R1: *"non-conservative flag"* : In **practice** the algorithm should **always** be run non-conservatively to obtain a regret bound as that is the more flexible/noise-tolerant bound. We **only** include the conservative case for presentation issues as the bound is *simpler*, analogous to Novikoff's (perceptron) bound, and in order to give "easy" upper and lower bound insights in that case (e.g., lines 244-249, 350-353). We will update the manuscript to reflect this advice.

" For Theorem 1, ... non-conservative = 0? ... don't depend on the value of γ at all, so why does it show up in the bounds?" : See line 172 - 173, 1) the algorithm's predictions depends on γ through η ($\eta = \gamma$). 2) the bound fails hold if for the "true" comparator matrix $\boldsymbol{U}$ the margin complexity is large i.e., $\mathrm{mc}(\boldsymbol{U}) > 1/\gamma$.

"modification upon MEG ... (e.g. were any new proof techniques required)." : See 109-112 which then refers to Appendix B.2 (esp. lines 520-536) which discusses our new proof techniques wrt MEG.

R2: *"Why does mc($\boldsymbol{U}$) ... resembles the margin in SVM"*: In lines 2-8 of the abstract we discuss the term $1/\gamma^2$ which serves as the algorithmic proxy for $\mathrm{mc}^2(\boldsymbol{U})$ i.e., the tightest bound is obtained if $1/\gamma^2 = \mathrm{mc}^2(\boldsymbol{U})$ and it is in those lines that we make the explicit connection to the "SVM margin." This initial discussion is then built on in lines 54-61 where we further refer to the three references [5,6,7].

"The latent block structure seems not well-motivated from a practical part. In Netflix, it should be a five rated problem ... , the latent block structure may be a strong assumption and not easy to be tested in reality." : See lines 40-53 (esp. 51-53). Observe that the latent block structure (LBS) assumption is actually a *weaker* assumption than the common low-rank assumption. Note that Netflix does not meet the low-rank assumption as it is experimentally known that human ratings are only *ordinally* qualitative. Finally the definition of LBS trivially generalizes to categorical data as do our algorithms.

R3: *"Lower bounds"* : We agree that there is value in formalizing the setting. Our casual mistake lower bounds generalize to lower bounds for regret using [*Agnostic Online Learning*, 2009, Ben-David et al; Lem. 14].

R4: *"(c) The exposition ... Section 4.1, can be simplified ... matrices M and N directly in Section 4.1 rather than arriving at them from a graph perspective).* Yes, an alternate presentation is to assume feature vectors associated with each row and column and let $\boldsymbol{M}$ and $\boldsymbol{N}$ be the inverse of the gram (kernel) matrices (also see 251-256). However we focus on the graph perspective because of the smaller bound on $\mathcal{D}$ wrt graphs given in Thm 3. Eq (7). We note in the batch setting different methods of using graph side information were considered in [15,16] (lines 87-94). In the inductive setting (Sec. 5.1) we give an example of using non-graph side-information. Finally we note that 4.1 serves as the necessary background for 4.2 which has two important observations: 1) a bound from [22] can be recovered and extended, 2) An example of a class of $k \times k$ biclustered matrices (whose rank are k) for which max norm is $O(1)$.

"(d), (e)" : We will move related work and use forward references for undefined notation.
" (f) What are $\mathcal{R}_{\boldsymbol{M}}, \mathcal{R}_{\boldsymbol{N}}, \mathcal{R}_{\boldsymbol{L}}$ ..." : See line 129. The *"Iff"* (147-148) .. follows from definition of $\|\boldsymbol{U}\|_{\max}$.

"(2) The transductive setting seems unnatural. Why is it reasonable to assume that $\boldsymbol{M}$ and $\boldsymbol{N}$ will be available beforehand?" : We argue side-information is the norm rather than the exception. I.e., in the Netflix example we may have demographic info on the *users* as well as categorization, actor lists, etc on the *movies*. $\boldsymbol{M}$ and $\boldsymbol{N}$ may then be constructed from feature vectors by selecting kernels and inverting the kernel matrices. Or as in graph-based semi-supervised learning, a graph may be constructed using the feature vectors and using the corresponding Laplacian.

