1 Empirical Risk Minimization

Deep learning models are general function estimators that are usually trained under the empirical risk minimization framework. In supervised image classification, we want to fit hypothesis f to an existing but unavailable oracle function $h : \mathbf{X} \longrightarrow \mathbf{Y}$. Given a loss function $\mathcal{L}(\hat{y}, y)$, where $\hat{y}$ is the model prediction and y is the target, the true risk of an arbitrary hypothesis f is defined as

$$\mathcal{R}(f) = \mathbb{E}_{(x,y)\sim\mathcal{P}(x,y)}\Big[\mathcal{L}(f(x), y)\Big] = \int \mathcal{L}(f(x), y) dP(x, y) \tag{1}$$

However, since the true risk can usually not be computed, we instead sample repeatedly from joint probability distribution $\mathcal{P}(x, y)$ to create a dataset $\mathcal{D} : (x_1, y_1)$, (x_2, y_2), ..., (x_n, y_n) and test the hypothesis using the empirical risk:

$$\hat{\mathcal{R}}(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\mathcal{L}(f(x), y)\Big] = \sum_{i}^{n} \mathcal{L}(f(x_i), y_i) \tag{2}$$

The optimization problem under ERM takes the shape of

$$f^* = \arg\min_{f\in\mathcal{F}} \hat{\mathcal{R}}(f) \tag{3}$$

where the goal is to find the optimal hypothesis in a larger family of possible solutions that minimizes the empirical risk. It is assumed that minimizing the empirical risk will often deliver an acceptable solution to the true risk. **How realistic this claim is will depend on how well the training data captures the underlying generative distribution** $P(x, y)$.

2 Adversarial attacks

Deep learning models trained under the **ERM** framework have been shown to be vulnerable to adversarial attacks. To counter this, a proposed approach from [1] is to reformulate the previous optimization problem as

$$f_\theta^* = arg\min_\theta \ \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\max_\delta \mathcal{L}(f_\theta(x+\delta), y)\Big] = arg\min_\theta \frac{1}{n}\sum_{i}^{n} \max_\delta \mathcal{L}(f_\theta(x_i+\delta), y_i) \tag{4}$$

1. Solution takes the shape $(\theta, \delta_1, \delta_2, ..., \delta_n)$
2. Protects against the average worse case scenario specific to each sample (x_i, y_i).
3. Mutually high information context. Attacker has information about its target and its features. Defender must defend against the *most dangerous* attack possible against said target.
4. The formulation attributes high priority to highly damaging attacks and does not consider the average attack, which may be more common (formulation does not take frequentist nature of attacks into account). May be suited for scenarios where the average attack is not very damaging, but the worse case scenario is absolutely catastrophic.

The previous formulation is contrasted by the more general case:

$$f_\theta^* = arg\min_\theta \max_\delta \ \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\mathcal{L}(f_\theta(x+\delta), y)\Big] = arg\min_\theta \max_\delta \frac{1}{n}\sum_{i}^{n} \mathcal{L}(f_\theta(x_i+\delta), y_i) \tag{5}$$

1. Solution takes the shape (θ, δ)
2. Protects against the average worse case scenario relating to the entire dataset.

3. Mutually low information context. The attacker has information about the generative distribution but not about the specific target sample. The defender does not have prior information about the attack type, nor does he know about the targeted sample.

4. The formulation attributes the same priority to all attacks. May be better suited to less critical systems where the range in consequences of attacks is less significant.

An important note is that these two formulations do not penalize perturbations with higher norms (that move the adversarial image further away from the original image), which is usually an important consideration in adversarial settings.

3 Optimisation problem

- The number of samples N is large. Deep learning datasets usually reach the tens of thousands of examples.

- Inputs are typically composed of 3 integer channels ranging from [0-256] (RGB). However, inputs are often normalized to the [0-1] range, which moves the problem from integer to real programming.

  Model parameters θ are real valued and continuous.

- The dimensionality of θ is very large. Neural networks often reach hundreds of thousands of parameters (or more). The dimensionality of δ is equal to the input size, which is comparatively small but can still easily reach the thousands or millions (especially in high definition image classification).

- We do technically have access to the analytical form of f_c, as neural networks are compositions of scalar products and differentiable non-linear functions. Differentiability is required for conventional backpropagation, although some activation functions are non-differentiable in 1 point or another (ReLU at x=0).

