Sujet de Thèse – PhD Research Project

Starting **September 2020**.

- **Subject:** Reliability of Deep Learning networks with Rare Event simulation algorithms. Theoretical and practical issues.
- **Collaboration:** Thèse Cifre-Defense with La Ruche Thales Group.
- **Locations:**
  - Thales Group - La Ruche - Rennes
  - Inria Rennes (teams SIMSMART & LinkMedia)
- **Starting date**: September 2020.
- **Field:** Machine Learning, Statistical Reliability Engineering, Computational Probability and Statistics.
- **Key Words**: Deep Neural Networks, Large Deviations, Sequential Monte Carlo, Particle Methods.
- **Supervisors:**
  - Academic: Mathias Rousset (mathias.rousset@inria.fr), Teddy Furon (teddy.furon@inria.fr)
  - Industrial: Louis-Marie Traonouez (louis-marie.traonouez@thalesgroup.com).
- **Conditions:** French citizenship required.

1 Context of the PhD thesis

The topic of the PhD is the assessment of the reliability of some machine learning algorithms, especially fully supervised classifiers. Thanks to a training dataset composed of observations and associated class labels, a classifier is learned to predict the class of any observation. Reliability is usually considered as the generalization ability of the classifier to unseen observations. This is measured by the *accuracy*, *i.e.* the empirical probability of predicting the correct class over a testing dataset.

This thesis goes beyond this adhoc procedure by making the connection with the field of *Statistical Reliability Engineering*. This field assumes that the inputs (load, strength, and stress) of a complex system are corrupted by uncertainties which may lead to a failure. Given a statistical model, the goal is then to estimate the probability of failure. This framework makes sense in our application, classification of noisy observations, or when there is a mismatch between the statistical distributions of the training and testing data.

The application of Statistical Reliability Engineering tools to Machine Learning is however not trivial for the following reasons:

- State-of-the-art classifiers (like deep neural networks) are very complex systems. This prevents understanding the *physics of failure*. The knowledge of the main root cause failure mechanisms is out of reach.
- Classifiers tackle high dimensional observations where tiny uncertainties can snowball in a wrong prediction. We foresee that the estimation of the probability of failure will be more difficult as the dimension of the space is larger.

These difficulties are indeed shared with the field of *Rare Event Simulations*.

The goal of the PhD thesis is to invent a new reliability measurement tool specific to Deep Neural Networks. This tool will be applied to many setups ranging from image classification (ImageNet, MNIST, CIFAR) to airplane traffic monitoring. This last application is the core of the collaboration with the industrial partner La Ruche from Thales Group.

2 Scientific background

The Phd thesis is at the crossroad of three domains: Deep Learning, Rare Event Simulations, and Statistical Reliability Engineering. To clarify the topic, we introduce the following notations and backgrounds. Thanks to a training dataset $(\mathbf{x}_i, y_i)_i$ where $\mathbf{x}_i$ denotes one observation and y_i its associated label y_i, a classifier is learned to predict the class $\hat{y} = \mathsf{predict}(\mathbf{x})$ of any observation.

2.1 Uncertainties, Reliability and Security of Deep Neural Networks.

Consider a particular observation $\mathbf{x}$ which is well classified.

Is there any perturbation $d\mathbf{x}$ *of norm* $\|d\mathbf{x}\| < \epsilon$ *which leads to a failure:*
$\mathsf{predict}(\mathbf{x}) \neq \mathsf{predict}(\mathbf{x} + d\mathbf{x})$*?*

The question is less about discovering such perturbations than just proving they exist and investigating how the number of such perturbations grows with ϵ. As far as Deep Neural Networks are concerned, very few works [14, 2] are able to answer this problem thanks to Satisfiability Modulo Theory solvers. This task being NP-complete, their approach is tractable only for very small networks with a specific norm (L_∞).

We believe that a probabilistic formulation is more insightful and tractable: What is the probability $\mathbb{P}(\mathsf{predict}(\mathbf{x}) \neq \mathsf{predict}(\mathbf{x}+\mathbf{U}))$ where $\mathbf{U}$ models random uncertainties perturbating $\mathbf{x}$ along the acquisition chain. They can also model a source mismatch where hot data to test might come from a different acquisition chain hence follow a different distribution than cold data used in training.

Estimating this probability is useful in adversarial sampling too. It gauges the security of a given classifier against an attacker who perturbates the observation to induce a wrong prediction. So far, attacks and defenses fuel an endless cat and mouse game [18, 16, 17, 5]. Demonstrating that one defense breaks an attack does not mean that this defense is robust to anything more

elaborate. In contrast, our security assessment based on probability estimation will be regardless to any attack.

