We thank all reviewers for their comments and suggestions!

Reviewer 1: Q1. About the benefits of the newly-proposed algorithms.
A. First of all, we emphasize that our goal is not to develop an algorithm for solving ℓ_0 norm constrained problem and prove an exact recovery result, but rather to analyze stochastic proximal gradient (SPG) for handling a general non-convex regularizer under minimal assumptions about the problem. The benefits of the newly-proposed algorithms is that **it is applicable to a much broader family of problems**. First, we are not restricted to ℓ_0 norm constrained or regularized problems. As long as the regularizer's proximal mapping can be efficiently computed, our algorithms and their convergence guarantee are applicable (c.f. our experiments for learning with quantization). Second, we **do not impose stringent condition** on the data matrix or the loss function, such as restricted isometry property or restricted eigenvalue or restricted strong convexity that is typical for traditional sparse recovery algorithms (e.g., IHT, StoIHT). Third, our results are applicable to any smooth loss functions even if they are non-convex, while most previous results are restricted to convex loss. We believe adding such stringent conditions one could derive much stronger result of SPG for ℓ_0 norm constrained problems following existing works (e.g., [R1,R2]). But it is not the focus of this paper.

Reviewer 1: Q2. About Theorem 5 and convergence.
A. Thanks for this great question! Please note that this is not an error. We will make the statement of Theorem 5 more clear in the revision (somehow the current upper bound in Thm. 5 is to capture the online setting). In fact, for the finite-sum setting, the second term $(\gamma + 4\theta L)\sigma^2/(2\theta L|S_1|)$ will disappear in the upper bound since it is caused by the variance of stochastic gradient $\nabla f_{\mathcal{S}_1}(\mathbf{x}_t)$ (c.f. Line 440 of supplement). We have briefly explained in the proof of Corollary 6 for the finite-sum setting (c.f. Line 478 of supplement) and will add more details. We will present Theorem 5 in a better way by considering the online and finite-sum setting separately. For the online setting, the current bound holds without any change, for the finite-sum setting the upper bound only includes the first term. Thanks again!

Reviewer 2: Q. How to justify the Assumption 1 (ii)?
A. This assumption is quite standard and has been used in many non-convex literatures (see references [18, 19, 29, 31, 35, 41]). As long as the objective function is lower bounded, the assumption holds without assuming a compact domain. In most machine learning applications the objective function is non-negative, i.e., $F(\mathbf{x}) \geq 0$. Hence, one can simply set $\Delta = F(\mathbf{x}_0)$.

Reviewer 2: Specific Comments and Improvements.
A. We thank the reviewer for all comments. We will improve the paper following on the reviewer's comments and add more discussion on the bounded variance assumption in connection with [29]. Thanks for the positive rating!

Reviewer 3: Q. About the constant learning rate with a large mini-batch size vs decreasing learning rate with a small mini-batch size.
A. While we agree with the reviewer that an algorithm with a decreasing learning rate and small mini-batch size is interesting, it might be unfair to say that an algorithm with large mini-batch size and constant learning rate is not practical. At least, in the distributed setting it is more natural to consider a large mini-batch size rather than a small batch size [R3]. Indeed, we have presented a variant with an increasing sequence of mini-batch sizes rather than a large mini-batch size from the beginning. It is still an open problem to prove the **non-asymptotic convergence** of SPG without using a large mini-batch size for a non-convex regularized problem (An asymptotic analysis of SPG without a large batch size is presented in [15]).

Reviewer 3: Improvements.
A. We will formally define the practical algorithm. Thanks for the positive rating!

Reference:

[R1]. Linear Convergence of Stochastic Iterative Greedy Algorithms with Sparse Constraints. Nguyen et al. 2014.

[R2]. On Iterative Hard Thresholding Methods for High-dimensional M-Estimation. Jain et al, 2014.

[R3]. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. Goyal et al. 2017.