Deep Neural Networks Are Congestion Games: From Loss Landscape to Wardrop Equilibrium and Beyond

Nina Vesseron
ENS de Lyon,
F-69000, Lyon, France

Ievgen Redko **Charlotte Laclau**
Univ Lyon, UJM-Saint-Etienne, CNRS
Institut d Optique Graduate School,
Laboratoire Hubert Curien UMR 5516,
F-42023, Saint-Etienne, France

Abstract

The theoretical analysis of deep neural networks (DNN) is arguably among the most challenging research directions in machine learning (ML) right now, as it requires from scientists to lay novel statistical learning foundations to explain their behaviour in practice. While some success has been achieved recently in this endeavour, the question on whether DNNs can be analyzed using the tools from other scientific fields outside the ML community has not received the attention it may well have deserved. In this paper, we explore the interplay between DNNs and game theory (GT), and show how one can benefit from the classic readily available results from the latter when analyzing the former. In particular, we consider the widely studied class of congestion games, and illustrate their intrinsic relatedness to both linear and non-linear DNNs and to the properties of their loss surface. Beyond retrieving the state-of-the-art results from the literature, we argue that our work provides a very promising novel tool for analyzing the DNNs and support this claim by proposing concrete open problems that can advance significantly our understanding of DNNs when solved.

1 INTRODUCTION

Since the very seeding of the machine learning (ML) field, the ML researchers have constantly drawn inspiration from other areas of science both to develop novel approaches and to better understand the existing ones. One such notable example is a longstanding fruitful relationship of ML with game theory (GT) that manifested itself by the novel insights regarding the analysis of such different learning settings as reinforcement learning [Peshkin et al., 2000, Hu and Wellman, 2003, Claus and Boutilier, 1998], boosting [Freund and Schapire, 1996] and adversarial classification [Liu and Chawla, 2009, Brückner and Scheffer, 2011, Dritsoula et al., 2017] to name a few. While the interplay between ML and GT in the above-mentioned cases is natural, *ie*, reinforcement learning is a game played between the agent and the environment, boosting is a repeated game with rewards and adversarial learning can be seen as a traditional minimax game, very few works studied the connection between the deep neural networks (DNNs) and GT despite the omnipresence of the former in the ML field. Indeed, in recent years, deep learning has imposed itself as the state of the art ML method in many real-world tasks, such as computer vision or natural language processing to name a few [Goodfellow et al., 2016]. While achieving impressive performance in practice, training DNNs requires optimizing a non-convex non-concave objective function even in the case of linear activation functions and can potentially lead to local minima that are arbitrary far from global minimum. This, however, is not the typical behaviour observed in practice, as several works [Dauphin et al., 2014, Goodfellow and Vinyals, 2015] showed empirically that even in the case of training the state-of-the-art convolutional or fully-connected feedforward neural networks one does not converge to suboptimal local minima. Such a mysterious behaviour made studying the loss surface of DNNs and characterizing their local minima one of the topics of high scientific importance for the ML community.

In this paper, we propose a novel approach for analyzing DNNs' behaviour by modelling them as congestion games, a popular class of games first studied by

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

[Rosenthal, 1973] in the context of traffic routing. To this end, we first prove that linear DNNs can be cast as special instances of non-atomic congestion games defined entirely in terms of the DNNs main characteristics. This result allows us to successfully draw the parallel between local minima of the loss function of a linear DNN and the Wardrop equilibria of the corresponding non-atomic congestion game under some mild assumptions on the considered loss function. As a consequence, we prove the well-known result provided by [Kawaguchi, 2016] for linear networks regarding the equivalence between the local minima of the loss function optimized by a linear DNN and its global optimum. Second, we study the case of non-linear DNNs with rectified linear activation function (ReLU) by considering the model proposed in the seminal work of [Choromanska et al., 2014]. We model such networks as congestion games where some resources available to the agents in the game can fail. In this latter setting, we show that the seminal model of [Choromanska et al., 2014] is essentially equivalent to the linear DNN model studied before and thus enjoys the same guarantees. To the best of our knowledge, the proposed approach for the analysis of DNNs has never been studied in the literature before and we expect it to have a very strong scientific impact due to the established formal connection between one of the most studied ML models and one of the most rich areas of GT in terms of the number of available results.

The rest of this paper is organized as follows. In Section 2, we review other existing works on the analysis of DNNs and their loss surface and convergence properties. In Section 3, we provide the required preliminary knowledge related to both DNNs and congestion games. Section 4 contains our main contributions that analyze both linear and non-linear DNNs. Finally, in Section 5 we emphasize the importance of the established connection between DNNs and congestion games and pose several open problems.

2 RELATED WORK

Below, we briefly review the main related works with a particular emphasis on contributions analyzing the loss surface of DNNs[1] and those linking DNNs to GT.

Analysis of DNNs' loss surface While strong empirical performance of DNNs make of them a number one choice for many ML practitioners, it has been shown that training a neural network is NP-hard [Blum and Rivest, 1992] as it requires finding a global minimum of a non-convex function of high dimensionality. To circumvent this difficulty, the methods for convex optimization are widely used to train DNNs, but the reasons why these methods work well in practice remain unknown as, in principle, nothing restricts them from converging to poor local minima arbitrary far from the global minimum. To shed light on this, several works adapted a geometric approach to provide a justification for optimizing DNNs using convex optimization methods. This latter consists in studying the general class of non-convex optimization problems with desired geometric properties, i.e., equivalence of local minima to global optimum and negative curvature for every saddle point, and showing that DNNs belong to this class. In the case of the linear networks, such notable result was provided in the works of [Baldi and Hornik, 1989] and [Kawaguchi, 2016] who proved that local minima are global minimum when a squared loss function is considered. This statement was proved for non-linear networks as well, first by [Choromanska et al., 2014] who showed that the number of bad local minima decreases quickly with the distance to the global optimum and then by several other more recent follow-up works considering different NN's configurations [Hardt and Ma, 2017, Freeman and Bruna, 2017, Soudry and Hoffer, 2018, Safran and Shamir, 2018]. An important consequence for DNNs having these properties is that (perturbed) gradient descent provably converges to a global optimum in this case [Ge et al., 2015, Jin et al., 2017, Du et al., 2017]. Contrary to the works using the above-mentioned geometric approach, we obtain the same results for a family of loss functions resembling the squared loss relying solely on the properties of non-atomic congestion games.

