Asymptotic Errors for Teacher-Student Convex Generalized Linear Models (Or: How to Prove Kabashima's Replica Formula)

Cedric Gerbelot, Alia Abbara, and Florent Krzakala

***Abstract*— There has been a recent surge of interest in the study of asymptotic reconstruction performance in various cases of generalized linear estimation problems in the teacher-student setting, especially for the case of i.i.d standard normal matrices. Here, we go beyond these matrices, and prove an analytical formula for the reconstruction performance of convex generalized linear models with rotationally-invariant data matrices with arbitrary bounded spectrum, rigorously confirming, under suitable assumptions, a conjecture originally derived using the replica method from statistical physics. The proof is achieved by leveraging on message passing algorithms and the statistical properties of their iterates, allowing to characterize the asymptotic empirical distribution of the estimator. For sufficiently strongly convex problems, we show that the two-layer vector approximate message passing algorithm (2-MLVAMP) converges, where the convergence analysis is done by checking the stability of an equivalent dynamical system, which gives the result for such problems. We then show that, under a concentration assumption, an analytical continuation may be carried out to extend the result to convex (non-strongly) problems. We illustrate our claim with numerical examples on mainstream learning methods such as sparse logistic regression and linear support vector classifiers, showing excellent agreement between moderate size simulation and the asymptotic prediction.**

***Index Terms*— Parametric statistics, estimation error, message passing, expectation-maximization algorithms, optimization, convergence, convex functions, linear matrix inequalities.**

Manuscript received 12 January 2021; revised 1 October 2022; accepted 24 October 2022. Date of publication 17 November 2022; date of current version 16 February 2023. This work was supported in part by the French Agence Nationale de la Recherche under Grant ANR-17-CE23-0023-01 PAIL and Grant ANR-19-P3IA-0001 PRAIRIE, in part by the Swiss National Science Foundation through Swiss National Science Foundation (SNSF) OperaGOST under Grant 200021_200390, and in part by the "Chaire de recherche sur les modèles et sciences des données," Fondation Capital Fund Management (CFM) pour la Recherche-Ecole Normale Supérieure (ENS). *(Corresponding author: Florent Krzakala.)*

Cedric Gerbelot was with the Laboratoire de Physique de l'Ecole normale supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, 75005 Paris, France. He is now with the Courant Institute of Mathematical Sciences, NYU, New York City, NY 10012 USA (e-mail: cedric.gerbelot@cims.nyu.edu).

Alia Abbara was with the Laboratoire de Physique de l'Ecole normale supérieure, ENS, 75005 Paris, France. She is now with the Laboratory of Computational Biology and Theoretical Biophysics, EPFL, 1015 Lausanne, Switzerland (e-mail: alia.abbara@epfl.ch).

Florent Krzakala is with the IdePhics Laboratory, EPFL, 1015 Lausanne, Switzerland (e-mail: florent.krzakala@epfl.ch).

Communicated by R. Venkataramanan, Associate Editor for Machine Learning and Statistics, Communications, Signal Processing and Source Coding.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2022.3222913.

Digital Object Identifier 10.1109/TIT.2022.3222913

I. Introduction

A. Background and Motivation

In the modern era of statistics and machine learning, data analysis often requires solving high-dimensional estimation problems with a very large number of parameters. Developing algorithms for this task and understanding their limitations has become a major challenge. In this paper, we consider this question in the framework of supervised learning under the teacher-student scenario: (i) the data is synthetic and labels are generated by a "teacher"rule and (ii) training is done with a convex Generalized Linear Model (GLM). Such problems are ubiquitous in machine learning, statistics, communications, and signal processing.

The study of asymptotic (i.e. large-dimensional) reconstruction performance of generalized linear estimation in the teacher-student setting has been the subject of a significant body of work over the past few decades [1], [2], [3], [4], [5], [6], [7], and is currently witnessing a renewal of interest, especially for the case of identically and independently distributed (i.i.d.) standard normal data matrices, see e.g. [8], [9], [10]. The aim of this paper is to provide a general analytical formula describing the reconstruction performance of such convex generalized linear models, but for a broader class of more adaptable matrices.

The problem is defined as follows: we aim at reconstructing a given i.i.d. weight vector $\mathbf{x}_0 \in \mathbb{R}^N$ from outputs $\mathbf{y} \in \mathbb{R}^M$ generated using a training set $(\mathbf{f}_\mu)_{\mu=1,\dots,M}$ and the "teacher" rule:

$$\mathbf{y} = \varphi(\mathbf{F}\mathbf{x}_0, \boldsymbol{\omega_0}) \tag{1}$$

where φ is a proper, closed, continuous function and $\boldsymbol{\omega_0} \sim \mathcal{N}(0, \delta_0 \mathrm{Id})$ is an i.i.d. noise vector. To go beyond the Gaussian i.i.d. case tackled in a majority of theoretical works, we shall allow matrices of arbitrary spectrum. We consider the data matrix $\mathbf{F} \in \mathbb{R}^{M\times N}$, obtained by concatenating the vectors of the training set, to be *rotationally invariant*: its singular value decomposition reads $\mathbf{F} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ where $\mathbf{U} \in \mathbb{R}^{M\times M}$,$\mathbf{V} \in \mathbb{R}^{N\times N}$ are uniformly sampled from the orthogonal groups $O(M)$ and $O(N)$ respectively. $\mathbf{D} \in \mathbb{R}^{M\times N}$ contains the singular values of $\mathbf{F}$ on its diagonal. Our analysis encompasses any singular value distribution with compact support. We place ourselves in the so-called high-dimensional regime, so that $M, N \to \infty$ while the ratio $\alpha \equiv M/N$ is kept finite. Our goal is to study the reconstruction performance of

the generalized linear estimation method:

$$\hat{\mathbf{x}} \in \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min} \{g(\mathbf{F}\mathbf{x}, \mathbf{y}) + f(\mathbf{x})\} \tag{2}$$

where g and f are proper, closed, convex and separable functions. This type of procedure is an instance of empirical risk minimization and is one of the building blocks of modern machine learning. It encompasses several mainstream methods such as logistic regression, the LASSO or linear support vector machines. More precisely, the quantities of interest representing the reconstruction performance are the mean squared error $E = \mathbb{E}\left[\frac{1}{N}\|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2\right]$ for regression problems, and the reconstruction angle $\theta_x = \arccos \frac{\mathbf{x}_0^T \hat{\mathbf{x}}}{\|\mathbf{x}_0\|_2 \|\hat{\mathbf{x}}\|_2}$ for classification problems.

B. Main Contributions

- We provide a set of equations characterizing the asymptotic statistical properties of the estimator defined by problem (2) with data generated by (1) in the asymptotic setup, for separable, convex losses and penalties (including for instance Logistic, Hinge, LASSO and Elastic net), for rotationally invariant sequences of matrices $\mathbf{F}$. For sufficiently strongly convex problems (in the sense of Lemma 3), our assumptions are classical with respect to earlier work. To extend the result to convex problems however, we require a concentration assumption that we discuss further in section III.
- By doing so, we give, under the aforementioned set of assumptions, a mathematically rigorous proof, of a replica formula obtained heuristically through statistical physics for this problem, notably by Kabashima [11]. This is a significant step beyond the setting of most rigorous work on replica results, which assume matrices to be i.i.d. random Gaussian ones.
- Our proof method builds on a detailed mapping between alternating directions descent methods [12] from convex optimization and a set of algorithms called multi-layer vector approximate message-passing algorithms [13], [14]. This enables us to use convergence results from convex analysis and dynamical systems to study the trajectories of vector approximate message-passing algorithms.
- Beyond the high-dimensional result on the estimator defined by the GLM, our convergence analysis provides a generic condition for the convergence of 2-layer MLVAMP, regardless of the randomness of the design matrix and of the dimensions of the problem, for sufficiently strongly convex problems.

C. Related Work

The simplest case of the present question, when both f and g are quadratic functions, can be mapped to a random matrix theory problem and solved rigorously, as in e.g. [9]. Handling non-linearity is, however, more challenging. A long history of research tackles this difficulty in the high-dimensional limit, especially in the statistical physics literature where this setup is common. The usual analytical approach in statistical physics of learning [1], [2], [3] is a heuristic, non-rigorous but very adaptable technique called the replica method [15], [16]. In particular, it has been applied on many variations of the present problem, and laid the foundation of a large number of deep, non-trivial results in machine learning, signal processing and statistics, e.g. [17], [18], [19], [20], [21], [22], [23], [24], [25]. Among them, a generic formula for the present problem has been conjectured by Y. Kabashima, providing sharp asymptotics for the reconstruction performance of the signal $\mathbf{x_0}$ [11].

Proving the validity of a replica prediction is a difficult task altogether. There has been recent progress in the particular case of Gaussian data, where the matrix $\mathbf{F}$ is made of i.i.d. standard Gaussian coefficients. In this case, the asymptotic performance of the LASSO was rigorously derived in [26], and the existence of the logistic estimator discussed in [8]. A set of papers managed to extend this study to a large set of convex losses g, using the so-called Gordon comparison theorem [27]. We broaden those results here by proving the Kabashima formula, valid for the set of rotationally invariant matrices introduced above and any convex, separable loss g and sufficiently strongly convex regularizer f under classical conditions. We extend this result to any convex, separable g and f under stronger assumptions.

Our proof strategy is based on the use of approximate-message-passing [28], [29], as pioneered in [4], and is similar to a recent work [30] on a simpler setting. This family of algorithms is a statistical physics-inspired variant of belief propagation [31], [32], [33] where local beliefs are approximated by Gaussian distributions. A key feature of these algorithms is the existence of the state evolution equations, a scalar equivalent model which allows to track the asymptotic statistical properties of the iterates at every time step. A series of groundbreaking papers initiated with [26] proved that these equations are exact in the large system limit, and extended the method to treat nonlinear problems [29] and handle rotationally invariant matrices [34], [35]. We shall use a variant of these algorithms called multi-layer vector approximate message-passing (MLVAMP) [14], [36]. The key technical point in our approach is an analysis of the convergence of MLVAMP. This is achieved by phrasing the algorithm as a dynamical system, and then determining sufficient conditions for convergence with linear rate. Our analysis guarantees converging trajectories above a threshold value of the strong convexity parameter of the problem, which is sufficient to complete the proof in that region. We use an analytic continuation to extend the result to convex problems, at the cost of an additional condition discussed after stating our main set of assumption.

II. Background on MLVAMP

In this section, we present background on the multi-layer vector approximate message-passing algorithm developed in [36]. In doing so, we will introduce the key quantities involved in our main theorem. MLVAMP was initially designed as a probabilistic inference algorithm in multilayer architectures. Here, we only focus on the 2-layer version for inference in GLMs, and use the notations of [35]. The algorithm can be derived in several ways, notably from expectation-consistent variational inference frameworks such as expectation propagation [37], where the target posterior distribution is approximated by a simpler one with moment

matching constraints. In the maximum a posteriori setting (MAP), the frequentist optimization framework is recovered, with additional parameter prescriptions due to the probabilistic models, as we will see below. The derivation of the algorithm is, however, not our point of interest. We focus on providing a self-contained interpretation from the convex optimization point of view, in particular in terms of variable splitting.

A. Link With Variable Splitting and Proximal Descent

A common procedure to tackle nonlinear optimization problems involving several functions is variable splitting, so that each non-linearity may be treated independently. Augmenting the Lagrangian with a square penalty on the slack variable equality constraint leads to the family of alternating direction methods of multipliers (ADMM) [12], where the objective is iteratively minimized in the direction of each initial variable and slack variable. The descent steps then take the form of proximal operators of the non-linearities. For example, on problem (2), adding a slack variable $\mathbf{z} = \mathbf{F}\mathbf{x}$ would lead to the augmented Lagrangian:

$$g(\mathbf{z}, \mathbf{y}) + f(\mathbf{x}) + \theta^T(\mathbf{z} - \mathbf{F}\mathbf{x}) + \frac{\alpha}{2}\|\mathbf{z} - \mathbf{F}\mathbf{x}\|_2^2 \qquad (3)$$

where $\alpha > 0$ is a free parameter that can enforce strong convexity of the objective if large enough and θ is a Lagrange multiplier. Updating $\mathbf{x}$ from an update on $\mathbf{z}$ amounts to a linear estimation problem, which can be solved by least squares. This is implemented, for example, in linearized ADMM [12], where the proximal descent steps are coupled to least-square ones.

MLVAMP solves problem (2) by introducing the same splitting as in (3) with an additional trivial splitting for each variable: $\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}_1, \mathbf{z}_2$ such that $\mathbf{x}_1 = \mathbf{x}_2,\ \mathbf{z}_1 = \mathbf{F}\mathbf{x}_1,\ \mathbf{z}_2 = \mathbf{F}\mathbf{x}_2$. In the convex optimization framework, parameters like gradient step sizes, or proximal parameters need to be chosen. In the expectation propagation framework, they are prescribed by expectation-consistency constraints, which leads to additional steps in the algorithm. MLVAMP thus consists in four descent steps on $\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}_1, \mathbf{z}_2$, and the updates on the parameters of the functions corresponding to those descent steps. This is shown in the MLVAMP iterations (see (1) further), where $\mathbf{x}_1, \mathbf{z}_1$ are updated using the proximal operators of the loss and regularizer, while $\mathbf{z}_2$ and $\mathbf{x}_2$ are obtained through least-squares. As mentioned above, the parameters of proximal operators (or denoisers in the signal processing literature) and least-squares are set by probabilistic inference rules (here moment-matching of marginal distributions). It is shown in [38] that, in the MAP setting, these updates amount to adapting the parameters to the local curvature of the cost function.

B. 2-Layer MLVAMP and Its State Evolution

We lay out the full iterations of the MLVAMP algorithm from [36] applied to a 2-layer network in Algorithm 1. For a given operator $T : \mathcal{X} \to \mathbb{R}^d$ where d is M or N in our setting, the brackets $\langle T(\mathbf{x})\rangle = \frac{1}{d}\sum_{i=1}^d T(\mathbf{x})_i$ denote element-wise averaging operations. For a given matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$, the brackets amount to $\langle \mathbf{M}\rangle = \frac{1}{d}\mathbf{Tr}(\mathbf{M})$. For a given function,

Algorithm 1 2-Layer MLVAMP

Require: Initialize $\mathbf{h}_{1x}^{(0)}, \mathbf{h}_{2z}^{(0)}, \hat{Q}_{1x}^{(0)}, \hat{Q}_{2z}^{(0)}$, number of iterations T.

for t=0,1…,T **do**

// Denoising $\mathbf{x}$

$\hat{\mathbf{x}}_1^{(t)} = g_{1x}(\mathbf{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)})$
$\chi_{1x}^{(t)} = \left\langle \partial_{\mathbf{h}_{1x}^{(t)}} g_{1x}(\ldots)\right\rangle / \hat{Q}_{1x}^{(t)}$
$\hat{Q}_{2x}^{(t)} = 1/\chi_{1x}^{(t)} - \hat{Q}_{1x}^{(t)}$
$\mathbf{h}_{2x}^{(t)} = (\hat{\mathbf{x}}_1^{(t)}/\chi_{1x}^{(t)} - \hat{Q}_{1x}^{(t)}\mathbf{h}_{1x}^{(t)})/\hat{Q}_{2x}^{(t)}$

// LMMSE estimation of $\mathbf{z}$

$\hat{\mathbf{z}}_2^{(t)} = g_{2z}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t)})$
$\chi_{2z}^{(t)} = \left\langle \partial_{\mathbf{h}_{2z}^{(t)}} g_{2z}(\ldots)\right\rangle / \hat{Q}_{2z}^{(t)}$
$\hat{Q}_{1z}^{(t)} = 1/\chi_{2z}^{(t)} - \hat{Q}_{2z}^{(t)}$
$\mathbf{h}_{1z}^{(t)} = (\hat{\mathbf{z}}_2^{(t)}/\chi_{2z}^{(t)} - \hat{Q}_{2z}^{(t)}\mathbf{h}_{2z}^{(t)})/\hat{Q}_{1z}^{(t)}$

// Denoising $\mathbf{z}$

$\hat{\mathbf{z}}_1^{(t)} = g_{1z}(\mathbf{h}_{1z}^{(t)}, \hat{Q}_{1z}^{(t)})$,
$\chi_{1z}^{(t)} = \left\langle \partial_{\mathbf{h}_{1z}^{(t)}} g_{1z}(\ldots)\right\rangle / \hat{Q}_{1z}^{(t)}$
$\hat{Q}_{2z}^{(t+1)} = 1/\chi_{1z}^{(t)} - \hat{Q}_{1z}^{(t)}$
$\mathbf{h}_{2z}^{(t+1)} = (\hat{\mathbf{z}}_1^{(t)}/\chi_{1z}^{(t)} - \hat{Q}_{1z}^{(t)}\mathbf{h}_{1z}^{(t)})/\hat{Q}_{2z}^{(t+1)}$

// LMMSE estimation of $\mathbf{x}$

$\hat{\mathbf{x}}_2^{(t+1)} = g_{2x}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t+1)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t+1)})$
$\chi_{2x}^{(t+1)} = \left\langle \partial_{\mathbf{h}_{2x}^{(t)}} g_{2x}(\ldots)\right\rangle / \hat{Q}_{2x}^{(t)}$
$\hat{Q}_{1x}^{(t+1)} = 1/\chi_{2x}^{(t+1)} - \hat{Q}_{2x}^{(t)}$
$\mathbf{h}_{1x}^{(t+1)} = (\hat{\mathbf{x}}_2^{(t+1)}/\chi_{2x}^{(t+1)} - \hat{Q}_{2x}^{(t)}\mathbf{h}_{2x}^{(t)})/\hat{Q}_{1x}^{(t+1)}$

end for

return $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2$

for example g_{1x}, we use the shorthand $g_{1x}(\ldots)$ when the arguments have been made clear in a line above and are left unchanged. The denoising functions g_{1x} and g_{1z} can be written as proximal operators in the MAP setting:

$$\begin{aligned} g_{1x}(\mathbf{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)}) &= \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min}\left\{ f(\mathbf{x}) + \frac{\hat{Q}_{1x}^{(t)}}{2}\|\mathbf{x} - \mathbf{h}_{1x}^{(t)}\|_2^2 \right\} \\ &= \mathrm{Prox}_{f/\hat{Q}_{1x}^{(t)}}(\mathbf{h}_{1x}^{(t)}) \qquad (4) \\ g_{1z}(\mathbf{h}_{1z}^{(t)}, \hat{Q}_{1z}^{(t)}) &= \underset{\mathbf{z}\in\mathbb{R}^M}{\arg\min}\left\{ g(\mathbf{y}, \mathbf{z}) + \frac{\hat{Q}_{1z}^{(t)}}{2}\|\mathbf{z} - \mathbf{h}_{1z}^{(t)}\|_2^2 \right\} \\ &= \mathrm{Prox}_{g(\cdot,\mathbf{y})/\hat{Q}_{1z}^{(t)}}(\mathbf{h}_{1z}^{(t)}). \qquad (5) \end{aligned}$$

The LMMSE denoisers g_{2z} and g_{2x} in the MAP setting read (see [14]):

$$g_{2z}(\ldots) = \mathbf{F}\mathbf{M}_1^{(t)}(\hat{Q}_{2x}^{(t)}\mathbf{h}_{2x}^{(t)} + \hat{Q}_{2z}^{(t)}\mathbf{F}^T\mathbf{h}_{2z}^{(t)}) \qquad (6)$$

$$g_{2x}(\ldots) = \mathbf{M}_2^{(t)}(\hat{Q}_{2x}^{(t)}\mathbf{h}_{2x}^{(t)} + \hat{Q}_{2z}^{(t+1)}\mathbf{F}^T\mathbf{h}_{2z}^{(t+1)}) \qquad (7)$$

where we defined the matrices $\mathbf{M}_1^{(t)} = (\hat{Q}_{2z}^{(t)}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}^{(t)}\mathrm{Id})^{-1}$, and $\mathbf{M}_2^{(t)} = (\hat{Q}_{2z}^{(t+1)}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}^{(t)}\mathrm{Id})^{-1}$. As mentioned in the previous section, MLVAMP returns at each iteration two sets of estimators $(\hat{\mathbf{x}}_1^{(t)}, \hat{\mathbf{x}}_2^{(t)})$ and $(\hat{\mathbf{z}}_1^{(t)}, \hat{\mathbf{z}}_2^{(t)})$ which respectively

aim at reconstructing the minimizer $\hat{\mathbf{x}}$ and $\hat{\mathbf{z}} = \mathbf{F}\hat{\mathbf{x}}$. At the fixed point, we have $\hat{\mathbf{x}}_1^{(t)} = \hat{\mathbf{x}}_2^{(t)}$ and $\hat{\mathbf{z}}_1^{(t)} = \hat{\mathbf{z}}_2^{(t)}$, as proven in [39]. The intermediate vectors $\mathbf{h}_{1x}^{(t)}$, $\mathbf{h}_{2x}^{(t)}$, $\mathbf{h}_{1z}^{(t)}$ and $\mathbf{h}_{2z}^{(t)}$ have the key feature that they behave asymptotically as Gaussian centered around $\mathbf{x}_0$ and $\mathbf{z}_0 = \mathbf{F}\mathbf{x}_0$, under the set of assumptions given in appendix E-B. More precisely, at each iteration, they converge empirically with second order moment (PL2) towards Gaussian variables:

$$\begin{aligned}
\lim_{M,N\to\infty} \hat{Q}_{1x}^{(t)}\mathbf{h}_{1x}^{(t)} - \hat{m}_{1x}^{(t)}\mathbf{x_0} &\overset{PL(2)}{=} \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_{1x}^{(t)} \\
\lim_{M,N\to\infty} \mathbf{V}^T(\hat{Q}_{2x}^{(t)}\mathbf{h}_{2x}^{(t)} - \hat{m}_{2x}^{(t)}\mathbf{x_0}) &\overset{PL(2)}{=} \sqrt{\hat{\chi}_{2x}^{(t)}}\xi_{2x}^{(t)} \\
\lim_{M,N\to\infty} \mathbf{U}^T(\hat{Q}_{1z}^{(t)}\mathbf{h}_{1z}^{(t)} - \hat{m}_{1z}^{(t)}\mathbf{z_0}) &\overset{PL(2)}{=} \sqrt{\hat{\chi}_{1z}^{(t)}}\xi_{1z}^{(t)} \\
\lim_{M,N\to\infty} \hat{Q}_{2z}^{(t)}\mathbf{h}_{2z}^{(t)} - \hat{m}_{2z}^{(t)}\mathbf{z_0} &\overset{PL(2)}{=} \sqrt{\hat{\chi}_{2z}^{(t)}}\xi_{2z}^{(t)} \qquad (8)
\end{aligned}$$

where $\xi_{1x}^{(t)}, \xi_{2x}^{(t)}, \xi_{1z}^{(t)}, \xi_{2z}^{(t)}$ are i.i.d standard normal random variables independent of all other quantities. The definition of PL(2) convergence is reminded in Appendix A, and we use the notation $\overset{PL(2)}{=}$ following [34], [36]. We can roughly say that the $\hat{Q}, \hat{m}, \hat{\chi}$'s parameters characterize the distributions of the $\mathbf{h}$'s. Using the representation (8) in the iterations of MLVAMP results in a scalar recursion that tracks the evolution of the parameters of the aforementioned Gaussian distributions. This recursion provides the so-called state evolution equations. The existence of state evolution equations is the reason why we use 2-layer MLVAMP in our proof. Indeed, they allow the construction of iterate paths that lead to the solution of problem (1), while knowing their statistical properties.

III. Main Result

Our main result characterizes the asymptotic empirical distribution of the estimator $\hat{\mathbf{x}}$ defined in (2) with data generated by (1), and of $\hat{\mathbf{z}} = \mathbf{F}\hat{\mathbf{x}}$. We start by stating the necessary assumptions.

Assumption 1:

(a) the functions f and g are proper, closed, convex and separable functions.

(b) the cost function $g(\mathbf{F}.,\mathbf{y}) + f(.)$ is coercive, i.e. $\lim_{\|\mathbf{x}\|\to\infty} g(\mathbf{F}\mathbf{x},\mathbf{y}) + f(\mathbf{x}) = +\infty$.

(c) there exists a finite constant B_1 such that $\frac{1}{N}\|\hat{\mathbf{x}}\|_2^2 \leqslant B_1$ almost surely as $N \to \infty$. We also assume that, for any pseudo-Lipschitz function of order 2, if there exists a finite constant B_2 such that $\forall N \in \mathbb{N}, \frac{1}{N}\sum_{i=1}^N \phi(\hat{x}_i) \leqslant B_2$, then the limit $\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^N \phi(\hat{x}_i)$ exists.

(d) for any $\mathbf{x} \in \mathrm{dom}(f)$ and any $\mathbf{x}' \in \partial f(\mathbf{x})$, there exists a constant C such that $\|\mathbf{x}'\|_2 \leqslant C(1 + \|\mathbf{x}\|_2)$. The same holds for g on its domain.

(e) there exist sequences of real analytic functions g_ϵ, f_ϵ such that for any x, $\lim_{\epsilon\to 0} g_\epsilon(x) = g(x)$, $\lim_{\epsilon\to 0} f_\epsilon(x) = f(x)$, and for all $\epsilon > 0$, g_ϵ'' and f_ϵ'' belong to the Schwartz space.

(f) the empirical distributions of the underlying truth $\mathbf{x}_0$, eigenvalues of $\mathbf{F}^T\mathbf{F}$, and noise vector w_0, respectively converge empirically with second order moments, as defined in appendix A, to independent scalar random variables x_0, w_0, λ with distributions p_{x_0}, p_λ, p_{w_0}. We assume that the distribution p_λ is not all-zero and has compact support.

(g) the design matrix $\mathbf{F} = \mathbf{U}\mathbf{D}\mathbf{V}^\top \in \mathbb{R}^{M\times N}$ is rotationally invariant, as defined in the introduction, where the elements of the Haar distributed matrices $\mathbf{U}, \mathbf{V}$ are independent of the elements of the ground truth vector $\mathbf{x}_0$, noise $\boldsymbol{\omega}_0$ and elements of $\mathbf{D}$.

(h) the solution to the set of fixed point equations (9) exists and is unique, for any convex g and f verifying the assumptions above

(i) finally assume that $M, N \to \infty$ with fixed ratio $\alpha = M/N$.

The coercivity assumption (b) ensures that the minimization problem Eq.(2) is feasible and that the estimator exists. Most machine learning cost functions verify this assumption, including any convex loss which is bounded below and regularized with a coercive term such as the ℓ_1 or ℓ_2 norm, see [40] Corollary 11.15. Non-coercive problems include unregularized logistic regression and unregularized, underspecified least-squares for example. The scaling assumptions (d) are required for the state evolution equations of the MLVAMP iteration corresponding to the optimization problem Eq.(2) to hold, as discussed in appendix E-B. Such conditions are often encountered in high dimensional analysis of M-estimators, see, e.g. [27], and are verified by the setups proposed in the experiments section. The convergence of averaged sumes of PL2 observables in assumption (c) and the analytic approximation in assumption (e) are required for our analytic continuation to hold, and we show that any combination of hinge, logistic and square loss with ℓ_1 or ℓ_2 regularization verifies the latter in Appendix H, subsection H-F. We show in Lemma 4 that, for sufficiently strongly convex problems, these two assumptions are not required. The concentration assumption we require has been proven to hold for a number of convex problems with Gaussian random design regardless of the strong convexity of the problem (see the related work section), and we believe rotationally invariant matrices do not change this behaviour. However, since we are unable to prove it below the threshold value of the strong convexity parameter, it remains an assumption. Additional detail on the notion of empirical convergence is given in appendix A. This analysis framework is mainly due to [26] and is related to convergence in Wasserstein metric as pointed out in [25]. We are now ready to state our main theorem.

