Notes on Game redefinition

Taking into account past meetings and papers I have read, I came to conclusion the following shall be redefined:

- A defender plays a mixed strategy over pure strategies (a pure strategy is a particular classifier $h \in \mathcal{H}$, $h : \mathbb{X} \mapsto \mathbb{C}$). However, this in effect is used as a randomised classifier in which first a classifier h is drawn and then is used to classify a sample $x \in \mathbb{X}$. Thus, we can very well model the mixed strategy thoroughly by defining a defender's goal as a search for a distribution $D_\theta(d|x)$ where $d \in \mathbb{C}$ stands for a decision ruled by the defender.
- An attacker may choose to not employ an attack if it finds out its less costly. For example, a particular critical set of n queries $Q^{\mathsf{CR}} = \{q_1, \ldots, q_n\}$ might be too hard to obfuscate so it is rational to not employ the queries at all. For this reason, an attacker's action set is extended by $\texttt{NO} - \texttt{ATTACK}$. This means that given a critical query set Q^{CR}, an attacker either obfuscates the critical query set and carries out the attack, or it chooses not to attack.

The stated imposes two implications:

- The attacker's model must be revisited to incorporate both the mixed strategy of the defender and the extension of its action space.
- From the defender's view, $\texttt{NO} - \texttt{ATTACK}$ can be considered a detected attack in terms of loss function, i.e. a defender's loss is also zero in this case.

Defender Model

Loss function

The loss of a defender ℓ_{-1} is defined in the table bellow. Note that,

	$\texttt{NO} - \texttt{ATTACK}$	Obfuscation by g
B	ϵ	$\Omega_{-1}(g(Q^{\mathsf{CR}}), Q^{\mathsf{CR}})$
M	ϵ	$\Omega_{-1}(g(Q^{\mathsf{CR}}), Q^{\mathsf{CR}}) + L_0$

This extension also leaves room for further enhancements of the attacker's action space. For example, a partial objectives may be considered meaning only a subset of Q^{CR} may be used.

Task

Again, the defender's task is to minimise the expected risk R_{-1}:

$$R_{-1}(a, D_\theta) = \mathbb{E}\, \ell_{-1}(d, a)$$

Recall that d is generated by $D_\theta(d|x)$ and a is an action played by the defender.

However, the task reduces to the following form. For a given Q^{CR}, the attacker solves the task:

$$\min_{x \in S(Q^{\mathsf{CR}})} L_0 \cdot D_\theta(\mathsf{M}|x) + \Omega_{-1}(x, Q^{\mathsf{CR}})$$

If the optimal value of the criterion is lower than ϵ then the attacker does not attack, playing the action $\texttt{NO} - \texttt{ATTACK}$. Otherwise, the optimal argument x^* is played.

Descent

Since the criterion of the defender's task is fully differentiable, it can be optimised with gradient descent. In our domain, however, $S(Q^{\mathsf{CR}})$ stands for a discrete set of plausible obfuscations of a critical set Q^{CR}. We already defined that only additive changes are allowed (we can only query the service with more urls). In conclusion, a projected gradient descent or any other discrete descent method may be used to optimise the criterion (we discussed a linear relaxation of the problem that I proposed).

Optimal solutions

We know that the set of optimal solutions X^* of the problem is a subset of arguments that set the first derivative of the criterion to zero.

$$X^* \subseteq \{x | L_0 \cdot \frac{\partial D_\theta(\mathsf{M}|x)}{\partial x} + \frac{\partial \Omega_{-1}(x, Q^{\mathsf{CR}})}{\partial x} = 0\}$$

The aforementioned descent then converges to x^* that belongs to the just defined set or is very close to it.

Attacker model

The attacker's expected risk that is being minimised is defined as:

$$R_{+1}(\theta) = \mathbb{E}\,[\,\ell_{+1}(d, c)\,]$$

where the expectation is over input samples x, decisions $d \sim D_\theta$ and ground-truth classes c.

The nature of the task implies the priori class probabilities are not known. Therefore, we choose to redefine the task to comply with the Neyman-Pearson Problem. This means, instead of minimising $R_{+1}(D_\theta)$, we solve the following problem:

$$\min_\theta \mathbb{E}\,[\,\ell_{+1}(d, \mathsf{M})\,|\,\mathsf{M}\,] \qquad \text{s.t.} \quad \mathbb{E}\,[\,\ell_{+1}(d, \mathsf{B})\,|\,\mathsf{B}\,] \leq \tau_{\mathsf{max}}$$

Employing Langrange multiplicators, we arrive at:

$$\max_{\lambda \geq 0} \min_\theta \mathbb{E}\,[\,\ell_{+1}(d, \mathsf{M})\,|\,\mathsf{M}\,] + \lambda \cdot (\,\mathbb{E}\,[\,\ell_{+1}(d, \mathsf{B})\,|\,\mathsf{B}\,] - \tau_{\mathsf{max}}\,)$$

Notice that $\lambda = \frac{p(\mathsf{B})}{p(\mathsf{M})}$. Conveniently, the inner minimisation task can be rewritten to the following form:

$$\min_\theta \mathbb{E}\,[\,\ell_{+1}(d, \mathsf{M})\,|\,\mathsf{M}\,] \cdot p(\mathsf{M}) + \mathbb{E}\,[\,\ell_{+1}(d, \mathsf{B})\,|\,\mathsf{B}\,] \cdot p(\mathsf{B})$$

which reassembles minimisation of $R_{+1}(\theta)$.