Game theory and ML To the best of our knowledge, only two other studies have reduced learning of DNNs to game playing. In [Balduzzi, 2016], the author studied DNNs with non-differentiable activation functions in order to explain why methods designed for convex optimization are guaranteed to converge on modern convnets with non-convex loss functions[2]. On the other hand, the authors of [Schuurmans and Zinkevich, 2016] showed how supervised learning of a DNN with differentiable convex gates can be seen as a simultaneous move two-person zero-sum game in order to further establish the equivalence between the Karush-Kuhn-Tucker (KKT) points of a DNN and Nash equilibria of the corresponding game. With these results in hand, the authors illustrated empirically that a well-known regret matching algorithm often used to find

[1] For more results on the theoretical analysis of gradient descent convergence for over-parametrized models including Recurrent and Convolutional NNs, we refer the interested reader to [Du et al., 2017, Section 2] and [Allen-Zhu et al., 2019, Section 1.2].

[2] While being highly insightful, this work was shown to have several flaws [Schuurmans and Zinkevich, 2016, Supp. material, Section J] that remain unaddressed up to now.

coarse-correlated Nash equilibria can be used successfully to train DNNs. It is worth noticing that in both papers, the games considered by the authors were designed for their specific purpose and have not been studied independently in the game theory field. On the contrary, in this paper we aim to study DNNs as instances of arguably one of the most studied classes of games in order to make the rich body of existing theoretical results proved for them readily available for the ML researchers. Also, unlike the two other papers, we study non-atomic games which are infinite-person games with each player having an infinitesimal impact on the game's analysis and Wardrop equilibria specific to such games contrary to games with a finite number of players and Nash equilibria considered before.

3 BACKGROUND KNOWLEDGE

In this section, we briefly review the main definitions related to DNNs and congestion games.

Deep Neural Networks Let us consider a DNN defined as $N = (V, E, I, O, F)$, where 1) V is a set of vertices, i.e., the total number of units in the neural network; 2) $E \subseteq V \times V$ is a set of edges; 3) $I = \{i_1, \dots, i_d\} \subset V$ is a set of input vertices equal to the number of input features; 4) $O = \{o_1, \dots, o_C\} \subset V$ is a set of output vertices of size equal to number of outputs and 5) $F = \{f_v : v \in V\}$ is a set of activation functions, where $f_v : \mathbb{R} \to \mathbb{R}$.

In the graph defined by $G = (V, E)$ and having a layered structure with L layers, a path $p = (v_1, ..., v_L)$ with $v_1 \in I$ and $v_L \in O$ consists of a sequence of vertices such that $(v_j, v_{j+1}) \in E$ for all j. We assume that G is directed and contains no cycles, the input vertices have no incoming edges and the output vertices have no outgoing edges. We let n_l denote the number of neurons at each layer $l \in [1, \dots, L]$ where $n_1 = d$ and $n_L = C$. We further associate a (trainable) weight $w_{ij}^{(l)}$ to an edge between vertex $v_i^{(l)}$ of layer l and $v_j^{(l-1)}$ of layer $l-1$ and denote by $w^{(l)}$ the matrix of all weights between the two layers. $W = \{w^{(\ell)}, \forall \ell\}$ is the set of all parameters associated to the network. For each vertex $v_i^{(l)}$, we also associate a value (activation function) $g_i^{(l)} = f_{v_i^{(l)}}(z_i^{(l)})$ with $z_{i=1}^{(l)} = \sum_k^{n_{l-1}} w_{ik}^{(l)} g_k^{(l-1)}$.

Given a training set $\mathcal{L} = \{(x^j, y^j)\}_{j=1}^M$ drawn from distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} \subseteq \mathbb{R}^d$ and $|\mathcal{Y}| = \mathcal{C}$, the task of the neural network is to produce a predictor $h : \mathcal{X} \to \mathcal{Y}$ that assigns a label close to $y^j \in \mathcal{Y}$ to each $x^j \in \mathcal{X}$. This is done by solving the following optimization problem:

$$\min_W \text{loss}(W) = \min_W \frac{1}{M} \sum_{j=1}^{M} \ell(o^L(x^j), y^j), \qquad (1)$$

where $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is a convex loss function. Stochastic gradient descent (SGD) is commonly used to solve Problem (1) where the weights are updated either for each example x or for a mini-batch.

Congestion Games We consider a non-atomic version of the congestion games [Schmeidler, 1973] that were first defined in [Rosenthal, 1973] to model road traffic. All along the paper, we use the definition of non-atomic congestion games from [Roughgarden and Éva Tardos, 2004] and use some of the results established in this paper. A non-atomic congestion game illustrated in Figure 1 is composed of the following five elements:

- n: the size of each population of players. In non-atomic game, the number of players is infinite and the significance of one player is negligible. Consequently, players are distributed into populations and we denote by d the number of such populations. Each population $i \in [[d]]$ has a size n_i and must be seen as a flow of players.
- $E(G)$: the set of resources of the game which are available for players when choosing a strategy. In the setting we study, the resources are the edges of a graph G that players can use when choosing a path from their starting point to the ending one.
- S: the set of strategies. Each population of players needs to travel from the starting to the ending point in the graph using the graph paths between them. Then, for a population i with a starting point d_i and an ending point f_i, the set of strategies S_i of the players from population i are the set of paths that link d_i to f_i.
- c: the cost functions. To each edge of the graph $e \in E(G)$, we associate a non-negative, continuous cost function $c_e(\cdot)$ on $\mathbb{R}^+$ which denotes the cost paid by the players using this edge. The cost function depends on the flow of players that use e and can be viewed as time spent by players to travel so that more players using the same edge leads to them spending more time on it.
- a: the rates of consumption. As non-atomic congestion games were designed to model road traffic, they can also take into account the possible types of roads (i.e., edges) and the types of users in the game. The non-negative coefficients $a_{S,e}$ are created for this aim (with the convention

