in \partial(e^h)(z_2),$$

which, by Proposition 1, implies strong convexity of $e^h$ with modulus $\gamma'$. ■

The implications of connecting MEM problems to empirical risk minimization are twofold:

First, we may leverage the existing literature on $M$-estimation to establish statistical properties of the empirical MEM problem. As noted above, there is a vast literature on analyzing $M$-estimators under different assumptions on the underlying problem. A recent example is given in the work of Brunel (2025) which establishes limit laws for $M$-estimators in a general setting. While rates of convergence (in expectation) for $M$-estimators are accounted for in the literature (cf., e.g., Birgé and Massart 1993), these results are stated in greater generality than is required here, and hence require the verification of a number of technical conditions. By contrast, the rates obtained in Section 3 follow from first principles of convex analysis and probability theory, enabling a simpler, and more self-contained exposition.

Second, the alternative formulation of the dual MEM problem is amenable to optimization via stochastic gradient methods. As such, even in the case where a population-level prior, $\mu$, is known, but does not admit a tractable log moment generating function, we may approximate the gradient via a sample-based approximation and derive meaningful convergence results as expounded next.

### 4.2 Clipped SGD and convergence

As noted above, the dual MEM problem with prior $\mu \in \mathcal{P}(\mathcal{X})$ can be reformulated as

$$\min_{z\in\mathbb{R}^d} \mathbb{E}_\mu \left[\exp\left(\alpha g^*(-z/\alpha) - \langle b, z\rangle + \langle A^\top z, X\rangle\right)\right] =: \min_{z\in\mathbb{R}^d} \mathbb{E}_\mu \left[\ell(X, z)\right]. \tag{$D'_\mu$}$$

As such, the corresponding empirical MEM dual problem is

$$\min_{z\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \exp\left(\alpha g^*(-z/\alpha) - \langle b, z\rangle + \langle A^\top z, X_i\rangle\right) = \min_{z\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \ell(X_i, z). \tag{$D'_{\hat\mu_n}$}$$

Both of these problems are thus amenable to optimization via stochastic gradient methods. The main distinction between these problems is that optimizing $(D'_\mu)$ using such methods requires a mechanism to draw new i.i.d. samples from the population measure $\mu$ at each iteration to approximate the population gradient via mini-batching. On the other hand, if we only have a fixed finite number of samples from $\mu$ (e.g., a dataset) we may also consider solving $(D'_{\hat\mu_n})$ using stochastic methods by taking random subsets of the finite collection of samples to form mini-batched gradients. This is particularly useful in the case that the number of samples is so large that storing the entire dataset in memory is prohibitively expensive or that summing over all of the samples is costly. The results of the previous section establish that $x_{\hat\mu_n}$ can serve as a good proxy for $x_\mu$ (on average) if the samples used to form $\hat\mu_n$ are i.i.d. from $\mu$ and $n$ is sufficiently large.

This perspective is thus advantageous in cases where excessively large amounts of data are available, as it enables scalability via mini-batching gradients and the ability to stream the data in place of storing it. On the other hand, dropping the outer logarithm introduces numerical instability as the gradients involve exponential terms.

For instance, if $g = g^* = \frac{1}{2}\|\cdot\|^2$,

$$\nabla \mathbb{E}_\mu \left[\ell(X, z)\right] = \mathbb{E}_\mu \left[\left(\frac{1}{\alpha}z - b + AX\right)\exp\left(\frac{1}{2\alpha}\|z\|^2 + \langle AX - b, z\rangle\right)\right],$$

which can lead to numerical overflow if $\|z\|$ is even moderately large.

To account for this, we propose to use a stochastic gradient method with clipped gradients as discussed in Mai and Johansson (2021). The crux of this approach consists of normalizing the stochastic gradient before taking a step to avoid taking excessively large steps. We note, however, that the method analyzed in that work appears to tacitly assume that the function being minimized has full domain[†] and that subgradients are chosen in a measurable manner. To account for functions with restricted domain, we introduce a projection step into the clipped SGD method and work under appropriate conditions guaranteeing that all iterates of the algorithm are in $\operatorname{int}(\operatorname{dom}(g^*(-\cdot/\alpha)))$ which guarantees that subgradients at each iterate are bounded. We provide a sufficient condition for this assumption to hold in (A2) ahead.

[†]In effect, the clipped SGD update (Mai and Johansson, 2021, Equation 2) does not enforce that iterates lie in the domain of the objective. Further, their assumption A4 requires the expectation of the squared norm of a subgradient to be uniformly bounded. As the subdifferential of a convex function $f$ is non-empty and bounded only at points in $\operatorname{int}(\operatorname{dom}(f))$ such conditions generally cannot hold at boundary points of $\operatorname{dom}(f)$.

#### 4.2.1 Projected Clipped SGD

We now state and explain the proposed algorithm. As aforementioned, there are two situations to consider, namely

(S1) we have access to a mechanism which generates fresh collections of i.i.d. samples of a prescribed size from the population $\mu$ which are conditionally independent from the previously generated collections.

(S2) we have a fixed dataset $\mathcal{D} = (x_i)_{i=1}^n$ of samples from $\mu$ and have access to a mechanism which generates fresh collections of i.i.d. samples from the corresponding empirical measure $\hat{\mu}_n$ which are conditionally independent from the previous collections and from the original dataset.

**Remark 20** (Generating batches under (S2))**.** *Given a fixed dataset $\mathcal{D}$, a standard approach for generating batches of samples satisfying the conditions of (S2) consists of simply resampling $m$ samples from $\mathcal{D}$ uniformly at random with replacement. That is, we generate an i.i.d. collection of indices $\{\sigma_1, \dots, \sigma_m\}$ from the uniform distribution on $\{1, \dots, n\}$ which is independent from previous draws and the original dataset, and take the mini-batch to be $\{x_{\sigma_1}, \dots, x_{\sigma_m}\}$. With this, $\{x_{\sigma_1}, \dots, x_{\sigma_m}\}$ is conditionally i.i.d. from $\hat{\mu}_n$ and independent from the previously drawn mini-batches conditionally on $\mathcal{D}$ so that (S2) holds.*

Notably, when applying Algorithm 1 ahead using mini-batched gradients with collections generated according to (S1) the aim is to approximately solve $(D'_\mu)$, i.e., the dual MEM problem with prior $\bar{\mu} = \mu$. If the collections are generated according to (S2), we approximately solve $(D'_{\hat{\mu}_n})$, i.e., the dual MEM problem with prior $\bar{\mu} = \hat{\mu}_n$. Here and in the sequel, we use $\bar{\mu}$ to denote the population prior $\mu$ under (S1) or the empirical prior $\hat{\mu}_n$ under (S2) to simplify notation. To clearly state and motivate the projected clipped SGD method, we make the following assumption

$$\exists r > 0 : \ B_r \subset \operatorname{int}(\operatorname{dom}(\mathbb{E}_{\bar{\mu}}[\ell(X, \cdot)])) = \operatorname{int}(\operatorname{dom}(g^*(-(\cdot)/\alpha))) \text{ and } z_{\bar{\mu}} \in B_r. \quad \text{(A2)}$$

As noted in Theorem 8 (ii), $\|z_{\bar{\mu}}\| \leq \frac{\alpha}{\beta}\big(\min_{v \in \partial g^*(0)} \|v\| + \|b\| + \|A\||\mathcal{X}|_\infty\big) =: r$, where we underscore that $\partial g^*(0)$ is non-empty and bounded as $0 \in \operatorname{int}(\operatorname{dom}(g^*))$ by assumption, see Rockafellar (1970, Theorem 23.4). It follows that the above assumption holds if $\operatorname{int}(B_{\bar{r}/\alpha}) \subset \operatorname{int}(\operatorname{dom}(g^*(\cdot)))$ for some $r < \bar{r}$ which is equivalent to the following condition on $g$ itself.

