

Article

# Multi-Task Deep Learning Games: Investigating Nash Equilibria and Convergence Properties

Minhyeok Lee 

School of Electrical and Electronics Engineering, Chung-Ang University, Seoul 06974, Republic of Korea; mlee@cau.ac.kr

**Abstract:** This paper conducts a rigorous game-theoretic analysis on multi-task deep learning, providing mathematical insights into the dynamics and interactions of tasks within these models. Multi-task deep learning has attracted significant attention in recent years due to its ability to leverage shared representations across multiple correlated tasks, leading to improved generalization and reduced training time. However, understanding and examining the interactions between tasks within a multi-task deep learning system poses a considerable challenge. In this paper, we present a game-theoretic investigation of multi-task deep learning, focusing on the existence and convergence of Nash equilibria. Game theory provides a suitable framework for modeling the interactions among various tasks in a multi-task deep learning system, as it captures the strategic behavior of learning agents sharing a common set of parameters. Our primary contributions include: casting the multi-task deep learning problem as a game where each task acts as a player aiming to minimize its task-specific loss function; introducing the notion of a Nash equilibrium for the multi-task deep learning game; demonstrating the existence of at least one Nash equilibrium under specific convexity and Lipschitz continuity assumptions for the loss functions; examining the convergence characteristics of the Nash equilibrium; and providing a comprehensive analysis of the implications and limitations of our theoretical findings. We also discuss potential extensions and directions for future research in the multi-task deep learning landscape.

**Keywords:** multi-task deep learning; multi-task; multiple losses; deep learning; game theory; Nash equilibrium; learning agents

**MSC:** 68T27



**Citation:** Lee, M. Multi-Task Deep Learning Games: Investigating Nash Equilibria and Convergence Properties. *Axioms* **2023**, *12*, 569. <https://doi.org/10.3390/axioms12060569>

Academic Editor: Freddy A. Lucay and Wenceslao Palma

Received: 13 May 2023

Revised: 4 June 2023

Accepted: 7 June 2023

Published: 8 June 2023



**Copyright:** © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The emergence of deep learning, a transformative shift in the field of machine learning, has ushered in an era of innovation and discovery [1–3]. The surge in research and development has given birth to a plethora of novel methodologies, such as generative adversarial networks (GANs) [4–6], which are capable of generating synthetic data that mirrors the structure of real-world data, providing a new tool for data augmentation and anomaly detection. Other notable advancements include diffusion models [7–10], a type of generative model that transforms a simple noise distribution into a complex data distribution through a sequence of diffusion steps. Further expanding the boundaries of deep learning, the innovative neural radiance fields (NeRFs) [11–14] have proven effective in synthesizing novel views of complex 3D scenes from sparse input views, contributing substantially to the field of computer graphics.

These pioneering techniques have substantially broadened the applicability of deep learning, allowing for its integration into a wide range of fields. For instance, in biomedical data analysis [15], deep learning techniques have played a crucial role in accelerating the discovery of complex patterns and hidden structures within high-dimensional medical data. Similarly, deep learning has made notable contributions in the domain of engineering

design [16] by enabling data-driven optimization of system parameters, portfolio optimization [17–19] through efficient modeling of financial data, and in computer vision [20,21], revolutionizing the ability of machines to interpret and understand visual data.

In the contemporary era of machine learning, multi-task deep learning methodologies have emerged as highly potent tools. Zhao et al. [22] have shown how these methodologies are reshaping the landscape by facilitating superior performance across a range of inter-related tasks. Similarly, Samant et al. [23] have explored the potential of multi-task deep learning in diverse fields, enhancing our understanding of its practical applications.

The primary idea underpinning these methodologies, as explained by Vithayathil et al. [24], is the concurrent training of a single model on multiple tasks. This unique approach leverages the interplay of similarities and differences among these tasks, often leading to improved performance over isolated models trained independently. A comprehensive review by Zhou et al. [25] provides further insights into this concept in medical image applications. Despite its promise, the intricate inter-task dynamics in a multi-task deep learning system pose a significant challenge. Xu et al. [26] illustrate this complexity and highlight the necessity for refined analytical tools to navigate this domain.

In an attempt to meet this challenge, our work explores multi-task deep learning through the approach of game theory [27–29]. Game theory provides a potent toolset for understanding the interactions between different tasks in a multi-task learning scenario [30]. Our primary motivation stems from the understanding that while multi-task deep learning models are being extensively used in contemporary research, the true complexity of their internal interactions is often not fully appreciated. By adopting a game-theoretic perspective, we explore the strategic behavior of learning tasks, as if they were players in a game, each striving to minimize their specific loss functions while sharing a common set of parameters. This novel perspective allows us to cast multi-task deep learning as a game, opening up new avenues for exploring the existence and convergence of Nash equilibria within these models. Understanding these aspects is crucial as it provides a robust theoretical framework to establish novel multi-task models.

In this game-theoretic framework, tasks act as players sharing a common parameter space, each striving to minimize its unique loss function, as explained in Gavidia et al. [31]. Furthermore, we investigate the mathematical underpinnings of the existence and convergence of Nash equilibria within the multi-task deep learning paradigm. Reny [32] provided a robust groundwork in understanding the fundamental principles of Nash equilibria, which we attempt to extend to the context of deep learning.

The portrayal of multi-task deep learning as a game involving multiple agents provides us with a robust platform for the formal analysis of inter-task interactions as they vie for shared resources, such as model parameters and computational capacity, throughout the learning process. The game-theoretic paradigm presents a novel contribution to the understanding of multi-task deep learning dynamics, which have been traditionally investigated from an optimization perspective.

Specifically, the game-theoretic conceptualization of multi-task deep learning empowers us to confirm the existence and convergence of Nash equilibria, given certain conditions on the loss functions, such as convexity and Lipschitz continuity. This novel finding paves the way for a more systematic exploration of the intricate balance between competition and cooperation among tasks in a multi-task environment. Moreover, it guides the development of learning algorithms that account for the strategic conduct of tasks to enhance overall performance.

Our findings also lay a solid groundwork for future research into the game-theoretic properties of multi-task deep learning. For example, extending the current framework to include cooperative games, where tasks can form strategic coalitions to optimize their joint performance, is a viable research direction. Such an extension could reveal novel strategies for the design and training of multi-task deep learning models that capitalize on the inherent correlations among tasks.

Our key contributions are as follows:

- We cast the multi-task deep learning problem as a game where each task is treated as a player aiming to minimize its task-specific loss function.
- We lay the groundwork for the necessary mathematical concepts and introduce the notion of a Nash equilibrium for the multi-task deep learning game, which corresponds to a strategy profile in which no player can unilaterally deviate from their strategy and achieve a lower task-specific loss.
- Under specific convexity and Lipschitz continuity assumptions for the loss functions, we demonstrate the existence of at least one Nash equilibrium for the multi-task deep learning game.
- We examine the convergence characteristics of the Nash equilibrium and ascertain conditions under which the iterative updates of task-specific and shared parameters converge to the Nash equilibrium.
- We present a thorough analysis of the implications and limitations of our theoretical findings and discuss potential extensions and avenues for future research.

The rest of this manuscript is organized as follows. Section 2 introduces the necessary preliminaries, covering the fundamentals of deep learning, multi-task deep learning, and the game theory of multi-task deep learning in Sections 2.1–2.3, respectively. Following that, Section 3 presents our main results and their proofs. Specifically, Section 3.1 provides a proof for the existence of Nash equilibria in multi-task deep learning, while Section 3.2 offers an interpretation of this theorem and discusses its implications. Further, Section 3.3 investigates the proof of Nash equilibria's convergence properties. Section 4 discusses the challenges associated with the interpretation of this study and outlines potential directions for future work. Finally, Section 5 concludes the manuscript, summarizing the key findings and their implications.

## 2. Preliminaries

### 2.1. Deep Learning

Deep learning is a subfield of machine learning that deals with the construction and training of neural networks composed of multiple layers, enabling the learning of hierarchical representations [33–35]. Given an input  $x \in \mathcal{X}$ , where  $\mathcal{X}$  denotes the input space, a deep learning model  $\mathcal{N}$  with parameters  $\Theta$  maps the input to an output  $\hat{y} \in \mathcal{Y}$ , where  $\mathcal{Y}$  denotes the output space. We represent the neural network as a composition of functions  $f_1, f_2, \dots, f_L$ , where  $L$  is the number of layers:

$$\mathcal{N}(x; \Theta) = f_L \circ f_{L-1} \circ \dots \circ f_1(x; \Theta). \quad (1)$$

This sequence of transformations, as defined in Equation (1), allows the neural network to learn intricate and abstract representations of the input data. Each function  $f_l$  in the composition encapsulates a transformation at the  $l$ -th layer of the network, transforming the information passed from the preceding layer, or the raw input data for  $l = 1$ , into a higher-level representation.

Each layer  $f_l$  is defined by a weight matrix  $W_l \in \mathbb{R}^{n_l \times n_{l-1}}$ , a bias vector  $b_l \in \mathbb{R}^{n_l}$ , and an activation function  $g_l : \mathbb{R}^{n_l} \rightarrow \mathbb{R}^{n_l}$ , with  $n_l$  denoting the number of units in layer  $l$ . The layer function  $f_l$  can be represented as:

$$f_l(h_{l-1}; W_l, b_l) = g_l(W_l h_{l-1} + b_l), \quad (2)$$

where  $h_{l-1} \in \mathbb{R}^{n_{l-1}}$  is the output of the previous layer or the input for  $l = 1$ . The parameters  $\Theta$  of the neural network are given by  $\Theta = \{W_l, b_l\}_{l=1}^L$ .