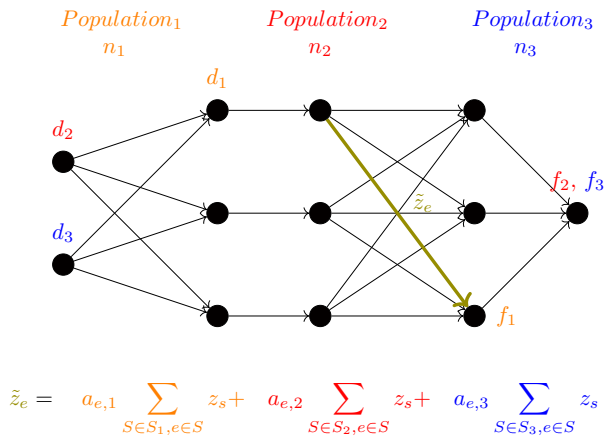

Figure 1: Example of a non-atomic congestion game with three populations of players.

$a_{S,e} = 0$ if $e \notin S$ and $a_{S,e} > 0$ otherwise). Then, for each population i, a strategy $S_i \in S$ and an edge e, we associate a coefficient $a_{S,e}$ which is used while defining the flow of users on this edge and the cost a player pays.

These five elements define a non-atomic congestion game $\text{NCG} = (E(G), c, S, n, a)$.

The decisions of players are modelled through the action distribution $z = (z_S)_{\substack{S\in S_i \\ i\in[[d]]}}$ that lists all possible strategies and is given by a vector of non-negative reals such that $\sum_{S\in S_i} z_S = n_i$ for each player type i. One can see z_S as the measure of the set of players that selects strategy S. We call $\tilde{z}_e$ the total amount of congestion on element e produced by the action distribution z:

$$\tilde{z}_e = \sum_{i=1}^{d} \sum_{S\in S_i} a_{S,e} z_S.$$

The cost $c_S(z)$ incurred by a player of type i selecting strategy $S \in S_i$ is defined with respect to the action distribution z as follows:

$$c_S(z) = \sum_{e\in S} a_{S,e} c_e(\tilde{z}_e).$$

One would have noticed that this cost is the sum over all edges used by the player of the costs of these edges. The social cost $\text{SC}(z)$ w.r.t. an action distribution z and the social optimum SO of a game are given respectively by:

$$\text{SC}(z) = \sum_{i=1}^{d} \sum_{S\in S_i} c_S(z) z_S, \quad \text{SO} = \min_z \text{SC}(z).$$

The social cost can be seen as the sum over all players of the costs payed the players while the social optimum is the optimal social cost. In what follows, when we speak about the value of an action distribution, we mean the value of the social cost associated to this distribution.

An action distribution z is a **Wardrop equilibrium** (WE) if for each player type $i = 1, 2, ..., d$ and strategies S_a, $S_b \in S_i$ such that $z_{S_a} > 0$, we have $c_{S_a}(z) \leq c_{S_b}(z)$. The Wardrop equilibrium is a situation in which no player intends to switch to another strategy because each of the players has already chosen the cheaper strategy with respect to the choices of the other players.

The main results about non-atomic congestion games needed further are the followings:

P1. Social cost $\text{SC}(z)$ can be rewritten as:

$$\text{SC}(z) = \sum_{e\in E} c_e(\tilde{z}_e)\tilde{z}_e \text{ with } \tilde{z}_e = \sum_i \sum_{S\in S_i} a_{S,e} z_S.$$

P2. Each NCG admits a Wardrop equilibrium.

P3. All Wardrop equilibria have the same value.

P4. For a given game NCG, we define the price of anarchy (PoA) of a game as:

$$\text{PoA}(\text{NCG}) = \frac{\text{WE}(\text{NCG})}{\text{SO}(\text{NCG})}.$$

4 MAIN CONTRIBUTIONS

We start by introducing the assumptions needed to model DNNs as non-atomic congestion games. We argue that these assumptions are not restrictive in practice and have been used in the literature before. Then, we proceed by formally proving the equivalence between DNNs and non-atomic congestion games and by relating the local minima of the former to the Wardrop equilibria of the latter.

4.1 Problem Setup

Hereafter, we consider the following assumptions:

A1. $\forall i, j, l$, $w_{ji}^{(l)} \geq 0$ and $\forall i, l$, $\sum_j w_{ji}^{(l)} = 1$.

A2. $\mathcal{X} \subseteq \mathbb{R}_+^d$, ie, all learning samples are positive vectors.

A3. $\forall l \geq 2$, $n_l \geq C$, ie, all hidden layers are wider than the output layer.

A4. The loss can be written as:

$$\text{loss}(W) = \sum_j \sum_k \ell(o_k^j, y_k^j)$$

with o_k^j the value of the k^{th} output of the DNN for the instance x^j.

Regarding **A1**, one should note that the normalization of the weights have been commonly used both to study DNNs' properties and even to accelerate their training (see [Salimans and Kingma, 2016]). On the other hand, the non-negativity constraint, which at first glance might seem too restrictive, has also been used by [Gautier et al., 2016] where the authors empirically demonstrate the lack of its negative impact on the NNs' expressiveness. **A2** is naturally satisfied by numerous real-world data sets used to train DNNs, such as image collections or text corpora. **A3** implies for the hidden layers to be wider than the output layer and was used in [Nguyen and Hein, 2017] under the name of pyramidal structure assumption. While the importance of depth is often required for DNNs to have good approximation properties, the width of DNNs should be constrained to be wide enough to achieve disconnected decision regions as shown in [Nguyen et al., 2018]. Finally, **A4** restricts us to consider only those losses that are computed output-wise thus including many popular norm-based loss function. Note that considering this assumption is less restrictive than many other previous works on the subject that explicitly analyze only the least square loss.

4.2 Analysis of Linear DNNs

We start by proving a first result related to linear DNNs. We recall that while being quite restrictive in practice, the theoretical analysis of this setting is still challenging as it represents a non-convex optimization problem. Furthermore, we use it as a cornerstone result for our future developments as it allows to illustrate our proposed construction.

