Asymptotic Errors for High-Dimensional Convex Penalized Linear Regression beyond Gaussian Matrices

Cedric Gerbelot CEDRIC.GERBELOT@ENS.FR
Alia Abbara ALIA.ABBARA@ENS.FR
Florent Krzakala FLORENT.KRZAKALA@ENS.FR
Laboratoire de Physique de l'Ecole normale supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, F-75005 Paris, France

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We consider the problem of learning a coefficient vector $\mathbf{x_0} \in \mathbb{R}^{\mathbf{N}}$ from noisy linear observations $\mathbf{y} = \mathbf{F}\mathbf{x}_0 + \mathbf{w} \in \mathbb{R}^M$ in the high dimensional limit $M, N \to \infty$ with $\alpha \equiv M/N$ fixed. We provide a rigorous derivation of an explicit formula —first conjectured using heuristic methods from statistical physics— for the asymptotic mean squared error obtained by penalized convex regression estimators such as the LASSO or the elastic net, for a class of very generic random matrices corresponding to rotationally invariant data matrices with arbitrary spectrum. The proof is based on a convergence analysis of an oracle version of vector approximate message-passing (oracle-VAMP) and on the properties of its state evolution equations. Our method leverages on and highlights the link between vector approximate message-passing, Douglas-Rachford splitting and proximal descent algorithms, extending previous results obtained with i.i.d. matrices for a large class of problems. We illustrate our results on some concrete examples and show that even though they are asymptotic, our predictions agree remarkably well with numerics even for very moderate sizes.

Keywords: Linear models, Compressed sensing, Replica formula, Vector approximate message-passing, Proximal operator, LASSO, High-Dimensional problems, Asymptotics

1. Introduction

Solving a regression problem with convex penalty in high dimension is certainly one of the most fundamental questions in a number of disciplines, ranging from statistical learning to signal processing. We shall consider this standard quadratic minimization problem on a given input space $\mathcal{X} \subset \mathbb{R}^N$ with M samples:

$$\mathbf{x}^* = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min} \frac{1}{2}\|\mathbf{y} - \mathbf{F}\mathbf{x}\|_2^2 + f(\mathbf{x}) \tag{1}$$

where $\mathbf{F} \in \mathbb{R}^{M\times N}$ is a known data matrix (in statistics/machine learning Graybill (1976)) or a a known measurement matrix (in signal processing/compressed sensing Donoho (2006)), and f a proper, closed, convex and separable regularization function. Examples includes ridge Regression

Marquardt and Snee (1975), the LASSO Tibshirani (1996), or Elastic nets Zou and Hastie (2005). Here, we assume the vector $\mathbf{y}$ has been obtained according to a noisy linear process as

$$\mathbf{y} = \mathbf{F}\mathbf{x}_0 + \mathbf{w} \tag{2}$$

where all elements from the vector $\mathbf{x}_0 \in \mathbb{R}^N$ are identically and independently distributed (i.i.d.) according to an arbitrary given distribution $\phi_0(x_0)$, and $\mathbf{w} \in \mathbb{R}^M$ is an i.i.d. Gaussian white noise of zero mean and variance Δ_0, independent of $\mathbf{F}$ and $\mathbf{x}_0$. We shall present asymptotically exact expressions for the mean squared error on the recovery of $\mathbf{x}_0$, that is on the error

$$\text{MSE} = \mathbb{E}\left[\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2\right]. \tag{3}$$

Our asymptotic result will hold almost surely for a specific type of random matrices: we shall consider sequences of random matrices $\mathbf{F}$ with fixed aspect ratio $\alpha \equiv M/N$ as $M, N \to \infty$.

In a pioneering paper, Bayati and Montanari (2011b) considered this case for matrices $\mathbf{F}$ with independent Gaussian entries, and provided a rigorous derivation of an explicit formula for the asymptotic mean squared error of the LASSO estimator. Our goal here is to go beyond independent and Gaussian matrices, and to give instead an asymptotic formula for generic matrices. Our matrices $\mathbf{F}$ will be assumed to be rotationally invariant: their singular value decomposition can be written $\mathbf{F} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ where $\mathbf{U}, \mathbf{V}$ are Haar distributed (i.e. uniformly sampled over the orthogonal group) and $\mathbf{D}$ is an arbitrary diagonal matrix containing the singular values of $\mathbf{F}$. While this setting is certainly specific, there is a long standing tradition for such problems in signal processing Rangan et al. (2009), statistical physics Guo and Verdú (2005), random matrix theory Guionnet (2009) and communications theory Tulino et al. (2004). From the point of view of statistical learning, this model allows to give "typical-case" results, that represent an alternative and appealing approach to the worst-case analysis Mohri et al. (2018). Although the latter is both mathematically rigorous and robust, it sometimes fails to predict interesting behaviors, like the recently observed double descent effects in generalization performance curves in overparametrized neural networks Belkin et al. (2019). Over the last few decades a considerable body of theoretical work has been made on "typical-case" scenarios, especially within the framework of statistical mechanics Seung et al. (1992); Watkin et al. (1993); Mézard and Montanari (2009); Advani et al. (2013); Zdeborová and Krzakala (2016) and this direction is currently witnessing a burst of activity, see e.g. Sur et al. (2019); Hastie et al. (2019); Mei and Montanari (2019).

1.1. Main contributions

Our main contributions are the following:

- We provide an analytical formula for the reconstruction error of problem (1) in the asymptotic setup, for all convex penalties (including for instance LASSO and Elastic net), for all rotationally invariant sequences of matrices $\mathbf{F}$, extending the results of Bayati and Montanari (2011b); El Karoui et al. (2013) beyond Gaussian matrices.
- In doing so, we give a mathematically rigorous proof of the replica formula obtained heuristically from statistical physics for this problem Rangan et al. (2009); Kabashima et al. (2012); Kabashima and Vehkaperä (2014). To the best of our knowledge, this is the first proof of such formulas for generic rotationally invariant matrices.

- Our proof analysis has an interest of its own, and builds on a detailed mapping between proximal descent algorithms and maximum a posteriori forms of message-passing algorithms. In particular, it gives upper bounds on the convergence rates of an oracle version of vector approximate message-passing Rangan et al. (2019), and discusses a simple method to enforce convergence.
- Finally, we also show a rigorous proof for the statistical distribution of the estimator defined by (1) for sufficiently strongly convex f, and conjecture its validity for any convex f.

1.2. Related works

Asymptotic distribution of M-estimators and reconstruction error — The initial effort to prove asymptotic reconstruction error in this type of problem originates in Bayati and Montanari (2011b) for the LASSO with Gaussian matrices. Their proof is based on a sequence of specifically designed iterates, whose statistical properties are analytically tracked in an asymptotic setting. The key idea is built on a modification of the celebrated iterative soft thresholding algorithm (ISTA) Daubechies et al. (2004) inspired by statistical physics. This additional term allows the exact computation of the conditional expectation of key quantities at each iteration based on the σ-algebra generated by the previous observations. This idea, initially presented for spin glasses in Bolthausen (2009), was later transposed to statistics as a leave-one-out method for ridge regularized M-estimators in El Karoui et al. (2013), among others. In a recent paper Sur et al. (2019), the probabilistic setting with a converging sequence is once again used to prove the statistical distribution of logistic regression problems. This technique, now part of the modern theoretical tools of learning theory and high-dimensional statistics, is yet to be extended to problems involving correlated matrices.

Replica heuristic and their rigorous proofs — As the reconstruction error of problems akin to (1) is of interest in the machine learning and signal processing literature, so are overlap parameters between estimators and underlying truth in statistical physics. In the statistical physics literature, these quantities are calculated using the reknown replica method Mézard et al. (1987), a powerful heuristic calculation of the log partition function of a Bayesian problem based on the identity $\log Z = \lim_{n\to 0} \frac{Z^n - 1}{n}$. This method has led to a number of predictions in various fields of computer science Mézard and Montanari (2009), and in particular in machine learning Seung et al. (1992); Watkin et al. (1993) regarding the generalization error of single neurons Opper and Kinzel (1996), support vector machines Dietrich et al. (1999) or fundamental quantities such as capacity of neural nets Gardner (1988). Although it is adaptable to many machine learning problems, see for example Zdeborová and Krzakala (2016) for a review, the replica method suffers from the fact that it is a non-rigorous approach. While replica calculations are often restricted to i.i.d. problems, they have been extended to rotationally invariant matrices, see for instance Kabashima and Vehkaperä (2014), using the Harish-Chandra-Itzykson-Zuber Collins (2003) formula. A substantial effort has been dedicated to prove the replica formula in specific settings, for problems originating in both statistical physics Talagrand (2003) and machine learning, such as low rank matrix factorization Dia et al. (2016). Generic methods have also been proposed based on the Guerra interpolation technique Guerra (2003), and later extended to modern Bayesian inference Barbier and Macris (2019). In particular, rigorous proof for the replica formula in Bayes MMSE version of solving the inverse problem (2) has been given in Barbier et al. (2016); Reeves and Pfister (2016); Barbier et al. (2019) for Gaussian matrices, and in Barbier et al. (2018) for (a large part of) rotationally invariant matrices.

Message-passing algorithms — Both the replica approach and ongoing body of work on the asymptotic distribution of M-estimators can be linked to variational inference Wainwright et al. (2008) through approximate message-passing algorithms Mézard and Montanari (2009). This family of algorithms is a statistical physics inspired variant of belief propagation, where local beliefs are approximated by Gaussian distributions. In a probabilistic framework, this leads to a powerful alternative to Monte Carlo methods that scale well with data dimension. One of the key properties of these algorithms are the so-called state evolution equations, an iterative scalar equivalent model which allows to track the asymptotic statistical properties of the iterates. A series of groundbreaking papers initiated with Bayati and Montanari (2011a) proved the exactness of these equations in the large system limit, and extended the method to treat nonlinear problems Rangan (2011) and handle rotationally invariant matrices Rangan et al. (2019). The latter involves vector approximate message-passing, and lies at the center of our approach.

The maximum a posteriori (MAP) forms of approximate message-passing algorithms are closely linked to proximal descent methods Parikh et al. (2014). As pointed out above, the original MAP approximate message-passing amounts to writing the ISTA with a second order correction term Montanari et al. (2012) based on mean values of previous iterates, sometimes referred to as the Onsager reaction term in theoretical physics. Vector-approximate message-passing and the related class of expectation consistent inference algorithms Minka (2001), Opper and Winther (2005); Fletcher et al. (2016) yield adaptative versions of Douglas-Rachford/ADMM Parikh et al. (2014) descent methods where the step sizes match the local curvature of the cost function at each iteration. The overall result is a family of faster algorithms than proximal descent ones, with asymptotically exact analytical forms for the statistical properties of the iterates. Their main drawback is the restricted class of matrices to which they are applicable, and the somewhat more complicated structure they present due to the adaptative terms. Although this impedes message-passing methods from becoming mainstream optimization procedures, they remain formidable theoretical tools.

2. Main Results

Our main result is a rigorous proof of a replica conjecture left open by Rangan et al. (2009); Vehkaperä et al. (2016). Here we state our main theorems, show how they agree with simulations at finite size and give a brief sketch of proof. The framework of assumptions is the same as the one introduced in Bayati and Montanari (2011a) and later used in Rangan et al. (2019). We provide details on this framework in appendix E.

