Algorithm S1: Computation of a targeted UAP

Input: Set $\boldsymbol{X}$ of input images, target class y, classifier C, set $\boldsymbol{Q}$ of search directions, attack strength ϵ, cap ξ on L_p norm of the perturbation, norm type p (1, 2, or ∞), maximum number $i_{\max}$ of iterations.

Output: non-targeted UAP vector $\boldsymbol{\delta}$

$\boldsymbol{\delta} \leftarrow \mathbf{0}$, $r \leftarrow 0$, $i \leftarrow 0$
while $r < 1$ and $i < i_{\max}$ **do**
    Pick a direction randomly: $\boldsymbol{q} \in \boldsymbol{Q}$
    for $\alpha \in \{-\epsilon, \epsilon\}$ **do**
        $\boldsymbol{\delta}' \leftarrow \text{project}(\boldsymbol{\delta} + \alpha\boldsymbol{q}, p, \xi)$
        if $\sum_{\boldsymbol{x} \in \boldsymbol{X}} p_C(y|\boldsymbol{x} + \boldsymbol{\delta}') > \sum_{\boldsymbol{x} \in \boldsymbol{X}} p_C(y|\boldsymbol{x} + \boldsymbol{\delta})$ **then**
            $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta}'$
            break
        end if
    end for
    $r \leftarrow |\boldsymbol{X}|^{-1} \sum_{\boldsymbol{x} \in \boldsymbol{X}} \mathbb{I}(C(\boldsymbol{x} + \boldsymbol{\delta}) = y)$
    $i \leftarrow i + 1$
end while