The mathematical encapsulation of each layer function, as presented in Equation (2), serves to highlight the primary operations occurring at each layer of the deep learning model. This encapsulation underscores the linear transformation, parameterized by  $W_l$  and  $b_l$ , and the non-linear transformation, governed by the activation function  $g_l(\cdot)$ . This

interplay between linear and non-linear operations is what equips deep learning models with their remarkable expressivity and learning capability.

**Assumption 1.** The activation functions  $g_l(\cdot)$  are Lipschitz continuous with a Lipschitz constant  $K_l > 0$  for all layers  $l \in \{1, 2, \dots, L\}$ .

**Remark 1.** Common activation functions, such as the hyperbolic tangent, rectified linear unit (ReLU), and Gaussian error linear unit (GELU) functions satisfy the Lipschitz continuity assumption.

Given a dataset  $\mathcal{D} = (x_i, y_i)_{i=1}^N$ , the objective in deep learning is to minimize a loss function  $L(\Theta)$ , which measures the discrepancy between the true labels  $y_i \in \mathcal{Y}$  and the predicted labels  $\hat{y}_i = \mathcal{N}(x_i; \Theta)$ . The loss function can be written as:

$$L(\Theta) = \frac{1}{N} \sum_{i=1}^N \ell(y_i, \hat{y}_i), \quad (3)$$

where  $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_{\geq 0}$  is a task-specific loss function, such as the mean squared error for regression tasks or the cross-entropy loss for classification tasks.

With Equation (3), we encapsulate the objective of the learning process in the context of supervised learning. The objective is to find the parameters  $\Theta$  that minimize the discrepancy, as measured by the loss function  $L(\Theta)$ , between the model's predictions and the true labels. This mathematical representation underlines the essence of the empirical risk minimization principle that guides the learning process in deep learning models.

The optimization problem can be expressed as:

$$\Theta^* = \min_{\Theta} L(\Theta). \quad (4)$$

**Proposition 1.** Let  $L$  be a continuously differentiable function with respect to  $\Theta$ , and let  $\nabla L(\Theta)$  denote its gradient. The optimization problem can be solved using gradient-based methods, such as Adam and RMSprop, which update the parameters  $\Theta$  iteratively using the following rule:

$$\Theta^{(t+1)} = \Theta^{(t)} - \alpha_t \nabla L(\Theta^{(t)}), \quad (5)$$

where  $t$  is the iteration index, and  $\alpha_t > 0$  is the learning rate at iteration  $t$ .

**Assumption 2.** The learning rate  $\alpha_t$  satisfies the following conditions:  $\sum_{t=1}^{\infty} \alpha_t < \infty$  and  $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$ .

**Remark 2.** In this paper, we assume the use of a learning rate scheduler to ensure a decreased learning rate throughout the training process. The learning rate  $\alpha_t$  is determined by the scheduler, which adjusts the value of  $\alpha_t$  at some pre-defined  $t$ . The purpose of this assumption is to guarantee the convergence of the optimization algorithm. It is common practice in machine learning to employ learning rate schedulers to facilitate effective training and convergence. By gradually reducing the learning rate over time, the model can navigate the optimization landscape more efficiently and avoid large oscillations in the parameter updates. This approach has been shown to improve the convergence behavior and stability of training algorithms in various settings. It is important to note that the use of a learning rate scheduler is a common practice and does not introduce any additional assumptions or complexity to the underlying problem. Rather, it is a technique employed to enhance the convergence behavior and performance of the optimization process.

**Theorem 1.** Under Assumptions 1 and 2, the optimization problem converges to a stationary point.

**Proof.** We prove Theorem 1 by establishing that  $\{\Theta^{(t)}\}_{t \in \mathbb{N}}$  forms a Cauchy sequence under certain conditions.

Given Assumption 1, the activation functions  $g_l(\cdot)$  are Lipschitz continuous with a Lipschitz constant  $K_l > 0$  for all layers  $l \in 1, 2, \dots, L$ . This implies that the neural network function  $\mathcal{N}(\cdot; \Theta)$  is Lipschitz continuous, which in turn implies that the loss function  $L(\Theta)$  is Lipschitz continuous in its argument, given that the other components of  $L(\Theta)$  are either constants or deterministic functions of the data.

Let  $\{\Theta^{(t)}\}_{t \in \mathbb{N}}$  denote the sequence of parameters generated by the optimization algorithm specified in Proposition 1. We will show that this sequence is a Cauchy sequence.

For any  $\varepsilon > 0$ , there exists a  $T \in \mathbb{N}$  such that for all  $t, s \geq T$ , we have:

$$|\Theta^{(t)} - \Theta^{(s)}| \leq \varepsilon. \tag{6}$$

Since  $L(\Theta)$  is Lipschitz continuous, there exists a Lipschitz constant  $K > 0$  such that for all  $\Theta_1, \Theta_2 \in \mathbb{R}^d$ , we have:

$$|L(\Theta_1) - L(\Theta_2)| \leq K|\Theta_1 - \Theta_2|. \tag{7}$$

By the mean value theorem, for all  $t \geq 1$ , there exists a  $\Theta' \in \mathbb{R}^d$  such that:

$$|L(\Theta^{(t)}) - L(\Theta^{(t-1)})| = |\nabla L(\Theta')^T(\Theta^{(t)} - \Theta^{(t-1)})|. \tag{8}$$

Now, combining Equations (7) and (8), we obtain:

$$|\nabla L(\Theta')| \leq K. \tag{9}$$

Given Assumption 2, the learning rate  $\alpha_t$  satisfies the conditions  $\sum_{t=1}^{\infty} \alpha_t < \infty$  and  $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$ . This, combined with Equation (9), implies that the sequence  $\{\Theta^{(t)}\}_{t \in \mathbb{N}}$  forms a Cauchy sequence.

Let us consider any two elements in the sequence  $\{\Theta^{(t)}\}_{t \in \mathbb{N}}$  and  $\{\Theta^{(s)}\}_{s \in \mathbb{N}}$  where  $s > t \geq T$ . We can write:

$$|\Theta^{(s)} - \Theta^{(t)}| \leq \sum_{k=t}^{s-1} |\Theta^{(k+1)} - \Theta^{(k)}|. \tag{10}$$

The right-hand side of Equation (10) can be interpreted as the sum of the distances between consecutive elements in the sequence from  $t$  to  $s$ . By the update rule in Proposition 1, we have:

$$|\Theta^{(k+1)} - \Theta^{(k)}| = \alpha_k |\nabla L(\Theta^{(k)})|. \tag{11}$$

Substituting Equation (11) into Equation (10) and applying the condition from Equation (9), we obtain:

$$|\Theta^{(s)} - \Theta^{(t)}| \leq K \sum_{k=t}^{s-1} \alpha_k. \tag{12}$$

Since the sequence  $\sum_{t=1}^{\infty} \alpha_t < \infty$ , for any  $\varepsilon > 0$ , there exists a  $T \in \mathbb{N}$  such that for all  $s > t \geq T$ , we have:

$$\sum_{k=t}^{s-1} \alpha_k < \frac{\varepsilon}{K}. \tag{13}$$

Substituting Equation (13) into Equation (12), we obtain:

$$|\Theta^{(s)} - \Theta^{(t)}| < \varepsilon, \tag{14}$$

which fulfills the condition for a Cauchy sequence as specified in Equation (6).

Therefore, the sequence  $\{\Theta^{(t)}\}_{t \in \mathbb{N}}$  is a Cauchy sequence. Since  $\mathbb{R}^d$  is a complete metric space, every Cauchy sequence in  $\mathbb{R}^d$  converges to a limit in  $\mathbb{R}^d$ . Hence, the sequence  $\{\Theta^{(t)}\}_{t \in \mathbb{N}}$  converges to a limit  $\Theta^* \in \mathbb{R}^d$ . By the continuity of the gradient  $\nabla L(\Theta)$ , we have  $\lim_{t \rightarrow \infty} \nabla L(\Theta^{(t)}) = \nabla L(\Theta^*)$ .  $\square$

**Remark 3.** Theorem 1 shows that under certain assumptions, the optimization problem of a deep learning model converges to a stationary point. This result is instrumental in explaining the empirical success of deep learning. However, it is worth noting that this result does not guarantee convergence to a global minimum, or even a local minimum. In fact, the optimization problem of deep learning is known to be non-convex, and in general, may have many saddle points and local minima.

### 2.2. Multi-Task Deep Learning

In this paper, our focus is specifically on a generalized scenario of multi-task deep learning, wherein each task is modelled as a player in a game, aiming to minimize its task-specific loss function. This game-theoretic perspective can be extended to a broad range of multi-task deep learning models that have distinct task-specific loss functions and shared representations or parameters. While our mathematical investigations primarily pertain to a certain class of multi-task deep learning models, the game-theoretic framework and the identified properties of Nash equilibria and their convergence are largely universal, given that the models adhere to the aforementioned conditions. It is important to note that the specific performance and behavior of the models may vary depending on the intricacies of the individual tasks, the structure of the shared and task-specific components, and the nature of the data. Thus, the findings presented in this paper should be interpreted considering these factors.

Now, we introduce the concept of multi-task learning in the context of deep learning. Suppose we have  $M$  related tasks, each with its own dataset  $\mathcal{D}_m = (x_{i,m}, y_{i,m})_{i=1}^{N_m}$ , for  $m \in \{1, 2, \dots, M\}$ . The objective in multi-task deep learning is to learn a shared representation that benefits all tasks, while also learning task-specific components. We achieve this by introducing task-specific output layers and shared hidden layers.

The notion of shared representations in multi-task learning stems from the fundamental premise that related tasks share underlying patterns and structures. By learning a shared representation, the model leverages this commonality to enhance its performance across tasks. It is essential to note that, while the shared layers capture the commonalities across tasks, the task-specific layers, parameterized by  $\Theta_m$ , enable the model to learn and adapt to the idiosyncrasies of each individual task.

Let  $\mathcal{N}_m$  denote the neural network corresponding to task  $m$ , and let  $\Theta_m$  and  $\Psi$  denote its task-specific and shared parameters, respectively. The output of  $\mathcal{N}_m$  is given by:

$$\mathcal{N}_m(x; \Theta_m, \Psi) = f_{L_m} \circ f_{L_m-1} \circ \dots \circ f_1(x; \Theta_m, \Psi), \tag{15}$$

where  $L_m$  is the number of layers for task  $m$ , and  $f_1, \dots, f_{L_s}$ , with  $1 \leq L_s \leq L_m$ , are shared layers across tasks. We assume the task-specific loss functions  $\ell_m : \mathcal{Y}_m \times \mathcal{Y}_m \rightarrow \mathbb{R}_{\geq 0}$ , where  $\mathcal{Y}_m$  is the output space for task  $m$ . The overall multi-task loss function is given by:

In Equation (15), we describe the functional form of each task-specific model in the multi-task learning scenario. It highlights the two key components of multi-task learning: shared layers, which capture common features across tasks, and task-specific layers, which cater to the unique aspects of each task.

$$L_{MT}(\Theta_1, \dots, \Theta_M, \Psi) = \sum_{m=1}^M \frac{1}{N_m} \sum_{i=1}^{N_m} \ell_m(y_{i,m}, \hat{y}_{i,m}), \tag{16}$$

where  $\hat{y}_{i,m} = \mathcal{N}_m(x_{i,m}; \Theta_m, \Psi)$ . The multi-task optimization problem can be written as:

$$\min_{\Theta_1, \dots, \Theta_M, \Psi} L_{MT}(\Theta_1, \dots, \Theta_M, \Psi). \tag{17}$$

### 2.3. Game Theory of Multi-Task Deep Learning

To apply game-theoretic concepts to this multi-task optimization problem, we treat each task as a player in a game. Each player  $m$  aims to minimize its task-specific loss  $\ell_m$

subject to the shared layers  $\Psi$ . This leads us to the concept of Nash equilibria in the context of multi-task deep learning.

The game-theoretic perspective adopted here provides a rigorous and insightful framework to study the dynamics of multi-task learning. By framing each task as a player striving to minimize its loss, we capture the inherent tension and cooperation in multi-task learning. Each player, or task, cooperates by sharing representations through  $\Psi$ , while also competing to tailor these shared representations to minimize its own loss. This delicate balance between cooperation and competition is at the heart of the multi-task learning process.

**Definition 1** (Nash Equilibrium). *A strategy profile  $(\Theta_1^*, \dots, \Theta_M^*, \Psi^*)$  is a Nash equilibrium of the multi-task deep learning game if, for each player  $m$ , we have*

$$L_{MT}(\Theta_1^*, \dots, \Theta_{m-1}^*, \Theta_m, \Theta_{m+1}^*, \dots, \Theta_M^*, \Psi^*) \geq L_{MT}(\Theta_1^*, \dots, \Theta_M^*, \Psi^*), \quad (18)$$

for all  $\Theta_m \in \mathbb{R}^{n_m}$ .

Definition 1 formalizes the notion of an equilibrium in the context of the multi-task deep learning game. At a Nash equilibrium, no player or task, can unilaterally reduce its loss by deviating from the equilibrium strategy. This signifies a state of balance where each task has adapted the shared representations to best minimize its own loss, given the strategies of the other tasks. This equilibrium concept provides a useful theoretical tool to study the performance and convergence properties of multi-task learning algorithms.

To study the existence and convergence of Nash equilibria in the multi-task deep learning game, we introduce the following assumptions.

**Assumption 3.** *The task-specific loss functions  $\ell_m$  are convex and continuously differentiable with respect to  $\Theta_m$  and  $\Psi$ , for all  $m \in \{1, 2, \dots, M\}$ .*

**Assumption 4.** *The gradients of the task-specific loss functions  $\ell_m$  with respect to  $\Theta_m$  and  $\Psi$  are Lipschitz continuous with Lipschitz constants  $C_m > 0$ , for all  $m \in \{1, 2, \dots, M\}$ .*

**Theorem 2** (Existence of Nash Equilibrium). *Under Assumptions 3 and 4, there exists at least one Nash equilibrium for the multi-task deep learning game.*

**Theorem 3** (Convergence of Nash Equilibrium). *Under Assumptions 1–4, if each player  $m$  updates its task-specific parameters  $\Theta_m$  using a gradient-based method satisfying Assumption 2, and the shared parameters  $\Psi$  are updated using a consensus-based approach, the iterative updates converge to a Nash equilibrium of the multi-task deep learning game, i.e.,*

$$\lim_{t \rightarrow \infty} \nabla_{\Theta_m} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) = 0, \quad \forall m \in \{1, 2, \dots, M\}. \quad (19)$$

In the context of our work, the Nash equilibrium pertains to a state in the multi-task learning game where no single task can reduce its loss by independently changing its strategy, given the strategies of all of the other tasks. It represents a stable state of the system where each task’s strategy is optimal with respect to the strategies of all other tasks. It is important to note that a Nash equilibrium does not necessarily correspond to the optimal state of the entire system, but to a state of stability where no task has an incentive to deviate unilaterally.

Theorems 2 and 3, which are the main contributions of this paper, illustrate that the multi-task deep learning model trained under certain assumptions will eventually converge to a Nash equilibrium. This clarifies why many existing studies, which utilize multi-task losses for their models, generally achieve convergence even if the loss function of the model is composed of the sum of multiple losses. These theorems essentially provide the theoretical underpinning explaining the observed behavior of these models.

### 3. Main Results and Proofs

#### 3.1. Proof of Theorem 2

In this section, we provide a rigorous proof for Theorem 2 concerning the existence of a Nash equilibrium in the multi-task deep learning game under Assumptions 3 and 4. These steps include defining a pseudo-gradient and establishing its crucial properties under certain assumptions. We then follow this with a detailed lemma and its proof, which prove the Lipschitz continuity of the pseudo-gradient.

First, let us define the pseudo-gradient  $\tilde{\nabla}$  of the overall multi-task loss function  $L_{MT}$  with respect to  $\Theta_m$  and  $\Psi$  as follows:

$$\tilde{\nabla}_{\Theta_m} L_{MT}(\Theta_1, \dots, \Theta_M, \Psi) = \frac{1}{N_m} \sum_{i=1}^{N_m} \nabla_{\Theta_m} \ell_m(y_{i,m}, \hat{y}_{i,m}), \tag{20}$$

$$\tilde{\nabla}_{\Psi} L_{MT}(\Theta_1, \dots, \Theta_M, \Psi) = \frac{1}{M} \sum_{m=1}^M \frac{1}{N_m} \sum_{i=1}^{N_m} \nabla_{\Psi} \ell_m(y_{i,m}, \hat{y}_{i,m}). \tag{21}$$

The following lemma establishes a crucial property of the pseudo-gradient.

**Lemma 1.** Under Assumption 4, the pseudo-gradient  $\tilde{\nabla}$  of  $L_{MT}$  with respect to  $\Theta_m$  and  $\Psi$  is Lipschitz continuous with Lipschitz constants  $\tilde{C}_m = C_m$  and  $\tilde{C}_{\Psi} = \max_{1 \leq m \leq M} C_m$ , respectively.

**Proof.** Consider two distinct points  $(\Theta_1, \dots, \Theta_M, \Psi)$  and  $(\Theta'_1, \dots, \Theta'_M, \Psi')$  in the parameter space. We have:

$$\begin{aligned} & |\tilde{\nabla}_{\Theta_m} L_{MT}(\Theta_1, \dots, \Theta_M, \Psi) - \tilde{\nabla}_{\Theta_m} L_{MT}(\Theta'_1, \dots, \Theta'_M, \Psi')| \\ &= \left| \frac{1}{N_m} \sum_{i=1}^{N_m} (\nabla_{\Theta_m} \ell_m(y_{i,m}, \hat{y}_{i,m}) - \nabla_{\Theta_m} \ell_m(y_{i,m}, \hat{y}'_{i,m})) \right| \\ &\leq \frac{1}{N_m} \sum_{i=1}^{N_m} |\nabla_{\Theta_m} \ell_m(y_{i,m}, \hat{y}_{i,m}) - \nabla_{\Theta_m} \ell_m(y_{i,m}, \hat{y}'_{i,m})| \\ &\leq \frac{1}{N_m} \sum_{i,m} C_m |\hat{y}_{i,m} - \hat{y}'_{i,m}| \\ &\leq C_m \max_{1 \leq i \leq N_m} |\hat{y}_{i,m} - \hat{y}'_{i,m}| \\ &\leq C_m |(\Theta_1, \dots, \Theta_M, \Psi) - (\Theta'_1, \dots, \Theta'_M, \Psi')|, \end{aligned}$$

where the first inequality follows from the triangle inequality, the second inequality follows from Assumption 4, and the last inequality follows from the Lipschitz continuity of the neural networks.