2.1. Main theorem

Theorem 1 *Consider problem* (1) *with a proper closed, convex and separable* f*. Consider that the empirical distributions of the underlying truth* $\mathbf{x}_0$ *and singular values of the rotationally invariant sensing matrix respectively converge with second order moments, as defined in appendix E, to given distributions* p_{X_0} *and* p_S*. Assume that the distribution* p_S *is non-trivial and has compact support. Finally consider the limit* $M, N \to \infty$ *with fixed ratio* $M/N = \alpha$*. Then the average mean squared error* $\mathrm{MSE} = \frac{1}{N}\mathbb{E}\left[\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2\right]$ *for the estimator prescribed by* (1) *is given by the fixed point* $\tilde{E}$ *of*

the equations:

$$\tilde{V} = \mathbb{E}\left[\frac{1}{\mathcal{R}_{\mathbf{C}}(-\tilde{V})}\mathrm{Prox}'_{f/\mathcal{R}_{\mathbf{C}}(-\tilde{V})}\left(x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}\left(-\tilde{V}\right)}\sqrt{\left(\tilde{E} - \Delta_0\tilde{V}\right)\mathcal{R}'_{\mathbf{C}}\left(-\tilde{V}\right) + \Delta_0\mathcal{R}_{\mathbf{C}}\left(-\tilde{V}\right)}\right)\right] \tag{4a}$$

$$\tilde{E} = \mathbb{E}\left[\left\{\mathrm{Prox}_{f/\mathcal{R}_{\mathbf{C}}(-\tilde{V})}\left(x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}\left(-\tilde{V}\right)}\sqrt{\left(\tilde{E} - \Delta_0\tilde{V}\right)\mathcal{R}'_{\mathbf{C}}\left(-\tilde{V}\right) + \Delta_0\mathcal{R}_{\mathbf{C}}\left(-\tilde{V}\right)}\right) - x_0\right\}^2\right], \tag{4b}$$

where $\mathbf{C} = \mathbf{F}^T\mathbf{F}$, $\mathcal{R}_{\mathbf{C}}$ is the R-transform with respect to the spectral distribution of $\mathbf{F}^T\mathbf{F}$, which is defined in appendix A, and expectations are over $z \sim \mathcal{N}(0,1)$ and $x_0 \sim p_{X_0}$. Prox is the proximal operator defined as:

$$\forall\gamma \in \mathbb{R}^+, x, y \in \mathbb{R} \quad \mathrm{Prox}_{\gamma f}(y) \equiv \arg\min_x \left\{f(x) + \frac{1}{2\gamma}(x-y)_2^2\right\}. \tag{5}$$

Additionally, for any instance of problem (1), consider the regularized problem where f is replaced by $h = f + \frac{\lambda_2}{2}\|.\|_2^2$ and $\mathbf{x}^*_{\lambda_2}$ the corresponding solution. We then have the following result on the element wise distribution of this solution:

$$\exists\ \lambda_2^*\ \textit{s.t.}\ \forall\ \lambda_2 > \lambda_2^*:$$
$$x^*_{\lambda_2} \sim \mathrm{Prox}_{h/\mathcal{R}_{\mathbf{C}}(-\tilde{V})}\left(x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}\left(-\tilde{V}\right)}\sqrt{\left(\tilde{E} - \Delta_0\tilde{V}\right)\mathcal{R}'_{\mathbf{C}}\left(-\tilde{V}\right) + \Delta_0\mathcal{R}_{\mathbf{C}}\left(-\tilde{V}\right)}\right). \tag{6}$$

For completeness, the replica computation leading to (4), that appeared in Rangan et al. (2009); Vehkaperä et al. (2016), is given in appendix A where we used the notations of Krzakala et al. (2012). While we believe and conjecture that the second property (6) holds for any $\lambda_2 \geqslant 0$, we could prove it for an open subset of λ_2. The asymptotic error from Theorem 1 can be equivalently written:

Theorem 2 *Under the assumptions presented above, the average mean squared error of $\mathbf{x}^*$ is equivalently given by the fixed point of the state evolution equations of vector approximate message-passing Rangan et al. (2019).*

$$\alpha_{1k} = \mathbb{E}\left[\mathrm{Prox}'_{\frac{1}{A_{1k}}f}(x_0 + P_{1k})\right] \qquad V_{1k} = \frac{\alpha_{1k}}{A_{1k}} \tag{7a}$$

$$A_{2k} = \frac{1}{V_{1k}} - A_{1k} \qquad \tau_{2k} = \frac{1}{(1-\alpha_{1k})^2}\left[\mathcal{E}_1(A_{1k},\tau_{1k}) - \alpha_{1k}^2\tau_{1k}\right] \tag{7b}$$

$$\alpha_{2k} = \mathbb{E}\left[\frac{A_{2k}}{\lambda_{\mathbf{C}} + A_{2k}}\right] \qquad V_{2k} = \frac{\alpha_{2k}}{A_{2k}} \tag{7c}$$

$$A_{1,k+1} = \frac{1}{V_{2k}} - A_{2k} \qquad \tau_{1,k+1} = \frac{1}{(1-\alpha_{2k})^2}\left[\mathcal{E}_2(A_{2k},\tau_{2k}) - \alpha_{2k}^2\tau_{2k}\right] \tag{7d}$$

where $\mathcal{E}_1$ and $\mathcal{E}_2$ are function defined by:

$$\mathcal{E}_1(A_{1k},\tau_{1k}) = \mathbb{E}\left[\left(\mathrm{Prox}_{\frac{1}{A_{1k}}f}(x_0 + P_{1k}) - x_0\right)^2\right], \quad \mathcal{E}_2(A_{2k},\tau_{2k}) = \mathbb{E}\left[\frac{\Delta_0\lambda_{\mathbf{C}} + \tau_{2k}A_{2k}^2}{(\lambda_{\mathbf{C}} + A_{2k})^2}\right] \tag{8}$$

and

$$P_{1k} \sim \mathcal{N}(0,\tau_{1k}). \tag{9}$$

At the fixed point, $\mathcal{E}_1 = \mathcal{E}_2 = \frac{1}{N}\mathbb{E}\left[\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2\right]$.

When solving these equations numerically with an explicit distribution of singular values, we found that Theorem 2 was sometimes better suited for numerical evalutation than Theorem 1 as it gives a more stable numerical scheme upon (damped) iteration. Both theorems draw a strong connection between asymptotic formulas for the MSE of Problem 1 and the various transforms used in free probability and random matrix theory Tulino et al. (2004).

2.2. Applications and numerical experiments

We compare the result of Theorem 2 with numerics on two typical problems. In both, the underlying truth vector $\mathbf{x_0}$ is an i.i.d. one where each element is pulled from a Gauss-Bernoulli distribution with sparsity parameter $\rho \in \mathbb{R}_+$:

$$\phi_0(x_0) = (1-\rho)\delta(x_0) + \rho \frac{1}{\sqrt{2\pi}} \exp\left(-x_0^2/2\right), \tag{10}$$

and the training vector $\mathbf{y}$ is obtained from (2). All experimental points are obtained using the Scikit-learn Pedregosa et al. (2011) implementation of the LASSO, which uses a coordinate descent method with duality gap convergence control Friedman et al. (2010); Kim et al. (2007).

Signal recovery with row orthogonal matrices — In the first model, we consider a setting popular in signal processing and use row orthogonal matrices. Such random matrices are very similar to subsampled Fourier and Hadamard matrices, and play a fundamental role in e.g. compressed sensing Tsaig and Donoho (2006) and communication Guo and Verdú (2005). We aim to recover the underlying sparse vector using a LASSO regression and tune the regularization parameter. We want to compare the performance using an i.i.d. Gaussian matrix pulled from $\mathcal{N}(0, \frac{1}{N})$ and a row-orthogonal one, i.e. where the singular values of $\mathbf{F}$ are set to one, which gives the following distribution for the eigenvalues of $\mathbf{C}$:

$$\lambda_{\mathbf{C}} \sim \max(0, 1-\alpha)\delta(0) + \min(1, \alpha)\delta(1). \tag{11}$$

We take $M, N = 200, 100$ ($\alpha = 2$), $\Delta_0 = 0.01$ and $\rho = 0.3$. Each point is an average over 10^4 realizations. The error bars in this case are vanishingly small ($\sim 10^{-5}$). We see that an excellent agreement is obtained with the asymptotics of Theorem 1 although the simulation matrices are rather small, indicating that the prediction remains very good at finite values of M, N.

Overparametrization and double descent — In the second setup, we consider the effect of the aspect ratio on the reconstruction performance of a sparse vector. We want to reproduce the double descent phenomenon that was observed and discussed recently in several papers Belkin et al. (2019); Hastie et al. (2019); Mitra (2019); Mei and Montanari (2019) in linear regression (but appeared in some form already in Opper and Kinzel (1996)). In order to provide a minimal model of such a phenomenon, we follow the intuition proposed in Advani and Saxe (2017), underlining that the eigenvalue distribution of $\mathbf{C}$ must be divergent (but still be integrable) at $\lambda = 0$ for $\alpha = 1$. While the Marchenko-Pastur density is typical of Gaussian data, we can here use random matrices with any spectrum. We choose 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 $\mathbf{C}$:

$$\lambda_{\mathbf{C}} \sim \max(0, 1-\alpha)\delta(0) + \min(1, \alpha) \left(\frac{1}{2((1+\alpha)^2 - (1-\alpha)^2)} \mathbb{I}_{\{\sqrt{y} \in [(1-\alpha)^2, (1+\alpha)^2]\}} \frac{1}{\sqrt{y}} \right), \tag{12}$$

where $\mathbb{I}$ is the indicator function. Our results are shown in Fig. 2 using $M = \lfloor \alpha N \rfloor, N = 250$, $\Delta_0 = 0.05$, for two values of the regularization parameter $\lambda_1 = 10^{-4}, 10^{-1}$. Each point is an average over a hundred realizations. We recover the double descent with the vanishingly small regularization (blue curve). Note that the error peak can be moved to any point p on the x-axis by taking $\mathcal{U}(\left[(p-\alpha)^2, (p+\alpha)^2\right])$. Multiple descents can also be obtained by adding several distributions of the form (12) with different shifts p. Augmenting the regularization to reach a realistic LASSO, however, is found to remove the error peak (green curve). As before, one observed striking agreement between the asymptotics and the simulation. Our formulas generalize here the results of Mitra (2019) for any distribution of singular values.

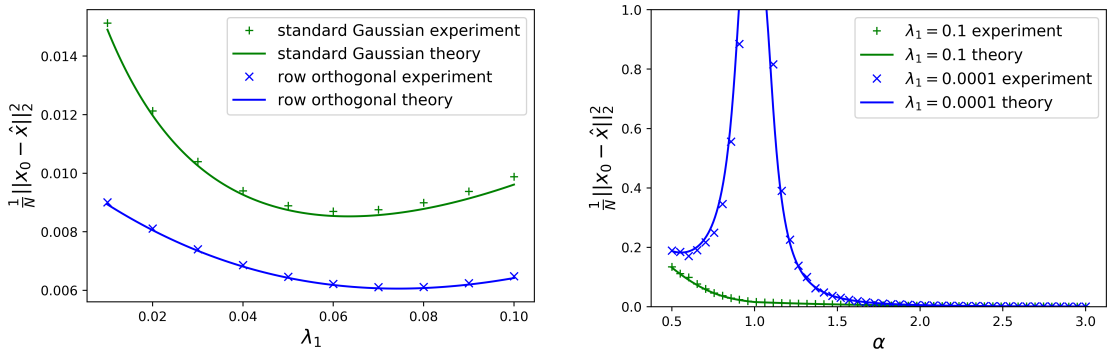

Figure 1: Reconstruction error for different realization of model (1) using the asymptotic formula and numerical simulation with scikit-learn Pedregosa et al. (2011). The asymptotic predictions are observed to be extremely accurate even at finite low dimension. Left: Reconstruction error for a sparse vector using Gaussian and row-orthogonal matrices with LASSO as regularization function using $M, N = 200, 100$. Right: An example of a double descent Belkin et al. (2019) behavior and its overfitting peak at $\alpha = M/N = 1$ as a function of the sampling ratio. Regularization (here ℓ_1) explicitly removes the peak (green) to give a smooth curve.

2.3. Sketch of Proof

We briefly sketch here the proof techniques that we use. Consider the following modification of problem (1), where an additional ℓ_2 penalty is added to enforce strong convexity and the potentially non-differentiable penalty function f is replaced by its twice differentiable approximation $\tilde{f}$, using for example Azagra et al. (2007):

$$\mathbf{x}^*_{\lambda_2} = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min} \frac{1}{2}\|\mathbf{y} - \mathbf{F}\mathbf{x}\|_2^2 + \tilde{f}(\mathbf{x}) + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2. \tag{13}$$

The main idea behind the proof is a variation of the approach pioneered in Bayati and Montanari (2011a), and leverages on the results of Rangan et al. (2019) who computed the errors obtained by vector approximate message-passing (VAMP) on this problem. The key steps of the proof are: (i) find a sequence whose iterates can be statistically characterized by equations that match (4) at their

fixed point. (ii) show that the sequence converges to its fixed point and that this fixed point solves problem (13). These first two steps prove the asymptotic error for the estimator (13). (iii) show that the dependence in the additional ridge parameter λ_2 allows an analytic continuation of the result.

