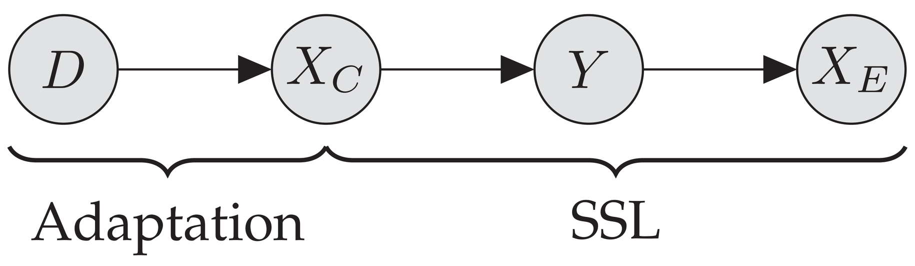
Semi-Generative Modelling: Covariate-Shift Adaptation with Cause and Effect Features

Julius von Kügelgen, Alexander Mey, Marco Loog

TU Delft
University of Cambridge

Objective

Improving an adapted model with unlabelled data when labelled data is scarce, that is, combining covariate-shift (CS) adaptation and semi-supervised learning (SSL):



Background

- CS assumption: $P(X)$ changes, but $P(Y|X)$ remains domain invariant.
- Current CS approaches use unlabelled data for importance reweighting [3, 5] or learning domain-invariant features [1].
- Classifier is trained on labelled data only.
- When amount of labelled data is the bottleneck, also use unlabelled data for SSL.
- Combining CS and SSL requires learning with both cause and effect features [2, 4].

Problem Setting

Given:

- small labelled source-domain ($D = 0$) sample, $(x_C^i, y^i, x_E^i) \sim P(X_C, Y, X_E|D = 0)$
- large unlabelled target-domain ($D = 1$) sample, $(x_C^j, x_E^j) \sim P(X_C, X_E|D = 1)$

Goal:

- minimise expected target-domain loss, $\mathbb{E}_{P(X_C,Y,X_E|D=1)}[L(\hat{Y}(X_C, X_E), Y)]$

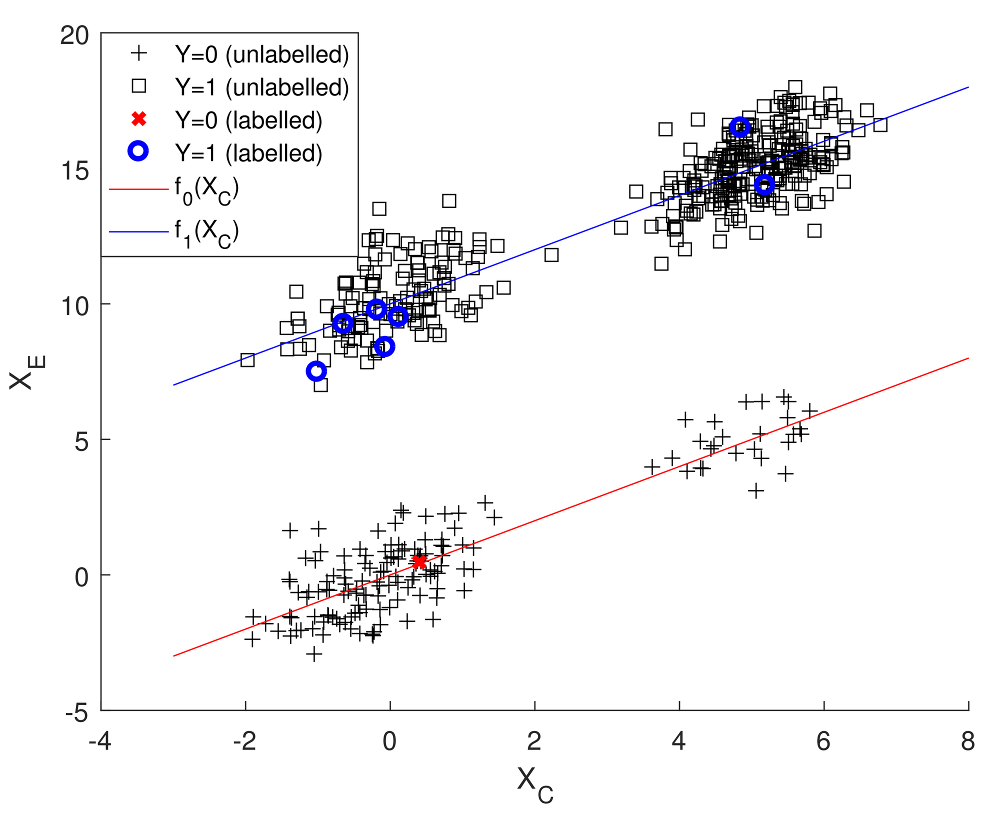
Assumption:

- underlying structural causal model is known to be of the form (see Figure 1):

$$X_C := f_C(D, N_C)$$
$$Y := f_Y(X_C, N_Y)$$
$$X_E := f_E(Y, N_E)$$

Future Work

- Relax assumptions to the more general setting by allowing $X_C \to X_E$.
- Incorporate common assumptions such as clustering or low density separation



Causal and Anticausal Learning [2]

For many learning settings it matters whether a feature X is a cause or an effect of a target Y!

1. Causal Learning: $X \to Y$

- $P(X)$ and $P(Y|X)$ are independent, i.e. share no information
- Covariate shift holds: changes in $P(X)$ should have no effect on $P(Y|X)$
- SSL impossible: $P(X)$ does not contain information about $P(Y|X)$

2. Anticausal Learning: $Y \to X$

- $P(Y)$ and $P(X|Y)$ are independent, but $P(X)$ and $P(Y|X)$ are *dependent*
- Covariate shift does not hold: changes in $P(X)$ can have an effect on $P(Y|X)$
- SSL possible: $P(X)$ contains information about $P(Y|X)$

Semi-Generative Modelling Approach

Main Idea: Condition on causal features, but explicitly model the distribution of effect features.

Discriminative Model	**Semi-Generative Model**	**Generative Model**
$P(Y\|X_C, X_E, \theta)$	$P(Y, X_E\|X_C, \theta)$	$P(X_C, Y, X_E\|D, \theta)$
domain-invariant	domain-invariant	not domain-invariant
cannot use unlabelled data (x_C, x_E) for SSL	can use unlabelled data (x_C, x_E) for SSL	can use unlabelled data (x_C, x_E) for SSL

Supervised source-domain log-likelihood:

$$\ell_S(\theta) = \frac{1}{n_S} \sum_{i=1}^{n_S} \left(\log P(y^i|x_C^i, \theta) + \log P(x_E^i|y^i, \theta) \right)$$

Unsupervised target-domain log-likelihood:

$$\ell_T(\theta) = \frac{1}{n_T} \sum_{j=n_S+1}^{n_S+n_T} \log \left(\sum_{y \in \mathcal{Y}} P(y|x_C^j, \theta) P(x_E^j|y, \theta) \right)$$

Interpolated pooled log-likelihood with $\lambda \in (0, 1)$:

$$\ell_P^\lambda(\theta) = \lambda\, \ell_S(\theta) + (1 - \lambda)\, \ell_T(\theta)$$

Factorisation of semi-generative model:

