Non-Negative Bregman Divergence Minimization for Deep Direct Density Ratio Estimation

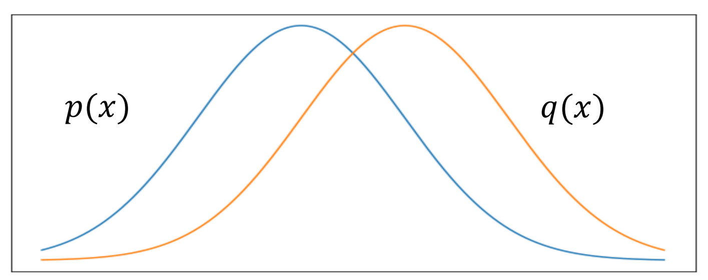
Masahiro Kato, Cyberagent, Inc.
Takeshi Teshima, The University of Tokyo

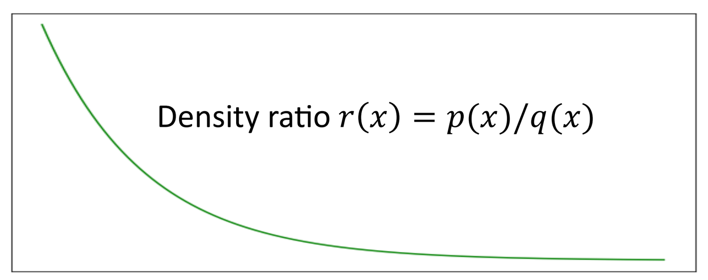
CyberAgent AI Lab

Abstract

Density Ratio:

- Density Ratio
$$r^*(x) = \frac{p(x)}{q(x)}$$
- The ratio of the two probability densities $p(x)$ and $q(x)$.
- The density ratio appears in many tasks in machine learning.
  - Anomaly detection
  - Domain adaptation etc.





Goal:

- Density ratio estimation with deep neural networks.

Issue:

- Train loss often diverges when we use neural networks.

Contributions:

- We detect the cause of this problem.
- We propose an empirical risk correction to mitigate this problem.
- Proposed method performs well in anomaly detection.

1. Density Ratio Estimation (DRE)

How to estimate the density ratio?:

- Samples from two datasets:
$$\{x_j^{\mathrm{nu}}\}_{j=1}^{n^{\mathrm{nu}}} \sim p(x) \text{ and } \{x_i^{\mathrm{de}}\}_{i=1}^{n^{\mathrm{de}}} \sim q(x)$$
- A naive method is to estimate the probability densities separately.
- Then, we construct an estimator as their fraction: $\hat{r}(x) = \frac{\hat{p}(x)}{\hat{q}(x)}$.
- However, estimating the probability densities is not easy.

→ Various methods for **direct DRE** have been proposed.

- Ex. Hastie et al., (2001), Gretton et al., (2009), etc.
- Sugiyama et al. (2011) unified them from the Bregman divergence (BD) minimization perspective.

Objective function of direct DRE with BD minimization:

$$\widehat{\mathrm{BD}}_f(r) := \hat{\mathbb{E}}_{\mathrm{de}}\left[\partial f(r(X_i))r(X_i) - f(r(X_i))\right] - \hat{\mathbb{E}}_{\mathrm{nu}}\left[\partial f\left(r(X_j)\right)\right],$$

- $\hat{\mathbb{E}}_{\mathrm{de}}\ (\hat{\mathbb{E}}_{\mathrm{nu}})$: sample averages over $\{x_i^{\mathrm{de}}\}_{i=1}^{n^{\mathrm{de}}} \sim q(x)\ \left(\{x_j^{\mathrm{nu}}\}_{j=1}^{n^{\mathrm{de}}} \sim p(x)\right)$.
- $f(t)$ is a twice continuously differentiable convex function.

Table 1. Summary of DRE methods (Sugiyama et al., 2011b). For PULogLoss, we use $C < \frac{1}{\overline{R}}$.