3. Vector approximate message-passing (VAMP)

We briefly present the VAMP algorithm Rangan et al. (2019) along with its properties and oracle version that will be fundamental for our proof, which builds on results established mostly in Rangan et al. (2019) and Fletcher et al. (2016). VAMP is also linked with the expectation-propagation strategy Minka (2001), as well as other algorithms Cakmak et al. (2014); Ma and Ping (2017). Rangan et al. (2019), however, has the significant trait to provide rigourously derived state evolution equations

Maximum a posteriori formulation of VAMP — In a Bayesian framework, problem (1) corresponds to a maximum a posteriori (MAP) problem. The MAP formulation of vector approximate message-passing for (1) reads:

Choose initial A_{10} *and isotropically distributed* $\mathbf{B}_{10}$

$$\hat{\mathbf{x}}_{1k} = \mathrm{Prox}_{\frac{1}{A_{1k}}f}\left(\frac{\mathbf{B}_{1k}}{A_{1k}}\right) \qquad \hat{\mathbf{x}}_{2k} = (\mathbf{F}^T\mathbf{F} + A_{2k}\mathrm{Id})^{-1}(\mathbf{F}^{\mathrm{T}}\mathrm{y} + \mathbf{B}_{2\mathrm{k}}) \tag{14a}$$

$$V_{1k} = \frac{\langle \mathrm{Prox}'_{\frac{1}{A_{1k}}f}\rangle}{A_{1k}} \qquad V_{2k} = \frac{1}{N}\mathrm{Tr}\left[(\mathbf{F}^T\mathbf{F} + A_{2k}\mathrm{Id})^{-1}\right] \tag{14b}$$

$$A_{2k} = \frac{1}{V_{1k}} - A_{1k} \qquad A_{1,k+1} = \frac{1}{V_{2k}} - A_{2k} \tag{14c}$$

$$\mathbf{B}_{2k} = \frac{\hat{\mathbf{x}}_{1k}}{V_{1k}} - \mathbf{B}_{1k} \qquad \mathbf{B}_{1,k+1} = \frac{\hat{\mathbf{x}}_{2k}}{V_{2k}} - \mathbf{B}_{2k} \tag{14d}$$

where $\langle\cdot\rangle$ is an element-wise averaging operator $\langle\mathbf{x}\rangle = \frac{1}{N}\sum_{i=1}^{N} x_i$, and the vector valued proximal operator is defined as :

$$\forall\gamma \in \mathbb{R}^+, \mathbf{x}, \mathbf{y} \in \mathcal{X} \quad \mathrm{Prox}_{\gamma f}(\mathbf{y}) \equiv \arg\min_{\mathbf{x}}\left\{f(\mathbf{x}) + \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{y}\|_2^2\right\}. \tag{15}$$

It is akin to a projection on the level sets of f tuned with the parameter γ and can be evaluated even when the objective function is non-differentiable. Proximal descent methods enjoy a long lasting success in machine learning and signal processing Combettes and Pesquet (2011) because of their stability, simplicity to implement and solid theoretical anchoring, notably from a monotone operator theory point of view Bauschke et al. (2011). A popular algorithm for solving composite convex optimization problems of the form $\arg\min_{\mathbf{x}}\{f(\mathbf{x})+g(\mathbf{x})\}$ is the Douglas-Rachford splitting method Parikh et al. (2014), which roughly amounts to successively applying the proximal of f and the one of g. It is shown in Fletcher et al. (2016), a connection pursued in Manoel et al. (2018), that VAMP is akin to a Douglas-Rachford descent with parameters that adapt to the local curvature of the cost function.

The state evolution equations of VAMP give the statistical distribution of the iterates in (14). The quantities $\mathbf{B_1}$ and $\mathbf{B_2}$ behave as noisy Gaussian estimates of $\mathbf{x}_0$:

$$\mathbf{B}_{1k} = A_{1k}(\mathbf{x}_0 + \mathbf{P}_{1k}) \qquad \mathbf{B}_{2k} = A_{2k}(\mathbf{x}_0 + \mathbf{P}_{2k}), \tag{16}$$

$$\text{where} \quad P_{1k} \sim \mathcal{N}(0, \tau_{1k}) \qquad P_{2k} \sim \mathcal{N}(0, \tau_{2k}) \tag{17}$$

with τ_{1k} and τ_{2k} the variances of the estimates. The other key quantities are the variances V_{1k} and V_{2k} of the estimates $\hat{\mathbf{x}}_{1k}$ and $\hat{\mathbf{x}}_{2k}$, and their respective MSEs $\mathcal{E}_1$ and $\mathcal{E}_2$ from Theorem 2, which can also be written (appendix C from Rangan et al. (2019)):

$$\mathcal{E}_2 = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_{x_0, P_{2k}} \left[\| (\mathbf{F}^T\mathbf{F} + A_{2k}\mathrm{Id})^{-1}(\mathbf{F}^{\mathrm{T}}\mathrm{y} + \mathbf{B}_{2\mathrm{k}}) - \mathbf{x}_0 \|^2 \right] \tag{18}$$

The thresholds A_{1k} and A_{2k} of the proximal operators are adjusted to the variance of the noisy estimates of the teacher vector. The full state evolution (SE) equations are then given by (7).

These equations can be solved analytically if the teacher distribution $\phi(x_0)$ is known. In practice, all the averages are empirical and τ_{10} is initialized with the empirical variance of $\mathbf{B}_{10}$. Note that the SE equations only hold if VAMP is initialized with an isotropically distributed vector $\mathbf{B}_{10}$ which empirically converges with second order moment. Three additional assumptions on the denoiser function are required for the state evolution theorem in Rangan et al. (2019) to hold. These are automatically verified in the convex MAP case, as properties of the proximal mapping. This is reminded in appendix E. The next lemma characterizes the fixed point of (14).

Lemma 3 *For any pair (A_1, A_2) such that $A_1 + A_2 = \frac{1}{V}$, the fixed point of iterations (14) solves the minimization problem (1).*

Proof. See appendix C.1.

From VAMP to oracle-VAMP — We remind the reader that our goal is to find a convergent sequence of VAMP iterates that obey the state evolution equations, and reaches a fixed point solving (1). To simplify the analysis, we want to choose fixed values for the variance parameters of iterations (14), i.e. A_1, A_2, V_1 and V_2. To do so, we initialize A_1 in its value given by the fixed point of the state evolution equations. Taking $\mathbf{B_{10}}$ an isotropically distributed vector, VAMP will then yield iterates that obey the SE equations while keeping constant variance parameters. Equations (14) turn into a simpler algorithm, that we call oracle-VAMP:

$$\hat{\mathbf{x}}_{1k} = \mathrm{Prox}_{\frac{1}{A_1}f}\left(\frac{\mathbf{B}_{1k}}{A_1}\right) \qquad \hat{\mathbf{x}}_{2k} = \mathrm{Prox}_{\frac{1}{2A_2}||y-Fx||_2^2}\left(\frac{\mathbf{B}_{2k}}{A_2}\right) \tag{19a}$$

$$\mathbf{B}_{2k} = \frac{\hat{\mathbf{x}}_{1k}}{V_1} - \mathbf{B}_{1k} \qquad \mathbf{B}_{1,k+1} = \frac{\hat{\mathbf{x}}_{2k}}{V_2} - \mathbf{B}_{2k}, \tag{19b}$$

where the coefficients A_1 and A_2 verify:

$$V = \mathcal{S}_{\mathbf{F}^T\mathbf{F}}(-A_2), \; A_1 + A_2 = \frac{1}{V} \tag{20}$$

with $\mathcal{S}_{\mathbf{C}}$ the Stieltjes transform with respect to the spectral measure of $\mathbf{C}$.
(19) can then be rewritten as a single iteration on $\mathbf{B_2}$:

$$\mathbf{B}_{2,k+1} = \mathcal{O}_1 \circ \mathcal{O}_2(\mathbf{B}_{2k}) \tag{21}$$

$$\text{where} \quad \mathcal{O}_1 = \frac{1}{V}\text{Prox}_{\frac{1}{A_1}f}(\frac{\cdot}{A_1}) - \text{Id}, \quad \text{and} \quad \mathcal{O}_2 = \left(\frac{1}{\text{V}}\text{Prox}_{\frac{1}{2A_2}||\mathbf{y}-\mathbf{Fx}||_2^2}(\frac{\cdot}{\text{A}_2}) - \text{Id}\right). \tag{22}$$

which becomes the Peaceman-Rachford operator Peaceman and Rachford (1955) if $A_1 = A_2 = \frac{1}{2V}$ is artificially prescribed (note that such a manipulation renders the state evolution equations invalid).

We point out that, for an ℓ_2-penalty, oracle-VAMP iterations reduce to a one step process which is identical to a ridge regression with parameter A_2 (see appendix C.2). To generalize this result to any convex regularization, we derive upper bounds for the Lipschitz constants of the iteration.

Lipschitz constants of oracle-VAMP iteration — To characterize oracle-VAMP iterations, we will make use of convenient proximal operator properties, such as firm nonexpansiveness: for all $\mathbf{x}, \mathbf{y}$ in the input space, the following inequality holds

$$\langle \mathbf{x} - \mathbf{y}, \text{Prox}_f(\mathbf{x}) - \text{Prox}_f(\mathbf{y}) \rangle \geqslant \|\text{Prox}_f(\mathbf{x}) - \text{Prox}_f(\mathbf{y})\|^2. \tag{23}$$

We remind a few useful definitions from convex analysis in the appendix B.1. Moreover, we will use the following result from Giselsson and Boyd (2016):

Proposition 4 *(Proposition 2 from Giselsson and Boyd (2016)) Assume that f is σ-strongly convex and β-smooth and that $\gamma \in]0, \infty[$. Then $\text{Prox}_{\gamma f} - \frac{1}{1+\gamma\beta}\text{Id}$ is $\frac{1}{\frac{1}{1+\gamma\beta} - \frac{1}{1+\gamma\sigma}}$-cocoercive if $\beta > \sigma$ and 0-Lipschitz if $\beta = \sigma$.*

Let (σ_1, β_1) be the strong convexity and smoothness constants of the regularization function. Let (σ_2, β_2) be the corresponding constants of the squared loss $(\mathbf{x} \mapsto \frac{1}{2}\|\mathbf{y} - \mathbf{Fx}\|_2^2)$. We easily find $\sigma_2 = \lambda_{min}(\mathbf{C})$ and $\beta_2 = \lambda_{max}(\mathbf{C})$, the minimal and maximal eigenvalue of $\mathbf{C}$. Using these results and the properties of the fixed point of the state evolution equations, we get the following upper bounds on the Lipschitz constant of the iteration (21), depending on the aspect ratio $\alpha = M/N$ and constants $(\sigma_{1,2}, \beta_{1,2})$.

Lemma 5 *(Lipschitz constants of iteration (21))*

Lipschitz constant of $\mathcal{O}_1$ — *The Lipschitz constant $\mathcal{L}_1$ of the operator $\mathcal{O}_1$ in the cases $0 < \sigma_1 < \beta_1$, $0 < \sigma_1 = \beta_1, 0 = \sigma_1 = \beta_1$ respectively reads:*

$$\mathcal{L}_1 = \max\left(\frac{|A_2 - \sigma_1|}{A_1 + \sigma_1}, \frac{|\beta_1 - A_2|}{A_1 + \beta_1}\right), \ \mathcal{L}_1 = \sqrt{\left(\frac{(A_2^2 - A_1^2)}{(A_1 + \sigma_1)^2} + 1\right)}, \ \mathcal{L}_1 = \max\left(1, \frac{A_2}{A_1}\right). \tag{24}$$

Lipschitz constant of $\mathcal{O}_2$ — *The Lipschitz constant $\mathcal{L}_2$ of the operator $\mathcal{O}_2$ reads*

$$\mathcal{L}_2 = \max\left(\frac{|A_1 - \lambda_{min}(\mathbf{F}^T\mathbf{F})|}{A_2 + \lambda_{min}(\mathbf{F}^T\mathbf{F})}, \frac{|\lambda_{max}(\mathbf{F}^T\mathbf{F}) - A_1|}{A_2 + \lambda_{max}(\mathbf{F}^T\mathbf{F})}\right). \tag{25}$$

Proof. See appendix C.2.

The case $0 < \sigma_1 < \beta_1, \alpha > 1$ yields the same constant as the one derived in Fletcher et al. (2016), which studies a more general version of VAMP, and where the proof method relies on the analysis of the Jacobian of the prescribed iteration. Note that all those constants reduce to 1, i.e. to non-expansive operators if $A_1 = A_2$ is set, which is consistent with the non-expansiveness of the Peaceman-Rachford operator Peaceman and Rachford (1955).

4. Proof of Theorem 2

To prove Theorem 2, we will theoretically solve problem (13) with the oracle-VAMP (19) algorithm. In addition to its SE equations, it enjoys the following useful property: the fixed point of oracle-VAMP solves the minimization problem we are interested in, i.e. problems (1) or (13) depending on the choice of penalty function. This is proven in appendix C.1.

Knowing those two points, it is now tempting – but incorrect – to conclude the proof here by stating that the desired MAP estimator's properties are described by SE equations since it trivially belongs to the trajectory of VAMP initialized at its solution. However this reasoning is flawed : for the SE equations to hold, we need to initialize oracle-VAMP with an isotropically distributed vector. Having no such information on the solution, we need to find at least one converging trajectory starting from an isotropically distriubted vector.

