Deep Neural Networks Are Congestion Games: Supplementary Materials

Deep Neural Networks (DNN) We start with a quick recall about the notations and the background used in the main paper. To this end, we consider a deep neural network which architecture is specified by $N = (V, E, I, O, F)$, where:

- V is a set of vertices, i.e., the total number of units in the neural network;
- $E \subseteq V \times V$ is a set of edges;
- $I = \{i_1, \ldots, i_d\} \subset V$ is a set of input vertices;
- $O = \{o_1, \ldots, o_C\} \subset V$ is a set of output vertices
- $F = \{f_v : v \in V\}$ 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. The network is related to the training data by assuming $|I| = d$, the number of input vertices corresponds to the number of input features, and $|O| = C$, the number of output vertices corresponds to the number of output dimensions. We let n_l denote the number of neurons at each layer $l \in [1, \ldots, L]$ 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 $z_i^{(l)} = \sum_{k=1}^{n_{l-1}} w_{ik}^{(l)} g_k^{(l-1)}$.

Given a training set $S = \{x^j, y^j\}_{j=1}^M$ 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 $h : \mathcal{X} \to Y$ that assigns a label close to $y^j \in Y$ to each $x^j \in \mathcal{X}$. This is done by solving the following optimization problem:

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

where $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is some convex loss function such as squared ℓ_2 norm. Note that one usually uses stochastic gradient descent to optimize the loss function of a neural network where the weights are updated either for each example x or for a mini-batch of examples.

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 **??** 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.

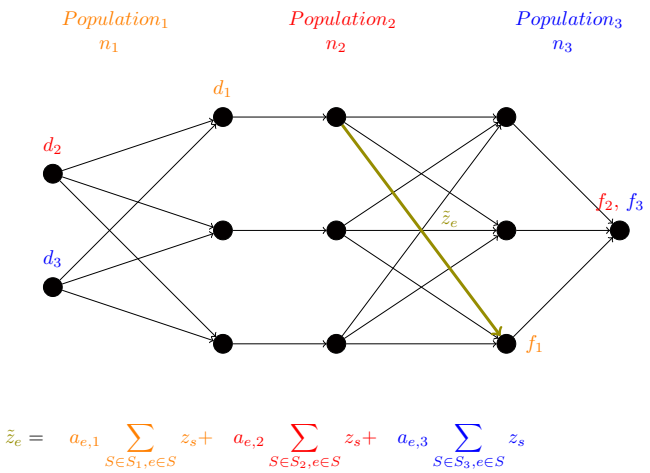

Figure 1: Example of a non-atomic congestion game with three populations of players

- 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 $a_{S,e} = 0$ 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 $\mathrm{SC}(z)$ can be rewritten as:

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

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:

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

Our first result for which the full proof is given in the main paper is stated below.

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* $\mathrm{loss}(\cdot)$ *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* $\mathrm{NCG}_N^{\mathrm{loss}} = (E, c, S, n, a)$ *fully defined in terms of N,* $\mathrm{loss}(\cdot)$ *and $\mathcal{L}$.*

FULL PROOF OF THEOREM 1

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* $\mathrm{loss}(\cdot)$ *associated to N is a Wardrop equilibrium of the associated congestion game* $\mathrm{NCG}_N^{\mathrm{loss}}$.

Proof. In what follows, we write $c_k^j := c_{e_k^j}$ and similarly for all quantities that have such subscripts to avoid cumbersome notations. The proof of this theorem relies on several lemmas with the first one showing how the weight matrix W of the neural network N is related 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.

Lemma. *Under the assumption of Lemma 1, a configuration W of the neural network N defines an action distribution z_W of the associated congestion game* $\mathrm{NCG}_N^{\mathrm{loss}}$ *such that* $\mathrm{loss}(W) = \mathrm{SC}(z_W)$. *Similarly, for every action distribution z_W of the associated congestion game* $\mathrm{NCG}_N^{\mathrm{loss}}$ *there exists a set of weights W of N such that* $\mathrm{loss}(W) = \mathrm{SC}(z_W)$.

