R1: Q1: It is hard to understand what Remark 6 conveys.
A: Yes, the error bound condition refers to the inequality in Lemma 1. Lemma 1 implies that the error bound condition is weaker than the PL condition. Remark 6 means that our analysis only requires the error bound condition to hold though we assume the PL condition at the beginning since it is more widely known in the deep learning literature. The reason that we include this remark is that our experiments directly verifies the error bound condition.

R1: Q2: how the bound can affect or guide a specific choice of K in stagewise SGD.
A: The value of K depends on the choice of ϵ (i.e., optimization error). Theoretically, the testing error bound (e.g., Theorem 5) allows us to find an ϵ that balances the optimization error and the generalization error. However, it only affects K in a logarithmic way. In practice, it is just a small number.

R1: Q3: It might be better to use “START" in the paper title and Figure 1, instead of “SGD".
A: Thanks for the suggestion! We will make the change. Indeed, "stagewise SGD (Vx)" are variants of START with different algorithmic choices. Their common feature is the stagewise step size scheme. stagewise SGD (V1) does not use algorithmic regularization (with $\gamma = \infty$). SGD($c/\sqrt{t}$) and SGD(c/t) refer to the vanilla SGD (not covered by START) using two different polynomially decreasing step sizes. The comparison between stagewise SGD (Vx) and the vanilla SGD demonstrates the importance of the stagewise step size scheme, which is the key point of this paper.

R1: Q4: This paper mentions “Corollary 2" several times.
A: Sorry for the confusion. Corollary 2 should be Theorem 2.

R1: Q5: In Line 49, definition of μ. In Line 123, the assumption $|f(w, z)| \leq 1$ is strong.
A: μ refers to the constant in the inequality (PL condition) just before line 49 (between line 46 and line 47). We will make it clear. The boundness assumption on $f(\mathbf{w}, \mathbf{z})$ is only used in the stability analysis for non-convex loss functions. This is following the analysis in [13]. Since a general upper bound $|f(w, z)| \leq M$ only affects the result by a constant factor, we said without loss of generality.

R2: We thank R2 for all comments.
R2: Q1: what is the main take-away message for me after reading this paper.
A: In this paper, we focus on comparing two different step size schemes instead of challenging the classical framework that either analyzes the optimization error convergence or the generalization error of SGD with a particular step size scheme. Most existing theoretical analysis of SGD uses a polynomially decreasing step size or a small step size. However, in practice people mostly use a stagewise step size for SGD, which decreases in a stagewise fashion geometrically. **The main takeway message of this paper is that we give the first theory to justify why the widely used stagewise step size scheme gives faster convergence than a polynomially decreasing step size, i.e., the stagewise step size scheme can adapt to the nice properties of deep neural networks.** That is why we compare the results in Theorem 5 and Theorem 9 (using a stagewise step size scheme) with the result in Theorem 2 (using a polynomially decreasing step size), and in Figure 1 we compare stagewise SGD with SGD with a polynomially decreasing step size.

R2: Q2: Is $\mathcal{A}(S)$ a random variable, or a random probability measure.
A: We use $\mathcal{A}(S)$ to denote a randomized model returned by the algorithm $\mathcal{A}$ based on the dataset $\mathcal{S}$. Basically it is a random variable. Please refer to line 89. So $f(\mathcal{A}(S), Z)$ means the loss of the randomized model found by algorithm $\mathcal{A}$ on a random data Z.

R2: Q3: Why is there a <= in the error decomposition after line 96?
A: Yes, it is indeed an equality.

R2: Q4: It seems that F_w^γ is never used in this routine? What does the function O represent at stage 5 of Algorithm 2?
A: We will find a better way to present it. The structure of F_w^γ that decomposes to the original objective and a quadratic regularizer is used in Algorithm 2. The function $\mathcal{O}$ returns a solution given a sequence of intermediate solutions.

R2: Q5: Line 257: I am confused by the way you want to choose theta. I assume you want theta to be very large?
A: We expect a larger θ in order to explain the advantage of stagewise SGD compared with the vanilla SGD with a polynomially decreasing step size, since the vanilla SGD has a complexity of $O(1/(\mu^2\epsilon))$ for reaching an ϵ-level of optimization error and the considered stagewise SGD has a complexity of $O(1/(\theta^2\mu\epsilon))$. Our results in Theorem 6 and Theorem 9 indicate that the larger the θ, the faster the convergence of optimization error and the smaller of the testing error. Nevertheless, θ is a property of the function. A convex function has $\theta \geq 1$. Our experiments verify that for deep neural networks θ is also around 1.

Response to R3: Thanks for the comments. The green curve is actually for the μ values across all iterations (the same as θ). We just add a number to mark its average value so that readers can have a sense how small is the μ as the curve is almost on zero. We will add discussion to discuss the limitation of the presented results.