Theorem 1 (Fixed Point Equations): Under assumption 1, consider the ground-truth $\mathbf{x_0}$ and let $\mathbf{z_0} = \mathbf{F}\mathbf{x_0}$, $\rho_x \equiv \|\mathbf{x_0}\|_2^2/N$ and $\rho_z \equiv \|\mathbf{z_0}\|_2^2/M$. For a strictly convex instance of problem (2), let $\hat{\mathbf{x}}$ be its unique solution. For a convex (non-strictly) instance of problem (2), let $\hat{\mathbf{x}}$ be its unique least ℓ_2 norm solution. Then let $\hat{\mathbf{z}} = \mathbf{F}\hat{\mathbf{x}}$. Then, for any real analytic, pseudo-Lipschitz function of order 2 ϕ whose second derivative belongs to the Schwartz space, the following holds:

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^N \phi(x_{0,i}, \hat{x}_i) \overset{a.s.}{=} \mathbb{E}[\phi(x_0, \mathrm{Prox}_{f/\hat{Q}_{1x}^{(*)}}(H_x))]$$

$$\lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M}\phi(z_{0,i},\hat{z}_i)\overset{a.s.}{=}\mathbb{E}[\phi(z_0,\text{Prox}_{f/\hat{Q}_{1z}^{(*)}}(H_z))]$$

where $H_x=\frac{\hat{m}_{1x}^* x_0+\sqrt{\hat{\chi}_{1x}^*}\xi_{1x}}{\hat{Q}_{1x}^*}$, $H_z=\frac{\hat{m}_{1z}^* z_0+\sqrt{\hat{\chi}_{1z}^*}\xi_{1z}}{\hat{Q}_{1z}^*}$ and expectations are taken with respect to the random variables $x_0\sim p_{x_0}$, $z_0\sim\mathcal{N}(0,\sqrt{\rho_z})$, $\xi_{1x},\xi_{1z}\sim\mathcal{N}(0,1)$. The parameters $\hat{Q}_{1x}^*,\hat{Q}_{1z}^*,\hat{m}_{1x}^*,\hat{m}_{1z}^*,\hat{\chi}_{1x}^*,\hat{\chi}_{1z}^*$ are determined by the fixed point of the system:

$$\begin{aligned}
\hat{Q}_{2x}&=\hat{Q}_{1x}(\mathbb{E}\left[\eta'_{f/\hat{Q}_{1x}}(H_x)\right]^{-1}-1)\\
\hat{Q}_{2z}&=\hat{Q}_{1z}(\mathbb{E}\left[\eta'_{g(.,y)/\hat{Q}_{1z}}(H_z)\right]^{-1}-1)\\
\hat{m}_{2x}&=\frac{\mathbb{E}\left[x_0\eta_{f/\hat{Q}_{1x}}(H_x)\right]}{\rho_x\chi_x}-\hat{m}_{1x}\\
\hat{m}_{2z}&=\frac{\mathbb{E}\left[z_0\eta_{g(.,y)/\hat{Q}_{1z}}(H_z)\right]}{\rho_z\chi_z}-\hat{m}_{1z}\\
\hat{\chi}_{2x}&=\frac{\mathbb{E}\left[\eta^2_{f/\hat{Q}_{1x}}(H_x)\right]}{\chi_x^2}-\rho_x(\hat{m}_{1x}+\hat{m}_{2x})^2-\hat{\chi}_{1x}\\
\hat{\chi}_{2z}&=\frac{\mathbb{E}\left[\eta^2_{g(.,y)/\hat{Q}_{1z}}(H_z)\right]}{\chi_z^2}-\rho_z(\hat{m}_{1z}+\hat{m}_{2z})^2-\hat{\chi}_{1z}\\
\hat{Q}_{1x}&=\mathbb{E}\left[\frac{1}{\hat{Q}_{2x}+\lambda\hat{Q}_{2z}}\right]^{-1}-\hat{Q}_{2x}\\
\hat{Q}_{1z}&=\alpha\mathbb{E}\left[\frac{\lambda}{\hat{Q}_{2x}+\lambda\hat{Q}_{2z}}\right]^{-1}-\hat{Q}_{2z}\\
\hat{m}_{1x}&=\frac{1}{\chi_x}\mathbb{E}\left[\frac{\hat{m}_{2x}+\lambda\hat{m}_{2z}}{\hat{Q}_{2x}+\lambda\hat{Q}_{2z}}\right]-\hat{m}_{2x}\\
\hat{m}_{1z}&=\frac{\rho_x}{\alpha\chi_z\rho_z}\mathbb{E}\left[\frac{\lambda(\hat{m}_{2x}+\lambda\hat{m}_{2z})}{\hat{Q}_{2x}+\lambda\hat{Q}_{2z}}\right]-\hat{m}_{2z}\\
\hat{\chi}_{1x}&=\frac{1}{\chi_x^2}\mathbb{E}\left[\frac{\hat{\chi}_{2x}+\lambda\hat{\chi}_{2z}+\rho_x(\hat{m}_{2x}+\lambda\hat{m}_{2z})^2}{(\hat{Q}_{2x}+\lambda\hat{Q}_{2z})^2}\right]\\
&\qquad-\rho_x(\hat{m}_{1x}+\hat{m}_{2x})^2-\hat{\chi}_{2x}\\
\hat{\chi}_{1z}&=\frac{1}{\alpha\chi_z^2}\mathbb{E}\left[\frac{\lambda(\hat{\chi}_{2x}+\lambda\hat{\chi}_{2z}+\rho_x(\hat{m}_{2x}+\lambda\hat{m}_{2z})^2)}{(\hat{Q}_{2x}+\lambda\hat{Q}_{2z})^2}\right]\\
&\qquad-\rho_z(\hat{m}_{1z}+\hat{m}_{2z})^2-\hat{\chi}_{2z},
\end{aligned}\tag{9}$$

where $\chi_x=(\hat{Q}_{1x}+\hat{Q}_{2x})^{-1}$, $\chi_z=(\hat{Q}_{1z}+\hat{Q}_{2z})^{-1}$, and expectations are taken with respect to the random variables $x_0\sim p_{x_0}$, $z_0\sim\mathcal{N}(0,\sqrt{\rho_z})$, $y\sim\varphi(z_0,\omega_0)$, $\xi_{1x},\xi_{1z}\sim\mathcal{N}(0,1)$, and eigenvalues $\lambda\sim p_\lambda$. η is a shorthand for the scalar proximal operator:

$$\eta_{\gamma f}(z)=\underset{x\in\mathcal{X}}{\arg\min}\left\{\gamma f(x)+\frac{1}{2}(x-z)^2\right\}.$$

The set of fixed point equations from Theorem 1 naturally stems from the "replica-symmetric" free energy commonly used in the statistical physics community [15], [16]. The free energy depends on a set of parameters, and extremizing it with respect to all parameters, i.e. writing the zero gradient condition for each parameter, provides the set of equations (9). We state this correspondence in the following corollary to Theorem 1:

Corollary 1 (The Kabashima formula): The fixed point equations from theorem 1 can equivalently be rewritten as the solution of the extreme value problem (10), shown at the bottom of the next page, defined by the replica free energy from [35].

β is a parameter that corresponds in the physics approach to an inverse temperature. In the $\beta\to\infty$ limit (the so-called zero temperature limit), the integrals defining ϕ_x and ϕ_z concentrate on their extremal value. Note that they are closely related to the Moreau envelopes $\mathcal{M}$ [40], [41] of f and g, which represent a smoothed form of the objective function with the same minimizers:

$$\begin{aligned}
&\phi_x(\hat{m}_{1x},\hat{Q}_{1x},\hat{\chi}_{1x};x_0,\xi_{1x})=\frac{\hat{Q}_{1x}}{2}H_x^2-\mathcal{M}_{\frac{f}{\hat{Q}_{1x}}}(H_x)\\
&\text{where }\forall\,\gamma\geqslant 0,\ \mathcal{M}_{\gamma f}(z)=\inf_x\left\{f(x)+\frac{1}{2\gamma}\|x-z\|_2^2\right\}.
\end{aligned}$$

We provide details on this correspondence in appendix C. In the zero-temperature limit we consider, it is possible to have more precise information on the geometry of the cost function defining the optimization problem in Corollary 1. Indeed, it is composed of functions whose convexity or concavity are staightforward to establish: linear terms, inverses, logarithms, squares and expectation of Moreau envelopes. The convexity of the latter is well documented in [27]. First, note that the parameters $\chi_x,\chi_z,\hat{\chi}_{1x},\hat{\chi}_{2x},\hat{\chi}_{1z},\hat{\chi}_{2z},q_x,q_z,\hat{Q}_{1x},\hat{Q}_{2x},\hat{Q}_{1z},\hat{Q}_{2z}$ are positive so we may restrict their feasibility set to $\mathbb{R}^+$, while $m_x,m_z,\hat{m}_{1x},\hat{m}_{1z},\hat{m}_{2x},\hat{m}_{2z}$ can take any value in $\mathbb{R}$. Then, $q_x^*=\frac{1}{N}\|\hat{\mathbf{x}}\|^2$ and $m_x^*=\frac{1}{N}\mathbf{x}_0^\top\hat{\mathbf{x}}$. The Cauchy-Schwarz inequality thus gives

$$q_x^*\geqslant\frac{(m_x^*)^2}{\rho_x}.$$

Similarly with $\hat{\mathbf{z}}$,

$$q_z^*\geqslant\frac{(m_z^*)^2}{\rho_z}.$$

We may thus restrict the feasibility sets of q_x,q_z,m_x,m_z such that they verify these inequalities. In these regions, the function g_s is convex in χ_x,χ_z, linear in q_x,q_z and concave in m_x,m_z. The terms involving $q_x,q_z,m_x,m_z,\chi_x,\chi_z$ in g_G and g_F are all linear. Moving to g_g, the cost function defining it is convex in $\hat{Q}_{2x},\hat{Q}_{2z}$ (negative logarithm and inverse function on $\mathbb{R}^+$), linear in $\hat{\chi}_{2x},\hat{\chi}_{2z}$ and convex in $\hat{m}_{2x},\hat{m}_{2z}$. Regarding g_F, all terms are linear except for the replica potentials. Using Moreau's identity, we may write

$$\phi_x(\hat{m}_{1x},\hat{Q}_{1x},\hat{\chi}_{1x};x_0,\xi_{1x})=\mathcal{M}_{\hat{Q}_{1x}f^*}\left(\hat{m}_{1x}x_0+\sqrt{\hat{\chi}_{1x}}\xi_{1x}\right)$$

where f^* is the conjugate of f. Using the properties summarized in [27], the cost function defining g_F is convex in $\hat{m}_{1x},\hat{m}_{1z},\hat{Q}_{1x},\hat{Q}_{1z}$. The convexity with respect to χ_{1x},χ_{1z} is harder to characterize due to the composition of the Moreau envelope with the square root, and should be studied locally for more information. The extremization may then be rewritten as a maximization over the variables in which the cost function is concave and minimization over the variables in which the cost function is convex. Note that this does not give information

on the uniqueness of the solution, which would require joint strict convexity and strict concavity. As immediate corollaries to Theorem 1, we can determine the asymptotic errors of the GLM and the optimal value of the loss function. To characterize the asymptotic reconstruction errors and angles, we can define the norms of the estimators and their overlaps with the ground-truth vectors as the limits

$$m_x^* \equiv \lim_{N\to\infty} \frac{\hat{\mathbf{x}}^T \mathbf{x}_0}{N} \qquad m_z^* \equiv \lim_{M\to\infty} \frac{\hat{\mathbf{z}}^T \mathbf{z}_0}{M}$$
$$q_x^* \equiv \lim_{N\to\infty} \frac{\|\hat{\mathbf{x}}\|_2^2}{N} \qquad q_z^* \equiv \lim_{N\to\infty} \frac{\|\hat{\mathbf{z}}\|_2^2}{M}.$$

We then have:

Corollary 2: Under the set of Assumptions 1, the squared norms m_x^*, m_z^* of estimator $\hat{\mathbf{x}}$ defined by (2) and $\hat{\mathbf{z}} = \mathbf{F}\hat{\mathbf{x}}$, and their overlaps q_x^*, q_z^* with ground-truth vectors are almost surely given by:

$$m_x^* = \mathbb{E}\left[x_0 \eta_{\frac{f}{\hat{Q}_{1x}^*}}(H_x)\right],\ q_x^* = \mathbb{E}\left[\eta^2_{\frac{f}{\hat{Q}_{1x}^*}}(H_x)\right]$$
$$m_z^* = \mathbb{E}\left[z_0 \eta_{\frac{g(.,y)}{\hat{Q}_{1z}^*}}(H_z)\right],\ q_z^* = \mathbb{E}\left[\eta^2_{\frac{g(.,y)}{\hat{Q}_{1z}^*}}(H_z)\right]$$

with H_x and H_z defined as in Theorem 1.

With the knowledge of the asymptotic overlap m_x^*, and squared norms q_x^*, ρ_x, most quantities of interest can be determined. For instance, the quadratic reconstruction error is obtained from its definition as $\mathrm{E} = \rho_x + q_x^* - 2m_x^*$, while the angle between the ground-truth vector and the estimator is $\theta = \arccos(m_x^*/(\sqrt{\rho_x q_x^*}))$. One can also evaluate the generalization error for new random Gaussian samples, as advocated in [3], or compute similar errors for the denoising of $\mathbf{z_0}$.

IV. Numerical Results

Obtaining a stable implementation of the fixed point equations can be challenging. We provide simulation details in appendix F along with a link to the script we used to produce the figures. Theoretical predictions (full lines) are compared with numerical experiments (points) conducted using standard convex optimization solvers from [42]. The comparison with finite size ($N \approx$ a few hundreds) numerical experiments shows that, despite being asymptotic in nature, the predictions are accurate even at moderate system sizes. All experimental points were done with $N = 200$ and averaged one hundred times.

A. Validity of the Replica Prediction

We start with a simple verification of the replica prediction in Figure1, on a classification problem where data is generated as $\mathbf{y} = \mathrm{sign}(\mathbf{F}\mathbf{x_0})$. We consider two types of singular value distributions for $\mathbf{F}$ and three types of losses: a square loss, a linear support vector classification (SVC) loss and a logistic loss. Technical details and expressions are given in appendix F. We use ridge regularization with penalty $f = \frac{\lambda_2}{2}\|\cdot\|_2^2$. We plot the reconstruction angle θ as a function of the aspect ratio of the problem α in Figure 1. A first plot is done with a Marchenko-Pastur eigenvalue distribution for $\mathbf{F}^T\mathbf{F}$ corresponding to $\mathbf{F}$ being i.i.d Gaussian. We then move out of the Gaussian setting and change the eigenvalue distribution for (34), which has a qualitatively similar behaviour: it has bounded support, and includes vanishing singular values at a given value $\alpha = 1$ of the aspect ratio. We recover a result close to the i.i.d. Gaussian one, including the error peak for the square loss when $\alpha = 1$. In both cases, the SVC and the logistic regression perform similarly. Note that error peaks can also be obtained for the max-margin solution as shown in [43], using a more elaborate teacher.

B. Sparse Logistic Regression

We now use the replica prediction to study sparse logistic regression with i.i.d Gaussian and row-orthogonal data, the

$$f = -\underset{m_x,\chi_x,q_x,m_z,\chi_z,q_z}{\mathrm{extr}} \{g_\mathrm{F} + g_\mathrm{G} - g_\mathrm{S}\}, \tag{10}$$
$$g_\mathrm{F} = \underset{\hat{m}_{1x},\hat{\chi}_{1x},\hat{Q}_{1x},\hat{m}_{1z},\hat{\chi}_{1z},\hat{Q}_{1z}}{\mathrm{extr}} \left\{\frac{1}{2}q_x\hat{Q}_{1x} - \frac{1}{2}\chi_x\hat{\chi}_{1x} - \hat{m}_{1x}m_x - \alpha\hat{m}_{1z}m_z + \frac{\alpha}{2}\left(q_z\hat{Q}_{1z} - \chi_z\hat{\chi}_{1z}\right)\right.$$
$$\left. + \mathbb{E}\left[\phi_x(\hat{m}_{1x},\hat{Q}_{1x},\hat{\chi}_{1x};x_0,\xi_{1x})\right] + \alpha\mathbb{E}\left[\phi_z(\hat{m}_{1z},\hat{Q}_{1z},\hat{\chi}_{1z};z_0,\xi_{1z})\right]\right\},$$
$$g_\mathrm{G} = \underset{\hat{m}_{2x},\hat{\chi}_{2x},\hat{Q}_{2x},\hat{m}_{2z},\hat{\chi}_{2z},\hat{Q}_{2z}}{\mathrm{extr}} \left\{\frac{1}{2}q_x\hat{Q}_{2x} - \frac{1}{2}\chi_x\hat{\chi}_{2x} - m_x\hat{m}_{2x} - \alpha m_z\hat{m}_{2z} + \frac{\alpha}{2}\left(q_z\hat{Q}_{2z} - \chi_z\hat{\chi}_{2z}\right)\right.$$
$$\left. - \frac{1}{2}\left(\mathbb{E}\left[\log(\hat{Q}_{2x} + \lambda\hat{Q}_{2z})\right] - \mathbb{E}\left[\frac{\hat{\chi}_{2x} + \lambda\hat{\chi}_{2z}}{\hat{Q}_{2x} + \lambda\hat{Q}_{2z}}\right] - \mathbb{E}\left[\frac{\rho_x(\hat{m}_{2x} + \lambda\hat{m}_{2z})^2}{(\hat{Q}_{2x} + \lambda\hat{Q}_{2z})}\right]\right)\right\},$$
$$g_\mathrm{S} = \frac{1}{2}\left(\frac{q_x}{\chi_x} - \frac{m_x^2}{\rho_x\chi_x}\right) + \frac{\alpha}{2}\left(\frac{q_z}{\chi_z} - \frac{m_z^2}{\rho_z\chi_z}\right),$$

where ϕ_x and ϕ_z are the potential functions

$$\phi_x(\hat{m}_{1x},\hat{Q}_{1x},\hat{\chi}_{1x};x_0,\xi_{1x}) = \lim_{\beta\to\infty}\frac{1}{\beta}\log\int e^{-\frac{\beta\hat{Q}_{1x}}{2}x^2 + \beta(\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x})x - \beta f(x)}dx,$$
$$\phi_z(\hat{m}_{1z},\hat{Q}_{1z},\hat{\chi}_{1z};z_0,\chi_{1z}) = \lim_{\beta\to\infty}\frac{1}{\beta}\log\int e^{-\frac{\beta\hat{Q}_{1z}}{2}z^2 + \beta(\hat{m}_{1z}z_0 + \sqrt{\hat{\chi}_{1z}}\xi_{1z})z - \beta g(y,z)}dz.$$

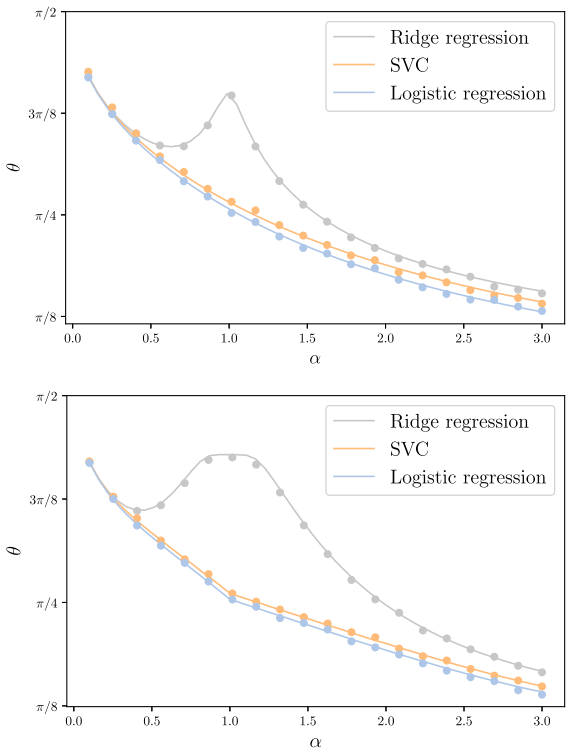

Fig. 1. Illustration of Theorem 1 in a binary classification problem with data generated as $\mathbf{y} = \phi(\mathbf{F}\mathbf{x}_0)$ with the data matrix $\mathbf{F}$ being **Left:** a Gaussian i.i.d. matrix and **Right:** a random orthogonal invariant matrix with a squared uniform density of singular values. We plot the angle between the estimator and the ground-truth vector $\theta = \arccos(m_x^*/(\sqrt{\rho_x q_x^*}))$ as a function of the aspect ratio $\alpha = M/N$ with three different losses: ridge regression, a Support Vector Machine with linear kernel and a logistic regression. f is a ℓ_2 penalty with parameter $\lambda_2 = 10^{-3}$. The theoretical prediction (full line) is compared with numerical experiments (points) conducted using standard convex optimization solvers from [42].

latter being ubiquitous in signal processing. Row-orthogonal data gives rise to a discrete eigenvalue distribution for $\mathbf{F}^T\mathbf{F}$ of zeroes and ones:

$$\lambda_{\mathbf{F}^T\mathbf{F}} \sim \max(0, 1-\alpha)\delta(0) + \min(1, \alpha)\delta(1)$$

and is often found to outperform Gaussian sensing matrices for recovery tasks, see e.g. [21] or [30]. In what follows, we define the sparsity ρ of the ground truth vector as the fraction of non-zero components which are sampled from a standard normal distribution. Labels are generated with $\mathbf{y} = \text{sign}(\mathbf{F}\mathbf{x_0})$ as for Figure 1.

1) Effect of Sparsity: In Figure 2, we start by plotting the reconstruction angle against the aspect ratio of the measurement matrix for different values of the sparsity of the teacher vector, for ℓ_2 regularization $f = \frac{\lambda_2}{2}\|\cdot\|_2^2$ and ℓ_1 regularization $f = \lambda_1\|\cdot\|_1$, and a fixed value of regularization parameters λ_1, λ_2. In the case of ℓ_2-regularization, we observe that the reconstruction performance remains the same whatever the sparsity of the original teacher vector as all curves collapse together (top and bottom left). The ridge regularization is thus unable to differentiate sparse and non-sparse problems. For ℓ_1, better performance is observed when the sparsity increases. Comparing the values for ℓ_2 and ℓ_1 also shows that, for a non-sparse signal, ℓ_2 and ℓ_1 reconstruction perform similarly. The largest difference is observed at $\rho = 0.1$, where the ℓ_1 penalized logistic regression significantly outperforms the ridge one. We thus keep this value of the sparsity parameter for the next figures.

2) Varying the Regularization Parameter at Constant Sparsity: In Figure 3, keeping the sparsity of the teacher constant at $\rho = 0.1$, we look to tune the regularization strength. An interesting effect appears in the ridge-regularized case with row-orthogonal measurements: the curves collapse to a single one when the aspect ratio goes below $\alpha = 1$. We find that the optimal regularization strength for the ℓ_2 penalty lies around $\lambda_2 = 0.01$, and for the ℓ_1-penalty around $\lambda_1 = 0.1$, for both types of matrices.

3) Comparing Case: In Figure 4, we directly compare the reconstruction performance of logistic regression on a sparse problem with previously tuned regularization parameter of ℓ_2 and ℓ_1 penalties, with the two types of measurement matrices. We naturally observe that the ℓ_1 penalty leads to better reconstruction of the sparse vector. Row-orthogonal matrices outperform the i.i.d. Gaussian ones with both regularization, although the gap is less significant with the ℓ_1 penalty.

4) Discussion: Several non-trivial effects are observed when studying the interplay between eigenvalue distribution of the design matrix, loss function, regularization and structure of the underlying teacher vector. Looking for analytical simplifications of the fixed point equations from Theorem 1 in specific cases would be interesting to understand how the key quantities interact and lead, for example, to the collapsing observed in ℓ_2-penalized problems. This further motivates the use of these equations to determine reconstruction limits of generalized-linear modeling. Some examples include limits of sparse recovery for different types of measurement matrices, or finding if optimal losses can be designed to achieve performances close to Bayes optimal errors.

V. Sketch of Proof of Theorem 1

Our proof follows an approach pioneered in [4] where the LASSO risk for i.i.d. Gaussian matrices is determined. The idea is to build a sequence of iterates that provably converges towards the estimator $\hat{\mathbf{x}}$, while also knowing the statistical properties of those iterates through a set of equations. We must therefore concern ourselves with three fundamental aspects:

(i) construct a sequence of iterates with a rigorous statistical characterization that matches their equations of Theorem 1 at the fixed point,
(ii) verify that the sequence's fixed point corresponds to the estimator $\hat{\mathbf{x}}$,
(iii) check that this sequence is provably convergent, otherwise the iterates might drift off on a diverging trajectory, and the fixed point would never be reached. We thus make sure the statistical characterization indeed applies to the point of interest $\hat{\mathbf{x}}$.

In short, we have a sequence of estimates $(\mathbf{x}_k)_{k\in\mathbb{N}}$ taking values in $\mathbb{R}^N$, and their exact asymptotic (in N) distribution for any $k > 0$. To show that these statistics extend to $\hat{\mathbf{x}}$, we need to show that $\lim_{k\to\infty} \mathbf{x}_k = \hat{\mathbf{x}}$. To do so, we need the sequence

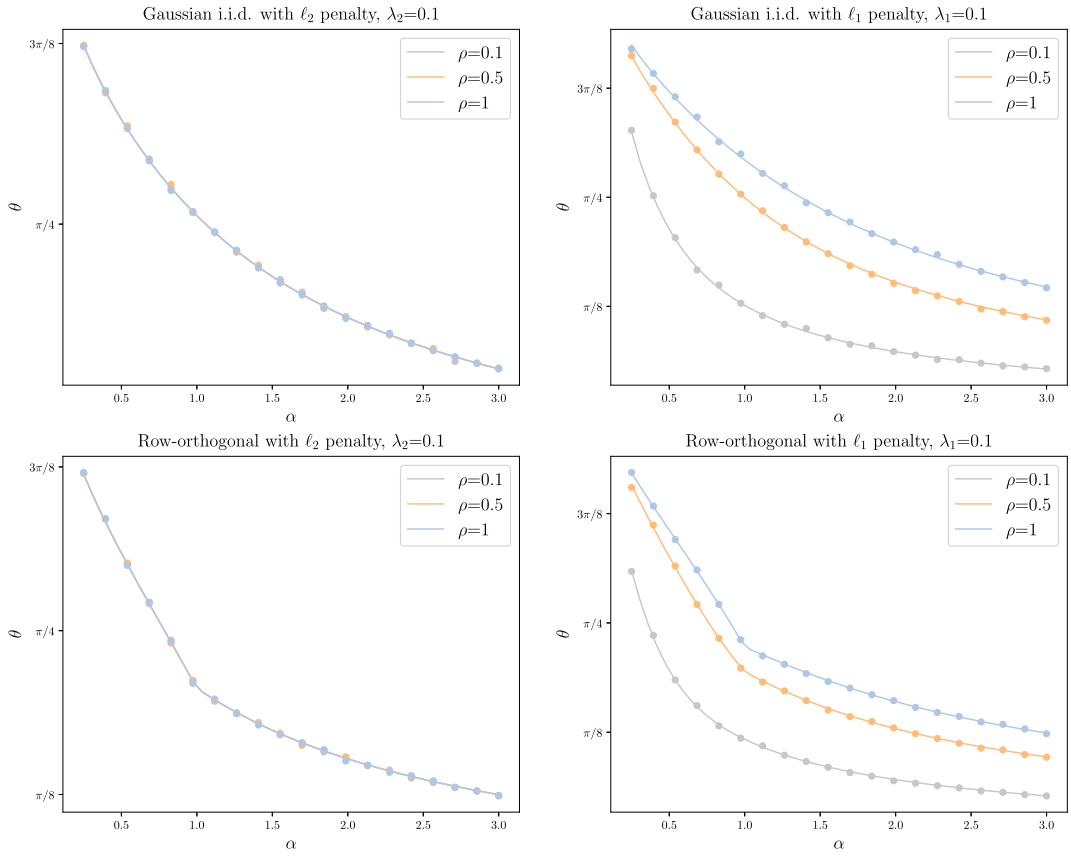

Fig. 2. Effect of the sparsity of the planted vector. We plot the angle between the estimator and the ground truth in a binary classification problem with $\mathbf{y} = \text{sign}(\mathbf{F}\mathbf{x}_0)$ as a function of $\alpha = M/N$, for different values of sparsity ρ. We use logistic regression. Figures in the top are for $\mathbf{F}$ Gaussian i.i.d., while figures in the bottom are for $\mathbf{F}$ row-orthogonal. **Left:** we use a ℓ_2 penalty with parameter $\lambda_2 = 0.1$, and notice that the angle is the same for any sparsity. **Right:** we use a ℓ_1 penalty with parameter $\lambda_1 = 0.1$. The theoretical prediction (full line) is compared with numerical experiments (points) conducted using standard convex optimization solvers from [42].

to converge (i.e. point iii), and its fixed point to be $\hat{\mathbf{x}}$ (point ii). As indicated in the introduction, we will use an instance of the 2-layer MLVAMP algorithm to construct this sequence. Note that, for the sake of brevity, we do not verify that limiting points of 2-layer MLVAMP trajectories $\lim_{k\to\infty} \mathbf{x}_k$ converge empirically to the Gaussian distribution prescribed by the state evolution equations. This point is treated explicitly in [25].

The following lemma establishes the link between the state evolution equations and our main theorem.

Lemma 1 (Fixed Point of 2-Layer MLVAMP State Evolution Equations): The state evolution equations of 2-layer MLVAMP from [36], reminded in appendix E, match the equations of Theorem 1 at their fixed point.

Proof: See appendix E. ■

This confirms that 2-layer MLVAMP is a good choice to design the sequences that we seek. We know that the iterates of 2-layer MLVAMP can be characterized by state evolution equations which correspond, at their fixed point, to the equations of Theorem 1 by virtue of Lemma 1. The necessary assumptions for the state evolution equations to hold are verified in appendix E-B. We must now show that the estimator of interest defined by (1) and (2) can be reached using 2-layer MLVAMP. We thus continue with point (ii).

Lemma 2 (Fixed Point of 2-layer MLVAMP): The fixed point of algorithm (1) matches the optimality condition of the unconstrained convex problem Eq.(2)

Proof: See appendix D. ■

This part is a consequence of the structure of the algorithm and properties of proximal operators. We now move to point (iii) and seek to characterize the convergence properties of 2-layer MLVAMP. Instead of directly tackling the convergence of 2-layer MLVAMP on any convex GLM, we take a detour and focus on a constrained problem, where functions f and g are augmented by a ℓ_2 norm with ridge parameters λ_2, $\tilde{\lambda}_2$. The called on intuition is that the algorithm will be more likely to converge in a strongly convex problem. We start by showing the convergence of MLVAMP in the constrained strongly convex setting, for values of λ_2 larger than a certain threshold, and any strictly positive $\tilde{\lambda}_2$.