Similarly, for the pseudo-gradient with respect to  $\Psi$ , we have:

$$\begin{aligned} & |\tilde{\nabla}_{\Psi} L_{MT}(\Theta_1, \dots, \Theta_M, \Psi) - \tilde{\nabla}_{\Psi} L_{MT}(\Theta'_1, \dots, \Theta'_M, \Psi')| \\ &= \left| \frac{1}{M} \sum_{m=1}^M \frac{1}{N_m} \sum_{i=1}^{N_m} (\nabla_{\Psi} \ell_m(y_{i,m}, \hat{y}_{i,m}) - \nabla_{\Psi} \ell_m(y_{i,m}, \hat{y}'_{i,m})) \right| \\ &\leq \frac{1}{M} \sum_{m=1}^M \frac{1}{N_m} \sum_{i=1}^{N_m} |\nabla_{\Psi} \ell_m(y_{i,m}, \hat{y}_{i,m}) - \nabla_{\Psi} \ell_m(y_{i,m}, \hat{y}'_{i,m})| \\ &\leq \frac{1}{M} \sum_{m=1}^M C_m \max_{1 \leq i \leq N_m} |\hat{y}_{i,m} - \hat{y}'_{i,m}| \\ &\leq \max_{1 \leq m \leq M} C_m |(\Theta_1, \dots, \Theta_M, \Psi) - (\Theta'_1, \dots, \Theta'_M, \Psi')|. \end{aligned}$$



Thus, the Lipschitz continuity of the pseudo-gradient  $\tilde{\nabla}$  with respect to  $\Theta_m$  and  $\Psi$  is established with Lipschitz constants  $\tilde{C}_m = C_m$  and  $\tilde{C}_\Psi = \max_{1 \leq m \leq M} C_m$ , respectively.  $\square$

Now, we proceed with the proof of Theorem 2. Due to the convexity of the task-specific loss functions  $\ell_m$  (Assumption 3) and the Lipschitz continuity of their gradients (Assumption 4), we have that the pseudo-gradient  $\tilde{\nabla}$  of the overall multi-task loss function  $L_{MT}$  with respect to  $\Theta_m$  and  $\Psi$  is Lipschitz continuous (Lemma 1). Consequently, the overall multi-task loss function  $L_{MT}$  is also convex and continuously differentiable with respect to  $\Theta_m$  and  $\Psi$ .

The existence of a Nash equilibrium follows from the Kakutani fixed-point theorem, which states that a set-valued mapping that maps a compact convex set into itself and has a closed graph admits a fixed point. In our case, we can define a set-valued mapping  $F : \mathbb{R}^{n_1} \times \dots \times \mathbb{R}^{n_M} \times \mathbb{R}^{n_\Psi} \rightarrow 2^{\mathbb{R}^{n_1} \times \dots \times \mathbb{R}^{n_M} \times \mathbb{R}^{n_\Psi}}$  as follows:

$$F(\Theta_1, \dots, \Theta_M, \Psi) = \bigcap_{m=1}^M \{(\Theta'_1, \dots, \Theta'_M, \Psi') \in \mathbb{R}^{n_1} \times \dots \times \mathbb{R}^{n_M} \times \mathbb{R}^{n_\Psi} : \tilde{\nabla}_{\Theta_m} L_{MT}(\Theta_1, \dots, \Theta_M, \Psi) = 0\}. \tag{22}$$

By construction,  $F$  maps a compact convex set into itself and has a closed graph. Hence, according to the Kakutani fixed-point theorem,  $F$  admits a fixed point  $(\Theta_1^*, \dots, \Theta_M^*, \Psi^*)$ . By Definition 1, this fixed point corresponds to a Nash equilibrium of the multi-task deep learning game. Therefore, Theorem 2 is proven.  $\square$

### 3.2. Interpretation and Implications of Theorem 2

**Remark 4.** Assumption 3 is an ideal condition in the context of deep learning since it requires the task-specific loss functions  $\ell_m$  to be convex with respect to  $\Theta_m$  and  $\Psi$ . In practice, the loss functions in deep learning models often exhibit non-convex behavior with multiple local minima. However, we adopt this assumption to facilitate the analysis of the existence of Nash equilibria and to provide theoretical insights into the multi-task deep learning game.

In the pursuit of understanding the behavior of multi-task deep learning systems, Assumption 3 plays a pivotal role. As outlined in Remark 4, it is indeed an ideal condition that aims to simplify the complexity inherent in such systems by assuming the task-specific loss functions  $\ell_m$  to be convex with respect to  $\Theta_m$  and  $\Psi$ . This assumption, albeit far from the intricate realities of deep learning models, serves as a useful abstraction that provides a tractable analytical framework to study the existence of Nash equilibria.

**Remark 5.** Theorem 2 establishes the existence of at least one Nash equilibrium for the multi-task deep learning game under Assumptions 3 and 4. This result indicates that, in the context of the multi-task deep learning game, there exists a strategy profile  $(\Theta_1^* \dots, \Theta_M^*, \Psi^*)$  such that no player can unilaterally deviate from their strategy and achieve a lower task-specific loss. In other words, each player's optimal strategy depends on the strategies of the other players, and no single player can improve their task-specific performance without affecting the performance of other tasks.

Theorem 2, as explicated in Remark 5, builds on these ideal assumptions to prove the existence of at least one Nash equilibrium in the multi-task deep learning game. This is a significant theoretical milestone, as it offers a stable point in the strategy space where no player can unilaterally deviate from their strategy to achieve a lower task-specific loss. This equilibrium encapsulates the competitive yet interdependent nature of the tasks in the multi-task deep learning game.

**Proposition 2.** Under Assumptions 3 and 4, the Nash equilibrium of the multi-task deep learning game is unique.

**Proof.** The proof of Proposition 2 is based on the fact that the overall multi-task loss function  $L_{MT}$  is strictly convex under Assumptions 3 and 4. Due to the strict convexity of  $L_{MT}$ , there can be at most one global minimum. Since a Nash equilibrium corresponds to a global minimum of  $L_{MT}$ , the uniqueness of the Nash equilibrium follows directly from the uniqueness of the global minimum.  $\square$

Proposition 2 further extends the implications of the convex loss assumption to claim uniqueness of the Nash equilibrium under these ideal conditions. The proof of this proposition relies on the strict convexity of the overall multi-task loss function  $L_{MT}$  under the stated assumptions. This proposition, when taken in tandem with Theorem 2, provides a theoretical framework for analyzing the equilibrium properties of multi-task deep learning systems.

**Remark 6.** *In practice, given the non-convex nature of deep learning loss functions, the uniqueness of the Nash equilibrium (Proposition 2) may not hold. However, this result provides a theoretical insight into the behavior of the multi-task deep learning game under ideal conditions, and it serves as a useful guideline when designing multi-task deep learning algorithms.*

The findings, as discussed in Remark 6, are not without limitations, especially considering the non-convex nature of deep learning loss functions in practical scenarios. The uniqueness of the Nash equilibrium, established under ideal conditions, may not hold in real-world settings. However, the insights derived from these theoretical results underpin the understanding of the behavior of multi-task deep learning systems under ideal conditions, thereby providing a benchmark for the design and analysis of such systems.

Theorem 2 and Proposition 2 together provide a strong foundation for the analysis of the multi-task deep learning game. However, it is essential to acknowledge the limitations imposed by the ideal conditions assumed in the analysis, such as the convexity of the loss functions. Despite these limitations, the results offer valuable insights into the behavior of multi-task deep learning systems and the interplay between different tasks sharing a common set of parameters.

In summary, the results from Theorem 2 and Proposition 2 pave the way for an elaborate understanding of the underlying dynamics of multi-task deep learning games. It is, however, crucial to be aware of the limitations posed by the assumed ideal conditions, which may deviate from the realities of deep learning loss functions. Despite these limitations, the theoretical findings illuminate the complex behavior of multi-task deep learning systems and the nuanced interactions among tasks sharing a common set of parameters.

### 3.3. Proof of Theorem 3

In order to prove Theorem 3, we first describe the gradient-based and consensus-based update rules for the task-specific parameters  $\Theta_m$  and shared parameters  $\Psi$ , respectively. To start with, we redefine our update rules, namely, the gradient-based and consensus-based updates, for the task-specific parameters  $\Theta_m$  and the shared parameters  $\Psi$ . These update rules essentially use gradients of the multi-task loss function to update parameters in an iterative manner. The shared parameters are updated using the consensus-based update rule, which averages the gradients of all tasks.

**Definition 2** (Gradient-Based Update). *For each player  $m$ , the task-specific parameters  $\Theta_m$  are updated using the following rule:*

$$\Theta_m^{(t+1)} = \Theta_m^{(t)} - \alpha_t^{(m)} \nabla_{\Theta_m} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}), \quad (23)$$

where  $t$  is the iteration index, and  $\alpha_t^{(m)} > 0$  is the learning rate for player  $m$  at iteration  $t$ .

**Definition 3** (Consensus-Based Update). *The shared parameters  $\Psi$  are updated using the following rule:*

$$\Psi^{(t+1)} = \Psi^{(t)} - \beta_t \sum_{m=1}^M \nabla_{\Psi} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}), \tag{24}$$

where  $\beta_t > 0$  is the consensus learning rate at iteration  $t$ .

In the course of our analysis, we establish a technical lemma pertaining to the consensus-based update rule. This lemma guarantees that the difference between the shared parameters at successive iterations tends to zero as the number of iterations goes to infinity. This result is crucial as it suggests the convergence of the shared parameters.

**Lemma 2.** *Under Assumptions 2 and 4, the consensus-based update rule for the shared parameters  $\Psi$  guarantees that*

$$\lim_{t \rightarrow \infty} |\Psi^{(t+1)} - \Psi^{(t)}| = 0. \tag{25}$$

In the proof of Theorem 3, we leverage these update rules. We consider the task-specific parameters and the shared parameters in the multi-task learning framework, and show that they are updated in such a way that they converge to a Nash equilibrium. The key steps involve demonstrating that the sequence of parameters generated by the iterative updates minimizes the multi-task loss function, thus establishing the convergence of a Nash equilibrium.

**Proof of Theorem 3.** From Definitions 2 and 3, we can rewrite the iterative updates for the task-specific parameters  $\Theta_m$  and shared parameters  $\Psi$  as follows:

$$\Theta_m^{(t+1)} = \Theta_m^{(t)} - \alpha_t^{(m)} \nabla_{\Theta_m} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}), \tag{26}$$

$$\Psi^{(t+1)} = \Psi^{(t)} - \beta_t \sum_{m=1}^M \nabla_{\Psi} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}). \tag{27}$$

By Assumption 4, the gradients of the task-specific loss functions are Lipschitz continuous with Lipschitz constants  $C_m > 0$ , for all  $m \in \{1, 2, \dots, M\}$ . Hence, the gradient-based update in (26) and the consensus-based update in (27) are well-defined. Now, we will show that the sequence of parameters generated by the iterative updates converges to a Nash equilibrium under the given assumptions.