Lemma 1. *Assume A1-4, let DNN be defined as* $N = (V, E, I, O, F)$ *with* $F = \{f : \forall z, f(z) = z\}$, *let* $\text{loss}(\cdot)$ *be its associated loss function and let* $\mathcal{L} = \{(x^j, y^j)\}_{j=1}^M$ *be the learning sample. Then, one can construct a non-atomic congestion game* $\text{NCG}_N^{\text{loss}} = (E, c, S, n, a)$ *fully defined in terms of* N, $\text{loss}(\cdot)$ *and* $\mathcal{L}$.

Proof. We start by defining E of the corresponding congestion game by applying a set of predefined rules to the network N as follows:

1. Each edge of N becomes an edge of $\text{NCG}_N^{\text{loss}}$. Denote by B the set of these edges.
2. Each node of N with an activation function (nodes of the hidden layers and of the output layer) becomes an edge of the $\text{NCG}_N^{\text{loss}}$. Denote by J the set of these edges.
3. Each node of the output layer becomes a concatenation of M edges of $\text{NCG}_N^{\text{loss}}$ which are added to the edge of J so that the last edge points to a common node of the congestion game F. Denote by T the set of these edges such that $T = \{e_k^j : e_k^j$ associated to o_k for a tuple $(x^j, y^j)\}$. The index k is used for the output number which is considered $1 \leq k \leq C$ and for a fixed k the index j is the number of the arc which is considered $1 \leq j \leq M$. Let p_k be the set of the paths of the neural network (seen as a DAG) that include the concatenation of the M edges associated to output k.
4. The i^{th} node of the input layer become a node of the congestion game named d_i for $1 \leq i \leq d$.

An illustrative example of such transformation is given in Figure 2. We now define S, n, c and a as follows.

S,n. One population of players i is created for each node i of the input layer. The set of strategies of the player i, S_i, is the set of the paths from d_i to F. The size of the population i is 1 for each i, ie, $n_i = 1$.

c. We define the cost of the edges as follows:

$$c_e(\xi) = \begin{cases} 0, & \text{if } e \in B \cup J \\ \ell(\xi, y_k^j)/\xi, & \text{if } e = e_k^j \in T. \end{cases}$$

a. We define the rate of consumption as follows:

$$\forall S \in S_i,\ a_{S,e} = \begin{cases} 1, & \text{if } e \in B \cup J \\ x_i^j, & \text{if } e = e_k^j \in T. \end{cases}$$

The congestion game is entirely defined. □

Remark 1. *We implicitly assume that* $c_{e_k^j}(\xi) = \ell(\xi, y_k^j)/\xi$ *is continuous, non-negative and non-decreasing in* ξ *so as to respect the non-negativity and the non-decrease of the cost functions of the associated congestion game. Moreover, it is assumed that* $\ell(\xi, y_k^j)/\xi$ *is well defined in 0. We discuss later how these assumption can be shown to cover some loss functions used in practice.*

Let us now give the main theorem of the paper.

Theorem 1. *Under the assumptions of Lemma 1, let* $\ell(\xi, y_k^j) = A_k^j \xi^\beta$ *with* $A_k^j \geq 0$, $\beta \geq 2$. *Then, given a neural network* N, *every local minimum of the loss function* $\text{loss}(\cdot)$ *associated to* N *is a Wardrop equilibrium of the associated congestion game* $\text{NCG}_N^{\text{loss}}$.

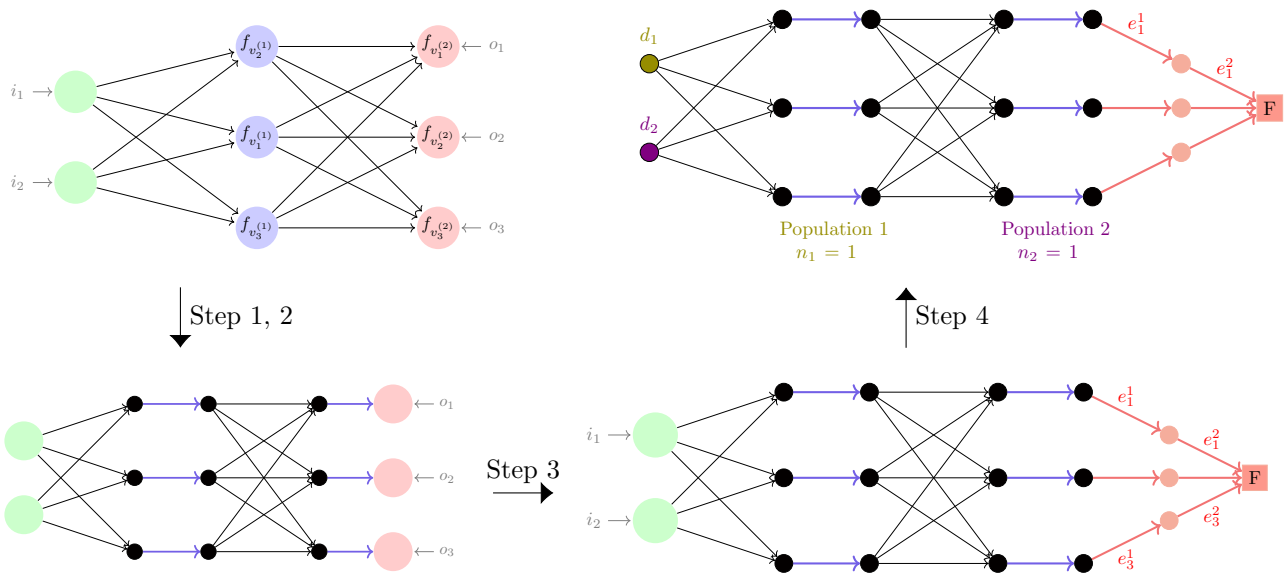

Figure 2: Illustration of how a non-atomic congestion game is constructed from a given DNN. **(upper left)** Example of a DNN with a number $d = 2$ of input neurons, $C = 3$ of output neurons and one hidden layer; **(upper right)** Graph associated to the DNN presented on the left for a learning sample of size $M = 2$. B is the set of black edges, J is the set of blue edges and T is the set of red edges. **(bottom left)** and **(bottom right)** are the intermediate steps.