The fidelity term $g$ satisfies $\liminf_{\|z\| \to \infty} \frac{g(z)}{\|z\|} \geq \frac{\bar{r}}{\alpha}$ for some $\bar{r} > r$.

Thus, if $\lim_{\|z\| \to \infty} \frac{g(z)}{\|z\|} = \infty$, i.e., $g$ is supercoercive, (A2) is satisfied. For instance, the prototypical fidelity term $g = \frac{1}{2}\|\cdot\|^2$ is supercoercive. We now state and prove this equivalence.

**Proposition 21.** *Suppose that $h : \mathbb{R}^d \to \overline{\mathbb{R}}$ is proper, lsc, and convex. Then,* $\operatorname{int}(\operatorname{dom}(h^*))$ *contains the open ball of radius $\bar{r} > 0$ centered at 0 if and only if* $\liminf_{\|z\| \to \infty} \frac{h(z)}{\|z\|} \geq \bar{r}$.

**Proof** By Rockafellar (1970, Corollary 13.3.4 (c)), $z \in \operatorname{int}(\operatorname{dom}(h^*))$ if and only if the recession function of $h - \langle \cdot, z \rangle$ is strictly greater than 0 on $\mathbb{R}^d \backslash \{0\}$. Following Rockafellar and Wets (1998, Theorem 3.26 (a)), this latter condition is equivalent to the property that $h - \langle \cdot, z \rangle$ is bounded below on bounded sets and

$$\liminf_{\|w\| \to \infty} \frac{h(w) - \langle w, z \rangle}{\|w\|} > 0.$$

Now, suppose that $\liminf_{\|w\| \to \infty} \frac{h(w)}{\|w\|} \geq \bar{r} > 0$, then, for any $z \in \mathbb{R}^d$ with $\|z\| < \bar{r}$,

$$\liminf_{\|w\| \to \infty} \frac{h(w) - \langle w, z \rangle}{\|w\|} \geq \bar{r} - \limsup_{\|w\| \to \infty} \frac{\langle w, z \rangle}{\|w\|} > \bar{r} - \bar{r} = 0.$$

Evidently $h - \langle \cdot, z \rangle$ is bounded below on any bounded set so that $z \in \operatorname{int}(\operatorname{dom}(h^*))$ by the previous deliberations. Thus the open ball of radius $\bar{r}$ is contained in $\operatorname{int}(\operatorname{dom}(h^*))$.

On the other hand, if the open ball of radius $\bar{r}$ is contained in $\operatorname{int}(\operatorname{dom}(h^*))$, we have from Rockafellar and Wets (1998, Theorems 3.26 and 11.5) that, for any $r' < \bar{r}$,

$$\liminf_{\|w\| \to \infty} \frac{h(w)}{\|w\|} = \inf_{\|w\|=1} \sup_{v \in \operatorname{dom}(h^*)} \langle w, v \rangle \geq r' \inf_{\|w\|=1} \|w\| = r',$$

since, for any $w \in \mathbb{R}^d$, $r' \frac{w}{\|w\|} \in \operatorname{dom}(h^*)$ and so $\sup_{v \in \operatorname{dom}(h^*)} \langle w, v \rangle \geq r' \|w\|$. Taking the limit $r' \to \bar{r}$ above proves the claimed result. ■

With this explanation of the projection step in hand, we now state the algorithm:

**Algorithm 1:** Projected Clipped SGD

**Data:** Fix a radius $r$ satisfying (A2). Initialize $z_0 \in B_r$, a stepsize schedule $(\beta_k)_{k \in \mathbb{N}}$, a sequence of batch sizes $(m_k)_{k \in \mathbb{N}}$, and a clipping parameter $\gamma > 0$.

**for** $k = 0, 1, 2, 3 \ldots$ **do**

Draw a collection of samples $x_{k_1}, \ldots, x_{k_{m_k}}$ according to (S1) or (S2)

$g_k \leftarrow \frac{1}{m_k} \sum_{i=1}^{m_k} \ell'(x_{k_i}, z_k)$, for a measurable selection $\ell'(x_{k_i}, z_k) \in \partial_z \ell(x_{k_i}, z_k)$

$z_{k+1} \leftarrow P_{B_r}\left(z_k - \beta_k \min\{1, \frac{\gamma}{\|g_k\|}\} g_k\right)$

The main feature of this algorithm is the final line, known as the clipping step, where instead of taking the usual gradient step $-\beta_k g_k$, the update is enforced to have norm $\beta_k \min\{\|g_k\|, \gamma\}$. This prevents instability if the gradient has a large magnitude. Our method also includes a projection onto $B_r$, denoted $P_{B_r}$, which ensures that the iterates remain in the interior of the domain of the objective; this step is not present in the original algorithm. As presented, the algorithm does not include a stopping criterion and, in our examples, is run up to a pre-specified allowance of iterations.

The remainder of this section is devoted to proving the almost sure convergence of Algorithm 1 for the empirical risk formulation of the MEM dual problem. We do so by adapting the proof techniques of Mai and Johansson (2021) for the projected version of the CSGD algorithm. As noted previously, the results of that work only appear to apply to objectives with full domain. We begin by compiling some useful properties of $\ell$.

**Lemma 22** (Properties of $\ell$). *Let $g, r$ satisfy assumptions* (A1) *and* (A2) *and, for $x \in \mathcal{X}$, let*

$$\ell'(x,z) = (v(z) - b + Ax)\exp(\alpha g^*(-z/\alpha) - \langle b, z\rangle + \langle A^\top z, x\rangle), \tag{26}$$

*where $v(z) \in -\partial g^*(-z/\alpha)$ is independent of $x$ and $z \in B_r \mapsto v(z)$ is measurable. Then $\ell'(x,z)$ is a measurable selection of subgradient of $\ell(x,\cdot)$ at $z \in B_r$ and, if $X \sim \bar{\mu}$, the following hold:*

*(P1) (Unbiased gradient estimates).* $\mathbb{E}[\ell'(X,z)] \in \partial\,\mathbb{E}[\ell(X,z)]$.

*(P2) (Quadratic growth). Let $\{z_{\bar{\mu}}\} = \operatorname{argmin}_{\mathbb{R}^d} \mathbb{E}[\ell(X,\cdot)]$. There exists $a > 0$ for which*

$$\mathbb{E}[\ell(X,z)] - \mathbb{E}[\ell(X,z_{\bar{\mu}})] \geq a\|z - z_{\bar{\mu}}\|^2 \text{ for all } z \in B_r$$

*(P3) (Finite variance). There exists $\sigma > 0$ such that*

$$\mathbb{E}[\|\ell'(X,z) - \mathbb{E}[\ell'(X,z)]\|^2] \leq \sigma^2 \text{ for all } z \in B_r.$$

**Proof** We begin by proving that $\ell'(x,z)$ corresponds to a measurable selection of subgradient. From Proposition 19, for each $x \in \mathcal{X}$ and $z \in B_r \subset \operatorname{int}(\operatorname{dom}(\ell(x,\cdot)))$,

$$\partial_z \ell(x,z) = (-\partial g^*(-z/\alpha) - b + Ax)\exp(\alpha g^*(-z/\alpha) - \langle b, z\rangle + \langle A^\top z, x\rangle).$$

Hence, $\ell'(x,z)$ from (26) corresponds to a certain choice of subgradient of $\ell(x,\cdot)$. For fixed $z, v$, $\ell'(\cdot,z)$ is a continuous function on the compact set $\mathcal{X}$ and is, in particular, bounded and measurable and, in fact, integrable.
(P1) Given that $\ell'(x,z) \in \partial_z \ell(x,z)$,

$$\ell(x,z') - \ell(x,z) \geq \langle \ell'(x,z), z' - z\rangle, \text{ for each } z' \in \mathbb{R}^d.$$

Taking expectations on both sides of the inequality, we obtain that

$$\mathbb{E}[\ell(X,z')] - \mathbb{E}[\ell(X,z)] \geq \langle \mathbb{E}[\ell'(X,z)], z' - z\rangle, \text{ for each } z' \in \mathbb{R}^d.$$

This demonstrates that $\mathbb{E}[\ell'(X,z)]$ is a subgradient of $\mathbb{E}[\ell(X,\cdot)]$ at $z$ as desired.
(P2) Fix $z \in B_r$, and let $z_{\bar{\mu}} \in B_r$ be the unique minimizer of $\mathbb{E}[\ell(X,z)]$. Then

$$\mathbb{E}[\ell(X,z)] - \mathbb{E}[\ell(X,z_{\bar{\mu}})] = \mathbb{E}[\exp h_X(z)] - \mathbb{E}[\exp h_X(z_{\bar{\mu}})],$$

where $h_x(z) := \alpha g^*(-z/\alpha) - \langle b, z\rangle + \langle A^\top z, x\rangle$ for $x \in \operatorname{supp}(\bar{\mu})$. This function is proper, lsc, and $\beta/\alpha$-strongly convex for all $x$ (cf. (A1)), so that, for

$$\gamma_x := \frac{\beta}{\alpha} e^{\inf_{\mathbb{R}^d} h_x},$$

the function $\exp h_x$ is $\gamma_x$-strongly convex in $z$ by Proposition 19. Thus, for each $\lambda \in [0,1]$ and any $z, z' \in B_r$,

$$\exp h_x\left(\lambda z + (1-\lambda)z'\right) \leq \lambda \exp h_x(z) + (1-\lambda)\exp h_x(z') - \frac{\gamma_x}{2}\lambda(1-\lambda)\|z - z'\|^2.$$

Taking the expectation on both sides of the inequality shows that $\mathbb{E}[\ell(X,\cdot)]$ is strongly convex with modulus $\mathbb{E}[\gamma_X] \geq \frac{\beta}{\alpha} e^{\inf_{x\in\operatorname{supp}(\bar{\mu})} \inf_{\mathbb{R}^d} h_x} > 0$. This final lower bound is due to the fact that, setting $\bar{z}$ as the unique global minimizer of $g^*(-\cdot/\alpha)$,

$$\begin{aligned} h_x(z) &\geq \alpha g^*(-\bar{z}/\alpha) + \frac{\beta}{2\alpha}\|z-\bar{z}\|^2 - \|z\|\|b\| - \|A\||\mathcal{X}|_\infty\|z\| \\ &\geq \alpha g^*(-\bar{z}/\alpha) + \frac{\beta}{2\alpha}(\|z\|^2 - 2\|z\|\|\bar{z}\| + \|\bar{z}\|^2) - \|z\|\|b\| - \|A\||\mathcal{X}|_\infty\|z\|, \end{aligned}$$

where the final lower bound is a convex parabola in $\|z\|$ which is independent of $x$ and hence admits a finite minimum. By (9) and using the fact that $z_{\bar{\mu}}$ is the minimizer of $\mathbb{E}[\ell(X,\cdot)]$,

$$\mathbb{E}[\ell(X,z)] - \mathbb{E}[\ell(X,z_{\bar{\mu}})] \geq \frac{\mathbb{E}[\gamma_X]}{2}\|z-z_{\bar{\mu}}\|^2,$$

proving the claim.
(P3) Given $z \in B_r$, we have from (26) that

$$\ell'(x,z) = (v(z) - b + Ax)\exp(\alpha g^*(-z/\alpha) - \langle b, z\rangle + \langle A^\top z, x\rangle)$$

for some $v(z) \in -\partial g^*(-z/\alpha)$. Thus,

$$\begin{aligned} &\ell'(x,z) - \mathbb{E}[\ell'(X,z)] \\ &= (v(z)-b)\left(\exp(\alpha g^*(-z/\alpha) - \langle b,z\rangle + \langle A^\top z, x\rangle) - \mathbb{E}\left[\exp(\alpha g^*(-z/\alpha) - \langle b,z\rangle + \langle A^\top z, X\rangle)\right]\right) \\ &+ A\left(x\exp(\alpha g^*(-z/\alpha) - \langle b,z\rangle + \langle A^\top z, x\rangle) - \mathbb{E}\left[X\exp(\alpha g^*(-z/\alpha) - \langle b,z\rangle + \langle A^\top z, X\rangle)\right]\right). \end{aligned}$$

Taking the squared norm on both sides of the equality followed by the expectation, we have

$$\begin{aligned} \mathbb{E}\left[\|\ell'(X,z) - \mathbb{E}[\ell'(X,z)]\|^2\right] \leq{}& 2\|v(z)-b\|^2 \operatorname{var}\left(\exp(\alpha g^*(-z/\alpha) - \langle b,z\rangle + \langle A^\top z, X\rangle)\right) \\ &+ 2\|A\|^2 \operatorname{tr}\left(\operatorname{cov}\left(X\exp(\alpha g^*(-z/\alpha) - \langle b,z\rangle + \langle A^\top z, X\rangle)\right)\right). \end{aligned}$$

Applying standard bounds for the variance and trace of the covariance,

$$\begin{aligned} \mathbb{E}\left[\|\ell'(X,z) - \mathbb{E}[\ell'(X,z)]\|^2\right] \leq{}& 2\|v(z)-b\|^2\mathbb{E}\left[\exp(\alpha g^*(-z/\alpha) - \langle b,z\rangle + \langle A^\top z, X\rangle)^2\right] \\ &+ 2\|A\|^2\mathbb{E}\left[\left\|X\exp(\alpha g^*(-z/\alpha) - \langle b,z\rangle + \langle A^\top z, X\rangle)\right\|^2\right] \end{aligned}$$

and so

$$\begin{aligned} &\mathbb{E}\left[\|\ell'(X,z) - \mathbb{E}[\ell'(X,z)]\|^2\right] \\ &\leq 2\left(\sup_{v\in-\partial g^*(-z/\alpha)}\|v-b\|^2 + \|A\|^2|\mathcal{X}|_\infty^2\right)\exp\left(2\left(\sup_{z\in B_r}\alpha g^*(-z/\alpha) + \|b\|r + |\mathcal{X}|_\infty\|A\|r\right)\right). \end{aligned}$$

Note that since $B_{\frac{r}{\alpha}} \subset \operatorname{int}(\operatorname{dom} g^*)$ by assumption, Rockafellar (1970, Theorem 24.7) asserts that $\partial g^*(B_{\frac{r}{\alpha}}) = \cup_{z\in B_{\frac{r}{\alpha}}}\partial g^*(z)$ is bounded. Therefore $\sup_{z\in B_r}\sup_{v\in-\partial g^*(-z/\alpha)}\|v-b\|^2 < \infty$.

Furthermore, $\sup_{z\in B_r} g^*(-z/\alpha) < +\infty$ as any convex function is continuous on the interior of its domain (Rockafellar, 1970, Theorem 10.1) and $B_r$ is compact. It follows that $\mathbb{E}\left[\|\ell'(X,z) - \mathbb{E}[\ell'(X,z)]\|^2\right] \leq \sigma^2$ for a finite constant $\sigma^2 > 0$. ■

We note that a measurable selection of subgradient always exists under (A2) (cf. e.g., Rockafellar and Wets, 1998, Corollary 14.6 and Theorem 14.56) and, in cases where $g^*$ is differentiable, the subdifferential is a singleton so that no choice is required. In any case subgradient selection rule which can be expressed in terms of compositions of Borel measurable operations will be measurable recalling Remark 5. Furthermore by taking measurable subgradients it is guaranteed that the iterates of the Algorithm 1 are also measurable.

The properties proved in Lemma 22 are paramount in showing the convergence of Algorithm 1 as stated next.

**Theorem 23** (Convergence of Algorithm 1)**.** *Let $g$ and $r$ satisfy assumptions A1 and A2 and suppose that the stepsizes $\beta_k \geq 0$ satisfy $\sum_{k=0}^{\infty}\beta_k = +\infty$, $\sum_{k=0}^{\infty}\beta_k^2 < \infty$ and $\sum_{k=0}^{\infty}\beta_k/m_k < +\infty$ Then, any sequence of iterates $z_k$ generated by Algorithm 1 as applied to $f(x,z) = \ell(x,z)$ converges almost surely to $\{z_{\bar{\mu}}\} = \operatorname{argmin}_{\mathbb{R}^d}\mathbb{E}_{X\sim\bar{\mu}}[\ell(X,\cdot)]$.*

**Proof** The proof of this result follows similar lines to that of Mai and Johansson (2021, Theorem 1), but accounts for extended real-valued functions. For notational brevity, we introduce the quantity $\rho_k = \min\left\{1, \frac{\gamma}{\|g_k\|}\right\}$. Then, as $P_{B_r}$ is non-expansive,

$$\begin{aligned}\|z_{k+1} - z_{\bar{\mu}}\|^2 &= \|P_{B_r}(z_k - \beta_k\rho_k g_k) - z_{\bar{\mu}}\| \\ &\leq \|z_k - \beta_k\rho_k g_k - z_{\bar{\mu}}\|^2 \\ &= \|z_k - z_{\bar{\mu}}\|^2 - 2\beta_k\rho_k\langle g_k - H_k + H_k, z_k - z_{\bar{\mu}}\rangle + \beta_k^2\rho_k^2\|g_k\|^2. \end{aligned} \tag{27}$$

Here, $z_k \in B_r$, by the projection step combined with the assumption $z_0 \in B_r$, and $H_k = \mathbb{E}[g_k|\mathcal{F}_k]$, where $\mathcal{F}_k$ is the $\sigma$-algebra generated by the first $k$ batches[‡] under (S1) and, under (S2), $\mathcal{F}_k$ also includes the dataset $\mathcal{D}$ as a generator. From Lemma 22 (P2),

$$a\|z_k - z_{\bar{\mu}}\|^2 \leq \mathbb{E}[\ell(X,z_k)] - \mathbb{E}[\ell(X,z_{\bar{\mu}})] \leq \langle H_k, z_k - z_{\bar{\mu}}\rangle, \tag{28}$$

where the second inequality follows from the subgradient inequality for convex functions. Indeed, noting that conditionally on $\mathcal{F}_k$, $z_k$ is fixed and the mini-batch used to obtain $g_k$ is independent of $\mathcal{F}_k$, $\mathbb{E}[g_k|\mathcal{F}_k] = \mathbb{E}_{X\sim\bar{\mu}}[\ell'(X,z_k)] \in \partial(\mathbb{E}[\ell(X,\cdot)])(z_k)$ by Lemma 22 (P1). Applying the Cauchy-Schwarz and $\varepsilon$-Young inequalities with $\varepsilon = a$, we obtain that

$$\begin{aligned}|\langle g_k - H_k, z_k - z_{\bar{\mu}}\rangle| &\leq \|g_k - H_k\|\cdot\|z_k - z_{\bar{\mu}}\| \\ &\leq \frac{1}{2a}\|g_k - H_k\|^2 + \frac{a}{2}\|z_k - z_{\bar{\mu}}\|^2 \\ &\leq \frac{1}{2a}\|g_k - H_k\|^2 + \frac{1}{2}\langle H_k, z_k - z_{\bar{\mu}}\rangle,\end{aligned}$$

[‡]i.e. $\mathcal{F}_k = \sigma(x_{0_1}, x_{0_2}, \ldots x_{0_{m_0}}, \ldots, x_{k-1_1}, \ldots, x_{k-1_{m_{k-1}}})$ is the smallest $\sigma$-algebra containing the first $k$ batches, where $x_{i_j}$ as defined in Algorithm 1 is the $j$-th mini-batch sample at iteration $i$.

where we have substituted (28) in the final step. Applying this bound in (27) yields that

$$\|z_{k+1} - z_{\bar{\mu}}\|^2 \leq \|z_k - z_{\bar{\mu}}\|^2 - \beta_k \rho_k \langle H_k, z_k - z_{\bar{\mu}} \rangle + \frac{\beta_k \rho_k}{a} \|g_k - H_k\|^2 + \beta_k^2 \rho_k^2 \|g_k\|^2$$

and so, taking the conditional expectation and using the fact that $\rho_k \|g_k\| \leq \gamma$ and $\rho_k \leq 1$,

$$\begin{aligned} &\mathbb{E}\Big[\|z_{k+1} - z_{\bar{\mu}}\|^2 | \mathcal{F}_k\Big] \\ &\leq \|z_k - z_{\bar{\mu}}\|^2 - \beta_k \langle H_k, z_k - z_{\bar{\mu}} \rangle \mathbb{E}[\rho_k | \mathcal{F}_k] + \frac{\beta_k}{a} \mathbb{E}\big[\|g_k - H_k\|^2 | \mathcal{F}_k\big] + \gamma^2 \beta_k^2. \end{aligned} \tag{29}$$

We now invoke the classical almost supermartingale convergence theorem of Robbins and Siegmund (1971, Theorem 1), which states that if $X_k, Y_k, U_k, V_k, W_k$ are non-negative $\mathcal{F}_k$-measurable random variables satisfying

$$\mathbb{E}[X_{k+1} | \mathcal{F}_k] \leq X_k(1 + U_k) + V_k - W_k,$$

for each $k \in \mathbb{N}$, then $\lim_{k \to \infty} X_k$ exists and is almost surely finite and $\sum_{k=1}^{\infty} W_k < \infty$ almost surely on the event that $\sum_{k=1}^{\infty} U_k$ and $\sum_{k=1}^{\infty} V_k$ are finite. Evidently (29) is of this form with $X_k = \|z_k - z_{\bar{\mu}}\|^2, U_k = 0, V_k = \frac{\beta_k}{a} \mathbb{E}\big[\|g_k - H_k\|^2 | \mathcal{F}_k\big] + \gamma^2 \beta_k^2$, and $W_k = \beta_k \langle H_k, z_k - z_{\bar{\mu}} \rangle \mathbb{E}[\rho_k | \mathcal{F}_k]$ (which was shown to be non-negative in (28)). It follows that there exists an almost surely finite random variable $Z_\infty$ for which

$$\|z_k - z_{\bar{\mu}}\|^2 \text{ converges a.s. to } Z_\infty \text{ and } \sum_{k=0}^{\infty} \beta_k \langle H_k, z_k - z_{\bar{\mu}} \rangle \mathbb{E}[\rho_k | \mathcal{F}_k] < \infty \text{ a.s.}$$

on the event that $\sum_{k=0}^{\infty} \left( \frac{\beta_k}{a} \mathbb{E}\big[\|g_k - H_k\|^2 | \mathcal{F}_k\big] + \gamma^2 \beta_k^2 \right) < \infty$. We now establish that this event occurs with probability 1.