Consider the following Lyapunov function:

$$V(\Theta_1, \dots, \Theta_M, \Psi) = \sum_{m=1}^M |\Theta_m - \Theta_m^*|^2 + |\Psi - \Psi^*|^2, \tag{28}$$

where  $(\Theta_1^*, \dots, \Theta_M^*, \Psi^*)$  is a Nash equilibrium. We aim to show that  $\lim_{t \rightarrow \infty} V(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) = 0$ .

Applying the gradient-based and consensus-based updates, we have:

$$\begin{aligned} & V(\Theta_1^{(t+1)}, \dots, \Theta_M^{(t+1)}, \Psi^{(t+1)}) - V(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) \\ &= \sum_{m=1}^M \left[ |\Theta_m^{(t+1)} - \Theta_m^*|^2 - |\Theta_m^{(t)} - \Theta_m^*|^2 \right] + |\Psi^{(t+1)} - \Psi^*|^2 - |\Psi^{(t)} - \Psi^*|^2. \end{aligned} \tag{29}$$

For each  $m$ , let us expand the term  $|\Theta_m^{(t+1)} - \Theta_m^*|^2 - |\Theta_m^{(t)} - \Theta_m^*|^2$ :

$$\begin{aligned} & |\Theta_m^{(t+1)} - \Theta_m^*|^2 - |\Theta_m^{(t)} - \Theta_m^*|^2 \\ &= \left| \Theta_m^{(t)} - \alpha_t^{(m)} \nabla_{\Theta_m} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) - \Theta_m^* \right|^2 - |\Theta_m^{(t)} - \Theta_m^*|^2 \\ &= -2\alpha_t^{(m)} \langle \Theta_m^{(t)} - \Theta_m^*, \nabla_{\Theta_m} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) \rangle \\ &\quad + (\alpha_t^{(m)})^2 |\nabla_{\Theta_m} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)})|^2. \end{aligned} \tag{30}$$

Now, we examine the term  $|\Psi^{(t+1)} - \Psi^*|^2 - |\Psi^{(t)} - \Psi^*|^2$ :

$$\begin{aligned} & |\Psi^{(t+1)} - \Psi^*|^2 - |\Psi^{(t)} - \Psi^*|^2 \\ &= \left| \Psi^{(t)} - \beta_t \sum_{m=1}^M \nabla_{\Psi} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) - \Psi^* \right|^2 - |\Psi^{(t)} - \Psi^*|^2 \\ &= -2\beta_t \langle \Psi^{(t)} - \Psi^*, \sum_{m=1}^M \nabla_{\Psi} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) \rangle + \beta_t^2 \left| \sum_{m=1}^M \nabla_{\Psi} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) \right|^2. \end{aligned} \tag{31}$$

Substituting (30) and (31) into (29), we obtain:

$$\begin{aligned} & V(\Theta_1^{(t+1)}, \dots, \Theta_M^{(t+1)}, \Psi^{(t+1)}) - V(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) \\ &= -2 \sum_{m=1}^M \alpha_t^{(m)} \langle \Theta_m^{(t)} - \Theta_m^*, \nabla_{\Theta_m} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) \rangle \\ &\quad + \sum_{m=1}^M (\alpha_t^{(m)})^2 |\nabla_{\Theta_m} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)})|^2 \\ &\quad - 2\beta_t \langle \Psi^{(t)} - \Psi^*, \sum_{m=1}^M \nabla_{\Psi} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) \rangle \\ &\quad + \beta_t^2 \left| \sum_{m=1}^M \nabla_{\Psi} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) \right|^2. \end{aligned} \tag{32}$$

By Assumption 3, the task-specific loss functions are convex, and thus, we have:

$$\begin{aligned} \langle \Theta_m^{(t)} - \Theta_m^*, \nabla_{\Theta_m} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) \rangle &\geq L_{MT}(\Theta_1^{(t)}, \dots, \Theta_m^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) \\ &\quad - L_{MT}(\Theta_1^{(t)}, \dots, \Theta_m^*, \dots, \Theta_M^{(t)}, \Psi^{(t)}). \end{aligned} \tag{33}$$

This inequality is derived from the convexity of the task-specific loss functions, as stated in Assumption 3. We defined the function  $g(t) = L_{MT}(t\Theta_m^* + (1-t)\Theta_m^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)})$ , which is convex in  $t$ . The derivative of  $g(t)$  with respect to  $t$  evaluated at  $t = 1$  is the inner product on the left-hand side of the inequality. Since  $g(t)$  is convex, its derivative is non-decreasing, which leads to the inequality (33).

By summing (33) over  $m = 1, 2, \dots, M$ , we have:

$$\begin{aligned} \sum_{m=1}^M \langle \Theta_m^{(t)} - \Theta_m^*, \nabla_{\Theta_m} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) \rangle &\geq \sum_{m=1}^M L_{MT}(\Theta_1^{(t)}, \dots, \Theta_m^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) \\ &\quad - L_{MT}(\Theta_1^{(t)}, \dots, \Theta_m^*, \dots, \Theta_M^{(t)}, \Psi^{(t)}). \end{aligned} \tag{34}$$

Applying (34) to (32), we obtain:

$$\begin{aligned}
 & V(\Theta_1^{(t+1)}, \dots, \Theta_M^{(t+1)}, \Psi^{(t+1)}) - V(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) \\
 & \leq -2 \sum_{m=1}^M \alpha_t^{(m)} (L_{MT}(\Theta_1^{(t)}, \dots, \Theta_m^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) - L_{MT}(\Theta_1^{(t)}, \dots, \Theta_m^*, \dots, \Theta_M^{(t)}, \Psi^{(t)})) \\
 & \quad + \sum_{m=1}^M (\alpha_t^{(m)})^2 |\nabla_{\Theta_m} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)})|^2 \\
 & \quad - 2\beta_t \left\langle \Psi^{(t)} - \Psi^*, \sum_{m=1}^M \nabla_{\Psi} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) \right\rangle \\
 & \quad + \beta_t^2 \left| \sum_{m=1}^M \nabla_{\Psi} L_{MT}(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) \right|^2. \tag{35}
 \end{aligned}$$

Since the learning rates  $\alpha_t^{(m)}$  and  $\beta_t$  satisfy the conditions in Assumption 2, we have:

$$\sum_{t=0}^{\infty} \alpha_t^{(m)} < \infty, \quad \sum_{t=0}^{\infty} (\alpha_t^{(m)})^2 < \infty, \quad \sum_{t=0}^{\infty} \beta_t < \infty, \quad \sum_{t=0}^{\infty} \beta_t^2 < \infty. \tag{36}$$

From Lemma 2, we have:

$$\lim_{t \rightarrow \infty} |\Psi^{(t+1)} - \Psi^{(t)}| = 0. \tag{37}$$

By applying (36) and (37) to (35), we obtain:

$$\lim_{t \rightarrow \infty} V(\Theta_1^{(t+1)}, \dots, \Theta_M^{(t+1)}, \Psi^{(t+1)}) - V(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) \leq 0. \tag{38}$$

Since the Lyapunov function  $V(\Theta_1, \dots, \Theta_M, \Psi)$  is non-negative, it follows from (38) that the sequence of parameters generated by the iterative updates converges to a Nash equilibrium:

$$\lim_{t \rightarrow \infty} V(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) = 0. \tag{39}$$

Thus, we have proved Theorem 3.  $\square$

### 3.4. Interpretation and Implications of Theorem 3

We begin by analyzing the implications of Theorem 3 on the convergence properties of multi-task deep learning systems. Specifically, we discuss the role of the assumptions and their impact on the convergence rates. Additionally, we provide remarks on the limitations and possible extensions of the results presented in this theorem.

**Assumption 5** (Convergence Rate Condition). *Let  $\Theta_m^{(t)}$  and  $\Psi^{(t)}$  be the sequences of task-specific parameters and shared parameters generated by the iterative updates described in Definitions 2 and 3, respectively. We assume that the convergence rate of the parameters towards the Nash equilibrium satisfies the following condition:*

$$\lim_{t \rightarrow \infty} \frac{|(\Theta_1^{(t)}, \dots, \Theta_M^{(t)}, \Psi^{(t)}) - (\Theta_1^*, \dots, \Theta_M^*, \Psi^*)|}{t} = 0. \tag{40}$$

In analyzing Theorem 3, one finds the critical role played by the involved assumptions in determining the convergence rate of parameters in multi-task deep learning systems. Assumption 5 sets the stage by providing the generic condition for convergence. It warrants that the sequences of task-specific parameters and shared parameters generated iteratively, eventually converge towards the Nash equilibrium. However, it provides no insight into the rate of such convergence.

**Remark 7.** Assumption 5 is a relatively weak condition that implies the convergence of the sequences  $\Theta_m^{(t)}$  and  $\Psi^{(t)}$  to their respective Nash equilibrium values. However, it does not provide any information about the rate at which this convergence occurs. To obtain more precise convergence rates, one may require additional assumptions on the smoothness of the loss function or the structure of the multi-task deep learning game.

The subsequent remark points out the inherent weakness in Assumption 5, which is its silence on the rate of convergence. It suggests the need for additional assumptions to ascertain more precise rates of convergence. This requirement stems from the fact that the rate at which the sequences  $\Theta_m^{(t)}$  and  $\Psi^{(t)}$  converge to their Nash equilibrium values can be crucial for the practical application of these theorems.