Proof sketch. We start by relating the weight W of the neural network N to the flow in the associated congestion game such that the loss of the neural network becomes equal to the social cost of the associated congestion game. Then, we show how a local minimum W of the loss function induces a distribution z_W which is a Wardrop equilibrium of the associated congestion game. To this end, we use the result from [Kinderlehrer and Stampacchia, 2000] showing that every local minimum x^* of a function h belonging to class C^1 and defined on a closed and convex subset $X \subseteq \mathbb{R}^n$ verifies the following variational inequality:

$$\langle \nabla h(x^*), x - x^* \rangle \geq 0, \quad \forall x \in X.$$

On the other hand, we can characterize the Wardrop equilibrium of a non-atomic congestion game by proving that a distribution z^* is a Wardrop equilibrium of $\text{NCG}_N^{\text{loss}}$ if and only if, by denoting $c_k^j := c_{e_k^j}$ we have :

$$\sum_i \sum_k \sum_j x_i^j c_k^j(\tilde{z}_k^{j*})(z_{k,i} - z_{k,i}^*) \geq 0, \quad \forall z \in Z,$$

where $z_{k,i} = \sum_{S \in S_i \cap p_k} z_S$ with p_k being the set of paths which include $(e_k^{j'})_{1 \leq j' \leq M}$ and $\tilde{z}_k^j = \sum_i x_i^j z_{k,i}$. For the considered family of loss functions, we further show that variational characterization of a local minimum implies that of a Wardrop equilibrium. The desired result is obtained by establishing that the flow associated to a local minimum W is a Wardrop equilibrium of the associated game. □

This theorem is important as it allows to deduce two corollaries, one on neural networks and the second about congestion games.

Corollary 1. *Under the assumptions of Theorem 1, every local minimum of* loss$(\cdot)$ *is a global optimum.*

Proof. For a loss of this type, we have shown that a local minimum of the linear neural network is a Wardrop equilibrium of the associated congestion game. As a local minimum, the global minimum is also a Wardrop equilibrium. It is known that for non-atomic congestion games, the Wardrop equilibria have the same social cost. Because the value of the loss function at W is equal to the value of the social cost of the associated congestion game at z_W, local minimums and global ones have the same value which is the value of the Wardrop equilibria. □

We now comment on the differences between our results and those obtained in [Kawaguchi, 2016]. First, we note that Corollary 1 and Theorem 2.3 from [Kawaguchi, 2016] both establish the equivalence of local minima to the global optimum for linear DNNs (the same holds for Corollary 4 from our paper proved later for non-linear DNNs and Corollary 3.2 of [Kawaguchi, 2016]). This, however, is achieved under the assumptions that require 1) XX^T and YY^T to be full rank ($X \in \mathbb{R}^{d_x \times m}$ is the learning sample, $Y \in \mathbb{R}^{d_y \times m}$ are labels), 2) $YX^T(XX^T)^{-1}XY^T$ to have d_y distinct eigenvalues, and 3) $d_y \geq d_x$. One may note that these

assumptions cannot be compared directly with ours in general. For instance, our assumptions are satisfied for any positive data sample, a network with non-negative normalized weights for the task of binary classification with squared loss, while the assumptions of [Kawaguchi, 2016] are violated whenever the data fed to the neural-network lies in a low-dimensional manifold. As for the loss function considered, [Kawaguchi, 2016] considers squared loss only while our loss is more general and includes squared loss as a special case.

Corollary 2. *Under the assumptions of Theorem 1,* $\mathrm{PoA}(\mathrm{NCG}_N^{\mathrm{loss}}) = 1$.

Proof. A global minimum of the loss of N is a social optimum of the associated congestion game. Global minima are also Wardrop equilibria. Then, because the value of the loss function at W is equal to the value of the social cost of the associated congestion game at z_W, we get WE = SO and PoA = 1. □

4.3 Learning with Loss Function from Theorem 1

We now explain how we can use the loss studied in Theorem 1 in practice when dealing with classification task where y^j is a binary vector with only one coordinate equal to 1 and the rest being equal to 0. For each j, let us denote by e_j the coordinate of the vector y^j which is equal to 1, i.e., $y^j_{\mathsf{e}_j} = 1$ and $y^j_k = 0$ for $k \neq \mathsf{e}_j$. Moreover, we consider normalized inputs such that $\sum_i x_i^j = 1$ for all j. Then, we can use our loss functions in the following way:

1. We fix $\beta \geq 2$.
2. For each j, we impose $A^j_k = 1$ for $k \neq \mathsf{e}_j$ and $A^j_k = 0$ if $k = \mathsf{e}_j$. By this way, we penalize the outputs of the network which have to be equal to 0, while putting no penalty on the outputs that need to be equal to 1.

We can deduce the following corollary.

Corollary 3. *Under the assumptions of Theorem 1, let* $C = 2$ *and let* ℓ *be the squared loss. Then, a local minimum of* $\mathrm{loss}(\cdot)$ *is a global minimum.*

Proof sketch. We rewrite our loss and the squared loss and show that for $C > 2$ they differ by a constant. We then analyze this constant and prove that for $C = 2$ it reduces to the squared loss thus leading to the same optimization objective. □

4.4 Extension to DNNs with ReLUs

We now proceed to a study of non-linear DNNs with activation functions F given by ReLUs defined as:

$$f_v : \left| \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \max(0, x). \end{array} \right. \tag{2}$$

Such a non-linear DNN can be seen as a linear DNN where some paths of the graph underlying the network fail and thus can be also seen as non-atomic congestion game where edges of J, which represent the activation functions, can fail depending on the congestion that occurs on them. More precisely, one can show that for a non-linear DNN with ReLUs, its k^{th} output for the instance x^j is given by:

$$o^j_k = \sum_{p \in p_k} z^j_p x^j_p w_p,$$

where p_k is the set of paths that end to output k, w_p is the product of the weights on the path p and x^j_p is the value of the coordinate of x^j from which the path p starts. As for z^j_p, it is a variable that is equal to 1 if all ReLUs f_v encountered in the path p are such that $f_v(g^j_v) = g^j_v$ where g_v is the value of the node v on the example x^j and 0, otherwise. In other words, the variable z^j_p reflects whether the path p is active ($z^j_p = 1$) or not ($z^j_p = 0$) depending on the ReLU activation on the path p for the instance x^j.