2.2 Rare Events Simulation

The simulation of a rare event, as well as the estimation of its (very small) probability, is a crucial scientific computing problem arising for various models of a random physical phenomena or statistical contexts. This problem is key in several application domains like risk and reliability analysis (air traffic control, neutrons in nuclear safety), prediction of extreme events (meteorology, climate science), or the simulation of "hopping" between physical meta-stable states (chemical reactions in molecular simulation).

In such complex systems, the analytic study is out of reach, and a simulation must be used. We are interested in the so-called static case, where we want to estimate $\mathbb{P}(s(\mathbf{U}) > l^\star)$, with $\mathbf{U}$ a random variable perturbating the stable state of the system, and s a score function. It reflects the deviation from the stable state of the system and triggers a failure when bigger than level $l^\star$. We suppose we know how to simulate the law π of $\mathbf{U}$. When the event is really rare (say 10^{-10}), a naïve Monte Carlo approach does not work.

Importance Splitting [7] is a state-of-the-art technique, also known as *Subset Sampling*. It consists in describing the rare event by nested increasingly rare events $A_k = \{\mathbf{u} : s(\mathbf{u}) > l_k\}$ for $l_1 < \cdots < l_n = l^\star$, and to estimate the probabilities $\mathbb{P}(A_k|A_{k-1})$. The most standard version of the algorithm falls within the framework of sequential Monte-Carlo type particle methods developed, among others, by P. Del Moral [19] and A. Doucet. However, there are many optimized variants that do not fit into this framework and whose analysis is either incomplete or non-existent.

A simple example of an efficient variant of these algorithms is the so-called *last particle adaptive splitting algorithm* [12]. In a nutshell, it assumes that we know how to simulate a Markov probability transition P reversible for the law π. We then consider N particles whose measure initially is $\pi = \mathcal{L}(\mathbf{U})$, and at iteration k $\mathcal{L}(\mathbf{U}|\, s(\mathbf{U}) > l_k)$. At each stage, i) the lowest score particle is killed, ii) another randomly selected particle is duplicated, and iii) is randomly modified ($\mathbf{U} \to \mathbf{U}'$ using the Markov transition P, unless the proposed state has a smaller score. We stop the algorithm when all the particles have a score higher than $l^\star$.

2.3 Statistical Reliability Engineering

A canonical approach in Statistical Reliability Engineering [10] is to transform the problem in order to deal with Gaussian latent random variables. In our context, suppose the above-mentioned π is a standardized Gaussian distribution. Then, the First Order Reliability Method assumes that the score function s is linear. The probability of failure $\mathbb{P}(s(\mathbf{U}) > l^\star)$ is approximated as $\hat{P} = \Phi(-\|\mathbf{u}^\star\|_2)$ where $\mathbf{u}^\star$, so-called the design point, is the closest point from the origin satisfying $s(\mathbf{u}^\star) = l^\star$. In other words, $\mathbf{u}^\star$ is the failure point with the biggest probability density and the surface is approximated as the affine

hyperplane $\{\mathbf{x} : \mathbf{x}^\top \mathbf{u}^\star = \|\mathbf{u}^\star\|^2\}$. Its norm $\|\mathbf{u}^\star\|$ usually defines the reliability index β [13].

There are plenty of algorithms for finding the design point $\mathbf{u}^\star$ on the surface. The most popular techniques are [15]:

- First order algorithms needing the gradient of function s: Hasofer-Lind-Rackwitz-Fiessler (HLRF) steepest descend [13, 20] and its improved version with convergence assessment [21],
- Second order algorithms needing the Hessian of function s (Sequential Quadratic Programming) or a cheap approximation (Broyden-Fletcher-Goldfarb-Shanno) [4],
- Hybrid algorithms like Abdo-Rackwitz [1].

A classical refinement is *Importance Sampling*. In this context, importance sampling estimates the probability of failure using an i.i.d. sample with a new law centered at $\mathbf{u}^\star$. It outputs a new estimation of the probability $\hat{P}$ which is larger or smaller than $\Phi(-\|\mathbf{u}^\star\|_2)$ depending on the local curvature of the surface $\{\mathbf{x} : s(\mathbf{x}) = l^\star\}$ around $\mathbf{u}^\star$. This in turn is mapped back to a new reliability index $\beta = -\Phi^{-1}(\hat{P})$. This method suffers from two weaknesses:

- Multiple design points: there might exist other points on the surface as close as $\mathbf{u}^\star$. Discarding them amounts to underestimate the probability of failure and hence to overestimate the reliability index. Some methods (like Restarted iHLRF [8]) iteratively exclude known design points in order to discover new ones, if any. Multiple simulations are launched around them and the estimated probabilities are aggregated into one global reliability index.
- High dimension: Importance Sampling is known to be inaccurate in high dimensional spaces due to the vanishing weighting function.