Now, we present a proposition that explores the relationship between the convergence rate condition and the various assumptions made in Theorem 3.

**Proposition 3.** Under Assumptions 1–4, the convergence rate condition in Assumption 5 holds if and only if

$$\lim_{t \rightarrow \infty} \frac{\sum_{m=1}^M |\Theta_m^{(t)} - \Theta_m^*|^2 + |\Psi^{(t)} - \Psi^*|^2}{t} = 0. \tag{41}$$

**Proof.** The proof follows directly from the definition of the convergence rate condition and the results of Theorem 3. Since the Lyapunov function  $V(\Theta_1, \dots, \Theta_M, \Psi)$  converges to zero as  $t \rightarrow \infty$ , it is sufficient to show that the convergence rate condition holds if and only if the limit of the ratio of the Lyapunov function to  $t$  is zero. This completes the proof.  $\square$

Proposition 3 is an important extension of the discussions thus far. It establishes a direct relationship between the convergence rate condition and the various assumptions integral to Theorem 3. This proposition stands as a testament to the intricate and intertwined nature of these assumptions and their collective impact on the convergence rate of the multi-task deep learning system.

The proof of this proposition follows logically from the definition of the convergence rate condition and the results of Theorem 3. It brings to light the critical role of the Lyapunov function  $V(\Theta_1, \dots, \Theta_M, \Psi)$ , with its convergence to zero as  $t \rightarrow \infty$  being the key to understanding the convergence rate condition.

**Remark 8.** Proposition 3 highlights the significance of the assumptions made in Theorem 3. Specifically, it reveals that the convergence rate condition is closely related to the convergence properties of the Lyapunov function. Moreover, the proposition suggests that the optimal strategy for a given player in a multi-task deep learning scenario can be characterized by the Nash equilibrium.

#### 4. Challenges and Future Work

Although our results provide important insights into the existence and convergence properties of Nash equilibria in multi-task deep learning games, several challenges remain to be addressed.

The first pressing challenge, detailed in Remark 9, pertains to the non-convex nature of most deep learning loss functions [20,36]. In the current state of our work, we have made extensive use of the suppositions of convexity and Lipschitz continuity, as laid out in Assumptions 3 and 4. These suppositions were instrumental in the establishment of Theorem 2, which confirms the existence of Nash equilibria. However, most real-world deep learning models incorporate non-convex loss functions, a fact that complicates the process of investigating Nash equilibria’s existence and convergence properties. Thus, it is of paramount importance that we undertake further research into this aspect of multi-task deep learning games.

The non-convexity issue extends beyond the mere fact that real-world models utilize non-convex loss functions. From a theoretical standpoint, non-convexity significantly perturbs the mathematical tractability of our problem, as the elegant properties of convex

functions, which were harnessed in our derivations, are lost. As a consequence, the analysis of convergence properties becomes exceedingly complex. Furthermore, the interplay between non-convexity and Lipschitz continuity, a condition that we have also presupposed, necessitates careful scrutiny. The Lipschitz condition, which was crucial in ensuring the boundedness of the gradients in our analysis, may be compromised in the face of non-convex loss functions. The interrelationship between non-convexity and Lipschitz continuity in the multi-task deep learning game domain is still largely uncharted territory, thus opening a rich vein of research to be pursued.

**Remark 9** (Non-convex Loss Functions). *The assumptions of convexity and Lipschitz continuity (Assumptions 3 and 4) played a pivotal role in proving the existence of Nash equilibria (Theorem 2). However, in practice, many deep learning models involve non-convex loss functions. Investigating the existence and convergence properties of Nash equilibria in multi-task deep learning games with non-convex loss functions remains an open challenge.*

The issue of parameter initialization is, indeed, a thorny one, as it pertains not only to the convergence of the iterative updates but also to the stability of the system. Specifically, in multi-task deep learning games, a poor initialization could potentially destabilize the entire system, leading to oscillatory dynamics or even divergence of the iterative updates. The crux of the matter lies in understanding the correlation between the initial choice of parameters and the resultant equilibrium point, and how the latter affects the convergence and stability of the system. The dynamics of the game could possibly be influenced by the initial conditions, leading to an intricate system behavior that needs meticulous examination.

Our next challenge, as highlighted in Remark 10, is the potential sensitivity of Nash equilibria's convergence to the initial selection of parameters [37]. The pivotal question here pertains to the influence of parameter initialization on the convergence of iterative updates, as encapsulated in Equations (26) and (27). We must make an in-depth exploration of this question, bearing in mind its potential implications for effective initialization strategies in multi-task deep learning settings.

**Remark 10** (Parameter Initialization). *The convergence of Nash equilibria may be sensitive to the initial choice of parameters. It is important to explore how parameter initialization affects the convergence of the iterative updates (Equations (26) and (27)) and to identify strategies for effective initialization in the multi-task deep learning setting.*

We have made the assumption in our analysis that both task-specific and shared parameters will have decreased learning rates, as given in Definitions 2 and 3. This is stated in Remark 11. However, in deep learning practices, adaptive learning rate techniques such as RMSprop and Adam [38] have been empirically shown to boost convergence properties. Thus, extending our theoretical investigation to include adaptive learning rate schemes presents an enticing direction for future research.

In our current study, we made a simplifying assumption of decreased learning rates. However, this assumption might limit the applicability of our results to practical scenarios where adaptive learning rate techniques are prevalent. The adaptive learning rates, by dynamically adjusting the step size during the optimization process, have the potential to significantly alter the game dynamics. The interplay between adaptive learning rates and the convergence of Nash equilibria warrants detailed exploration. We need to examine how adaptive learning rates affect the trajectory and stability of the Nash equilibria, and how they interact with other elements of the multi-task deep learning game.

**Remark 11** (Adaptive Learning Rates). *Our analysis assumes constant learning rates for both task-specific and shared parameters (Definitions 2 and 3). However, adaptive learning rate techniques, such as RMSprop and Adam, have been shown to improve convergence properties in deep*

*learning. Extending our analysis to include adaptive learning rate schemes would be a valuable direction for future work.*

Another challenge, elucidated in Remark 12, is the need to understand the finite-sample performance of the multi-task deep learning game and its convergence rate. Our current analysis of Nash equilibria's convergence, as outlined in Theorem 3, is focused on the asymptotic behavior as the number of iterations tends to infinity. However, it is of practical consequence to understand how the game behaves with the finite number of samples typically encountered in real-world applications.

The current focus on the asymptotic behavior of the convergence of Nash equilibria might overlook crucial finite-sample effects that may arise in practice. It is noteworthy that in real-world applications, we are typically not afforded the luxury of infinite iterations, and thus the asymptotic analysis might not provide a complete picture of the game's dynamics. The behavior of the game during the finite-sample phase could be significantly different from its asymptotic behavior, thereby necessitating a comprehensive examination of the finite-sample performance. Furthermore, the rate of convergence in the finite-sample scenario is of prime importance, as it directly impacts the efficiency of the learning process.

**Remark 12** (Finite-Sample Performance). *The analysis of the convergence of Nash equilibria in this work considers the asymptotic behavior as the number of iterations goes to infinity (Theorem 3). However, understanding the finite-sample performance of the multi-task deep learning game and the convergence rate is of practical significance for real-world applications.*

Furthermore, as posited in Remark 13, we consider the possibility of multiple Nash equilibria existing within the multi-task deep learning game. Such a situation could lead to disparate performance outcomes for the game, warranting an in-depth analysis of the implications of the existence of multiple Nash equilibria. A thorough characterization of the conditions leading to the manifestation of multiple equilibria and the formulation of strategies to select the most beneficial equilibrium for a specific multi-task learning problem emerge as imperative future research directions.

**Remark 13** (Multiple Nash Equilibria). *The existence of multiple Nash equilibria may lead to different performance outcomes for the multi-task deep learning game. Analyzing the implications of multiple Nash equilibria, characterizing the conditions that give rise to them, and devising strategies to select the most suitable equilibrium for a given multi-task learning problem are essential future research directions.*

The existence of multiple Nash equilibria (see Remark 13) invites a plethora of potential performance outcomes. This multiplicity, often observed in complex game-theoretic scenarios, introduces an element of stochasticity, as the selection of equilibrium, and hence the ensuing performance, might be dependent on the initial conditions or the idiosyncrasies of the iterative process. The multiplicity of equilibria also alludes to the possibility of a diverse set of optimal strategies under various circumstances, a factor that can significantly affect the interpretation and applicability of the outcomes. Given these complexities, the systematic characterization of conditions that give rise to multiple Nash equilibria, along with the construction of strategies to select the most beneficial equilibrium, become crucial research directions.

**Remark 14** (Complexity of the Game). *In our analysis, we have assumed a relatively simple multi-task deep learning game structure, with players aiming to minimize their task-specific loss functions. However, more complex game-theoretic settings, such as hierarchical or cooperative games, may provide additional insights into the interplay between tasks in multi-task deep learning. Investigating these settings remains a promising avenue for future work.*



Our examination of the multi-task deep learning game, as detailed in Remark 14, has been predicated on a relatively simplistic model, with each player striving to minimize task-specific loss functions. However, real-world scenarios often manifest more intricate structures, mandating an exploration of more complex game-theoretic settings. For instance, the paradigm of hierarchical or cooperative games, where tasks cooperate in a certain structure or compete within coalitions, may offer a rich repository of insights into the task interplay dynamics in multi-task deep learning. These complex game structures might unveil nuanced strategies and trade-offs, enriching our understanding of multi-task deep learning games.