  The analytical form can rapidly become very complex, especially in image classification where the scalar products are heavily indexed (2d scalar product).

- The classifier is trained using binary cross entropy:

$$\mathcal{L} = -\frac{1}{N}\sum_{i}^{N}[y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i)]$$

- We can modify the overall optimization problem to add constraints on θ and δ. For example, a popular practice is to add a weight decay term to favor generalization to out of sample data. The same constraint could be applied to the l-norm of δ in order to produce less perceptible attacks. Equations 4 and 5 become

$$f^* = arg\min_{\theta} \frac{1}{n}\sum_{i}^{n} \max_{\delta} \mathcal{L}(f_\theta(x_i+\delta), y_i) + \epsilon_1 ||\theta||_{s_1} + \epsilon_2 ||\delta||_{s_2} \tag{6}$$

$$f^* = arg\min_{\theta} \max_{\delta} \frac{1}{n}\sum_{i}^{n} \mathcal{L}(f_\theta(x_i+\delta), y_i) + \epsilon_1 ||\theta||_{s_1} + \epsilon_2 ||\delta||_{s_2} \tag{7}$$

- The landscape of the loss in the space of θ and δ is more than likely highly non-convex.

4 Foundational literature

4.1 Towards Deep Learning Models Resistant to Adversarial Attacks [1], 2017, 6663 citations

The authors claims that the classical ERM framework does not yield models that are robust against **adversarially crafted examples**. They offer equation 4 as a way to obtain an upper bound on all possible attacks. They claim that obtaining a small average loss gives a guarantee that no strong attacks are possible since the loss is small for all allowed perturbations.

4.2 Universal adversarial perturbations [2], 2017, 2075 citations

The authors show the existence of a **single universal and small perturbation vector** that causes natural images to be misclassified with high probability. They impose a norm constraint ϵ on the perturbation vector v

$$||v||_p \leq \epsilon \tag{8}$$

and impose a *fooling rate* δ on the classifier $\hat{k}$:

$$\mathcal{P}_{x\sim D}[\hat{k}(x+v) \neq \hat{k}(x)] \geq 1-\delta \tag{9}$$

Although not modelled the same way, the objective remains very similar to equation 5, with added constraints. We seek a single attack vector that will fool the average target most of the time. The following is the proposed algorithm by the authors

Algorithm 1 Computation of universal perturbations.

input: Data points X, classifier $\hat{k}$, desired ℓ_p norm of the perturbation ξ, desired accuracy on perturbed samples δ.
output: Universal perturbation vector v.
Initialize $v \leftarrow 0$.
while $\text{Err}(X_v) \leq 1-\delta$ **do**
  for each datapoint $x_i \in X$ **do**
    if $\hat{k}(x_i+v) = \hat{k}(x_i)$ **then**
      Compute the *minimal* perturbation that sends $x_i + v$ to the decision boundary:

$$\Delta v_i \leftarrow \arg\min_r \|r\|_2 \text{ s.t. } \hat{k}(x_i+v+r) \neq \hat{k}(x_i).$$

      Update the perturbation:

$$v \leftarrow \mathcal{P}_{p,\xi}(v+\Delta v_i).$$

    end if
  end for
end while

Figure 1: Caption

4.3 Relevant derivative works (direct citations)

- [2] Universal Adversarial Perturbations Against Semantic Image Segmantation, 2017, 251 citations
- [2] Fast Feature Fool: A data independent approach to universal adversarial perturbations, 2017, 178 citations
- [2] Defense against Universal Adversarial Peturbations, 2018, 190 citations
- [1] [2] Adversarially Robust Generalization Requires More Data, 2018, 578 citations

- [1] Robustness May Be at Odds with Accuracy, 2018, 1080 citations
- [1] Scaling provable adversarial defenses, 2018, 365 citations
- [1] Exploring the Landscape of Spatial Robustness, 2019, 314 citations
- [1] On Evaluating Adversarial Robustness, 2019, 609 citations
- [1] Theoretically principled trade-off between robustness and accuracy, 2019, 1293 citations
- [1] Overfitting in adversarially robust deep learning, 2020, 356 citations

References

[1] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[2] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.