Proof. ($\longrightarrow$) Given a flow z, the social cost of a congestion game can be written as:

$$\mathrm{SC}(z) = \sum_{e \in E} c_e(\tilde{z}_e)\tilde{z}_e.$$

For the studied congestion games, the following holds:

$$\begin{aligned}\mathrm{SC}(z) &= \sum_{e \in E} c_e(\tilde{z}_e)\tilde{z}_e \\ &= \sum_k \sum_j c_k^j(\tilde{z}_k^j)\tilde{z}_k^j\end{aligned}$$

$$= \sum_k \sum_j \ell(\tilde{z}_k^j, y_k^j).$$

Given a set of weights W of a given network N, the loss of the network can be written as:

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

with $b_{k,i} = \sum_{p \in S_i \cap p_k} w_p$ where w_p is the product of the weights encountered in the path p and p_k is the set of paths that include the M edges associated to the output k, $\tilde{b}_k^j = \sum_i x_i^j b_{k,i}$.

In order to relate $\tilde{z}_k^j$ and $\tilde{b}_k^j$, we let $k \in [1, \dots, C]$ and note that a player of type i who wishes to travel on the edge e_k^j travels on all the edges $(e_k^{j'})_{1 \leq j' \leq M}$. Thus, the measure of population i using the edge e_k^j is equal to the measure of population i that uses the edge $e_k^{j'}$ and is denoted by $z_{k,i} = \sum_{S \in S_i \cap p_k} z_S$ where p_k is the set of paths which include $(e_k^{j'})_{1 \leq j' \leq M}$. One can verify that

$$\tilde{z}_k^j = \sum_i x_i^j z_{k,i}.$$

Relating $\tilde{z}_k^j$ and $\tilde{b}_k^j$ now boils down to establishing the equality between $z_{k,i}$ and $b_{k,i}$. To this end, we note that for an action distribution z defined such that for each i and $S \in S_i$, $z_S = w_S$ with w_S being the product of the weights encountered in the path S, $z_{k,i} = \sum_{S \in S_i \cap p_k} z_S = \sum_{S \in S_i \cap p_k} w_S = b_{k,i}$ which implies $\tilde{z}_k^j = \tilde{b}_k^j$ so that

$$\begin{aligned} \text{loss}(W) &= \sum_j \sum_k \ell(\tilde{b}_k^j, y_k^j) = \sum_j \sum_k \ell(\tilde{z}_k^j, y_k^j) \\ &= \text{SC}(z_W). \end{aligned}$$

Note that z is a valid distribution since $\sum_{S \in S_i} z_S = 1$ as each subgraph of the network with d_i as root is a probability tree with non-negative and normalized weights.

($\longleftarrow$) We prove the second statement of this lemma below. From the Assumption 1, it follows that all the matrices we consider in this proof are non-negative and verify that the sum of the coefficients on each column is equal to 1. Let $w^{(1)}, ..., w^{(L)}$ be the matrices of weights of the neural network such that, for $1 \leq \ell \leq L$, $w^{(\ell)}$ is the matrix of the weights which stands between layer $\ell - 1$ and layer ℓ. Concerning the dimensions of the matrices, we have that $w^{(1)} \in M_{n_1,d}(\mathbb{R}_+)$ and for $1 \leq \ell \leq L$, $w^{(\ell)} \in M_{n_\ell, n_{\ell-1}}(\mathbb{R}_+)$. Let $w' \in M_{C,d}(\mathbb{R}_+)$ be the following matrix: $w' = (z_{k,i})_{\substack{1 \leq k \leq C \\ 1 \leq i \leq d}}$. Let w'' be the matrix such that $w'' = (b_{k,i})_{\substack{1 \leq k \leq C \\ 1 \leq i \leq d}}$. One can verify that $w'' = w^{(L)} * ... * w^{(1)}$ Now, we see that our problem boils down to whether for each matrix w', there exists $w^{(1)}, ..., w^{(L)}$ such that $w' = w^{(L)} * ... * w^{(1)}$. We will show that the answer is positive by using the fact that each matrix $w^{(\ell)}$ has dimensions superior to C due to the Assumption 3. Let w' be a non-negative and normalized matrix. Let us take $w^{(1)}, ..., w^{(L)}$ such that:

$$w^{(1)} = \begin{pmatrix} 0 \\ w' \end{pmatrix}, \quad \text{and for } \ell = 2, \dots, L: \quad w^{(\ell)} = \begin{pmatrix} H_\ell & 0 \\ 0 & I_C \end{pmatrix} \quad \text{with } H_\ell = \begin{pmatrix} 1 & \dots & 1 \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{pmatrix}, \ I_C \in M_C(\mathbb{R}).$$

One can verify that $w^{(1)}, ..., w^{(L)}$ are non-negative and normalized matrices. Then, we have $w^{(1)}, ..., w^{(L)}$ admissible matrices such that $w' = w^{(L)} * ... * w^{(1)}$. □

We now proceed by showing 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:

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

Given W, the loss of the network only depends on $\{b_{k,i}\}$ (see lemma above) which induces the outputs $\{\tilde{b}_k^j\}$. Then, we can define the loss on the set of admissible families $B = \{b_{k,i}\}$ and verify that B is convex and closed. Given the particular shape of the studied loss functions, one can also verify that the loss is C^1 on a neighborhood of B. Then, we get that every local minimum b^* of the loss function on B verifies the variational inequality:

$$\langle \nabla \text{loss}(b^*), b - b^* \rangle \geq 0, \quad \forall b \in B.$$

We can now prove the following lemma.

Lemma. *For any b and b^* in B, the following holds:*

$$\begin{aligned}\langle \nabla \text{loss}(b^*), b - b^* \rangle = &\sum_i \sum_k \sum_j x_i^j c_k^{j\prime}(\tilde{b}_k^{j*}) \tilde{b}_k^{j*} (b_{k,i} - b_{k,i}^*) \\ &+ \sum_i \sum_k \sum_j x_i^j c_k^j (\tilde{b}_k^{j*})(b_{k,i} - b_{k,i}^*).\end{aligned}$$

Proof. Be $b \in B$, we have:

$$\begin{aligned}\text{loss}(b) &= \sum_k \sum_j \ell(\tilde{b}_k^j, y_k^j) \\ &= \sum_k \sum_j c_k^j(\tilde{b}_k^j)\tilde{b}_k^j\end{aligned}$$

with $\tilde{b}_k^j = \sum_i x_i^j b_{k,i}$. For $b \in B$ we can compute that:

$$\begin{aligned}\frac{\partial \text{loss}}{\partial b_{k,i}}(b) &= \sum_j \frac{\partial c_k^j}{\partial b_{k,i}}(\tilde{b}_k^j)\tilde{b}_k^j + c_k^j(\tilde{b}_k^j)\frac{\partial \tilde{b}_k^j}{\partial b_{k,i}} \\ &= \sum_j c_k^{j\prime}(\tilde{b}_k^j) x_i^j \tilde{b}_k^j + c_k^j(\tilde{b}_k^j) x_i^j \\ &= \sum_j c_k^{j\prime}(\tilde{b}_k^j) x_i^j \tilde{b}_k^j + \sum_j c_k^j(\tilde{b}_k^j) x_i^j.\end{aligned}$$

For $b \in B$ and $b^* \in B$, the previous calculations lead to the desired result.

$$\begin{aligned}\langle \nabla \text{loss}(b^*), b - b^* \rangle &= \sum_i \sum_k \frac{\partial \text{loss}}{\partial b_{k,i}}(b^*)(b_{k,i} - b_{k,i}^*) \\ &= \sum_i \sum_k \sum_j x_i^j c_k^{j\prime}(\tilde{b}_k^{j*}) \tilde{b}_k^{j*} (b_{k,i} - b_{k,i}^*) + \sum_i \sum_k \sum_j x_i^j c_k^j (\tilde{b}_k^{j*})(b_{k,i} - b_{k,i}^*).\end{aligned}$$

□

On the other hand, we can characterize the Wardrop equilibrium of a non-atomic congestion game using the following variational inequality. Let Z be the set of admissible flows for $\text{NCG}_N^{\text{loss}}$.

