

Why Transfer Learning Fails Under Target Non-Identifiability

Yuki Kobayashi

Kyoto University

ABSTRACT

Transfer learning (TL) improves model performance in a target domain with limited data by leveraging data from a source domain. When the source–target discrepancy is large, TL can degrade target-domain performance, a phenomenon known as negative transfer (NT). In linear regression, the coefficient vector is only partially identifiable when the target design matrix is rank-deficient. Although TL can exploit source information to address such non-identifiability, it may also amplify coefficient estimation error. However, the mechanisms and conditions underlying this type of NT have not been fully characterized. Frustratingly easy domain adaptation (FEDA) is a TL method that has been successfully applied in the process industry. This study derives the mechanism and conditions of NT in FEDA with linear regression. Because the derived NT condition involves the unob-

servable true target coefficient, we further construct a proxy condition that can be evaluated from observed data by assuming an upper bound on the source–target coefficient discrepancy. Synthetic experiments confirm the mechanism of NT and examine the effects of coefficient-shift sparsity, coefficient-shift magnitude, and target sample size. This study contributes to a theoretical understanding of the NT mechanism under target non-identifiability in FEDA with linear regression.

© 2026 by the authors. Licensed to PSEcommunity.org and PSE Press. This is an open access article under the creative commons CC-BY-SA licensing terms. Credit must be given to creator and adaptations must be shared under the same terms. See <https://creativecommons.org/licenses/by-sa/4.0/>