Lemma 3 (Linear Convergence of 2-Layer MLVAMP for Strongly Convex Problems): Assume f and g are twice differentiable. Define the constrained problem

$$\hat{\mathbf{x}}(\lambda_2, \tilde{\lambda}_2) = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min} \left\{ \tilde{g}(\mathbf{F}\mathbf{x}, \mathbf{y}) + \tilde{f}(\mathbf{x}) \right\} \tag{11}$$

where $\tilde{f}(\mathbf{x}) = f(\mathbf{x}) + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2$, $\tilde{g}(\mathbf{x}, \mathbf{y}) = g(\mathbf{x}, \mathbf{y}) + \frac{\tilde{\lambda}_2}{2}\|\mathbf{x}\|_2^2$. Consider 2-layer MLVAMP applied to find (11), from which

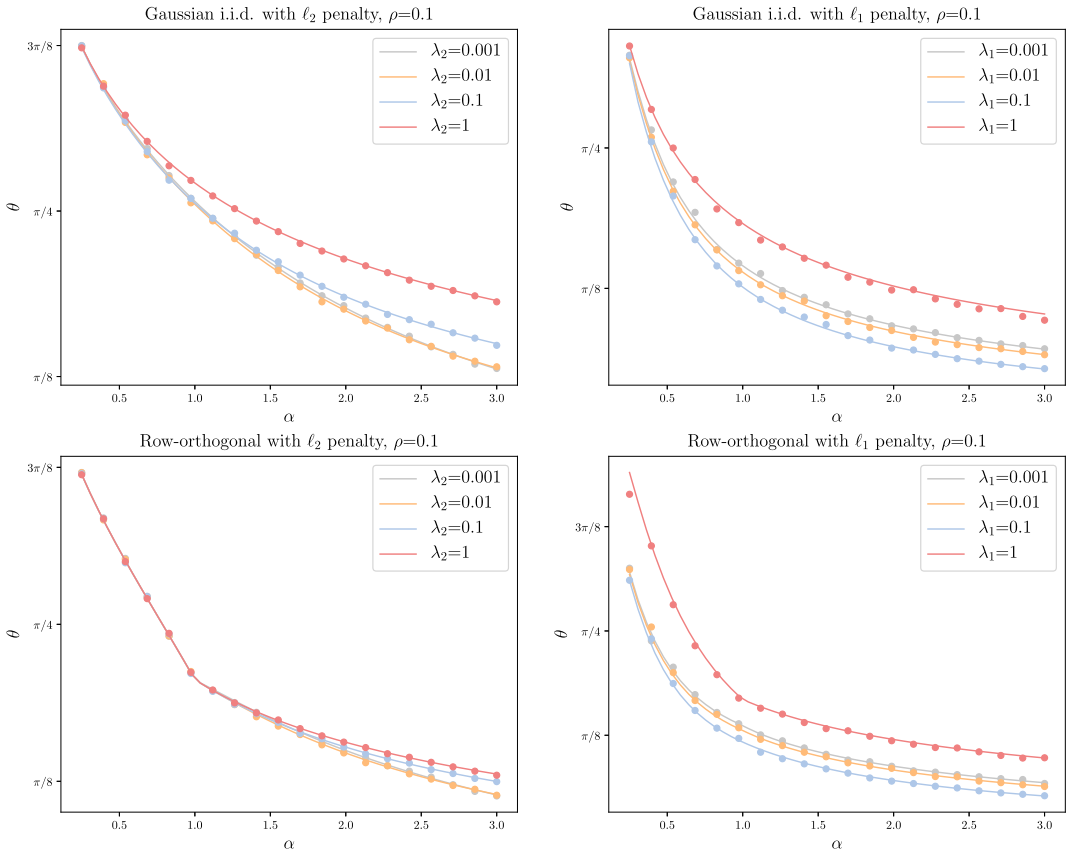

Fig. 3. Tuning the regularization parameter. We still plot the angle between the estimator and the ground truth in a binary classification problem with $\mathbf{y} = \text{sign}(\mathbf{F}\mathbf{x}_0)$ as a function of $\alpha = M/N$, for a fixed sparsity of planted vector $\rho = 0.1$, for different values of regularization parameters. Figures in the top are for $\mathbf{F}$ Gaussian i.i.d., while figures in the bottom are for $\mathbf{F}$ row-orthogonal. **Left:** ℓ_2 penalty with different values of regularization parameter λ_2. **Right:** ℓ_1 penalty with different values of regularization parameter λ_1.

we extract at each iteration the vector $\mathbf{h}^{(t)} = \left[\mathbf{h}_{2z}^{(t)}, \mathbf{h}_{1x}^{(t)}\right]^T$. Let $\mathbf{h}^*$ be its value at the fixed point of algorithm (1). We then have that, for any $\tilde{\lambda}_2 > 0$, there exists a value λ_2^* such that, for any $\lambda_2 > \lambda_2^*$, there exists a strictly positive constant c verifying $0 < c < \lambda_2$, such that for any $t \in \mathbb{N}$:

$$\|\mathbf{h}^{(t)} - \mathbf{h}^*\|_2^2 \leqslant \left(\frac{c}{\lambda_2}\right)^t \|\mathbf{h}^{(0)} - \mathbf{h}^*\|_2^2,$$

The convergence of $\mathbf{h}^{(t)}$ implies that estimators $\hat{\mathbf{x}}_1^{(t)}$ and $\hat{\mathbf{x}}_2^{(t)}$ returned by 2-layer MLVAMP also converge to the desired $\hat{\mathbf{x}}(\lambda_2, \tilde{\lambda}_2)$, i.e., under the conditions listed above

$$\lim_{t\to\infty} \|\hat{\mathbf{x}}^{(t)} - \hat{\mathbf{x}}(\lambda_2, \tilde{\lambda}_2)\|_2^2 = 0.$$

Proof: See appendix G. ■

For a loss function $\tilde{g}$ with any non-zero strong convexity constant, and a regularization $\tilde{f}$ with a sufficiently strong convexity, 2-layer MLVAMP converges linearly towards its unique fixed point. Note that this convergence result is independent from the dimension. We elaborate on this lemma in the next section. An immediate consequence is the following lemma, which claims that Theorem 1 holds when 2-layer MLVAMP converges. Since this result does not rely on an analytic continuation, the assumptions on the concentration of

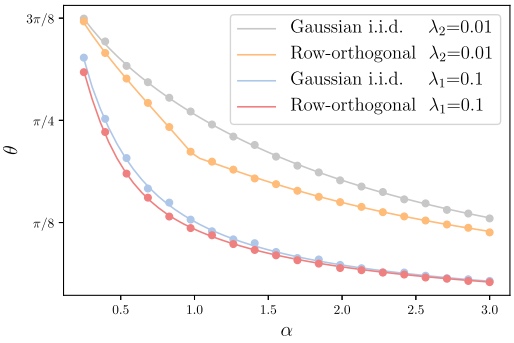

Fig. 4. Comparing reconstruction performance for Gaussian i.i.d. and row-orthogonal matrices. In this figure, we compare the reconstruction angles between the estimator and the ground-truth for binary classification obtained with ℓ_1 and ℓ_2 penalties. We use logistic regression. The sparsity of the sparse vector is fixed to $\rho = 0.1$. For both Gaussian i.i.d. and row-orthogonal data matrices, we see that ℓ_1 penalty with $\lambda_1 = 0.1$ performs better than the ℓ_2 penalty with $\lambda_2 = 0.01$. For those two penalties, row-orthogonal matrices allow to obtain smaller reconstruction angles than Gaussian i.i.d. matrices.

PL2 observables of $\hat{\mathbf{x}}$, given by the state evolution property, and approximation of the cost function by analytic functions with fast decaying higher order derivatives are not required. The result can also be stated for any PL2 observable, with no restriction on its derivability and decay of higher order

derivatives. We summarize the necessary assumptions in the following list:

Assumption 2:

(a) the functions f and g are proper, closed, convex and separable functions.
(b) the cost function $g(\mathbf{F}.,\mathbf{y}) + f(.)$ is coercive, i.e. $\lim_{\|\mathbf{x}\|\to\infty} g(\mathbf{F}\mathbf{x},\mathbf{y}) + f(\mathbf{x}) = +\infty$.
(c) there exists a constant B_1 such that $\frac{1}{N}\|\hat{\mathbf{x}}\|_2^2 \leqslant B_1$ almost surely as $N \to \infty$.
(d) for any $\mathbf{x} \in \text{dom}(f)$ and any $\mathbf{x}' \in \partial f(\mathbf{x})$, there exists a constant C such that $\|\mathbf{x}'\|_2 \leqslant C(1 + \|\mathbf{x}\|_2)$. The same holds for g on its domain.
(e) the empirical distributions of the underlying truth $\mathbf{x}_0$, eigenvalues of $\mathbf{F}^T\mathbf{F}$, and noise vector w_0, respectively converge empirically with second order moments, as defined in appendix A, to independent scalar random variables x_0, w_0, λ with distributions p_{x_0}, p_λ, p_{w_0}. We assume that the distribution p_λ is not all-zero and has compact support.
(f) the design matrix $\mathbf{F} = \mathbf{U}\mathbf{D}\mathbf{V}^\top \in \mathbb{R}^{M\times N}$ is rotationally invariant, as defined in the introduction, where the elements of the Haar distributed matrices $\mathbf{U}, \mathbf{V}$ are independent of the elements of the ground truth vector $\mathbf{x}_0$, noise $\boldsymbol{\omega}_0$ and elements of $\mathbf{D}$.
(g) the solution to the set of fixed point equations (9) exists and is unique for any convex functions f, g verifying the
(h) finally assume that $M, N \to \infty$ with fixed ratio $\alpha = M/N$.

Lemma 4 (Asymptotic Error for the Twice Differentiable, Sufficiently Strongly Convex Problem): Consider the strongly convex minimization problem with twice differentiable f and g (11). Under the set of assumptions 2, for any $\tilde{\lambda}_2 > 0$, there exists a λ_2^* such that, for any $\lambda_2 > \lambda_2^*$, Then, for any pseudo-Lipschitz function of order 2 ϕ, the following holds:

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \phi(x_{0,i}, \hat{x}_i) \overset{a.s.}{=} \mathbb{E}[\phi(x_0, \text{Prox}_{f/\hat{Q}_{1x}^{(t)}}(H_x))]$$
$$\lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^{M} \phi(z_{0,i}, \hat{z}_i) \overset{a.s.}{=} \mathbb{E}[\phi(z_0, \text{Prox}_{f/\hat{Q}_{1z}^{(t)}}(H_z))]$$

where the scalars $\hat{Q}_{1x}, \hat{Q}_{1z}$ and the random variables H_x, H_z are defined as in Theorem 1.

Proof: Using the result from Lemma 3, we have $\lim_{t\to\infty}\lim_{N\to\infty} \frac{1}{N}\|\mathbf{x}^{(t)} - \hat{\mathbf{x}}(\lambda_2, \tilde{\lambda}_2)\|_2^2 = 0$. As proven in [25], the state evolution parameters will converge to those of the fixed point of the state evolution equations along a converging trajectory of 2-layer MLVAMP. Using the assumption on the bounded averaged norm of $\hat{\mathbf{x}}$, the state evolution equations to show that the averaged norm of the iterates are bounded along a converging trajectory, and the state evolution equations to obtain the exact asymptotics of each iterate along the converging trajectory, an identical argument to that of the proof of Theorem 1.5 from [26] gives Lemma 4. ■

We are now left to prove Theorem 1, for any range of parameters $(\lambda_2, \tilde{\lambda}_2)$. $\tilde{\lambda}_2$ can already be chosen arbitrarily small. This means we need to relax the threshold value on λ_2 for the validity of the scalar quantities involved in Theorem 1. To do so, we start by introducing another modification of the original problem, where the objective functions are assumed to be real analytic. Lemma 4 naturally holds for real analytic convex functions. Proving Theorem 1 on the real analytic problem then boils down to performing an analytic continuation on the λ_2 parameter, and is detailed in Appendix H. We thus have the following intermediate result:

Lemma 5 (Asymptotics of the Real Analytic Problem): Consider assumption 1 is verified. Suppose additionally that f and g are real analytic. Then Theorem 1 holds for any $\tilde{\lambda}_2 > 0$ and any $\lambda_2 > 0$.

Theorem 1 can then be proven from Lemma 5 by showing that the solutions of the original problem and of its real analytic approximation are arbitrarily close, and by carefully studying the limits $\tilde{\lambda}_2 \to 0$ and $\lambda_2 \to 0$. This is deferred to Appendix H. Note that the proof of the analytic continuation presented here makes the one from [30], which was incomplete, rigorous.

The remaining technical part is the proof of the convergence Lemma 3. For this purpose, we use a dynamical system reformulation of 2-layer MLVAMP and a result from control theory, adapted to machine learning in [44] and more specifically to ADMM in [45].

VI. Convergence Analysis of 2-Layer MLVAMP

The key idea of the approach pioneered in [44] is to recast any non-linear dynamical system as a linear one, where convergence will be naturally characterized by a matrix norm. For a given non-linearity $\tilde{\mathcal{O}}$ and iterate $\mathbf{v}$, we define the variable $\mathbf{u} = \tilde{\mathcal{O}}(\mathbf{v})$ and rewrite the initial algorithm in terms of this trivial transform. Any property of $\tilde{\mathcal{O}}$ is then summarized in a constraint matrix linking $\mathbf{v}$ and $\mathbf{u}$. For example, if $\tilde{\mathcal{O}}$ has Lipschitz constant ω, then for all t:

$$\|\mathbf{u}^{(t+1)} - \mathbf{u}^{(t)}\|_2^2 \leqslant \omega^2\|\mathbf{v}^{(t+1)} - \mathbf{v}^{(t)}\|_2^2,$$

which can be rewritten in matrix form:

$$\mathbf{U}^T \begin{bmatrix} \omega^2\mathbf{I}_{d_v} & 0 \\ 0 & -\mathbf{I}_{d_u} \end{bmatrix} \mathbf{U} \succcurlyeq 0$$
$$\text{where } \mathbf{U} = \begin{bmatrix} \mathbf{v}^{(t+1)} - \mathbf{v}^{(t)} \\ \mathbf{u}^{(t+1)} - \mathbf{u}^{(t)} \end{bmatrix}$$

where $\mathbf{I}_{d_v}, \mathbf{I}_{d_u}$ are the identity matrices with dimensions of $\mathbf{v}, \mathbf{u}$, i.e. M or N in our case. Any co-coercivity property (verified by proximal operators) can be rewritten in matrix form but yields non block diagonal constraint matrices. We will thus directly use the Lipschitz constants for our proof, as they lead to simpler derivations and suffice to prove the required result. The main theorem from [44], adapted to ADMM in [45], then establishes a sufficient condition for convergence with a linear matrix inequality, involving the matrices defining the linear recast of the algorithm and the constraints. Let us now detail how this approach can be used on 2-layer MLVAMP.

A. 2-Layer MLVAMP as a Dynamical System: Sketch of Proof of Lemma 3

We start by rewriting 2-layer MLVAMP in a more compact form:

$$\begin{aligned}
&\text{Initialize } \mathbf{h}_{1x}^{(0)}, \mathbf{h}_{2z}^{(0)}\\
\mathbf{h}_{1x}^{(t+1)} &= \mathbf{W}_1^{(t)}\tilde{\mathcal{O}}_1^{(t)}\mathbf{h}_{1x}^{(t)}\\
&\quad + \mathbf{W}_2^{(t)}\tilde{\mathcal{O}}_2^{(t)}(\mathbf{W}_3^{(t)}\mathbf{h}_{2z}^{(t)} + \mathbf{W}_4^{(t)}\tilde{\mathcal{O}}_1^{(t)}(\mathbf{h}_{1x}^{(t)}))\\
\mathbf{h}_{2z}^{(t+1)} &= \tilde{\mathcal{O}}_2^{(t)}(\mathbf{W}_3^{(t)}\mathbf{h}_{2z}^{(t)} + \mathbf{W}_4^{(t)}\tilde{\mathcal{O}}_1^{(t)}(\mathbf{h}_{1x}^{(t)}))
\end{aligned} \tag{12}$$

where

$$\begin{aligned}
\mathbf{W_1}^{(t)} &= \frac{\hat{Q}_{2x}^{(t)}}{\hat{Q}_{1x}^{(t+1)}}\left(\frac{1}{\chi_{2x}^{(t+1)}}(\hat{Q}_{2z}^{(t+1)}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}^{(t)}\mathrm{Id})^{-1} - \mathrm{Id}\right)\\
\mathbf{W_2}^{(t)} &= \frac{\hat{Q}_{2z}^{(t+1)}}{\chi_{2x}^{(t+1)}\hat{Q}_{1x}^{(t+1)}}(\hat{Q}_{2z}^{(t+1)}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}^{(t)}\mathrm{Id})^{-1}\mathbf{F}^T\\
\mathbf{W_3}^{(t)} &= \frac{\hat{Q}_{2z}^{(t)}}{\hat{Q}_{1z}^{(t)}}\left(\frac{1}{\chi_{2z}^{(t)}}\mathbf{F}(\hat{Q}_{2z}^{(t)}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}^{(t)}\mathrm{Id})^{-1}\mathbf{F}^T - \mathrm{Id}\right)\\
\mathbf{W_4}^{(t)} &= \frac{\hat{Q}_{2x}^{(t)}}{\hat{Q}_{1z}^{(t)}\chi_{2z}^{(t)}}\mathbf{F}(\hat{Q}_{2z}^{(t)}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}^{(t)}\mathrm{Id})^{-1}\\
\tilde{\mathbf{O}}_\mathbf{1}^{(t)} &= \frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}}\left(\frac{1}{\chi_{1x}^{(t)}\hat{Q}_{1x}^{(t)}}\mathrm{Prox}_{\mathbf{f}/\hat{Q}_{1x}^{(t)}}(\cdot) - \mathrm{Id}\right)\\
\tilde{\mathbf{O}}_\mathbf{2}^{(t)} &= \frac{\hat{Q}_{1z}^{(t)}}{\hat{Q}_{2z}^{(t+1)}}\left(\frac{1}{\chi_{1z}^{(t)}\hat{Q}_{1z}^{(t)}}\mathrm{Prox}_{\mathbf{g}(.,y)/\hat{Q}_{1z}^{(t)}}(\cdot) - \mathrm{Id}\right).
\end{aligned} \tag{13}$$

For the linear recast, we then define the variables:

$$\begin{aligned}
&\mathbf{u}_1^{(t)} = \tilde{\mathcal{O}}_1^{(t)}(\mathbf{h}_{1x}^{(t)}),\ \mathbf{v}^{(t)} = \mathbf{W}_3^{(t)}\mathbf{h}_{2z}^{(t)} + \mathbf{W}_4^{(t)}\mathbf{u}_1^{(t)},\\
&\mathbf{u}_2^{(t)} = \tilde{\mathcal{O}}_2^{(t)}(\mathbf{v}^{(t)}),\\
&\text{s.t. } \mathbf{h}_{2z}^{(t+1)} = \mathbf{u}_2^{(t)}, \mathbf{h}_{1x}^{(t+1)} = \mathbf{W}_1^{(t)}\mathbf{u}_1^{(t)} + \mathbf{W}_2^{(t)}\mathbf{u}_2^{(t)}.
\end{aligned}$$

where $\mathbf{u}_1, \mathbf{h}_{1x} \in \mathbb{R}^N$; and $\mathbf{v}, \mathbf{u}_2, \mathbf{h}_{2z} \in \mathbb{R}^M$. We then define as new variables the vectors

$$\mathbf{h}^{(t)} = \begin{bmatrix}\mathbf{h}_{2z}^{(t)}\\ \mathbf{h}_{1x}^{(t)}\end{bmatrix}, \quad \mathbf{u}^{(t)} = \begin{bmatrix}\mathbf{u}_2^{(t)}\\ \mathbf{u}_1^{(t)}\end{bmatrix},$$
$$\mathbf{w}_1^{(t)} = \begin{bmatrix}\mathbf{h}_{1x}^{(t)}\\ \mathbf{u}_1^{(t)}\end{bmatrix}, \quad \mathbf{w}_2^{(t)} = \begin{bmatrix}\mathbf{v}^{(t)}\\ \mathbf{u}_2^{(t)}\end{bmatrix}.$$

This leads to the following linear dynamical system recast of (12):

$$\begin{aligned}
\mathbf{h}^{(t+1)} &= \mathbf{A}^{(t)}\mathbf{h}^{(t)} + \mathbf{B}^{(t)}\mathbf{u}^{(t)}\\
\mathbf{w}_1^{(t)} &= \mathbf{C}_1^{(t)}\mathbf{h}^{(t)} + \mathbf{D}_1^{(t)}\mathbf{u}^{(t)}\\
\mathbf{w}_2^{(t)} &= \mathbf{C}_2^{(t)}\mathbf{h}^{(t)} + \mathbf{D}_2^{(t)}\mathbf{u}^{(t)}
\end{aligned} \tag{14}$$

where

$$\begin{aligned}
\mathbf{A}^{(t)} &= \mathbf{0}_{(M+N)\times(M+N)}\ \mathbf{B}^{(t)} = \begin{bmatrix}\mathbf{I}_M & \mathbf{0}_{M\times N}\\ \mathbf{W}_2^{(t)} & \mathbf{W}_1^{(t)}\end{bmatrix}\\
\mathbf{C}_1^{(t)} &= \begin{bmatrix}\mathbf{0}_{N\times M} & \mathbf{I}_N\\ \mathbf{0}_{N\times M} & \mathbf{0}_{N\times N}\end{bmatrix}\mathbf{D}_1^{(t)} = \begin{bmatrix}\mathbf{0}_{N\times M} & \mathbf{0}_{N\times N}\\ \mathbf{0}_{N\times M} & \mathbf{I}_N\end{bmatrix}\\
\mathbf{C}_2^{(t)} &= \begin{bmatrix}\mathbf{W}_3^{(t)} & \mathbf{0}_{M\times N}\\ \mathbf{0}_{M\times M} & \mathbf{0}_{M\times N}\end{bmatrix}\mathbf{D}_2^{(t)} = \begin{bmatrix}\mathbf{0}_{M\times M} & \mathbf{W}_4^{(t)}\\ \mathbf{I}_M & \mathbf{0}_{M\times N}\end{bmatrix}.
\end{aligned}$$

$\mathbf{O}$ denotes a matrix with only zeros. The next step is to impose the properties of the non-linearities $\tilde{\mathcal{O}}_1^{(t)}, \tilde{\mathcal{O}}_2^{(t)}$ through constraint matrices. The Lipschitz constants $\omega_1^{(t)}, \omega_2^{(t)}$ of $\tilde{\mathcal{O}}_1^{(t)}, \tilde{\mathcal{O}}_2^{(t)}$ can be determined using properties of proximal operators [46] and are directly linked to the strong convexity and smoothness of the cost function and regularization. The relevant properties of proximal operators are reminded in appendix B, while the subsequent derivation of the Lipschitz constants is detailed in appendix G and yields:

$$\begin{aligned}
\omega_1^{(t)} &= \frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}}\sqrt{1 + \frac{(\hat{Q}_{2x}^{(t)})^2 - (\hat{Q}_{1x}^{(t)})^2}{(\hat{Q}_{1x}^{(t)} + \lambda_2)^2}}\\
\omega_2^{(t)} &= \frac{\hat{Q}_{1z}^{(t)}}{\hat{Q}_{2z}^{(t)}}\sqrt{1 + \frac{(\hat{Q}_{2z}^{(t)})^2 - (\hat{Q}_{1z}^{(t)})^2}{(\hat{Q}_{1z}^{(t)} + \tilde{\lambda}_2)^2}}.
\end{aligned} \tag{15}$$

We thus define the constraints matrices

$$\mathbf{M}_1^{(t)} = \begin{bmatrix}(\omega_1^{(t)})^2 & 0\\ 0 & -1\end{bmatrix} \otimes \mathbf{I}_N, \quad \mathbf{M}_2^{(t)} = \begin{bmatrix}(\omega_2^{(t)})^2 & 0\\ 0 & -1\end{bmatrix} \otimes \mathbf{I}_M$$

where $\otimes$ denotes the Kronecker product. We then use a time dependent form of Theorem 4 from [44] in the appropriate form for 2-layer MLVAMP, as was done in [45] for ADMM.

Proposition 1 (Time Dependent Version of Theorem 4 from [44]): Consider, at each time step $t \in \mathbb{N}$, the following linear matrix inequality with $\tau_{(t)} \in [0, 1]$:

$$\begin{aligned}
0 \succeq &\begin{bmatrix}(\mathbf{A}^{(t)})^T\mathbf{P}\mathbf{A}^{(t)} - (\tau_{(t)})^2\mathbf{P} & (\mathbf{A}^{(t)})^T\mathbf{P}\mathbf{B}^{(t)}\\ (\mathbf{B}^{(t)})^T\mathbf{P}\mathbf{A}^{(t)} & (\mathbf{B}^{(t)})^T\mathbf{P}\mathbf{B}^{(t)}\end{bmatrix}\\
&+ \begin{bmatrix}\mathbf{C}_1^{(t)} & \mathbf{D}_1^{(t)}\\ \mathbf{C}_2^{(t)} & \mathbf{D}_2^{(t)}\end{bmatrix}^T\begin{bmatrix}\beta_1^{(t)}\mathbf{M}_1^{(t)} & \mathbf{0}_{2N\times 2M}\\ \mathbf{0}_{2M\times 2N} & \beta_2^{(t)}\mathbf{M}_2^{(t)}\end{bmatrix}\begin{bmatrix}\mathbf{C}_1^{(t)} & \mathbf{D}_1^{(t)}\\ \mathbf{C}_2^{(t)} & \mathbf{D}_2^{(t)}\end{bmatrix}
\end{aligned} \tag{16}$$

If, at each time step, (16) is feasible for some $\mathbf{P} \succ 0$ and $\beta_1^{(t)}, \beta_2^{(t)} \geqslant 0$, then for any initialization $\mathbf{h}^{(0)}$, $\mathbf{h}^{(t)}$ converges to $\mathbf{h}^*$, the fixed point of (14):

$$\forall t, \quad \|\mathbf{h}^{(t)} - \mathbf{h}^*\| \leqslant \sqrt{\kappa(\mathbf{P})}(\tau^*)^t\|\mathbf{h}^{(0)} - \mathbf{h}^*\|$$

where $\kappa(\mathbf{P})$ is the condition number of $\mathbf{P}$ and we defined $\tau^* = \sup_t \tau_{(t)}$.

Proof: see appendix G-A ■

We show in appendix G how the additional ridge penalties from the constrained problem (11) parametrized by $\lambda_2, \tilde{\lambda}_2$ can be used to make (16) feasible and prove Lemma 3. The core idea is to leverage on the Lipschitz constants (15), the operator norms of the matrices defined in (13) and the following upper and lower bounds on the $\hat{Q}$ parameters defined by the fixed point of state evolution equations:

$$\begin{aligned}
\lambda_{min}(\mathcal{H}_f) \leqslant \hat{Q}_{2x}^{(t)} &\leqslant \lambda_{max}(\mathcal{H}_f)\\
\lambda_{min}(\mathcal{H}_g) \leqslant \hat{Q}_{2z}^{(t+1)} &\leqslant \lambda_{max}(\mathcal{H}_g)\\
\hat{Q}_{2z}^{(t)}\lambda_{min}(\mathbf{F}^T\mathbf{F}) \leqslant \hat{Q}_{1x}^{(t+1)} &\leqslant \hat{Q}_{2z}^{(t)}\lambda_{max}(\mathbf{F}^T\mathbf{F})\\
\frac{\hat{Q}_{2x}^{(t)}}{\lambda_{max}(\mathbf{F}\mathbf{F}^T)} \leqslant \hat{Q}_{1z}^{(t)} &\leqslant \frac{\hat{Q}_{2x}^{(t)}}{\lambda_{min}(\mathbf{F}\mathbf{F}^T)},
\end{aligned}$$

where $\mathcal{H}_f, \mathcal{H}_g$ are the Hessian of the loss and regularization functions taken at the fixed point. These bounds are obtained

from the definitions of χ_x, χ_z in the state evolution equations (or equivalently in Theorem 1), and the fact that the derivative of a proximal operator reads, for a twice differentiable function:

$$\mathcal{D}\eta_{\gamma f}(\mathbf{x}) = (\mathrm{Id} + \gamma \mathcal{H}_f(\eta_{\gamma f}(\mathbf{x})))^{-1}.$$

Detail of this derivation can also be found in appendices B and G. For the constrained problem (11), the maximum and minimum eigenvalues of the Hessians are directly augmented by $\tilde{\lambda}_2, \lambda_2$, which allows us to control the scaling of the $\hat{Q}$ parameters. The rest of the convergence proof is then based on successive application of Schur's lemma [47] on the linear matrix inequality (16); and translating the resulting conditions on inequalities which can be verified by choosing the appropriate $\tilde{\lambda}_2, \lambda_2, \beta_1^{(t)}, \beta_2^{(t)}$. Convergence of gradient-based descent methods for sufficiently strongly-convex objectives is a coherent result from an optimization point of view. This is corroborated by the symbolic convergence rates derived for ADMM in [45], where a sufficiently strongly convex objective is also considered.

B. Numerical Experiments for Lemma 3

Here we provide numerical evidence for the linear convergence condition proved in Lemma 3. We consider a logistic regression penalized with the ℓ_1 norm ($\lambda_1 = 0.1$) with an ill-conditioned design matrix, with i.i.d. standard normal elements. This corresponds to the setting of Figure 3. Since the logistic loss is strongly convex on any compact space, we do not need to add $\tilde{\lambda}_2$. We follow the convergence of 2-layer MLVAMP for this problem for increasing values of an additional ridge penalty $\lambda_2 = 0, 0.01, 0.05, 0.1$ and plot the average distance between successive iterates $\frac{1}{N}\|\mathbf{h}_{1x}^{(t+1)} - \mathbf{h}_{1x}^{(t+1)}\|_2^2$ and the evolution of the reconstruction angle θ as a function of the number of iterations. We perform two experiments with aspect ratios $\alpha = 1$ and $\alpha = 0.2$. For $\alpha = 1$, 2-layer MLVAMP converges without any additional ridge penalty, and convergence is accelerated by larger values of λ_2. As a sanity check, note that the reconstruction angle of the estimator returned by the algorithm for $\lambda_2 = 0$ (grey line on the lower left plot) converges to the value predicted at Figure 3 for $\alpha = 1, \lambda_1 = 0.1$ and a Gaussian matrix. For $\alpha = 0.2$n the design matrix is ill-conditioned and we see that 2-layer MLVAMP diverges. Adding the ridge penalty leads to converging trajectories for a sufficiently large value of λ_2, as shown on the upper right block. Larger values of λ_2 again lead to faster convergence.

APPENDIX A
CONVERGENCE OF VECTOR SEQUENCES

This section is a brief summary of the framework originally introduced in [26] and used in [36] and [34]. We review the key definitions and verify that they apply in our setting. We remind the full set of state evolution equations from [36] at (27), when applied to learning a GLM, in appendix E, along with the required assumptions for them to hold in appendix E-B.

The main building blocks are the notions of *vector sequence* and *pseudo-Lipschitz function*, which allow to define the *empirical convergence with p-th order moment*. Consider a vector of the form

$$\mathbf{x}(N) = (\mathbf{x}_1(N), \dots, \mathbf{x}_N(N))$$

where each sub-vector $\mathbf{x}_n(N) \in \mathbb{R}^r$ for any given $r \in \mathbb{N}^*$. For r=1, which we use in Theorem 1, $\mathbf{x}(N)$ is denoted a *vector sequence*.