**Remark 15** (New Perspective on Multi-Task Deep Learning). *By framing multi-task deep learning as a game with multiple agents, this work contributes a novel perspective that highlights the interplay between tasks and the shared resources they optimize. This approach enables us to formally analyze the existence and convergence of Nash equilibria, which may provide a deeper understanding of the trade-offs and interactions between tasks. Our results (Theorems 2 and 3) lay the foundation for further study of the game-theoretic properties of multi-task deep learning and their implications for practical applications.*

The contribution of this work, encapsulated in Remark 15, hinges on the novel perspective it offers by framing multi-task deep learning as a game involving multiple agents. This conceptualization underscores the interplay between tasks and the shared resources they optimize, a perspective that not only enriches our theoretical understanding but also provides practical insights into the design and deployment of multi-task deep learning models. The game-theoretic formalism employed allows for a rigorous analysis of the existence and convergence of Nash equilibria, thereby paving the way for a more profound understanding of the trade-offs and interactions between tasks.

**Example 1** (Cooperative Games). *One potential avenue for future work is to explore cooperative games in the context of multi-task deep learning. In such a setting, tasks can form coalitions and collaborate to improve their joint performance, while still competing for shared resources. Investigating the formation of optimal coalitions and their impact on the convergence and performance of multi-task deep learning may lead to new strategies for designing and training models that better leverage the synergies between tasks.*

The exploration of cooperative games, as suggested in Example 1, constitutes a compelling avenue for future research. Cooperative games, where tasks form coalitions and collaborate while competing for shared resources, introduce a layer of complexity that mirrors the multi-faceted nature of real-world tasks. The formation of optimal coalitions, driven by both competition and cooperation, and their impact on the convergence and performance of multi-task deep learning, may offer intriguing insights into the delicate balance between synergy and competition among tasks.

**Assumption 6** (Game-Theoretic Regularization). *Another direction for future work is to incorporate game-theoretic regularization techniques into the multi-task deep learning game. By adding regularization terms that penalize the deviation of the task-specific parameters from their Nash equilibrium values, we may be able to improve the convergence properties and performance of multi-task deep learning models. Investigating the trade-off between regularization strength and model performance, as well as the impact of different regularization techniques on the convergence of Nash equilibria, constitutes a valuable research direction.*

The proposal to incorporate game-theoretic regularization techniques into the multi-task deep learning game, as proposed in Assumption 6, presents an exciting direction for future work. By introducing regularization terms that penalize the deviation of task-specific parameters from their Nash equilibrium values, we might be able to enhance the convergence properties and performance of multi-task deep learning models. This

enhancement is based on the premise that these penalties encourage tasks to “stay close” to their Nash equilibria, thereby promoting stability and improving convergence.

**Conjecture 4** (Global Convergence). *Under certain conditions, it may be possible to prove global convergence of the multi-task deep learning game to a unique Nash equilibrium. Establishing these conditions and deriving a global convergence result would be an important contribution, as it would provide stronger guarantees on the performance and convergence of multi-task deep learning models in practice.*

The conjecture presented in Conjecture 4 postulates the potential of establishing global convergence to a unique Nash equilibrium in a multi-task deep learning game, under certain specified conditions. This conjecture opens up intriguing possibilities since global convergence, if proven, would provide stronger guarantees on the performance and convergence of multi-task deep learning models. This would not only significantly contribute to the theoretical underpinnings of the field but could also have important implications for practical applications, by offering more predictable and reliable models in real-world settings.

**Assumption 7** (Game-Theoretic Stability). *As a future research direction, it would be valuable to analyze the stability of the multi-task deep learning game under various perturbations. By considering different noise models and their impact on the convergence and performance of the game, we can better understand the robustness of multi-task deep learning models and potentially design training algorithms that are more resilient to various sources of uncertainty.*

As the future research direction suggested in Assumption 7 implies, analyzing the stability of the multi-task deep learning game under different types of perturbations could provide important insights into the robustness of these models. Noise, an inescapable aspect of real-world data, might impact the convergence and performance of the game. Therefore, understanding the interaction between different noise models and the multi-task learning game is of paramount importance. A detailed examination of these interactions could lead to the design of more resilient training algorithms, which could withstand various forms of uncertainty, thus enhancing the practical applicability of multi-task deep learning models.

**Assumption 8** (Heterogeneous Tasks). *Our current analysis assumes a relatively homogeneous set of tasks in the multi-task deep learning game. However, real-world applications often involve heterogeneous tasks with different complexities, objectives, and data distributions. Extending the game-theoretic framework to handle heterogeneous tasks would provide a more realistic model for multi-task deep learning and enable the analysis of additional challenges and trade-offs that arise in such settings.*

Our analysis thus far, as delineated in Assumption 8, has been based on a relatively homogeneous set of tasks in the multi-task deep learning game. Yet, real-world applications frequently involve a diverse array of tasks, each with its own set of complexities, objectives, and data distributions. This suggests that a more realistic model for multi-task deep learning would need to incorporate this heterogeneity, thereby requiring an extension of the game-theoretic framework. Such an extension could provide a more nuanced understanding of the various challenges and trade-offs that emerge in settings involving heterogeneous tasks.

As we further explore the fascinating field of multi-task deep learning, we see that there are many potential areas where our research may be of great relevance. This study, for example, can greatly contribute to our understanding of the theoretical underpinnings of these models. For instance, several studies [39,40] apply game theory in enhancing the security and trustworthiness of wireless sensor networks in IoT, as well as in the design of efficient earthquake early-warning systems. These are insightful demonstrations of

the potency of game theory in real-world applications. Moreover, the power of deep learning in the discrimination of earthquakes and quarry blasts has been investigated [41], addressing a significant challenge in seismic hazard assessment. Such applications resonate with our assertion that understanding the game-like nature of multi-task deep learning systems could greatly assist in the development of new models. Understanding how these multi-task learning systems function as a ‘game’ between tasks could greatly assist in the development of new models. This is especially true considering that the game-theoretic approach to general multi-task deep learning models has not been precisely proved, leaving much room for exploration and improvement.

With a more profound understanding of these models, researchers are empowered to create more efficient and effective systems. They can utilize the insights from our study when designing the structures of their own models. The possibilities are vast and could encompass applications in fields such as computer vision, natural language processing, healthcare, and beyond, where multi-task deep learning models are commonly employed. In essence, our research provides the foundational theory that can stimulate the evolution of multi-task deep learning models in numerous applications, contributing to the advancements in this exciting and ever-expanding field.

## 5. Conclusions

In this paper, we have explored the intricate interplay between multiple tasks in a multi-task deep learning system through the lens of game theory. By casting the multi-task deep learning problem as a game wherein each task operates as a player with the objective of minimizing its task-specific loss function, we have managed to delve deeper into the underlying mechanics governing the interactions among various tasks sharing a common set of parameters.

This study introduces a game-theoretic approach to multi-task deep learning dynamics, providing a fresh alternative to the traditional optimization perspective. Through this approach, we illuminate the intricate interdependencies between tasks and the consequential dynamics inherent to multi-task learning, recognizing their cooperative and competitive aspects [42]. This paradigm aligns with emerging trends in deep learning research, where network interactions are becoming increasingly critical [43]. Consequently, our work not only enhances the theoretical comprehension of these dynamics but also promotes novel strategies to harness the cooperative–competitive nexus in multi-task learning.

Utilizing a formal and mathematical approach, we have introduced the concept of the Nash equilibrium for the multi-task deep learning game and investigated its existence and convergence properties. Under certain convexity and Lipschitz continuity assumptions on the loss functions, we have proven the existence of at least one Nash equilibrium for the multi-task deep learning game. Moreover, we have examined the convergence characteristics of the Nash equilibrium and ascertained conditions under which the iterative updates of task-specific and shared parameters converge to the Nash equilibrium.

Although our analysis has been carried out under idealized assumptions, such as the convexity of the loss functions, we believe that the insights gleaned from our study offer valuable contributions to the understanding of multi-task deep learning systems. Furthermore, the establishment of the existence and uniqueness of Nash equilibria in such systems provides a robust foundation for future research in this area.

We acknowledge that our study’s limitations arise from the idealized assumptions, which may not hold in real-world scenarios. However, our theoretical findings pave the way for potential extensions and research directions that could further enrich our understanding of multi-task deep learning systems. For instance, future research could explore the role of non-convex loss functions, study the implications of different types of gradient-based methods, and investigate the impact of varying degrees of task similarity on the Nash equilibrium and its convergence properties.

**Funding:** This work was supported by a research grant funded by Generative Artificial Intelligence System Inc. (GAIS).

**Data Availability Statement:** No new data were created or analyzed in this study.

**Conflicts of Interest:** The author declares no conflict of interest.

## Abbreviations

Notation	Description
$\mathcal{X}$	Input space
$\mathcal{Y}$	Output space
$x$	Input data
$\hat{y}$	Predicted output
$\mathcal{N}$	Neural network model
$\Theta$	Set of parameters
$\Theta_m$	Task-specific parameters for player $m$
$\Psi$	Shared parameters
$f_i$	Function representing the $i$ -th layer
$L$	Total number of layers in the model
$W_i$	Weight matrix of the $i$ -th layer
$b_i$	Bias vector of the $i$ -th layer
$g_i$	Activation function of the $i$ -th layer
$n_i$	Number of units in the $i$ -th layer
$h_i$	Output of the $i$ -th layer
$K_i$	Lipschitz constant of the $i$ -th layer
$\mathcal{D}$	Dataset
$N$	Number of data points in the dataset
$\ell$	Task-specific loss function
$L(\Theta)$	Overall loss function
$\alpha_t$	Learning rate at iteration $t$
$\alpha_t^{(m)}$	Learning rate for player $m$ at iteration $t$
$\beta_t$	Consensus learning rate at iteration $t$
$L_{MT}$	Multi-task loss function
$\tilde{\nabla}$	Pseudo-gradient
$\tilde{C}_m$	Lipschitz constant for the pseudo-gradient
$\tilde{C}_\Psi$	Lipschitz constant for the pseudo-gradient with respect to shared parameters
Abbreviation	Full Form
ReLU	Rectified linear unit
GELU	Gaussian error linear unit
RMSprop	Root mean square propagation
Adam	Adaptive moment estimation

