We consider a binary strategic classification task.

let be a population and let D be a distribution over X.

let be a target classifier and let be a non-negative cost function.

In this setting, there is a judge who wishes to publish a classifier which maximizes his prediction accuracy over a strategically modified dataset. Formally, the judge wishes to publish a classifier f which maximizes , where .

We focus on the case where f is linear (i.e., of the form for some )

given a set of labeled examples , drawn from the distribution D we denote the judge’s loss as follows:

To minimize this loss, we use the help of a differentiable optimization layer. This layer can solve convex optimization problems while still allowing backpropagation gradients to flow through it. We use it to solve the argmax problem of .However, the sign function is not convex.

For this reason, we use a sigmoid-like function that smoothens the sign function. Then, for each sample x, we approximate a convex function which has the same solution to the argmax problem as the non-convex sigmoid-like function.

To smooth the sign function, we use the function:

To find the approximated function, we use the convex-concave procedure (ccp).

Then, we use gradient decent techniques to minimize the loss. The loss of batch of samples B is calculated as follows: