Convex Relaxations for DNNs and Connection to LASSO

Faris Qadan
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332
fqadan3@gatech.edu

Abstract

Despite their exceptional performance on various complex machine learning tasks, deep neural networks (DNNs) become increasingly data-hungry, with larger datasets often leading to more powerful models. However, this over-reliance on large datasets introduces the issue of model interpretability. Various approaches have been proposed to tackle this issue, including a recent connection established between the non-convex training problems of DNNs, and the problems' convex counterparts. This work begins by exploring a foundational result in this area and introduces a framework for characterizing DNN training loss by revising an important assumption underpinning that result. Finally, suggestions for future work are provided to tailor this extension to more specific, nuanced learning applications.

1 Introduction

As data has become more available (and complex), efficient optimization has become a central problem in the field of machine learning. It is clear that deep neural networks (DNNs) [can] achieve remarkable results in all things image processing, language processing, and beyond [1]. However, these DNNs are traditionally optimized from a non-convex lens; this motivates the notion of convex relaxations for deep networks. It is clear that convex optimization is tractable, and often provides a simpler guarantee of global optimality. Naturally, the least absolute shrinkage and selection operator (LASSO) [7] presents itself as a powerful regularization tool in this process.

Throughout this project, we will explore these relationships as defined by one of the cornerstone pieces of literature connecting them. *Convex Geometry and Duality of Over-parametrized Neural Networks* [2] provides a unique interpretation of the training of two-layer rectified linear unit (ReLU) networks via the lens of convex duality.

Specifically, the authors investigate the behavior of over-parametrized two-layer ReLU networks to formally relate the network's optimal solution to the extrema of convex sets. Crucially, the latter is a very well-studied (and well-behaved!) area of mathematics; we thus greatly benefit from being able to reformulate the regularized training of these networks as convex problems.

Despite the paper's (a) novelty, and (b) relatively narrow scope, its list of contributions is significant. For the purpose of this project, however, we will focus on *Theorem 1*, which outlines how over-parametrized networks can be reformulated as dual optimization problems. Occasionally, other notes from the paper may be used, but this result – at least in the context of this exploration – holds the most weight.

The motivations for this project align closely with those of the original work. By mapping a computationally expensive and less interpretable problem into a more tractable and efficiently

Preprint. Under review.

solvable one, we stand to gain both theoretical and practical benefits in systems relying on this framework.

To be precise, the work below will focus on these convex relaxations and their connections to a LASSO regularization scheme. We will build on the paper's results by (a) exploring their robustness and context relative to the paper's overarching goal, and (b) adjusting a key assumption in these relaxations to analyze how different tools may be used in tandem with the authors' findings. Lastly, we will explore how our extensions can provide new insights or applications beyond those considered by the original authors.

2 Convex Relaxations for DNNs and Relation to LASSO

To begin, we look to understand the many results of this paper at a deeper level. Specifically, they are all underpinned by Theorem 1, which outlines how over-parametrized networks can be reformulated as dual optimization problems.

Of course, this is exciting at a high level; this section will revise the definitions and assumptions to be used in upcoming sections. While it is not meant to be a comprehensive review of the paper leading up to Theorem 1 – and certainly not one of the entire paper – it lays the groundwork for the extension of this theorem shown later in this report.

2.1 Why Reformulate Such a Well-Studied Problem?

While we have briefly motivated the need for this work above, we look to do so more formally here. Despite the benefits that come with faster training, reduced compute, and decreased carbon emissions are obvious, an emerging problem in today's largest, most complex models is their training interpretability. That is asking the questions, how can we better understand data movement through complex deep learning pipelines, and how can we use this information to ensure that (a) data remains anonymous through, (b) models remain efficient, and (c) models do not reinforce any potential data biases?

Of course, this list is not comprehensive. However, much work has been done recently, specifically targeting the interpretability of networks similar to that which we explore here. Approaches have been varied, including furthering the connection between DNNs and LASSO [5] and applying convex relaxations similar to these to more complex generative models [3]. Unsurprisingly, the latter includes Ergen and Pilanci as co-authors, explicitly referencing the results of this work as a building block towards their newer results.