As non-linear DNNs are notoriously hard to study, most papers introduce simplifications to model non-linearities in order to analyze the simplified models afterwards. One prominent example of such modelization was introduced by [Choromanska et al., 2014] and further improved by [Kawaguchi, 2016] who successfully discarded several of the unrealistic assumptions of the original model and lightened others. For our analysis, we use some of the lighter assumptions made in the latter paper. The first assumption, denoted by $A1p_m$ in the corresponding paper, states that z^j_p are Bernoulli random variables with the same probability of success ρ. The second assumption, called $A5u_m$, states that z^j_p are independent from the inputs $\{x^j\}_{j=1}^M$ and the weight parameters W. These assumptions, although remaining unrealistic in case of $A5u_m$, allow us to write the expected output o^j_k as follows:

$$\mathbb{E}(o^j_k) = \sum_{p \in p_k} \rho x^j_p w_p = \rho \sum_{p \in p_k} x^j_p w_p. \tag{3}$$

One can remark that the output of the network has simply been multiplied by ρ. Given a non-linear DNN N, let $\mathrm{lin}(N)$ be the linear DNN associated to N where all activation functions are replaced by the function:

$$f_v : \left| \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & x. \end{array} \right.$$

We now show how we can reduce non-linear DNNs to congestion games with failures by adapting the results obtained for atomic congestion games with failures in the paper [Li et al., 2017] to a non-atomic case.

Lemma 2. *Assume A1-4, A1p_m, A5u_m, let $N = (V, E, I, O, F)$ with O and F defined as in (3) and (2), respectively. Let* loss$(\cdot)$ *be its associated loss function and let $\mathcal{L} = \{(x^j, y^j)\}_{j=1}^M$ be the available learning sample. Then, N can be reduced to a non-atomic congestion game with failures* $\text{NCGF}_N^{\text{loss}} = (E, c, S, n, a)$ *fully defined in terms of N, $\mathcal{L}$ and* loss$(\cdot)$.

Proof. E, S, n, a remain the same as for $\text{NCG}_N^{\text{loss}}$ from Lemma 1. The only modification is that if a player chooses a path with no failures, then its cost is $c_S(z) = \sum_{e \in S} a_{S,e} c_e(z_e)$ where z is the flow of the game such that failures are taken into account. Otherwise, we impose $c_S(z) = w_i$ where w_i is a constant associated to a player type i. $\text{NCGF}_N^{\text{loss}}$ is now defined. □

Given W and our set of assumptions, we study the same loss as in the linear case but on the expected outputs of the neural networks, ie,

$$\text{loss}(W) = \sum_j \sum_k \ell(\mathbb{E}(o_k^j), y_k^j), \tag{4}$$

where o_k^j is the k^{th} output of N for instance j. We further let $\beta = 2$ in the definition of ℓ as done in [Kawaguchi, 2016] for the squared loss. We now establish the equivalence between the local minima of the non-linear model to those of the linear one.

Lemma 3. *Under the assumptions of Lemma 2, let $\beta = 2$ and let the loss function be as in (4). Then, a local minimum of* loss(W) *of N is a local minimum of the loss function of the corresponding linear network from Lemma 1.*

The final result obtained through the transitivity of properties shown in Theorem 1 is stated as follows.

Corollary 4. *Under the assumptions of Lemma 3, a local minimum of a loss function in (4) is a Wardrop equilibrium of the game associated to the linear network from Lemma 1.*

Note that out of 7 assumptions used in [Choromanska et al., 2014] to prove a similar result, we keep only two in addition to Assumptions A1-4. Also, their main result is different from ours as it states that the number of poor local minimum may be not too large, while our result states that there is no local minima in such setting. Finally, [Choromanska et al., 2014] consider squared loss with $d_y = 1$ (note that it does not correspond to commonly used one-hot label encodings). On a higher level, this corollary suggests that the model introduced for non-linear DNNs in [Choromanska et al., 2014] is equivalent to studying a linear DNN.

Overall, our contributions establish the state-of-the-art results related to the analysis of the loss surface of both linear and non-linear DNNs following a completely novel approach. While this is an important contribution in itself, we believe that it further paves the way to several other highly promising research directions presented below.

5 OPEN PROBLEMS

In the introduction we claimed that our analysis can be used as a tool to prove other fundamental results about DNNs by modelling them as congestion games. Below, we aim to highlight this claim by rigorously formulating three open problems for future research.

Impact of DNNs architecture Characterizing PoA depending on the network topology and the used cost function is commonly done in the field of congestion games and we expect it to be very useful for DNNs as well. Indeed, it is known that PoA of non-atomic congestion games is independent of network topology [Colini-Baldeschi et al., 2017] when cost functions are polynomials of an arbitrary degree. Consequently, one may wonder whether our results can be extended beyond multi-layer networks to take into account other network architectures, such as U-Nets [Ronneberger et al., 2015], with potentially different activation and/or loss functions. More formally, we propose the following open problem.

Open Problem 1. *For a DNN $N = (V, E, I, O, F)$, let* NCG *be a congestion game such that every local minimum/critical point and global minima of* loss$(\cdot)$ *associated to N are Wardrop equilibrium and social optimums of* NCG*, respectively. Then,* PoA(NCG) $=$ 1 *and does not depend on $G = (V, E)$, F and* loss$(\cdot)$.

We note that this open problem can also lead to a negative result where the local minima inefficiency can be proved to be arbitrary high. To obtain such result, one will have to consider *atomic* congestion games CG for which, contrary to Wardrop equilibria, different Nash equilibria can lead to outcomes with different costs. If there exists a DNN such that its associated atomic congestion game has PoA(CG) > 1 for a particular choice of G, F and loss$(\cdot)$, then such neural architectures may exhibit an arbitrary bad behaviour during the optimization. One such example is the PoA in atomic splittable games with polynomial cost functions of degree d for which there exists particular network topologies with PoA behaving as $((1+\sqrt{d+1})/2)^{d+1}$ [Roughgarden and Schoppmann, 2011].