Given $p \geqslant 1$, a function $\mathbf{f} : \mathbb{R}^r \to \mathbb{R}^s$ is said to be *pseudo-Lipschitz continuous of order p* if there exists a constant $C > 0$ such that for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^s$:

$$\|\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_2)\| \leqslant C\|\mathbf{x}_1 - \mathbf{x}_2\| \left[1 + \|\mathbf{x}_1\|^{p-1} + \|\mathbf{x}_2\|^{p-1}\right].$$

Then, a given vector sequence $\mathbf{x}(N)$ *converges empirically with p-th order moment* if there exists a random variable $X \in \mathbb{R}^r$ such that:

- $\mathbb{E}\|X\|_p^p < \infty$; and
- for any scalar-valued pseudo-Lipschitz continuous $\mathbf{f} : \mathbb{R}^r \to \mathbb{R}$ of order p,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbf{f}(x_n(N)) = \mathbb{E}[f(X)].$$

Note that defining an empirically converging singular value distribution implicitly defines a sequence of matrices $\mathbf{F}(N)$ using the definition of rotational invariance from the introduction. This naturally brings us back to the original definitions from [26]. An important point is that the almost sure convergence of the second condition holds for random vector sequences, such as the ones we consider in the introduction. Note that the noise vector ω_0 must also satisfy these conditions, and naturally does when it is an i.i.d. Gaussian one. We also remind the definition of *uniform Lipschitz continuity*.

For a given mapping $\phi(\mathbf{x}, A)$ defined on $\mathbf{x} \in \mathcal{X}$ and $A \in \mathbb{R}$, we say it is *uniformly Lipschitz continuous* in $\mathbf{x}$ at $A = \bar{A}$ if there exists constants L_1 and $L_2 \geqslant 0$ and an open neighborhood U of $\bar{A}$ such that:

$$\|\phi(\mathbf{x}_1, A) - \phi(\mathbf{x}_2, A)\| \leqslant L_1\|\mathbf{x}_1 - \mathbf{x}_2\|$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $A \in U$; and

$$\|\phi(\mathbf{x}, A_1) - \phi(\mathbf{x}, A_2)\| \leqslant L_2(1 + \|\mathbf{x}\|)|A_1 - A_2|$$

for all $\mathbf{x} \in \mathcal{X}$ and $A_1, A_2 \in U$.

We discuss the required assumptions for the state evolution equations to hold in detail, and why they are verified in our setting, in appendix E-B.

APPENDIX B
CONVEX ANALYSIS AND PROPERTIES OF PROXIMAL OPERATORS

We start this section with a few useful definitions from convex analysis, which can all be found in textbooks such as [40]. We then remind important properties of proximal operators, which we use in appendix G to derive upper bounds on the Lipschitz constants of the non-linear operators $\tilde{\mathcal{O}}_1, \tilde{\mathcal{O}}_2$.

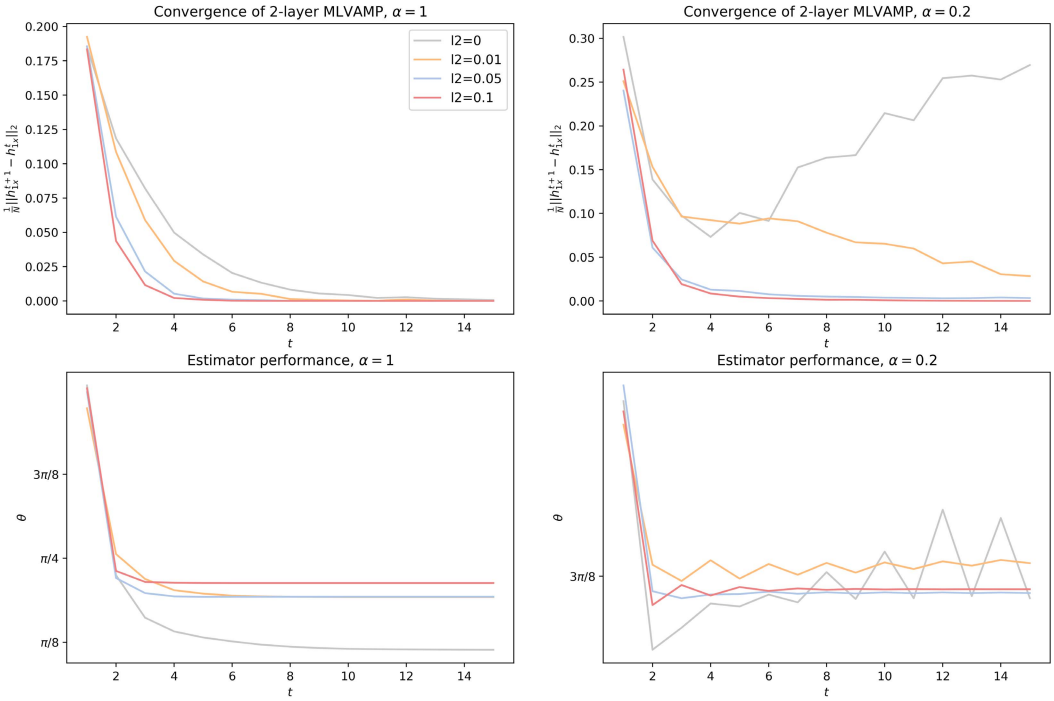

Fig. 5. Convergence of 2-layer MLVAMP on a logistic regression with ℓ_1 penalty with $\lambda_1 = 0.1$, a Gaussian design matrix and two values of the aspect ratio $\alpha = 1$ (left) and $\alpha = 0.2$ (right). For $\alpha = 1$, the algorithm converges regardless of the additional ridge penalty and we recover the performance predicted by Theorem 1 for the plain ℓ_1 regularization. For $\alpha = 0.2$, the plain ℓ_1 leads to an unstable iteration and a sufficiently large additional ridge indeed leads to convergence. In both cases, the larger the additional ridge, the faster the algorithm converges.

In what follows, we denote $\mathcal{X}$ the Hilbert space with scalar inner product serving as input and output space, here $\mathbb{R}^N$ or $\mathbb{R}^M$. For simplicity, we will write all operators as going from $\mathcal{X}$ to $\mathcal{X}$.

Definition 1 (Strong Convexity): A proper closed function is σ-strongly convex with $\sigma > 0$ if $f - \frac{\sigma}{2}\|.\|^2$ is convex. If f is differentiable, the definition is equivalent to

$$f(x) \geqslant f(y) + \langle \nabla f(y), x - y \rangle + \frac{\sigma}{2}\|x - y\|^2$$

for all $x, y \in \mathcal{X}$.

Definition 2 (Smoothness for Convex Functions): A proper closed function f is β-smooth with $\beta > 0$ if $\frac{\beta}{2}\|.\|^2 - f$ is convex. If f is differentiable, the definition is equivalent to

$$f(x) \leqslant f(y) + \langle \nabla f(y), x - y \rangle + \frac{\beta}{2}\|x - y\|^2$$

for all $x, y \in \mathcal{X}$.

An immediate consequence of those definitions is the following second order condition: for twice differentiable functions, f is σ-strongly convex and β-smooth if and only if:

$$\sigma \mathrm{Id} \preceq \mathcal{H}_\mathrm{f} \preceq \beta \mathrm{Id}.$$

Definition 3 (Co-coercivity): Let $T : \mathcal{X} \to \mathcal{X}$ and $\beta \in \mathbb{R}_+^*$. Then T is β co-coercive if βT is firmly-nonexpansive, i.e.

$$\langle \mathbf{x} - \mathbf{y}, T(\mathbf{x}) - T(\mathbf{y}) \rangle \geqslant \beta \|T(\mathbf{x}) - T(\mathbf{y})\|_2^2$$

for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

Proximal operators are 1 co-coercive or equivalently firmly-nonexpansive.

Corollary 3 (Remark 4.24 [40]): A mapping $T : \mathcal{X} \to \mathcal{X}$ is β-cocoercive if and only if βT is half-averaged. This means that T can be expressed as:

$$T = \frac{1}{2\beta}(\mathrm{Id} + \mathrm{S})$$

where S is a nonexpansive operator.

Proposition 2 (Resolvent of the Sub-Differential [40]): The proximal mapping of a convex function f is the resolvent of the sub-differential ∂f of f:

$$\mathrm{Prox}_{\gamma f} = (\mathrm{Id} + \gamma \partial \mathrm{f})^{-1}.$$

The following proposition is due to [46], and is useful to determine upper bounds on the Lipschitz constant of update functions involving proximal operators.

Proposition 3 (Proposition 2 from [46]): Assume that f is σ-strongly convex and β-smooth and that $\gamma \in]0, \infty[$. Then $\mathrm{Prox}_{\gamma f} - \frac{1}{1+\gamma\beta}\mathrm{Id}$ is $\frac{1}{\frac{1}{1+\gamma\beta} - \frac{1}{1+\gamma\sigma}}$-cocoercive if $\beta > \sigma$ and 0-Lipschitz if $\beta = \sigma$. If f has no smoothness constant, the same holds by taking $\beta = +\infty$.

We will use these definitions and properties to derive the Lipschitz constants of $\tilde{\mathcal{O}}_1, \tilde{\mathcal{O}}_2$ in appendix G.

Lemma 6 (Jacobian of the Proximal): Using proposition 2, the proximal operator can be written, for any parameter $\gamma \in \mathbb{R}^+$ and $\mathbf{x}$ in the input space $\mathcal{X}$:

$$\mathrm{Prox}_{\gamma f}(\mathbf{x}) = (\mathrm{Id} + \gamma \partial \mathrm{f})^{-1}(\mathbf{x}).$$

For any convex and differentiable function f, we have:

$$\mathrm{Prox}_{\gamma f}(\mathbf{x}) + \gamma \nabla f(\mathrm{Prox}_{\gamma f}(\mathbf{x})) = \mathbf{x}.$$

For a twice differentiable f, applying the chain rule then yields:

$$\mathcal{D}_{\text{Prox}_{\gamma f}}(\mathbf{x}) + \gamma \mathcal{H}_f(\text{Prox}_{\gamma f}(\mathbf{x}))\mathcal{D}_{\text{Prox}_{\gamma f}}(\mathbf{x}) = \text{Id}$$

where $\mathcal{D}$ is the Jacobian matrix and $\mathcal{H}$ the Hessian. Since f is a convex function, its Hessian is positive semi-definite, and knowing that γ is strictly positive, the matrix $(\text{Id} + \gamma \mathcal{H}_{\text{f}}(\text{Prox}_{\gamma f}))$ is invertible. We thus have:

$$\mathcal{D}_{\text{Prox}_{\gamma f}}(\mathbf{x}) = (\text{Id} + \gamma \mathcal{H}_{\text{f}}(\text{Prox}_{\gamma f}(\mathbf{x})))^{-1}.$$

Lemma 7 (Proximal of Ridge Regularized Functions): Since we consider only separable functions, we can work with scalar version of the proximal operators. The scalar proximal of a given function with an added ridge regularization can be written:

$$\begin{aligned}\text{Prox}_{\gamma(f+\frac{\lambda_2}{2}\|.\|_2^2)}(x) &= (\text{Id} + \gamma(\partial \text{f} + \lambda_2))^{-1}(\text{x}) \\ &= ((1+\gamma\lambda_2)Id + \gamma f')^{-1}(x)\end{aligned}$$

where the second equality is true only for differentiable f. If f is real analytic, we can apply the analytic inverse function theorem [48] and verify analyticity in λ_2 of the proximal.

Finally, we remind a result from [40] describing the limiting behavior of regularized estimators for vanishing regularization.

Proposition 4 (Theorem 26.20 from [40]): Let f and h be proper, lower semi-continuous, convex functions defined on $\mathcal{X}$. Suppose that $\arg\min f \cap \text{dom}(h) \neq \emptyset$ and that h is coercive and strictly convex. Then h admits a unique minimizer $\mathbf{x}_0$ over $\arg\min f$ and, for every $\epsilon \in]0,1[$, the regularized problem

$$\arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + \epsilon h(\mathbf{x})$$

admits a unique solution $\mathbf{x}_\epsilon$. If we assume further that h is uniformly convex on any closed ball of the input space, then $\lim_{\epsilon\to 0} \mathbf{x}_\epsilon = \mathbf{x}_0$.

Appendix C
From Replica Potentials to Moreau Envelopes

Here we show how the potentials defined for the replica free energy of corollary 1 can be mapped to Moreau envelopes in the zero temperature limit, i.e. $\beta \to \infty$ where β is the inverse temperature. We consider the scalar case since the replica expressions are scalar. All functions are separable here, so any needed generalization to the multidimensional case is immediate. We start by reminding the definition of the Moreau envelope [40], [41] $\mathcal{M}_{\gamma f}$ of a proper, closed and convex function f for a given $\gamma \in \mathbb{R}_+^*$ and any $z \in \mathbb{R}$:

$$\mathcal{M}_{\gamma f}(z) = \inf_{x\in\mathbb{R}} \left\{ f(x) + (1/2\gamma)\|x - z\|_2^2 \right\}.$$

The Moreau envelope can be interpreted as a smoothed version of a given objective function with the same minimizer. For ℓ_1 minimization for example, it allows to work with a differentiable objective. By definition of the proximal operator we have the following identity:

$$\begin{aligned}\text{Prox}_{\gamma f}(z) &= \arg\min_{x\in\mathbb{R}} \left\{ f(x) + (1/2\gamma)\|x - z\|_2^2 \right\}, \\ \mathcal{M}_{\gamma f}(z) &= f(\text{Prox}_{\gamma f}(z)) + \frac{1}{2}\|\text{Prox}_{\gamma f}(z) - z\|_2^2.\end{aligned}$$

We can now match the replica potentials with the Moreau envelope. We start from the definition of said potentials, to which we apply Laplace's approximation:

$$\begin{aligned}&\phi_x(\hat{m}_{1x}, \hat{Q}_{1x}, \hat{\chi}_{1x}; x_0, \xi_{1x}) \\ &= \lim_{\beta\to\infty} \frac{1}{\beta} \log \int e^{-\frac{\beta\hat{Q}_{1x}}{2}x^2 + \beta(\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x})x - \beta f(x)} dx \\ &= -\frac{\hat{Q}_{1x}}{2}(x^*)^2 + (\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x})x^* - f(x^*)\end{aligned}$$

where

$$x^* = \arg\min_x \left\{ -\frac{\hat{Q}_{1x}}{2}x^2 + (\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x})x - f(x) \right\}.$$

This is an unconstraint convex optimization problem, thus its optimality condition is enough to characterize its set of minimizers:

$$\begin{aligned}&-\hat{Q}_{1x}x^* + (\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x}) - \partial f(x^*) = 0 \\ &\iff x^* = (Id + \frac{1}{\hat{Q}_{1x}}\partial f)^{-1}\left(\frac{\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x}}{\hat{Q}_{1x}}\right) \\ &\iff x^* = \text{Prox}_{\frac{f}{\hat{Q}_{1x}}}\left(\frac{\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x}}{\hat{Q}_{1x}}\right).\end{aligned}$$

Replacing this in the replica potential and completing the square, we get:

$$\begin{aligned}&\phi_x(\hat{m}_{1x}, \hat{Q}_{1x}, \hat{\chi}_{1x}; x_0, \xi_{1x}) \\ &= -f(\text{Prox}_{\gamma f}(X)) - \frac{\hat{Q}_{1x}}{2}\|X - \text{Prox}_{\gamma f}(X)\|_2^2 + \frac{X^2}{2}\hat{Q}_{1x} \\ &= \hat{Q}_{1x}\frac{X^2}{2} - \mathcal{M}_{\frac{1}{\hat{Q}_{1x}}f}(X)\end{aligned}$$

where we used the shorthand $X = \frac{\hat{m}_{1x}x_0 + \sqrt{\hat{\chi}_{1x}}\xi_{1x}}{\hat{Q}_{1x}}$.

Appendix D
Fixed Point of Multilayer Vector Approximate Message Passing

Here we show that the fixed point of 2-layer MLVAMP coincides with the optimality condition of the convex problem 2, proving Lemma 2. Writing the fixed point of the scalar parameters of algorithm (1), we get the following prescriptions on the scalar quantities:

$$\begin{aligned}\frac{1}{\chi_x} \equiv \frac{1}{\chi_{1x}} &= \frac{1}{\chi_{2x}} = \hat{Q}_{1x} + \hat{Q}_{2x} \\ \frac{1}{\chi_z} \equiv \frac{1}{\chi_{1z}} &= \frac{1}{\chi_{2z}} = \hat{Q}_{1z} + \hat{Q}_{2z} \qquad (17) \\ \hat{Q}_{1x}\chi_{1x} &+ \hat{Q}_{2x}\chi_{2x} = 1 \\ \hat{Q}_{1z}\chi_{1z} &+ \hat{Q}_{2z}\chi_{2z} = 1, \qquad (18)\end{aligned}$$

and the following ones on the estimates, as proved in [39] section III:

$$\begin{aligned}\hat{\mathbf{x}}_1 &= \hat{\mathbf{x}}_2 \qquad \hat{\mathbf{z}}_1 = \hat{\mathbf{z}}_2 \\ \hat{\mathbf{z}}_1 &= \mathbf{F}\hat{\mathbf{x}}_1 \qquad \hat{\mathbf{z}}_2 = \mathbf{F}\hat{\mathbf{x}}_2.\end{aligned}$$

We would like the fixed point of MLVAMP to satisfy the following first-order optimality condition

$$\partial f(\hat{\mathbf{x}}) + \mathbf{F}^T \partial g(\mathbf{F}\hat{\mathbf{x}}) = 0, \tag{19}$$

which characterizes the unique minimizer of the unconstraint convex problem (2). Replacing $\mathbf{h}_{1x}$'s expression inside $\mathbf{h}_{2x}$ reads

$$\begin{aligned}\mathbf{h}_{2x} &= \left(\frac{\hat{\mathbf{x}}_1}{\chi_x} - \hat{Q}_{1x}\mathbf{h}_{1x}\right)/\hat{Q}_{2x} \\ &= \left(\frac{\hat{\mathbf{x}}_1}{\chi_x} - \left(\frac{\hat{\mathbf{x}}_2}{\chi_x} - \hat{Q}_{2x}\mathbf{h}_{2x}\right)\right)/\hat{Q}_{2x},\end{aligned}$$

and using (17) we get $\hat{\mathbf{x}}_1 = \hat{\mathbf{x}}_2$, and a similar reasoning gives $\hat{\mathbf{z}}_2 = \hat{\mathbf{z}}_1$. From (6) and (7), we clearly find $\hat{\mathbf{z}}_2 = \mathbf{F}\hat{\mathbf{x}}_2$. Inverting the proximal operators in (4) and (5) yields

$$\begin{aligned}\hat{\mathbf{x}}_1 + \frac{1}{\hat{Q}_{1x}}\partial g(\hat{\mathbf{x}}_1) &= \mathbf{h}_{1x} \\ \hat{\mathbf{z}}_1 + \frac{1}{\hat{Q}_{1z}}\partial g(\hat{\mathbf{z}}_1) &= \mathbf{h}_{1z}.\end{aligned} \tag{20}$$

Starting from the MLVAMP equation on $\mathbf{h}_{1x}$, we write

$$\begin{aligned}\mathbf{h}_{1x} &= \left(\frac{\hat{\mathbf{x}}_2}{\chi_x} - \hat{Q}_{2x}\mathbf{h}_{2x}\right)/\hat{Q}_{1x} \\ &= \frac{\left(\frac{\hat{\mathbf{x}}_2}{\chi_x} - (\hat{Q}_{2z}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}\mathrm{Id})\hat{\mathbf{x}}_2 + \hat{Q}_{2z}\mathbf{F}^T\mathbf{h}_{2z}\right)}{\hat{Q}_{1x}} \\ &= -\frac{\left(\hat{Q}_{2z}\mathbf{F}^T\mathbf{F} + \hat{Q}_{2x}\left(1 - \frac{1}{\chi_x\hat{Q}_{2x}}\right)\mathrm{Id}\right)\hat{\mathbf{x}}_2}{\hat{Q}_{2x}} \\ &\quad + \mathbf{F}^T\left(\hat{Q}_{1z}\left(\frac{1}{\chi_z\hat{Q}_{1z}} - 1\right)\hat{\mathbf{z}}_1 - \partial\mathbf{g}(\hat{\mathbf{z}}_1)\right)\end{aligned}$$

which is equal to the left-hand term in (20). Using this equality, as well as $\hat{\mathbf{z}}_1 = \mathbf{F}\hat{\mathbf{x}}_2$ and relations (17) and (18) yields

$$\partial f(\hat{\mathbf{x}}_2) + \mathbf{F}^T\partial g(\mathbf{F}\hat{\mathbf{x}}_2) = 0.$$

Hence, the fixed point of MLVAMP satisfies the optimality condition (19) and is indeed the desired estimator: $\hat{\mathbf{x}}_1 = \hat{\mathbf{x}}_2 = \hat{\mathbf{x}}$.

APPENDIX E
STATE EVOLUTION EQUATIONS

This appendix is intended mainly for completeness, to show that the fixed point equations from Theorem 1, stemming from the heuristic state evolution written in [35] are indeed made rigorous by the results presented in [36].

A. Heuristic State Evolution Equations

The state evolution equations track the evolution of MLVAMP (1) and provide statistical properties of its iterates. They are derived in [35] taking the heuristic assumption that $\mathbf{h_{1x}}, \mathbf{h_{1z}}, \mathbf{h_{2x}}, \mathbf{h_{2z}}$ behave as Gaussian estimates, which comes from the physics cavity approach:

$$\begin{aligned}\hat{Q}_{1x}^{(t)}\mathbf{h}_{1x}^{(t)} - \hat{m}_{1x}^{(t)}\mathbf{x_0} &\overset{PL2}{=} \sqrt{\hat{\chi}_{1x}^{(t)}}\boldsymbol{\xi}_{1x}^{(t)} \\ \mathbf{V}^T(\hat{Q}_{2x}^{(t)}\mathbf{h}_{2x}^{(t)} - \hat{m}_{2x}^{(t)}\mathbf{x_0}) &\overset{PL2}{=} \sqrt{\hat{\chi}_{2x}^{(t)}}\boldsymbol{\xi}_{2x}^{(t)} \\ \mathbf{U}^T(\hat{Q}_{1z}^{(t)}\mathbf{h}_{1z}^{(t)} - \hat{m}_{1z}^{(t)}\mathbf{z_0}) &\overset{PL2}{=} \sqrt{\hat{\chi}_{1z}^{(t)}}\boldsymbol{\xi}_{1z}^{(t)} \\ \hat{Q}_{2z}^{(t)}\mathbf{h}_{2z}^{(t)} - \hat{m}_{2z}^{(t)}\mathbf{z_0} &\overset{PL2}{=} \sqrt{\hat{\chi}_{2z}^{(t)}}\boldsymbol{\xi}_{2z}^{(t)}\end{aligned} \tag{21}$$

where $\overset{PL2}{=}$ denotes $PL2$ convergence. $\mathbf{U}$ and $\mathbf{V}$ come from the singular value decomposition $\mathbf{F} = \mathbf{UDV}^T$ and are Haar-sampled; $\xi_{1x}^{(t)}, \xi_{2x}^{(t)}, \xi_{1z}^{(t)}, \xi_{2z}^{(t)}$ are normal Gaussian vectors, independent from $\mathbf{x_0}, \mathbf{z_0}, \mathbf{V}^T\mathbf{x_0}$ and $\mathbf{U}^T\mathbf{z_0}$. Parameters $\hat{Q}_{1x}^{(t)}, \hat{Q}_{1z}^{(t)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t)}$ are defined through MLVAMP's iterations (1); while parameters $\hat{m}_{1x}^{(t)}, \hat{m}_{1z}^{(t)}, \hat{m}_{2x}^{(t)}, \hat{m}_{2z}^{(t)}$ and $\hat{\chi}_{1x}^{(t)}, \hat{\chi}_{1z}^{(t)}, \hat{\chi}_{2x}^{(t)}, \hat{\chi}_{2z}^{(t)}$ are prescribed through SE equations. Other useful variables are the overlaps and squared norms of estimators, for $k \in \{1, 2\}$:

$$\begin{aligned}m_{kx}^{(t)} &= \frac{\mathbf{x}_0^\top\hat{\mathbf{x}}_k^{(t)}}{N} \qquad q_{kx}^{(t)} = \frac{\|\hat{\mathbf{x}}_k^{(t)}\|_2^2}{N} \\ m_{kz}^{(t)} &= \frac{\mathbf{z}_0^\top\hat{\mathbf{z}}_k^{(t)}}{M} \qquad q_{kz}^{(t)} = \frac{\|\hat{\mathbf{z}}_k^{(t)}\|_2^2}{M}.\end{aligned}$$

Starting from assumptions (21), and following the derivation of [35] adapted to the iteration order from (1), the heuristic state evolution equations read:

$$\text{Initialize} \quad \hat{Q}_{1x}^{(0)}, \hat{Q}_{2z}^{(0)}, \hat{m}_{1x}^{(0)}, \hat{m}_{2z}^{(0)}, \hat{\chi}_{1x}^{(0)}, \hat{\chi}_{2z}^{(0)} > 0.$$

$$m_{1x}^{(t)} = \mathbb{E}\left[x_0\eta_{f/\hat{Q}_{1x}^{(t)}}\left(\frac{\hat{m}_{1x}^{(t)}x_0 + \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_{1x}^{(t)}}{\hat{Q}_{1x}^{(t)}}\right)\right] \tag{22a}$$

$$\chi_{1x}^{(t)} = \frac{1}{\hat{Q}_{1x}^{(t)}}\mathbb{E}\left[\eta'_{f/\hat{Q}_{1x}^{(t)}}\left(\frac{\hat{m}_{1x}^{(t)}x_0 + \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_{1x}^{(t)}}{\hat{Q}_{1x}^{(t)}}\right)\right] \tag{22b}$$

$$q_{1x}^{(t)} = \mathbb{E}\left[\eta^2_{f/\hat{Q}_{1x}^{(t)}}\left(\frac{\hat{m}_{1x}^{(t)}x_0 + \sqrt{\hat{\chi}_{1x}^{(t)}}\xi_{1x}^{(t)}}{\hat{Q}_{1x}^{(t)}}\right)\right] \tag{22c}$$

$$\hat{Q}_{2x}^{(t)} = \frac{1}{\chi_{1x}^{(t)}} - \hat{Q}_{1x}^{(t)} \tag{22d}$$

$$\hat{m}_{2x}^{(t)} = \frac{m_{1x}^{(t)}}{\rho_x\chi_{1x}^{(t)}} - \hat{m}_{1x}^{(t)} \tag{22e}$$

$$\hat{\chi}_{2x}^{(t)} = \frac{q_{1x}^{(t)}}{(\chi_{1x}^{(t)})^2} - \frac{(m_{1x}^{(t)})^2}{\rho_x(\chi_{1x}^{(t)})^2} - \hat{\chi}_{1x}^{(t)} \tag{22f}$$

$$m_{2z}^{(t)} = \frac{\rho_x}{\alpha}\mathbb{E}\left[\frac{\lambda(\hat{m}_{2x}^{(t)} + \lambda\hat{m}_{2z}^{(t)})}{\hat{Q}_{2x}^{(t)} + \lambda\hat{Q}_{2z}^{(t)}}\right] \tag{22g}$$

$$\chi_{2z}^{(t)} = \frac{1}{\alpha}\mathbb{E}\left[\frac{\lambda}{\hat{Q}_{2x}^{(t)} + \lambda\hat{Q}_{2z}^{(t)}}\right] \tag{22h}$$

$$\begin{aligned}q_{2z}^{(t)} = \frac{1}{\alpha}\mathbb{E}&\left[\frac{\lambda(\hat{\chi}_{2x}^{(t)} + \lambda\hat{\chi}_{2z}^{(t)})}{(\hat{Q}_{2x}^{(t)} + \lambda\hat{Q}_{2z}^{(t)})^2}\right] \\ &+ \frac{\rho_x}{\alpha}\mathbb{E}\left[\frac{\lambda(\hat{m}_{2x}^{(t)} + \lambda\hat{m}_{2z}^{(t)})^2}{(\hat{Q}_{2x}^{(t)} + \lambda\hat{Q}_{2z}^{(t)})^2}\right]\end{aligned} \tag{22i}$$

$$\hat{Q}_{1z}^{(t)} = \frac{1}{\chi_{2z}^{(t)}} - \hat{Q}_{2z}^{(t)} \tag{22j}$$

$$\hat{m}_{1z}^{(t)} = \frac{m_{2z}^{(t)}}{\rho_z \chi_{2z}^{(t)}} - \hat{m}_{2z}^{(t)} \tag{22k}$$

$$\hat{\chi}_{1z}^{(t)} = \frac{q_{2z}^{(t)}}{(\chi_{2z}^{(t)})^2} - \frac{(m_{2z}^{(t)})^2}{\rho_z (\chi_{2z}^{(t)})^2} - \hat{\chi}_{2z}^{(t)} \tag{22l}$$

$$m_{1z}^{(t)} = \mathbb{E}\left[z_0 \eta_{g(y,.)/\hat{Q}_{1z}^{(t)}} \left(\frac{\hat{m}_{1z}^{(t)} z_0 + \sqrt{\hat{\chi}_{1z}^{(t)}} \xi_{1z}^{(t)}}{\hat{Q}_{1z}^{(t)}} \right) \right] \tag{22m}$$

$$\chi_{1z}^{(t)} = \frac{1}{\hat{Q}_{1z}^{(t)}} \mathbb{E}\left[\eta'_{g(y,.)/\hat{Q}_{1z}^{(t)}} \left(\frac{\hat{m}_{1z}^{(t)} z_0 + \sqrt{\hat{\chi}_{1z}^{(t)}} \xi_{1z}^{(t)}}{\hat{Q}_{1z}^{(t)}} \right) \right] \tag{22n}$$

$$q_{1z}^{(t)} = \mathbb{E}\left[\eta^2_{g(y,.)/\hat{Q}_{1z}^{(t)}} \left(\frac{\hat{m}_{1z}^{(t)} z_0 + \sqrt{\hat{\chi}_{1z}^{(t)}} \xi_{1z}^{(t)}}{\hat{Q}_{1z}^{(t)}} \right) \right] \tag{22o}$$

$$\hat{Q}_{2z}^{(t+1)} = \frac{1}{\chi_{1z}^{(t)}} - \hat{Q}_{1z}^{(t)} \tag{22p}$$

$$\hat{m}_{2z}^{(t+1)} = \frac{m_{1z}^{(t)}}{\rho_z \chi_{1z}^{(t)}} - \hat{m}_{1z}^{(t)} \tag{22q}$$

$$\hat{\chi}_{2z}^{(t+1)} = \frac{q_{1z}^{(t)}}{(\chi_{1z}^{(t)})^2} - \frac{(m_{1z}^{(t)})^2}{\rho_z (\chi_{1z}^{(t)})^2} - \hat{\chi}_{1z}^{(t)} \tag{22r}$$

$$m_{2x}^{(t+1)} = \rho_x \mathbb{E}\left[\frac{\hat{m}_{2x}^{(t)} + \lambda \hat{m}_{2z}^{(t+1)}}{\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t+1)}} \right] \tag{22s}$$

$$\chi_{2x}^{(t+1)} = \mathbb{E}\left[\frac{1}{\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t+1)}} \right] \tag{22t}$$

$$q_{2x}^{(t+1)} = \mathbb{E}\left[\frac{\hat{\chi}_{2x}^{(t)} + \lambda \hat{\chi}_{2z}^{(t+1)}}{(\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t+1)})^2} \right] + \rho_x \mathbb{E}\left[\frac{(\hat{m}_{2x}^{(t+1)} + \lambda \hat{m}_{2z}^{(t+1)})^2}{(\hat{Q}_{2x}^{(t)} + \lambda \hat{Q}_{2z}^{(t+1)})^2} \right] \tag{22u}$$

$$\hat{Q}_{1x}^{(t+1)} = \frac{1}{\chi_{2x}^{(t+1)}} - \hat{Q}_{2x}^{(t)} \tag{22v}$$

$$\hat{m}_{1x}^{(t+1)} = \frac{m_{2x}^{(t+1)}}{\rho_x \chi_{2x}^{(t+1)}} - \hat{m}_{2x}^{(t)} \tag{22w}$$

$$\hat{\chi}_{1x}^{(t+1)} = \frac{q_{2x}^{(t+1)}}{(\chi_{2x}^{(t+1)})^2} - \frac{(m_{2x}^{(t+1)})^2}{\rho_x (\chi_{2x}^{(t+1)})^2} - \hat{\chi}_{2x}^{(t)}. \tag{22x}$$

We are interested in the fixed point of these state evolution equations, where $\chi_{1x}^{(t)} = \chi_{2x}^{(t)} = \chi_x$, $q_{1x}^{(t)} = q_{2x}^{(t)} = q_x$, $m_{1x}^{(t)} = m_{2x}^{(t)} = m_x$, $\chi_{1z}^{(t)} = \chi_{2z}^{(t)} = \chi_z$, $q_{1z}^{(t)} = q_{2z}^{(t)} = q_z$, and $m_{1z}^{(t)} = m_{2z}^{(t)} = m_z$ are achieved. From there we easily recover eq. (9). However, these equations are not rigorous since the starting assumptions are not proven. Therefore, we will turn to a rigorous formalism to consolidate those results.

