We thank the reviewers for their detailed comments and insightful suggestions. Below we address some specific comments; we apologize for conciseness due to space constraints. (We thank Reviewers 1 and 8 for supportive reviews!).

Reviewer 3:

We thank the reviewer for a detailed and thorough review. Again, we apologize for our brevity; the nature of this topic warrants a much more careful and elaborate discussion, and it is unfortunate that we cannot do that here.

1. We were indeed inaccurate here: we will add an explicit projection step to the exposition of SGD and GD. As we stress below, this does not affect our technical development in any crucial way.
2. As stated, there is indeed a formal projection. Note that in all our constructions the trajectory remains at a bounded constant sized-ball. Thus, for a bounded and closed W, the projection step is avoided and does not affect our constructions—we will highlight this in the proofs where needed.
3. Indeed, Thm. 1 does not imply that generalization cannot be explained via implicit bias (we do not make such a claim). It only discusses strongly convex regularizers. As such, it definitely does not rule out the possibility of a bias towards solutions with *small enough norm*. Thm. 2, on the other hand, shows that for any (admissible) regularizer we can construct an instance where SGD converges to a point w_* even though there exists another point w_r that has the same empirical error but (strictly) better regularization penalty. We follow here the intuition that if r models the implicit bias of SGD then, given two solutions with same empirical error, SGD needs to choose (approximately) the one with smaller regularization penalty.
4. When comparing the output of SGD after enough iterations over the unregularized regression loss vs. ridge regression solution with fixed λ: in this case the output of SGD and the output of ridge regression solution are incomparable (in the Pareto-optimality sense). Namely, SGD will have a larger regularization penalty whereas ridge regression will have a larger empirical error; as such, it does not prove Thm. 1. The theorem states the existence of a problem where SGD converges to a solution that is not Pareto-efficient w.r.t. the empirical loss and ℓ_2 norm (not even approximately).
5. The reviewer is correct here that this requires further explanation. First, note that the assumptions of admissibility are used (and stated) only in Thm. 2, where we rule out *distribution-independent* regularizer. Any regularizer where $r(0) \neq \min r(w)$ can be ruled out in this setting by simply considering the zero function (i.e., $f(w, z) = 0$ for all z). In this case SGD converges to zero, which by assumption is r-suboptimal. Hence, we may assume here that $r(0) = \min r(w)$ and we only need to normalize by choosing $\min r(w) = 0$. (again, we emphasize that this concerns only distribution independent bias; in the distribution dependent section we are not making any assumption on the regularizer to begin with). Following this discussion we will revise the assumption to $\min r(w) = 0$ (and not $r(0) = 0$) and incorporate the discussion above in Thm. 2 where necessary.

Several other remarks:

- We thank the reviewer for indicating some important related work. The suggested lines of work are indeed relevant and should be discussed. Thank you for pointing those out!
- Two works which we discuss that give evidence for training neural networks without regularizations are [14,22]. Specifically, [14] shows how training to zero training error with overcapacitated networks can improve test performance.
- The remark on where SGD is identical to introducing ℓ_2 regularization: note that this claim is only for *linear* optimization (hence not relevant for regression). Perhaps, a better reference for this fact is Shalev Shwartz "Online Learning and Online Convex Optimization" (see Examples 2.3 therein).

Reviewer 7:

- *Typo in the notion of statistically complex set?*: No, the quantifiers are okay. Note that we want to understand if the structure of the set K can explain generalization, and without further considerations of the specific problem at hand. For that one must choose a complexity measure that is independent of D. Note that any (successful) learning algorithm converges, on a specific distribution, to a set of solutions that are okay for that distribution – thus changing the quantifiers will lead to a tautology. What we desire here is to measure the complexity, or understand the structure of the set of solutions w.r.t. any acceptable distribution. Specifically, Thm. 4 shows that on a given instance the set of solution can be "too rich" in the sense that the capacity of the set cannot explain generalization.
- *Remarks on quantification and improving clarity of Thm. 4*: We will elaborate and clarify here. The quantification over the regularizer is before the sample (i.e., the regularizer may depend on the distribution D but is independent of the sample S). We will also clarify the relationship between d and T; in a nutshell, the asymptotic variable is T (and $d = O(T)$, $\eta = 1/\sqrt{T}$). Thanks for these suggestions!