adversarial_regression

We consider a similar scenario to the above adversarial problem. Let $D \in \mathbb{R}^{n \times p}$ be a static set of n samples of p features with corresponding labels $\gamma \in \mathbb{R}^n$. The learner, in the upper-level, seeks to construct a prediction model with weights $w \in \mathbb{R}^p$ on this data. Meanwhile, suppose there's any adversary who creates dataset $X \in \mathbb{R}^{m \times p}$ containing m samples of the same features with corresponding labels $Y \in \mathbb{R}^m$. The adversary seeks to have their data mislabelled as some target labels $Z = Y + \mu$ for some $\mu \in \mathbb{R}^m$. For example, suppose the learner is training a model to predict the expected insurance payouts for new customers. An adversary might lie on their application form to achieve lower insurance premiums.

$$\begin{aligned} \underset{w,X}{\text{minimise}} \quad & \frac{1}{n}\|w^T D - \gamma\|_2^2 + \frac{1}{m}\|w^T X - Y\|_2^2 + \frac{1}{\rho}\|w\|_2^2 \\ \text{subject to} \quad & X \in \arg\min_y \begin{cases} \frac{1}{m}\|w^T X - Z\|_2^2 \\ \text{s.t.} \quad \delta - \frac{X_i \cdot X_i^0}{||X_i|| ||X_i^0||} \leq 0, \quad i = 1, \ldots, m \end{cases} \end{aligned}$$

Where $\rho \in \mathbb{R}$ is a regularisation parameter and $\delta \in \mathbb{R}$ is the similarity threshold.