To this end, recall from the previous deliberations that, conditionally on $\mathcal{F}_k$, $z_k$ is fixed and, since the mini-batch samples used for $g_k$ are independent of $\mathcal{F}_k$ and independent of each other,

$$\begin{aligned} \mathbb{E}\big[\|g_k - H_k\|^2 | \mathcal{F}_k\big] &= \mathbb{E}\left[ \left\| {m_k}^{-1} \sum_{i=1}^{m_k} \left( \ell'(x_{k_i}, z_k) - \mathbb{E}[\ell'(X, z_k)] \right) \right\|^2 \Big| \mathcal{F}_k \right] \\ &= \frac{1}{m_k^2} \sum_{i=1}^{m_k} \mathbb{E}_{X \sim \bar{\mu}} \left[ \left\| \ell'(X, z_k) - \mathbb{E}[\ell'(X, z_k)] \right\|^2 \right] \leq \sigma^2 / m_k, \end{aligned} \tag{30}$$

where the final inequality is due to Lemma 22 (P3). Conclude that

$$\sum_{k=0}^{\infty} \left( \frac{\beta_k}{a} \mathbb{E}\big[\|g_k - H_k\|^2 | \mathcal{F}_k\big] + \gamma^2 \beta_k^2 \right) \leq \sum_{k=0}^{\infty} \left( \frac{\beta_k \sigma^2}{a m_k} + \gamma^2 \beta_k^2 \right) < \infty$$

as follows from the assumptions that $\sum_{k=0}^{\infty} \beta_k^2 < \infty$ and $\sum_{k=0}^{\infty} \frac{\beta_k}{m_k} < \infty$. That is, this event occurs with probability 1.

It follows that $\sum_{k=0}^{\infty} \beta_k \langle H_k, z_k - z_{\bar{\mu}} \rangle \mathbb{E}[\rho_k | \mathcal{F}_k] < \infty$ a.s. and $\|z_k - z_{\bar{\mu}}\|^2$ converges a.s. to $Z_\infty$ which is almost surely finite. It remains to be shown that $Z_\infty = 0$. Following the proof of (Mai and Johansson, 2021, Theorem 1) we have that

$$\mathbb{E}[\rho_k | \mathcal{F}_k] \geq \frac{\gamma}{\gamma + (\mathbb{E}[\|g_k\|^2 | \mathcal{F}_k])^{\frac{1}{2}}}. \tag{31}$$

We now note that

$$\mathbb{E}[\|g_k\|^2 | \mathcal{F}_k] = \|H_k\|^2 + \mathbb{E}[\|g_k - H_k\|^2 | \mathcal{F}_k] \leq \|H_k\|^2 + \frac{\sigma^2}{m_k}, \tag{32}$$

by applying (30). By the conditional Jensen's inequality,

$$\Big\| \mathbb{E}\Big[ \frac{1}{m_k} \sum_{i=1}^{m_k} \ell'(x_{k_i}, z_k) | \mathcal{F}_k \Big] \Big\| \leq \mathbb{E}\Big[ \frac{1}{m_k} \sum_{i=1}^{m_k} \|\ell'(x_{k_i}, z_k)\| | \mathcal{F}_k \Big]$$

and, recalling (26),

$$\begin{aligned} \|\ell'(x_{k_i}, z_k)\| &= \|v_{k_i} - b + A x_{k_i}\| \exp(\alpha g^*(-z_k/\alpha) - \langle b, z_k \rangle + \langle A^\top z_k, x_{k_i} \rangle) \\ &\leq (\|v_{k_i}\| + \|b\| + \|A\| |\mathcal{X}|_\infty) \exp\left( \sup_{B_r} \alpha g^*(-(\cdot)/\alpha) + \|b\| r + \|A\| |\mathcal{X}|_\infty r \right) \end{aligned}$$

where $v_{k_i} \in -\partial g^*(-z_k/\alpha)$. Combining these two displayed equations, we obtain that

$$\|H_k\| \leq \Big( \sup_{z \in B_r} \sup_{v \in \partial g^*(-\frac{z}{\alpha})} \|v\| + \|b\| + \|A\| |\mathcal{X}|_\infty \Big) \exp\left( \sup_{B_r} \alpha g^*(-(\cdot)/\alpha) + \|b\| r + \|A\| |\mathcal{X}|_\infty r \right),$$

and underscore that $\sup_{z \in B_r} \sup_{v \in \partial g^*(-\frac{z}{\alpha})} \|v\| < \infty$ as argued in the proof of Lemma 22 (P3). Letting $S$ denote the right-hand side of the above display, insert this bound in (31) which yields that

$$\infty > \sum_{k=0}^{\infty} \beta_k \langle H_k, z_k - z_{\bar{\mu}} \rangle \mathbb{E}[\rho_k | \mathcal{F}_k] \geq a \frac{\gamma}{\gamma + \sqrt{S^2 + \sigma^2}} \sum_{k=0}^{\infty} \beta_k \|z_k - z_{\bar{\mu}}\|^2, \text{ almost surely,}$$

where we have also applied (28) and (32) and used the fact that $m_k \geq 1$. It follows that $\sum_{k=0}^{\infty} \beta_k \|z_k - z_{\bar{\mu}}\|^2 < \infty$ almost surely. On the other hand, $\|z_k - z_{\bar{\mu}}\|^2$ converges a.s. to some $Z_\infty < \infty$.

Suppose now that $Z_\infty > 0$ on a set with positive probability (note that $Z_\infty$ is non-negative by construction). Then, $Z_\infty > 0$ and $\|z_k - z_{\bar{\mu}}\|^2 \to Z_\infty$, $\sum_{k=0}^{\infty} \beta_k \|z_k - z_{\bar{\mu}}\|^2 = \infty$ since $\sum_{k=0}^{\infty} \beta_k = +\infty$, but this contradicts the fact that $\sum_{k=0}^{\infty} \beta_k \|z_k - z_{\bar{\mu}}\|^2 < \infty$. Conclude that $Z_\infty = 0$ with probability 1 so that $\|z_k - z_{\bar{\mu}}\| \to 0$ a.s., as required. ■

## 5 Numerical Experiments

Having established the convergence of projected clipped SGD we now study the performance of this method on image denoising problems. Throughout we set the fidelity term to $g = \frac{1}{2}\|\cdot\|_2^2$ so that $g^* = \frac{1}{2}\|\cdot\|_2^2$ has full domain and $g^*$ is smooth so that no measurability issues arise in the subgradient selection. Moreover, (A2) is satisfied for any choice of $r > 0$ such that $z_\mu \in B_r$. We emphasize that Theorem 8 (ii) ensures that such $r$ is available.

