Transformers from an Optimization Perspective

Yongyi Yang*
University of Michigan
yongyi@umich.edu

Zengfeng Huang
Fudan University
huangzf@fudan.edu.cn

David Wipf
Amazon Web Services
davidwipf@gmail.com

Abstract

Deep learning models such as the Transformer are often constructed by heuristics and experience. To provide a complementary foundation, in this work we study the following problem: Is it possible to find an energy function underlying the Transformer model, such that descent steps along this energy correspond with the Transformer forward pass? By finding such a function, we can view Transformers as the unfolding of an interpretable optimization process across iterations. This unfolding perspective has been frequently adopted in the past to elucidate more straightforward deep models such as MLPs and CNNs; however, it has thus far remained elusive obtaining a similar equivalence for more complex models with self-attention mechanisms like the Transformer. To this end, we first outline several major obstacles before providing companion techniques to at least partially address them, demonstrating for the first time a close association between energy function minimization and deep layers with self-attention. This interpretation contributes to our intuition and understanding of Transformers, while potentially laying the ground-work for new model designs.

1 Introduction

Although deep learning has achieved extensive practical success [25, 26], model architectures are often motivated by heuristics and at times remain difficult to understand. As such, significant effort has recently be devoted to various ways of analyzing and reinterpreting popular deep network structures. Of particular note for our purposes here are assorted attempts to create a one-to-one association between deep model layers and the unfolded iterations of some optimization process designed to minimize an interpretable energy function [14, 15, 52, 54, 40]. In this way the forward pass can be viewed as computing, at least approximately, a representation with minimal energy, where model parameters dictate the form of this energy and can be trained via the backward pass for a downstream goal of interest, e.g., classification or regression, etc.

This so-called *unfolded optimization* perspective can provide insights into the nature of the inductive biases introduced by different architectures, while potentially serving as a guide for bespoke design choices instantiated through properties of the underlying energy function involved. In fact, modifications of the latter can lead to predictable model behaviors, which have in the past been tailored to incorporate useful regularization factors [58], avoid subtle degenerate solutions [54], or design effective new models altogether [1, 14].

Despite these successes, the majority of prior work in this area has either addressed relatively simple network structures such as MLPs [52], or else restricted consideration to a single model component in isolation from a larger system [40]. Consequently, complex models such as the widely-used Transformer [48] have been mostly ignored, in large part because of the difficulty involved in simultaneously mapping both the self-attention and feed-forward network (FFN) modules to an integrated optimization process that minimizes a single, unified energy function.

*Work completed during an internship at the AWS Shanghai AI Lab.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

We take significant steps towards extending the unfolded optimization perspective to a full Transformer stack via the following contributions:

- After providing background details and existing examples of unfolded optimization, Section 2 formalizes the four key challenges to mapping Transformer layers to optimization steps descending a well-specified energy function. These include: (i) handling self-attention, (ii) integrating heterogeneous Transformer layer modules (i.e., self-attention and FFN), (iii) accounting for non-linear activations, and (iv) allowing for arbitrary/unstructured weight matrices.
- Later, Section 3 derives an energy function that closely reproduces self-attention per Challenge (i), Section 4 rigorously derives convergence results which demonstrate how to approximately handle Challenge (ii) via a novel form of alternating minimization, and Section 5 provides explicit details for solving Challenge (iii) via proximal methods; consideration of Challenge (iv) is deferred to the supplementary along with additional supporting analyses.
- The above contributions culminate in an energy function that is minimized by Transformer layers when certain technical conditions hold. To provide further motivation, Section 6 empirically demonstrates that indeed this energy is minimized with real-world data even when technical conditions are difficult to formally verify.

Collectively, these results provide a complementary foundation for understanding Transformer models, and suggest plausible avenues for future enhancements.

2 Brief Intro to Deep Architectures Formed from Unfolded Optimization

The basic idea of the unfolded optimization perspective is to create a one-to-one correspondence between the layers of a deep model and the iterations of some algorithmic process designed to minimize a parameterized, model-specific energy function. In this way, the forward pass can be viewed as computing an approximation to the minimum energy solution. Meanwhile, assuming each forward iteration is differentiable w.r.t. the energy function parameters, we can pass gradients of some meta objective function (e.g., classification, regression, etc.) through the approximate minimization steps for the backward pass, which forms an interpretable bilevel-optimization process [11, 14, 24]. Overall then, unfolded optimization produces minimal energy representations, whereby the specific structure of the energy function is trainable per the relevant downstream application domain.

2.1 Mathematical Formulation

The bilevel optimization process described above can be expressed more formally as:

$$Y^*(W,X) = \arg\min_{Y} E(Y;W,X), \tag{1}$$

$$W^* = \arg\min_{W} \frac{1}{S}\sum_{i=1}^{S} \ell_i(\psi[Y^*(W,X_i)]). \tag{2}$$

Here S is the size of dataset, ℓ_i is the i-th loss function and $X_i \in \mathbb{R}^{n\times d}$ is the i-th input instance (e.g., input sentence to a language model), where n is the number of tokens and d is the token dimensionality. We refer to $Y \in \mathbb{R}^{n\times d}$ as the hidden representation of the model, where Y^* is the (ideal) model output (e.g. set of word representations output by a Transformer encoder) that minimizes to lower-level energy function[2] $E : \mathbb{R}^{n\times d} \to \mathbb{R}$ whose form is dictated by parameters W (e.g., these could also correspond with the trainable parameters in a Transformer model). And finally, ℓ is a meta-loss function (e.g., the loss function used to train a Transformer) while $\psi : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$ is an arbitrary transformation that converts hidden representations to the model output, which can also potentially have its own learnable parameters. Assuming $\partial Y^*(W,X)/\partial W$ and $\partial \ell/\partial Y^*(W,X)$ are computable, where the former involves passing gradients through the iterative steps used to approximate $Y^*(W,X)$, then the combined bilevel system can be optimized w.r.t. W.