B. Necessary Assumptions for the Rigorous State Evolution Equations

Here we remind the main assumptions needed for the rigorous state evolution equations to hold, as they are listed for Theorem 1 of [36], and show they are verified in our setting.

Assumption 3:

- the empirical distributions of the underlying truth $\mathbf{x}_0$, eigenvalues of $\mathbf{F}^T\mathbf{F}$, and noise vector w_0, respectively converge with second order moments, as defined in appendix A, to independent scalar random variables x_0, w_0, λ with distributions p_{x_0}, p_λ, p_{w_0}. We assume that the distribution p_λ is not all-zero and has compact support.
- the design matrix $\mathbf{F} = \mathbf{U}\mathbf{D}\mathbf{V}^\top \in \mathbb{R}^{M\times N}$ is rotationally invariant, as defined in the introduction, where the elements of the Haar distributed matrices $\mathbf{U}, \mathbf{V}$ are independent of the random variables x_0, w_0, λ
- assume that $M, N \to \infty$ with fixed ratio $\alpha = M/N$ independent of M, N.
- the activation function $\phi(., \mathbf{w}_0)$ from Eq.(1) is pseudo-Lipschitz of order 2.
- the constants $\left\langle \partial_{\mathbf{h}_{1x}^{(t)}} g_{1x}(\mathbf{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)}) \right\rangle$, $\left\langle \partial_{\mathbf{h}_{1z}^{(t)}} g_{1z}(\mathbf{h}_{1z}^{(t)}, \hat{Q}_{1z}^{(t)}) \right\rangle$ $\left\langle \partial_{\mathbf{h}_{2x}^{(t)}} g_{2x}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t+1)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t+1)}) \right\rangle$ $\left\langle \partial_{\mathbf{h}_{2x}^{(t)}} g_{2z}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t)}) \right\rangle$ from algorithm (1) are all in $[0, 1]$.
- the component estimation functions $g_{1x}(\mathbf{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)})$, $g_{1z}(\mathbf{h}_{1z}^{(t)}, \hat{Q}_{1z}^{(t)}), g_{2x}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t+1)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t+1)})$, $g_{2z}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t)})$ from algorithm (1) are uniformly Lipschitz continuous, at all time steps t, respectively in $\mathbf{h}_{1x}^{(t)}$ at $\hat{Q}_{1x}^{(t)}$, in $\mathbf{h}_{1z}^{(t)}$ at $\hat{Q}_{1z}^{(t)}$, $\mathbf{h}_{2x}^{(t)}$ at $\hat{Q}_{2x}^{(t)}$ and in $\mathbf{h}_{2z}^{(t)}$ at $\hat{Q}_{2z}^{(t)}$.

The first four points are included in the set of assumptions 1 and are therefore verified. We need to check the last two points, starting with the function $g_{1x}(\mathbf{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)}) = \text{Prox}_{f/\hat{Q}_{1x}^{(t)}}(\mathbf{h}_{1x}^{(t)})$. Since proximal operators are firmly nonexpansive, they are 1-Lipschitz and we thus have, using the separability of the function f:

$$\left\langle \partial_{\mathbf{h}_{1x}^{(t)}} g_{1x}(\mathbf{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)}) \right\rangle = \frac{1}{N} \sum_{i=1}^{N} \text{Prox}'_{f_i/\hat{Q}_{1x}^{(t)}}(\mathbf{h}_{1x,i}^{(t)}) \in [0, 1]$$

where each $f_i : \mathbb{R} \to \mathbb{R}$ is the same function applied to each coordinates. Now consider the restriction of $g_{1x}(\mathbf{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)})$ to its second argument. Its gradient w.r.t. $\hat{Q}_{1x}^{(t)}$ at a given point $\mathbf{h}_{1x}^{(t)}$ verifies, assuming the function f is differentiable:

$$\begin{aligned} &\|\nabla_{\hat{Q}_{1x}^{(t)}} \text{Prox}_{f/\hat{Q}_{1x}^{(t)}}(\mathbf{h}_{1x}^{(t)})\|_2 \\ &\quad = \|(Id + \frac{1}{\hat{Q}_{1x}^{(t)}} \mathcal{H}_f(\text{Prox}_{f/\hat{Q}_{1x}^{(t)}}(\mathbf{h}_{1x}^{(t)})))^{-1} \nabla f(\mathbf{h}_{\mathbf{1x}}^{(\mathbf{t})})\|_2 \\ &\quad \leqslant \|\nabla f(\mathbf{h}_{1x}^{(t)})\|_2 \\ &\quad \leqslant C(1 + \|\mathbf{h}_{1x}^{(t)}\|_2) \end{aligned}$$

where the last line is obtained using the scaling conditions on the subdifferential of f from assumption 1. Then, for any $\hat{Q}_{1x}^{(t)}, \hat{Q}_{1x}^{(t')}$,

$$\|\text{Prox}_{f/\hat{Q}_{1x}^{(t)}} - \text{Prox}_{f/\hat{Q}_{1x}^{(t')}}\|_2 \leqslant C(1 + \|\mathbf{h}_{1x}^{(t)}\|_2)|\hat{Q}_{1x}^{(t)} - \hat{Q}_{1x}^{(t')}|$$

and $g_{1x}(\mathbf{h}_{1x}^{(t)}, \hat{Q}_{1x}^{(t)})$ is uniformly Lipschitz in $\mathbf{h}_{1x}^{(t)}$ at $\hat{Q}_{1x}^{(t)}$, at any time index t. The argument is identical for $g_{1z}(\mathbf{h}_{1z}^{(t)}, \hat{Q}_{1z}^{(t)}) = \text{Prox}_{f/\hat{Q}_{1z}^{(t)}}(\mathbf{h}_{1z}^{(t)})$. The functions

$g_{2x}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t+1)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t+1)}), g_{2z}(\mathbf{h}_{2x}^{(t)}, \mathbf{h}_{2z}^{(t)}, \hat{Q}_{2x}^{(t)}, \hat{Q}_{2z}^{(t)})$ have explicit expressions and it is straightforward to check the last two points using linear algebra and the assumptions on the spectrum of $\mathbf{F}^\top\mathbf{F}$.

C. Rigorous State Evolution Formalism

We now look into the state evolution equations derived for MLVAMP in [14]. Those equations are proven to be exact in the asymptotic limit, and follow the same algorithm as (1). In particular, they provide statistical properties of vectors $\mathbf{h}_{1x}, \mathbf{h}_{2x}, \mathbf{h}_{1z}, \mathbf{h}_{2z}$. We can read relations from [36] using the following dictionary between our notations and theirs, valid at each iteration of the algorithm:

$$\begin{aligned}
\hat{Q}_{1x}, \hat{Q}_{2x}, \hat{Q}_{1z}, \hat{Q}_{2z} &\longleftrightarrow \gamma_0^-, \gamma_0^+, \gamma_1^+, \gamma_1^- \\
\chi_{1x}\hat{Q}_{1x}, \chi_{2x}\hat{Q}_{2x} &\longleftrightarrow \alpha_0^-, \alpha_0^+ \\
\chi_{1z}\hat{Q}_{1z}, \chi_{2z}\hat{Q}_{2z} &\longleftrightarrow \alpha_1^-, \alpha_1^+ \\
\mathbf{x_0}, \mathbf{z_0}, \rho_x, \rho_z &\longleftrightarrow \mathbf{Q}_0^0, \mathbf{Q}_1^0, \tau_0^0, \tau_1^0 \\
\mathbf{h}_{1x}, \mathbf{h}_{2x}, \mathbf{h}_{1z}, \mathbf{h}_{2z} &\longleftrightarrow \mathbf{r}_0^-, \mathbf{r}_0^+, \mathbf{r}_1^+, \mathbf{r}_1^-.
\end{aligned} \tag{23}$$

Placing ourselves in the asymptotic limit, [36] shows the following equalities:

$$\begin{aligned}
\mathbf{r}_0^- &= \mathbf{Q}_0^0 + \mathbf{Q}_0^- \\
\mathbf{r}_0^+ &= \mathbf{Q}_0^0 + \mathbf{Q}_0^+ \\
\mathbf{r}_1^- &= \mathbf{Q}_1^0 + \mathbf{Q}_1^- \\
\mathbf{r}_1^+ &= \mathbf{Q}_1^0 + \mathbf{Q}_1^+
\end{aligned} \tag{24}$$

where $\mathbf{Q}_0^- \sim \mathcal{N}(0, \tau_0^-)^N$ and $\mathbf{Q}_1^- \sim \mathcal{N}(0, \tau_1^-)^N$ are i.i.d. Gaussian vectors. $\mathbf{Q}_0^+$, $\mathbf{Q}_1^+$ have the following norms and non-zero correlations with ground-truth vectors $\mathbf{Q}_0^0, \mathbf{Q}_1^0$:

$$\begin{aligned}
\tau_0^+ &\equiv \frac{\|\mathbf{Q}_0^+\|_2^2}{N} & c_0^+ &\equiv \frac{\mathbf{Q}_0^{0T}\mathbf{Q}_0^+}{N} \\
\tau_1^+ &\equiv \frac{\|\mathbf{Q}_1^+\|_2^2}{M} & c_1^+ &\equiv \frac{\mathbf{Q}_1^{0T}\mathbf{Q}_1^+}{M}.
\end{aligned}$$

With simple manipulations, we can rewrite (24) as:

$$\begin{aligned}
\mathbf{r}_0^- &\overset{\mathrm{d}}{=} \mathbf{Q}_0 + \mathbf{Q}_0^- \\
\mathbf{V}^T\mathbf{r}_0^+ &\overset{\mathrm{d}}{=} \left(1 + \frac{c_0^+}{\tau_0^0}\right)\mathbf{V}^T\mathbf{Q}_0^0 + \mathbf{V}^T\tilde{\mathbf{Q}}_0^+ \\
\mathbf{r}_1^- &\overset{\mathrm{d}}{=} \mathbf{Q}_1^0 + \mathbf{Q}_1^- \\
\mathbf{U}^T\mathbf{r}_1^+ &\overset{\mathrm{d}}{=} \left(1 + \frac{c_1^+}{\tau_1^0}\right)\mathbf{U}^T\mathbf{Q}_1^0 + \mathbf{U}^T\tilde{\mathbf{Q}}_1^+
\end{aligned} \tag{25}$$

where for $k \in \{1, 2\}$ vectors

$$\tilde{\mathbf{Q}}_k^+ = -\frac{c_k^+}{\tau_k^0}\mathbf{Q}_k^0 + \mathbf{Q}_k^+$$

and $\mathbf{Q}_0^-, \mathbf{Q}_1^-$ have no correlation with ground-truth vectors $\mathbf{Q}_0^0$, $\mathbf{Q}_1^0$, $\mathbf{U}^T\mathbf{Q}_0^0$, $\mathbf{V}^T\mathbf{Q}_1^0$. Besides, Lemma 5 from [34] states that $\mathbf{V}^T\tilde{\mathbf{Q}}_0^+$ and $\mathbf{U}^T\tilde{\mathbf{Q}}_1^+$ have components that converge empirically to Gaussian variables, respectively $\mathcal{N}(0, \tau_0^+)$ and $\mathcal{N}(0, \tau_1^+)$. Let us now translate this in our own terms, using the following relations that complete our dictionary with state evolution parameters:

$$\begin{aligned}
\frac{\hat{m}_{1x}}{\hat{Q}_{1x}} &\longleftrightarrow 1 & \frac{\hat{m}_{2z}}{\hat{Q}_{2z}} &\longleftrightarrow 1 \\
\frac{\hat{m}_{2x}}{\hat{Q}_{2x}} &\longleftrightarrow 1 + \frac{c_0^+}{\tau_0^0} & \frac{\hat{m}_{1z}}{\hat{Q}_{1z}} &\longleftrightarrow 1 + \frac{c_1^+}{\tau_1^0} \\
\frac{\hat{\chi}_{1x}}{\hat{Q}_{1x}^2} &\longleftrightarrow \tau_0^- & \frac{\hat{\chi}_{2z}}{\hat{Q}_{2z}^2} &\longleftrightarrow \tau_1^- \\
\frac{\hat{\chi}_{2x}}{\hat{Q}_{2x}^2} &\longleftrightarrow \tau_0^+ - \frac{(c_0^+)^2}{\tau_0^0} & \frac{\hat{\chi}_{1z}}{\hat{Q}_{1z}^2} &\longleftrightarrow \tau_1^+ - \frac{(c_1^+)^2}{\tau_1^0}.
\end{aligned} \tag{26}$$

Simple bookkeeping transforms equations (25) into a rigorous statement of starting assumptions (24) from [35]. Since those assumptions are now rigorously established in the asymptotic limit, the remaining derivation of state evolution equations (22) holds and provides a mathematically exact statement.

D. Scalar Equivalent Model of State Evolution

For the sake of completeness, we will provide an overview of the explicit matching between the state evolution formalism from [36] which was developed in a series of papers, and the replica formulation from [35] which relies on statistical physics methods. Although not necessary to our proof, it is interesting to develop an intuition about the correspondence between those two faces of the same coin. We have seen in the previous subsection that [36] introduces ground-truth vectors $\mathbf{Q}_0^0, \mathbf{Q}_1^0$, estimates $\mathbf{r}_0^\pm, \mathbf{r}_1^\pm$ which are related to vectors $\mathbf{Q}_0^\pm, \mathbf{Q}_1^\pm$. Let us introduce a few more vectors using matrices from the singular value decomposition $\mathbf{F} = \mathbf{U}\mathbf{D}\mathbf{V}^T$. Let $\mathbf{s}_\nu \in \mathbb{R}^N$ be the vector containing all square roots of eigenvalues of $\mathbf{F}^T\mathbf{F}$ with p_ν its element-wise distribution; and $\mathbf{s}_\mu \in \mathbb{R}^M$ the vector containing all square roots of eigenvalues of $\mathbf{F}\mathbf{F}^T$ with p_μ its element-wise distribution. Note that those two vectors contain the singular values of $\mathbf{F}$, but one of them also contains $\max(M, N) - \min(M, N)$ zero values. p_μ and p_ν are both well-defined since p_λ is properly defined in Assumptions 1. We also define

$$\begin{aligned}
\mathbf{P}_0^0 &= \mathbf{V}^T\mathbf{Q}_0^0 & \mathbf{P}_0^+ &= \mathbf{V}^T\mathbf{Q}_0^+ & \mathbf{P}_0^- &= \mathbf{V}^T\mathbf{Q}_0^- \\
\mathbf{P}_1^0 &= \mathbf{U}\mathbf{Q}_1^0 & \mathbf{P}_1^+ &= \mathbf{U}\mathbf{Q}_1^+ & \mathbf{P}_1^- &= \mathbf{U}\mathbf{Q}_1^-.
\end{aligned}$$

By virtue of Lemma 5 from [34], the six previous vectors have elements that converge empirically to a Gaussian variable. Hence, all defined vectors have an element-wise separable distribution, and we can write the state evolution as a scalar model on random variables sampled from those distributions. To do so, we will simply write the variables without the bold font: for instance $Z_0^0 \sim p_{x_0}$, $s_\nu \sim p_\nu$, and Q_0^- refers to the random variable distributed according to the element-wise distribution of vector $\mathbf{Q}_0^-$. The scalar random variable state evolution from [36] now reads:

$$\text{Initialize } \gamma_1^{-(0)}, \gamma_0^{-(0)}, \tau_0^{-(0)}, \tau_1^{-(0)}, \tag{27a}$$
$$Q_0^{-(0)} \sim \mathcal{N}(0, \tau_0^{-(0)}), Q_1^{-(0)} \sim \mathcal{N}(0, \tau_1^{-(0)}), \alpha_0^{-(0)}, \alpha_1^{-(0)}$$
$$\textit{Initial pass (ground truth only)}$$
$$s_\nu \sim p_\nu, \qquad s_\mu \sim p_\mu, \qquad Q_0^0 \sim p_{x_0} \tag{27b}$$

$$\tau_0^0 = \mathbb{E}[(Q_0^0)^2] \qquad P_0^0 \sim \mathcal{N}(0, \tau_0^0) \tag{27c}$$

$$Q_1^0 = s_\mu P_0^0 \qquad \tau_1^0 = \mathbb{E}[(s_\mu P_0^0)^2] = \mathbb{E}[(s_\mu)^2]\tau_0^0 \tag{27d}$$

$$P_1^0 \sim \mathcal{N}(0, \tau_1^0) \tag{27e}$$

Forward Pass (estimation):

$$\alpha_0^{+(t)} = \mathbb{E}\left[\eta'_{f/\gamma_0^{-(t)}}(Q_0^0 + Q_0^{-(t)})\right] \tag{27f}$$

$$\gamma_0^{+(t)} = \frac{\gamma_0^{(t)}}{\alpha_0^{+(t)}} - \gamma_0^{-(t)} \tag{27g}$$

$$Q_0^{+(t)} = \frac{1}{1-\alpha_0^{+(t)}}\Big\{\eta_{f/\gamma_0^{-(t)}}(Q_0^0 + Q_0^{-(t)}) - Q_0^0 - \alpha_0^+ Q_0^{-(t)}\Big\} \tag{27h}$$

$$\mathbf{K}_0^{+(t)} = Cov\left(Q_0^0, Q_0^{+(t)}\right) \tag{27i}$$

$$\left(P_0^0, P_0^{+(t)}\right) \sim \mathcal{N}\left(0, \mathbf{K}_0^{+(t)}\right) \tag{27j}$$

$$\alpha_1^{+(t)} = \mathbb{E}\left[\frac{s_\mu^2 \gamma_1^{-(t)}}{\gamma_1^{-(t)} s_\mu^2 + \gamma_0^{+(t)}}\right] \tag{27k}$$

$$\gamma_1^{+(t)} = \frac{\gamma_1^{-(t)}}{\alpha_1^{+(t)}} - \gamma_1^{-(t)} \tag{27l}$$

$$Q_1^{+(t)} = \frac{1}{1-\alpha_1^{+(t)}}\Big\{\frac{s_\mu^2\gamma_1^{-(t)}}{\gamma_1^{-(t)} s_\mu^2 + \gamma_0^{+(t)}}(Q_1^{-(t)} + Q_1^0) + \frac{s_\mu \gamma_0^{+(t)}}{\gamma_1^{-(t)} s_\mu^2 + \gamma_0^{+(t)}}(P_0^{+(t)} + P_0^0) - Q_1^0 - \alpha_1^{+(t)} Q_1^{-(t)}\Big\} \tag{27m}$$

$$\mathbf{K}_1^{+(t)} = Cov\left(Q_1^0, Q_1^{+(t)}\right) \tag{27n}$$

$$\left(P_1^0, P_1^{+(t)}\right) \sim \mathcal{N}\left(0, \mathbf{K}_1^{+(t)}\right) \tag{27o}$$

Backward Pass (estimation):

$$\alpha_1^{-(t+1)} = \mathbb{E}\left[\eta_{g(y,.)/\gamma_1^{+(t)}}(P_1^0 + P_1^{+(t)})\right] \tag{27p}$$

$$\gamma_1^{-(t+1)} = \frac{\gamma_1^{+(t)}}{\alpha_1^{-(t+1)}} - \gamma_1^{+(t)} \tag{27q}$$

$$P_1^{-(t+1)} = \frac{1}{1-\alpha_1^{-(t+1)}}\Big\{\eta_{g(y,.)/\gamma_1^{+(t)}}(P_1^0 + P_1^{+(t)}) - P_1^0 - \alpha_1^{-(t+1)} P_1^{+(t)}\Big\} \tag{27r}$$

$$\tau_1^{-(t+1)} = \mathbb{E}\left[(P_1^{-(t+1)})^2\right] \tag{27s}$$

$$Q_1^{-(t+1)} \sim \mathcal{N}(0, \tau_1^{-(t+1)}) \tag{27t}$$

$$\alpha_0^{-(t+1)} = \mathbb{E}\left[\frac{\gamma_0^{+(t)}}{\gamma_1^{-(t+1)} s_\nu^2 + \gamma_0^{+(t)}}\right] \tag{27u}$$

$$\gamma_0^{-(t+1)} = \frac{\gamma_0^{+(t)}}{\alpha_0^{-(t+1)}} - \gamma_0^{+(t)} \tag{27v}$$

$$P_0^{-(t+1)} = \frac{1}{1-\alpha_0^{-(t+1)}}\Big\{\frac{s_\nu \gamma_1^{-(t)}}{\gamma_1^{-(t+1)} s_\nu^2 + \gamma_0^{+(t)}}(Q_1^{-(t+1)} + Q_1^0) + \frac{\gamma_0^{+(t)}}{\gamma_1^{-(t+1)} s_\nu^2 + \gamma_0^{+(t)}}(P_0^{+(t)} + P_0^0) - P_0^0 - \alpha_0^{-(t+1)} P_0^{+(t)}\Big\} \tag{27w}$$

$$\tau_0^{-(t+1)} = \mathbb{E}\left[(P_0^{-(t+1)})^2\right] \tag{27x}$$

$$Q_0^{-(t+1)} \sim \mathcal{N}(0, \tau_0^{-(t+1)}). \tag{27y}$$

E. Direct Matching of the State Evolution Fixed Point Equations

To be consistent, we should be able to show that equations (27) allow us to recover equations (22) at their fixed point. Although somewhat tedious, this task is facilitated using dictionaries (23) and (26). We shall give here an overview of this matching through a few examples.

- Recovering Eq. (22e)

Let us start from the rigorous scalar state evolution, in particular Eq. (27h) that defines variable Q_0^+. We get rid of time indices here since we focus on the fixed point. We first compute the correlation

$$\begin{aligned} c_0^+ &= \mathbb{E}\left[Q_0^0 Q_0^+\right] \\ &= \frac{1}{1-\alpha_0^+}\left\{\mathbb{E}\left[Q_0^0 \eta_{f/\gamma_0^-}(Q_0^0 + Q_0^-)\right] - \tau_0^0\right\} \end{aligned} \tag{28}$$

where we have used $\mathbb{E}[(Q_0^0)^2] = \tau_0^0$. At the fixed point, we know from MLVAMP or simply translating equations (17), (18) that

$$1-\alpha_0^+ = \alpha_0^-, \qquad \frac{1}{\alpha_0^-} = \frac{\gamma_0^- + \gamma_0^+}{\gamma_0^+}, \qquad \gamma_0^+\alpha_0^+ = \gamma_0^-\alpha_0^-.$$

Simple manipulations take us to

$$\begin{aligned} c_0^+ &= \frac{\mathbb{E}\left[Q_0^0 \eta_{f/\gamma_0^-}(Q_0^0 + Q_0^-)\right]}{\alpha_0^-} - \tau_0^0\left(1 + \frac{\gamma_0^-}{\gamma_0^+}\right) \\ \left(1 + \frac{c_0^+}{\tau_0^0}\right)\gamma_0^+ &= \frac{\mathbb{E}\left[Q_0^0 \eta_{f/\gamma_0^-}(Q_0^0 + Q_0^-)\right]\gamma_0^+}{\tau_0^0 \alpha_0^-} - \gamma_0^-. \end{aligned} \tag{29}$$

Now let us translate this back into our notations. The term $\mathbb{E}\left[Q_0^0 \eta_{f/\gamma_0^-}(Q_0^0 + Q_0^-)\right]$ simply translates into m_{1x}, and the rest of the terms can all be changed according to our dictionary. (29) exactly becomes

$$\hat{m}_{2x} = \frac{m_{1x}}{\rho_x \chi_x} - \hat{m}_{1x},$$

hence we perfectly recover equations (22e) at the fixed point.

- Recovering Eq. (22f)

We start again from (27h) and square it:

$$\begin{aligned} \mathbb{E}\left[(Q_0^+)^2\right] &= \frac{1}{(1-\alpha_0^+)^2}\Big\{\mathbb{E}\left[\eta^2_{f/\gamma_0^-}(Q_0^0 + Q_0^-)\right] \\ &+ (\alpha_0^+)^2 \mathbb{E}\left[(Q_0^-)^2\right] - 2\mathbb{E}\left[Q_0^0 \eta_{f/\gamma_0^-}(Q_0^0 + Q_0^-)\right] \\ &- 2\alpha_0^+ \mathbb{E}\left[Q_0^- \eta^2_{f/\gamma_0^-}(Q_0^0 + Q_0^-) + \mathbb{E}\left[(Q_0^0)^2\right]\right]\Big\} \\ \tau_0^+ &= \frac{1}{(1-\alpha_0^+)^2}\Big\{\mathbb{E}\left[\eta^2_{f/\gamma_0^-}(Q_0^0 + Q_0^-)\right] + \tau_0^0 \\ &+ (\alpha_0^+)^2 \tau_0^- - 2\mathbb{E}\left[Q_0^0 \eta_{f/\gamma_0^-}(Q_0^0 + Q_0^-)\right] \\ &- 2\alpha_0^+ \mathbb{E}\left[Q_0^- \eta^2_{f/\gamma_0^-}(Q_0^0 + Q_0^-)\right]\Big\}. \end{aligned} \tag{30}$$

Since Q_0^- is a Gaussian variable, independent from Q_0^0, we can use Stein's lemma and use Eq. (27f) to get

$$\mathbb{E}\left[Q_0^- \eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right] = \alpha_0^+\tau_0^-. \tag{31}$$

Moreover, from (28) we have

$$\begin{aligned}(c_0^+)^2(\alpha_0^-)^2 &= \left(\mathbb{E}\left[Q_0^0\eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]-\tau_0^0\right)^2,\\ \frac{(c_0^+)^2(\alpha_0^-)^2}{\tau_0^0} &- \frac{(\mathbb{E}\left[Q_0^0\eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right])^2}{\tau_0^0}\\ &= -2\mathbb{E}\left[Q_0^0\eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]+\tau_0^0. \end{aligned}\tag{32}$$

Replacing (31) and (32) into (30), we reach

$$\begin{aligned}\left(\tau_0^+ - \frac{(c_0^+)^2}{\tau_0^0}\right)(\alpha_0^-)^2 &= \mathbb{E}\left[\eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]\\ &- \frac{\left(\mathbb{E}\left[Q_0^0\eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]\right)^2}{\tau_0^0} - (\alpha_0^+)^2\tau_0^-, \end{aligned}\tag{33}$$

$$\begin{aligned}\left(\tau_0^+ - \frac{(c_0^+)^2}{\tau_0^0}\right)(\gamma_0^+)^2 &= \frac{\mathbb{E}\left[\eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right](\gamma_0^+)^2}{(\alpha_0^-)^2}\\ &- \frac{\left(\mathbb{E}\left[Q_0^0\eta_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]\right)^2(\gamma_0^+)^2}{\tau_0^0(\alpha_0^-)^2} - (\gamma_0^-)^2\tau_0^-.\end{aligned}$$

Notice that $\mathbb{E}\left[\eta^2_{f/\gamma_0^-}(Q_0^0+Q_0^-)\right]$ simply translates into our variable q_{1x} from its definition (22c), and our dictionary directly transforms (33) into Eq. (22f):

$$\hat{\chi}_{2x} = \frac{q_{1x}}{\chi_{1x}^2} - \frac{m_{1x}^2}{\rho_x\chi_{1x}^2} - \hat{\chi}_{1x}.$$

- Recovering Eq. (22s)

We first note that for any function h,

$$\mathbb{E}[h(s_\nu)] = \min(1,\alpha)\mathbb{E}[h(s_\mu)] + \max(0,1-\alpha)h(0).$$

and $s_\nu^2 \sim p_\lambda$. Applying this to $h(s) = \dfrac{\gamma_1^- s^2}{\gamma_1^- s^2+\gamma_0^+}$ and starting from (27m), we rewrite

$$\begin{aligned}\alpha_1^+ &= \mathbb{E}\left[\frac{\gamma_1^- s_\mu^2}{\gamma_1^- s_\mu^2+\gamma_0^+}\right]\\ &= \frac{1}{\alpha}\mathbb{E}\left[\frac{\gamma_1^-\lambda}{\gamma_1^-\lambda+\gamma_0^+}\right]\end{aligned}$$

with $\lambda \sim p_\lambda$, which translates into Eq. (22s):

$$\chi_{2z} = \frac{1}{\alpha}\mathbb{E}\left[\frac{\lambda}{\hat{Q}_{2x}+\lambda\hat{Q}_{2z}}\right].$$

In a similar fashion, we can recover all equations (22) by writing variances and correlations between scalar random variables defined in (27), and using the independence properties established in [36]; thus directly showing the matching between the two state evolution formalisms at their fixed point.