Convergence of oracle-VAMP sequence — The convergence analysis of oracle-VAMP is based on its similarities with the Douglas-Rachford algorithm . We have derived in Lemma 5 upper bounds on the Lipschitz constant of the oracle-VAMP iterations. We now turn to problem (13) and start by proving that, for sufficiently large regularization parameter λ_2, the operator $\mathcal{O}_1 \circ \mathcal{O}_2$ in iteration (21) is a contraction. We first state two useful lemmas that will help in bounding the Lipschitz constants.

Lemma 6 *At the fixed point of the state evolution equations, the coefficients A_1 and A_2 verify:*

$$V = \mathcal{S}_{\mathbf{C}}(-A_2) \qquad V = \mathcal{S}_{\mathcal{H}_f(\hat{\mathbf{x}})}(-A_1) \tag{26}$$

where $\mathcal{H}_f(\hat{\mathbf{x}})$ is the Hessian of the penalty function taken at the fixed point of the algorithm.

Proof. See appendix C.3.

A direct consequence of Lemma 6 is to give upper and lower bounds on A_1 and A_2, which are the constituents of the upper bounds on the Lipschitz constants of iteration (21).

Lemma 7 *At the fixed point of the state evolution equations, we have:*

$$\lambda_{min}(\mathbf{F}^T\mathbf{F}) \leqslant A_1 \leqslant \lambda_{max}(\mathbf{F}^T\mathbf{F}), \qquad \sigma_1 \leqslant A_2 \leqslant \beta_1. \tag{27}$$

Proof. See appendix C.4.

We now consider problem (13), which involves $h = \tilde{f} + \frac{\lambda_2}{2}\|x\|_2^2$ as regularization function. Let $(\tilde{\sigma}_1, \tilde{\beta}_1)$ be the strong convexity and smoothness constants of $\tilde{f}$. Using the second-order definition of strong convexity and smoothness, it is straightforward to obtain $\sigma_1 = \tilde{\sigma}_1 + \lambda_2, \beta_1 = \tilde{\beta}_1 + \lambda_2$. By

arbitrarily increasing λ_2, we can thus accordingly augment σ_1, β_1. Using Lemma 7, we see that A_2 grows with λ_2, while A_1 remains bounded.

We can easily see that the Lipschitz constant of operator $\mathcal{O}_1$ given in (24) is bounded by a constant C. We focus on the Lipschitz constant (25) of $\mathcal{O}_2$. Using inequalities (27), we get

$$\mathcal{L}_2 \leqslant \frac{\lambda_{max}(\mathbf{F}^T\mathbf{F}) - \lambda_{min}(\mathbf{F}^T\mathbf{F})}{\tilde{\sigma}_1 + \lambda_2 + \lambda_{min}(\mathbf{F}^T\mathbf{F})}. \tag{28}$$

We choose λ_2 large enough, for instance $\lambda_2 > C(\lambda_{max}(\mathbf{F}^T\mathbf{F}) - \lambda_{min}(\mathbf{F}^T\mathbf{F})) - \tilde{\sigma}_1 - \lambda_{min}(\mathbf{F}^T\mathbf{F})$, which induces

$$\mathcal{L}_2 \leqslant \frac{\lambda_{max}(\mathbf{F}^T\mathbf{F}) - \lambda_{min}(\mathbf{F}^T\mathbf{F})}{\tilde{\sigma}_1 + \lambda_2 + \lambda_{min}(\mathbf{F}^T\mathbf{F})} < 1. \tag{29}$$

Therefore the Lipschitz constant $\mathcal{L}$ of $\mathcal{O}_1 \circ \mathcal{O}_2$ verifies

$$\mathcal{L} \leqslant \mathcal{L}_1\mathcal{L}_2 < 1. \tag{30}$$

For problem (13) with sufficiently large λ_2, iteration (21) becomes a contraction and we force the convergence. We provide plots in appendix D to illustrate this claim on the elastic net problem. Any oracle-VAMP iterates sequence will thus be convergent in this setting. This sequence can be properly initialized and will be described by state evolution equations. According to Lemma 3, the estimator returned by oracle-VAMP does solve our modified minimization problem. Moreover, the MSE and variance of this estimator obey the state evolution fixed point equations: this concludes the proof for Theorem 2, for high enough λ_2. Theorem 2 is equivalent to Theorem 1, as shown in C.5.

Analytic continuation — We now need to continuate the result for any λ_2, which is only possible because of the convexity of the problem and its analyticity. We first invoke the optimality condition on the convex problem (13) which gives the following prescription for the solution:

$$\left(\mathbf{F}^T\mathbf{F} + \lambda_2 \mathrm{Id} + \nabla \mathrm{f}\right) \mathbf{x} = \mathbf{F}^T\mathbf{y}. \tag{31}$$

Using the analytic inverse function theorem Krantz and Parks (2002), this clearly prescribes an analytic solution for $\mathbf{x}$ in λ_2. We then turn to the SE equations (7), which can also be written as (4) using the replica formalism, as highlighted in C.5. Appendix **??** helps isolate the additional ridge contribution, and shows that equations (4) are analytic in λ_2. The implicit function theorem Krantz and Parks (2012) ensures that the scalar quantities defined by the equations, including the mean squared error, are analytic in λ_2. We can conclude using the analytic continuation property Krantz and Parks (2002) that the replica formula and all the SE quantities hold true whatever the value of λ_2. In particular, taking $\lambda_2 = 0$ provides the MSE of the modified problem which only differs of the original problem (1) by the use of a twice differentiable penalty function $\tilde{f}$. Going from the differentiable relaxation to the real problem only relies on finding an appropriate sequence of twice differentiable functions $(f_n)_{n\in\mathbb{N}}$ converging towards f and taking the limit $n \to \infty$ inside well-defined scalar quantities. It will not be detailed here as it remains intuitive: non-differentiable f in (1) can be naturally approximated by appropriate sequences of differentiable functions, see for example the remark in El Karoui et al. (2013). Figure 2 indeed shows that the prediction holds for a plain LASSO, with no additional ridge or differentiable approximation. Although all scalar quantities can be continuated, we have no theorem for the continuation of the Gaussian property. We conjecture it to be true for all λ_2 in the second part of Theorem 1.

Acknowledgments

We thank Marc Lelarge, Antoine Maillard, Sundeep Rangan and Lenka Zdeborová for insightful discussions. This work benefited from state aid managed by the Agence Nationale de la Recherche under the "Investissements d'avenir" program with the reference ANR-19-P3IA-0001, and from the ANR PAIL. We also acknowledge support from the chaire CFM-ENS "science des données.

References

Madhu Advani, Subhaneil Lahiri, and Surya Ganguli. Statistical mechanics of complex neural systems and high dimensional data. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(03):P03014, 2013.

Madhu S Advani and Andrew M Saxe. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*, 2017.

Daniel Azagra, Juan Ferrera, Fernando López-Mesas, and Yenny Rangel. Smooth approximation of lipschitz functions on riemannian manifolds. *Journal of Mathematical Analysis and Applications*, 326(2):1370–1378, 2007.

Jean Barbier and Nicolas Macris. The adaptive interpolation method: a simple scheme to prove replica formulas in bayesian inference. *Probability Theory and Related Fields*, 174(3-4):1133–1185, 2019.

Jean Barbier, Mohamad Dia, Nicolas Macris, and Florent Krzakala. The mutual information in random linear estimation. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 625–632. IEEE, 2016.

Jean Barbier, Nicolas Macris, Antoine Maillard, and Florent Krzakala. The mutual information in random linear estimation beyond iid matrices. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1390–1394. IEEE, 2018.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

Heinz H Bauschke, Patrick L Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011a.

Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2011b.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

E Bolthausen. On the high-temperature phase of the sherrington-kirkpatrick model. In *Seminar at EURANDOM, Eindhoven*, 2009.

Burak Cakmak, Ole Winther, and Bernard H Fleury. S-amp: Approximate message passing for general matrix ensembles. In *2014 IEEE Information Theory Workshop (ITW 2014)*, pages 192–196. IEEE, 2014.

Benoît Collins. Moments and cumulants of polynomial random variables on unitarygroups, the itzykson-zuber integral, and free probability. *International Mathematics Research Notices*, 2003 (17):953–982, 2003.

Patrick L Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, Lenka Zdeborová, et al. Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In *Advances in Neural Information Processing Systems*, pages 424–432, 2016.

Rainer Dietrich, Manfred Opper, and Haim Sompolinsky. Statistical mechanics of support vector networks. *Physical review letters*, 82(14):2975, 1999.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110 (36):14557–14562, 2013.

Alyson Fletcher, Mojtaba Sahraee-Ardakan, Sundeep Rangan, and Philip Schniter. Expectation consistent approximate inference: Generalizations and convergence. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 190–194. IEEE, 2016.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

Elizabeth Gardner. The space of interactions in neural network models. *Journal of physics A: Mathematical and general*, 21(1):257, 1988.

Pontus Giselsson and Stephen Boyd. Linear convergence and metric selection for douglas-rachford splitting and admm. *IEEE Transactions on Automatic Control*, 62(2):532–544, 2016.

Franklin A Graybill. *Theory and application of the linear model*, volume 183. Duxbury press North Scituate, MA, 1976.

Francesco Guerra. Broken replica symmetry bounds in the mean field spin glass model. *Communications in mathematical physics*, 233(1):1–12, 2003.

Alice Guionnet. Asymptotics of harish-chandra-itzykson-zuber integrals and of schur polynomials. In *Large Random Matrices: Lectures on Macroscopic Asymptotics*, pages 211–216. Springer, 2009.

Dongning Guo and Sergio Verdú. Randomly spread cdma: Asymptotics via statistical physics. *IEEE Transactions on Information Theory*, 51(6):1983–2010, 2005.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Yoshiyuki Kabashima and Mikko Vehkaperä. Signal recovery using expectation consistent approximation for linear observations. In *2014 IEEE International Symposium on Information Theory*, pages 226–230. IEEE, 2014.

Yoshiyuki Kabashima, Mikko Vehkaperä, and Saikat Chatterjee. Typical l1-recovery limit of sparse vectors represented by concatenations of random orthogonal matrices. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(12):P12003, 2012.

Seung-Jean Kim, Kwangmoo Koh, Michael Lustig, Stephen Boyd, and Dimitry Gorinevsky. An interior-point method for large-scale ℓ_1-regularized least squares. *IEEE journal of selected topics in signal processing*, 1(4):606–617, 2007.

Steven G Krantz and Harold R Parks. *A primer of real analytic functions*. Springer Science & Business Media, 2002.

Steven G Krantz and Harold R Parks. *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media, 2012.

Florent Krzakala, Marc Mézard, Francois Sausset, Yifan Sun, and Lenka Zdeborová. Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08009, 2012.

Junjie Ma and Li Ping. Orthogonal amp. *IEEE Access*, 5:2020–2033, 2017.

Andre Manoel, Florent Krzakala, Gaël Varoquaux, Bertrand Thirion, and Lenka Zdeborová. Approximate message-passing for convex optimization with non-separable penalties. *arXiv preprint arXiv:1809.06304*, 2018.

Donald W Marquardt and Ronald D Snee. Ridge regression in practice. *The American Statistician*, 29(1):3–20, 1975.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

Thomas Peter Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

Partha P Mitra. Understanding overfitting peaks in generalization error: Analytical risk curves for l_2 and l_1 penalized interpolation. *arXiv preprint arXiv:1906.03667*, 2019.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Andrea Montanari, YC Eldar, and G Kutyniok. Graphical models concepts in compressed sensing. *Compressed Sensing: Theory and Applications*, pages 394–438, 2012.

Marc Mézard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

Manfred Opper and Wolfgang Kinzel. Statistical mechanics of generalization. In *Models of neural networks III*, pages 151–209. Springer, 1996.

Manfred Opper and Ole Winther. Expectation consistent approximate inference. *Journal of Machine Learning Research*, 6(Dec):2177–2204, 2005.

Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

Giorgio Parisi and Marc Potters. Mean-field equations for spin models with orthogonal interaction matrices. *Journal of Physics A: Mathematical and General*, 28(18):5267, 1995.

Donald W Peaceman and Henry H Rachford, Jr. The numerical solution of parabolic and elliptic differential equations. *Journal of the Society for industrial and Applied Mathematics*, 3(1):28–41, 1955.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2168–2172. IEEE, 2011.

Sundeep Rangan, Vivek Goyal, and Alyson K Fletcher. Asymptotic analysis of map estimation via the replica method and compressed sensing. In *Advances in Neural Information Processing Systems*, pages 1545–1553, 2009.

Sundeep Rangan, Philip Schniter, and Alyson K Fletcher. Vector approximate message passing. *IEEE Transactions on Information Theory*, 2019.

Galen Reeves and Henry D Pfister. The replica-symmetric prediction for compressed sensing with gaussian matrices is exact. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 665–669. IEEE, 2016.

Hyunjune Sebastian Seung, Haim Sompolinsky, and Naftali Tishby. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.

Pragya Sur, Yuxin Chen, and Emmanuel J Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields*, 175(1-2):487–558, 2019.