Note that hereinafter if the energy function does not include inter-token interactions, we view it as a function of vectors, and we use lower-case characters $\mathbf{y}$ and $\boldsymbol{x}$ to represent an arbitrary row of Y and

[2]Note that although we frequently the term "energy", the concept is different from what has been called "energy-based learning" in the past, where the energy is defined on the joint input and output spaces [2, 27]; in our case energy is just a function for explaining the forward pass.

X for simplicity. Also, when clear from context, we omit writing the parameters of a function, for example we may write $E(Y)$ to reference $E(Y; W, X)$.

2.2 Related Work and Limitations

In this section we present related work on unfolded optimization, and in particular, analyze some of the relevant limitations that serve as motivation for our efforts and underscore the challenges involved.

Optimization Induced Feed-Forward Networks A wide variety of prior work has investigated feed-forward structures from an unfolded optimization perspective [14, 16, 20, 22, 46, 52]. Here we take a closer look at [52] which is based on proximal methods. The basic energy function in [52] is

$$E(\mathbf{y}) = \mathbf{1}^\top \tilde{\sigma}^*(W^{-\top}\mathbf{y}) - \left\langle f(\boldsymbol{x}), W^{-\top}\mathbf{y} \right\rangle - \frac{1}{2}\|\mathbf{y}\|^2, \tag{3}$$

where $\mathbf{1}$ is a vector with all entries being 1, $\tilde{\sigma}^*$ is the convex conjugate [41][3] of $\tilde{\sigma} : a \mapsto \int_0^a \sigma(r)\mathrm{d}r$, σ is an activation function and $f(\boldsymbol{x})$ is some transformation of input prepresentation $\boldsymbol{x}$. The proximal operator [12] of (3) is

$$\mathbf{y}^{(t+1)} = \mathrm{prox}_E\left(\mathbf{y}^{(t)}\right) = W^\top \sigma\left(W\mathbf{y}^{(t)} + f(\boldsymbol{x})\right). \tag{4}$$

Note that although W can in principle be arbitrary matrix as long as the dimensionality is properly aligned, after stacking multiple layers of this model, the effective transformation is actually forced to be constrained:

$$\mathbf{y}^{(t+2)} = W^\top \sigma\left[WW^\top \sigma\left(W\mathbf{y}^{(t)} + f(\boldsymbol{x})\right) + f(\boldsymbol{x})\right], \tag{5}$$

where clearly after the first layer the actual feed-forward transformation becomes $WW^\top$, which is necessarily positive semi-definite (PSD). Hence unconstrained layer weights are not actually possible within this paradigm.

Optimization Induced Attention Mechanisms There is also limited work that attempts to interpret attention mechanisms from the unfolded optimization perspective [15, 40]. For example, the energy function

$$E(\mathbf{y}) = -\beta^{-1}\log\left(\sum_{i=1}^{d} e^{\beta S_i \mathbf{y}}\right) + \frac{1}{2}\|\mathbf{y}\|^2, \tag{6}$$

was proposed in [40], with subsequent updating using the concave convex procedure [56] leading to iterations of the form

$$\mathbf{y}^{(t+1)} = S\,\mathrm{softmax}(\beta S \mathbf{y}^{(t)}), \tag{7}$$

where $S \in \mathbb{R}^{n\times n}$ corresponds with the attention keys while $\mathbf{y}$ maps to the attention query, and β is a constant scalar. Critically though, this attention mechanism is actually *cross-attention* (using $\mathbf{y}$ to attend to S), which does not well-align with the typical self-attention use-case of Transformers. Moreover, this work does not account for the aggregated Transformer feed-forward network module and attendant nonlinearity.

Optimization Induced Graph Neural Networks A variety of graph neural network architectures have also been developed from a similar optimization perspective [8, 30, 33, 36, 53, 54, 57, 59]. For example, graph attention mechanisms were derived using the iterative reweighted least squares (IRLS) algorithm in [54]. While this result is related to self-attention, which can be viewed as graph attention on fully connected graphs [5], it fails to produce the Transformer softmax term or the combined self-attention/feedforward Transformer stack. Nonetheless, we will later demonstrate how ideas from [54] can be leveraged to derive self-attention with softmax.