Lemma. *A distribution z^* is a Wardrop equilibrium of $\text{NCG}_N^{\text{loss}}$ if and only if:*

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

Proof. ($\longrightarrow$) In the following, we will use [Roughgarden and Éva Tardos, 2004, Proposition 2.7]. It states that if z is a Wardrop equilibrium, then for each player type i there is a real number $c_i(z)$ such that all strategies $S \in S_i$ with $z_S > 0$ verify $c_S(z) = c_i(z)$. Then, the social cost of z is:

$$\text{SC}(z) = \sum_{i=1}^{d} c_i(z) n_i.$$

We can add that if z is a Wardrop equilibrium, then for each player type i and strategy $S \in S_i$, if $z_S = 0$ then $c_S(z) \geq c_i(z)$. This addition is immediate from the definition of the Wardrop equilibrium because it would be in the interest of the players of type i to use the strategy S otherwise.

We will now apply this proposition to our case. Let z^* be a Wardrop equilibrum of $\text{NCG}_N^{\text{loss}}$. For each player type i, there exists a number $c_i(z^*)$ such that for all k, if $z^*_{k,i} > 0$ then

$$\forall S \in S_i \cap p_k, c_S(z^*) = \sum_j x_i^j c_k^j(\tilde{z}_k^{j*}) = c_i(z^*).$$

Moreover, if $z^*_{k,i} = 0$ then for all $S \in S_i \cap p_k$, we have

$$c_S(z^*) = \sum_j x_i^j c_k^j(\tilde{z}_k^{j*}) \geq c_i(z^*)$$

and in general

$$\sum_j x_i^j c_k^j(\tilde{z}_k^{j*}) \geq c_i(z^*).$$

We can compute that:

$$\begin{aligned}
\sum_i \sum_k \sum_j x_i^j c_k^j(\tilde{z}_k^{j*})(z_{k,i} - z^*_{k,i}) &= \sum_i \sum_k \sum_j x_i^j c_k^j(\tilde{z}_k^{j*}) z_{k,i} - \sum_i \sum_k \sum_j x_i^j c_k^j(\tilde{z}_k^{j*}) z^*_{k,i} \\
&= \sum_i \sum_k z_{k,i} \sum_j x_i^j c_k^j(\tilde{z}_k^{j*}) - \sum_k \sum_j c_k^j(\tilde{z}_k^{j*}) \sum_i x_i^j z^*_{k,i} \\
&= \sum_i \sum_k z_{k,i} \sum_j x_i^j c_k^j(\tilde{z}_k^{j*}) - \sum_k \sum_j c_k^j(\tilde{z}_k^{j*}) \tilde{z}_k^{j*} \\
&= \sum_i \sum_k z_{k,i} \sum_j x_i^j c_k^j(\tilde{z}_k^{j*}) - \text{SC}(z^*) \\
&= \sum_i \sum_k z_{k,i} \sum_j x_i^j c_k^j(\tilde{z}_k^{j*}) - \sum_i c_i(z^*) n_i \\
&\geq \sum_i \sum_k z_{k,i} c_i(z^*) - \sum_i c_i(z^*) n_i \\
&= \sum_i c_i(z^*) \sum_k z_{k,i} - \sum_i c_i(z^*) n_i \\
&= \sum_i c_i(z^*) n_i - \sum_i c_i(z^*) n_i \\
&= 0.
\end{aligned}$$