In the following experiments we set $r = 500$ for simplicity and note that the iterates in each run all lie in the interior of this ball. In the sequel, we adopt setting (S2) wherein a dataset $\mathcal{D}$ of fixed sized $n$ is given and the corresponding prior is $\hat{\mu}_n$. In this context, the primal-dual recovery can be accomplished as follows.

**Algorithm 2:** Primal Solution Recovery

**Data:** Output $z^*$ of Algorithm 1.
Initialize $\ell_1 = 0 \in \mathbb{R}^m$, $\ell_2 = 0 \in \mathbb{R}$.
**for** $k = 1, 2, 3, \ldots, n$ **do**
  $\ell_1 \leftarrow \ell_1 + x_k \exp\langle A^\top z^*, x_k\rangle$
  $\ell_2 \leftarrow \ell_2 + \exp\langle A^\top z^*, x_k\rangle$
Return $x^* = \ell_1/\ell_2$.

For our experiments, we apply some common *ad hoc* engineering techniques to improve the performance of the method. Concretely, we apply a "cosine learning rate",as implemented in, e.g., PyTorch which consists of taking $\beta_k$ to behave like a cosine curve and is common in deep learning problems, see the specific hyperparameters for each experiment for explicit expressions. Moreover, for certain larger experiments we incorporate momentum to this algorithm as is standard in the SGD literature, see Section 5.2 for details.

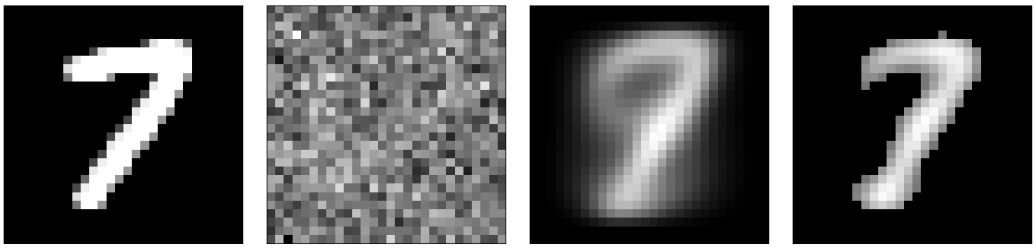
The numerics are arranged as follows. We begin with an experiment on the extended MNIST (EMNIST) dataset (Cohen et al., 2017) which consists of $28 \times 28$ black and white pictures of letters and digits, see Section 5.1 for details and Figures 1 and 2 for the results. We next apply our method to the more realistic Street View House Numbers (SVHN) dataset (Netzer et al., 2011) which consists of $32 \times 32$ black and white and colored cropped images of house number digits. Exact details are provided in Section 5.2. We use this dataset to compare the performance of stochastic gradient methods with deterministic methods, see Figures 4 to 6 for examples. We conclude with a case where the method fails to accurately recover a noisy image which illustrates the importance of having representative data in the prior, see Figure 7.

### 5.1 Extended MNIST

For the first host of experiments, we use the Extended MNIST dataset which consists of 800,000 hand drawn 28x28 pixel images including the numbers 0-9, and letters A-Z. We underscore that directly loading this complete dataset into memory requires 4GB of RAM, so that stochastic gradient methods which allow streaming data will be required for larger real-life datasets. Importantly, optimizing the original objective with the empirical prior on the entire dataset using gradient-based methods requires evaluating the full gradient

$\nabla\phi_{\mu_n}(z)$ at any point $z$ and hence requires a full data-pass of $800,000$ inner products. While one could in principle store each $Ax_i$ to expedite this process (which is already prohibitive), this does not sidestep the need to compute the sum $\sum_{i=1}^{n} e^{\langle z, Ax_i\rangle}$ at each step of optimization, as is required to evaluate $\nabla\phi_{\mu_n}(z)$. In the following experiments, the ground truth images are hand drawn by the authors and hence are not present in the original dataset which is used to construct the prior.

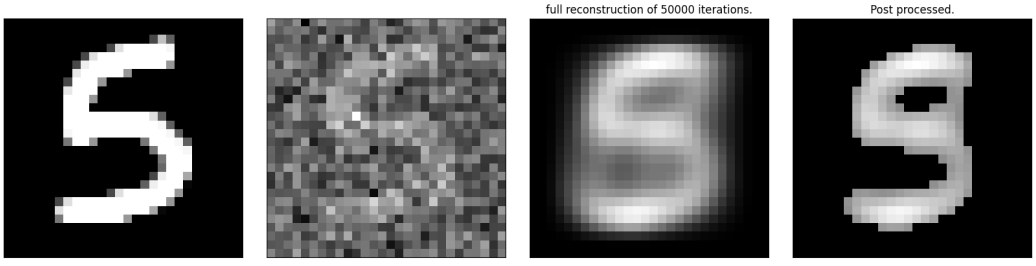
For our first experiment, Figure 1, we denoise and deblur an image generated as $b = Ax + \mathcal{N}(0, (0.1\|x\|)^2)$ where $A$ is a tridiagonal matrix with constant diagonal entries $0.25, 0.5, 0.25$ and $\mathcal{N}(0, (0.1\|x\|)^2)$ is the mean-zero normal distribution with variance $(0.1\|x\|)^2$. For this experiment, we set the following hyperparameters for the projected clipped SGD : $\alpha = 10$, $\beta_k = 10^{-4} + \frac{1}{2}(10^{-2} - 10^{-4})(1 + \cos(\frac{k+1}{60000}\pi))$, $m_k = 5$, $\gamma = 100$, $z_0 = 0$, and a maximum iteration count of 2000. Moreover in the final display we postprocess the output by setting pixels with intensity above 0.8 to 1 and those below 0.2 to 0.



(a) Ground truth (b) Observed image (c) CSGD output (2000 iterations) (d) Postprocessed

Figure 1: Result of solving the MEM problem using Projected Clipped SGD using a mini-batched stochastic gradient.

For Figure 2 the same blurring kernel $A$ is used along with the following hyperparameters for the projected clipped SGD : $\alpha = 2$, $\beta_k = 10^{-4} + \frac{1}{2}(10^{-2} - 10^{-4})(1 + \cos(\frac{k+1}{60000}\pi))$, $m_k = 5$, $\gamma = 10$, and $z_0 = 0$ with a maximum iteration count of 50000.




(a) Ground truth (b) Observed image (c) CSGD output (50000 iterations) (d) Postprocessed

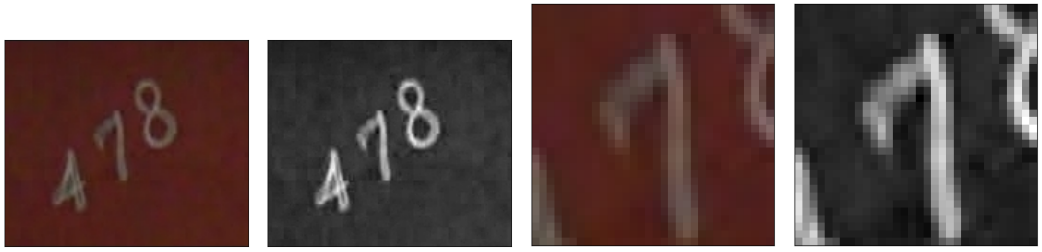
Figure 2: Result of solving the MEM problem using Projected Clipped SGD using a mini-batched stochastic gradient.

In both cases, the postprocessed output is close to the ground truth image. While the hyperparameters used above are tuned for each example, we underscore that, in testing, once the hyperparameters fall within a reasonable range further tuning does not significantly improve the quality of the recovered image except for the batchsize. Indeed, if the batchsize $m_k$ is set to be too small with the same amount of iterations this can result in significantly worse results.

