Sugiyama and many others [1, 2, 3] describe covariate shift to be fully specified by the density ratio $w(x) = dQX/dPX(x)$ that they refer to as *importance weighting*. We propose to estimate this quantity from data. The estimation is formalized by the f-divergence, given P and Q be two probability distributions over a space Ω such that P is continuous with respect to Q. For a convex function f such that $f(1) = 0$, the f-divergence of P from Q is defined as $D_f(P||Q) = E_Q f(dP/dQ)]$, with dP/dQ interpreted as the Radon-Nikodym derivative [].

The usual assumption is for f to be closed and convex, with $f(1) = 0$ and $f(x) < +\infty \quad \text{for} \quad x > 0$. The relevant choice of f include $f(x) = xlogx$, which yields the Kullback–Leibler (KL) divergence. This provides a metric on the space of probability distributions in order to measure the amount of shift between the distribution of different batches. Among other f measures, we prefer the KL-divergence (or D_{KL}) because of its theoretical proximity with the cross-entropy loss of the archetypal classification network.

References

[1] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

[2] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *Algorithmic Learning Theory: 19th International Conference, ALT 2008, Budapest, Hungary, October 13-16, 2008. Proceedings 19*, pages 38–53. Springer, 2008.

[3] Tongtong Fang, Nan Lu, Gang Niu, and Masashi Sugiyama. Rethinking importance weighting for deep learning under distribution shift. *Advances in neural information processing systems*, 33:11996–12007, 2020.