Method	$f(t)$	Lower bound of $\widehat{\mathrm{BD}}_f$	Reference
LSIF	$(t-1)^2/2$	Not bounded	Kanamori et al. (2009)
Kernel Mean Matching	$(t-1)^2/2$	Not bounded	Gretton et al. (2009)
UKL	$t\log(t) - t$	Not bounded	Nguyen et al. (2010)
KLIEP	$t\log(t) - t$	Not bounded	Sugiyama et al. (2008)
BKL (LR)	$t\log(t) - (1+t)\log(1+t)$	Bounded	Hastie et al. (2001)
PULogLoss	$C\log(1-t) + Ct(\log(t) - \log(1-t))$ for $0 < t < 1$	Not bounded	Kato et al. (2019)

- Existing studies mainly estimate r^* with linear models.

↔ Recently, neural networks are shown to be effective in many tasks.

3. Train Loss Hacking

Empirical BD minimization with neural networks:

→ The train loss often goes to $-\infty$ due to $-\hat{\mathbb{E}}_{\mathrm{nu}}\left[\partial f\left(r(X_j)\right)\right]$.

$$\min_{r \in \mathcal{H}} \hat{\mathbb{E}}_{\mathrm{de}}\left[\partial f(r(X_i))r(X_i) - f(r(X_i))\right] - \hat{\mathbb{E}}_{\mathrm{nu}}\left[\partial f\left(r(X_j)\right)\right].$$

$$\longrightarrow -\infty$$

- We call this phenomenon <u>train loss hacking</u>.

The causes of this problem are

- (i) <u>a too flexible model</u> and (ii) <u>finite samples</u>.

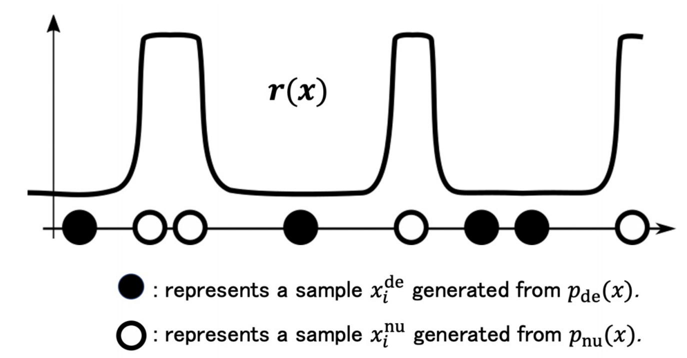
→ <u>The flexible model can overfit and leads the train loss to $-\infty$.</u>

- $\hat{\mathbb{E}}_{\mathrm{de}}\left[\partial f(r(X_i))r(X_i) - f(r(X_i))\right]$

→ Keep around 0.

- $\hat{\mathbb{E}}_{\mathrm{nu}}\left[\partial f\left(r(X_j)\right)\right]$.

→ Goes to $-\infty$.



4. Upper Bound of the Density Ratio

How to prevent train loss hacking?:

- $-\hat{\mathbb{E}}_{\mathrm{nu}}\left[\partial f\left(r(X_j)\right)\right] \to -\infty$ means $r(X_j) \to \infty$.

→ For x_j^{nu}, we want to prevent $r(x_j^{\mathrm{nu}}) \to \infty$.

The Role of the upper bound of the density ratio:

- Suppose there exists a constant $\overline{R} > 0$ such that $\forall\, x\ r^*(x) < \overline{R}$
- Use a model satisfying $r(x) < \overline{R}$? Ex. $r(x) = \frac{\overline{R}}{1+\exp(-f(x))}$.
- Even when $r(X_j)$ is a bounded function,
  - $\partial f\left(r(X_j)\right)$ sticks to the upper bound because it is monotonically increasing function
- Let $C > 0$ be a constant such that $C > 1/\overline{R}$
- Let us decompose the empirical BD as
$$\hat{\mathbb{E}}_{\mathrm{de}}\left[\partial f(r(X_i))r(X_i) - f(r(X_i))\right] - \hat{\mathbb{E}}_{\mathrm{nu}}\left[\partial f\left(r(X_j)\right)\right]$$
$$= \left(\hat{\mathbb{E}}_{\mathrm{de}}[\ell_1(r(X_i))] - C\hat{\mathbb{E}}_{\mathrm{nu}}[\ell_1(r(X_i))]\right) + \hat{\mathbb{E}}_{\mathrm{nu}}\left[\ell_2\left(r(X_j)\right)\right].$$
  - $\ell_1(t)$ and $\ell_2(t)$ are components of empirical BD.
- If $r^*(x) < \overline{R}$,
$$\mathbb{E}_{\mathrm{de}}[\ell_1(r(X_i))] - C\mathbb{E}_{\mathrm{nu}}[\ell_1(r(X_i))]$$