Inspiration by Stackgrad

Building on the basis of Amin et al. (STACKGRAD), we first fix λ and compute correspondingly the prior class probabilities. This reduces the task back to the general definition with now estimated prior class probabilities. The problem can now be solved with gradient descent. The gradient of the criterion is as follows:

$$\nabla_\theta R_{+1}(\theta) = \mathbb{E}\,[\;\ell_{+1}(d, c) \cdot \frac{\nabla_\theta D_\theta(d|x)}{D_\theta(d|x)}\,]$$

Notice the gradient is defined as an expectation over random variables c, x, d. Since its full analytical form is intractable, we can estimate it by sampling the triple.

First we draw c according to $p(c)$ (determined by λ). If a benign class is drawn, then x is taken from the training dataset of benign samples T^{B}. But if a malicious class is drawn, we randomly select Q^{CR} and perform the attacker optimisation $g^*(Q^{\mathsf{CR}})$. If the attacker chooses not to carry out any attack, then simply a correct classification is assumed and the procedure stops. But if the attacker attacks with $x = g^*(\ Q^{\mathsf{CR}})$, a decision d is drawn from $D_\theta(d|x)$.

Gradient of Mixed Strategy

$\nabla_\theta D_\theta(d|x)$ is different for each class based on the way x is generated.

For a benign class, the situation is fairly simple. The gradient consists of partial derivatives:

$$\frac{\partial D_\theta(\mathsf{B}, x)}{\partial \theta_i}$$

However, this gets more complicated in case of a malicious class due to the presence of attacker's best response. Since $g^*(Q^{\mathsf{CR}})$ is also dependant on θ through out the distribution D_θ, we employ a chain rule in differentiation of D_θ:

$$\frac{\partial D_\theta(\mathsf{M}, g^*(Q^{\mathsf{CR}}))}{\partial \theta_i} = \frac{\partial D_\theta}{\partial \theta_i}(\mathsf{M}, g^*(Q^{\mathsf{CR}})) + \sum_k \frac{\partial D_\theta}{\partial x_k}\frac{\partial x_k}{\partial \theta_i}$$

The term $\frac{\partial x_k}{\partial \theta_i}$ is interesting. It stands for a derivative of x with respect to a parameter θ_i.

Recall that we have already derived that the optimal attack feature vector x^* is a member of a set of optimal solutions X^* that is a subset of arguments that set the first derivative of the attacker's criterion to zero.

And since we compute the partial derivative $\frac{\partial x_k}{\partial \theta_i}$ at the point x^*, the first derivatives set can be utilised.

$$x^* \in \{x | L_0 \cdot \frac{\partial D_\theta(\mathsf{M}|x)}{\partial x} + \frac{\partial \Omega_{-1}(x, Q^{\mathsf{CR}})}{\partial x} = 0\}$$

Using the Implicit Function Theorem, which define derivatives of variables expressed by implicit functions, we are able to arrive at a close form of the partial derivative $\frac{\partial x_k}{\partial \theta_i}$. For simplicity, we define:

$$f_k(x, \theta) = L_0 \cdot \frac{\partial D_\theta(\mathsf{M}|x)}{\partial x_k} + \frac{\partial \Omega_{-1}(x, Q^{\mathsf{CR}})}{\partial x_k} = 0$$

Now, we use f to derive the partial derivate:

$$\frac{\partial x_k}{\partial \theta_i} = -\frac{\frac{\partial f_k}{\partial \theta_i}}{\frac{\partial f_k}{\partial x_k}}$$

The problem with this solution comes with infinitesimal shifts $\partial\theta_i$ that cause the attacker change its action to $\mathtt{NO-ATTACK}$. The function f is only defined at points whose corresponding risk is lower than ϵ. However, this might not be a critical problem of the solution.

Monte Carlo Estimaton

The gradient can be estimated from m sampling cycles based on the current $\theta^{(t)}$, each yielding $\gamma^{(i)}$:

$$\gamma^{(i)} = \ell(d^{(i)}(x^{(i)}), c^{(i)}) \cdot \frac{\nabla_\theta D_{\theta^{(t)}}(d^{(i)} \mid x^{(i)})}{D_{\theta^{(t)}}(d^{(i)} \mid x^{(i)})}$$