($\longleftarrow$) Let z^* be a distribution that verifies

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

By contradiction, let us suppose that z^* is not a Wardrop equilibrium. Then by definition, there exists a player type i and two strategies $S_1 \in S_i$, $S_2 \in S_i$ such that $z^*_{S_1} > 0$ and $c_{S_1}(z^*) > c_{S_2}(z^*)$. We construct the distribution z such that $z_S = z^*_S$ for all S such that $S \neq S_1$ and $S \neq S_2$. We impose $z_{S_1} = 0$ and $z_{S_2} = z^*_{S_1} + z^*_{S_2}$. One can verify that z is an acceptable distribution. Let us suppose that $S_1 \in p_k$ and $S_2 \in p_{k'}$. Then, we have $k \neq k'$ because $c_{S_1}(z^*) > c_{S_2}(z^*)$ (otherwise we would have $c_{S_1}(z^*) = c_{S_2}(z^*)$). We also have that $z_{k',i} = z^*_{k',i} + z^*_{S_1}$ while $z_{k,i} = z^*_{k,i} - z^*_{S_1}$. Furthermore, it holds that $c_{S_1}(z^*) = \sum_j x_i^j c_k^j(\tilde{z}_k^{j*}) > c_{S_2}(z^*) = \sum_j x_i^j c_{k'}^j(\tilde{z}_{k'}^{j*})$. If $i'' \neq i$ or $k'' \neq k \neq k'$, we have $z_{k'',i''} - z^*_{k'',i''} = 0$. We can compute that:

$$\begin{aligned}
\sum_i \sum_k \sum_j x_i^j c_k^j(\tilde{z}_k^{j*})(z_{k,i} - z^*_{k,i}) &= \sum_j x_i^j c_{k'}^j(\tilde{z}_{k'}^{j*})(z_{k',i} - z^*_{k',i}) + \sum_j x_i^j c_k^j(\tilde{z}_k^{j*})(z_{k,i} - z^*_{k,i}) \\
&= \sum_j x_i^j c_{k'}^j(\tilde{z}_{k'}^{j*}) z^*_{S_1} + \sum_j x_i^j c_k^j(\tilde{z}_k^{j*})(-z^*_{S_1})
\end{aligned}$$

$$
\begin{aligned}
&= c_{S_2}(z^*)z^*_{S_1} - c_{S_1}(z^*)z^*_{S_1} \\
&= z^*_{S_1}(c_{S_2}(z^*) - c_{S_1}(z^*)) \\
&< 0
\end{aligned}
$$

that leads to a contradiction. □

Let us remind that we consider loss functions of the form $\text{loss}(W) = \sum_j \sum_k \ell(\xi, y^j_k)$ where $\ell(\xi, y^j_k) = A^j_k \xi^\beta$ with $A^j_k \geq 0$, $\beta \geq 2$. For such loss functions, we can establish the following result.

Lemma. *Let* $\text{loss}(W) = \sum_j \sum_k \ell(\xi, y^j_k)$ *where* $\ell(\xi, y^j_k) = A^j_k \xi^\beta$ *with* $A^j_k \geq 0$, $\beta \geq 2$ *and* $b^* \in B$. *Then,*

$$\forall b \in B,\ \langle \nabla \text{loss}(b^*), b - b^* \rangle \geq 0 \text{ if and only if } \forall z \in Z,\ \sum_i \sum_k \sum_j x^j_i c^j_k(\tilde{b}^{j*}_k)(z_{k,i} - b^*_{k,i}) \geq 0.$$

Proof. We start by showing that the considered loss functions ensure that the cost functions of the associated congestion game are solutions of the differential equation:

$$(\beta - 1)c^j_k(t) = c^{j\prime}_k(t)t.$$

In fact, the solutions of the equation $(\beta - 1)U(\xi) = U'(\xi)\xi$ which verify $U(0) = 0$ are functions U such that $U(\xi) = A_U \xi^{(\beta-1)}$ with $\beta \geq 2$. We restrict the set of solutions to functions U such that $A_U \geq 0$ so that U is non-decreasing. Then we have, for all k and j, $c^j_k(\xi) = A^j_k \xi^{\beta-1}$ with $A^j_k > 0$, $\beta \geq 2$. The fact that $\ell(\xi, y^j_k) = c^j_k(\xi)\xi$ imposes that ℓ has the form: $\ell(\xi, y^j_k) = A^j_k \xi^\beta$ with $A^j_k > 0$, $\beta \geq 2$. The loss function $\text{loss}(W) = \sum_j \sum_k \ell(\xi, y^j_k)$ where $\ell(\xi, y^j_k) = A^j_k \xi^\beta$ with $A^j_k \geq 0$ respects all the conditions, i.e., $\frac{\ell}{x}$ is non-decreasing, non-negative and continuous. It also allows the cost functions of the associated congestion game to verify the differential equation written above. For this type of loss, we can rewrite the condition:

$$
\begin{aligned}
\langle \nabla \text{loss}(b^*), b - b^* \rangle &= \sum_i \sum_k \sum_j x^j_i c^{j\prime}_k(\tilde{b}^{j*}_k)\tilde{b}^{j*}_k(b_{k,i} - b^*_{k,i}) + \sum_i \sum_k \sum_j x^j_i c^j_k(\tilde{b}^{j*}_k)(b_{k,i} - b^*_{k,i}) \\
&= \sum_i \sum_k \sum_j x^j_i (\beta - 1) c^j_k(\tilde{b}^{j*}_k)(b_{k,i} - b^*_{k,i}) + \sum_i \sum_k \sum_j x^j_i c^j_k(\tilde{b}^{j*}_k)(b_{k,i} - b^*_{k,i}) \\
&= \beta \sum_i \sum_k \sum_j x^j_i c^j_k(\tilde{b}^{j*}_k)(b_{k,i} - b^*_{k,i}).
\end{aligned}
$$

Then, using the fact that $B = Z$ (proved in the first lemma of the proof of Theorem 1), we have:

$$\forall b \in B,\ \langle \nabla \text{loss}(b^*), b - b^* \rangle \geq 0 \iff \forall z \in Z,\ \sum_i \sum_k \sum_j x^j_i c^j_k(\tilde{b}^{j*}_k)(z_{k,i} - b^*_{k,i}) \geq 0.$$

□

The proof of the theorem is now straightforward. If W is such that the family b^* induced by W is a local minimum of the loss function $\text{loss}(W)$ of N, then

$$\langle \nabla \text{loss}(b^*), b - b^* \rangle \geq 0, \quad \forall b \in B$$

and

$$\sum_i \sum_k \sum_j x^j_i c^j_k(\tilde{b}^{j*}_k)(z_{k,i} - b^*_{k,i}) \geq 0, \quad \forall z \in Z$$

which implies that z_W, the flow associated to W such that $z^*_{k,i} = b^*_{k,i}$, is a Wardrop equilibrium of the associated congestion game. This concludes the proof of the main theorem. □

PROOF OF COROLLARY 3

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

Proof. For a set of weights W, we can express the squared loss denoted by Sloss as follows:

$$S\text{loss}(W) = \sum_j \left(\sum_{k \neq \mathsf{e}_j} \left(\tilde{b}^j_k\right)^2 + \left(1 - \tilde{b}^j_{\mathsf{e}_j}\right)^2 \right)$$

while for loss$(\cdot)$ we have:

$$\text{loss}(W) = \sum_j \sum_{k \neq \mathsf{e}_j} (\tilde{b}^j_k)^2.$$

One can see that:

$$S\text{loss}(W) = \text{loss}(W) + const,$$

where $const = \sum_j (1 - \tilde{b}^j_{\mathsf{e}_j})^2$. The squared loss penalizes the outputs equal to 1. Let us rewrite *const* recalling that $\sum_i x^j_i = 1$ and $\sum_k b_{k,i} = 1$.

$$\begin{aligned}
\tilde{b}^j_{\mathsf{e}_j} &= \sum_i x^j_i b_{\mathsf{e}_j,i} = \sum_i x^j_i \left(1 - \sum_{k \neq \mathsf{e}_j} b_{k,i}\right) \\
&= \sum_i x^j_i - \sum_i x^j_i \sum_{k \neq \mathsf{e}_j} b_{k,i} = 1 - \sum_{k \neq \mathsf{e}_j} \sum_i b_{k,i} x^j_i \\
&= 1 - \sum_{k \neq \mathsf{e}_j} \tilde{b}^j_k.
\end{aligned}$$