Speed of convergence Apart from the geometric approach that relies on the equivalence between local minima and the global optimum, another way to analyze DNN's behaviour is to study directly the optimization dynamics of SGD and its variations. In the context of games, this latter can be modelled by repeating the one-shot game, i.e., the game with one data point or a mini batch, over T time steps and analyzing the average of the costs associated with the outcomes of each step. In our work, we provided a characterization of the PoA for one-shot non-atomic congestion game, but this analysis can be further applied to repeated games using the extension theorems and the notion of $(\lambda, \mu)-$smoothness developed in [Roughgarden, 2015]. In line with what we discussed above, we now consider atomic congestion games to allow for Nash equilibria with different costs. Such games can be characterized by the notion of a robust PoA ρ defined in terms of the parameters λ and μ and coinciding with the traditional PoA in several cases of interest. For these games, extension theorems ensure that the sequence of outcomes of a smooth game where every player experiences vanishing average (external) regret converges to the optimal outcome times the robust PoA. This is due to the fact that for every pure/mixed/correlated/coarse-correlated equilibrium z', $\frac{\mathbb{E}_{z\sim z'}[\mathrm{SC}(z)]}{\mathrm{SO(CG)}} \leq \rho$. Then, as a sequence of outcomes with vanishing external regret converges to a correlated equilibrium, we have the following.

Open Problem 2. *For a deep neural network $N = (V, E, I, O, F)$, let* CG *be its corresponding $(\lambda, \mu)-$smooth congestion game such that*

1. $\mathrm{loss}(W)$ *is equal to the social cost* $\mathrm{SC}(z_W)$.
2. $\mathrm{SO(CG)}$ *is the global optimum of* $\mathrm{loss}(\cdot)$.

Then, if $\mathrm{loss}(W^i)$ *are social costs associating to a vanishing average (external) regret, the following holds:*

$$\frac{1}{T}\sum_{i=1}^{T}\mathrm{loss}(W^i) \leq (\rho(\mathrm{CG}) + o(1))\mathrm{loss}(W^*) \text{ as } T \mapsto \infty$$

where W^i is the outcome of iteration i and W^* is a global optimum.

Moreover, if each critical point W of the loss function is either pure, mixed, correlated or coarse-correlated equilibrium of CG*, then* $\frac{\mathrm{loss}W}{\mathrm{loss}W^*} \leq \rho(\mathrm{CG})$.

Note that [Balduzzi, 2016] related the coarse-correlated equilibrium to critical points of a non-linear DNN, but their arguments were shown to be flawed in [Schuurmans and Zinkevich, 2016]. This latter work managed to draw the equivalence between critical points and Nash equilibrium correctly and used a popular regret matching algorithm to successfully learn DNNs. Unfortunately, their proposed learning strategy is applied to games solved on each vertex of a DNN and thus is not guaranteed to converge to a globally optimal strategy. In this regard, congestion games offer a more convenient alternative as they benefit from the convergence guarantees and allow to characterize the speed of this convergence based on the characteristics of the considered game.

Beyond backprop As discussed above, game-theoretical interpretation of DNNs had already led to new learning strategies used to find equilibrium points in the corresponding games. As shown in [Schuurmans and Zinkevich, 2016], regret matching algorithm outperforms widely-used SGD and Adam methods and leads to sparser networks with higher accuracy when deployed on the test set. In the context of our work, we would like to make a step further and go beyond the regret matching algorithm mentioned above by proposing a new learning strategy based on optimal transport [Villani, 2009]. Optimal transport considers a problem of transforming one probability measure into another following the principle of the least effort. While traditionally optimal transport does not account for congestion effects, several recent works studied this variation and showed that in this case the solution can be related to the notion of the equilibria of Wardrop type [Carlier et al., 2008, Blanchet and Carlier, 2016]. This leads to the following open problem.

Open Problem 3. *For a deep neural network $N = (V, E, I, O, F)$, let* NCG *be its corresponding congestion game with equilibria given by:*

$$\eta^* \in \mathrm{argmin}_{\eta} W_c(\mu, \eta) + \mathcal{E}(\eta), \tag{5}$$

where W_c is the Wasserstein distance between a measures defined on the input flow of NCG *and $\mathcal{E}(\eta)$ is a function of congestion for an action distribution η. Then, η^* is a critical point of* $\mathrm{loss}(\cdot)$.

We note that several papers [Chizat and Bach, 2018, Rotskoff and Vanden-Eijnden, 2018, Mei et al., 2018] used the notion of the Wasserstein gradient flow to show the convergence of convex optimization methods for overparametrized models. Their underlying idea was to consider a problem of learning a measure that minimizes the DNN's loss function and to study the dynamics of the gradient descent performed on its weights and positions. One question to answer thus is whether the Wasserstein gradient flow of solving (5) is naturally linked to the Wasserstein gradient flow considered in previous work? Such an equivalence may well indicate that the game-theoretical interpretation of DNNs reconciles both geometric approaches for studying DNNs loss surface and those based on analyzing their optimization dynamics.

References

[Allen-Zhu et al., 2019] Allen-Zhu, Z., Li, Y., and Song, Z. (2019). A convergence theory for deep learning via over-parameterization. In *ICML*, pages 242–252.

[Baldi and Hornik, 1989] Baldi, P. and Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Netw.*, 2(1):53–58.

[Balduzzi, 2016] Balduzzi, D. (2016). Deep online convex optimization with gated games.

[Blanchet and Carlier, 2016] Blanchet, A. and Carlier, G. (2016). Optimal transport and cournot-nash equilibria. *Math. Oper. Res.*, 41(1):125–145.

[Blum and Rivest, 1992] Blum, A. L. and Rivest, R. L. (1992). Training a 3-node neural network is np-complete. *Neural Networks*, 5(1):117 – 127.