becomes positive because

$$q(x) - \frac{p(x)}{\overline{R}} = q(x)\left(1 - \frac{r^*(x)}{\overline{R}}\right) > 0\ \forall\, x$$

holds from $r^*(x) < \overline{R}$, $\ell_1(t) > 0$, and

$$\mathbb{E}_{\mathrm{de}}[\ell_1(r(X_i))] - C\mathbb{E}_{\mathrm{nu}}[\ell_1(r(X_i))] = \int \ell_1(r(X_i))\left(q(x) - \frac{p(x)}{\overline{R}}\right)\mathrm{d}x > 0.$$

5. Non-negative BD Minimization

Nonnegative BD (nnBD):

- We find the relationship between empirical BD and $\overline{R}$.

→ Based on this finding, we propose the nonnegative correction:

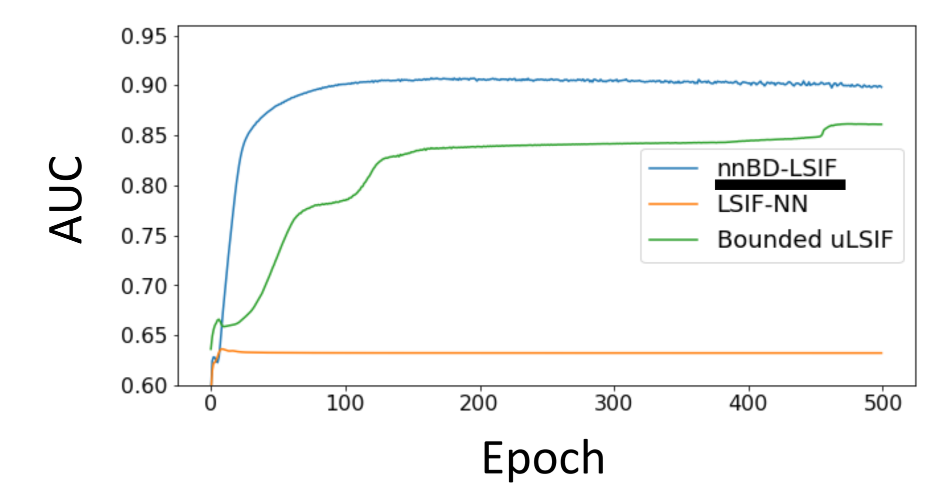
$$\widehat{\mathrm{nnBD}}_f(r) = \left(\hat{\mathbb{E}}_{\mathrm{de}}[\ell_1(r(X_i))] - C\hat{\mathbb{E}}_{\mathrm{nu}}[\ell_1(r(X_j))]\right)_+ + \hat{\mathbb{E}}_{\mathrm{nu}}\left[\ell_2\left(r(X_j)\right)\right].$$

- $\ell_1(r(X))$ and $\ell_2(r(X))$ are components of empirical BD.
- $\overline{R}$: <u>The upper bound of the density ratio.</u>
- C is a constant such that $C > 1/\overline{R}$.

➢ **We call the corrected empirical BD nonnegative BD (nnBD).**

- Direct DRE based on nnBD minimization: **deep direct DRE (D3RE).**

➢ <u>D3RE significantly mitigates the train loss hacking problem.</u>



6. Experiments

Inlier-based outlier detection

- One of the settings of semi-supervised anomaly detection.
- Compute the AUROC for CIFAR-10 and FMNIST datasets.