Michel Talagrand. *Spin glasses: a challenge for mathematicians: cavity and mean field models*, volume 46. Springer Science & Business Media, 2003.

T Tanaka. Asymptotics of harish-chandra-itzykson-zuber integrals and free probability theory. *Journal of Physics: Conference Series*, 95:012002, jan 2008. doi: 10.1088/1742-6596/95/1/012002.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Yaakov Tsaig and David L Donoho. Extensions of compressed sensing. *Signal processing*, 86(3): 549–571, 2006.

Antonia M Tulino, Sergio Verdú, et al. Random matrix theory and wireless communications. *Foundations and Trends® in Communications and Information Theory*, 1(1):1–182, 2004.

Mikko Vehkaperä, Yoshiyuki Kabashima, and Saikat Chatterjee. Analysis of regularized ls reconstruction and random matrix ensembles in compressed sensing. *IEEE Transactions on Information Theory*, 62(4):2100–2124, 2016.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Timothy LH Watkin, Albrecht Rau, and Michael Biehl. The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65(2):499, 1993.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

Appendix A. Heuristic replica derivation from statistical physics

For completeness, we give here the heuristic derivation from the replica method. To characterize the MAP estimator properties, it is useful to compute the posterior distribution normalization factor $\mathcal{Z}(\mathbf{y}, \mathbf{F})$. In a physics perspective, $\mathcal{Z}$ is the partition function of the problem, and we can define the corresponding free energy averaged on the size of the signal $\Phi = \frac{1}{N}\log \mathcal{Z}$. This quantity is known to be self-averaging: when N goes to infinity, Φ concentrates on its average with respect to the distribution of data matrix elements. Hence we will focus on the averaged free energy $\frac{1}{N}\mathbb{E}_{\mathbf{F},\mathbf{x}_0}(\log \mathcal{Z})$. However, directly computing the average of the logarithm of $\mathcal{Z}$ is analytically intractable. We will replace this computation by an easier integral following the so-called replica trick:

$$\lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{\mathbf{F},\mathbf{x}_0,\mathbf{w}}(\log \mathcal{Z}) = \lim_{N\to\infty} \frac{1}{N} \lim_{n\to 0} \frac{\mathbb{E}_{\mathbf{F},\mathbf{x}_0,\mathbf{w}}(\mathcal{Z}^n) - 1}{n}. \tag{32}$$

The partition function reads

$$\mathcal{Z}(\mathbf{y}, \mathbf{F}, \mathbf{w}) = \int \prod_{i=1}^{N} dx_i \prod_{i=1}^{N} p(x_i) \prod_{\mu=1}^{M} \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{1}{2\Delta}\left(y_\mu - \sum_{i=1}^{N} F_{\mu i} x_i\right)^2}. \tag{33}$$

Introducing n replicas of the system, we now consider

$$\mathbb{E}_{\mathbf{F},\mathbf{x}_0,\mathbf{w}}(Z^n) = \int \prod_{i,a} dx_i^a \prod_{i,a} p(x_i^a) \prod_{\mu} \mathbb{E}_{\mathbf{F},\mathbf{x}_0,\mathbf{w}} \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{1}{2\Delta}\sum_{a=1}^{n}\left(\sum_{i=1}^{N} F_{\mu i} x_{0,i} + w_\mu - \sum_{i=1}^{N} F_{\mu i} x_i^a\right)^2}. \tag{34}$$

Following the statistical physics tradition, the computation goes on by introducing the following order parameters:

$$m^a = \frac{1}{N}\sum_{i=1}^{N} x_i^a x_{0,i} \quad a = 1, 2, ...n \tag{35}$$

$$Q^a = \frac{1}{N}\sum_{i=1}^{N} (x_i^a)^2 \quad a = 1, 2, ...n \tag{36}$$

$$q^{ab} = \frac{1}{N}\sum_{i=1}^{N} x_i^a x_i^b \quad a, b = 1, 2, ...n. \tag{37}$$

The parameters m and q respectively quantify the overlap between the teacher and the student weights, and the overlap between the replicas and the student weights. Q is a norm-like parameter for the replicas of the student weights. Those order parameters carry physical meaning and will eventually provide information about the MAP estimator. A key step of the computation is to carry out the average on the matrix elements, which depends on the statistics assumed for the data matrix. In the case of i.i.d. elements, this can be done using the central limit theorem, see Krzakala et al. (2012). An extension from the i.i.d. case has been proposed both in the context of statistical physics Parisi and Potters (1995) and signal processing Kabashima and Vehkaperä (2014). In that case, $\mathbf{F}$ is considered rotationally invariant and $\mathbf{C} = \mathbf{F}^T\mathbf{F}$ has an arbitrary and well-defined singular value distribution $\mu(\lambda)$ with compact support. We can define its minimum λ_{min} and its maximum λ_{max}. Let us recall useful transform definitions. The Stieltjes transform associated to $\mu(\lambda)$ is

$$\mathcal{S}_{\mathbf{C}}(z) = \int_{\lambda_{min}}^{\lambda_{max}} \frac{d\lambda\mu(\lambda)}{\lambda - z} = \mathbb{E}\left[\frac{1}{\lambda - z}\right] \tag{38}$$

and is correctly defined outside of μ's support. The corresponding R-transform is

$$\mathcal{R}_{\mathbf{C}}(x) = \mathcal{S}_{\mathbf{C}}^{-1}(-x) - \frac{1}{x}. \tag{39}$$

Throughout this paper, the support of the considered matrices is always comprised in $\mathbb{R}_+$. The Stieltjes transform is hence well-defined on strictly negative values, and the R-transform is defined in a neighborhood of 0, included in $\mathbb{R}_-^*$. We first perform the averages over $\mathbf{x}_0$ and $\mathbf{w}$ in (34), we can then average on $\mathbf{F}$ using the asymptotic form of the Harish-Chandra-Itzykson-Zuber integral Tanaka (2008). Since $\mathbf{F}$ is rotationally invariant, for any function ϕ of $\mathbf{F}$:

$$\mathbb{E}_{\mathbf{F}}\left[\phi(\mathbf{F})\right] = \mathbb{E}_{\mathbf{F}}\left[\mathcal{D}\mathbf{U}\ \mathcal{D}\mathbf{V}\ \phi(\mathbf{U}\mathbf{F}\mathbf{V}^T)\right] \tag{40}$$

where integrating on $\mathcal{D}\mathbf{U}$, $\mathcal{D}\mathbf{V}$ represents averages over the ensemble of orthogonal matrices using the Haar measure. The Harish-Chandra-Itzykson-Zuber integral then allows to write the result as a function that depends only on the singular value asymptotic distribution $\mu(\lambda)$. This step can only be done when the matrix $\mathbf{F}$ is rotationally invariant, which is an assumption of Theorem 1 and Theorem 2.

Finally, we take the so-called replica symmetric ansatz which assumes that order parameters are the same for every replica. We can thus remove the subscripts in (37) and only keep three order parameters and their conjugate parameters. The replicated partition function reads

$$\mathbb{E}_{\mathbf{F},\mathbf{x}_0,\mathbf{w}}(Z^n) = \int dQ\ d\hat{Q}\ dq\ d\hat{q}\ dm\ d\hat{m}\ e^{nN\Phi(Q,q,m,\hat{Q},\hat{q},\hat{m})}. \tag{41}$$

Assuming that we can take the limit $N \to \infty$ before the limit $n \to 0$, the integral will concentrate on its saddle-point

$$\mathbb{E}_{\mathbf{F},\mathbf{x}_0,\mathbf{w}}(Z^n) = e^{nN\Phi^*}. \tag{42}$$

The final form of the average free energy is given in Kabashima and Vehkaperä (2014) and reads

$$\begin{aligned}\Phi(Q,q,m,\hat{Q},\hat{q},\hat{m}) &= \mathcal{G}_{\mathbf{C}}\left(-\frac{Q-q}{\Delta}\right) + \left(-\frac{\mathbb{E}(x_0^2) - 2m + q}{\Delta} + \frac{\Delta_0(Q-q)}{\Delta^2}\right)\mathcal{G}'_{\mathbf{C}}\left(-\frac{Q-q}{\Delta}\right) \\ &+ \frac{Q\hat{Q}}{2} - m\hat{m} + \frac{q\hat{q}}{2} + \int dx_0\ \phi(x_0) \int Dz \log\left\{\int dx\ p(x) e^{-\frac{1}{\Delta}f(x) - \frac{\hat{Q}+\hat{q}}{2}x^2 + \hat{m}xx_0 + z\sqrt{\hat{q}}x}\right\}\end{aligned} \tag{43}$$

where $Dz = dz\frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$ is the Gaussian measure, and $\mathcal{G}_{\mathbf{C}}$ appears when performing the Harish-Chandra-Itzykson-Zuber integral, being defined with respect to $\mu(\lambda)$ as

$$\mathcal{G}_{\mathbf{C}}(x) = \frac{1}{2}\mathrm{Sup}_{\Lambda}\left\{-\int d\lambda\mu(\lambda)\log|\Lambda-\lambda| + \Lambda x\right\} - \frac{1}{2}\log|x| - \frac{1}{2}. \tag{44}$$

Note that in the domain of definition of $\mathcal{R}_{\mathbf{C}}$, we have $\mathcal{G}'_{\mathbf{C}}(x) = \frac{1}{2}\mathcal{R}_{\mathbf{C}}(x)$. In this paper, $\mathcal{R}_{\mathbf{C}}$ is applied to strictly negative values. For simplicity of notation, we will assume that it is well-defined in the considered range and use the R-transform notation. Otherwise, it can simply be replaced by $2\mathcal{G}'_{\mathbf{C}}$, which is always valid.

The desired parameters $Q, q, m, \hat{Q}, \hat{q}, \hat{m}$ are solutions of the saddle-point equations, and describe the MAP estimator properties since we have maximized the integral over all distributions of $\mathbf{x}$. In particular, notice that the mean squared error $\tilde{E}$ and variance $\tilde{V}$ of the MAP estimator read at the saddle-point

$$\tilde{E} = q - 2m + \mathbb{E}(x_0^2) \tag{45}$$

$$\tilde{V} = Q - q. \tag{46}$$

At this point, we differentiate the free energy (43) with respect to its parameters. For instance, differentiating with respect to $\hat{m}$ and using, (46) gives

$$m = \int dx_0 \phi(x_0) x_0 \int Dz \int \frac{dx}{\tilde{Z}} p(x) x \exp\left\{ -\frac{f(x)}{\Delta} + \frac{x x_0}{\Delta} \mathcal{R}_{\mathbf{C}}\left(-\frac{\tilde{V}}{\Delta}\right) - \frac{x^2}{2\Delta} \mathcal{R}_{\mathbf{C}}\left(-\frac{\tilde{V}}{\Delta}\right) + z\sqrt{\hat{q}} x \right\} \tag{47}$$

with $\tilde{Z} = \int dx\ p(x) e^{-\frac{1}{\Delta} f(x) - \frac{\hat{Q}+\hat{q}}{2} x^2 + \hat{m} x x_0 + z\sqrt{\hat{q}} x}$. Combining saddle-point equations obtained by differentiating on m, q, and $Q - q$, and using (45), (46) we also know

$$\hat{q} = \frac{\Delta_0}{2\Delta^2} \mathcal{R}_{\mathbf{C}}\left(-\frac{\tilde{V}}{\Delta}\right) + \frac{1}{2\Delta}\left(\tilde{E} - \frac{\Delta_0}{\Delta}\tilde{V}\right) \mathcal{R}'_{\mathbf{C}}\left(-\frac{\tilde{V}}{\Delta}\right). \tag{48}$$

which can be inserted into (47). To take the limit $\Delta \to 0$, we rescale the variance parameter $\tilde{V}$ into $\tilde{V}/\Delta$, but keep the same name for simplicity. We are now left to do a Laplace approximation in the integral term of (47), to reach

$$\begin{aligned} m = &\lim_{\Delta \to 0} \int dx_0 \phi(x_0) x_0 \int Dz \int \frac{dx}{\tilde{Z}} p(x) x \\ &\exp\left\{ -\frac{1}{\Delta}\left[f(x) + \frac{x^2}{2}\mathcal{R}_{\mathbf{C}}(-\tilde{V}) - \frac{x x_0}{2}\mathcal{R}_{\mathbf{C}}(-\tilde{V}) - zx\sqrt{\Delta_0 \mathcal{R}_{\mathbf{C}}(-\tilde{V}) + (\tilde{E} - \Delta_0 \tilde{V})\mathcal{R}'_{\mathbf{C}}(-\tilde{V})} \right] \right\} \\ = &\mathbb{E}\left[x_0 \arg\min_x \left\{ f(x) + \frac{\mathcal{R}_{\mathbf{C}}(-\tilde{V})}{2}\left(x - \left[x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}\left(-\tilde{V}\right)} \sqrt{\left(\tilde{E} - \Delta_0 \tilde{V}\right)\mathcal{R}'_{\mathbf{C}}\left(-\tilde{V}\right) + \Delta_0 \mathcal{R}_{\mathbf{C}}\left(-\tilde{V}\right)} \right] \right)^2 \right\} \right]. \end{aligned} \tag{49}$$