APPENDIX F
NUMERICAL IMPLEMENTATION DETAILS

The plots were generated using the toolbox available at https://github.com/cgerbelo/Replica_GLM_orth.inv.git

Here we give a few derivation details for implementation of the equations presented in Theorem 1. We provide the Python script used to produce the figures in the main body of the paper as an example. The experimental points were obtained using the convex optimization tools of [42], with a data matrix of dimension $N = 200$, $M = \alpha N$, for $\alpha \in [0.1, 3]$. Each point is averaged 100 times to get smoother curves. The theoretical prediction was simply obtained by iterating the equations from Theorem 1. This can lead to unstable numerical schemes, and we include a few comments about stability in the code provided with this version of the paper. For Gaussian data, the design matrices were simply obtained by sampling a normal distribution $\mathcal{N}(0, \sqrt{1/M})$, effectively yielding the Marchenko-Pastur distribution [49] for averaging on the eigenvalues of $\mathbf{F}^T\mathbf{F}$ in the state evolution equations:

$$\lambda_{\mathbf{F}^T\mathbf{F}} \sim \max(0,1-\alpha)\delta(\lambda-0) + \alpha\frac{\sqrt{(0,\lambda-a)^+(0,b-\lambda)^+}}{2\pi\lambda}$$

where

$$a = \sqrt{1-\left(\frac{1}{\alpha}\right)^2}, \quad b = \sqrt{1+\left(\frac{1}{\alpha}\right)^2},$$

and $(0,x)^+ = \max(0,x)$. For the example of orthogonally invariant matrix with arbitrary spectrum, we chose to sample the singular values of $\mathbf{F}$ from the uniform distribution $\mathcal{U}(\left[(1-\alpha)^2,(1+\alpha)^2\right])$. This leads to the following distribution for the eigenvalues of $\mathbf{F}^T\mathbf{F}$:

$$\lambda_{\mathbf{F}^T\mathbf{F}} \sim \max(0,1-\alpha)\delta(0) + \min(1,\alpha)d(\lambda,\alpha) \tag{34}$$

where $\mathbb{I}$ is the indicator function and

$$d(\lambda,\alpha) = \left(\frac{1}{2((1+\alpha)^2-(1-\alpha)^2)}\mathbb{I}_{\{\sqrt{\lambda}\in[(1-\alpha)^2,(1+\alpha)^2]\}}\frac{1}{\sqrt{\lambda}}\right).$$

The only quantities that need additional calculus are the averages of proximals, squared proximals and derivatives of proximals. Here we give the corresponding expressions for the losses and regularizations that were used to make the figures. Note that the stability and convergence of the state evolution equations closely follow the result of Lemma 3. For example, a ridge regularized logistic regression, which is a strongly convex objective in both the loss (on compact spaces) and regularization will lead to more stable iterations than a LASSO SVC.

A. Regularization: Elastic Net

For the elastic net regularization, we can obtain an exact expression, avoiding any numerical integration. The proximal of the elastic net reads:

$$\text{Prox}_{\frac{1}{\hat{Q}_{1x}}(\lambda_1|\mathbf{x}|_1+\frac{\lambda_2}{2}\|\mathbf{x}\|_2^2)}(.) = \frac{1}{1+\frac{\lambda_2}{\hat{Q}_{1x}}}\,s\left(.,\frac{\lambda_1}{\hat{Q}_{1x}}\right)$$

where $s\left(.,\frac{\lambda_1}{\hat{Q}_{1x}}\right)$ is the soft-thresholding function:

$$s\left(r_{1k},\frac{\lambda_1}{\hat{Q}_{1x}}\right)=\begin{cases} r_{1k}+\frac{\lambda_1}{\hat{Q}_{1x}} & \text{if} \quad r_{1k}<-\frac{\lambda_1}{\hat{Q}_{1x}} \\ 0 & \text{if} \quad -\frac{\lambda_1}{\hat{Q}_{1x}}<r_{1k}<\frac{\lambda_1}{\hat{Q}_{1x}} \\ r_{1k}-\frac{\lambda_1}{\hat{Q}_{1x}} & \text{if} \quad r_{1k}>\frac{\lambda_1}{\hat{Q}_{1x}}. \end{cases}$$

We assume that the ground-truth x_0 is pulled from a Gauss-Bernoulli law of the form:

$$\phi(x_0)=(1-\rho)\delta(0)+\rho\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-x_0^2/(2\sigma^2)\right).$$

Note that we did our plots with $\rho=1$, but this form can be used to study the effect of sparsity in the model. Writing $X=(\hat{m}_{1x}x_0+\sqrt{\hat{\chi}_{1x}}\xi_{1x})/\hat{Q}_{1x}$, and remembering that $\xi_{1x}\sim\mathcal{N}(0,1)$, some calculus then shows equations (35)-(37), shown at the bottom of the next page. We now turn to the loss functions.

B. Loss Functions

The loss functions sometimes have no closed form, as is the case for the logistic loss. In that case, numerical integration cannot be avoided, and we recommend marginalizing all the possible variables that can be averaged out. In the present model, if the teacher y is chosen as a sign, one-dimensional integrals can be reached, leading to stable and reasonably fast implementation (a few minutes to generate a curve comparable to those of Figure 1 for the non-linear models, the ridge regression being very fast). The interested reader can find the corresponding marginalized prefactors in the code jointly provided with this paper.

a) Square loss: The square loss is defined as:

$$f(x,y)=\frac{1}{2}(x-y)^2,$$

its proximal and partial derivative then read:

$$\begin{aligned}\text{Prox}_{\frac{1}{\gamma}f}(p)&=\frac{\gamma}{1+\gamma}p+\frac{1}{1+\gamma}y\\ \frac{\partial}{\partial p}\text{Prox}_{\frac{1}{\gamma}f}(p)&=\frac{\gamma}{1+\gamma}.\end{aligned}$$

Using this form with a plain ridge penalty (elastic net with $\ell_1=0$) leads to great simplification in the equations of Theorem 1 and we recover the classical expressions obtained for ridge regression in papers such as [9], [30].

b) Hinge loss: The hinge loss reads:

$$f(x,y)=\max(0,1-yx).$$

Assuming $y\in\{-1,+1\}$, its proximal and partial derivative then read:

$$\begin{aligned}\text{Prox}_{\frac{1}{\gamma}f}(p)&=\begin{cases} p+\frac{y}{\gamma} & \text{if} \quad \gamma(1-yp)\geqslant 1 \\ y & \text{if} \quad 0\leqslant\gamma(1-yp)\leqslant 1 \\ p & \text{if} \quad \gamma(1-yp)\leqslant 0 \end{cases}\\ \frac{\partial}{\partial p}\text{Prox}_{\frac{1}{\gamma}f}(p)&=\begin{cases} 1 & \text{if} \quad \gamma(1-yp)\geqslant 1 \\ 0 & \text{if} \quad 0\leqslant\gamma(1-yp)\leqslant 1 \\ 1 & \text{if} \quad \gamma(1-yp)\leqslant 0. \end{cases}\end{aligned}$$

c) Logistic loss: The logistic loss reads:

$$f(x,y)=\log(1+\exp(-yx))$$

Its proximal (at point p) is the solution to the fixed point problem:

$$x=p+\frac{y}{\gamma(1+\exp(yx))},$$

and its derivative, given that the logistic loss is twice differentiable, reads:

$$\begin{aligned}\frac{\partial}{\partial p}\text{Prox}_{\frac{1}{\gamma}f}(p)&=\frac{1}{1+\frac{1}{\gamma}\frac{\partial^2}{\partial p^2}f(\text{Prox}_{\frac{1}{\gamma}f}(p))}\\ &=\frac{1}{1+\frac{1}{\gamma}(2+2\text{cosh}(\text{Prox}_{\frac{1}{\gamma}f}(p)))^{-1}}.\end{aligned}$$

APPENDIX G
PROOF OF LEMMA 3: CONVERGENCE ANALYSIS OF 2-LAYER MLVAMP

In this section, we give the detail of the convergence proof of 2-layer MLVAMP.

A. Proof of Proposition 1

This proof is quite straightforward and close to the one of Theorem 4 from [44]. Let $\delta\mathbf{h}^t=\delta\mathbf{h}^{(t)}-\mathbf{h}^{(t-1)}$, $\delta\mathbf{u}^t=\delta\mathbf{u}^{(t)}-\mathbf{u}^{(t-1)}$, $\delta\mathbf{w}^t=\mathbf{w}^{(t)}-\mathbf{w}^{(t-1)}$.

Multiplying Eq.(16) on the left and right by $[(\delta\mathbf{h}^t)^\top\ \ (\delta\mathbf{u}^t)^\top]$ and its transpose respectively, we get

$$\begin{aligned}&(\mathbf{A}^{(t)}(\delta\mathbf{h}^t)+\mathbf{B}^{(t)}(\delta\mathbf{u}^t))^\top\mathbf{P}(\mathbf{A}^{(t)}(\delta\mathbf{h}^t)+\mathbf{B}^{(t)}(\delta\mathbf{u}^t))\\ &-(\tau_{(t)})^2(\delta\mathbf{h}^t)^\top\mathbf{P}(\delta\mathbf{h}^t)\\ &+\beta_1^{(t)}(\mathbf{C}_1^{(t)}(\delta\mathbf{h}^t)+\mathbf{D}_1(\delta\mathbf{u}^t))^\top\mathbf{M}_1^{(t)}(\mathbf{C}_1^{(t)}(\delta\mathbf{h}^t)+\mathbf{D}_1(\delta\mathbf{u}^t))\\ &+\beta_2^{(t)}(\mathbf{C}_2^{(t)}(\delta\mathbf{h}^t)+\mathbf{D}_2(\delta\mathbf{u}^t))^\top\mathbf{M}_2^{(t)}(\mathbf{C}_2^{(t)}(\delta\mathbf{h}^t)+\mathbf{D}_2(\delta\mathbf{u}^t))\\ &\leqslant 0.\end{aligned}$$

Using the definition of iteration (14), this simplifies to

$$\begin{aligned}&(\delta\mathbf{h}^{t+1})^\top\mathbf{P}(\delta\mathbf{h}^{t+1})-(\tau_{(t)})^2(\delta\mathbf{h}^t)^\top\mathbf{P}(\delta\mathbf{h}^t)\\ &+\beta_1(\delta\mathbf{w}_1^t)^\top\mathbf{M}_1^{(t)}(\delta\mathbf{w}_1^t)+\beta_2(\delta\mathbf{w}_2^t)^\top\mathbf{M}_2^{(t)}(\delta\mathbf{w}_1^t)\leqslant 0.\end{aligned}$$

Owing to the Lipschitz properties of $\tilde{\mathcal{O}}_1^{(t)},\tilde{\mathcal{O}}_2^{(t)}$ and the definitions of $\mathbf{w}_1^{(t)},\mathbf{w}_2^{(t)}$, the terms factoring β_1,β_2 are both non-negative. We thus have, at each time step t:

$$(\delta\mathbf{h}^{t+1})^\top\mathbf{P}(\delta\mathbf{h}^{t+1})\leqslant\tau_{(t)}(\delta\mathbf{h}^t)^\top\mathbf{P}(\delta\mathbf{h}^t).$$

Letting $\tau^*=\sup_t\tau_{(t)}$, an immediate induction concludes the proof.

B. Bounds on $\hat{Q}_{1x}^{(t+1)},\hat{Q}_{1z}^{(t)},\hat{Q}_{2x}^{(t)},\hat{Q}_{2z}^{(t+1)}$

We remind that, since the functions f and g are separable, their Hessians are diagonal matrices. For any time index t, the following bounds hold:

- On $\hat{Q}_{2x}^{(t)}$

$$\begin{aligned}&\hat{Q}_{2x}^{(t)}=1/\chi_{1x}^{(t)}-\hat{Q}_{1x}^{(t)}\ \text{ where }\ \chi_{1x}^{(t)}=\left\langle\partial_{\mathbf{h}_{1x}^{(t)}}g_{1x}(\ldots)\right\rangle/\hat{Q}_{1x}^{(t)},\\ &\text{then}\ \ \frac{1}{\hat{Q}_{2x}^{(t)}+\hat{Q}_{1x}^{(t)}}=\frac{1}{N}\left(\text{Tr}\left[(\hat{Q}_{1x}^{(t)}Id+\mathcal{H}_f(\text{Prox}))^{-1}\right]\right),\\ &\hat{Q}_{1x}^{(t)}+\lambda_{min}(\mathcal{H}_f)\leqslant\hat{Q}_{1x}^{(t)}+\hat{Q}_{2x}^{(t)}\leqslant\hat{Q}_{1x}^{(t)}+\lambda_{max}(\mathcal{H}_f).\end{aligned}$$

- On $\hat{Q}_{2z}^{(t+1)}$

$\hat{Q}_{2z}^{(t+1)} = 1/\chi_{1z}^{(t)} - \hat{Q}_{1z}^{(t)}$ where $\chi_{1z}^{(t)} = \left\langle \partial_{\mathbf{h}_{1z}^{(t)}} g_{1z}(\ldots) \right\rangle / \hat{Q}_{1z}^{(t)}$,
then $\dfrac{1}{\hat{Q}_{2z}^{(t+1)} + \hat{Q}_{1z}^{(t)}} = \dfrac{1}{M}\left(\text{Tr}\left[(\hat{Q}_{1z}^{(t)} Id + \mathcal{H}_g(\text{Prox}))^{-1}\right]\right)$,
$\hat{Q}_{1z}^{(t)} + \lambda_{min}(\mathcal{H}_g) \leqslant \hat{Q}_{1z}^{(t)} + \hat{Q}_{2z}^{(t+1)} \leqslant \hat{Q}_{1z}^{(t)} + \lambda_{max}(\mathcal{H}_g)$.

- On $\hat{Q}_{1z}^{(t)}$

$\hat{Q}_{1z}^{(t)} = 1/\chi_{2z}^{(t)} - \hat{Q}_{2z}^{(t)}$ where $\chi_{2z}^{(t)} = \left\langle \partial_{\mathbf{h}_{2z}^{(t)}} g_{2z}(\ldots) \right\rangle / \hat{Q}_{2z}^{(t)}$,
then $\dfrac{1}{\hat{Q}_{1z}^{(t)} + \hat{Q}_{2z}^{(t)}} = \dfrac{1}{M}\text{Tr}\left[\mathbf{F}\mathbf{F}^\top \left(\hat{Q}_{2z}^{(t)} \mathbf{F}\mathbf{F}^\top + \hat{Q}_{2x}^{(t)} Id\right)^{-1}\right]$.

The matrices on the r.h.s. of the previous equation are all diagonalizable in the same basis. Then each eigenvalue has the form

$$\frac{\lambda_k(\mathbf{F}\mathbf{F}^\top)}{\hat{Q}_{2z}^{(t)}\lambda_k(\mathbf{F}\mathbf{F}^\top) + \hat{Q}_{2x}^{(t)}}$$

which leads to the bound

$$\hat{Q}_{2z}^{(t)} + \frac{\hat{Q}_{2x}^{(t)}}{\lambda_{max}(\mathbf{F}\mathbf{F}^\top)} \leqslant \hat{Q}_{1z}^{(t)} + \hat{Q}_{2z}^{(t)} \leqslant \hat{Q}_{2z}^{(t)} + \frac{\hat{Q}_{2x}^{(t)}}{\lambda_{min}(\mathbf{F}\mathbf{F}^\top)}.$$

- On $\hat{Q}_{1x}^{(t+1)}$

$\hat{Q}_{1x}^{(t+1)} = 1/\chi_{2x}^{(t+1)} - \hat{Q}_{2x}^{(t)}$ where $\chi_{2x}^{(t+1)} = \left\langle \partial_{\mathbf{h}_{2x}^{(t)}} g_{2x}(\ldots) \right\rangle / \hat{Q}_{2x}^{(t)}$,
then $\dfrac{1}{\hat{Q}_{1x}^{(t+1)} + \hat{Q}_{2x}^{(t)}} = \dfrac{1}{N}\text{Tr}\left[\left(\hat{Q}_{2z}^{(t+1)} \mathbf{F}^\top\mathbf{F} + \hat{Q}_{2x}^{(t)} Id\right)^{-1}\right]$,

which leads to

$$\begin{aligned}\hat{Q}_{2x}^{(t)} + \lambda_{min}(\mathbf{F}^\top\mathbf{F})\hat{Q}_{2z}^{(t+1)} &\leqslant \hat{Q}_{1x}^{(t+1)} + \hat{Q}_{2x}^{(t)} \\ &\leqslant \hat{Q}_{2x}^{(t)} + \lambda_{max}(\mathbf{F}^\top\mathbf{F})\hat{Q}_{2z}^{(t+1)}.\end{aligned}$$

C. Operator Norms and Lipschitz Constants

1) Operator Norms of $\mathbf{W_1}^{(t)}, \mathbf{W_2}^{(t)}, \mathbf{W_3}^{(t)}, \mathbf{W_4}^{(t)}$: The norms of the linear operators $\mathbf{W_1}^{(t)}$, $\mathbf{W_2}^{(t)}$, $\mathbf{W_3}^{(t)}$, $\mathbf{W_4}^{(t)}$ can be computed or bounded with respect to the singular values of the matrix $\mathbf{F}$. The derivations are straightforward and do not require any specific mathematical result. Denoting $\|\mathbf{W}\|$ the operator norm of a given matrix $\mathbf{W}$, we have the following:

$$\begin{aligned}\|\mathbf{W_1}^{(t)}\| &= \frac{\hat{Q}_{2x}^{(t)}}{\hat{Q}_{1x}^{(t+1)}} \max\left(\frac{|\hat{Q}_{1x}^{(t+1)} - \hat{Q}_{2z}^{(t+1)}\lambda_{min}(\mathbf{F}^T\mathbf{F})|}{\hat{Q}_{2x}^{(t)} + \hat{Q}_{2z}^{(t+1)}\lambda_{min}(\mathbf{F}^T\mathbf{F})}, \right. \\ &\qquad \left. \frac{|\hat{Q}_{1x}^{(t+1)} - \hat{Q}_{2z}^{(t+1)}\lambda_{max}(\mathbf{F}^T\mathbf{F})|}{\hat{Q}_{2x}^{(t)} + \hat{Q}_{2z}^{(t+1)}\lambda_{max}(\mathbf{F}^T\mathbf{F})}\right) \\ \|\mathbf{W_2}^{(t)}\| &= \frac{\hat{Q}_{2z}^{(t+1)}}{\chi_{2x}^{(t+1)}\hat{Q}_{1x}^{(t+1)}} \frac{\sqrt{\lambda_{max}(\mathbf{F}^T\mathbf{F})}}{\hat{Q}_{2x}^{(t)} + \hat{Q}_{2z}^{(t+1)}\lambda_{min}(\mathbf{F}^T\mathbf{F})} \\ \|\mathbf{W_3}^{(t)}\| &= \frac{\hat{Q}_{2z}^{(t)}}{\hat{Q}_{1z}^{(t)}} \max\left(\frac{|\hat{Q}_{2x}^{(t)} - \hat{Q}_{1z}^{(t)}\lambda_{min}(\mathbf{F}\mathbf{F}^T)|}{\hat{Q}_{2x}^{(t)} + \hat{Q}_{2z}^{(t)}\lambda_{min}(\mathbf{F}\mathbf{F}^T)}, \right. \\ &\qquad \left. \frac{|\hat{Q}_{2x}^{(t)} - \hat{Q}_{1z}^{(t)}\lambda_{max}(\mathbf{F}\mathbf{F}^T)|}{\hat{Q}_{2x}^{(t)} + \hat{Q}_{2z}^{(t)}\lambda_{max}(\mathbf{F}\mathbf{F}^T)}\right) \\ \|\mathbf{W_4}^{(t)}\| &= \frac{\hat{Q}_{2x}^{(t)}}{\chi_{2z}^{(t)}\hat{Q}_{1z}^{(t)}} \frac{\sqrt{\lambda_{max}(\mathbf{F}^T\mathbf{F})}}{\hat{Q}_{2x}^{(t)} + \hat{Q}_{2z}^{(t)}\lambda_{min}(\mathbf{F}^T\mathbf{F})}.\end{aligned}$$

2) Lispchitz Constants of $\tilde{\mathcal{O}}_1^{(t)}, \tilde{\mathcal{O}}_2^{(t)}$: We now derive upper bounds of the Lipschitz constants of $\tilde{\mathcal{O}}_1^{(t)}, \tilde{\mathcal{O}}_2^{(t)}$ using the convex analysis reminder in appendix B. We give detail for $\tilde{\mathcal{O}}_1^{(t)}$, the derivation is identical for $\tilde{\mathcal{O}}_2^{(t)}$. Let $(\sigma_1, \beta_1) \in \mathbb{R}_+^{*2}$ be the strong-convexity and smoothness constants of f, if they exist. If f has no strong convexity constant, we set $\sigma_1 = 0$, and if it holds no smoothness assumption, we set $\beta_1 = +\infty$. Note that, from the upper and lower bounds obtained in appendix G-B, we have $\sigma_1 \leqslant \hat{Q}_{2x}^{(t)} \leqslant \beta_1$.

$$\begin{aligned}&\mathbb{E}[\text{Prox}^2_{\mathbf{f}/\hat{Q}_{1x}}(X)] \\ &= \left(\frac{1}{1+\frac{\lambda_2}{\hat{Q}_{1x}}}\right)^2 \left[(1-\rho)\left(\frac{\lambda_1^2 + \hat{\chi}_{1x}}{(\hat{Q}_{1x})^2}\,\text{erfc}\left(\frac{\lambda_1}{\sqrt{2\hat{\chi}_{1x}}}\right) - \frac{\lambda_1\sqrt{2\hat{\chi}_{1x}}\exp(-\frac{\lambda_1^2}{2(\hat{\chi}_{1x})})}{\sqrt{\pi}}\right)\right. \\ &\left. + \rho\left(\frac{\lambda_1^2 + \hat{\chi}_{1x} + \sigma^2\hat{m}_{1x}^2}{(\hat{Q}_{1x})^2}\,\text{erfc}\left(\frac{\lambda_1}{\sqrt{2(\hat{\chi}_{1x} + \sigma^2\hat{m}_{1x}^2)}}\right) - \frac{\lambda_1\sqrt{2(\hat{\chi}_{1x} + \sigma^2\hat{m}_{1x}^2)}\exp(-\frac{\lambda_1^2}{2(\hat{Q}_{1x})^2(\hat{\chi}_{1x} + \sigma^2\hat{m}_{1x}^2)})}{\sqrt{\pi}}\right)\right].\end{aligned} \tag{35}$$

Similarly, we have

$$\mathbb{E}[\text{Prox}'_{\mathbf{f}/\hat{Q}_{1x}}(X)] = \frac{1}{1+\frac{\lambda_2}{\hat{Q}_{1x}}}\left[(1-\rho)\,\text{erfc}\left(\frac{\lambda_1}{\sqrt{2\hat{\chi}_{1x}}}\right) + \rho\,\text{erfc}\left(\frac{\lambda_1}{\sqrt{2(\hat{\chi}_{1x} + \sigma^2\hat{m}_{1x}^2)}}\right)\right] \tag{36}$$

and

$$\mathbb{E}[x_0\text{Prox}_{\mathbf{f}/\hat{Q}_{1x}}(X)] = \frac{\rho|\sigma\hat{m}_{1x}|}{\hat{Q}_{1x} + \lambda_2}\,\text{erfc}\left(\frac{\lambda_1}{\sqrt{2(\hat{\chi}_{1x} + \sigma^2\hat{m}_{1x}^2)}}\right). \tag{37}$$

a) Case 1: $0 < \sigma_1 < \beta_1$: Proposition 3 gives the following expression:

$$\text{Prox}_{\frac{1}{\hat{Q}_{1x}^{(t)}} f} = \frac{1}{2}\left(\frac{1}{1+\sigma_1/\hat{Q}_{1x}^{(t)}} + \frac{1}{1+\beta_1/\hat{Q}_{1x}^{(t)}}\right)\text{Id} + \frac{1}{2}\left(\frac{1}{1+\sigma_1/\hat{Q}_{1x}^{(t)}} - \frac{1}{1+\beta_1/\hat{Q}_{1x}^{(t)}}\right) S_1$$

where S_1 is a non-expansive operator. Replacing in the expression of $\tilde{\mathcal{O}}_1$ leads to:

$$\tilde{\mathcal{O}}_1^{(t)} = \frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}}\left(\left(\frac{1}{2\chi_{1x}^{(t)}}\left(\frac{1}{\hat{Q}_{1x}^{(t)}+\sigma_1} + \frac{1}{\hat{Q}_{1x}^{(t)}+\beta_1}\right) - 1\right)\text{Id} + \frac{1}{2\chi_{1x}^{(t)}}\left(\frac{1}{\hat{Q}_{1x}^{(t)}+\sigma_1} - \frac{1}{\hat{Q}_{1x}^{(t)}+\beta_1}\right) S_1\right). \tag{38}$$

Knowing that $\hat{Q}_{1x}^{(t)} + \hat{Q}_{2x}^{(t)} = 1/\chi_{1x}^{(t)}$, and separating the case where the first term of the sum in Eq.(38) is negative or positive, $\tilde{\mathcal{O}}_1$ has Lipschitz constant:

$$\omega_1^{(t)} = \frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}} \max\left(\frac{\hat{Q}_{2x}^{(t)}-\sigma_1}{\hat{Q}_{1x}^{(t)}+\sigma_1}, \frac{\beta_1-\hat{Q}_{2x}^{(t)}}{\hat{Q}_{1x}^{(t)}+\beta_1}\right). \tag{39}$$

b) Case 2: $0 < \sigma_1 = \beta_1$: In this case, we have from Proposition 3:

$$\|\text{Prox}_{\frac{1}{\hat{Q}_{1x}^{(t)}} f}(x) - \text{Prox}_{\frac{1}{\hat{Q}_{1x}^{(t)}} f}(y)\|_2^2 = \left(\frac{1}{1+\sigma_1/\hat{Q}_{1x}^{(t)}}\right)^2 \|x-y\|_2^2.$$

With the firm non-expansiveness of the proximal operator gives, we reach for any $x, y \in \mathbb{R}$ Eq. (40), shown at the bottom of the next page, detailed below. The upper bound on the Lipschitz constant is therefore:

$$\omega_1 = \frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}}\sqrt{1 + \frac{((\hat{Q}_{2x}^{(t)})^2 - (\hat{Q}_{1x}^{(t)})^2)}{(\hat{Q}_{1x}^{(t)}+\sigma_1)^2}}. \tag{41}$$

c) Case 3: no strong convexity or smoothness assumption: This setting is not necessary for our proof, because we only handle penalty functions which have a strictly positive strong convexity constant, by adding a ridge term. However, we list it for completeness. In this case, the only information we have is the firm nonexpansiveness of the proximal operator, which leads us to the same derivation as the previous one up to (40), where the first term in the sum can be positive or negative. This yields the Lipschitz constant:

$$\omega_1^{(t)} = \frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}} \max\left(1, \frac{\hat{Q}_{2x}^{(t)}}{\hat{Q}_{1x}^{(t)}}\right).$$

d) Recovering (15): In our proof, we make no assumption on the strong-convexity or smoothness of the function, but adding the ridge penalties $\lambda_2, \tilde{\lambda}_2$ brings us for both $\tilde{\mathcal{O}}_1^{(t)}$ and $\tilde{\mathcal{O}}_2^{(t)}$ to either the first of the second case above. It is straightforward to see that the Lipschitz constant (41) is an upper bound of (39). We thus use (41) for generality, and recover the expressions (15) shown in the main body of the paper.

$$\omega_1^{(t)} = \frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}}\sqrt{1 + \frac{(\hat{Q}_{2x}^{(t)})^2 - (\hat{Q}_{1x}^{(t)})^2}{(\hat{Q}_{1x}^{(t)}+\lambda_2)^2}}$$

$$\omega_2^{(t)} = \frac{\hat{Q}_{1z}^{(t)}}{\hat{Q}_{2z}^{(t)}}\sqrt{1 + \frac{(\hat{Q}_{2z}^{(t)})^2 - (\hat{Q}_{1z}^{(t)})^2}{(\hat{Q}_{1z}^{(t)}+\tilde{\lambda}_2)^2}}.$$

D. Dynamical System Convergence Analysis

We are now ready to prove Lemma 3.