"(3) ... experiments (4) ...time complexity ..." : We agree that experiments would be useful and that the time complexity is large. However, we note that natural heuristics include maintaining only a low-rank approximation to $\tilde{\boldsymbol{W}}^t$ and maintaining a fixed number of indices in $\mathbb{U}$ (e.g., decaying old indices) for Algs. 1 and 2, respectively.

"... counterintuitive ... inversely proportional to gamma. ... Shouldn't more mistakes be made when the margin requirements are higher?" : See lines 3-4 of the Abstract. This is the usual intuition behind perceptron, SVM, and other "largin margin" classifiers. I.e., the further that data in classes are apart the easier it is separate and thus (online) fewer mistakes (batch) better generalization as opposed to the case where the classes are arbitrarily near one another.

"How is the max norm block invariant? I believe that as the sizes of the matrices are different, they will be on different scales." : **Theorem: max-norm is block invariant.** *Pf. Sketch.* We first show $\|\boldsymbol{X}\|_{\max} \geq \|\boldsymbol{R}\boldsymbol{X}\boldsymbol{C}^\top\|_{\max}$ for all $m, k, n, \ell \in \mathbb{N}^+$ with $m \geq k$, $n \geq \ell$, $\boldsymbol{R} \in \mathcal{B}^{m,k}$ and $\boldsymbol{C} \in \mathcal{B}^{n,\ell}$ (where $\mathcal{B}^{m,k}$ is the set of all $m \times k$ block expansion matrices (cf lines 140-143)). WLOG. assume $\boldsymbol{X}$ is $k \times \ell$. If $\boldsymbol{P}\boldsymbol{Q}^\top = \boldsymbol{X}$ then $(\boldsymbol{R}\boldsymbol{P})(\boldsymbol{C}\boldsymbol{Q})^\top = \boldsymbol{R}\boldsymbol{X}\boldsymbol{C}^\top$. Observe that $\max_{i \in [k]} \|\boldsymbol{P}_i\| = \max_{s \in [m]} \|(\boldsymbol{R}\boldsymbol{P})_s\|$ since every row in $\boldsymbol{P}$ is duplicated by (1+) rows in $(\boldsymbol{R}\boldsymbol{P})$ and there are no distinct rows in $(\boldsymbol{R}\boldsymbol{P})$ that are not in $\boldsymbol{P}$. Recall $\|\boldsymbol{X}\|_{\max} := \min_{\boldsymbol{P}\boldsymbol{Q}^\top = \boldsymbol{X}} \{\max_{1 \leq i \leq m} \|\boldsymbol{P}_i\| \times \max_{1 \leq j \leq n} \|\boldsymbol{Q}_j\|\}$ then since for every decomposition $\boldsymbol{P}\boldsymbol{Q}^\top = \boldsymbol{X}$ there exists a decomposition $(\boldsymbol{R}\boldsymbol{P})(\boldsymbol{C}\boldsymbol{Q})^\top = \boldsymbol{R}\boldsymbol{X}\boldsymbol{C}^\top$ thus $\|\boldsymbol{X}\|_{\max} \geq \|\boldsymbol{R}\boldsymbol{X}\boldsymbol{C}^\top\|_{\max}$. We have $\|\boldsymbol{X}\|_{\max} \leq \|\boldsymbol{R}\boldsymbol{X}\boldsymbol{C}^\top\|_{\max}$ since trivially the max-norm of a sub-matrix cannot be larger than that of the matrix ■.

"Where is the minimum in (7)?" : See line 205. The minimum is before the large '{' in (7) and is over the matrices $\boldsymbol{R}, \boldsymbol{C}, \boldsymbol{U}^*$ s.t. $\boldsymbol{U} = \boldsymbol{R}\boldsymbol{U}^*\boldsymbol{C}^\top$.