We start by thanking the reviewers for their detailed reviews and comments that will help improving the final version of the paper. Below, we address the different remarks made by the reviewers.

——- Theoretical aspects ——-

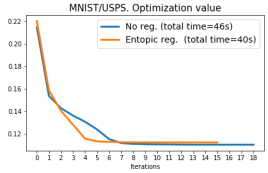
(**R1: COOT is a distance**) COOT is a distance in general in the permutational sense. When $n \neq n'$, $d \neq d'$ we have indeed COOT > 0. That is what we awkwardly meant by "identity of indiscernibles cannot be proven". We will clarify it. (**R2: COOT as a tight convex relaxation of BAP**) The reviewer is correct: the problem is not jointly convex in (π_s, π_v). By convex relaxation, we target specifically the set of constraints but keeping the latter tight as we still recover a solution of the original BAP problem. (**R3: COOT and GW**) COOT includes GW as special case and both are the same when the problem is concave (*e.g* when $\mathbf{X}, \mathbf{X}'$ are squared Euclidean distance matrices) as discussed in Prop 3. (**R3: missing theoretical grounding**) As mentioned in the conclusion, the continuous formulation of COOT is indeed of high interest. We chose to focus on studying its discrete version with use-cases that are more relevant for the ML community. We hope this work will pave the way for more theoretical studies on this particular novel instance of OT problem.

——- Experimental settings ——-

(R1: choice of L) In all experiments we found $L = |\cdot|^2$ to be efficient, but we agree that a deeper analysis on its choice can be relevant for future works. **(R1: which regularization?)** For co-clustering, we use entropic regularization on features and samples to obtain soft clustering assignments. For HDA experiment, we use entropic regularization on features only as the number of samples is relatively low, and following practices of OT in domain adaptation where the entropic regularization proved to be efficient for handling such cases. **(R2: low scores for EGW)** We confirm the low scores for EGW. While we acknowledge that the choice of the hyperparameter might not be optimal, we observed that the score on the test set remained low for most of the values tested. Contrary to [16], the features here are more high-dimensional (DeCAF and GoogleNet). We suspect that EGW cannot handle the cases where n is low and d is large. **(R3: scores on Olivetti and MovieLens)** Our goal for these two datasets was to highlight qualitatively the COOT's ability to find meaningful solutions to a quantization problem. A quantitative study of COOT w.r.t. other co-clustering baselines is given on simulated datasets with known ground truth. **(R4: 20 samples per class in HDA)** This is the classical setting for this experiment. It was introduced in Yan et al. "Learning Discriminative Correlation Subspace for HDA", IJCAI'17 and used in [16]. We will add this citation.

——- Timings & Computational complexities ——-

We will detail both time and memory complexities of COOT in the final version.



(R1&R2&R4: runtime details) For an 1e-20 precision, the obtained runtime characteristics for the MNIST/USPS example and co-clustering experiments are given in figure on the left and table below. As one can see, the number of iterations for the BCD do not generally exceed 20. This means that the complexity of COOT mostly depends on the complexity of the used OT solver. Also, for HDA the timing of COOT is comparable to the one of SGW ($\sim 10s$), but superior to the one of KCCA ($\sim 0.1s$) to solve for one pair. We will include a more general study on simulated data with different values of n and d, as suggested by R1, in the Supplementary material.

Data set	Characteristics			
	Runtime(s)	BCD #iter. (COOT+X_c)	BCD #iter. (COOT)	COOT value
D1	4.72±6	21.5±24.57	3.16±0.37	0.46±0.25
D2	0.64±0.81	9.77±11.53	3.4±0.58	1.35±0.16
D3	0.95±1.55	8.47±11.11	3.01±0.1	2.52±0.24
D4	6.27±5.13	33.15±23.75	4.21±0.41	0.06±0.005

Table 1: Mean ($\pm$ standard-deviation) of different runtime characteristics of COOT.

(**R2: initialization's impact**) We conducted a study regarding the convergence properties of COOT in the co-clustering application when the π_s, π_v and X_c are initialized randomly over 100 trials. This leads to a certain variance in the obtained value of the COOT distance as expected when solving a non-convex problem. The obtained CCEs remain largely in line with the obtained results even for different random initializations. The quantitative results will be included in the paper, following the recommendation of R4. (**R4: scalability of COOT**) While our current implementation relies on solvers that compute couplings solutions to the primal OT problem (near linear time complexity for entropic regularization [23] but with a quadratic memory overhead), stochastic solvers working solely with dual variables could be used to efficiently deal with large datasets such as CelebA (eg, with neural networks as dual potentials). As suggested by R2 (thanks for the insights), warm starting the solvers inbetween BCD iterations can also accelerate our code and is an exciting avenue for scaling up COOT computation that we are currently working on.