We will use the bounds derived above to prove the convergence lemma. Since we have proved the required bounds at any time step, we drop the time indices in the remainder of this proof for simplicity. The choice of additional regularization is λ_2 arbitrarily large, and $\tilde{\lambda}_2$ fixed but finite and non-zero. $\hat{Q}_{2x}, \hat{Q}_{1z}$ can thus be made arbitrarily large, and $\hat{Q}_{2z}, \hat{Q}_{1x}$ remain finite. We write the corresponding linear matrix inequality (16) and expand the constraint term. Some algebra shows that:

$$\mathbf{C}_1^T\mathbf{M}_1\mathbf{C}_1 = \begin{bmatrix}\mathbf{0}_{M\times M} & 0_{M\times N}\\ \mathbf{0}_{N\times M} & \omega_1^2 I_{N\times N}\end{bmatrix}$$

$$\mathbf{C}_2^T\mathbf{M}_2\mathbf{C}_2 = \begin{bmatrix}\omega_2^2\mathbf{W}_3^T\mathbf{W}_3 & \mathbf{0}_{M\times N}\\ \mathbf{0}_{N\times M} & \mathbf{0}_{N\times N}\end{bmatrix}$$

$$\mathbf{C}_1^T\mathbf{M}_1\mathbf{D}_1 = \mathbf{0}_{(M+N)\times(M+N)}$$

$$\mathbf{D}_1^T\mathbf{M}_1\mathbf{C}_1 = \mathbf{0}_{(M+N)\times(M+N)}$$

$$\mathbf{C}_2^T\mathbf{M}_2\mathbf{D}_2 = \begin{bmatrix}\mathbf{0}_{M\times M} & \omega_2^2\mathbf{W}_3^T\mathbf{W}_4\\ \mathbf{0}_{N\times M} & \mathbf{0}_{N\times N}\end{bmatrix}$$

$$\mathbf{D}_2^T\mathbf{M}_2\mathbf{C}_2 = \begin{bmatrix}\mathbf{0}_{M\times M} & \mathbf{0}_{M\times N}\\ \omega_2^2\mathbf{W}_4^T\mathbf{W}_3 & \mathbf{0}_{N\times N}\end{bmatrix}$$

$$\mathbf{D}_1^T\mathbf{M}_1\mathbf{D}_1 = \begin{bmatrix}\mathbf{0}_{M\times M} & \mathbf{0}_{M\times N}\\ \mathbf{0}_{N\times M} & -\mathbf{I}_{N\times N}\end{bmatrix}$$

$$\mathbf{D}_2^T\mathbf{M}_2\mathbf{D}_2 = \begin{bmatrix}-\mathbf{I}_{M\times M} & \mathbf{0}_{M\times N}\\ \mathbf{0}_{N\times M} & \omega_2^2\mathbf{W}_4^T\mathbf{W}_4\end{bmatrix}$$

where all the matrices constituting the blocks have been defined in section VI. This gives the following form for the constraint matrix:

$$\begin{bmatrix}\mathbf{H}_1 & \mathbf{H}_2\\ \mathbf{H}_2^T & \mathbf{H}_3\end{bmatrix}$$

where

$$\mathbf{H}_1 = \begin{bmatrix}\beta_1\omega_2^2\mathbf{W}_3^T\mathbf{W}_3 & \mathbf{0}_{M\times N}\\ \mathbf{0}_{N\times M} & \beta_0\omega_1^2\mathbf{I}_{N\times N}\end{bmatrix}$$

$$\mathbf{H}_2 = \begin{bmatrix}\mathbf{0}_{M\times M} & \beta_1\omega_2^2\mathbf{W}_3^T\mathbf{W}_4\\ \mathbf{0}_{N\times M} & \mathbf{0}_{N\times N}\end{bmatrix}$$

$$\mathbf{H}_3 = \begin{bmatrix}-\beta_1\mathbf{I}_{M\times M} & \mathbf{0}_{M\times N}\\ \mathbf{0}_{N\times M} & -\beta_0\mathbf{I}_{N\times N} + \beta_1\omega_2^2\mathbf{W}_4^T\mathbf{W}_4\end{bmatrix}/$$

Thus the LMI (16) becomes:

$$0 \succeq \begin{bmatrix}-\tau^2\mathbf{P} + \mathbf{H}_1 & \mathbf{H}_2\\ \mathbf{H}_2^T & \mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3\end{bmatrix}.$$

We take $\mathbf{P}$ as block diagonal:

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_1 & \mathbf{0}_{M\times N} \\ \mathbf{0}_{N\times M} & \mathbf{P}_2 \end{bmatrix}$$

where $\mathbf{P}_1 \in \mathbb{R}^{M\times M}$ and $\mathbf{P}_2 \in \mathbb{R}^{N\times N}$ are positive definite (no zero eigenvalues) and diagonalizable in the same basis as $\mathbf{F}^T\mathbf{F}$, which is also the eigenbasis of $\mathbf{W}_1, \mathbf{W}_3, \mathbf{W}_2^T\mathbf{W}_2, \mathbf{W}_4^T\mathbf{W}_4$. We then have:

$$\mathbf{B}^T\mathbf{P}\mathbf{B} = \begin{bmatrix} \mathbf{P}_1 + \mathbf{W}_2^T\mathbf{P}_2\mathbf{W}_2 & \mathbf{W}_2^T\mathbf{P}_2\mathbf{W}_1 \\ \mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_2 & \mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_1 \end{bmatrix}.$$

We are then trying to find the conditions for the following problem to be feasible with $0 < \tau < 1$:

$$\begin{bmatrix} \tau^2\mathbf{P} - \mathbf{H}_1 & -\mathbf{H}_2 \\ -\mathbf{H}_2^T & -(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3) \end{bmatrix} \succeq 0. \tag{42}$$

Schur's lemma then gives that the strict version of (42), which we will consider, is equivalent [47] to:

$$\begin{aligned} &-(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3) \succ 0 \quad \text{and} \\ &\tau^2\mathbf{P} - \mathbf{H}_1 + \mathbf{H}_2(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3)^{-1}\mathbf{H}_2^T \succ 0. \end{aligned} \tag{43}$$

Let us first consider $-(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3)$.

1) Conditions for $-(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{H}_3) \succ 0$: Expanding $-(\mathbf{B}^T\mathbf{P}\mathbf{B}+\mathbf{H}_3) \succ 0$ and applying Schur's lemma again gives the equivalent problem:

$$\beta_1\mathbf{I}_{N\times N} - \beta_2\omega_2^2\mathbf{W}_4^T\mathbf{W}_4 - \mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_1 \succ 0 \quad \text{and} \tag{44}$$

$$\beta_2\mathbf{I}_{M\times M} - \mathbf{P}_1 - \mathbf{W}_2^T\mathbf{P}_2\mathbf{W}_2 - \mathbf{W}_2^T\mathbf{P}_2\mathbf{W}_1\mathbf{K}_1\mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_2 \succ 0. \tag{45}$$

where $\mathbf{K}_1 = (\beta_1\mathbf{I}_{N\times N} - \beta_2\omega_2^2\mathbf{W}_4^T\mathbf{W}_4 - \mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_1)^{-1}$. We start with (44). A sufficient condition for it to hold true is:

$$\beta_1 > \beta_2\omega_2^2\lambda_{max}(\mathbf{W}_4^T\mathbf{W}_4) + \lambda_{max}(\mathbf{P}_2)\lambda_{max}(\mathbf{W}_1^T\mathbf{W}_1).$$

Using the bounds from appendix G-C, we have:

$$\begin{aligned} \lambda_{max}(\mathbf{W}_1^T\mathbf{W}_1) &\leqslant \left(\frac{\hat{Q}_{2x}}{\hat{Q}_{1x}}\right)^2 \\ &\times\max\left(\frac{|\hat{Q}_{1x}-\hat{Q}_{2z}\lambda_{min}(\mathbf{F}^T\mathbf{F})|}{\hat{Q}_{2x}+\hat{Q}_{2z}\lambda_{min}(\mathbf{F}^T\mathbf{F})}, \frac{|\hat{Q}_{1x}-\hat{Q}_{2z}\lambda_{max}(\mathbf{F}^T\mathbf{F})|}{\hat{Q}_{2x}+\hat{Q}_{2z}\lambda_{max}(\mathbf{F}^T\mathbf{F})}\right)^2 \\ &\leqslant \max\left(\left(1-\frac{\hat{Q}_{2z}}{\hat{Q}_{1x}}\lambda_{min}(\mathbf{F}^T\mathbf{F})\right)^2, \left(1-\frac{\hat{Q}_{2z}}{\hat{Q}_{1x}}\lambda_{max}(\mathbf{F}^T\mathbf{F})\right)^2\right) \\ &= b_1, \end{aligned}$$

and

$$\begin{aligned} \omega_2^2\lambda_{max}(\mathbf{W}_4^T\mathbf{W}_4) &\leqslant \left(\frac{\hat{Q}_{1z}}{\hat{Q}_{2z}}\right)^2\left(\frac{\hat{Q}_{2x}}{\chi_{2z}\hat{Q}_{1z}}\right)^2 \\ &\times\left(1+\frac{(\hat{Q}_{2z})^2-(\hat{Q}_{1z})^2}{(\hat{Q}_{1z}+\tilde{\lambda}_2)^2}\right)\frac{\lambda_{max}(\mathbf{F}^T\mathbf{F})}{(\hat{Q}_{2x}+\hat{Q}_{2z}\lambda_{min}(\mathbf{F}^T\mathbf{F}))^2} \\ &\leqslant \hat{Q}_{1z}\left(2\tilde{\lambda}_2 + \frac{\tilde{\lambda}_2^2}{\hat{Q}_{1z}} + \frac{(\hat{Q}_{2z})^2}{\hat{Q}_{1z}}\right) \\ &\qquad\times\left(\frac{\hat{Q}_{1z}+\hat{Q}_{2z}}{\hat{Q}_{2z}(\hat{Q}_{1z}+\tilde{\lambda}_2)}\right)^2\lambda_{max}(\mathbf{F}^T\mathbf{F}). \end{aligned}$$

For arbitrarily large $\hat{Q}_{1z}$, the quantity

$$\left(2\tilde{\lambda}_2 + \frac{\tilde{\lambda}_2^2}{\hat{Q}_{1z}} + \frac{(\hat{Q}_{2z})^2}{\hat{Q}_{1z}}\right)\left(\frac{\hat{Q}_{1z}+\hat{Q}_{2z}}{\hat{Q}_{2z}(\hat{Q}_{1z}+\tilde{\lambda}_2)}\right)^2\lambda_{max}(\mathbf{F}^T\mathbf{F})$$

is trivially bounded above whatever the value of $\tilde{\lambda}_2, \hat{Q}_{2z}$. Let b_2 be such an upper bound independent of $\lambda_2, \hat{Q}_{2x}, \hat{Q}_{1z}$. The sufficient condition for (44) to hold thus becomes:

$$\beta_1 > \beta_2\hat{Q}_{1z}b_2 + \lambda_{max}(\mathbf{P}_2)b_1 \tag{46}$$

where b_1, b_2 are constants independent of $\lambda_2, \hat{Q}_{2x}, \hat{Q}_{1z}$.

We now turn to (45). A sufficient condition for it to hold is:

$$\begin{aligned} &\beta_2 > \lambda_{max}(\mathbf{P}_1) + \lambda_{max}(\mathbf{W}_2^T\mathbf{W}_2)\lambda_{max}(\mathbf{P}_2) \\ &+\frac{(\lambda_{max}(\mathbf{P}_2))^2\lambda_{max}(\mathbf{W}_2^T\mathbf{W}_2)\lambda_{max}(\mathbf{W}_1^T\mathbf{W}_1)}{\beta_1 - \beta_2\omega_2^2\lambda_{max}(\mathbf{W}_4^T\mathbf{W}_4) - \lambda_{max}(\mathbf{P}_2)\lambda_{max}(\mathbf{W}_1^T\mathbf{W}_1)}. \end{aligned} \tag{47}$$

$$\begin{aligned} \|\tilde{\mathcal{O}}_1^{(t)}(x) - \tilde{\mathcal{O}}_1^{(t)}(y)\|_2^2 &= \left(\frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}}\right)^2\Bigg(\frac{1}{(\hat{Q}_{1x}^{(t)})^2(\chi_{1x}^{(t)})^2}\|\mathrm{Prox}_{\frac{1}{\hat{Q}_{1x}^{(t)}}f}(x) - \mathrm{Prox}_{\frac{1}{\hat{Q}_{1x}^{(t)}}f}(y)\|_2^2 \\ &\qquad - 2\frac{1}{\hat{Q}_{1x}^{(t)}\chi_{1x}^{(t)}}\left\langle x-y, \mathrm{Prox}_{\frac{1}{\hat{Q}_{1x}^{(t)}}f}(x) - \mathrm{Prox}_{\frac{1}{\hat{Q}_{1x}^{(t)}}f}(y)\right\rangle + \|x-y\|_2^2\Bigg) \\ &\leqslant \left(\frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}}\right)^2\left(\left(\frac{1}{(\hat{Q}_{1x}^{(t)})^2(\chi_{1x}^{(t)})^2} - 2\frac{1}{\hat{Q}_{1x}^{(t)}\chi_{1x}^{(t)}}\right)\|\mathrm{Prox}_{\frac{1}{\hat{Q}_{1x}}f}(x) - \mathrm{Prox}_{\frac{1}{\hat{Q}_{1x}^{(t)}}f}(y)\|_2^2 + \|x-y\|_2^2\right) \\ &= \left(\frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}}\right)^2\left(\left(\frac{1}{(\hat{Q}_{1x}^{(t)})^2(\chi_{1x}^{(t)})^2} - 2\frac{1}{\hat{Q}_{1x}^{(t)}\chi_{1x}^{(t)}}\right)\left(\frac{1}{1+\sigma_1/\hat{Q}_{1x}^{(t)}}\right)^2 + 1\right)\|x-y\|_2^2 \\ &= \left(\frac{\hat{Q}_{1x}^{(t)}}{\hat{Q}_{2x}^{(t)}}\right)^2\left(\frac{(\hat{Q}_{2x}^{(t)})^2 - (\hat{Q}_{1x}^{(t)})^2}{(\hat{Q}_{1x}^{(t)}+\sigma_1)^2} + 1\right)\|x-y\|_2^2. \end{aligned} \tag{40}$$

Note that condition (44) ensures that the denominator in (47) is non-zero. We then have:

$$\lambda_{max}(\mathbf{W}_2^T\mathbf{W}_2) \leqslant \left(\frac{\hat{Q}_{2z}}{\chi_{2x}\hat{Q}_{1x}}\right)^2 \frac{\lambda_{max}(\mathbf{F}^T\mathbf{F})}{(\hat{Q}_{2x}+\hat{Q}_{2z}\lambda_{min}(\mathbf{F}^T\mathbf{F}))^2} \leqslant \left(\frac{\hat{Q}_{2z}(1+\frac{\hat{Q}_{1x}}{\hat{Q}_{2x}})}{\hat{Q}_{1x}}\right)^2 \lambda_{max}(\mathbf{F}^T\mathbf{F}).$$

This quantity can be bounded above by a constant independent of $\lambda_2, \hat{Q}_{2x}, \hat{Q}_{1z}$ for arbitrarily large $\hat{Q}_{2x}$. Let b_3 be such a constant. Then a sufficient condition for condition (45) to hold is:

$$\beta_2 > \lambda_{max}(\mathbf{P}_1) + b_3\lambda_{max}(\mathbf{P}_2) + \frac{b_1 b_3(\lambda_{max}(\mathbf{P}_2))^2}{\beta_1 - \beta_2\hat{Q}_{1z}b_2 - \lambda_{max}(\mathbf{P}_2)b_1}. \quad (48)$$

We see that β_1 must scale linearly with $\hat{Q}_{1z}$ which is one of the parameters that is made arbitrarily large. Then β_1 also needs to become arbitrarily large for the conditions to hold. We choose $\beta_1 = 2\beta_2\hat{Q}_{1z}b_2 + \lambda_{max}(\mathbf{P}_2)b_1$ for the rest of the proof. Condition (46) is then verified, and β_2 needs to be chosen according to condition (48), which becomes:

$$\beta_2 > \lambda_{max}(\mathbf{P}_1) + b_3\lambda_{max}(\mathbf{P}_2) + \frac{b_1 b_3\lambda_{max}^2(\mathbf{P}_2)}{\beta_2\hat{Q}_{1z}b_2}.$$

This has a bounded solution for large values of $\hat{Q}_{1z}$. We now turn to the second part of (43).

2) Conditions for $\tau^2\mathbf{P} - \mathbf{H}_1 + \mathbf{H}_2(\mathbf{B}^T\mathbf{P}\mathbf{B}+\mathbf{H}_3)^{-1}\mathbf{H}_2^T \succ 0$: We need to study the term $-\mathbf{H}_2(\mathbf{B}^T\mathbf{P}\mathbf{B}+\mathbf{H}_3)^{-1}\mathbf{H}_2^T$ (with the minus sign since the middle matrix is negative definite from conditions (44,45) which are now verified). As we will see, because of the form of $\mathbf{H}_2$, we don't need to explicitly compute the whole inverse. Let

$$\mathbf{Z} = -(\mathbf{B}^T\mathbf{P}\mathbf{B}+\mathbf{H}_3)^{-1} = \begin{bmatrix}\mathbf{Z}_1 & \mathbf{Z}_2\\ \mathbf{Z}_2^T & \mathbf{Z}_3\end{bmatrix}.$$

$\mathbf{Z}$ has the same block dimensions as $(\mathbf{B}^T\mathbf{P}\mathbf{B}+\mathbf{H}_3)$. We then have:

$$-\mathbf{H}_2(\mathbf{B}^T\mathbf{P}\mathbf{B}+\mathbf{H}_3)^{-1}\mathbf{H}_2^T = \mathbf{H}_2\mathbf{Z}\mathbf{H}_2^T = \begin{bmatrix}\beta_2^2\omega_2^4\mathbf{W}_3^T\mathbf{W}_4\mathbf{Z}_3\mathbf{W}_4^T\mathbf{W}_3 & \mathbf{0}_{M\times N}\\ \mathbf{0}_{N\times M} & \mathbf{0}_{N\times N}\end{bmatrix}.$$

We thus only need to characterize the lower right block of $\mathbf{Z}$. It is easy to see that conditions (44) and (45) also enforce that both the Schur complements associated with the upper left and lower right blocks of $-(\mathbf{B}^T\mathbf{P}\mathbf{B}+\mathbf{H}_3)$ are invertible, thus giving the following form for $\mathbf{Z}_3$ using the block matrix inversion lemma [47]:

$$\mathbf{Z}_3 = (\beta_1\mathbf{I}_N - \beta_2\omega_2^2\mathbf{W}_4^T\mathbf{W}_4 - \mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_1 - \mathbf{W}_1^T\mathbf{P}_2\mathbf{W}_2\mathbf{K}_2\mathbf{W}_2^T\mathbf{P}_2\mathbf{W}_1)^{-1}.$$

where $\mathbf{K}_2 = (\beta_1\mathbf{I}_M - \mathbf{P}_1 - \mathbf{W}_2^T\mathbf{P}_2\mathbf{W}_2)^{-1}$. We thus have the following upper bound on the largest eigenvalue of $\mathbf{Z}_3$:

$$\lambda_{max}(\mathbf{Z}_3) \leqslant \frac{1}{\beta_1 - \beta_2\hat{Q}_{1z}b_2 - \lambda_{max}(\mathbf{P}_2)b_1 - k},$$

where

$$k = \frac{b_1 b_3\lambda_{max}^2(\mathbf{P}_2)}{\beta_2 - \lambda_{max}(\mathbf{P}_1) - b_2\lambda_{max}(\mathbf{P}_2)}.$$

Using the prescription $\beta_1 = 2\beta_2\hat{Q}_{1z}b_2 + \lambda_{max}(\mathbf{P}_1)b_1$, we get:

$$\lambda_{max}(\mathbf{Z}_3) = \frac{1}{\beta_1\hat{Q}_{1z}b_2 - \frac{b_1 b_3\lambda_{max}^2(\mathbf{P}_2)}{\beta_1 - \lambda_{max}(\mathbf{P}_1) - b_2\lambda_{max}(\mathbf{P}_2)}} \leqslant \frac{b_4}{\hat{Q}_{1z}}$$

where b_4 is a constant independent of the arbitrarily large parameters $\lambda_2, \hat{Q}_{2x}, \hat{Q}_{1z}$. Thus $\lambda_{max}(\mathbf{Z}_3)$ can be made arbitrarily small by making λ_2 arbitrarily large.

We now want to find conditions to ensure

$$\tau^2\mathbf{P} - \mathbf{H}_1 + \mathbf{H}_2(\mathbf{B}^T\mathbf{P}\mathbf{B}+\mathbf{H}_3)^{-1}\mathbf{H}_2^T \succ 0,$$

which is equivalent to:

$$\tau^2\mathbf{P}_1 - \beta_2\omega_2^2\mathbf{W}_3^T\mathbf{W}_3 - \beta_2^2\omega_2^4\mathbf{W}_3^T\mathbf{W}_4\mathbf{Z}_3\mathbf{W}_4^T\mathbf{W}_3 \succeq 0$$
$$\tau^2\mathbf{P}_2 - \beta_1\omega_1^2\mathbf{I}_N \succeq 0. \quad (49)$$

We start with the upper matrix inequality, for which a sufficient condition is:

$$\tau^2\lambda_{min}(\mathbf{P}_1) - \beta_2\omega_2^2\lambda_{max}(\mathbf{W}_3^T\mathbf{W}_3) - \beta_2^2\omega_2^4\lambda_{max}(\mathbf{W}_3^T\mathbf{W}_3)\lambda_{max}(\mathbf{W}_4^T\mathbf{W}_4)\lambda_{max}(\mathbf{Z}_3) > 0.$$

Using the bounds from appendix G-C, we have:

$$\begin{aligned}&\omega_2^2\lambda_{max}(\mathbf{W}_3^T\mathbf{W}_3)\\ &\leqslant \left(\frac{\hat{Q}_{1z}}{\hat{Q}_{2z}}\right)^2\left(1+\frac{(\hat{Q}_{2z})^2-(\hat{Q}_{1z})^2}{(\hat{Q}_{1z}+\tilde{\lambda}_2)^2}\right)\lambda_{max}(\mathbf{W}_3^T\mathbf{W}_3)\\ &\leqslant \frac{2\tilde{\lambda}_2\hat{Q}_{1z}+\tilde{\lambda}_2^2+(\hat{Q}_{2z})^2}{(\hat{Q}_{1z}+\tilde{\lambda}_2)^2}\\ &\times\max\left(\left(1-\frac{\hat{Q}_{1z}}{\hat{Q}_{2x}}\lambda_{min}(\mathbf{F}^T\mathbf{F})\right)^2,\left(1-\frac{\hat{Q}_{1z}}{\hat{Q}_{2x}}\lambda_{max}(\mathbf{F}^T\mathbf{F})\right)^2\right)\\ &\leqslant \frac{1}{\hat{Q}_{1z}}\left(2\tilde{\lambda}_2+\frac{(\tilde{\lambda}_2^2+(\hat{Q}_{2z})^2)}{\hat{Q}_{1z}}\right)\\ &\times\max\left(\left(1-\frac{\hat{Q}_{1z}}{\hat{Q}_{2x}}\lambda_{min}(\mathbf{F}^T\mathbf{F})\right)^2,\left(1-\frac{\hat{Q}_{1z}}{\hat{Q}_{2x}}\lambda_{max}(\mathbf{F}^T\mathbf{F})\right)^2\right).\end{aligned}$$

Thus there exists a constant b_5, independent of $\lambda_2, \hat{Q}_{1z}, \hat{Q}_{2x}$ such that, for sufficiently large $\hat{Q}_{1z}$:

$$\omega_2^2\lambda_{max}(\mathbf{W}_3^T\mathbf{W}_3) \leqslant \frac{b_5}{\hat{Q}_{1z}}.$$

Remember that we had:

$$\omega_2^2\lambda_{max}(\mathbf{W}_4^T\mathbf{W}_4) \leqslant \hat{Q}_{1z}b_2,$$

which gives the following sufficient condition for the upper left block in (49):

$$\tau^2\lambda_{min}(\mathbf{P}_1) - \beta_2\frac{b_5}{\hat{Q}_{1z}} - \beta_2^2\frac{b_2 b_5 b_4}{\hat{Q}_{1z}} > 0.$$

A sufficient condition for the lower right block in (49) then reads:

$$\tau^2\lambda_{min}(\mathbf{P}_2) - \beta_1\omega_1^2 > 0,$$

where we have:

$$\begin{aligned}\beta_1\omega_1^2 &= \left(\frac{\hat{Q}_{1x}}{\hat{Q}_{2x}}\right)^2\left(1+\frac{(\hat{Q}_{2x})^2-(\hat{Q}_{1x})^2}{(\hat{Q}_{1x}+\lambda_2)^2}\right)\\ &\quad\times(2\beta_1\hat{Q}_{1z}b_2+\lambda_{max}(\mathbf{P}_2)b_1)\\ &= \frac{1}{\hat{Q}_{2x}}(\hat{Q}_{1x})^2\left(1+\frac{(\hat{Q}_{2x})^2-(\hat{Q}_{1x})^2}{(\hat{Q}_{1x}+\lambda_2)^2}\right)\\ &\quad\times\left(2\beta_1\frac{\hat{Q}_{1z}}{\hat{Q}_{2x}}b_2+\lambda_{max}(\mathbf{P}_2)\frac{b_1}{\hat{Q}_{2x}}\right).\end{aligned}$$

We remind the reader that $\hat{Q}_{1z}, \hat{Q}_{2x}$ grow linearly with λ_2. Thus the dominant scaling at large λ_2 is (exchanging $\hat{Q}_{2x}$ with $\hat{Q}_{1z}$ up to a constant) reads:

$$\beta_1\omega_1^2 \leqslant \frac{b_6}{\hat{Q}_{1z}},$$

where b_6 is a constant independent of the arbitrarily large quantities. The final condition becomes:

$$\begin{aligned}\tau^2\lambda_{min}(\mathbf{P}_1)-\beta_2\frac{b_5}{\hat{Q}_{1z}}-\beta_2^2\frac{b_2b_5b_4}{\hat{Q}_{1z}}>0\\ \tau^2\lambda_{min}(\mathbf{P}_2)-\frac{b_6}{\hat{Q}_{1z}}>0\end{aligned}$$

where we want $\tau < 1$. We now choose $\tau^2 = \tilde{c}/\hat{Q}_{1z}$ with a constant $\tilde{c}$ independent of $\lambda_2, \hat{Q}_{1z}, \hat{Q}_{2x}$ that verifies

$$\tilde{c} > \max\left(\frac{\beta_2 b_5+\beta_2^2 b_2 b_5 b_4}{\lambda_{min}(\mathbf{P}_1)}, \frac{b_6}{\lambda_{min}(\mathbf{P}_2)}\right),$$

such that:

$$\begin{aligned}\frac{\tilde{c}}{\hat{Q}_{1z}}\lambda_{min}(\mathbf{P}_1)-\beta_2\frac{b_5}{\hat{Q}_{1z}}-\beta_2^2\frac{b_2b_5b_4}{\hat{Q}_{1z}}>0\\ \frac{\tilde{c}}{\hat{Q}_{1z}}\lambda_{min}(\mathbf{P}_2)-\frac{b_6}{\hat{Q}_{1z}}>0.\end{aligned}$$

Since β_2 is bounded for large values of $\hat{Q}_{1z}$, and the b_i and c are constants independent of $\lambda_2, \hat{Q}_{2x}, \hat{Q}_{1z}$, we can then enforce $\tilde{c} < \hat{Q}_{1z}$ using the additional ridge penalty parametrized by λ_2 on the regularization to obtain $\tau < 1$ and a linear convergence rate proportional to $\sqrt{\tilde{c}/\lambda_2}$. We see that the eigenvalues of the matrix $\mathbf{P}$ are of little importance as long as they are non-vanishing. We choose $\mathbf{P}$ as the identity. In the statement of Lemma 3, we write c the exact constant which comes linking $\hat{Q}_{1z}$ to λ_2.

This proves Lemma 3.

APPENDIX H
ANALYTIC CONTINUATION

In this section, we prove the validity of the analytic continuation and approximation argument used to prove Theorem 1, under the required set of assumptions 1. According to Lemma 4, for any $\tilde{\lambda}_2 > 0$ and $\lambda_2 > \lambda_2^*$, any scalar pseudo-Lipschitz observable of order 2 ϕ, we have almost surely

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\phi(x_{0,i},\hat{x}_i(\lambda_2))=\mathbb{E}[\phi(x_0,\text{Prox}_{f/\hat{Q}_{1x}^{(t)}}(H_x))] \quad (50)$$

where $H_x = (\hat{m}_{1x}^* x_0 + \sqrt{\hat{\chi}_{1x}^*}\xi_{1x})/\hat{Q}_{1x}$ is defined in Theorem 1. We would like to show that this equality still holds for any $\lambda_2 > 0$. To do so we will show that, for a real analytic approximation of problem Eq.(2), both sides of Eq.(50) are real analytic in λ_2. We may then use the real analytic continuation theorem, as given in [48] to extend to any $\lambda_2 > 0$. We will treat the case $\lambda_2 = 0$ separately. In what follows, we will write the dependency in λ_2 of the estimator explicitly, i.e., $\hat{\mathbf{x}} = \hat{\mathbf{x}}(\lambda_2)$.

A. Real Analyticity of the Left Hand Side of Eq.(50)

We remind a useful characterization of real analytic functions from [48]:

Proposition 5 (Proposition 1.2.10 from [48]): Let $f \in \mathcal{C}^\infty(I)$ for some open interval I. The function f is in fact real analytic on I if and only if, for each $\alpha \in I$, there are an open interval J, with $\alpha \in J \subset I$, and finite constants $C > 0$ and $R > 0$ such that the derivatives of f satisfy:

$$|f^{(j)}(\alpha)| \leqslant C\frac{j!}{R^j}, \quad \forall\alpha \in J.$$

We also remind the formula for the higher order derivatives of a composition of two infinitely differentiable functions:

Proposition 6 (Faa di Bruno's Formula, [48] Theorem 1.3.2.): Consider two scalar functions f and g defined on an open interval $I \in \mathbb{R}$. Assume that both functions are infinitely differentiable on I and taking value in I. Then the derivatives of $h = g \circ f$ are given by

$$\begin{aligned}h^{(n)}(t) = \sum \frac{n!}{k_1!k_2!\dots k_n!}g^{(k)}(f(t))\\ \left(\frac{f^{(1)}(t)}{1!}\right)^{k_1}\left(\frac{f^{(2)}(t)}{2!}\right)^{k_2}\dots\left(\frac{f^{(n)}(t)}{n!}\right)^{k_n}\end{aligned}$$

where $k = k_1 + k_2 + \ldots + k_n$ and the sum is taken over all $k_1, k_2, \ldots, k_n$ for which $k_1 + 2k_2 + \ldots + nk_n = n$.

The following lemma establishes bounds on the higher order derivatives of $\hat{\mathbf{x}}(\lambda_2)$ with respect to λ_2.

Lemma 8: $\hat{\mathbf{x}}(\lambda_2)$ is infinitely differentiable w.r.t. λ_2 and, for any integer p, there exists a constant K' such that its elementwise p-th derivative, denoted $D_{\lambda_2}^{(p)}\hat{\mathbf{x}}(\lambda_2)$ verifies, almost surely

$$\frac{1}{N}\|D_{\lambda_2}^{(p)}\hat{\mathbf{x}}(\lambda_2)\|_2^2 \leqslant K' \quad (51)$$

Furthermore, $D_{\lambda_2}^{(p)}\hat{\mathbf{x}}(\lambda_2)$ is a Lipschitz function of $\hat{\mathbf{x}}(\lambda_2)$.