## References

- Greener, J.G.; Kandathil, S.M.; Moffat, L.; Jones, D.T. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 40–55. [\[CrossRef\]](#)
- Liu, B.; Ding, M.; Shaham, S.; Rahayu, W.; Farokhi, F.; Lin, Z. When machine learning meets privacy: A survey and outlook. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–36. [\[CrossRef\]](#)
- Carleo, G.; Cirac, I.; Cranmer, K.; Daudet, L.; Schuld, M.; Tishby, N.; Vogt-Maranto, L.; Zdeborová, L. Machine learning and the physical sciences. *Rev. Mod. Phys.* **2019**, *91*, 045002. [\[CrossRef\]](#)
- Aggarwal, A.; Mittal, M.; Battineni, G. Generative adversarial network: An overview of theory and applications. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100004. [\[CrossRef\]](#)
- Chen, Y.; Yang, X.H.; Wei, Z.; Heidari, A.A.; Zheng, N.; Li, Z.; Chen, H.; Hu, H.; Zhou, Q.; Guan, Q. Generative adversarial networks in medical image augmentation: A review. *Comput. Biol. Med.* **2022**, *114*, 105382. [\[CrossRef\]](#)
- Yeom, T.; Lee, M. DuDGAN: Improving Class-Conditional GANs via Dual-Diffusion. *arXiv* **2023**, arXiv:2305.14849.
- Croitoru, F.A.; Hondru, V.; Ionescu, R.T.; Shah, M. Diffusion models in vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**. [\[CrossRef\]](#)
- Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; Norouzi, M. Palette: Image-to-image diffusion models. In Proceedings of the ACM SIGGRAPH 2022 Conference Proceedings, Vancouver, BC, Canada, 7–11 August 2022; pp. 1–10.
- Blattmann, A.; Rombach, R.; Oktay, K.; Müller, J.; Ommer, B. Retrieval-augmented diffusion models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 15309–15324.

10. Wolleb, J.; Sandkühler, R.; Bieder, F.; Valmaggia, P.; Cattin, P.C. Diffusion models for implicit image segmentation ensembles. In Proceedings of the International Conference on Medical Imaging with Deep Learning, Zurich, Switzerland, 6–8 July 2022; pp. 1336–1348.
11. Kim, J.; Lee, M. Class-Continuous Conditional Generative Neural Radiance Field. *arXiv* **2023**, arXiv:2301.00950.
12. Barron, J.T.; Mildenhall, B.; Verbin, D.; Srinivasan, P.P.; Hedman, P. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5470–5479.
13. Fridovich-Keil, S.; Yu, A.; Tancik, M.; Chen, Q.; Recht, B.; Kanazawa, A. Plenoxels: Radiance fields without neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5501–5510.
14. Guo, Y.C.; Kang, D.; Bao, L.; He, Y.; Zhang, S.H. Nerfren: Neural radiance fields with reflections. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18409–18418.
15. Lee, S.; Ku, H.; Hyun, C.; Lee, M. Machine Learning-Based Analyses of the Effects of Various Types of Air Pollutants on Hospital Visits by Asthma Patients. *Toxics* **2022**, *10*, 644. [[CrossRef](#)] [[PubMed](#)]
16. Coli, G.M.; Boattini, E.; Filion, L.; Dijkstra, M. Inverse design of soft materials via a deep learning-based evolutionary strategy. *Sci. Adv.* **2022**, *8*, eabj6731. [[CrossRef](#)]
17. Du, J. Mean-variance portfolio optimization with deep learning based-forecasts for cointegrated stocks. *Expert Syst. Appl.* **2022**, *201*, 117005. [[CrossRef](#)]
18. Kim, J.; Lee, M. Portfolio Optimization using Predictive Auxiliary Classifier Generative Adversarial Networks with Measuring Uncertainty. *arXiv* **2023**, arXiv:2304.11856.
19. Sharma, M.; Shekhawat, H.S. Portfolio optimization and return prediction by integrating modified deep belief network and recurrent neural network. *Knowl.-Based Syst.* **2022**, *250*, 109024. [[CrossRef](#)]
20. Tian, Y.; Su, D.; Lauria, S.; Liu, X. Recent advances on loss functions in deep learning for computer vision. *Neurocomputing* **2022**, *497*, 129–158. [[CrossRef](#)]
21. Zvarikova, K.; Horak, J.; Bradley, P. Machine and Deep Learning Algorithms, Computer Vision Technologies, and Internet of Thingsbased Healthcare Monitoring Systems in COVID-19 Prevention, Testing, Detection, and Treatment. *Am. J. Med. Res.* **2022**, *9*, 145–160.
22. Zhao, Y.; Wang, X.; Che, T.; Bao, G.; Li, S. Multi-task deep learning for medical image computing and analysis: A review. *Comput. Biol. Med.* **2022**, *153*, 106496. [[CrossRef](#)]
23. Samant, R.M.; Bachute, M.R.; Gite, S.; Kotecha, K. Framework for deep learning-based language models using multi-task learning in natural language understanding: A systematic literature review and future directions. *IEEE Access* **2022**, *10*, 17078–17097. [[CrossRef](#)]
24. Vithayathil Varghese, N.; Mahmoud, Q.H. A survey of multi-task deep reinforcement learning. *Electronics* **2020**, *9*, 1363. [[CrossRef](#)]
25. Zhou, T.; Ruan, S.; Canu, S. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array* **2019**, *3*, 100004. [[CrossRef](#)]
26. Xu, Q.; Wang, N.; Wang, L.; Li, W.; Sun, Q. Multi-task optimization and multi-task evolutionary computation in the past five years: A brief review. *Mathematics* **2021**, *9*, 864. [[CrossRef](#)]
27. De Giovanni, P.; Zaccour, G. A selective survey of game-theoretic models of closed-loop supply chains. *Ann. Oper. Res.* **2022**, *314*, 77–116. [[CrossRef](#)]
28. Dasari, V.S.; Kantarci, B.; Pouryazdan, M.; Foschini, L.; Girolami, M. Game theory in mobile crowdsensing: A comprehensive survey. *Sensors* **2020**, *20*, 2055. [[CrossRef](#)] [[PubMed](#)]
29. Habib, M.A.; Moh, S. Game theory-based routing for wireless sensor networks: A comparative survey. *Appl. Sci.* **2019**, *9*, 2896. [[CrossRef](#)]
30. Piraveenan, M. Applications of game theory in project management: A structured review and analysis. *Mathematics* **2019**, *7*, 858. [[CrossRef](#)]
31. Gavidia-Calderon, C.; Sarro, F.; Harman, M.; Barr, E.T. Game-theoretic analysis of development practices: Challenges and opportunities. *J. Syst. Softw.* **2020**, *159*, 110424. [[CrossRef](#)]
32. Reny, P.J. Nash equilibrium in discontinuous games. *Annu. Rev. Econ.* **2020**, *12*, 439–470. [[CrossRef](#)]
33. Celard, P.; Iglesias, E.; Sorribes-Fdez, J.; Romero, R.; Vieira, A.S.; Borrajo, L. A survey on deep learning applied to medical images: From simple artificial neural networks to generative models. *Neural Comput. Appl.* **2023**, *35*, 2291–2323. [[CrossRef](#)]
34. Armeniakos, G.; Zervakis, G.; Soudris, D.; Henkel, J. Hardware approximate techniques for deep neural network accelerators: A survey. *ACM Comput. Surv.* **2022**, *55*, 1–36. [[CrossRef](#)]
35. Iliina, O.; Ziyadinov, V.; Klenov, N.; Tereshonok, M. A Survey on Symmetrical Neural Network Architectures and Applications. *Symmetry* **2022**, *14*, 1391. [[CrossRef](#)]
36. Wang, Q.; Ma, Y.; Zhao, K.; Tian, Y. A comprehensive survey of loss functions in machine learning. *Ann. Data Sci.* **2022**, *9*, 187–212. [[CrossRef](#)]
37. Arpit, D.; Campos, V.; Bengio, Y. How to initialize your network? robust initialization for weightnorm & resnets. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 10902–10911.

38. Zou, F.; Shen, L.; Jie, Z.; Zhang, W.; Liu, W. A sufficient condition for convergences of adam and rmsprop. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11127–11135.
39. Abdalzaher, M.S.; Muta, O. A game-theoretic approach for enhancing security and data trustworthiness in IoT applications. *IEEE Internet Things J.* **2020**, *7*, 11250–11261. [[CrossRef](#)]
40. Abdalzaher, M.S.; Soliman, M.S.; El-Hady, S.M.; Benslimane, A.; Elwekeil, M. A deep learning model for earthquake parameters observation in IoT system-based earthquake early warning. *IEEE Internet Things J.* **2021**, *9*, 8412–8424. [[CrossRef](#)]
41. Abdalzaher, M.S.; Moustafa, S.S.; Hafiez, H.A.; Ahmed, W.F. An optimized learning model augment analyst decisions for seismic source discrimination. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5920212. [[CrossRef](#)]
42. Yang, L.; Sun, Q.; Zhang, N.; Li, Y. Indirect Multi-Energy Transactions of Energy Internet with Deep Reinforcement Learning Approach. *IEEE Trans. Power Syst.* **2022**, *37*, 4067–4077. [[CrossRef](#)]
43. She, C.; Sun, C.; Gu, Z.; Li, Y.; Yang, C.; Poor, H.V.; Vucetic, B. A Tutorial on Ultrareliable and Low-Latency Communications in 6G: Integrating Domain Knowledge Into Deep Learning. *Proc. IEEE* **2021**, *109*, 204–246. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.