We thank all reviewers for their helpful feedback. Below we address the questions and comments individually.

R1,R2,R3 - Writing. We will correct typos in the main text and bibliography, and refer to Figure 1 in the introduction.

R1 - Finite Data Analysis. We remark that: (i) while our results are asymptotic in nature, our experiments showed that they give accurate description of the prediction risk for dataset with moderate size, as demonstrated in Figure 2 ($n = 300$); (ii) we believe that similar to previous works on double descent, our asymptotic characterization can be translated to non-asymptotic guarantees using standard concentration tools.

R1 - Definition of "Same Order". We apologize for the confusion. As we specify a joint relation between covariances, we use the term "order" to describe the increasing or decreasing trend of a vector, i.e. "same order" of $\boldsymbol{a}$ and $\boldsymbol{b}$ implies $a_i \geq a_j$ iff $b_i \geq b_j$ for all i,j. When $\boldsymbol{\Sigma}_x$, $\boldsymbol{\Sigma}_\beta$ are codiagonalizable, "same order" of eigenvalues suggests that the features are informative (learning is "easy"); this is analogous to a fast eigenvalue decay of the RKHS "source condition".

R1 - Extension to Neural Network. We believe that our analysis can be extended to neural nets that are well-described by a kernel or random features model (which is linear regression under different features). This includes a two-layer network with fixed 1st layer (RF model), or with trained 1st layer under overparameterization (the NTK model).

R2 - Novelty & Comparison with [1]. We thank the reviewer for mentioning this reference – we were not aware of this work (which is arxived in May) at the time of submission. However, we believe that the reviewer has misjudged the relation and overlap between [1] and our work. We first note that computing the prediction risk is only a small part of our contribution; we also provided precise characterization of the optimal ridge parameter and weighting matrix. More importantly, the setup and theory in [1] are quite different than and *do not* produce our results. In particular:

- **[1] may not cover the case when $\lambda \leq 0$.** When the regularization strength λ is negative, it is not guaranteed that limit of VAMP exists due to non-convexity of the objective in the overparameterized regime. Even if we assume VAMP converges for all λ and estimation of generalization error is accurate, it is still not clear that VAMP converges to the specific ridge solution studied in our paper when λ is negative or 0 (which corresponds to minimum ℓ_2 norm solution), since it is possible that the SE has multiple fixed points when $\lambda \leq 0$. We remark that establishing convergence and uniqueness can also be challenging for analysis based on the Convex Gordon Min-max Theorem.
- **The VAMP framework does not capture our "aligned" or "misalgined" cases.** The key assumption in the VAMP analysis in [1] is that the limiting distribution of $\beta_i^\star$, the components of the true signal, is independent to that of the features. As a result, random permutation of $\beta_i^\star$ does not effect the generalization error of the VAMP estimate. This corresponds to the "random order" case in our setup, for which we showed that the corresponding optimal ridge parameter is always non-negative[1]. In contrast, we specified a general joint distribution between the eigenvalues of $\boldsymbol{\Sigma}_x$ and the components of $\boldsymbol{d}_\beta$ (See exact definition in paper). As demonstrated in Figure 2, different joint distribution leads to different generalization error even when the limiting spectral distributions of $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_\beta$ are the same. It is precisely this extension that enables us to explain the "negative ridge" phenomenon.
- **The analysis in [1] alone cannot characterize the optimal ridge parameter and weighting matrix.** As mentioned above, the optimal ridge parameter is always non-negative under the assumption in VAMP (it is also not true that SE can always be simplified to exact expressions.). Furthermore, we decided the optimal weighting matrix for optimal λ or the bias and variance term separately, and also covered the case when it only depends on feature covariance.

R3 - Optimal Choice of Weighting Matrix. We provide the following clarification and comments on the optimal $\boldsymbol{\Sigma}_w$.

- The data covariance $\boldsymbol{\Sigma}_x$ considered in Section 6 is the population covariance matrix, *not* the empirical covariance (which is degenerate when $\gamma > 1$). While it can be difficult to determine $\boldsymbol{\Sigma}_x$ from labeled data alone, the quantity can be estimated from additional unlabeled data (i.e., a semi-supervised setting[2]). We agree with the reviewer that designing a weighting matrix solely based on the empirical covariance is an interesting problem.
- We agree that interpolating between weighting matrices is an intuitive strategy: Figure 4 and 5 consider various powers of data covariance ($\boldsymbol{\Sigma}_x^\alpha$ for different α), which can be seen as a geometric interpolation between $\boldsymbol{I}_p$ and $\boldsymbol{\Sigma}_x$.
- As mentioned in the paragraph starting from line 251, given the data covariance, a reasonable decision rule is to construct the weighting matrix based on certain polynomial transformations to $\boldsymbol{\Sigma}_x$, the parameter of which can be tuned via cross validation. Figure 5 and 8 show that such approach does indeed outperform standard ridge regression.

R3 - On Assumption 2. We agree with the reviewer that assumption 2 does not cover all possible weighting matrices. We remark that this assumption is *not* needed in our risk calculation, but only in the characterization of optimal weighting matrix (for convenient formulas). In addition, it is worth noting that similar codiagonalizability assumptions is not uncommon in the theoretical study of regression models. For instance, the standard source condition in RKHS regression is analogous to codiagonalizable $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_\beta$ with certain eigenvalue decay in our setup.

[1] This partially justifies the restriction to non-negative λ in the VAMP framework.

[2] For example see: Tony Cai, T., and Zijian Guo. "Semisupervised inference for explained variance in high dimensional linear regression and its applications." Journal of the Royal Statistical Society: Series B.