We thank all the reviewers for carefully reading of the manuscript and constructive comments. We addressed the issues raised in our response below, mostly focusing on the numerous questions raised by Reviewer 3.

Reviewer #1: Assumptions A.2 and A.3 used in Algorithm 2**.** Assumption A.2 was also used in Mairal et al. [1], while A.3 is a common assumption in linear regression analysis and it relates to the LARS problem in our online cvxMF setting. We agree with Reviewer #1 that the modifications described in detail on page 6 of the manuscript may have produced practical performance improvements in the proposed algorithm. We will report on these results in the revised paper. We wanted to avoid notational overload and exactly match our analysis with every step of the algorithm which is the reason why we did not dwell on testing the proposed modifications.

Reviewer #2: Improving the clarity of the technical sections. We agree with the reviewer that the notation may be hard to follow, but the problem setup is such that most of it is necessary. We will make every attempt to further improve readability through examples and detailed comments in the Supplement.

Reviewer #3: There seem to be several misunderstandings regarding the steps of our algorithm and its analysis.

1. MF versus NMF. Our results (both the proof and the algorithm) are independent of the non-negativity assumption and apply to both NMF and MF problems. To enforce non-negativity, several simple projection steps, akin to those described in Mairal et al. [1], suffice. For simplicity of exposition, and to be able to compare our results with the main result in [1], the derivations were presented for the classical MF problem only.

2. The role of K-means in the online algorithm. We would like to point out that K-means is used only once to initialize the representative sets and is not an intrinsic component of the online algorithm. Furthermore, K-means is performed on small subsampled datasets, as running it on complete datasets is time consuming and unnecessary.

3. The role of the constant N**.** A reasonable choice of N in the initialization and update phase depends on the size of the dataset n, the dimensions of the data points m and the number of clusters k. What is important to observe is that N is kept constant throughout in order to reduce the storage footprint and to ensure low-complexity online processing. Whenever a new point is fetched, it is compared against other points for inclusion into the representative regions. To maintain the list constant, for every added point another point is removed. Also, the reviewer is correct in observing that N does not feature in the convergence results, which are asymptotic and do not imply anything about the convergence rate. Clearly, if the point dimension m is large, it is beneficial to increase N. A final remark is that the N representative points are used to generate the convex bases for the space and not to "cover" the clusters.

4. The perfect assignment assumption for the restricted online cvxMF is unrealistic. First, we point out that both the algorithm and the convergence analysis in our paper mostly focus on the unrestricted online cvxMF problem. In this case, the cluster assignment for a newly retrieved data point is made uniformly at random over the set of all possible clusters; we proved that this random assignment suffices to find a stationary point of the problem. The set of representative points still contains N data samples, and the convex hull property is still valid, except that the bases are not required to lie in the convex hull of points from the same cluster. Clearly, this unrestricted case can only have better performance than the restricted setting as nothing in the algorithm or the proof relies on correct classification. This point is illustrated in Figure 1: Even when the new sample (triangle in second cluster) is misclassified as belonging to the bottom cluster, the basis representing the latter cluster (red star) remains in the convex hull of representative set; the newly added point is retained in the representative set if the objective decreases and discarded otherwise.

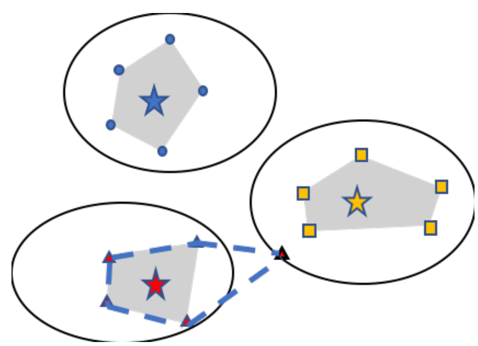

Figure 1: An incorrect assignment of a data point to a different cluster does not affect the "convex hull" constraint; the misclassified point (triangle) is still part of the representative set used to describe the basis (star). Hence, in this unrestricted representative region setting, classification errors do not influence the performance of the method nor do they compromise convergence guarantees.

The restricted version is proposed because of its practical utility and ease of interpretation as we explained in great detail on page 2, in the Introduction of the main text. In this setting, one requires the representative set to be partitioned into k representative subsets, each of which is restricted to be contained in its corresponding cluster. The basis is consequently restricted to be in the convex hull of data points from the same cluster. To satisfy these conditions we indeed require a "perfect assignment," which is possible in many supervised and semi-supervised learning tasks where the labels of points are known beforehand. We once again point out that the main goal of our algorithm is not to accurately classify the data point, but to find the optimal convex bases in an online manner and retain a small list of representative points from the data set. Even with the availability of labels or perfect cluster assignment it is nontrivial to compute bases that satisfy the restricted convexity constraint with an online algorithm. As shown by our experiments, the heuristic classification step that we proposed to use in practice for the restricted version works very well and provides excellent results for several real world tasks.