We thank the reviewers for the detailed comments. We note that reviewers 2 and 3 think highly of the quality of the paper. If our responses have addressed your concerns, we hope that you would give an accept recommendation.

Reviewer 2

1. Presentation of gapBoost: Thank you for the suggestion. We will re-organize Section 3.

2. Nonlinear extension: our analysis can be extended to (nonlinear) kernel models based on a reproducing kernel Hilbert space. The theoretical analysis of general nonlinear models (e.g., deep nets) is challenging since their loss landscape is usually non-convex. However, motivated by the empirical success of convex optimization methods for fitting complex deep nets, we hypothesize that we could still leverage the intuition behind our gap minimization principle to create novel deep transfer methods. In future work, we plan to empirically verify this conjecture.

3. In Section 3.5 of Appendix, we have shown that the $\mathcal{Y}$-discrepancy can be bounded from training data by constructing a classification problem, which may be used as a guideline to select parameters in a principled way. We choose $\rho_{\mathcal{T}} = 0$ as it corresponds to no punishment for the target data (the simplest setting). We have run additional experiments by varying both parameters. In Fig. 1, we can observe that by properly choosing both parameters (e.g., $\rho_{\mathcal{T}} = \log 2$, $\rho_{\mathcal{S}} = 0$), we may obtain even better results. As you point out, we could use a simple heuristic like choosing a relatively larger $\rho_{\mathcal{S}}$ when target data is small in order to leverage source data, as shown in Fig. 1(a). As the target data increase, the results are less sensitive to the parameter. As long as $\rho_{\mathcal{T}} > \rho_{\mathcal{S}}$, the performance of gapBoost is stable over a wide range of values of parameters, as shown in Fig. 1(b)–1(d). In Fig. 1 in the paper, we fixed $\rho_{\mathcal{T}} = 0$ and $\rho_{\mathcal{S}} = \log \frac{1}{2}$. This will be made more explicit in the revised version.

4. There are various measures for unlabeled data proposed in the literature (see the references in Line 57), which could be incorporated into our work. The notion of discrepancy [25] (the unsupervised version of $\mathcal{Y}$-discrepancy) is particularly relevant, due to its consistency with the notion used in our paper. We will also be working on generalizing the notion of gap to the unsupervised learning (domain adaptation) setting.

Reviewer 3

Thank you for your comments and pointing out the reference. We will add a qualitative comparison in our paper. Please note that the current baselines methods are all boosting-based approaches in order to make a fair comparison.

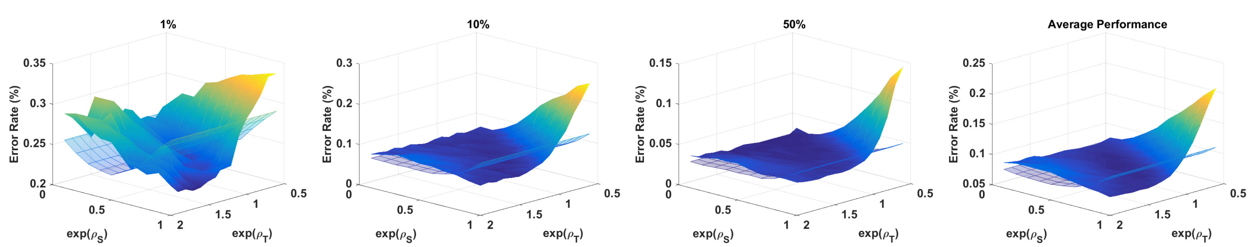
Reviewer 4

1. Vacuous bound: The inequality $||\Gamma||_2 \leq \sqrt{N}||\Gamma||_\infty$ is tight when we assign equal weights to all data points. **Since Γ is a probability simplex, we have $||\Gamma||_2 = \frac{1}{\sqrt{N}}$ and $||\Gamma||_\infty = \frac{1}{N}$. Then, after simplifying the multiplicative term $\sqrt{N}$, ε_Γ has a fast convergence rate of $\mathcal{O}(\frac{1}{\sqrt{N}})$ in this case,** which motivates Rule 2. In fact, we recover the learning bound of assigning equal weights on source and target instances [3] (i.e., pooling-task approach). See also Remark 3 for more discussions.

2. Moving parts: Thank you for noting that the trade-off between the multiple terms is intuitively reasonable, which motivates the proposed rules.

3. Line 182: As you correctly point out, the bound is controlled by the discrepancy—it is also shown in the last term of (2), which indeed motivates Rule 3. The convergence rate is in fact the convergence rate of ε_Γ. We will clarify this point in the revised version.

4. Tools are straightforward ... largely inspired by [20]: While the tools are commonly used, we extend the existing theoretical results in the following ways. First, we propose the novel notion of *performance gap*, revealing a new principle for transfer learning. Second, we extend existing tools to their "weighted" version (e.g., weighted Rademacher complexity/uniform stability/Hoeffding's inequality, see Appendix for details). Third, we develop the bounds for $\mathcal{Y}$-discrepancy in the supervised learning context (the notion of discrepancy in [3], [25] is designed in the unsupervised learning context). We also show that for 0-1 loss, the empirical $\mathcal{Y}$-discrepancy can be computed by constructing a new classification problem. See Section 3.5 of Appendix for more details. Finally, we only use [20] to derive the Rademacher bound after we have obtained the stability bound, and we extend it to our weighting setting.



(a) (b) (c) (d)

Figure 1: Test error rates (%) with varying $\rho_{\mathcal{S}}$ and $\rho_{\mathcal{T}}$. The valley curves are obtained by setting $\rho_{\mathcal{T}} = 0$ (i.e., the purple curves in Fig. 2 of main paper). Hence the areas below the curve indicate better parameter configurations.