# Density Ratio Estimation

Given two samples $X^+ \sim \rho^+$ and $X^- \sim \rho^-$, how can we estimate their density ratio

$$r(x) = \frac{\rho^+(x)}{\rho^-(x)}?$$

We provide estimators for a more general rational ratio of the form

$$\frac{b_+\rho^+(x) - b_-\rho^-(x)}{a_+\rho^+(x) + a_-\rho^-(x)},$$

where $a_\pm, b_\pm$ are constants.

## 1.1 Least Squares Estimators

The general idea is that we can recover information about the density ratio by fitting a function $f$ to different targets for data drawn from different distributions. Consider the following objective:

$$\min_f a\,\mathbb{E}_{X^+\sim\rho^+}[(f(X^+) - 1)^2] \;+\; b\,\mathbb{E}_{X^-\sim\rho^-}[(f(X^-) + 1)^2],$$

where $a, b > 0$ are positive coefficients. Here, we regress $f(X)$ to 1 if $X \sim \rho^+$, and to $-1$ if $X \sim \rho^-$.

**Theorem 1.1.1.** *The minimizer of the objective above is*

$$f^*(x) = \frac{a\rho^+(x) - b\rho^-(x)}{a\rho^+(x) + b\rho^-(x)}.$$

*Proof.* Expanding the expectations as integrals, the loss can be written as

$$L(f) = \int \left(a\rho^+(x) + b\rho^-(x)\right) f(x)^2 - 2\left(a\rho^+(x) - b\rho^-(x)\right) f(x)\,\mathrm{d}x + \text{const}.$$

It is clear that for each $x$, the value of $f(x)$ minimizing $L(f)$ is

$$f^*(x) = \frac{a\rho^+(x) - b\rho^-(x)}{a\rho^+(x) + b\rho^-(x)},$$

This completes the proof. □

**Remark** In general, we can fit $f$ to $m_+(x)$ for positive samples and to $m_-(x)$ for negative samples:

$$\min_f a\,\mathbb{E}_{X^+\sim\rho^+}[(f(X^+) - m_+(X^+))^2] + b\,\mathbb{E}_{X^-\sim\rho^-}[(f(X^-) - m_-(X^-))^2], \quad a + b > 0,$$

where $m_+$ and $m_-$ are given functions. The minimizer in this case is

$$f^*(x) = \frac{am_+(x)\rho^+(x) + bm_-(x)\rho^-(x)}{a\rho^+(x) + b\rho^-(x)}.$$

## 1.2 Convex $\phi$ Loss

We now extend the least squares loss to a more general form using a convex function $\phi$ to replace the $(\cdot)^2$ cost. Consider the objective:

$$\min_f \mathbb{E}_{X^+\sim\rho^+}[a_+\phi(f(X^+)) - b_+f(X^+)] + \mathbb{E}_{X^-\sim\rho^-}[a_-\phi(f(X^-)) + b_-f(X^-)],$$

where $\phi$ is a strictly convex function, $a_+, a_- \geq 0$, and $b_+, b_- \in \mathbb{R}$.

**Theorem 1.2.1.** *The optimal solution to the problem above satisfies*

$$\nabla\phi(f^*(x)) = \frac{b_+\rho^+(x) - b_-\rho^-(x)}{a_+\rho^+(x) + a_-\rho^-(x)}.$$

*Proof.* Expanding the expectations into integrals, the objective becomes

$$\int \big((a_+\rho^+(x) + a_-\rho^-(x))\phi(f(x)) - (b_+\rho^+(x) - b_-\rho^-(x))f(x)\big)\,\mathrm{d}x.$$

For each $x$, this is a pointwise convex optimization problem of the form

$$\min_{f(x)} A(x)\phi(f(x)) - B(x)f(x),$$

where $A(x) = a_+\rho^+(x) + a_-\rho^-(x)$ and $B(x) = b_+\rho^+(x) - b_-\rho^-(x)$. The unique minimizer is given by

$$A(x)\nabla\phi(f^*(x)) = B(x).$$

This yields the results. □

**Cross Entropy Loss** Let $\phi(x) = \log(\exp(x) + \exp(-x))$, the softplus of $|x|$. Then,

$$\nabla\phi(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} = \tanh(x).$$

Therefore, the optimal solution satisfies

$$\frac{\exp(2f^*(x)) - 1}{\exp(2f^*(x)) + 1} = \frac{b_+\rho^+(x) - b_-\rho^-(x)}{a_+\rho^+(x) + a_-\rho^-(x)}.$$

In particular, taking $a_+ = a_- = b_+ = b_- = 1$, and matching the two sides we get

$$2f^*(x) = \log\frac{\rho^+(x)}{\rho^-(x)}.$$

This reduces to the typical logistic regression estimator of density ratio:

$$\max_f \mathbb{E}_{X^+\sim\rho^+}[\log p_f(X^+)] \; + \; \mathbb{E}_{X^-\sim\rho^-}[\log(1 - p_f(X^+)],$$

where $\log p_f(x) = f(x) - \log(\exp(f(x)) + \exp(-f(x)))$, and $\log(1 - p_f(x)) = -f(x) - \log(\exp(f(x)) + \exp(-f(x)))$.