[Brückner and Scheffer, 2011] Brückner, M. and Scheffer, T. (2011). Stackelberg games for adversarial prediction problems. In *ACM SIGKDD*, pages 547–555.

[Carlier et al., 2008] Carlier, G., Jimenez, C., and Santambrogio, F. (2008). Optimal transportation with traffic congestion and wardrop equilibria. *SIAM J. Control Optim.*, 47(3):1330–1350.

[Chizat and Bach, 2018] Chizat, L. and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. In *NeurIPS*, pages 3040–3050.

[Choromanska et al., 2014] Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. (2014). The loss surfaces of multilayer networks.

[Claus and Boutilier, 1998] Claus, C. and Boutilier, C. (1998). The dynamics of reinforcement learning in cooperative multiagent systems. In *AAAI*, pages 746–752.

[Colini-Baldeschi et al., 2017] Colini-Baldeschi, R., Cominetti, R., Mertikopoulos, P., and Scarsini, M. (2017). The asymptotic behavior of the price of anarchy. In *WINE*, pages 133–145.

[Dauphin et al., 2014] Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *NIPS*, page 2933–2941.

[Dritsoula et al., 2017] Dritsoula, L., Loiseau, P., and Musacchio, J. (2017). A game-theoretic analysis of adversarial classification. *IEEE Transactions on Information Forensics and Security*, 12(12):3094–3109.

[Du et al., 2017] Du, S. S., Jin, C., Lee, J. D., Jordan, M. I., Singh, A., and Póczos, B. (2017). Gradient descent can take exponential time to escape saddle points. In *NIPS*, pages 1067–1077.

[Freeman and Bruna, 2017] Freeman, C. D. and Bruna, J. (2017). Topology and geometry of half-rectified network optimization. In *ICLR*.

[Freund and Schapire, 1996] Freund, Y. and Schapire, R. E. (1996). Game theory, on-line prediction and boosting. In *COLT*, page 325–332.

[Gautier et al., 2016] Gautier, A., Nguyen, Q., and Hein, M. (2016). Globally optimal training of generalized polynomial neural networks with nonlinear spectral methods.

[Ge et al., 2015] Ge, R., Huang, F., Jin, C., and Yuan, Y. (2015). Escaping from saddle points - online stochastic gradient for tensor decomposition. In *COLT*, pages 797–842.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

[Goodfellow and Vinyals, 2015] Goodfellow, I. J. and Vinyals, O. (2015). Qualitatively characterizing neural network optimization problems. In *ICLR*.

[Hardt and Ma, 2017] Hardt, M. and Ma, T. (2017). Identity matters in deep learning. In *ICLR*.

[Hu and Wellman, 2003] Hu, J. and Wellman, M. P. (2003). Nash q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, 4:1039–1069.

[Jin et al., 2017] Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. (2017). How to escape saddle points efficiently. In *ICML*, pages 1724–1732.

[Kawaguchi, 2016] Kawaguchi, K. (2016). Deep learning without poor local minima.

[Kinderlehrer and Stampacchia, 2000] Kinderlehrer, D. and Stampacchia, G. (2000). *An Introduction to Variational Inequalities and Their Applications*. SIAM.

[Li et al., 2017] Li, Y., Jia, Y., Tan, H., Wang, R., Han, Z., and Lau, F. (2017). Congestion game with agent and resource failures. *IEEE Journal on Selected Areas in Communications*, PP:1–1.

[Liu and Chawla, 2009] Liu, W. and Chawla, S. (2009). A game theoretical model for adversarial learning. In *ICDM Workshops*, pages 25–30.

[Mei et al., 2018] Mei, S., Montanari, A., and Nguyen, P. (2018). A mean field view of the landscape of two-layers neural networks. *CoRR*, abs/1804.06561.

[Nguyen and Hein, 2017] Nguyen, Q. and Hein, M. (2017). The loss surface of deep and wide neural networks. In *ICML*, pages 2603–2612.

[Nguyen et al., 2018] Nguyen, Q., Mukkamala, M. C., and Hein, M. (2018). Neural networks should be wide enough to learn disconnected decision regions. In *ICML*, pages 3737–3746.

[Peshkin et al., 2000] Peshkin, L., Kim, K.-E., Meuleau, N., and Kaelbling, L. P. (2000). Learning to cooperate via policy search. In *UAI*, pages 489–496.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241.

[Rosenthal, 1973] Rosenthal, R. W. (1973). A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory*, 2:65–67.

[Rotskoff and Vanden-Eijnden, 2018] Rotskoff, G. M. and Vanden-Eijnden, E. (2018). Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *CoRR*, abs/1805.00915.

[Roughgarden, 2015] Roughgarden, T. (2015). Intrinsic robustness of the price of anarchy. *J. ACM*, 62(5).

[Roughgarden and Schoppmann, 2011] Roughgarden, T. and Schoppmann, F. (2011). Local smoothness and the price of anarchy in atomic splittable congestion games. In *SODA*, SODA '11, page 255–267.

[Roughgarden and Éva Tardos, 2004] Roughgarden, T. and Éva Tardos (2004). Bounding the inefficiency of equilibria in nonatomic congestion games. *Games and Economic Behavior*, 47(2):389 – 403.

[Safran and Shamir, 2018] Safran, I. and Shamir, O. (2018). Spurious local minima are common in two-layer relu neural networks. In *ICML*, pages 4430–4438.

[Salimans and Kingma, 2016] Salimans, T. and Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *NIPS*, pages 901–909.

[Schmeidler, 1973] Schmeidler, D. (1973). Equilibrium points of nonatomic games. *Journal of Statistical Physics*, 7(4):295–300.

[Schuurmans and Zinkevich, 2016] Schuurmans, D. and Zinkevich, M. (2016). Deep learning games. In *NIPS*, pages 1678–1686.

[Soudry and Hoffer, 2018] Soudry, D. and Hoffer, E. (2018). Exponentially vanishing sub-optimal local minima in multilayer neural networks. In *ICLR, Workshop Track*.

[Villani, 2009] Villani, C. (2009). *Optimal transport: old and new*. Springer, Berlin.