2.3 Key Challenges Extending to General Transformers

In this paper, we focus on a more complex model than prior work, namely, the Transformer encoder. A Transformer layer [48] is typically composed of two primary components: the self-attention layer and the feed-forward network (FFN) layer. If we simplify the FFN to a single linear transformation

[3] Here σ^* is a scalar function, but for simplicity, we sometimes apply a scalar function to a vector or matrix in an entry-wise fashion for convenience.

with non-linear activation, and ignore the layer-norm and residual connections,[4] one Transformer layer reduces to the basic form

$$Y^{(t+1)} = \mathrm{ReLU}\left[\mathrm{softmax}\left(Y^{(t)}W_a Y^{(t)\top}\right)Y^{(t)}W_f\right], \tag{8}$$

where W_a and W_f represent self-attention and FFN layer weight matrices respectively. Regarding (8) and the limitations of prior work discussed above, we propose four major challenges in formally applying unfolded optimization to the Transformer setting:

1. *Token-Level Interactions*: As mentioned previously, Transformer models include not only a feed-forward process but also cross-token interactions, which is reflected by the self-attention mechanism considered to be the *sine qua non* of Transformers. And thus far, no prior work has derived Transformer-style self-attention, instead either relaxing to cross-attention [40], or failing to produce the ubiquitous softmax operator [54].
2. *Heterogeneous Layer Types*: Each Transformer encoder layer is composed of two totally different components: self-attention and a FFN. We refer to this structure as a "heterogeneous layer type," where each component has its own parameters and can be viewed as a distinct unfolded optimization process. However, it remains unknown whether or not it is possible to combine the corresponding energies of these two components to obtain a unified objective with any kind of convergence guarantees (even approximately) during the aggregated forward pass.
3. *Non-Linear Activations*: Many activation functions used in common neural network architectures can be understood from the perspective of proximal operators [28, 54], at least within the isolated context of some relatively simple feedfoward schema. However, when integrated within the heterogeneous Transformer layer types mentioned above, the optimization process becomes rather complex, and it is unknown if any convergence properties still hold after the inclusion of a proximal step.
4. *Asymmetry of Weights*: As alluded to in the discussion of (5), the scope of most work addressing feed-forward networks is actually limited to models with symmetric (or even tighter, PSD) weight transformations [1, 14, 52], which restricts the resulting universality. Additionally, related models from the graph neural network literature are generally predicated on undirected graphs [53, 54], where the graph propagation is also symmetric. Consequently, although to our knowledge not addressed/discussed previously, it remains unknown how to construct energy functions for more general asymmetric transformations.

In this paper, we propose techniques to at least partially solve Challenges 1, 2 and 3, while for 4, we defer preliminary discussion to the supplementary, leaving formal investigation as a future direction.

3 A New Derivation of Transformer Self-Attention

We now tackle Challenge 1 by constructing an energy function whose iterative optimization steps match Transformer-style softmax self-attention on fully connected tokens as is customary.

3.1 Basic Softmax Self-Attention via Unfolded Optimization Steps

Consider the energy function

$$E_1(Y) = \sum_{i=1}^{n}\sum_{j=1}^{n}\rho\left(\frac{1}{2}\|\mathbf{y}_i - \mathbf{y}_j\|^2\right) + R(Y), \tag{9}$$

where $\mathbf{y}_i \in \mathbb{R}^{d\times 1}$ is the i-th row of matrix $Y \in \mathbb{R}^{n\times d}$, $\rho : \mathbb{R}^+ \to \mathbb{R}$ is a concave non-decreasing function, and $R : \mathbb{R}^{n\times d} \to \mathbb{R}$ is a convex function. Interestingly, although E_1 is not necessarily convex because of the non-convexity of ρ, under specific choices of ρ and R, it can be optimized through a softmax-like structure as follows:

[4] Note that we are ignoring layer-norms and residual connections for simplicity of exposition; however, the analysis we present herein can naturally be generalized to accommodate layer-norm, and also certain forms of residual connections with additional assumptions; see supplementary for details.

Theorem 3.1. *Assume* $\rho(z) = -\exp\{-z\}$, $R(Y) = \frac{1}{2}\|Y\|_{\mathcal{F}}^2$, *and* $\boldsymbol{\beta}_i = \exp\left\{-\frac{1}{2}\left\|\mathbf{y}_i^{(t)}\right\|^2\right\}$, *and let* $Y^{(t)}$ *represent any fixed value for* Y. *Then the update rule*

$$\mathbf{y}_i^{(t+1)} = \frac{\sum_{j=1}^n \boldsymbol{\beta}_j \exp\left\{\mathbf{y}_i^{(t)\top}\mathbf{y}_j^{(t)}\right\}\mathbf{y}_j^{(t)}}{\sum_{j=1}^n \boldsymbol{\beta}_j \exp\left\{\mathbf{y}_i^{(t)\top}\mathbf{y}_j^{(j)}\right\}}, \quad \forall i, \tag{10}$$

satisfies $E_1\left(Y^{(t+1)}\right) \leq E_1\left(Y^{(t)}\right)$ *with equality iff* $Y^{(t)}$ *is a stationary point of* E_1.

It is worth mentioning that, although perhaps not obvious, the update step (10) can be generated via a majorization-minimization (MM) algorithm [47], where the majorization step produces a convex upper-bound, and the minimization step descends along the gradient of the upper-bound. Therefore, the core of this update is essentially a gradient step on a convex function, which will be relevant to the discussion in subsequent sections; see the proof and further details in the supplementary.

Remark 3.2. Several notable generalizations of Theorem 3.1 are possible. First, although in Theorem 3.1 we adopt a specific form for ρ and R to recover the softmax operator, with other selections of these functions the corresponding unfolded optimization algorithm can generate different/novel types of attention mechanisms. Secondly, obtaining (10) relies on a particular choice of gradient step size; however, for broader choices the resulting convergent updates interestingly lead to residual connections as a natural byproduct of the optimization trajectory (see supplementary). And finally, Theorem 3.1 can be easily modified to accommodate situations whereby full Transformer connectivity is constrained by some graph structure as in [10, 18] (again, see supplementary for details).