Proof: Recall the strongly convex problem, for any finite N,

$$\hat{\mathbf{x}}(\lambda_2,\tilde{\lambda}_2) = \underset{\mathbf{x}\in\mathcal{X}}{\arg\min}\,\tilde{g}(\mathbf{F}\mathbf{x},\mathbf{y}) + f(\mathbf{x}) + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2$$

where we absorbed $\tilde{\lambda}_2$ in $\tilde{g}$ as we are only interested in prolonging on λ_2.

The optimality condition then uniquely defines $\hat{\mathbf{x}}(\lambda_2)$ of each value of λ_2 and reads:

$$\mathbf{F}^\top\nabla\tilde{g}(\mathbf{F}\hat{\mathbf{x}}(\lambda_2),\mathbf{y}) + \nabla f(\hat{\mathbf{x}}(\lambda_2)) + \lambda_2\hat{\mathbf{x}}(\lambda_2) = 0.$$

The function $\mathbf{F}^\top\nabla\tilde{g}(\mathbf{F}\cdot,\mathbf{y}) + \nabla f(\cdot) + \lambda_2\cdot$ is real analytic in $\mathbb{R}^N$ and its Jacobian $\mathbf{F}^\top\mathcal{H}_{\tilde{g}}\mathbf{F} + \mathcal{H}_f + \lambda_2\mathbb{I}_N$ is non singular since f and $\tilde{g}$ are convex. The implicit function theorem [48]

then ensures that, at any finite $N > 0$, the function $\hat{\mathbf{x}}(\lambda_2)$ is elementwise real analytic in λ_2. We can now prove the lemma by induction.

a) Initialization: Owing to assumption 1, we have almost surely

$$\lim_{N\to\infty} \frac{1}{N}\|\hat{\mathbf{x}}(\lambda_2)\|_2^2 \leqslant K'$$

and the identity is a Lipshchitz function of $\hat{\mathbf{x}}(\lambda_2)$. The function of λ_2 defined by:

$$\lambda_2 \mapsto \nabla \tilde{g}(\mathbf{F}\hat{\mathbf{x}}(\lambda_2), \mathbf{y}) + \nabla f(\hat{\mathbf{x}}(\lambda_2)) + \lambda_2 \hat{\mathbf{x}}(\lambda_2)$$

is always zero valued from the definition of $\hat{\mathbf{x}}(\lambda_2)$, thus all its derivatives are zero. Taking the first derivative with respect to λ_2 yields:

$$(\mathbf{F}^T \mathcal{H}_{\tilde{g}}(\mathbf{F}\hat{\mathbf{x}}(\lambda_2), \mathbf{y})\mathbf{F} + \mathcal{H}_f(\hat{\mathbf{x}}(\lambda_2)) + \lambda_2 \mathbf{I}_N) D\hat{\mathbf{x}}(\lambda_2) + \hat{\mathbf{x}}(\lambda_2) = 0 \quad (52)$$

where D^p is the $(N \times 1)$ dimensional element-wise p-th differential of $\hat{\mathbf{x}}(\lambda_2)$. We then define the operator

$$\mathcal{O} : \begin{cases} \mathbb{R} \to \mathbb{R}^{N\times N} \\ \lambda_2 \mapsto \mathbf{F}^T \mathcal{H}_{\tilde{g}}(\mathbf{F}\hat{\mathbf{x}}(\lambda_2), \mathbf{y})\mathbf{F} + \mathcal{H}_f(\hat{\mathbf{x}}(\lambda_2)) + \lambda_2 \mathbf{I}_N. \end{cases}$$

We obtain a simple expression for $D\hat{\mathbf{x}}(\lambda_2)$

$$D\hat{\mathbf{x}}(\lambda_2) = -\mathcal{O}^{-1}(\lambda_2)\hat{\mathbf{x}}(\lambda_2).$$

Since f and g are convex, the operator norm of $\mathcal{O}^{-1}(\lambda_2)$ is bounded with probability one, and $D\hat{\mathbf{x}}(\lambda_2)$ is a Lipschitz function of $\hat{\mathbf{x}}(\lambda_2)$ where $\frac{1}{N}\|D\hat{\mathbf{x}}(\lambda_2)\|_2^2$ is almost surely bounded.

b) Induction step: Assume the property is verified up to $p-1$. For higher order derivatives, applying Leibniz's rule on Eq.(52) gives, denoting $\mathcal{O}^{(i)}(\lambda_2)$ the i-th derivative of $\mathcal{O}(\lambda_2)$, for the (p-1)-th derivative of (52):

$$\sum_{i=0}^{p-1} \binom{p-1}{i} \mathcal{O}^{(i)}(\lambda_2) D^{(p-i)}\hat{\mathbf{x}}(\lambda_2) + D^{(p-1)}\hat{\mathbf{x}}(\lambda_2) = 0,$$

such that

$$\sum_{i=1}^{p-1} \binom{p-1}{i} \mathcal{O}^{(i)}(\lambda_2) D^{(p-i)}\hat{\mathbf{x}}(\lambda_2) + \mathcal{O}(\lambda_2) D^{(p)}\hat{\mathbf{x}}(\lambda_2) + D^{(p-1)}\hat{\mathbf{x}}(\lambda_2) = 0.$$

We obtain the recursion on the differentials of $\hat{\mathbf{x}}(\lambda_2)$:

$$D^p\hat{\mathbf{x}}(\lambda_2) = -\mathcal{O}^{-1}(\lambda_2)\left(\sum_{i=1}^{p-1} \binom{p-1}{i} \mathcal{O}^{(i)}(\lambda_2) D^{(p-i)}\hat{\mathbf{x}}(\lambda_2) + D^{(p-1)}\hat{\mathbf{x}}(\lambda_2)\right),$$

where the matrix inverse $\mathcal{O}^{-1}(\lambda_2)$ is well defined for any $\lambda_2 > 0$ since f and g are convex. Using proposition 6, the assumption on the fast decay of the higher-order (larger than 2) derivatives of f and g, the bounded spectrum of the matrix $\mathbf{F}$, and the induction hypothesis, the operator norm of $\mathcal{O}^{(p)}(\lambda_2)$ is bounded with probability one for any $p \in \mathbb{N}$, $D^{(p)}\hat{\mathbf{x}}(\lambda_2)$ is a Lipschitz function of $\hat{\mathbf{x}}(\lambda_2)$ as a finite sum of Lipschitz functions of $\hat{\mathbf{x}}(\lambda_2)$, and its averaged squared norm is bounded almost surely. This concludes the induction. ■

Lemma 9: Under assumption 1, the function $\psi(\lambda_2)$ defined as

$$\psi : \mathbb{R} \to \mathbb{R}$$
$$\lambda_2 \to \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \phi(x_{0,i}, \hat{x}_i(\lambda_2))$$

is real analytic for $\lambda_2 > 0$.

Proof: Since ϕ is pseudo Lipschitz of order 2, there exists a constant C_ϕ such that, for any $x \in \mathbb{R}$, $\phi(x) \leqslant C_\phi(1 + x^2)$. Thus:

$$\lim_{N\to\infty} |\psi(\lambda_2)| \leqslant \lim_{N\to\infty} \frac{C_\phi}{N}(1 + \|\hat{\mathbf{x}}(\lambda_2)\|_2^2)$$

which is almost surely bounded. By assumption, the boundedness of ψ is enough to obtain its convergence. For the first derivative, the pseudo-Lipschitz property ensures that there exists a constant C'_ϕ such that, for any $x \in \mathbb{R}$,

$$|\frac{d\phi}{dx}(x)| \leqslant C'_\phi(1 + |x|).$$

Then

$$|\frac{d}{d\lambda_2}\phi(\hat{x}(\lambda_2))| \leqslant C'_\phi |\frac{d}{d\lambda_2}\hat{x}(\lambda_2)| \left(1 + |\hat{x}|(\lambda_2)\right),$$

so there exists a constant C'_ψ such that

$$\lim_{N\to\infty} D\psi(\lambda_2) \leqslant \lim_{N\to\infty} \frac{1}{N} C'_\psi \left(\|D\hat{\mathbf{x}}(\lambda_2)\|_2 + \|D\hat{\mathbf{x}}(\lambda_2)\|_2 \|\hat{\mathbf{x}}(\lambda_2)\|_2\right)$$

which is almost surely bounded. We have also proved in the previous lemma that $D\hat{\mathbf{x}}(\lambda_2)$ is a Lipschitz function of λ_2, thus $D\psi(\lambda_2)$ is a PL2 function of $\hat{\mathbf{x}}(\lambda_2)$ and its limit exists according to Assumption 1 (c). For the higher order derivatives, we use proposition 6 to obtain, for any coordinate $1 \leqslant i \leqslant n$:

$$|\frac{d^{(p)}}{d\lambda_2^{(p)}}\phi(\hat{x}_i(\lambda_2))| = \sum \frac{p!}{k_1!k_2!\dots k_p!}\phi^{(k)}\left(\hat{x}_i(\lambda_2)\right) \left(\frac{\hat{x}_i^{(1)}(\lambda_2)}{1!}\right)^{k_1} \left(\frac{\hat{x}_i^{(2)}(\lambda_2)}{2!}\right)^{k_2} \dots \left(\frac{\hat{x}_i^{(p)}(\lambda_2)}{p!}\right)^{k_p}.$$

The assumption on the higher order derivatives of ϕ from Theorem 1 and Lemma 8 implies that the term

$$\phi^{(k)}\left(\hat{x}_i(\lambda_2)\right) \left(\frac{\hat{x}_i^{(1)}(\lambda_2)}{1!}\right)^{k_1} \left(\frac{\hat{x}_i^{(2)}(\lambda_2)}{2!}\right)^{k_2} \dots \left(\frac{\hat{x}_i^{(p)}(\lambda_2)}{p!}\right)^{k_p}$$

has bounded absolute value with probability one, for all coordinates i. Using the characterization of real analytic functions and assumption 1 (c) from proposition 5, this concludes the proof. ■

B. Analytic Continuation to $(\tilde{\lambda}_2, \lambda_2) \in \mathbb{R}_+^* \times \mathbb{R}_+^*$

From assumption 1, the set of fixed point equations from Theorem 1 admit a unique solution for any $\lambda_2, \tilde{\lambda}_2$. Additionally, the implicit function theorem [48] can also be applied to the set of fixed point equations from Theorem 1 regarding the dependencies in $\lambda_2, \tilde{\lambda}_2$ to show that each quantity involved is real analytic in $\lambda_2, \tilde{\lambda}_2$. At this point, we have two analytic

functions, the observable and the one defined by the fixed point of the state evolution equations, that coincide for any $\lambda_2 \in [\lambda_2^*, +\infty[$ and any $\tilde{\lambda}_2 > 0$. We can now use the analytic continuation theorem [48] to show that these functions remain equal for any $\lambda_2 > 0$ and for $\tilde{\lambda}_2 > 0$. This concludes the proof of Lemma 5.

C. Real Analytic Approximation of Strongly Convex Problems

Consider

$$\hat{\mathbf{x}}_\epsilon(\lambda_2) = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min}\ \tilde{g}_\epsilon(\mathbf{F}\mathbf{x},\mathbf{y}) + f_\epsilon(\mathbf{x}) + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2$$

$$\hat{\mathbf{x}}(\lambda_2) = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min}\ \tilde{g}(\mathbf{F}\mathbf{x},\mathbf{y}) + f(\mathbf{x}) + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2$$

where g_ϵ, f_ϵ are real analytic approximations of the loss g and regularizer f verifying assumption 1(e). To relax the analytic approximation, we need to prove the following equality.

$$\lim_{\epsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\phi(\hat{x}_{\epsilon,i}(\lambda_2)) = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\phi(\hat{x}_i(\lambda_2)).$$

Under assumption 1 (c) and owing to the definition of PL2 functions, it is sufficient to prove

$$\lim_{\epsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\|\hat{\mathbf{x}}_\epsilon(\lambda_2) - \hat{\mathbf{x}}(\lambda_2)\|_2^2 = 0.$$

Denote $\mathcal{C}$ the cost function $\tilde{g}(\mathbf{F}.,\mathbf{y}) + f(.)$ and its real analytic counterpart $\mathcal{C}_\epsilon$ the cost function $\tilde{g}_\epsilon(\mathbf{F}.,\mathbf{y}) + f_\epsilon(.)$.

$$\forall \mathbf{x}\in\mathbb{R}^d \quad \lim_{\epsilon\to 0}\mathcal{C}_\epsilon(\mathbf{x}) = \mathcal{C}(\mathbf{x}).$$

Since minimizers of convex functions are fixed points of the corresponding proximity operators, it holds that

$$\begin{aligned}
&\frac{1}{N}\|\hat{\mathbf{x}}_\epsilon(\lambda_2) - \hat{\mathbf{x}}(\lambda_2)\|_2^2 \\
&= \frac{1}{N}\|\mathrm{Prox}_{\mathcal{C}_\epsilon(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}_\epsilon(\lambda_2)) - \mathrm{Prox}_{\mathcal{C}(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}(\lambda_2))\|_2^2 \\
&\leqslant \frac{1}{N}\|\mathrm{Prox}_{\mathcal{C}_\epsilon(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}_\epsilon(\lambda_2)) - \mathrm{Prox}_{\mathcal{C}_\epsilon(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}(\lambda_2))\|_2^2 \\
&+ \frac{1}{N}\|\mathrm{Prox}_{\mathcal{C}_\epsilon(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}(\lambda_2)) - \mathrm{Prox}_{\mathcal{C}(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}(\lambda_2))\|_2^2.
\end{aligned}$$

The results from appendix G-C.2 show that proximity operators of strongly convex functions are contractions, thus their exists a positive constant $L_{\lambda_2} < 1$ such that for any realisation of $\mathbf{F}, \mathbf{x}^0, \boldsymbol{\omega}_0$

$$\begin{aligned}
&\frac{1}{N}\|\hat{\mathbf{x}}_\epsilon(\lambda_2) - \hat{\mathbf{x}}(\lambda_2)\|_2^2 \leqslant \frac{1}{N}L_{\lambda_2}\|\hat{\mathbf{x}}_\epsilon(\lambda_2) - \hat{\mathbf{x}}(\lambda_2)\|_2^2 \\
&+ \frac{1}{N}\|\mathrm{Prox}_{\mathcal{C}_\epsilon(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}(\lambda_2)) - \mathrm{Prox}_{\mathcal{C}(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}(\lambda_2))\|_2^2.
\end{aligned}$$

Furthermore, the function $\mathrm{Prox}_{\mathcal{C}_\epsilon(.)+\frac{\lambda_2}{2}\|.\|_2^2}(.)$ converges uniformly to $\mathrm{Prox}_{\mathcal{C}(.)+\frac{\lambda_2}{2}\|.\|_2^2}(.)$ when $\epsilon\to 0$, and thus

$$\begin{aligned}
&\lim_{\epsilon\to 0}\lim_{N\to\infty}\frac{1}{N} \\
&\times \|\mathrm{Prox}_{\mathcal{C}_\epsilon(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}(\lambda_2)) - \mathrm{Prox}_{\mathcal{C}(.)+\frac{\lambda_2}{2}\|.\|_2^2}(\hat{\mathbf{x}}(\lambda_2))\|_2^2 = 0
\end{aligned}$$

which gives

$$\begin{aligned}
&\lim_{\epsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\|\hat{\mathbf{x}}_\epsilon(\lambda_2) - \hat{\mathbf{x}}(\lambda_2)\|_2^2 \\
&\qquad\leqslant L_{\lambda_2}\lim_{\epsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\|\hat{\mathbf{x}}_\epsilon(\lambda_2) - \hat{\mathbf{x}}(\lambda_2)\|_2^2.
\end{aligned}$$

Since $L_{\lambda_2} < 1$, this implies

$$\lim_{\epsilon\to 0}\lim_{N\to\infty}\frac{1}{N}\|\hat{\mathbf{x}}_\epsilon(\lambda_2) - \hat{\mathbf{x}}(\lambda_2)\|_2^2 = 0.$$

D. Continuous Extension to $\tilde{\lambda}_2 = 0$

Making the dependence on $\tilde{\lambda}_2$ explicit, define

$$\hat{\mathbf{x}}(\tilde{\lambda}_2,\lambda_2) = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min}\ g(\mathbf{F}\mathbf{x},\mathbf{y}) + f(\mathbf{x}) + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2 + \frac{\tilde{\lambda}_2}{2}\|\mathbf{F}\mathbf{x}\|_2^2$$

$$\hat{\mathbf{x}}(0,\lambda_2) = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min}\ g(\mathbf{F}\mathbf{x},\mathbf{y}) + f(\mathbf{x}) + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2.$$

Both cost functions defining $\hat{\mathbf{x}}(\tilde{\lambda}_2,\lambda_2), \hat{\mathbf{x}}(0,\lambda_2)$ are strongly convex for any $\lambda_2 > 0$. We can then use the same argument as in the previous subsection C to conclude

$$\lim_{\tilde{\lambda}_2\to 0}\lim_{N\to\infty}\frac{1}{N}\|\hat{\mathbf{x}}(\tilde{\lambda}_2,\lambda_2) - \hat{\mathbf{x}}(0,\lambda_2)\|_2^2 = 0.$$

E. Continuous Extension to $\lambda_2 = 0$

For $\tilde{\lambda}_2 = 0$, the estimator $\hat{\mathbf{x}}(\lambda_2)$ is still unique for any $\lambda_2 > 0$. We now need to study the limiting ridgeless estimator

$$\lim_{\lambda_2\to 0}\underset{\mathbf{x}\in\mathcal{X}}{\arg\min}\, g(\mathbf{F}\mathbf{x},\mathbf{y}) + f(\mathbf{x}) + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2$$

for functions f, g that may not be strictly convex. To do so we will use Theorem 26.20 from [40], which is reminded in appendix B, proposition 4. Under assumption 1 and since the l_2 norm is strongly convex thus uniformly convex, we have, denoting $\hat{\mathbf{x}}_0$ the unique least l_2 norm element in $\arg\min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{F}\mathbf{x},\mathbf{y}) + f(\mathbf{x})$,

$$\lim_{\lambda_2\to 0}\hat{\mathbf{x}}(\lambda_2) = \hat{\mathbf{x}}_0.$$

We can therefore uniquely define the continuous extension of any continuous observable ϕ of $\hat{\mathbf{x}}(\lambda_2)$ such that $\phi(\lambda_2 = 0) = \phi(\hat{\mathbf{x}}_0)$. Then this observable and the corresponding function implicitly defined by the set of fixed point equations are continuous on $[0,+\infty[$ and equal for any $\lambda_2 \in]0,+\infty[$, and thus also equal at $\lambda_2 = 0$ using the definition of continuity and the fact that $]0,+\infty[$ is dense in $[0,+\infty[$.

F. Real Analytic Approximation of Usual Cost Functions With Fast Decaying Higher-Order Derivatives

In this section, we show that any combination of the square, logistic and hinge loss with ℓ_1 or ℓ_2 verifies Assumption 1 (e), i.e. they can be approximated with real analytic functions whose second derivatives have higher-order derivatives that decrease faster than any polynomial. The square loss and ℓ_2 immediately verify these assumptions. Assuming $y = 1$ without loss of generality, the second derivative of the logistic loss is given by

$$g''(x) = \frac{\exp(x)}{(1+\exp(x))}.$$

All higher order derivatives will be a polynomial in $\exp(x)$ divided by a higher order polynomial in $\exp(x)$ plus one. Thus, for any sign of x, higher-order derivatives of the logistic loss will decrease exponentially fast when the absolute value of x goes to infinity. We now turn to the ℓ_1 penalty. Real

analytic approximations of functions may be constructed by considering their convolution with a Gaussian kernel, which is also known as the Weierstrass transform. Denoting $\mathcal{W}_\epsilon[f]$ the Weierstrass transform of a function f with parameter $\epsilon > 0$, we obtain for the ℓ_1 penalty

$$\mathcal{W}_\epsilon[|.|](x) = \frac{1}{\sqrt{2\pi\epsilon}} \int_{-\infty}^{+\infty} |u| \exp\left(-\frac{1}{2\epsilon}(u-x)^2\right) du$$
$$= \frac{1}{\sqrt{2\pi\epsilon}} \left(2\epsilon \exp\left(-\frac{1}{2\epsilon}x^2\right) + 2x \int_0^x \exp\left(-\frac{1}{2\epsilon}u^2\right) du\right)$$

whose second derivative reads

$$\frac{d^2}{dx^2}\mathcal{W}_\epsilon[|.|](x) = \frac{\sqrt{2}}{\sqrt{\pi\epsilon}} \exp\left(-\frac{1}{2\epsilon}x^2\right),$$

thus $\mathcal{W}_\epsilon[|.|)]$ is strongly convex and its higher order derivatives all decay faster than any finite order polynomial. A similar computation shows that, for the hinge loss,

$$\mathcal{W}_\epsilon[\max(0,1-.)](x)$$
$$= \frac{1}{\sqrt{2\pi\epsilon}} \int_{-\infty}^{+\infty} \max(0,1-u) \exp\left(-\frac{1}{2\epsilon}(u-x)^2\right) du$$
$$= \frac{1}{\sqrt{2\pi\epsilon}} \left((1-x)\sqrt{\frac{\pi\epsilon}{2}} + \epsilon \exp\left(-\frac{1}{2\epsilon}(1-x)^2\right) + (1-x)\int_0^x \exp\left(-\frac{1}{2\epsilon}(1-x)^2\right) du\right)$$

whose second derivative reads

$$\frac{d^2}{dx^2}\mathcal{W}_\epsilon[\max(0,1-.)](x) = \frac{1}{\sqrt{2\pi\epsilon}} \exp\left(-\frac{1}{2\epsilon}(1-x)^2\right).$$

Thus the hinge loss and ℓ_1 penalty verify Assumption 1 (e).

ACKNOWLEDGMENT

The authors would like to thank Andrea Montanari, Benjamin Aubin, Yoshiyuki Kabashima, and Lenka Zdeborová for discussions. They also thank three anonymous referees for their precise and valuable comments.

REFERENCES

[1] H. S. Seung, H. Sompolinsky, and N. Tishby, "Statistical mechanics of learning from examples," *Phys. Rev. A, Gen. Phys.*, vol. 45, no. 8, p. 6056, 1992.
[2] T. L. Watkin, A. Rau, and M. Biehl, "The statistical mechanics of learning a rule," *Rev. Modern Phys.*, vol. 65, no. 2, p. 499, 1993.
[3] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2001.
[4] M. Bayati and A. Montanari, "The LASSO risk for Gaussian matrices," *IEEE Trans. Inf. Theory*, vol. 58, no. 4, pp. 1997–2017, Apr. 2012.
[5] N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu, "On robust regression with high-dimensional predictors," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 36, pp. 14557–14562, 2013.
[6] D. Donoho and A. Montanari, "High dimensional robust M-estimation: Asymptotic variance via approximate message passing," *Probab. Theory Rel. Fields*, vol. 166, nos. 3–4, pp. 935–969, 2015.
[7] L. Zdeborová and F. Krzakala, "Statistical physics of inference: Thresholds and algorithms," *Adv. Phys.*, vol. 65, no. 5, pp. 453–552, 2016.
[8] P. Sur, Y. Chen, and E. J. Candès, "The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square," *Probab. Theory Rel. Fields*, vol. 175, pp. 487–558, Jan. 2019.
[9] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," *Ann. Statist.*, vol. 50, no. 2, pp. 949–986, Apr. 2022.
[10] S. Mei and A. Montanari, "The generalization error of random features regression: Precise asymptotics and the double descent curve," *Commun. Pure Appl. Math.*, vol. 75, no. 4, pp. 667–766, Apr. 2022.
[11] Y. Kabashima, "Inference from correlated patterns: A unified theory for perceptron learning and linear vector channels," *J. Phys., Conf. Ser.*, vol. 95, Jan. 2008, Art. no. 012001.
[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Nov. 2010.
[13] A. Manoel, F. Krzakala, M. Mezard, and L. Zdeborova, "Multi-layer generalized linear estimation," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 2098–2102.
[14] P. Schniter, S. Rangan, and A. K. Fletcher, "Vector approximate message passing for the generalized linear model," in *Proc. 50th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2016, pp. 1525–1529.
[15] M. Mézard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*, vol. 9. Singapore: World Scientific, 1987.
[16] M. Mezard and A. Montanari, *Information, Physics, and Computation*. London, U.K.: Oxford Univ. Press, 2009.
[17] E. Gardner and B. Derrida, "Three unfinished works on the optimal storage capacity of networks," *J. Phys. A, Math. Gen.*, vol. 22, no. 12, p. 1983, 1989.
[18] M. Opper, W. Kinzel, J. Kleinz, and R. Nehl, "On the ability of the optimal perceptron to generalise," *J. Phys. A, Math. Gen.*, vol. 23, no. 11, pp. L581–L586, Jun. 1990.
[19] M. Opper and W. Kinzel, "Statistical mechanics of generalization," in *Models of Neural Networks III*. Cham, Switzerland: Springer, 1996, pp. 151–209.
[20] T. P. Biehl, M. Caticha, N. Opper, and M. Villmann, "Statistical physics of learning and generalization," in *Adaptivity and Learning*. Berlin, Germany: Springer, 2003, pp. 77–88.
[21] Y. Kabashima, T. Wadayama, and T. Tanaka, "A typical reconstruction limit for compressed sensing based on Lp-norm minimization," *J. Stat. Mech., Theory Exp.*, vol. 2009, no. 9, Sep. 2009, Art. no. L09003.
[22] S. Ganguli and H. Sompolinsky, "Statistical mechanics of compressed sensing," *Phys. Rev. Lett.*, vol. 104, no. 18, 2010, Art. no. 188701.
[23] M. Advani and S. Ganguli, "An equivalence between high dimensional Bayes optimal inference and m-estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3378–3386.
[24] P. P. Mitra, "Compressed sensing and overparametrized networks: Overfitting peaks in a model of misparametrized sparse regression in the interpolation limit," Cold Spring Harbor Lab., Cold Spring Harbor, NY, USA, Tech. Rep., 2019.
[25] M. Emami, M. Sahraee-Ardakan, P. Pandit, S. Rangan, and A. Fletcher, "Generalization error of generalized linear models in high dimensions," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 2892–2901.
[26] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.
[27] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "Precise error analysis of regularized M-estimators in high dimensions," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5592–5628, Aug. 2018.
[28] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 45, pp. 18914–18919, Jul. 2009.
[29] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2011, pp. 2168–2172.
[30] C. Gerbelot, A. Abbara, and F. Krzakala, "Asymptotic errors for high-dimensional convex penalized linear regression beyond Gaussian matrices," in *Proc. Conf. Learn. Theory*, 2020, pp. 1682–1713.
[31] M. Mézard, "The space of interactions in neural networks: Gardner's computation with the cavity method," *J. Phys. A, Math. Gen.*, vol. 22, no. 12, p. 2181, 1989.
[32] Y. Kabashima, "A CDMA multiuser detection algorithm on the basis of belief propagation," *J. Phys. A, Math. Gen.*, vol. 36, no. 43, pp. 11111–11121, Oct. 2003.
[33] Y. Kabashima and S. Uda, "A bp-based algorithm for performing Bayesian inference in large perceptron-type networks," in *Proc. Int. Conf. Algorithmic Learn. Theory*. Cham, Switzerland: Springer, 2004, pp. 479–493.
[34] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," *IEEE Trans. Inf. Theory*, vol. 65, no. 10, pp. 6664–6684, Oct. 2019.

[35] T. Takahashi and Y. Kabashima, "Macroscopic analysis of vector approximate message passing in a model-mismatched setting," *IEEE Trans. Inf. Theory*, vol. 68, no. 8, pp. 5579–5600, Aug. 2022.
[36] A. K. Fletcher, S. Rangan, and P. Schniter, "Inference in deep networks in high dimensions," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 1884–1888.
[37] T. P. Minka, "A family of algorithms for approximate Bayesian inference," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2001.
[38] A. Fletcher, M. Sahraee-Ardakan, S. Rangan, and P. Schniter, "Expectation consistent approximate inference: Generalizations and convergence," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 190–194.
[39] P. Pandit, M. Sahraee-Ardakan, S. Rangan, P. Schniter, and A. K. Fletcher, "Inference with deep generative priors in high dimensions," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 336–347, May 2020.
[40] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, vol. 408. Cham, Switzerland: Springer, 2011.
[41] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, Jan. 2014.
[42] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2011.
[43] F. Gerace, B. Loureiro, F. Krzakala, M. Mézard, and L. Zdeborová, "Generalisation error in learning with random features and the hidden manifold model," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3452–3462.
[44] L. Lessard, B. Recht, and A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints," *SIAM J. Optim.*, vol. 26, no. 1, pp. 57–95, Jan. 2016.
[45] R. Nishihara, L. Lessard, B. Recht, A. Packard, and M. Jordan, "A general analysis of the convergence of ADMM," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 343–352.
[46] P. Giselsson and S. Boyd, "Linear convergence and metric selection for douglas-rachford splitting and ADMM," *IEEE Trans. Autom. Control*, vol. 62, no. 2, pp. 532–544, Feb. 2017.
[47] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2012.
[48] S. G. Krantz and H. R. Parks, *A Primer Real Analytic Functions*. Cham, Switzerland: Springer, 2002.
[49] A. M. Tulino and S. Verdú, *Random Matrix Theory and Wireless Communications*. Norwell, MA, USA: Now Publishers, 2004.

Cedric Gerbelot received the Engineering and M.Sc. degrees in physics from the Ecole Supérieure de Physique et de Chimie Industrielle, Paris, in 2018, the M.Sc. degree in applied mathematics from ENS Paris-Saclay in 2019, and the Ph.D. degree in mathematical physics and computer science from the Ecole Normale Supérieure de Paris in August 2022. He is currently a Courant Instructor at the Courant Institute of Mathematical Sciences, NYU, New York City, NY, USA.

Alia Abbara received the M.Sc. degree in theoretical physics in 2016 and the Ph.D. degree in statistical physics of inference at the crossroads of high-dimensional statistics from the Ecole Normale Supérieure de Paris in 2020. Since 2021, she has been a Postdoctoral Researcher at EPFL, Switzerland. Her research interests include mathematical models and statistical inference applied to complex biological systems.

Florent Krzakala is currently an EE and Physics Professor at EPFL, Switzerland.