It is a great honor for us to take your kind and meaningful feedback. We write some comments below.

**Major Comments.** We note our thinking about our experiments. Several reviewers pointed out that our experiments are insufficient. It is certainly interesting to see other comparative experiments. In this study, however, the main purpose is to propose a "simple" and "general" methodology with a link to transfer learning and concept drift. For example, the Group Lasso, Fused Lasso, Clustered Lasso, and so on, are in the same line. They have some concrete target problems, but these ideas can be easily applied to various problems. They were not the SOTA when they were proposed. This is natural because they offered general methodologies. At present, they are quite popular, because they are simple and easily applied to various problems. Our proposed method is motivated by transfer learning and concept drift, and the idea can be applied to various problems.

For the above reason, we restricted our experiments as follows. i) We focus on a high-dimensional regression problem and restrict our attention to a sparse regression model. Although this setting is important in many areas, it does not seem the mainstream of transfer learning or concept drift, so there exists limited related work. ii) We further excluded methods of transferring latent spaces, internal representations, or dictionaries in our experiments. These methods assume that source and target domains share some low-dimensional representations. This assumption is quite different from our assumption that source and target domains share model parameters (regression coefficients). We would emphasize that our method is superior to these methods in terms of interpretability and operability for continuously updating applications.

**Reviewer #1.** About weaknesses 2-4 and 6: Please see above. 5: The objective of Lasso screening such as safe screening is not accuracy but speed. These kinds of techniques can be applied to our method, but this is out of scope for this paper. 7: The data were divided into 30 batches without changing the order of the samples. 8: In the case of binary features, there are likely to exist variable pairs with very high correlations (or exactly the same variables). For $\alpha = 1/2$, the contour lines tend to be parallel to $b_j + b_k = const$ and the loss function remains equal under $b_j + b_k = const$ for $X_j \approx X_k$, making the global solution indefinite or unstable. 9: Each line is colored randomly for legibility and customary reasons (same as the `glmnet` package in R).

**Reviewer #2.** We use the ordinary coordinate descent algorithm. The formulation is convex and hence the solution does not depend on the initial value. The order of the variables to be optimized is arbitrary (can be even random). In our implementation, we used a warm-start technique for initialization and applied cyclic coordinate descent, but they do not affect results. Please check standard textbooks such as "The Element of Statistical Learning" (Section 3.8.4).

**Reviewer #3.** Thank you for the good point. It is not clear when the GRE condition will hold in general, but for $2\alpha - c - 1 > 0$, Theorem 1 and Corollary 1 in (Raskutti et al., 2010) imply that the GRE condition holds with high probability when $X_i$ is sampled from Gaussian distribution and its population covariance satisfies the GRE condition. Indeed, it can be easily shown by $\mathcal{B} \subset \{v : (2\alpha - c - 1)\|v_{S^c}\|_1 \leq (1 + c)\|v_S\|_1\}$.
About weakness 1: Theorem 2 supposes that $p$, $n$, and $s$ go to infinity and shows that $n > O(s \log(p))$ is required for convergence. 4: The search space of $\lambda$ for the ordinary Lasso is determined by $\lambda_{\max}$, the smallest value for which all coefficients are zero, and the smallest value for $\lambda$ as a fraction of $\lambda_{\max}$. In our method, we can determine $\lambda_{\max}$ as the smallest value for which all coefficients are zero or initial estimates using Theorem 4.

**Reviewer #4.** The reason why the source/target parameters should be close in $\ell_1$ norm is that we impose the assumption that the difference between source/target parameters is sparse. This sparsity assumption is reasonable under a high-dimensional online setting and informative parameters are not so many in a real-world environment, as is the Lasso assumption. We used the $\ell_1$ norm because it is a unique $\ell_q$ norm satisfying both sparsity and convexity.
We have assessed the influence of sample size, but we omitted the results due to space limitations. Figure 1 shows that Transfer Lasso was more effective than others in "4.2 Transfer Learning Simulation" under various sample sizes.

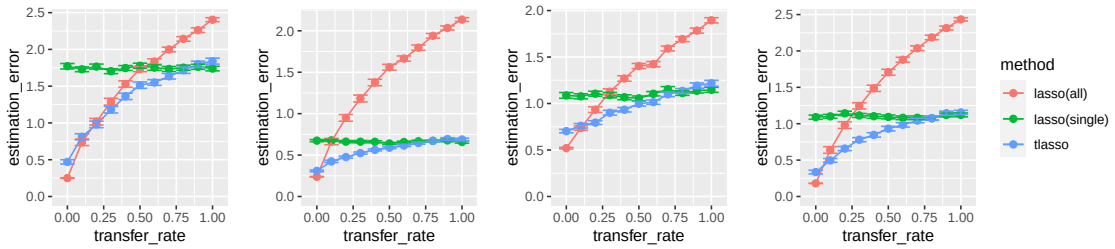


Figure 1: Estimation errors of transfer learning simulations (Section 4.2) under various source and target sample sizes $((n_s, n_t) = (500, 20), (500, 100), (100, 50)$, and $(1000, 50)$ from left to right).