CIFAR-10 Network	uLSIF-NN LeNet		nnBD-LSIF LeNet		nnBD-PU LeNet		nnBD-LSIF WRN		nnBD-PU WRN		Deep SAD LeNet		GT WRN	
Inlier Class	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
plane	0.745	0.056	0.934	0.002	**0.943**	0.001	0.925	0.004	0.923	0.001	0.627	0.066	0.697	0.009
car	0.758	0.078	0.957	0.002	**0.968**	0.001	0.965	0.002	0.960	0.001	0.606	0.018	0.962	0.003
bird	0.768	0.012	0.850	0.007	**0.878**	0.004	0.844	0.004	0.858	0.004	0.404	0.006	0.752	0.002
cat	0.745	0.037	0.820	0.003	**0.856**	0.002	0.810	0.009	0.841	0.002	0.517	0.018	0.727	0.014
deer	0.758	0.036	0.886	0.004	**0.909**	0.002	0.864	0.008	0.872	0.002	0.704	0.052	0.863	0.014
dog	0.728	0.103	0.875	0.004	**0.906**	0.002	0.887	0.005	0.896	0.002	0.490	0.025	0.873	0.002
frog	0.750	0.060	0.944	0.003	**0.958**	0.001	0.948	0.004	0.948	0.001	0.744	0.014	0.879	0.008
horse	0.782	0.048	0.928	0.003	0.948	0.002	0.921	0.007	0.927	0.002	0.519	0.015	**0.953**	0.001
ship	0.780	0.048	0.958	0.003	**0.965**	0.001	0.964	0.002	0.957	0.001	0.430	0.062	0.921	0.009
truck	0.708	0.081	0.939	0.003	**0.955**	0.001	0.952	0.003	0.949	0.001	0.393	0.008	0.911	0.003

FMNIST Network	uLSIF-NN LeNet		nnBD-LSIF LeNet		nnBD-PU LeNet		nnBD-LSIF WRN		nnBD-PU WRN		Deep SAD LeNet		GT WRN	
Inlier Class	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
T-shirt/top	0.960	0.005	0.981	0.001	**0.985**	0.000	0.984	0.001	0.982	0.000	0.558	0.031	0.890	0.007
Trouser	0.961	0.010	0.998	0.000	**1.000**	0.000	0.998	0.000	0.998	0.000	0.758	0.022	0.974	0.004
Pullover	0.944	0.012	0.976	0.001	0.980	0.001	**0.983**	0.002	0.972	0.001	0.617	0.046	0.902	0.005
Dress	0.973	0.006	0.986	0.001	**0.992**	0.000	0.991	0.001	0.986	0.000	0.525	0.038	0.843	0.014
Coat	0.958	0.006	0.978	0.001	**0.983**	0.000	0.981	0.002	0.974	0.000	0.627	0.029	0.885	0.003
Sandal	0.968	0.011	0.997	0.001	**0.999**	0.000	**0.999**	0.000	**0.999**	0.000	0.681	0.023	0.949	0.005
Shirt	0.919	0.005	0.952	0.001	**0.958**	0.001	0.944	0.005	0.932	0.001	0.618	0.015	0.842	0.004
Sneaker	0.991	0.001	0.994	0.002	**0.998**	0.000	**0.998**	0.000	**0.998**	0.000	0.802	0.054	0.954	0.006
Bag	0.980	0.005	0.994	0.001	**0.999**	0.000	0.998	0.000	**0.999**	0.000	0.447	0.034	0.973	0.006
Ankle boot	0.992	0.001	0.985	0.015	**0.999**	0.000	0.997	0.000	0.996	0.000	0.583	0.023	0.996	0.000

References

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. Covariate shift by kernel mean matching. Dataset Shift in Machine Learning, 131-160 (2009), 01 2009.

Hastie, T., Tibshirani, R., and Friedman, J. The elements of statistical learning: data mining, inference and prediction. Springer, 2001.

Kanamori, T., Hido, S., and Sugiyama, M. A least-squares approach to direct importance estimation. Journal of Machine Learning Research, 10(Jul.):1391–1445, 2009.

Kiryo, R., Niu, G., du Plessis, M. C., and Sugiyama, M. Positive-unlabeled learning with non-negative risk estimator. In NeurIPS, 2017.

Sugiyama, M., Suzuki, T., and Kanamori, T. Density ratio matching under the bregman divergence: A unified frame-work of density ratio estimation. Annals of the Institute of Statistical Mathematics, 64, 10 2011b