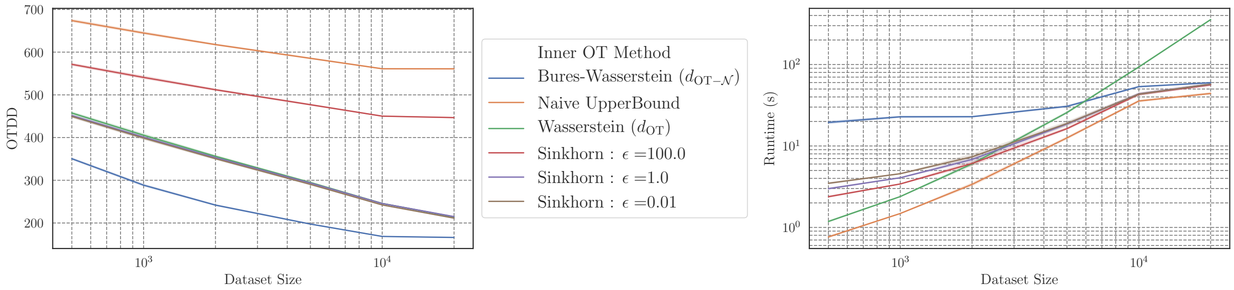

Figure 1: Comparison of methods to compute inner OT distance $d_{\mathcal{Y}}$ in OTDD between independent MNIST samples.

General Comments: We thank all reviewers for stupendous feedback, which has greatly improved the paper.

- **Gaussian Approximation** (R3/R5). Reviewers justifiably worry that this approximation could sometimes be too coarse. As R3 points out, in cases where the data is first embedded with some complex non-linear mapping (neural net or otherwise), there's empirical evidence that the first two moments capture enough relevant information for classification (Seddik et al, 2020). Note that this is the case for our text classification experiments. But we would argue that in general, despite the coarseness of this approximation, OTDD is objectively capturing relevant information (as shown by our experimental results), so any refinement over this approximation can only further improve the quality of the distance. For small datasets, computing the exact d_{OT} is feasible and—as predicted by Thm 5.1—sometimes even faster that $d_{\mathrm{OT}-\mathcal{N}}$, but for very large datasets ($n \gg d$) the latter will often be the only viable option (Fig 1).
- **Bounds** (R2/R5). We agree that the upper bound in Prop 4.1 is much less informative that the lower one (as is often the case in OT), but still, as R2 points out, having *any* upper bound allows one to e.g. bound the UB-LB band around the exact OTDD, as shown in Fig 1.
- **Baselines** (R2/R3). The lack of other distances that allow for datasets with non-identical label sets limits possible baselines except in simple settings. We will include those suggested by R2 plus some additional OTDD variants: (i) $d_{\mathcal{Z}} = d_{\mathcal{X}}$ (ignore labels), (ii) $d_{\mathcal{Z}} = d_{\mathcal{Y}}$ (ignore features), (iii) $d_{\mathcal{Y}} = \|\mu_y - \mu'_y\|$, (iv) $d_{\mathcal{Y}} = \mathcal{L}(y, y')$ (i.e., JDOT, possible if some pairs). Limited space precludes detailed results here, but we have now verified that these ablations weaken correlation with transferability on *NIST: e.g., ρ drops by 5-10% for (i)/(ii), and (i) loses significance (at $\alpha = .05$), though interestingly (ii) doesn't (c.f. R5's point). We will include these results in the final version.
- **Relation to JDOT/DeepJDOT**. We thank R3&R5 for raising this. It was an omission on our part not to discuss these very relevant works, but the revised version does. Summary: main difference with OTDD is that those rely on a classification loss to measure label similarity, which (i) requires same label sets across domains, (ii) depending on $\mathcal{L}$ might not yield a true metric. We will nevertheless compare against these where possible (i.e., MNIST↔USPS).
- **Scaling/weighting of $d_{\mathcal{X}}$ vs $d_{\mathcal{Y}}$** (R2/R5). Given the interpretation of $d_{\mathcal{Y}}(y, y')^p = \mathrm{W}_p^p(\alpha_y, \alpha_{y'})$ as an expectation of $d_{\mathcal{X}}^p$, the two terms in $d_{\mathcal{Z}}$ are by construction of the same order, and—we argue—should be equally weighted barring additional information. Yet, as mentioned by R2/R5, it might be interesting to weight these two terms differently in specific settings - we appreciate the suggestion and have included this option in our codebase.

Reviewer 1: Without assumptions on labels, cond. distrib is arguably the *only* info about y we have. We *do* use entropic OT for outer problem (cf. Thm 5.1) & can also be used for inner one, but Gaussian approx. is faster for large n (Fig 1).

Reviewer 2: We appreciate the many suggested baselines/ablations, some of which we have now implemented (Fig 1)

- Gromov / Hierarchical OT cannot—in their usual form—incorporate discrete and distinct label sets. We haven't observed much difference between various 'feature-only' metrics, but will include them nevertheless.
- The suggested ablation based on MMD is intriguing - we will try it and include in the revised version.
- $d_{\mathcal{X}}$ was always Euclidean here, but yes, it could be taken as OT distance! (additional cost/quality trade-off)

Reviewer 3: We hope the two points below clarify the doubts regarding experimental evaluation.

- "how is OTDD integrated into learning " → it is not. OTDD is computed directly on the datasets $\mathrm{D}_s, \mathrm{D}_t$. *Separately*, we pretrain net on D_s and fine-tune on D_t, w/ usual supervised training. Goal: show OTDD *predicts* transfer success.
- The bands in Figs 5-8 are 95% conf. intervals via bootstrap on 10 repetitions. "*wouldn't call them strongly correlated*": note that strength of correlation refers to the slope ρ. As for significance, a p-value of 0.02 might be considered 'mildly' significant, but the other three settings yield p-values $\in (10^{-5}, 10^{-2})$, significant by any standard.
- Complexity. Indeed, for $d \gg n$, d^3 dominates, which suggests (exact) d_{OT} is actually preferable (see Fig 1).

Reviewer 5: We are grateful for the detailed+acute observations/suggestions. We've adopted inner/outer terminology.

- As you suggest, following proof technique of JDOT paper we are able to show a similar adaptation bound - will add

[1] Seddik et al., *Random matrix theory Proves that Deep Learning representations of GAN-data behave as gaussian mixtures*, 2020.