This particular network formulation provides a fascinating balance between relative simplicity and nuance, making its analysis highly meaningful – and with these results, highly interpretable.

2.2 Overview and Definitions

As such, we reformulate the problem as it is presented by the authors, laying the foundation for our proposed expansion to come.

To begin, we define the problem as stated by the authors, where we look to optimize...

$$P^* = min_{\theta \in \Theta} R(\theta) \tag{1}$$

...for $R(\theta) = \|w\|_2^2 + \|U\|_F^2$, s.t. $f_\theta(A) = y$. This represents the general, non-convex DNN training problem.

Here, the two-layer ReLU network is 'applied' to A simply by $f(A) = (AU + 1b^T)_+ w$. Here, U is the hidden-layer's weight matrix, b is a bias vector, and w is the output weight vector. A here is defined as some data matrix whereby $A \in \mathbb{R}^{n \times d}$, and its columns are composed of some n samples whereby $a_i \in \mathbb{R}^d$.

Most importantly moving forward, we note that P^* as shown is non-convex, with $(.)_+ = \max(0, .)$ introducing a clear non-linearity to the problem.

Re-examining P^*, we note that it can be equivalently written as $P^* = min_{\theta \in \Theta} \|w\|_1$, s.t. $f_\theta(A)$, and $\left\|u_j\right\|_2 = 1, \forall j$. Already the connection between our ReLU characterization and the LASSO regularization term is clear.

Examining a lower-bound for P^* by some D^*, we note the dual of P^* (such that $P^* \geq D^*$). The crucial convex relaxation occurs when the authors leverage convex duality to arrive at a definition for the convex dual of P^*, whereby...

$$D^* = max_{v \in \mathbb{R}^n} v^T y \tag{2}$$

...subject to (s.t.) $\left|v^T (Au)_+\right| \leq 1 \forall u \in \mathcal{B}_2$. This result is analogous to that presented as Theorem 1 by the authors [2].

2.3 The Spike-Free Assumption and Its Implications

The authors make a key assumption about the nature of A for the purpose of their analysis: that A is spike-free. A is spike-free if and only if (iff)...

$$\mathcal{Q}_A = A\mathcal{B}_2 \cap \mathbb{R}^n_+$$

Here, we define the set $A\mathcal{B}_2 = \{Au | u \in \mathcal{B}_2\}$, and $\mathcal{B}_2$ simply as the l_2 unit ball. Note that $\mathcal{Q}_A$ is a convex set here. This is more thoroughly discussed in Lemma 3 of the paper.

The implication of this assumption for A is clear, and is one of the key building block throughout the authors' work. When A is spike-free, the convex relaxation of the ReLU function exactly matches the region defined by $\mathcal{Q}_A$. Thus, we are able to optimize over the convex set $\mathcal{Q}_A$ with relative ease, and plays a crucial role in mapping our non-convex formulation to our convex one.

This point will prove important in our later analysis of the problem.

3 Introducing Random Data and Re-characterizing ReLU Training Loss

As seen through various examples, randomness – in the context of high-dimensional statistics – is not the enemy that it is in traditional statistical settings. In fact, we stand to gain a great deal by assuming some form of randomness in our high-dimensional data. Working with random data gives us the ability to make very strong claims about different properties of our stochastic methods; most notably and relevantly to the work of Ergen and Pilanci, convergence and generalization guarantees.

One heavy hammer in the high-dimensional statistics toolbox that utilizes this fact is the Convex Gaussian Min-Max Theorem (CGMT) [6], which allows us to characterizes some random quantity $\Phi(X)$, given some relatively loose randomness requirements. To begin, we must make a slight alteration to the assumption that the authors make on the nature of their data matrix, A.

3.1 Why Reformulate Such a Well-Defined Reformulation of Such a Well-Studied Problem?

As seen in 2.1, it is clear that the initial mapping of the well-studied LASSO optimization for two-layer ReLU to a convex one is well-motivated. The extension we propose here bolsters the initial motivation, as will be shown below.