3.2 Extension to Include Trainable Parameters

After aggregating into matrix form, we have thus far shown that the iteration

$$Y^{(t+1)} = \mathrm{softmax}_{\boldsymbol{\beta}}\left(Y^{(t)}Y^{(t)\top}\right)Y^{(t)}, \tag{11}$$

will reduce (or leave unchanged) the energy from (9), where $\mathrm{softmax}_{\boldsymbol{\beta}}(\mathbf{y})_i = \frac{\boldsymbol{\beta}_i \exp\{\mathbf{y}_i\}}{\sum_j \boldsymbol{\beta}_j \exp\{\mathbf{y}_j\}}$ denotes a softmax operator with reweighting coefficient vector $\boldsymbol{\beta}$. If $\boldsymbol{\beta}_i$ is independent of i, i.e., $\left\|\mathbf{y}_i^{(t)}\right\|$ is constant (which is enforceable via layer normalization that can also be included within our framework as mentioned previously), then this reweighted softmax is equivalent to the canonical softmax used in Transformers.

Now consider the reparameterization $Y = ZW_a$, where $W_a \in \mathbb{R}^{d\times d}$ is an invertible matrix. It follows that $Z^{(t+1)}W_a = \mathrm{softmax}_{\boldsymbol{\beta}}\left(Z^{(t)}W_aW_a^\top Z^{(t)\top}\right)Z^{(t)}W_a$, leading to the revised update

$$Z^{(t+1)} = \mathrm{softmax}_{\boldsymbol{\beta}}\left(Z^{(t)}W_a^sZ^{(t)\top}\right)Z^{(t)}, \tag{12}$$

where $W_a^s = W_aW_a^\top$ and we have adopted the superscript 's' to indicate that this matrix is symmetric. Collectively then, the results of this section directly address Challenge 1, closely reproducing softmax-styled self-attention both with and without trainable parameters (the primary lingering limitation of symmetric weights being relegated to Challenge 4).

4 Combining Transformer Components via Alternating Inexact Minimization

We next address Challenge 2 and the heterogeneous Transformer layer types. With this goal in mind, we first introduce a general optimization scenario. Specifically, we ask the following question:

Given two (convex) objectives $f(\mathbf{y})$ *and* $g(\mathbf{y})$, *under what conditions, and/or to what extent, will alternatively taking separate gradient steps w.r.t.* f *and* g *optimize the aggregated function* $f+g$*?*

We refer to this optimization strategy as *alternating inexact minimization* (AIM), and we will shortly demonstrate conditions under which it converges to a ball with finite radius containing the optimal point of $f+g$. Later we discuss how these results contribute towards the resolution of Challenge 2.

We emphasize upfront that AIM as so-defined is quite different from what is commonly referred to as alternating minimization in the literature [19, 32]. The latter refers to scenarios whereby a unified objective function with multiple variables is minimized in an alternating fashion over one variable at a time with the others fixed, such that descent can be trivially achieved. In contrast, our AIM scenario involves multiple objective terms with a *shared* variable, and we alternate minimization over each term in isolation using the same variable, a much more challenging process to analyze.

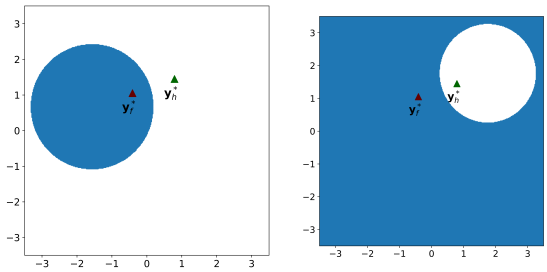

Figure 1: A two-dimensional illustration of $\mathcal{S}(\mathscr{C})$ with different values of $\mathscr{C}$. Left: $\frac{c_h \mathscr{C}}{L_f} = 0.7$. Right: $\frac{c_h \mathscr{C}}{L_f} = 1.5$. The blue area is $\mathcal{S}(\mathscr{C})$.

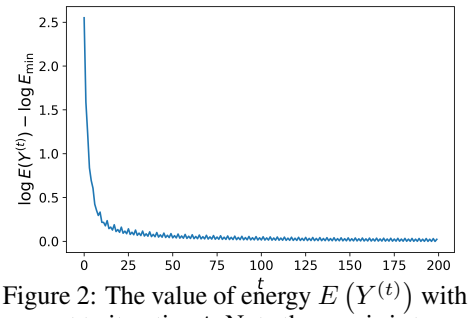

Figure 2: The value of energy $E\left(Y^{(t)}\right)$ with respect to iteration t. Note the y-axis is translated by a constant $\log E_{\min}$ to make the minimum align with zero.

4.1 General Alternating Inexact Minimization

Given two objectives $f, g : \mathbb{R}^d \to \mathbb{R}$, AIM is formalized as Algorithm 1, where α_1 and α_2 constitute step sizes. We will now investigate how Algorithm 1 relates to minimization of the combined objective defined by $h(\mathbf{y}) = f(\mathbf{y}) + g(\mathbf{y})$.

