Response for “Generalized Block-Diagonal Structure Pursuit: Learning Soft Latent Task Assignment against Negative Transfer ”
ID 3136

We thank all the reviewers for their valuable comments. We have fixed the typos pointed out by the reviewers. Below are the responses to each reviewer.

To Reviewer 1

Q1: **Is the framework limited only to linear models?** Our framework could be applied to more general model families in view of the following two perspectives. By the formulation $\boldsymbol{W} = \boldsymbol{L}\boldsymbol{S}$, our framework provides a latent representation of the task weights $\boldsymbol{W}$. It could be naturally extended to more complicated models such as DNNs via modifying the latent expression $\boldsymbol{L}$. From the generative point-of-view, our framework could also be extended to cover more complicated distribution families via changing $g_1(\cdot)$,$g_2(\cdot)$ and $g_3(\cdot)$.

Q2: **The convergence property and the convergence rate of the algorithm? What if only a local solution is found? 1)** Under mild assumptions, we can say that the proposed algorithm enjoys the global convergence property defined for non-convex problems, where *both the objective sequence and the parameter sequence converge to a critical point*. Moreover the convergence rate should be $\mathsf{O}(\frac{1}{T})$, which is *sublinear*. **2)** For non-convex problems, global convergence to a critical point is the best we can do in a general sense. The global minimum could be found only when the initial point is located in a local convex landscape covering the optimal solution. Nonetheless, according to Thm.3, the generalization ability will be promising if the loss is small (not necessarily only the optimal value) and the hypothesis space is properly chosen. In this sense, a local critical point would be a good candidate solution. This has also been suggested by our experimental results.

Q3: **Are the constraints in the Obj included in the class $\mathcal{H}(\boldsymbol{L}, \boldsymbol{S}, \widetilde{\boldsymbol{S}}, \boldsymbol{U})$?** The constraints are included in the hypothesis space. We will include them in the new version.

Q4: **The proof of Thm.3?** The proof follows naturally from Lem.3 and Thm. 13 in Ref.[23] mentioned in the main paper. There are only two key differences. We assume that $l(y, \cdot) : \hat{y} \mapsto [0, M]$ instead of $[0, 1]$. That is why we use $\frac{\Delta}{M}$ in our case. Moreover, we assume that $\breve{\ell}(\cdot) = l(y, \cdot)/M$ is ϕ-Lipschitz continuous, instead of 1-Lipschitz continuous. That is why the first two terms (which gives an upper bound for the Gaussian Average of $\breve{\ell} \circ \Big(\mathcal{H}(\boldsymbol{L}, \boldsymbol{S}, \widetilde{\boldsymbol{S}}, \boldsymbol{U})\Big)$) on the right hand side are multiplied by ϕ. We will provide a detailed proof in the new version.

Q5: **ξ_3 does not appear in Thm.3, how does it benefit the hypothesis space?** ξ_3 benefits the hypothesis space in the following sense. According to Thm.4, we see that decreasing ξ_3 reduces the sensitivity of numerical perturbation on the principle components of $\boldsymbol{S}$. More importantly, Thm.5 shows that shrinking ξ_3 results in a better recovery of the desired structure and helps to overcome the negative transfer issue. From the generalization perspective, the expected structure is sparse in most cases, this leads to an implicit control of the $\mathcal{VC}$ dimension of the space, which suggests that ξ_3 also contributes to the generalization ability.

To Reviewer 2

Q1: **For $\widetilde{S}$ subroutine, why adopt the dual problem? More discussion on the barycenter projection mapping.** 1) The dual problem could be solved much more efficiently than the primal. The dual problem only contains $\mathsf{O}(k + T)$ parameters while the primal requires $\mathsf{O}(kT)$ parameters. 2) Since $\widetilde{\boldsymbol{S}}$ is the solution of the regularized OT problem, we have $\widetilde{\boldsymbol{S}}_{ij} = \mathbb{P}(\boldsymbol{l} = i, \boldsymbol{o} = j)$. Then $\mathbb{E}_{\boldsymbol{l}|\boldsymbol{o}=i}(\boldsymbol{L}) = \frac{\boldsymbol{L}\widetilde{\boldsymbol{S}}^{(i)}}{T}$. In this sense , $\frac{\boldsymbol{L}\widetilde{\boldsymbol{S}}^{(i)}}{T}$ represents the output task as a barycenter in the latent task embedding space, since $\frac{\boldsymbol{L}\widetilde{\boldsymbol{S}}^{(i)}}{T} = \mathsf{argmin}_{\boldsymbol{z}}\, \mathbb{E}_{\boldsymbol{l}|\boldsymbol{o}=i}(d(\boldsymbol{L}^{(i)}, \boldsymbol{z}))$.

Q2: **The phrasing in Thm.2.** We will correct the phrasing issue as you suggested.

To Reviewer 3

Q1: **Discussion on Theorem 3.** Thm.3 states that if the magnitude of $\boldsymbol{COV}(\boldsymbol{X})$ is small, the difference between the population version of the task-averaged risk and the empirical risk Δ tends to zero asymptotically when $n \to +\infty$. Here $\boldsymbol{COV}(\boldsymbol{X})$ captures the correlation of the data points of all the samples. Please see the Reference [23] in the main paper for more details.

Q2: **Error bar in Fig.2.** The error bar in Fig.2 represents the standard deviation over all repetitions.

Q3: **Real-world dataset should be mentioned at the main paper.** We will add a brief introduction of the real-world dataset in a new version.