Equation (49) clearly yields a proximal operator such that

$$m = \mathbb{E}\left[x_0 \mathrm{Prox}_{f/\mathcal{R}_{\mathbf{C}}(-\tilde{V})}\left(x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}\left(-\tilde{V}\right)} \sqrt{\left(\tilde{E} - \Delta_0 \tilde{V}\right)\mathcal{R}'_{\mathbf{C}}\left(-\tilde{V}\right) + \Delta_0 \mathcal{R}_{\mathbf{C}}(-\tilde{V})\left(-\tilde{V}\right)} \right) \right]. \tag{50}$$

Similar computations on the saddle-point equations allow to rewrite them in terms of variables $(\tilde{E}, \tilde{V})$ and safely lead to Theorem 1's formulas:

$$\tilde{V} = \mathbb{E}\left[\frac{1}{\mathcal{R}_{\mathbf{C}}(-\tilde{V})} \mathrm{Prox}'_{f/\mathcal{R}_{\mathbf{C}}(-\tilde{V})}\left(x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}\left(-\tilde{V}\right)} \sqrt{\left(\tilde{E} - \Delta_0 \tilde{V}\right)\mathcal{R}'_{\mathbf{C}}\left(-\tilde{V}\right) + \Delta_0 \mathcal{R}_{\mathbf{C}}\left(-\tilde{V}\right)} \right) \right] \tag{51}$$

$$\tilde{E} = \mathbb{E}\left[\left\{ \mathrm{Prox}_{f/\mathcal{R}_{\mathbf{C}}(-\tilde{V})}\left(x_0 + \frac{z}{\mathcal{R}_{\mathbf{C}}\left(-\tilde{V}\right)} \sqrt{\left(\tilde{E} - \Delta_0 \tilde{V}\right)\mathcal{R}'_{\mathbf{C}}\left(-\tilde{V}\right) + \Delta_0 \mathcal{R}_{\mathbf{C}}\left(-\tilde{V}\right)} \right) - x_0 \right\}^2 \right]. \tag{52}$$

Solving those equations, we obtain the mean squared error $\tilde{E}$ of the MAP estimator. However, the replica formula for this problem has not been rigorously justified yet. We conclude this section by connecting the fixed point of the state evolution equations of VAMP with the fixed point of the replica equations.

Lemma 8 *The fixed point of the state evolution equations (7) yields the same MSE and variance as the replica prediction (4).*

Proof. See appendix C.5.

Appendix B. Properties of the proximal operator

B.1. Derivative of the proximal

The proximal operator can be written, for any parameter $\gamma \in \mathbb{R}^+$:

$$\mathrm{Prox}_{\gamma f}(\mathbf{x}) = (\mathrm{Id} + \gamma \partial \mathrm{f})^{-1}(\mathbf{x}). \tag{53}$$

For a separable and differentiable function f, we have:

$$\mathrm{Prox}_{\gamma f}(x) + \gamma f^{'}(\mathrm{Prox}_{\gamma f}(x)) = x \tag{54}$$

which, if f is twice differentiable, gives:

$$\mathrm{Prox}^{'}_{\gamma f}(x) = \frac{1}{1 + \gamma f''(\mathrm{Prox}_{\gamma f}(x))}. \tag{55}$$

B.2. 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:

$$\mathrm{Prox}_{\gamma(f + \frac{\lambda_2}{2}\|.\|_2^2)}(x) = (Id + \gamma(\partial f + \lambda_2))^{-1}(x) \tag{56}$$

$$= ((1 + \gamma\lambda_2)Id + \gamma f')^{-1}(x) \quad \text{for differentiable f} \tag{57}$$

the function inside the inverse is analytical in λ_2 with non-zero derivative (γ is strictly positive), so its inverse is also analytical in λ_2 according to the analytical inverse function theorem Krantz and Parks (2002).

Appendix C. Main lemmas and Theorem 1

C.1. Proof of Lemma 3

We start by reminding a useful identity on proximal operators:

Proposition 9 *(Resolvent of the sub-differential Bauschke et al. (2011)) The proximal mapping of a convex function f is the resolvent of the sub-differential of f:*

$$\underset{\gamma f}{\mathrm{Prox}} = (\mathrm{Id} + \gamma \partial \mathrm{f})^{-1}. \tag{58}$$

At the fixed point of the state evolution equations, we have $V_1 = V_2 = V$. Solving for the fixed point and replacing the proximal by the resolvent of ∂f in (14):

$$\mathbf{B}_1 = A_1 \hat{\mathbf{x}}_1 + \partial f(\hat{\mathbf{x}}_1) \qquad \mathbf{B}_2 = (\mathbf{F}^T\mathbf{F} + A_2 \mathrm{Id})\hat{\mathbf{x}}_2 - \mathbf{F}^{\mathrm{T}}\mathbf{y}. \tag{59}$$

Replacing in the second line of (19) (either of the two equations):

$$(\mathbf{F}^T\mathbf{F} + A_2\mathrm{Id})\hat{\mathbf{x}}_2 - \mathbf{F}^{\mathrm{T}}\mathrm{y} = \frac{\hat{\mathbf{x}}_1}{\mathrm{V}} - \mathrm{A}_1\hat{\mathbf{x}}_2 - \partial\mathrm{f}(\hat{\mathbf{x}}_1). \tag{60}$$

Knowing that $A_1 + A_2 = \frac{1}{V}$, and that $\hat{\mathbf{x}}_1 = \hat{\mathbf{x}}_2$ at the VAMP fixed point, it reduces to:

$$\mathbf{F}^T(y - \mathbf{F}x) = \partial f(\hat{\mathbf{x}}_1) \tag{61}$$

which is the optimality condition of problem (1). We see at the fixed point, the additional momentum terms and adaptative variances cancel out, giving the same fixed point as conventional proximal descent methods.

C.2. Oracle-VAMP Lispchitz constants: ℓ_2 case and proof of Lemma 5

C.2.1. A SIMPLE CASE: THE ℓ_2-PENALTY

For the ℓ_2-penalty case, oracle-VAMP's iteration (21) becomes a ridge regression with parameter A_2 which guarantees direct convergence of $\mathbf{B_2}$. Indeed, remember that the proximal operator of a ℓ_2-penalty with parameter λ_2 is:

$$\mathrm{Prox}_{\frac{\lambda_2}{2}||\mathbf{x}||_2^2} = 1/(1+\lambda_2). \tag{62}$$

Using (62) in (19) immediately shows that $\mathbf{B}_2^t$ cancels itself, leading to the fixed point:

$$\hat{\mathbf{x}}_1 = \hat{\mathbf{x}}_2 = (\mathbf{F}^T\mathbf{F} + A_2 Id)^{-1}(\mathbf{F}^T\mathbf{y}). \tag{63}$$

C.2.2. GENERAL CASE

We now turn to the general case and seek to establish Lipschitz bounds on operators $\mathcal{O}_1$ and $\mathcal{O}_2$. The approach is similar to that of Giselsson and Boyd (2016) for Peaceman/Douglas-Rachford splitting. We start by reminding a few useful definitions from convex analysis.

Definition 10 *(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*

$$f(x) \geqslant f(y) + \langle \nabla f(y), x - y\rangle + \frac{\sigma}{2}\|x - y\|^2 \tag{64}$$

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

Definition 11 *(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*

$$f(x) \leqslant f(y) + \langle \nabla f(y), x - y\rangle + \frac{\beta}{2}\|x - y\|^2 \tag{65}$$

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:

$$\sigma\mathrm{Id} \preceq \mathcal{H}_{\mathrm{f}} \preceq \beta\mathrm{Id}. \tag{66}$$

Corollary 12 *(Remark 4.24 Bauschke et al. (2011)) A mapping $T : \mathcal{X} \to \mathcal{D}$ (where $\mathcal{D}$ is a given output space) is β-cocoercive if and only if βT is half-averaged. This means that T can be expressed as:*

$$T = \frac{1}{2\beta}(\mathrm{Id} + \mathrm{S}) \tag{67}$$

where S is a nonexpansive operator.

The goal is now to determine the Lipschitz constants of $\mathcal{O}_1$ and $\mathcal{O}_2$ defined in (21).

C.2.3. LIPSCHITZ CONSTANT OF $\mathcal{O}_1$

Case 1: $0 < \sigma_1 < \beta_1$ Proposition 4 gives the following expression:

$$\mathrm{Prox}_{\frac{1}{A_1}f} = \frac{1}{2}\left(\frac{1}{1+\sigma_1/A_1} + \frac{1}{1+\beta_1/A_1}\right)\mathrm{Id} + \frac{1}{2}\left(\frac{1}{1+\sigma_1/\mathrm{A}_1} - \frac{1}{1+\beta_1/\mathrm{A}_1}\right)\mathrm{S}_1 \tag{68}$$

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

$$\mathcal{O}_1 = \left(\frac{1}{2V}\left(\frac{1}{A_1+\sigma_1} + \frac{1}{A_1+\beta_1}\right) - 1\right)\mathrm{Id} + \frac{1}{2\mathrm{V}}\left(\frac{1}{1+\sigma_1/\mathrm{A}_1} - \frac{1}{1+\beta_1/\mathrm{A}_1}\right)\mathrm{S}_1\left(\frac{.}{\mathrm{A}_1}\right) \tag{69}$$

which, knowing that $A_1 + A_2 = \frac{1}{V}$, $\mathcal{O}_1$ has Lipschitz constant:

$$\mathcal{L}_1 = \max\left(\frac{A_2 - \sigma_1}{A_1 + \sigma_1}, \frac{\beta_1 - A_2}{A_1 + \beta_1}\right). \tag{70}$$

Case 2: $0 < \sigma_1 = \beta_1$ In this case, we have from Proposition 4:

$$\|\mathrm{Prox}_{\frac{1}{A_1}f}(x) - \mathrm{Prox}_{\frac{1}{A_1}f}(y)\|_2^2 = \left(\frac{1}{1+\sigma_1/A_1}\right)^2 \|x-y\|_2^2 \tag{71}$$

which, with the firm non-expansiveness of the proximal operator gives:

$$\begin{aligned}
\|\mathcal{O}_1(x) - \mathcal{O}_1(y)\|_2^2 &= \frac{1}{V^2}\|\mathrm{Prox}_{\frac{1}{A_1}f}(x/A_1) - \mathrm{Prox}_{\frac{1}{A_1}f}(y/A_1)\|_2^2 && (72)\\
&\quad - 2\frac{A_1}{V}\left\langle \frac{x}{A_1} - \frac{y}{A_1}, \mathrm{Prox}_{\frac{1}{A_1}f}(x/A_1) - \mathrm{Prox}_{\frac{1}{A_1}f}(y/A_1)\right\rangle + \|x-y\|_2^2 && (73)\\
&\leqslant \left(\frac{1}{V^2} - 2\frac{A_1}{V}\right)\|\mathrm{Prox}_{\frac{1}{A_1}f}(x/A_1) - \mathrm{Prox}_{\frac{1}{A_1}f}(y/A_1)\|_2^2 + \|x-y\|_2^2 && (74)\\
&= \left(\left(\frac{1}{V^2} - 2\frac{A_1}{V}\right)\left(\frac{1}{A_1+\sigma_1}\right)^2 + 1\right)\|x-y\|_2^2 && (75)\\
&= \left(\frac{A_2^2 - A_1^2}{(A_1+\sigma_1)^2} + 1\right)\|x-y\|_2^2. && (76)
\end{aligned}$$

The upper bound on the Lipschitz constant is therefore:

$$\mathcal{L}_1 = \sqrt{\frac{(A_2^2 - A_1^2)}{(A_1+\sigma_1)^2} + 1}. \tag{77}$$

Case 3: no strong convexity or smoothness assumption In this case the only information we have is the firm nonexpansiveness of the proximal operator, which gives the same proof as in the previous case but stops at (74), immediately giving the upper bound:

$$\mathcal{L}_1 = \max\left(1, \frac{A_1}{A_2}\right). \tag{78}$$

Lipschitz constant of $\mathcal{O}_2$ Remember that we make the assumption that the data matrix is non-trivial, i.e. that $\lambda_{max}(\mathbf{F}^T\mathbf{F}) \neq 0$. In this case we use the explicit form of $\mathcal{O}_2$, which is linear:

$$\|\mathcal{O}_2(x) - \mathcal{O}_2(y)\|_2 = \|\left(\frac{1}{V}(\mathbf{F}^T\mathbf{F} + A_2^t\mathrm{Id})^{-1} - \mathbf{I}\right)(x-y)\|_2 \tag{79}$$

$$\leqslant \|\left(\frac{1}{V}(\mathbf{F}^T\mathbf{F} + A_2^t\mathrm{Id})^{-1} - \mathbf{I}\right)\|_2 \, \|x-y\|_2. \tag{80}$$

The spectral norm of the matrix in (79) gives the upper bound on the Lipschitz constant:

$$\mathcal{L}_2 = \max\left(\frac{A_1 - \lambda_{min}(\mathbf{F}^T\mathbf{F})}{A_2 + \lambda_{min}(\mathbf{F}^T\mathbf{F})}, \frac{\lambda_{max}(\mathbf{F}^T\mathbf{F}) - A_1}{A_2 + \lambda_{max}(\mathbf{F}^T\mathbf{F})}\right) \tag{81}$$

C.3. Proof of Lemma 6

The equation defining V_2 in (14) directly gives $V = \mathcal{S}_{\mathbf{C}}(-A_2)$ by the definition of the Stieltjes transform. For a separable and differentiable function, we have the element-wise identity (see appendix B.1):

$$\mathrm{Prox}'_{\gamma f}(x) = \frac{1}{1 + \gamma f''(\mathrm{Prox}_{\gamma f}(x))} \tag{82}$$

which, from the definition of the element wise averaging operator, gives:

$$\langle \mathrm{Prox}'_{\gamma f}(x) \rangle = \frac{1}{N}\mathrm{Trace}\left[(\mathrm{Id} + \gamma \mathcal{H}_{\mathrm{f}}(\hat{\mathbf{x}})\right]. \tag{83}$$

The prescription for V_1 in (14) then directly gives $V = \mathcal{S}_{\mathcal{H}_f(\hat{\mathbf{x}})}(-A_1)$.