Specifically, we propose that through this extension, one is able to characterize their network in relation to a significantly more interpretable one, depending on their application. As such, we remain in line with the original motivation, making meaningful strides towards it.

Further – and perhaps more specifically to the paper in question – numerous theorems that follow from Theorem 1 assume that 0 training error is achievable (see Theorem 2 for the most immediate instance).

While this assumption greatly eases the process of analyzing these networks, especially asymptotically, it is often not reflective of the case in real-world applications, where most training occurs in resource-scarce environments (whether that resource be compute, time, energy, money; or some combination of the four).

As such, arriving at a relatively tight loss guarantee is certainly favorable, and immediately applicable to this paper's added results.

3.2 Relaxing the Spike-Free Assumption

To proceed, we introduce a new assumption on the nature of the data matrix A. Specifically, we relax the deterministic spike-free assumption, and instead model A as a Gaussian random matrix

where each entry of A is now given by $A'_{ij} \sim N(0,1)$, for some $A' \in \mathbb{R}^{n \times d}$. We argue that this newly relaxed assumption offers us two main benefits in analysis, in a similar vein to other problems explored:

1. Randomness provides predictability. More explicitly, it was shown in numerous cases (e.g. work with Wigner matrices, singular value characterization, and beyond) that randomness induces a sort of structure into our problems, often yielding results of a very precise nature. The hope is that a similar outcome is seen here.
2. Ease of analysis. Somewhat following the above, assuming some randomness in our data opens the door to use an extensive arsenal of mathematical tools, including CGMT, for powerful results.

Of course, relaxations of this nature are seldom free. In this case, we are trading deterministic results for probabilistic ones, for the benefit of new characterizations of the final training loss of these DNNs.

Before proceeding, we must show that $\mathcal{Q}_{A'}$ remains convex with this new definition of A; if so, we are able to enjoy the main findings of this paper, while further characterizing these networks' behaviors for a slightly different matrix A.

We recall that $\mathcal{B}_2$ is convex by definition. A' can tautologically be viewed as a series of linear combinations, making $A'\mathcal{B}_2$ a linear transformation of the convex set $\mathcal{B}_2$. It must thus hold that $A'\mathcal{B}_2$ is also convex. $\mathbb{R}^n_+$ is also convex, as noted, and thus, $A'\mathcal{B}_2 \cap \mathbb{R}^n_+$, which is the definition of $\mathcal{Q}_{A'}$. This completes the proof of $\mathcal{Q}_{A'}$'s convexity.

3.3 Updated Problem Setup

Now, it is clear that a network with a random data matrix A' can yield the same relaxation guarantees as A, we proceed with an updated problem setup to facilitate our use of CGMT. To begin, we recall how Ergen and Pilanci formulated their training loss, regularized by LASSO...

$$min_{U,b,w} \frac{1}{2} \left\| f(A) - y \right\|_2^2 + \beta \|w\|_1$$

...where $y \in \mathbb{R}^n$ represents our target vector and $\beta \in \mathbb{R}_+$ is some regularization constant. Substituting, first, the expanded $f(A')$, and second, our definition of A' in, we get an analogous minimization, where we optimize for...

$$min_{U,b,w} \frac{1}{2} \left\| (A'U + 1b^T)_+ w - y \right\|_2^2 + \beta \|w\|_1$$

...noting the same definitions as above. Of course, this becomes reminiscent of a classic LASSO minimization problem. As such, we are able to rewrite it by...

$$D^* = min_{U,b,w} \frac{1}{2} \left\| f(A') - y \right\|_2^2 + \beta \|w\|_1$$

...where D^* is as shown in 2. Expanding from our above two definitions, we note that...

$$D^* = \frac{1}{2} \left\| A'Uw + 1b^T w - y \right\|_2^2 + \beta \|w\|_1$$

We note that since $A'U$ is linear and non-negative by definition, we are able to omit the secondary case of the ReLU function, which induces positivity to the terms above. This allows us to simplify our argument slightly.

