The original data set is $(X, \mathbf{y})$ where X is a $n \times 2$ matrix of (year, 1), and $\mathbf{y}$ is the vector of ice days. The attacker can modify $\mathbf{y}$ by adding a vector δ. The attacker's goal is to make ordinary least square solution on $(X, \mathbf{y} + \delta)$ to have a non-negative slope. The attacker also wants to minimize the p-norm of the change vector δ to hide the attack.

The bilevel problem is:

$$\begin{aligned} \min_{\delta \in \mathbb{R}^n, \alpha \in \mathbb{R}^2} \quad & \|\delta\|_p && (1) \\ \text{s.t.} \quad & \alpha_1 \geq 0 && (2) \\ & \alpha = \min_{\beta \in \mathbb{R}^2} \|\mathbf{y} + \delta - X\beta\|^2 && (3) \end{aligned}$$

References

[1] Shike Mei and Xiaojin Zhu. Using machine teaching to identify optimal training-set attacks on machine learners. In *The Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.