We sincerely appreciate the effort given by reviewers for the time to read, review and provide insightful comments. We will fix all the typos (in particular Eq.(2.5)). We thank the reviewers for pointing out the crucial bug in the display of citations. In general, we will improve the presentation, the experimental section and provide open source code. We now address the reviewers questions in detail.

Assumptions A and B are mostly satisfied in practice. Assumption A where C is the full space and $h = \frac{1}{2}\|\cdot\|^2$ is the standard setting of Proximal Gradient Descent. Examples such as QIP [8] and Cubic Regularization [45] satisfy the assumption for non-Euclidean function h. For the generalization to strict subsets C of $\mathbb{R}^d$, Assumption A uses a flexible formulation that does not require the function g to be smooth on the full space (only required on an open set containing $\operatorname{dom} h$). Also $\operatorname{dom} f \cap C$ has to be non-empty otherwise the optimization will take place over an empty set, which is not desirable. Moreover, one is often interested in reaching global minimum and do not expect to decrease the function value to negative infinity, hence the lower-boundedness. Assumption B ensures that the updates of BPG-methods are well-defined with mild requirements as explained on page 6 in [8], for example, under a classical constraint qualification.

Statistical evaluation experiments. From the empirical results given in the paper, BPG-methods are seemingly better than (or at least as good as) PALM-methods. We believe that the trajectories taken by both the methods are different and is source of good performance of BPG-methods over PALM-methods. The theoretical justification however is an open question, which we could not justify. We agree that statistical evaluation varied with initialization is an important experiment and we will add the proposed experiments. With these additional experiments, we hope to find some new insights (similar to preliminary results below) on the convergence behavior of BPG-methods vs PALM-methods. Our new preliminary statistical evaluation results with setting used in "Figure 2 with No regularization" show that PALM-methods get stuck at large objective values (≈ 96728.941 on average) however BPG-methods do not (≈ 378.173 on average). We will give detailed experiments along with precise settings in the next update.

Computational benefits and insights. We briefly remark some properties of the update steps of BPG-methods. Note that the updates are independent for $\mathbf{U}$ and $\mathbf{Z}$, for example, see the simple illustration on Page 2, where updates can be done in parallel blockwise (ignoring the 1D cubic equation). This should increase the speedup in practice, in particular for large matrices. Also, note that some terms in gradients overlap, so using temporary variables in implementation can increase the speedup. We will discuss these tricks, the time and computational complexity in detail in conjunction with the timing experiments along with a brief pointer in the main paper, as per suggestion by Reviewer 3. We now provide insights on why BPG-methods are a better choice over other methods, with focus on alternating methods.

- PALM-methods estimate a Lipschitz constant with respect to a block of coordinates in each iteration, which is expensive for large block matrices. BPG-methods use a global L-smad constant, which is efficient.
- PALM-methods cannot be parallelized block wise, for example, in the two block case, the computation of the Lipschitz constant of second block must wait for the first block to be updated, hence it is inherently serial.
- Alternating minimization methods do not converge for non-smooth regularization terms and can be inefficient (for, e.g., ALS) for some matrix factorization problems (see, for example, "Tensor Decompositions and Applications" by T. G. Kolda and B. W. Bader, also see "On search directions for minimization algorithms" by M.J.D. Powell). But, BPG-methods and PALM-methods converge (due to linearization).
- PALM is not applicable to the 2D function $g(x, y) = (x^3 + y^3)^2$, because the block-wise Lipschitz continuity of the gradients fails to hold even after fixing one variable. BPG-methods are applicable here.
- PALM is not applicable to, for example, symmetric Matrix Factorization as also pointed in [20] or the following penalty method based (relaxed) orthogonal NMF problem (similar to Eq.1.1)

$$\min_{\mathbf{U}\in\mathcal{U},\mathbf{Z}\in\mathcal{Z}} \left\{ \Psi \equiv \frac{1}{2}\|\mathbf{A}-\mathbf{U}\mathbf{Z}\|_F^2 + \frac{\rho}{2}\left\|\mathbf{U}^T\mathbf{U}-\mathbf{I}\right\|_F^2 + \mathbf{I}_{\mathbf{U}\geq\mathbf{0}} + \mathbf{I}_{\mathbf{Z}\geq\mathbf{0}} + \mathcal{R}_1(\mathbf{U}) + \mathcal{R}_2(\mathbf{Z}) \right\}, \qquad (1)$$

  where second term does not have a block-wise Lipschitz continuous gradient for any $\rho > 0$. But, here BPG-methods are applicable (similarly also for Projective NMF) with minor changes to the Bregman distance.
- The block wise Lipschitz continuity of gradients was the primary motive for PALM-methods. Now, with the BPG-methods, we can directly tackle the original problem with L-smad property.
- BPG-methods are very general so the choice of applications will increase substantially and we believe that this will open doors to design new losses and regularizers, without restricting to Lipschitz continuous gradients.

We agree that an expanded discussion of competing methods is crucial for the paper. We will also discuss the points mentioned in the previous answers in detail in the expanded discussion along with state of the art matrix factorization models and new perspectives with BPG-methods.