Considering the first term in the new formulation of D^*, and taking its convex conjugate, we can say that...

$$D_1^* = sup_v \langle v, A'Uw + 1b^T w - y \rangle - \frac{1}{2} \|v\|_2^2$$

...for $v \in \mathbb{R}^n$. Replacing this term with the first term above, we thus get that...

$$D^* = min_{U,b,w} sup_v \langle v, A'Uw + 1b^T w - y \rangle - \frac{1}{2} \|v\|_2^2 + \beta \|w\|_1$$

Relaxing this slightly, we see that this optimization becomes analogous to...

$$D^* = inf_{U,b,w} sup_v \langle v, A'Uw + 1b^T w - y \rangle - \frac{1}{2} \|v\|_2^2 + \beta \|w\|_1$$

Now, we are able to simplify by associativity of the inner product, to separate the data matrix term, such that...

$$D^* = inf_{U,b,w} sup_v \langle v, A'Uw \rangle + \langle v, 1b^T w - y \rangle - \frac{1}{2}\|v\|_2^2 + \beta \|w\|_1$$

Finally, we define some $u := Uw$, where $u \in \mathcal{U}$ and $\mathcal{U}$ is the set encompassing the hidden layers' effects on our output weights. More broadly, $\mathcal{U} \subseteq \mathbb{R}^d$.

Since A' is defined as a Gaussian standard matrix, $\mathcal{U} \subseteq \mathbb{R}^d$, and $v \in \mathbb{R}^n$, we are able to present our optimization problem in a more familiar form;

$$D^* = \Phi(A') = inf_u sup_v u^T A' v + Q(u, v) \tag{3}$$

Here, we note that...

$$Q(u, v) = \langle v, 1b^T w - y \rangle - \frac{1}{2}\|v\|_2^2 + \beta \|w\|_1 \tag{4}$$

...simply extracted from the results above.

3.4 Note on the Continuity of Q

It is desirable for $Q(u, v)$, shown in Equation 4 above, to be a continuous function. In this subsection, we prove that this is indeed the case. Breaking it into its three sub-terms, we note that...

1. $\langle v, 1b^T w - y \rangle$ forms a linear inner-product and thus, must be continuous.
2. $-\frac{1}{2}\|v\|_2^2 = -\left\|\frac{1}{4}v\right\|_2^2$, by homogeneity of norms. All norms are continuous, and thus, the same must hold here.
3. Similar to the above, $\beta\|w\|_1 = \|\beta \cdot w\|_1$, by homogeneity of norms; this term – being a norm – must also be continuous.

Hence, since Q is a linear combination of three continuous functions, it must be true that Q is also continuous.

3.5 Concentration Bounds on DNN Training Loss

Next, we look to characterize our over-parametrized DNN's training loss using the above formulation. In a minimization problem of this nature, we are particularly interested in devising an upper-bound on this loss value, in order to better understand a model's 'best' possible performance (or at least one valuable metric of it). Through CGMT, we derive a probabilistic upper-bound on the training loss, enabling a better characterization of DNN performance in such high-dimensional settings.

Following on from the above, we devise a simple Gaussian sub-process, given by $\|u\|_2 \cdot \langle g, v \rangle + \|v\|_2 \cdot \langle h, u \rangle$. Here, $g \sim N(0, I_d)$ and $h \sim N(0, I_n)$.

Using this, we note that a natural definition follows...

$$\phi(g, h) = inf_u sup_v \|u\|_2 \cdot \langle g, v \rangle + \|v\|_2 \cdot \langle h, u \rangle + Q(u, v) \tag{5}$$

Here, u, v, and Q are defined similar to the above, for Equation 3, where we recall the continuity of Q (see Section 3.4).

As such, we are able to say that...

$$P\{\Phi(A') \leq t\} \leq 2P\{\phi(g, h) \leq t\} \tag{6}$$

3.6 Result Discussion and Application

Similar to most CGMT-based results, the result in Equation 6 yields characterization of our DNN's loss when training on a random dataset, in relation to a less sophisticated, more predictable process. Following from the motivation in Section 2.1, this means that if we are able to bound the loss of a simpler, more interpretable deep neural network of a specific nature, we enjoy tight guarantees on our own network's training loss, with high probability (w.h.p.).