$$P(Y, X_E|X_C, \theta) = P(Y|X_C, \theta) P(X_E|Y, \theta)$$

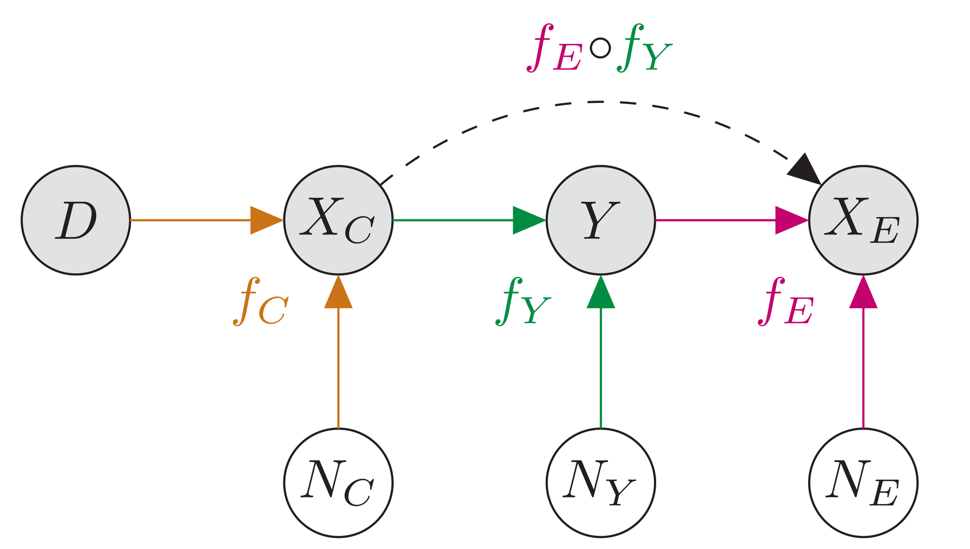


Figure 1: SSL by learning a noisy composition of f_Y and f_E from unlabelled data (x_C, x_E).

Results on Synthetic Classification Data

- $\theta_S = \arg\max_\theta \ell_S(\theta)$, supervised baseline
- $\theta_P^\lambda = \arg\max_\theta \ell_P^\lambda(\theta)$, pooled estimator
- θ_{WS}, importance-weighted form of θ_S [3]
- θ_{LR}, logistic-regression on (X_C, X_E)

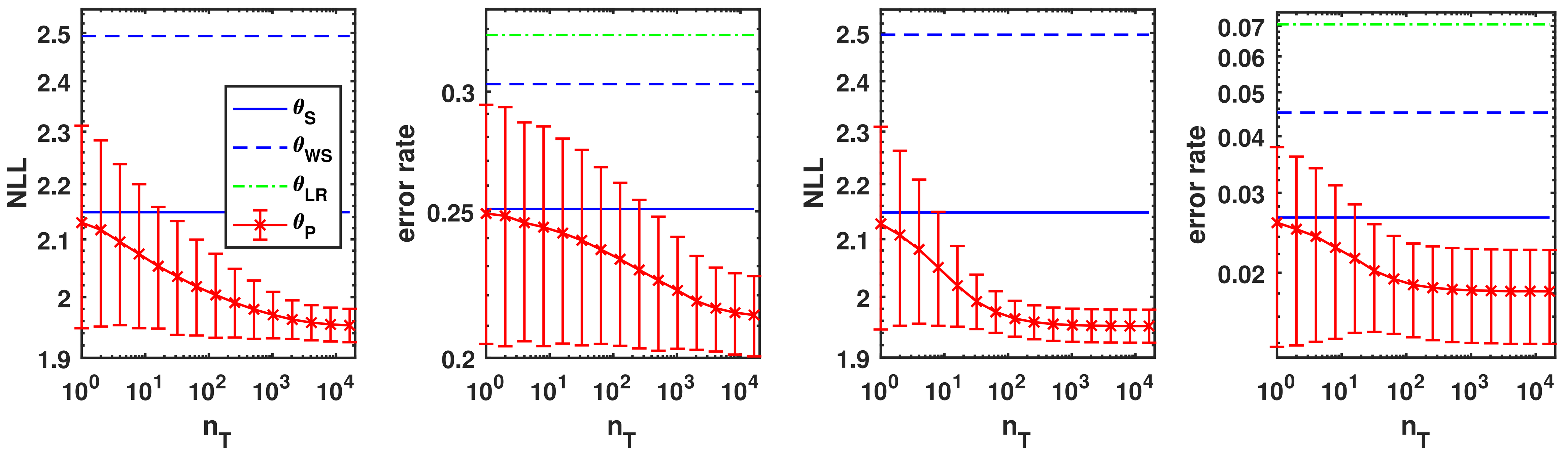


Figure 2: Results for $n_S = 8$ labelled examples using $\lambda = \frac{n_S}{n_S+n_T}$ for $\mu = 0.5$ (left), and $\mu = 2$ (right), where μ determines the amount of information X_E carries about Y: $X_E|(Y = \pm 1) \sim \mathcal{N}(\pm\mu, 1)$.

References

[1] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[2] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[3] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[4] A. Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 2009.

[5] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007.