C.4. Proof of Lemma 7

From the definition of the Stieltjes transform and the second order definition of strong convexity and smoothness:

$$\frac{1}{\lambda_{max}(\mathbf{F}^T\mathbf{F}) + A_2} \leqslant \frac{1}{V} \leqslant \frac{1}{\lambda_{min}(\mathbf{F}^T\mathbf{F}) + A_2} \qquad \frac{1}{\sigma_1 + A_1} \leqslant \frac{1}{V} \leqslant \frac{1}{\beta_1 + A_1} \tag{84}$$

which combined with $A_1 + A_2 = \frac{1}{V}$ gives the bound.

C.5. Proof of Lemma 8 and Theorem 1

The state evolution equations (7) follow a set of parameters through the VAMP iterations. Among those parameters, we find the mean squared errors $\mathcal{E}_1$ and $\mathcal{E}_2$ of the estimators $\hat{\mathbf{x}}_\mathbf{1}$ and $\hat{\mathbf{x}}_\mathbf{2}$ with respect to the true signal, and their variances V_1, V_2. At the fixed-point of VAMP, the estimators are equal, hence their errors and variances coincide with one mean squared error E and variance V.

Besides, the replica saddle point equations also close on the error $\tilde{E}$ and variance $\tilde{V}$. We would like to show that the replica prediction matches the state evolution fixed point conditions.

Starting from the state evolution fixed point, we notice that

$$V = \frac{\alpha_2}{A_2} = \mathcal{S}_{\mathbf{C}}(-A_2) \tag{85}$$

where S is the Stieltjes transform with respect to the spectral measure defined by $\mathbf{C} = \mathbf{F}^T\mathbf{F}$. Then

$$A_1 = \frac{1}{V} - A_2 = \frac{1}{V} + \mathcal{S}_{\mathbf{C}}^{-1}(-V) = \mathcal{R}_{\mathbf{C}}(-V). \tag{86}$$

Moreover,

$$V = \frac{1}{A_1}\mathbb{E}_{x_0,P_1}\left[\mathrm{Prox}'_{f/A_1}(x_0 + P_1)\right] \tag{87}$$

where P_1 is a Gaussian variable of variance τ_1. Looking at τ_2, we have

$$\tau_2 = \frac{1}{(1 - A_1V)^2}\left[E - \tau_1(A_1V)^2\right] = \frac{1}{(A_2V)^2}\left[E - \tau_1(1 - A_2V)^2\right]. \tag{88}$$

We rewrite the equation on τ_1, which involves an average on the eigenvalue distribution of $\mathbf{C} = \mathbf{F}^T\mathbf{F}$:

$$\tau_1 = \frac{1}{(1 - A_2V)^2}\left(\mathbb{E}\left[\Delta_0\frac{\lambda}{(\lambda + A_2)^2} + \tau_2\frac{A_2^2}{(\lambda + A_2)^2}\right] - \tau_2(A_2V)^2\right) \tag{89}$$

$$= \frac{1}{(1 - A_2V)^2}\left(\mathbb{E}\left[\frac{\Delta_0}{(\lambda + A_2)} - \frac{A_2\Delta_0}{(\lambda + A_2)^2} + \tau_2\frac{A_2^2}{(\lambda + A_2)^2}\right] - \tau_2(A_2V)^2\right) \tag{90}$$

$$\tau_1 = \frac{1}{(1 - A_2V)^2}\left(\Delta_0\mathcal{S}_{\mathbf{C}}(-A_2) - A_2\Delta_0\mathcal{S}'_{\mathbf{C}}(-A_2) + \tau_2A_2^2\mathcal{S}'_{\mathbf{C}}(-A_2) - \tau_2(A_2V)^2\right), \tag{91}$$

then plug in τ_2's expression (88) to reach

$$\begin{aligned}\tau_1 = &\frac{\Delta_0V^2}{(1 - A_2V)^2\mathcal{S}'_{\mathbf{C}}(-A_2)}(\mathcal{S}_{\mathbf{C}}(-A_2) - A_2\mathcal{S}'_{\mathbf{C}}(-A_2))\\ &+ \frac{E}{(1 - A_2V)^2\mathcal{S}'_{\mathbf{C}}(-A_2)}(\mathcal{S}'_{\mathbf{C}}(-A_2) - V^2).\end{aligned} \tag{92}$$

We have expressed the variance of the Gaussian variable P_1 as a function of E and V. We would like to match this with the variance of the Gaussian inside (51), namely

$$\tilde{\tau}_1 = \frac{1}{\mathcal{R}^2_{\mathbf{C}}\left(-\frac{\tilde{V}}{\Delta}\right)}\left(\left(\tilde{E} - \frac{\Delta_0}{\Delta}\tilde{V}\right)\mathcal{R}'_{\mathbf{C}}\left(-\frac{\tilde{V}}{\Delta}\right) + \Delta_0\mathcal{R}_{\mathbf{C}}\left(-\frac{\tilde{V}}{\Delta}\right)\right) \tag{93}$$

$$\begin{aligned}= &\frac{E}{\mathcal{R}^2_{\mathbf{C}}\left(-\frac{\tilde{V}}{\Delta}\right)}\left(\frac{-1}{\mathcal{S}'_{\mathbf{C}}(\mathcal{S}^{-1}_{\mathbf{C}}(\frac{\tilde{V}}{\Delta}))} + \frac{\Delta^2}{\tilde{V}^2}\right)\\ &+ \frac{\Delta_0}{\mathcal{R}_{\mathbf{C}^2}\left(-\frac{\tilde{V}}{\Delta}\right)}\left(\mathcal{R}_{\mathbf{C}}\left(-\frac{\tilde{V}}{\Delta}\right) - \frac{\tilde{V}}{\Delta}\left(\frac{-1}{\mathcal{S}'_{\mathbf{C}}(\mathcal{S}^{-1}_{\mathbf{C}}(\frac{\tilde{V}}{\Delta}))} + \frac{\Delta^2}{\tilde{V}^2}\right)\right).\end{aligned} \tag{94}$$

A few lines of computation show that $\tau_1 = \tilde{\tau}_1$. Therefore, the replica saddle point equation on $\tilde{V}$ (87) becomes exactly the same as the state evolution fixed point equation on V (4a). Similarly, we recall the definition of the fixed point value of E for SE equations

$$E = \mathbb{E}\left[\left(\text{Prox}_{\frac{1}{A_1}f}(x_0 + P_1) - x_0\right)^2\right] \tag{95}$$

where P_1 is also pulled from a Gaussian distribution with variance $\tau_1 = \tilde{\tau}_1$. (95) matches the replica equation on $\tilde{E}$ (4b). Finally the variables (E, V) from SE equations, and $(\tilde{E}, \tilde{V})$ from replica formalism are the same and satisfy the same relations, which proves Lemma 8.
It is hence straightforward to prove Theorem 1 after having shown Theorem 2. As shown above, Theorem 1 is simply a rewriting of state evolution equations from Theorem 2 in their replica fixed point form, i.e. removing some intermediate variables to obtain a more compact form.

Appendix D. State evolution equations for the elastic net problem

We solve the recursion (14) on an elastic net problem:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min}\left\{\frac{1}{2}\|\mathbf{y} - \mathbf{F}\mathbf{x}\|_2^2 + \lambda_1|\mathbf{x}|_1 + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2\right\}. \tag{96}$$

For a given parameter $\gamma \in \mathbb{R}^+$, the proximal operator of the corresponding regularization function reads:

$$\text{Prox}_{\frac{1}{A_{1k}}(\lambda_1|\mathbf{x}|_1 + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2)}(.) = \frac{1}{1 + \frac{\lambda_2}{A_{1k}}}\, s\left(., \frac{\lambda_1}{A_{1k}}\right) \tag{97}$$

where $s\left(., \frac{\lambda_1}{A_{1k}}\right)$ is the soft-thresholding function:

$$s\left(r_{1k}, \frac{\lambda_1}{A_{1k}}\right) = \begin{cases} r_{1k} + \frac{\lambda_1}{A_{1k}} & \text{if} \quad r_{1k} < -\frac{\lambda_1}{A_{1k}} \\ 0 & \text{if} \quad -\frac{\lambda_1}{A_{1k}} < r_{1k} < \frac{\lambda_1}{A_{1k}} \\ r_{1k} - \frac{\lambda_1}{A_{1k}} & \text{if} \quad r_{1k} > \frac{\lambda_1}{A_{1k}}. \end{cases} \tag{98}$$

We consider an i.i.d. teacher vector $\mathbf{x}_0$ pulled from the Gauss-Bernoulli distribution :

$$\phi(x_0) = (1-\rho)\delta(x_0) + \rho\frac{1}{\sqrt{2\pi}}\exp\left(-x_0^2/2\right). \tag{99}$$

Here we give the detail of the set of equations (7) for an elastic net minimization problem. The quantities that must be explicitly computed are the averages $\mathcal{E}_1$ and $\mathcal{E}_2$ and the ones on the derivatives.

$$\alpha_{1k} = \mathbb{E}\left[\frac{1}{1+\frac{\lambda_2}{A_{1k}}} s'\left(x_0 + P_{1k}, \frac{\lambda_1}{A_{1k}}\right)\right] \quad \text{where} \quad x_0 \sim \mathcal{N}(0,1) \quad p_{1k} \sim \mathcal{N}(0,\tau_{1k}) \tag{100}$$

$$= \frac{1}{1+\frac{\lambda_2}{A_{1k}}}(1-\rho)\left(\int_{-\infty}^{-\lambda_1/A_{1k}} dp \frac{1}{\sqrt{2\pi\tau_{1k}}} e^{-\frac{p^2}{2\tau_{1k}}} + \int_{\lambda_1/A_{1k}}^{+\infty} dp \frac{1}{\sqrt{2\pi\tau_{1k}}} e^{-\frac{p^2}{2\tau_{1k}}}\right) \tag{101}$$

$$+ \rho \frac{1}{1+\frac{\lambda_2}{A_1}}\left(\int_{-\infty}^{-\lambda_1/A_{1k}} dp \frac{1}{\sqrt{2\pi(\tau_{1k}+1)}} \exp(-\frac{p^2}{2(\tau_{1k}+1)})\right.$$

$$\left. + \int_{\lambda_1/A_{1k}}^{+\infty} dp \frac{1}{\sqrt{2\pi(\tau_{1k}+1)}} \exp\left(-\frac{p^2}{2(\tau_{1k}+1)}\right)\right)$$

$$= \frac{1}{1+\frac{\lambda_2}{A_{1k}}}\left[(1-\rho)\operatorname{erfc}\left(\frac{\lambda_1}{A_{1k}\sqrt{2\tau_{1k}}}\right) + \rho \operatorname{erfc}\left(\frac{\lambda_1}{A_{1k}\sqrt{2(\tau_{1k}+1)}}\right)\right]. \tag{102}$$

$$\mathcal{E}_1 = \mathbb{E}_{x_0,P_{1k}}\left[\left(\operatorname{Prox}_{\frac{1}{A_{1k}}f}(x_0+P_{1k}) - x_0\right)^2\right] \quad \text{where } x_0 \sim \mathcal{N}(0,1) \quad p_{1k} \sim \mathcal{N}(0,\tau_{1k}) \tag{103}$$

$$= \left(\frac{1}{1+\frac{\lambda_2}{A_{1k}}}\right)^2 (1-\rho)\int_{\mathbb{R}} dp \frac{1}{\sqrt{2\pi\tau_{1k}}} \exp\left(-\frac{p^2}{2\tau_{1k}}\right) s\left(p, \frac{\lambda_1}{A_{1k}}\right)^2 \tag{104}$$

$$+\rho \iint_{\mathbb{R}} dx_0 dp \frac{1}{\sqrt{2\pi\tau_{1k}}} \exp\left(-\frac{p^2}{2\tau_{1k}}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_0^2}{2}\right)\left(\frac{1}{1+\frac{\lambda_2}{A_{1k}}} s\left(x_0+p, \frac{\lambda_1}{A_{1k}}\right) - x_0\right)^2. \tag{105}$$