Algorithm 1

For the t-th iteration, execute $\quad \boldsymbol{u}^{(t)} = \mathbf{y}^{(t)} - \alpha_1 \nabla f(\mathbf{y}^{(t)}); \quad \mathbf{y}^{(t+1)} = \boldsymbol{u}^{(t)} - \alpha_2 \nabla g(\boldsymbol{u}^{(t)})$.

In the following, we assume that f, g are both Lipschitz smooth and strongly convex [41], with Lipschitz constants L_f and L_g respectively, and convexity parameter c_f and c_g respectively. Therefore, h will also be Lipschitz smooth and strongly convex, with Lipschitz constant L_h and convexity c_h. We denote the optimal points of f, g, and h as $\mathbf{y}_f^*$, $\mathbf{y}_g^*$, and $\mathbf{y}_h^*$ respectively.

Combining the two steps of Algorithm 1 gives

$$\mathbf{y}^{(t+1)} = \mathbf{y}^{(t)} - \alpha_1 \nabla f\left(\mathbf{y}^{(t)}\right) - \alpha_2 \nabla g\left[\mathbf{y}^{(t)} - \alpha_1 \nabla f\left(\mathbf{y}^{(t)}\right)\right], \tag{13}$$

while a canonical gradient descent step on objective h is $\mathbf{y}^{(t+1)} = \mathbf{y}^{(t)} - \alpha_2 \left[\nabla f\left(\mathbf{y}^{(t)}\right) + \nabla g\left(\mathbf{y}^{(t)}\right)\right]$. Comparing the two update rules it is apparent that Algorithm 1 can be viewed as a noisy gradient descent step with step size α_2 and noise factor

$$\Delta_t = \nabla h\left(\mathbf{y}^{(t)}\right) - \frac{\alpha_1}{\alpha_2} \nabla f\left(\mathbf{y}^{(t)}\right) - \nabla g\left[\mathbf{y}^{(t)} - \alpha_1 \nabla f\left(\mathbf{y}^{(t)}\right)\right]. \tag{14}$$

Noisy gradient descent is a well-studied problem in certain contexts [4, 35]. However, in our specific scenario we require a novel, different (actually tighter) bound than one can get from simply applying existing results.

Specifically, we demonstrate that when $\delta\left(\mathbf{y}^{(t)}\right) = \|\Delta_t\| / \left\|\nabla h\left(\mathbf{y}^{(t)}\right)\right\|$ is bounded, (13) is guaranteed to descend the objective h as follows:

Theorem 4.1. *When $\alpha_1 \leq \alpha_2 \leq L_h^{-1}$, suppose $\mathbf{y}^{(t)}$ and $\mathbf{y}^{(t+1)}$ are related by (13), with $\delta\left(\mathbf{y}^{(t)}\right) \leq \mathscr{C}$ and $\mathscr{C} = \frac{\alpha_2}{\alpha_2 - \alpha_1 + \alpha_1 \alpha_2 L_g}$. Then $h\left(\mathbf{y}^{(t+1)}\right) \leq h\left(\mathbf{y}^{(t)}\right)$.*

Remark 4.2. Although the definition of $\mathscr{C}$ may appear rather complicated, when one of α_1, α_2, L_f, or L_g is sufficiently small, $\mathscr{C}$ behaves as $\Omega\left([\alpha_1 \alpha_2 L_f L_g]^{-1}\right)$.

We further consider how to interpret the constraint $\delta(\mathbf{y}) \leq \mathscr{C}$ and the region within which (13) can optimize h as follows:

Lemma 4.3. *Let $\mathcal{S}(\mathscr{C}) = \left\{\mathbf{y} \,\middle|\, \frac{\|\mathbf{y} - \mathbf{y}_f^*\|}{\|\mathbf{y} - \mathbf{y}_h^*\|} \leq \frac{c_h \mathscr{C}}{L_f}\right\}$. When $\mathbf{y}^{(t)} \in \mathcal{S}(\mathscr{C})$, $\delta(\mathbf{y}) \leq \mathscr{C}$.*

Combining Theorem 4.1 and Lemma 4.3, we can therefore conclude that when $\mathbf{y}^{(t)} \in \mathcal{S}(\mathscr{C})$, $h\left(\mathbf{y}^{(t+1)}\right) \leq h\left(\mathbf{y}^{(t)}\right)$. Note that the boundary of $\mathcal{S}(\mathscr{C})$ given in Lemma 4.3 is called an Apollonian circle [42]. When $\mathscr{C} \leq \frac{L_f}{c_h}$, $\mathcal{S}(\mathscr{C})$ is a ball centered on $\mathbf{y}_f^*$, and when $\mathscr{C} \geq \frac{L_f}{c_h}$, $\mathcal{S}(\mathscr{C})$ is the whole space excluding a ball centered on $\mathbf{y}_h^*$. Figure 1 provides a 2D visualization for each case respectively.

Additionally, note that the same analysis can be done by switching the role of f and g (since the two processes are alternating) and further restrict the excepted region, although here we omit this for simplicity.

Remark 4.4. Our findings above can be summarized as follows: For sufficiently small values of α_1, α_2, Algorithm 1 reduces the combined objective h, at least provided that $\mathbf{y}$ is a certain distance away from the optimal point $\mathbf{y}_h^*$.

