We thank the reviewers for their time and feedback. To our knowledge, our work provides the first *practical* algorithm with *provable* iteration complexity for solving nonconvex optimization problems with nonlinear constraints which has numerous applications in machine learning. We argue that the proposed framework will be a staple for many such applications in the future, and our theory-based code will be made public, which will have sustained impact.

We first address the concerns of R3 on condition (15) and we wish R3 would reconsider their score:

1. To our knowledge, our condition (15) is new and inspired by our theoretical analysis. Mangasarian-Fromovitz constraint qualification, that is commonly assumed in the literature, is assumed on a specific region of the space, whereas our condition is algorithm dependent which we believe to be a weaker requirement. This is precisely the reason why the iteration count appears in the condition. We find this to be a strength rather than a weakness.
2. On the relation of (15) and β_k: We consider two cases, when g is the indicator of a convex set (or 0), the subdifferential set will be a cone (or 0), thus β_k will not have an effect. On the other hand, when g is a convex function defined on the whole space (please see Thm.3.1.13 of Nesterov's book), subdifferential set will be bounded. This will introduce an error term in (15) that is of the order $(1/\beta_k)$. One can see that b^k choice for β_k causes a linear decrease in this error term. In fact, all the examples in this paper fall into the first case. However, for generality, we will clarify these two cases in the main text.
3. The specific choice of $\beta_k = b^k$ was motivated by practical performance, and as argued above, compatible with (15).
4. We in fact validate (15) in Appendices E and F for some key examples. In the convex case, the geometric variant of (15) holds iff the Slater's condition holds (under additional assumptions).
5. The condition (15) is discussed in detail in the 'Regularity' paragraph in Section 4 with its connections to the existing conditions. We will add more intuition; however, we have been quite exhaustive in our literature review and strongly believe in the utility of (15) in the future.

Moreover, we respectfully disagree with R3 on the negativity of the numerical results. In fact,

1. We apply our algorithm to five diverse applications, which by itself is notable for a theoretical submission. In all cases, we perform head-to-head or better than the highly tailored algorithms for each application. These algorithms often do not have the generality and flexibility of our framework, which cannot be understated. Please see for instance the discussion in supplementary material, starting line 578.
2. We verify condition (15) for three key examples; clustering, basis pursuit and generalized eigenvalue decomposition.
3. We propose a nonconvex formulation of the standard basis pursuit template which can handle non-seperable priors (e.g., structured norms such as those arising in group Lasso) for which the baselines methods are not applicable. Please refer to 'discussion' paragraph in Appendix E.1.
4. We provide new state-of-the-art results for QAP, which is considered a difficult problem. We perform at least as good as the baseline for 18 datasets out of 19.

R2's double submission remark: We will clarify the differences in the camera ready. Specifically, 6483 considers strongly convex optimization subject to nonlinear constraints with an ADMM framework different from our general setting of nonconvex optimization and ALM instead of ADMM. The assumptions required for nonlinear operators are also different. In short, our submission is solving a more general framework than 6483, with a different algorithm. This requires different analysis and substantially different results. Moreover, results also do not reduce to each other in the specific cases.

Response to the remaining comments are in the sequel.

- **R1** & **R2:** As suggested by R1 and R2, we will summarize our literature review as a table for improved readability.
- **R2:** In clustering in section 6, $g \neq 0$ is the indicator function of a convex set, which justifies the BM example.
- **R2:** When $g = 0$, similar theoretical guarantees with more complicated (not simple to implement) algorithms are obtained in [7, 18]. Our proof idea is similar in the $g = 0$ case. We will include the idea of the proof in the main text.
- **R2:** The particular choice of σ_k is to keep the dual sequence bounded (see line 98). We will add more discussion for the differences in the nonconvex case.
- **R3:** Please see above for our clarification involving Condition (15).
- **R3:** The method we analyze is the classical ALM, the intuition for which is well-understood, and we will highlight this briefly in the final version. Condition (15) stems from the analysis, and carefully described in the text.
- **R3:** The technical results are self-contained, except $y_{\max}$, for which we had given a pointer to its expression for the sake of space. We will move back the expression for $y_{\max}$ to the text.
- **R3:** Please see above for practical remarks. We ran additional experiment for clustering on MNIST digits (in addition to fashion-MNIST) and converged at least 5 times faster with both apgm and lbfgs solvers. We can include these results in the camera ready.