# Response to Referees:<br>Simplified derivations for high-dimensional convex learning problems

David G. Clark, Haim Sompolinsky

**Color coding:** Referee comments are in red and revised text/changes are in blue.

We thank the referees for their constructive comments. We have implemented all of their suggestions. Below we describe our responses to each referee's comments.

## Response to Referee 1

"More explanations on the self-averaging of the self-responses (e.g. below (33)) could prove helpful."

We have added clarification after Eq. 36:

By self-averaging, we mean that the fluctuations of $F_{00}$ (defined by Eq. 33) around its $\mathcal{O}(1)$ expectation (given by Eq. 36) are $\mathcal{O}(1/\sqrt{N})$, as can be verified.

"To the best of my reading, the expression (58) for the number of supporting points is not established before (58), and could gain to be briefly discussed."

We now provide better explanation before Eq. 58:

Consulting Eq. 51, $\phi_0(\kappa)$ is the probability for a Lagrange multiplier to be nonzero. Thus, the weight response $S^w$ (Eq. 57) has a nice interpretation...

"Due to the bipartite structure, perturbations to other datum variables do not affect the [cavity variable]": is this statement true to leading order or in general? If the former, it would be clearer to make the precision."

We have clarified this point by expanding the explanation after Eq. 28:

One could include higher-order terms, the next of which would be smaller than the above term by a factor of $1/\sqrt{N}$ and given by $\sum_{\nu,\rho=1}^{P} \frac{d\lambda^\mu}{dI^\nu dI^\rho} x_0^\nu x_0^\rho w_0^2$. Tracking such higher-order terms is typically not necessary.
Upon introduction of $w_0$, there is also a perturbation to $w_i$ for $i = 1, \ldots, N$, however these

perturbations do not enter directly into the expression for $w_0$ due to the bipartite structure of the interactions. In particular, given the perturbation above, $w_0$ follows...

## Response to Referee 2

"It would be helpful to expand the discussion on extensions to more realistic settings, such as: (i) correlations across input dimensions (e.g., Gaussian mixtures with general covariances), (ii) correlations across patterns, and (iii) unbalanced datasets with unequal label distributions."

We have added the following discussion to the Conclusion section:

The method can accommodate additional structure in data and labels. When the separating hyperplane passes through the origin and the pattern distribution $P(x_i^\mu)$ is symmetric about zero, binary labels $y^\mu \in \{-1, +1\}$ do not affect the calculation, making class imbalance irrelevant [8]. However, introducing a bias term or using asymmetric pattern distributions breaks this symmetry and makes class imbalance relevant [5], requiring that the labels are tracked through the calculation.

For correlated features in point classification, a simple form of correlation arises when $P(x_i^\mu)$ is asymmetric (e.g., delta functions at ±1 with different weights), yielding nonzero $\langle x_i^\mu x_j^\mu \rangle$ for $i \neq j$. Since patterns remain i.i.d., the method we have presented applies straightforwardly by retaining labels and accounting for different statistics in the disorder averaging.

A different correlation structure involves patterns drawn from a multivariate Gaussian with anisotropic covariance $\mathbf{\Sigma}$. In this case one can rotate the data into the eigenbasis of $\mathbf{\Sigma}$. The makes the problem equivalent to using features that are statistically independent but not identical, with the variance of the $i$-th feature given by the eigenvalue $\lambda_i$ of $\mathbf{\Sigma}$. The capacity is equivalent to that for classifying $P$ i.i.d. (equal-variance) points in an effective number of dimensions given by the participation ratio of the spectrum,

$$\mathrm{PR} = \frac{\left(\sum_{i=1}^N \lambda_i\right)^2}{\sum_{i=1}^N \lambda_i^2}. \tag{1}$$

More complex structures, such as Gaussian mixtures, would require additional variables for cavity calculations.

For manifold capacity problems, one can analyze cases where manifolds exhibit correlations in their centers or orientations, as has been done using the replica method [33]. Again, doing this with the cavity method would require additional variables.

"It would be helpful to add a comment on the role of convexity in enabling the derivations, and discuss possible extensions to nonconvex settings."

We have added the following to the Conclusion:

The benefit of convexity is that any solution of the KKT conditions (in classification problems) or zero-gradient conditions (in regression problems) is guaranteed to be a global optimizer. In non-convex settings, such conditions are only necessary but not sufficient for global optimality. In problems with complex energy landscapes characterized by exponentially many hierarchically organized solutions (e.g., spin glasses), replica symmetry breaking or equivalent techniques may be required.

"It would be useful to add a comment on the validity of the spectrum assumption (right above Sec. 4.4) in real tasks."

We have clarified this in Section 4.3:

We analyze this system in a high-dimensional limit where both $N$ and $P$ approach infinity with their ratio fixed. That is, that the number of modes $N$ needed to diagonalize the kernel with respect to the full data distribution $p(x)$ is on the same order as the number of training data $P$. One way this can arise is if $p(x)$ is supported on a finite set of $N$ points that are uncorrelated or weakly correlated under the kernel, and the training set consists of $P = \mathcal{O}(N)$ randomly subsampled points from this support.

"Please double check the a,b,c,d indices on the right hand side of eq. 81."

We have corrected this error. Thank you for catching this.

"Some references are missing: Section 4.1: it would be appropriate to cite the seminal work [1] on the derivation of worst-case rates, and to include citations to [2,3] for the typical case. Discussion: a very detailed derivation of the dynamical cavity method for the perceptron model is presented in [4]."

We have added all the suggested references. In particular, we now include:

In contrast to the typical-case behavior, [30] (Caponnetto & De Vito) considered worst-case behavior. [31] (Cui et al.) showed that differences in kernel ridge regression decay rates previously attributed to typical-case vs. worst-case analyses actually result from noiseless vs. noisy data assumptions.

We have also added the recommended references to Spigler et al. (2020) on kernel ridge regression, as well as to Agoritsas et al. (2018) on the dynamical cavity method for the perceptron learning.