Reviewer 1 and 4

We thank for the reviews and will resolve the main concerns. We sincerely ask the reviewers to re-evaluate the rating.

- Reviewer 1: "It is pointless to study the gradient descent method in this setting. Also, the proposed method, on the contrary to standard weight normalization [52], can not generalize to nonlinear or higher dimension case."
- Reviewer 4:"The loss function for the theoretical analysis is over-simplified, even does not take weight decay into consideration. The statement that WN is equivalent to rPGD does not hold with weight decay since WN would suffer instability at w=0 while rPGD does not.

Linear regression is a fundamental theoretical problem. When applying WN to linear regression, it becomes *non-convex* optimization. Moreover, the most important part in our setting is "under-determined".

We apologize for our word "proposed" in the paper and will remove this word. We would like to emphasize that our paper is **NOT** about proposing a new method (i.e., rPGD) but to **theoretically** understand the implicit regularization effect of these methods. rPGD is an existing method [13]. We build a surprising connection between rPGD and WN, which is exact equivalence under some condition (see Lemma 2.2). With infinitely small step-size $\eta \to 0$ and initialization $\|w_0\| = 1$, the equivalence of gradient flows of the two methods under the nonlinear or high dimension case will be maintained. However, when the stepsize is not small, the two methods are not the same as the norm $\|w_t\|$ grows for WN, while $\|w_t\| = 1, \forall t > 0$ for rPGD. Since we focus on implicit regularization, we do not want to involve the growing norm $\|w_t\|$ and so study rPGD, not WN.

We did not consider "weight decay" as our motivation is to study implicit regularization (IR) along the lines of [22]. Understanding algorithms **without** explicit regularization is the starting point for studying IR. If weight decay is used for linear regression, the problem becomes strongly convex and has a unique solution. However, in future work the referee's suggestion may be interesting as WN makes this setting (linear regression with weight-decay) non-convex.

[13] Douglas, Amari, Kung. "On gradient adaptation with unit-norm constraints." IEEE TSP 48.6 (1998): 1843-1847.
[22] Gunasekar, Suriya, et al. "Implicit regularization in matrix factorization." NeurIPS 2017.

- Reviewer 1: "The experiments show that the proposed method derives a similar/smaller final norm compared to standard weight normalization. However, this does not prove that the method is useful. In fact, it only shows that this method provides a stronger constraint on the norm of the weight."
- Reviewer 4: "There are no empirical support for the conclusions."

We would like to explain that our experiments are to support the theory and not to show the "usefulness" of rPGD. We want to show the implicit regularization along the research line [22]. You are right that the rPGD is likely to provide a stronger constraint on the norm of the weight.

- Reviewer 4: "The implicit regularization effect of weight projection has been talked previously in e.g. arxiv:1710.02338. This study is only marginal. In the appendix A, the discussion on whether the term is larger than 0 is missing!"

Thanks for the reference. We do not agree that our study is only a marginal improvement. Thanks for pointing out the discussion on whether the term is larger than 0. We now address here and will add to the paper. The regularization parameters are highly dependent on g_t, g_{t+1} and the input matrix A. However, it is difficult to characterize the behavior of λ_t in general. In particular, we require the parameters g_t, g_{t+1}, w_t and w_{t+1} updated in a way that $\lambda_t > 0$. For the simpler setting of orthogonal A, we can see for rPGD that: 1) If the learning rate of g is small enough, we will have $g_{t+1} < g_t\|v_t\|$, which means that $\lambda_t > 0$; 2) When $g_t w_t$ is close to $g^* w^*$, we will have $\|v_t\| \approx 1$, and $g_{t+1} \approx g_t$, which means that $\lambda_t \approx 0$.

Reviewer 2 and 3:

We thank the reviewers for the positive evaluation.

- "Your analysis heavily depends on the data matrix $A^T A$. For SGD with WN, however, $A_i^T A_i$ might not commute thus cannot be diagonalized simultaneously. Due to this reason I guess it is quite non-trivial to extend your analysis to SGD with WN. Can you comment on the implicit bias of SGD with WN, which is a more practical optimizer?"

Indeed, it's not trivial to extend the continuous-time analysis to SGD with WN as we need to look at $w^\perp$, which depends on A. It is very challenging to analyse the discrete-time SGD case when updating both g and w, because g and w are random variable and, by taking expectation, their product is hard to analyze. A possible alternative may be to look at the stochastic Langevin dynamics or making g fixed.

- "The implicit bias for GD actually holds for quite a general class of losses in additional to l2-loss. Can you comment on your results for other losses, e.g., l4-loss and exponential loss?"

This is great question, we have thought about this. For L_p loss, we need to think about the what norm should be used for the weight norm algorithm. With L_4 norm for the WN algorithm and L_4 loss, then $w_t^{\circ(3)}$ ($\circ$ is Hadamard power) is involved and the norm $\|w_t\|_4$ is no longer constant, which makes the dynamics harder to analyse.