We remark that in Figure 2, the postprocessed image could be interpreted as a 5 or the character g. This visual artifact can be explained by noting that in many handwriting styles these characters differ only by a small stroke and so, in the presence of large noise, it is difficult to discern if that stroke was present or not.

### 5.2 Street View House Numbers

We now consider the more realistic Street View House Numbers (SVHN) dataset (Netzer et al., 2011) which contains 630000 images of house numbers taken directly from Google street view. This dataset contains both original images of the numbers as well as cropped images of each digit of size $32 \times 32$ for a total of 1.8 million digits. This dataset presents additional real-world challenges, such as confounding features, color channels, differing aspect ratios, rotations, and so on. See Figure 3 for an example of the original image and the corresponding cropped digit.



(a) Example data (b) Example data (c) 32×32 cropped (d) 32×32 cropped

Figure 3: Examples of original house numbers and the corresponding cropped digits from the SVHN dataset.

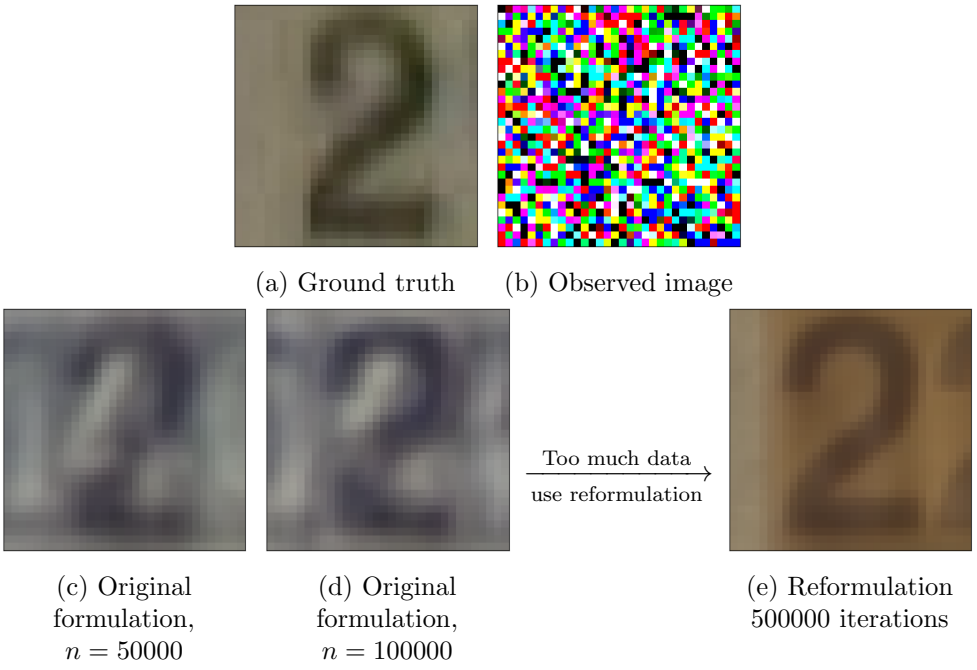
Throughout we consider only the cropped images so that all images are of the same size (i.e., $32 \times 32$). The corresponding problems are thus $d = 1024$ dimensional in black and white or $d = 3072$ dimensional in color. Furthermore, for each of the following experiments, the ground truth is taken as a point originally found in SVHN, but during recovery is removed from the dataset and hence not used to construct the prior.

We now demonstrate the advantage of clipped SGD over solving the original formulation of the problem. As before, we generate the observed image as $b = x + \mathcal{N}(0, (0.15\|x\|)^2)$ which corresponds to adding noise to the image only. To denoise the image we draw $n = 50000$ and $n = 100000$ samples and solve the corresponding full empirical MEM problem with the empirical prior via black-box BFGS applied to the original dual objective. It is critical to emphasize that even for $n = 500000$, it has become prohibitively expensive to compute the MEM solution for the full empirical problem, as simply storing the partial dataset already

requires 4GB of ram, with each gradient evaluation requiring a full data-pass, a problem that will only become worse as the ambient dimension increases. Furthermore, we solve the reformulated problem using the projected clipped method with the parameters $\alpha = 1$, $\beta_k = 10^{-4} + \frac{1}{2}(10^{-2} - 10^{-4})(1 + \cos(\frac{k+1}{60000}\pi))$, $m_k = 10$, $\gamma = 10$, and $z_0 = 0$ with a maximum iteration count of 500000. We also incorporate a momentum term which is seen to speed up convergence. Concretely, the update in Algorithm 1 for $g_k$ is replaced by

$$g_k \leftarrow 0.1 g_{k-1} + \frac{1}{m_k} \sum_{i=1}^{m_k} \ell'(x_{k_i}, z_k) \text{ for a measurable selection } \ell'(x_{k_i}, z_k) \in \partial_z \ell(x_{k_i}, z_k)$$

with the convention that $g_{-1} = 0$. This modification preserves some information about the direction of previous gradients before the clipping is performed. The results are compiled in Figure 4.



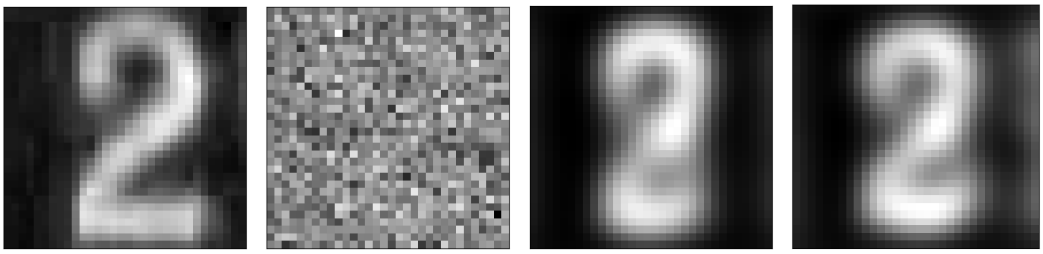

(a) Ground truth (b) Observed image

(c) Original formulation, $n = 50000$ (d) Original formulation, $n = 100000$ (e) Reformulation 500000 iterations

Figure 4: Comparison of solving the MEM problem with the prior $\hat{\mu}_n$ with $n = 50000$ and $n = 100000$ using full gradient-based methods and using 500000 iterations of the stochastic gradient method.

We underscore that, while the stochastic method is more scalable, a natural downside is that any given iteration may fail to decrease the objective value and it is prone to oscillate in some neighborhood of the minimum.

Next, we illustrate how the solution quality improves as the number of stochastic gradient steps increases. To this end, we plot the recovered solutions after a fixed number of iterations for a black and white image and a colored image. In both cases case, we set $\alpha = 2$, $\beta_k = 10^{-4} + \frac{1}{2}(10^{-2} - 10^{-4})(1 + \cos(\frac{k+1}{60000}\pi))$, $m_k = 10$, $\gamma = 10$, and $z_0 = 0$ with a maximum iteration count of 250000. The results are compiled in Figures 5 and 6.