3 Objectives of the Phd thesis

The main objective of the PhD is to set up the mathematical and methodological analysis of a reliability measurement tool dedicated to Deep Neural Networks, and to apply it to real cases such as airplane traffic monitoring. In particular, the open problems are the following ones.

3.1 Gradient-based Rare Event Simulations

Goal: Construct and analyze efficient smoothed variants of Importance Splitting when the gradient of the score ∇s is given.

We assume that the connection in between Deep Learning, Rare Events Simulation, and Statistical Reliability Engineering has been made by crafting a score function s that measures how close the network is from making a wrong prediction: $\mathsf{predict}(\mathbf{x}) \neq \mathsf{predict}(\mathbf{x} + \mathbf{U})$.

A key fact about Deep Neural Network is that, thanks to auto-differentiation, the value of the gradient ∇s is computed by backpropagation along the layers.

Its cost is not so expensive: 2 times the complexity of a call to the network function s. This extra information should improve Importance Splitting by offering faster estimators for a given accuracy level.

To do so we intend to design a variant of Importance Splitting where the 0-or-1 selection of the surviving particles is replaced by a probabilistic mechanism keeping particles with a probability proportional to $exp(\hat{\gamma}s(\mathbf{u}))$ where $\hat{\gamma}$ is adaptively chosen so that only few particles are killed. Then the transition $\mathbf{U} \to \mathbf{U}'$ uses a gradient-based Markov Chain Monte Carlo (a well-chosen discretization of Stochastic Differential Equation typically) which targets laws whose density is proportional to $\exp(\alpha s(\mathbf{u}))\pi(d\mathbf{u})$. The difficulty here is to adaptively adjust the parameters and to manage approximations.

3.2 Theoretical analysis

Goal: Study the consistency and the bias of the estimators obtained by the new algorithm w.r.t. its complexity, and the fluctuation analysis (central limit theorem) when its complexity asymptotically increases.

The complexity of the new algorithm is recorded by the number of samples N or the number of calls to the gradient / function s. In particular, this analysis is made difficult by the $1/N$ bias of the algorithm ; the stochastic analysis based on martingales which allows to obtain asymptotic normality being much more complicated [3].

We intend to use a method already obtained in [6], which consists of rewriting the algorithm as a particle system with mutation and selection (genetic algorithm) by indexing the particles according to the level value l [19, 11]. If we do that and we discretize the levels, we obtain a discrete system whose stochastic analysis in bias and martingale parts are simpler to analyze. Then the continuous limit can be approached.

The last study concerns the asymptotic efficiency of the algorithm as the probability to be estimated tends to 0. Assuming that the underlying model (*i.e.* the distribution π) satisfies a large deviation principle (small noise asymptotics) [9], the stochastic aspect is removed and the problem becomes simply an optimization problem. This point of view is helpful to understand in depth the dependence of the rare event with respect to other quantities. In Importance Splitting, this is a difficult open problem because of the selection between copies of the tendering system. We first intend to study the classical Importance Sampling method within this asymptotics, which should be relatively straightforward.

3.3 Applications

Goal: La Ruche in Rennes is developing solutions to analyze radar detections of aircrafts. Some of the techniques involve deep learning algorithms to classify different types of aircraft from the radio magnetic signals received from the aircrafts. The goal is to apply the reliability methods that invented and developed during the thesis to this case-study.

At first it will be possible to analyze public data from ADSB sensors (Automatic Depend Surveillance Broadcast) that has already been collected. Then we could extend it to radar data. One of the goals of the reliability analysis will be to quantify the noise of the signals that could lead to misclassification. Also some of these signals being dependent on the aircraft they could be jammed in order to fool the classifier. The goal of the reliability analysis would be to quantify the probability of this event.

3.4 Open issues

A fascinating question will be investigated: the experimental and theoretical analysis of the reliability when the underlying score function s comes from a classifier trained with supervision.

An experimental work will benchmark reliabilities of several networks differing from their structure, depth, and also learning process. For instance, adversarial training [17] is believed to provide more reliable networks. Is this really the case?