To illustrate this conclusion, we present a synthetic example with $f(Y) = \|SY\|_{\mathcal{F}}^2 + \|Y - B_1\|_{\mathcal{F}}^2$ and $g(Y) = \|YW\|_{\mathcal{F}}^2 + \|Y - B_2\|_{\mathcal{F}}^2$. Note that here we expand the variable $\mathbf{y}$ to a matrix Y and let f and g consist of left and right transformations of Y respectively, with an extra bias term to prevent the degenerate solution ($Y^* = 0$). We randomly set each entry of $\{S, W, B_1, B_2\}$ and execute Algorithm 1 with fixed step sizes and project the trace of $Y^{(t)}$ to a two-dimensional space with PCA for visualization. The trace of $Y^{(t)}$ is displayed in Figure 3 and the combined objective h across iterations in Figure 2. From these plots, the behaviour predicted by our theory can be verified: When $Y^{(t)}$ is a sufficient distance away from Y_h^* (the energy is relatively high), the optimization trajectory moves closer to Y_h^* (descending the energy), and after $Y^{(t)}$ is close enough to Y_h^* (and the h is at a relatively low level), the energy oscillates in a certain range about the optimal solution.

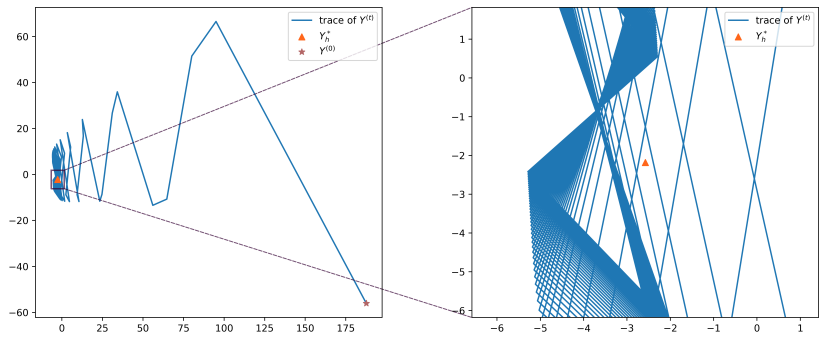

Figure 3: The trace of $Y^{(t)}$ projected to a plane. The figure on the right displays a zoomed local region around Y_h^* taken from the figure on the left.

4.2 Unfolded Optimization of Heterogeneous Layer Type Models

We now return to Transformer-specific models. Let $E_2(Y) = \frac{1}{2}\mathrm{Tr}\left(YW_fY^\top\right) + \frac{1}{2}\|Y\|_{\mathcal{F}}^2$ and consider the combined energy $E(Y) = E_1(YW_a) + E_2(Y)$ (analogous to h from the previous section) and alternatively perform the following AIM updates:

$$U^{(t)} = \mathrm{softmax}_{\boldsymbol{\beta}}\left(Y^{(t)\top}W_a^sY\right)Y^{(t)}; \tag{15}$$

$$Y^{(t+1)} = U^{(t)} - \alpha_2 \left.\frac{\partial E_2}{\partial Y}\right|_{Y=U^{(t)}} = U^{(t)}W_f^s, \tag{16}$$

where $W_f^s = (1-\alpha_2)I - \alpha_2\frac{W_f + W_f^\top}{2}$ provides an additional linear transformation after the softmax.

From Section 3, we know that (15) is optimizing E_1 using essentially a row-wise gradient descent step after majorization. Likewise, (16) is also a gradient step with step size α_2. Therefore, the combined rules fall into the scope of Algorithm 1, and from the analysis in Section 4.1, we can draw the following conclusion:

Corollary 4.5. *If $Y^{(t+1)}$ is computed via (15) and (16) with input $Y^{(t)}$, then $E\left(Y^{(t+1)}\right) \leq E\left(Y^{(t)}\right)$ when $Y^{(t)} \notin \mathcal{S}$, where $\mathcal{S}$ is a ball with finite radius containing $Y_{\hat{E}}^*$, the optimal point of $\hat{E}$, the convex upper bound of E.*

Note that after combining (15) and (16), the update aggregated update rule is already very similar to our target from (8), the only differences being the missing non-linearity and symmetric weights (note we use symmetric weight matrix W_a^s and W_f^s instead of W_a and W_f here), which correspond with Challenges 3 and 4 in Section 2.3 respectively; we address the former next.

5 Non-linear Activations Integrated within the Unfolding Paradigm

Handling non-linear activations within an unfolded optimization setting has been considered previously [16, 46, 50, 54]. However, prior work has largely relied on proximal operators to create non-linear activations paired with simple linear filters, and the analysis does not transfer to the more complex Transformer case under investigation here. While we will restrict our attention to adding ReLU activations in this section, both for simplicity of exposition and because of their ubiquity in Transformer models, our results generalize to broader selections.

Similar to Section 4, we first study a general form of the optimization problem by introducing a ReLU activations into Algorithm 1, formulating this modification as AIM with proximal steps, and then analyzing convergence under appropriate constraints. Later we apply these results to the Transformer.