Then,

$$\begin{aligned}
const &= \sum_j (1 - \tilde{b}^j_{\mathsf{e}_j})^2 \\
&= \sum_j \left(1 - \left(1 - \sum_{k \neq \mathsf{e}_j} \tilde{b}^j_k\right)\right)^2 \\
&= \sum_j \left(\sum_{k \neq \mathsf{e}_j} \tilde{b}^j_k\right)^2.
\end{aligned}$$

Finally, we have that

$$S\text{loss}(W) = \text{loss}(W) + \sum_j \left(\sum_{k \neq \mathsf{e}_j} \tilde{b}^j_k\right)^2.$$

In a general case of classification with a number of classes $C \geq 3$, nothing guarantees that Sloss$(\cdot)$ and loss$(\cdot)$ have the same local minima. However, in the case of binary classification with $C = 2$, we have that:

$$const = \sum_j \left(\sum_{k \neq \mathsf{e}_j} \tilde{b}^j_k\right)^2 = \sum_j \sum_{k \neq \mathsf{e}_j} (\tilde{b}^j_k)^2 = \text{loss}(W).$$

This result is obtained due to the fact that the set $\{k \neq \mathsf{e}_j\}$ has an only element ($1 \leq k \leq C = 2$). This implies that $S\text{loss}(W) = 2\text{loss}(W)$ and we have that Sloss and loss have the same local minima. □

PROOF OF LEMMA 3

We now briefly recall the background on non-linear DNNs with activation functions F given by ReLUs defined as:

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

One can show that for a non-linear DNN with ReLUs, its k^{th} output for the instance x^j is given by:

$$o_k^j = \sum_{p \in p_k} z_p^j x_p^j w_p,$$

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_p^j is the value of the coordinate of x^j from which the path p starts. As for z_p^j, it is a variable that is equal to 1 if all ReLUs f_v encountered in the path p are such that $f_v(g_v^j) = g_v^j$ where g_v is the value of the node v on the example x^j and 0, otherwise. In other words, the variable z_p^j reflects whether the path p is active ($z_p^j = 1$) or not ($z_p^j = 0$) depending on the ReLU activation on the path p for the instance x^j.

One prominent example of non-linear DNN's 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_p^j are Bernoulli random variables with the same probability of success ρ. The second assumption, called $A5u_m$, states that z_p^j 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_k^j as follows:

$$\mathbb{E}(o_k^j) = \sum_{p \in p_k} \rho x_p^j w_p = \rho \sum_{p \in p_k} x_p^j w_p. \tag{3}$$

One can remark that the output of the network has simply been multiplied by ρ. Given a non-linear DNN N, let $\text{lin}(N)$ 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 $\text{loss}(\cdot)$ be its associated loss function and let $\{(x^j, y^j)\}_{j=1}^M$ 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 and $\text{loss}(\cdot)$.*

The proof of this lemma can be found in the paper itself. 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 the loss function be as in (4). Then, a local minimum of $\text{loss}(W)$ of N is a local minimum of the loss function of the corresponding linear network from Lemma 1.*

Proof. Under the assumptions of Lemma 3, the loss of the non-linear DNN can be written as:

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

We further have that

$$\mathbb{E}(o_k^j) = \rho \sum_{p \in p_k} x_p^j w_p = \rho \tilde{b}_k^j.$$

Then,

$$\begin{aligned}
\text{loss}(W) &= \sum_j \sum_k \ell(\rho\, \tilde{b}_k^j, y_k^j) \\
&= \sum_j \sum_k A_k^j (\rho\, \tilde{b}_k^j)^\beta \\
&= \rho^\beta \sum_j \sum_k A_k^j (\tilde{b}_k^j)^\beta \\
&= \rho^\beta \text{loss}'(W),
\end{aligned}$$

where $\text{loss}'(W)$ is the loss of the linear network associated to N. It follows that loss and loss' have the same local minimums. □

References

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

[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.

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

[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.

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