(a) Ground truth (b) Observed image (c) 50000 iterations (d) 500000 iterations

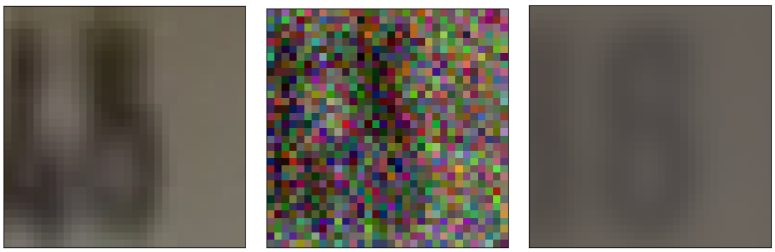
Figure 5: Result of applying the clipped SGD algorithm after 50000 and 500000 iterations. The observed image is generated via $b = x + \mathcal{N}(0, (0.2\|x\|)^2)$. After 50000 iterations the digit 2 is visible albeit with some blurring around the edges. After 500000 iterations the digit is clearer.

(a) Ground truth (b) Observed image (c) 1000 iterations (d) 100000 iterations (e) 250000 iterations

Figure 6: Result of applying the clipped SGD algorithm after 1000, 100000, and 2500000 iterations. The observed image is generated via $b = x + \mathcal{N}(0, (0.3\|x\|)^2)$. We note that the output at 100000 iterations is very similar to that at 250000 indicating that the convergence occurs relatively fast.

We note that the solutions obtained in Figure 6 have a discolored background; this can be understood by noting that the solution is obtained by taking a weighted combination of the datapoints so that the different background colors get averaged together. While color-valued postprocessing could potentially aid in the visual recovery, this introduces another layer of complexity beyond the scope of the MEM problem.

Finally, we note that this approach depends strongly on the dataset used to form the prior. Indeed, as illustrated in Figure 7, if the ground truth is not well-represented by elements of the dataset our method cannot accurately approximate it. In that figure, the ground truth is blurry and includes two digits due to poor cropping. Given that such images are outliers within the SVHN dataset it is unsurprising that the corresponding recovered image does not match the ground truth well. Here, this image is constructed with the hyperparameters $\alpha = 2$, $\beta_k = 10^{-4} + \frac{1}{2}(10^{-2} - 10^{-4})(1 + \cos(\frac{k+1}{60000}\pi))$, $m_k = 10$, $\gamma = 10$, and $z_0 = 0$.



(a) Ground truth (b) Observed image (c) Recovered image

Figure 7: Results of attempting to denoise an outlier within the SVHN dataset.

## 6 Conclusion

This work has advanced the statistical and computational study of the MEM problem. Namely, a parametric rate of convergence for empirical dual and primal solutions in expectation was derived which improves upon the $O(n^{-1/4})$ rate obtained in King-Roskamp et al. (2026). This statistical result was obtained by characterizing the stability properties of the primal and dual MEM solutions under perturbations of the prior distribution and follows from standard results in convex analysis and statistics. Furthermore, a connection between the MEM problem and conventional expected risk minimization was established. This connection was leveraged to propose and analyze new stochastic gradient methods for solving MEM problem which were shown to yield good results even in the presence of large amounts of noise.

Many promising research directions stem from this line of work. For instance, it is of interest to generalize the sample complexity results derived herein to the case of priors with unbounded support and to establish if these rates are minimax optimal. Establishing further statistical properties of these estimators such as central limit theorem or finite sample Gaussian approximation results à la Stein's method would also be useful for understanding the quality of approximations by empirical measures in the MEM framework and for calibrating statistical and optimization error.

The aforementioned connection to expected risk minimization also opens the door to a more engineering-based study of alternative stochastic optimization methods for the MEM problem with the aim of improving its empirical performance.

## Acknowledgments and Disclosure of Funding

RC and TH acknowledge the support of NSERC through its Discovery grants program.

### Disclosure on the use of Generative AI

In the preparation of this work Chat GPT-5.6 Sol was used for the purposes of copyediting and proofreading the near-complete manuscript. This model also suggested a correspondence of the domain and coercivity properties in Proposition 21, but the technical statement and corresponding proof of this result is entirely the original work of the authors. All original mathematical contributions are of the authors.

## References

C. Amblard, E. Lapalme, and J.-M. Lina. Biomagnetic source detection by maximum entropy and graphical models. *IEEE Trans. Biomed. Eng.*, 51(3):427–442, 2004.

L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probab. Theory Relat. Fields*, 97(1):113–150, 1993.

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

V.-E. Brunel. Asymptotics of constrained $M$-estimation under convexity, 2025. URL `https://arxiv.org/abs/2511.04612`.

J. V. Burke, T. Hoheisel, and Q. V. Nguyen. A study of convex convex-composite functions via infimal convolution with applications. *Math. Oper. Res.*, 46(4):1324–1348, 2021.

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. *Sci. Rep.*, 12(1), 2022. 2316.

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(2), 2013. E55969.

G. Cohen, S. Afshar, J. Tapson, and A. van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926, 2017. doi: 10.1109/IJCNN.2017.7966217.

D. Dacunha-Castelle and F. Gamboa. Maximum d'entropie et problème des moments. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 26(4):567–596, 1990.

R. M. Dudley. *Real analysis and probability*. CRC Press, 2018.

A. K. Fermin, J.-M. Loubes, and C. Ludena. Bayesian methods for a particular inverse problem seismic tomography. *Int. J. Tomogr. Stat*, 4(W06):1–19, 2006.

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

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(3):734–753, 2006.

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 Topogr.*, 29(1):162–181, 2016.

E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, 1957a.

E. T. Jaynes. Information theory and statistical mechanics. II. *Phys. Rev.*, 108:171–190, 1957b.

M. King-Roskamp, R. Choksi, and T. Hoheisel. Data-driven priors in the maximum entropy on the mean method for linear inverse problems. *SIAM J. Math. Data Sci.*, 8(1):108–140, 2026.

M. R. Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer, 2008.

S. Kullback and R. Leibler. On information and sufficiency. *Ann. Stat.*, 22(1):79–86, 1951.

G. Le Besnerais, J.-F. Bercher, and G. Demoment. A new look at entropy for solving linear inverse problems. *IEEE Trans. Inf. Theory*, 45(5):1565–1578, 1999.

V. Mai and M. Johansson. Stability and convergence of stochastic gradient clipping: Beyond Lipschitz continuity and smoothness. In *International Conference on Machine Learning*, pages 7325–7335. PMLR, 2021.

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 Probl.*, 13(1), 1997.

J. Navaza. On the maximum-entropy estimate of the electron density function. *Acta Crystallogr. Sect. A*, 41(3):232–244, 1985.

J. Navaza. The use of non-local constraints in maximum-entropy electron density reconstruction. *Acta Crystallogr. Sect. A*, 42(4):212–223, 1986.

Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 5, page 7, 2011.

E. Rietsch. The maximum entropy approach to inverse problems-spectral analysis of short data records and density structure of the earth. *J. Geophys.*, 42(1):489–506, 1976.

G. Rioux, R. Choksi, T. Hoheisel, P. Maréchal, and C. Scarvelis. The maximum entropy on the mean method for image deblurring. *Inverse Probl.*, 37(1):015011, 2020.

G. Rioux, C. Scarvelis, R. Choksi, T. Hoheisel, and P. Maréchal. Blind deblurring of barcodes via Kullback-Leibler divergence. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43 (1):77–88, 2021.

H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In J. S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233–257. Academic Press, 1971.

R. T. Rockafellar. *Convex Analysis*, volume 18. Princeton University Press, 1970.

R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*. Springer, 1998.

B. Urban. Retrieval of atmospheric thermodynamical parameters using satellite measurements with a maximum entropy method. *Inverse Probl.*, 12(5), 1996. 779.

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.*, 214(1):441–490, 2025.

A. W. van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.