5.1 Proximal-Alternating Inexact Minimization

Algorithm 2

For the t-th iteration, execute
$\boldsymbol{u}^{(t)} = \mathbf{y}^{(t)} - \alpha_1 \nabla f\left(\mathbf{y}^{(t)}\right); \quad \boldsymbol{v}^{(t)} = \boldsymbol{u}^{(t)} - \alpha_2 \nabla g\left(\boldsymbol{u}^{(t)}\right); \quad \mathbf{y}^{(t+1)} = \mathrm{ReLU}\left(\boldsymbol{v}^{(t)}\right).$

It has already been established [28, 46] that ReLU activations can be modeled as a proximal operator via

$$\mathrm{ReLU}(\mathbf{y}) = \arg\min_{\boldsymbol{z}} \frac{1}{2\lambda}\|\boldsymbol{z} - \mathbf{y}\|^2 + \phi(\boldsymbol{z}), \tag{17}$$

where ϕ is the indicator function for the first quadrant given by $\phi(z) = \begin{cases} +\infty & \text{if } z < 0 \\ 0 & \text{if } z \geq 0 \end{cases}$. Moreover, in Section 4, we have demonstrated how the steps of Algorithm 1 together form an inexact gradient descent iteration of the loss $h(\mathbf{y}) = f(\mathbf{y}) + g(\mathbf{y})$ with noisy term Δ_t defined in (14). Here, with the addition of the proximal step in Algorithm 2, we obtain an inexact version of proximal-gradient descent. In fact, one turn of Algorithm 2 is equivalent to

$$\mathbf{y}^{(t+1)} = \arg\min_{\boldsymbol{z}} \frac{1}{2\lambda}\left\|\boldsymbol{z} - \mathbf{y}^{(t)} + \alpha_2 \nabla h\left(\mathbf{y}^{(t)}\right) - \alpha_2 \Delta_t\right\|^2 + \phi(\boldsymbol{z}), \tag{18}$$

which is an inexact version of canonical proximal-gradient step with noise Δ_t, as compared to the exact version $\mathbf{y}^{(t+1)} = \arg\min_{\boldsymbol{z}} P\left(\boldsymbol{z}; \mathbf{y}^{(t)}\right)$, where $P(\boldsymbol{z}; \mathbf{y})$ is the proximal problem

$$P\left(\boldsymbol{z}; \mathbf{y}^{(t)}\right) = \frac{1}{2\lambda}\left\|\boldsymbol{z} - \mathbf{y}^{(t)} + \alpha_2 \nabla h\left(\mathbf{y}^{(t)}\right)\right\|^2 + \phi(\boldsymbol{z}). \tag{19}$$

While various forms of inexact proximal gradient descent have been studied in the past [13, 17, 43, 51], existing work either assumes constant noise [13], stochastic noise [51], or decreasing/convergent noise [17, 43]. Critically, no prior work that we are aware of applies to our case where the noise can potentially increase with iterations. Moreover, existing analysis in the literature is primarily concerned with convergence to a fixed point, while in our scenario, we instead consider entering a specific region formed around certain points.

In addition to bounding $\delta\left(\mathbf{y}^{(t)}\right)$ as in Section 4, we also need to bound the similarity between the current position $\mathbf{y}^{(t)}$ and the gradient $\alpha_2 \nabla h\left(\mathbf{y}^{(t)}\right)$, which is defined as $\mathfrak{D}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = \frac{1}{\|\boldsymbol{\xi}_1\|^2}\sum_{i=1}^{d} \min(\boldsymbol{\xi}_{2,i}^2 - \boldsymbol{\xi}_{1,i}^2, 0)$. Intuitively, $\mathfrak{D}(\boldsymbol{\xi}_1; \boldsymbol{\xi}_2)$ is defined such that each term of the summation is negative, but only close to 0 when $\boldsymbol{\xi}_{1,i}$ and $\boldsymbol{\xi}_{2,i}$ are both large. We then have the following:

Theorem 5.1. *If $\alpha_1 \leq \alpha_2 \leq L_h^{-1}$, $\mathfrak{D}\left(\alpha_2 \nabla h\left(\mathbf{y}^{(t)}\right); \mathbf{y}^{(t)}\right) \geq -\kappa$ for any $\kappa \in (0, 1)$, and $\delta\left(\mathbf{y}^{(t)}\right) \leq \mathscr{C}'$, where $\mathscr{C}' = \frac{\alpha_2 c_P \lambda \sqrt{1-\kappa}}{\sqrt{2}(\alpha_2 - \alpha_1 + \alpha_1 \alpha_2 L_g)}$, we have that $h\left(\mathbf{y}^{(t+1)}\right) + \phi\left(\mathbf{y}^{(t+1)}\right) \leq h\left(\mathbf{y}^{(t)}\right) + \phi\left(\mathbf{y}^{(t)}\right)$.*

Intuitively, Theorem 5.1 shows that the region where the descent of $h(\mathbf{y}) + \phi(\mathbf{y})$ is guaranteed is the intersection of $\mathcal{S}(\mathscr{C}')$ with $\mathcal{S}$ defined in Lemma 4.3, and the area $\mathcal{T}(\kappa) = \{\mathbf{y} \,|\, \mathfrak{D}(\alpha_2 \nabla h(\mathbf{y}); \mathbf{y}) \geq -\kappa\}$. While the shape of $\mathcal{T}(\kappa)$ remains difficult to specify in general, we note that $\mathcal{T}(\kappa)$ tends to the whole space when $\alpha_2 \to 0$ or $\kappa \to 1$. For example, we illustrate the area of $\mathcal{T}(\kappa)$ using a 2D synthetic example, where the energy function is $h(\mathbf{y}) = \|W\mathbf{y}\|_{\mathcal{F}}^2 + \|\mathbf{y} - \boldsymbol{b}\|^2$, and entries of W and $\boldsymbol{b}$ are randomly generated. See Figure 4 for the visualization using different values κ, which shows that when κ is sufficiently small, $\mathcal{T}(\kappa)$ is nearly the whole space (except a small area around the origin) which ensures the descent of $h(\mathbf{y}) + \phi(\mathbf{y})$ in most cases.