The first term (104) only involves one dimensional integrals and can easily be evaluated numerically:

$$(104) = (1-\rho)\left(\frac{1}{1+\frac{\lambda_2}{A_{1k}}}\right)^2 \left[\int_{-\infty}^{-\lambda_1/A_{1k}} dp\mathcal{N}(0,\tau_{1k})\left(p+\frac{\lambda_1}{A_{1k}}\right)^2 + \int_{\lambda_1/A_{1k}}^{+\infty} dp\mathcal{N}(0,\tau_{1k})\left(p-\frac{\lambda_1}{A_{1k}}\right)^2\right]$$

$$= (1-\rho)\left(\frac{1}{1+\frac{\lambda_2}{A_{1k}}}\right)^2 \left[\operatorname{erfc}\left(\frac{\lambda_1/A_{1k}}{\sqrt{2\tau_{1k}}}\right)\left(\left(\frac{\lambda_1}{A_{1k}}\right)^2 + \tau_{1k}\right) - e^{-\frac{(\lambda_1/A_{1k})^2}{2\tau_{1k}}}\sqrt{2\tau_{1k}/\pi}\frac{\lambda_1}{A_{1k}}\right]. \tag{106}$$

The second term (105) needs to be decomposed in order to avoid computing the two-dimensional integral numerically:

$$(105) = \rho\mathbb{E}_{x_0}\left[\int_{-\infty}^{-\lambda_1/A_{1k}-x_0} dp\mathcal{N}(0,\tau_{1k})\left(\frac{1}{1+\frac{\lambda_2}{A_{1k}}}\left(x_0+p+\frac{\lambda_1}{A_{1k}}\right) - x_0\right)^2\right.$$

$$\left. + \int_{\lambda_1/A_{1k}-x_0}^{+\infty} dp\mathcal{N}(0,\tau_{1k})\left(\frac{1}{1+\frac{\lambda_2}{A_{1k}}}\left(x_0+p-\frac{\lambda_1}{A_{1k}}\right) - x_0\right)^2 + \int_{-\lambda_1/A_{1k}-x_0}^{\lambda_1/A_{1k}-x_0} dp\mathcal{N}(0,\tau_{1k})x_0^2\right].$$

A little algebra allows to express (105) with error functions supported by most scientific coding libraries. We rewrite the shrinkage factor due to the ℓ_2 penalty $s = \frac{1}{1+\frac{\lambda_2}{A_{1k}}}$:

$$
\begin{aligned}
(105) = \rho \mathbb{E}_{x_0} \Bigg[& \frac{1}{2} x_0^2 \left(\operatorname{erf}\left(\frac{\lambda_1/A_{1k} - x_0}{\sqrt{2\tau_{1k}}} \right) + \operatorname{erf}\left(\frac{\lambda_1/A_{1k} + x_0}{\sqrt{2\tau_{1k}}} \right) \right) && (107) \\
& + x_0^2 - 2s x_0^2 + s^2(\tau_{1k} + (\lambda_1/A_{1k})^2 + x_0^2) && (108) \\
& + s\sqrt{\tau_{1k}/(2\pi)} \Bigg(\exp\left(-\frac{\lambda_1/A_{1k} - x_0^2}{2\tau_{1k}} \left((s-2)x_0 - s\frac{\lambda_1}{A_{1k}} \right) \right) && (109) \\
& + \exp\left(-\frac{\lambda_1/A_{1k} + x_0^2}{2\tau_{1k}} \left((2-s)x_0 - s\frac{\lambda_1}{A_{1k}} \right) \right) \Bigg) && (110) \\
& + \frac{1}{2} \Bigg((s^2(\tau_{1k} + (\lambda_1/A_{1k} - x_0)^2) + 2s(\lambda_1/A_{1k} - x_0)x_0 + x_0^2) \operatorname{erf}\left(\frac{\lambda_1/A_{1k} - x_0}{\sqrt{2\tau_{1k}}} \right) && (111) \\
& - (x_0^2 - 2s x_0(\lambda_1/A_{1k} + x_0) + s^2(\tau_{1k} + (\lambda_1/A_{1k} + x_0)^2)) \operatorname{erf}\left(\frac{\lambda_1/A_{1k} + x_0}{\sqrt{2\tau_{1k}}} \right) \Bigg) \Bigg]. && (112)
\end{aligned}
$$

We then invoke the appropriate expressions for α_{2k} and $\mathcal{E}_{2k}$ from Rangan et al. (2019):

$$
\alpha_{2k} = \mathbb{E}\left[\frac{A_{2k}}{\lambda_{\mathbf{F}^T\mathbf{F}} + A_{2k}} \right] \tag{113}
$$

$$
\mathcal{E}_2 = \mathbb{E}\left[\frac{\Delta_0 \lambda_{\mathbf{F}^T\mathbf{F}} + \tau_{2k} A_{2k}^2}{(\lambda_{\mathbf{F}^T\mathbf{F}} + A_{2k})^2} \right]. \tag{114}
$$

These forms can be used in the recursion (7) with the chosen values of λ_1 and λ_2 to find the right thresholding coefficients, errors and variances.

We used these forms with $\lambda_2 = 0$ for the LASSO simulations in the experiments in section 2.2. The elastic net also allows us to illustrate the convergence for large enough λ_2, as shown in the following figure. We launch oracle-VAMP on the elastic net problem for five values of the aspect ratio $\alpha = 0.1, 0.2, 0.5, 1, 2$. The problem setup is the same as in 2.2, with $\rho = 0.3$, $\Delta_0 = 0.01$. The choice of sensing matrix matters little here, as long as the eigenvalue spectrum has compact support. We used i.i.d. Gaussian matrices for simplicity. We plot the average squared distance between two successive iterates of (19) in a logarithmic scale on an elastic net problem with a LASSO parameter of $\lambda_1 = 0.1$ and varying ridge parameter $\lambda_2 = 0.1, 0.2, 0.3$. For low values of α, the data matrix is highly ill-conditioned and the algorithm diverges as shown on the first plot for $\alpha = 0.1, 0.2$. We then augment the ridge parameter on the second figure, which makes the $\alpha = 0.2$ curve converge. Pushing λ_2 further on the third plot makes the $\alpha = 0.1$ curve converge. We thus see that augmenting the ridge indeed enforces convergence.

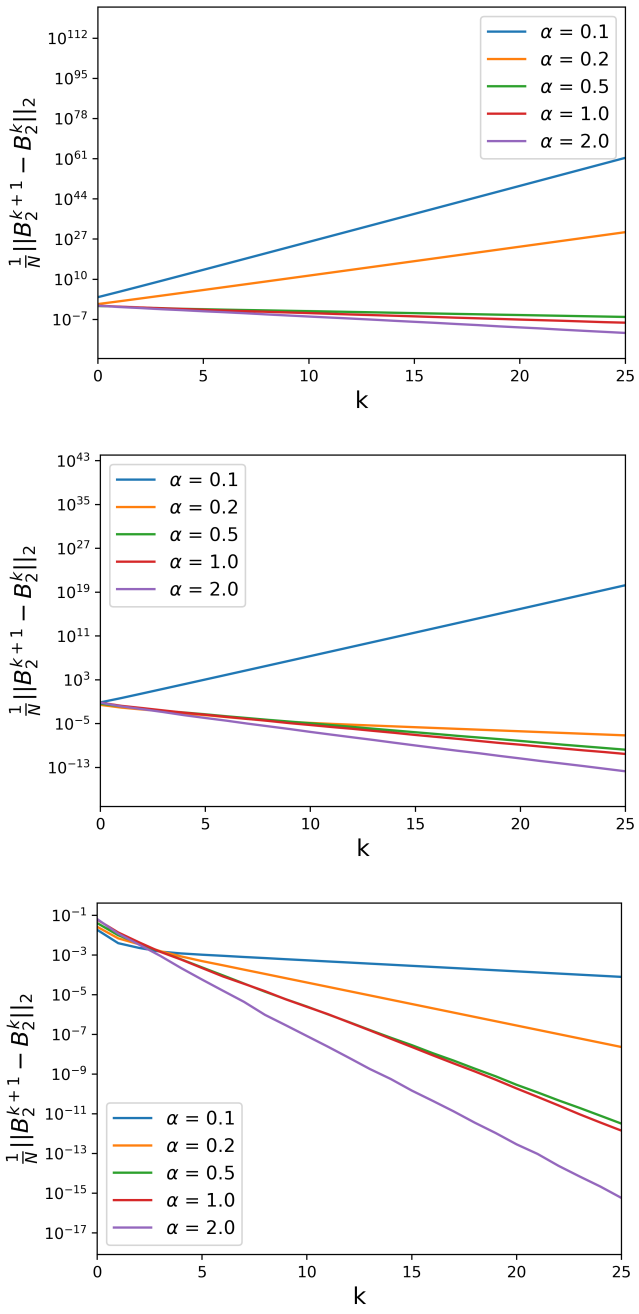

Figure 2: Logarithmic scale mean squared distance between successive iterates of oracle-VAMP on an elastic net problem with $\lambda_1 = 0.1$ and, from left to right $\lambda_2 = 0.1, 0.2, 0.3$. Each plot contains five trajectories with aspect ratio $\alpha = 0.1, 0.2, 0.5, 1.0, 2.0$. For low aspect ratios, the sensing matrix is highly ill-conditioned and diverging trajectories are observed (blue $\alpha = 0.1$ and orange $\alpha = 0.2$ curves, on the first plot). Augmenting the ridge parameter enforces the convergence: at $\lambda_2 = 0.2$ (second plot), the orange curve describes a converging trajectory, and at $\lambda_2 = 0.3$ (third plot), the blue curve converges as well. Plots are obtained with $N = 100$, $\rho = 0.3$ and $\Delta_0 = 0.01$.

Appendix E. Convergence of vector sequences

This is essentially a rewriting of appendix B of Rangan et al. (2019), which reviews the analysis framework from Bayati and Montanari (2011a).
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

$$\mathbf{x}(N) = (\mathbf{x}_1(N), ..., \mathbf{x}_N(N)) \tag{115}$$

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 $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^s$:

$$\|\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_2)\| \leqslant C\|\mathbf{x}_1 - \mathbf{x}_2\| \left[1 + \|\mathbf{x}_1\|^{p-1} + \|\mathbf{x}_1\|^{p-1}\right] \tag{116}$$

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

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbf{f}(x_n(N)) = \mathbb{E}[f(X)] \text{ a.s.} \tag{117}$$

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 Bayati and Montanari (2011a). 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. 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 *uniform 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:

$$\|\phi(\mathbf{x}_1, A) - \phi(\mathbf{x}_2, A)\| \leqslant L_1\|\mathbf{x}_1 - \mathbf{x}_2\| \tag{118}$$

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

$$\|\phi(\mathbf{x}, A_1) - \phi(\mathbf{x}, A_2)\| \leqslant L_2(1 + \|\mathbf{x}\|)|A_1 - A_2| \tag{119}$$

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

The additional conditions for the SE theorem (see Theorem 1 *(i-ii-iii)* from Rangan et al. (2019)) to hold are the following:
- α_{1k} must be in $[0, 1]$. This is always verified using B.1, knowing that $f'' \geqslant 0$ by convexity.
- The functions defining A_i and $\mathcal{E}_i$ must be continuous at the points prescribed by the SE equations. This holds true as well since proximals of convex functions are continuous.

- Finally the denoisers (here the proximals) and their derivatives need to be uniformly Lipschitz in their arguments at their parameters. This can be easily checked on a case by case basis, and holds true for most commonly used regularizations. Potentially pathological cases like logarithmic barriers can be handled by working on intervals where the function and its needed derivatives are bounded.