With regards to this bound itself, we consider a scenario where we have a simpler, more interpretable, mathematically 'compatible' (or rather, compliant to the above) model that is already pre-trained.

Our result dictates that if we can assure our data is correctly distributed, we are able to predict w.h.p. our final, worst-case loss value.

This can prove to be highly desirable with regards to resource allocation. As the carbon footprint of training such massive models increases rapidly, this framework enables one to train towards a specific loss 'milestone', rather than in a speculative fashion, as is done in many cases today[4].

4 Future Work

As was alluded to above, the primary utility of this extension will come by adapting the supporting function ϕ to the specific application in which this analysis is being carried out. To our knowledge, an extension to this work of this nature has not been done. As such, this work is designed to act as a framework for developing interpretable models (of the characteristics mentioned above), varying by application.

Following on from this, a future extension to this project could look to tie this work to a specific application whose model network is of a similar form to ϕ (whose general form is stated in 5). While it is clear that reducing complex networks and their training to simpler problems is by no means intuitive, a reasonable starting point could involve simple, well-understood networks that involve minimal non-linearities. As shown, efforts put into this reduction could pay dividends as networks grow larger and more complex at a rapid pace.

We note that similar techniques can also be employed to uncover different properties of DNN training. One such example could see a similar framework in use to quantify a model's generalizability given its dataset, compared to a better-studied one where this quantity is more thoroughly understood.

Lastly, this analysis can be expanded to further findings from the same paper. As mentioned previously many of the theorems provided build sequentially on one another. A new assumption (or set of assumptions, if combined with the former suggested future work) could provide entirely different perspectives on the results the authors arrived at.

5 Conclusion

Through this work, we explored the Ergen and Pilanci's reformulation of the non-convex DNN training objective in terms of its convex dual problem, such that the training objective became solvable by convex optimization. Further, we expanded on these findings by adjusting the spike-free data matrix assumption introduced by the author to a simpler Gaussian data matrix, instead. Proving that the solution set remained convex, we proceeded to map this updated formulation into one suitable for analysis with Convex Gaussian Min-Max Theorem, yielding a powerful result in the characterization of networks close to those proposed by the authors.

As stated before, much of this work was guided by two main motivations: better characterization of training loss for complex networks over time, and the search for more interpretable models in a time where data problems are becoming increasingly challenging and prevalant.

Our proposed extension shown above should be seen as a pleasant result for the former, and a first step towards the latter. As this specific field grows richer with work built atop Ergen and Pilanci's foundational work, these motivations will only grow clearer. We hope that the framework presented here serves as a valuable tool in traversing this rapidly growing body of work.

References

[1] Laith Alzubaidi, Jinglan Zhang, Amjad Jaleel Humaidi, Ayad Al-dujaili, Ye Duan, Omran Al-Shamma, José I. Santamaría, Mohammed Abdulraheem Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8, 2021.

[2] Tolga Ergen and Mert Pilanci. Convex geometry and duality of over-parameterized neural networks. *CoRR*, abs/2002.11219, 2020.

[3] Arda Sahiner, Tolga Ergen, Batu Ozturkler, Burak Bartan, John Pauly, Morteza Mardani, and Mert Pilanci. Hidden convexity of wasserstein gans: Interpretable generative models with closed-form solutions, 2022.

[4] Md Abu Bakar Siddik, Arman Shehabi, and Landon Marston. The environmental footprint of data centers in the united states. *Environmental Research Letters*, 16(6):064017, may 2021.

[5] Agus Sudjianto, William Knauth, Rahul Singh, Zebin Yang, and Aijun Zhang. Unwrapping the black box of deep relu networks: Interpretability, diagnostics, and simplification, 2020.

[6] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. The gaussian min-max theorem in the presence of convexity, 2015.

[7] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.