

1 We thank the referees for their detailed work and their thoughtful comments that we gratefully use to improve our paper.

2 **Answer to referee 1:** We will add and adjust the two mentioned references, thanks for letting us know.

3 1. We do not know whether there is a general criteria that would distinguish when the decay is  $1/\alpha$  or  $1/\sqrt{\alpha}$ . Providing  
4 such a generic criteria is definitely a line of research we would like investigate in the future.

5 2. Our analysis in this paper focuses on i.i.d Gaussian data and L2 regularization, as this case was very extensively  
6 studied in past works, in particular in statistical physics. As mentioned, the CGMT analysis used in this work can  
7 be generalized to non-isotropic Gaussian and any convex loss and separable regularization. We will add a related  
8 discussion in the final version to clarify.

9 3. We thank the referee for pointing out this very interesting work. First, note that the main focus of our work is on  
10 the test error including the constants for any  $\alpha$ , not only on the rate at very large  $\alpha$ . Secondly, we focus here on the  
11 generalization for a non Lipschitz function (namely the sign/misclassification loss) so that the suggested reference that  
12 focus on smooth loss functions does not readily apply. We shall, however, include the discussion of these bounds in the  
13 section where we compare to the Rademacher bounds, and investigate this interesting connection.

14 4. This is correct. The main reason why we use CGMT and not AMP is that we do not know how to prove that the state  
15 evolution of the AMP corresponds to the solution of the ERM. Only after having the CGMT proof in hand, it follows  
16 that the SE of AMP gives the same equations than the CGMT. We will add a comment to the paper.

17 **Answer to referee 2:** We shall indeed comment on the absence of the computational gap in the final version.

18 The set of fixed point equations is fully rigorous. The analytic solution of this set of equations is provided only in the  
19 ridge case, and also for vanishing  $\lambda \rightarrow 0$  that allows to obtain analytically the asymptotic generalization behaviour  
20 of the max-margin estimator. Unfortunately, in the others situations, we did not find a closed form expression of the  
21 fixed point, so that the generalization behaviour is evaluated numerically. We, however, note that these are fixed point  
22 equations on scalar variables so their numerical resolution posed no problem. The non-trivial part of the rigorous  
23 analysis is the reduction of the high-dimensional problem to the scalar fixed point equations.

24 We will clarify and distinguish the points of our work that are rigorously proven with this work, and the ones for which  
25 the proof can be extended. As correctly pointed out, in order to not overload the paper and as most of the numerics has  
26 been performed in this case, the set of fixed point equations (10) has been made rigorous only for L2 regularization  
27 for binary classification, even though the entropy for regression is proven as well in (SM III.1.1). We stated that our  
28 results are *valid* for a wide range of regularizers as the replica's prediction of the fixed point equations are provided for  
29 generic convex and separable loss and regularizer. They are believed to hold true and the generic Gordon's mini-max  
30 framework can be easily generalized to this case.

31 We only reproduced the results of [49] to bring to light interesting conclusions relating our work of the ERM estimations  
32 and the discussion of [49]. We will clarify this in the final version.

33 1. As briefly explained in the proof below Theorem 4.1, the GAMP algorithm, recalled in (S.M VI.1), is valid for ERM  
34 estimation with the corresponding updates  $f_{\text{out}}^{\text{erm}}(l, r)$ ,  $f_{\text{w}}^{\text{erm}}(l, r)$  in (S.M VI.2). Thm 4.1 relies basically therefore on  
35 the fact that the optimal loss  $l^{\text{opt}}$  and regularizer  $r^{\text{opt}}$  are designed in (S.M VI) such that at each time step the ERM  
36 denoisers match the Bayes-optimal ones:  $f_{\text{out}}^{\text{erm}}(l^{\text{opt}}, r^{\text{opt}}) = f_{\text{out}}^{\text{bayes}}$  and  $f_{\text{w}}^{\text{erm}}(l^{\text{opt}}, r^{\text{opt}}) = f_{\text{w}}^{\text{bayes}}$ . As these denoising  
37 steps are the only difference between ERM-AMP and Bayes-AMP, and as according to [53] AMP algorithm with  
38 Bayes-updates reaches Bayes-optimal performances, we obtain Theorem 4.1 and will clarify the discussion.

39 2. Indeed the Gordon analysis apply to  $\Phi$  (SM III.9) and  $\tilde{\Phi}$  (SM III.10). The free entropy of the problem is therefore  
40 given by  $\tilde{\Phi}$  that reads as an optimization problem over  $\mu, \delta, \tau$  in (SM III.4). As  $\lambda > 0$ , it can be shown that this problem  
41 has a unique solution  $(\mu^*, \delta^*)$  so that  $(\mu, \delta) \rightarrow (\mu^*, \delta^*)$ . As suggested, we will add a clarification of this technical step.

42 **Answer to referee 3:** Our motivation to focus on the sign-teacher with Gaussian data and ground truth vector is that it  
43 has been extensively studied in the past. Our analysis is valid more generically, and we will discuss this in the revised  
44 manuscript more clearly, see answer to referee 2. We will correct the reference to the Moreau-Yosida regularization.

45 **Answer to referee 5:** We will correct the typos, thank you very much. Concerning consequences for practitioners, we  
46 think that our result about approaching very closely the Bayes-optimal performance with simple regularized ERM is an  
47 interesting message for practitioners. Of course, shoving this in more realistic settings would be desired.

48 As correctly pointed out, the construction of the optimal loss and regularizer requires the knowledge of the teacher  
49 distribution related to  $\mathcal{Z}_{\text{out}}^*$ ,  $\mathcal{Z}_{\text{w}}^*$  in (17). Indeed, in the Bayes-optimal setting, we may directly use the Bayes-optimal  
50 AMP algorithm to achieve optimal performances as proven in [53]. Nevertheless, it seems to us interesting to point out  
51 that Bayes performances that require, in principle, to compute an intractable high-dimensional posterior sampling can  
52 be obtained instead by the easier, more common and practical ERM estimation.