From the theoretical point of view, we would like to compare the rare event of a real-life classifier *vs.* its ideal version learned over an infinite training set. We will therefore use asymptotic theories to find reliability guarantees on the quality of the score function. For instance, we may try to understand the topologies of the rare events in the large deviations asymptotical framework described above, and try to compare them with numerical experiments.

4 Roadmap

4.1 1st year

- Mastering the use case proposed by Thales Group: Airplane traffic monotoring. Develop, train, and test several classifiers from real datasets.
- Application and comparison of methods finding design points $\mathbf{x}^\star$. This benchmark includes recent algorithms designed by the Computer Vision (like C&W, PGD, DDN) and the old recipes due to the Statistical Reliability Engineering (HLRF [21], iHRF [8], AR [1]) communities. It is surprising that the connection between these two communities has not been explored yet.
- Application of classical estimators like FORM / SORM methods with Importance Sampling simulations refinement to estimate probabilities of failure. Verify that they are not suitable in high dimension spaces.

4.2 2nd year

- Application of one Importance Splitting algorithm to estimate probabilities of failure. Compare the efficiency - accuracy trade-off with the previous method.
- Design a new estimator based on Importance Splitting side-informed by the gradient. This will consume numerous back and forth travels between experimental and theoretical investigations.

4.3 3rd year

- Writing of the proofs for the consistency, the bias and the central limit theorem.
- Experimental and theoretical investigations about the dependence of the reliability to the quality of the data and the training procedure.

References

[1] T. Abdo and R. Rackwitz. Reliability of uncertain structural systems. In *Proc. Finite Elements in Engineering Applications*, pages 161–176, 1990.

[2] O. Bastani, Y. Ioannou, L. Lampropoulos, D. Vytiniotis, A. Nori, and A. Criminisi. Measuring neural net robustness with constraints. In *Neural Information Processing Systems*, 2016.

[3] C.-E. Bréhier, M. Gazeau, L. Goudenège, T. Lelièvre, and M. Rousset. Unbiasedness of some generalized adaptive multilevel splitting algorithms. *Annals of Applied Probability*, 26(6):3559 – 3601, 2016.

[4] C. G. Broyden. The convergence of a class of double-rank minimization algorithms. *Journal of the Institute of Mathematics and Its Applications*, 6:76–90, 1970.

[5] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Security and Privacy*, 2017.

[6] F. Cérou, B. Delyon, A. Guyader, and M. Rousset. On the Asymptotic Normality of Adaptive Multilevel Splitting. *SIAM/ASA Journal on Uncertainty Quantification*, 7(1):1–30, 2019. 38 pages, 5 figures.

[7] F. Cérou and A. Guyader. Adaptive multilevel splitting for rare event analysis. Research Report PI 1747, 2005.

[8] A. D. K. A. T. Dakessian. Multiple design points in first and second-order reliability. *Structural Safety*, 20(1):37–49, 1998.

[9] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Stochastic Modelling and Applied Probability. Springer-Verlag Berlin Heidelberg, 2010.

[10] O. Ditlevsen and H. O. Madsen. *Structural reliability methods*, volume 178. Wiley New York, 1996.

[11] E. Gobet and G. Liu. Rare event simulation using reversible shaking transformations. *SIAM Journal on Scientific Computing*, 37(5):A2295–A2316, 2015.

[12] A. Guyader, N. Hengartner, and E. Matzner-Løber. Simulation and estimation of extreme quantiles and extreme probabilities. *Applied Mathematics & Optimization*, 64(2):171–196, 2011.

[13] A. Hasofer and N. Lind. Exact and invariant second-moment code format. *Journal of the Engineering Mechanics Division*, 100(1):111–121, 1974.

[14] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In *Int. Conf. on Computer Aided Verification*, 2017.

[15] P.-L. Liu and A. D. Kiureghian. Optimization algorithms for structural reliability. *Structural Safety*, 9(3):161 – 177, 1991.

[16] J. Lu, T. Issaranon, and D. Forsyth. Safetynet: Detecting and rejecting adversarial examples robustly. *arXiv:1704.00103*, 2017.

[17] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *Int. Conf. on Learning Representations*, 2017.

[18] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff. On detecting adversarial perturbations. *arXiv:1702.04267*, 2017.

[19] P. D. Moral. *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Probability and Its Applications. Springer-Verlag New-York, 2004.

[20] R. Rackwitz and B. Flessler. Structural reliability under combined random load sequences. *Computers & Structures*, 9(5):489 – 494, 1978.

[21] S. Santos, L. Matioli, and A. Beck. New optimization algorithms for structural reliability analysis. *CMES-Computer Modeling in Engineering & Sciences*, 83(1):23–56, 2012.