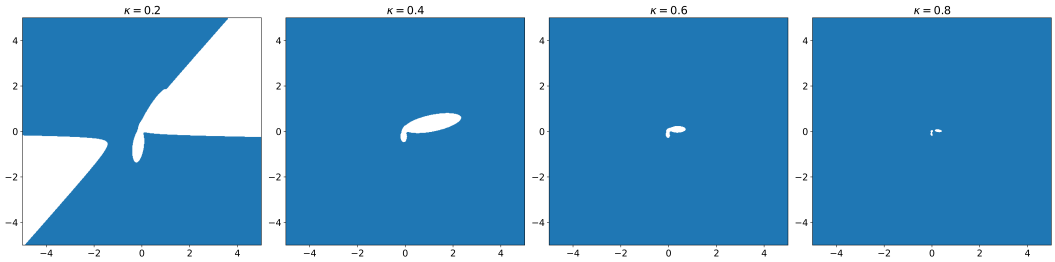

Figure 4: $\mathcal{T}(\kappa)$ region (blue area) of a 2D synthetic example with different values of κ.

5.2 Embedding the Non-Linearity within Transformer Models

By adding the penalty term to the previous energy, and with E_1 and E_2 obtained from Sections 3.1 and 4.2 respectively, we finally arrive at the total Transformer energy $E(Y) = E_1(Y) + E_2(Y) + \phi(Y)$. And the unfolded optimization of $E(Y)$ falls into the scope of Algorithm 2, such that the former analysis and Theorem 5.1 apply and we may conclude that the aggregated update rule

$$Y^{(t+1)} = \mathrm{ReLU}\left[\mathrm{softmax}_{\boldsymbol{\beta}}\left(Y^{(t)} W_a^s Y^{(t)\top}\right) W_f^s\right] \tag{20}$$

is a descent algorithm of E, except a region with finite measure. Moreover, if we modify the underlying gradient descent algorithm to make the step sizes α_1 and α_2 sufficiently small, the size of the exception region will tend to zero, and the update rule will be equipped with a residual term (see supplementary; similarly for a discussion about $\boldsymbol{\beta}$). Besides these issues, the unfolded update (20) only differs from (8) in its reliance on symmetric weights. This corresponds with Challenge 4 that we have not fully solved, although it has been shown that symmetric weights can mimic asymmetric weights if we enlarge the representation dimensions [23, 55].

6 Practical Verification

Although we have rigorously derived convergence criteria whereby Transformer layers descend a well-specified energy function to a region around optimal solutions, the analysis admittedly relies on conditions that would be difficult to formally verify on real-world datasets. However, our results are nonetheless amenable to targeted empirical corroboration, whereby we can check if the proposed energy does in fact descend during the Transformer forward pass on typical benchmarks.

To this end, we implement a Transformer model as described previously, up to known limitations like symmetric weights. We apply this model to two benchmarks, IMDB [34] and SST2 [45], which are both commonly-used sentiment classification datasets that rely on Glove-840b-300d [38] as the word embedding. Figures 5 and 6 display the energy of the output of each layer of a Transformer (as defined in (8)) averaged over 200 randomly chosen samples in the test set. Figure 5 uses randomly initialized weights while Figure 6 involves weights trained for 2000 steps with SGD and learning rate 0.01. Additionally, for the trained model we change the term $\|Y\|_{\mathcal{F}}^2$ in E_1 to $\|Y - X\|_{\mathcal{F}}^2$ in order to avoid degenerate representations (as can sometimes occur in trained Transformers [44]), noting that this modification is equally-well covered by our theory and leads to a commonly-used form of residual connection in the resulting Transformer architecture. See supplementary for details.

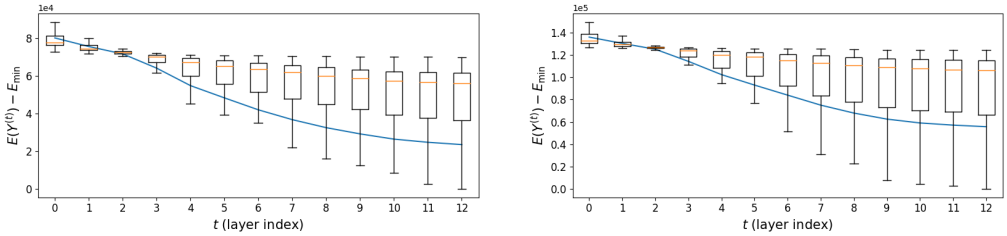

Figure 5: Box plot of the energy curves from randomly initialized 12-layer Transformers on IMDB (*left*) and SST2 (*right*) datasets; results are averaged over samples.

From these figures it is clear that even with real-world data, the the Transformer energy we have derived is (on average) monotonically decreasing across layers, matching the predictions of our

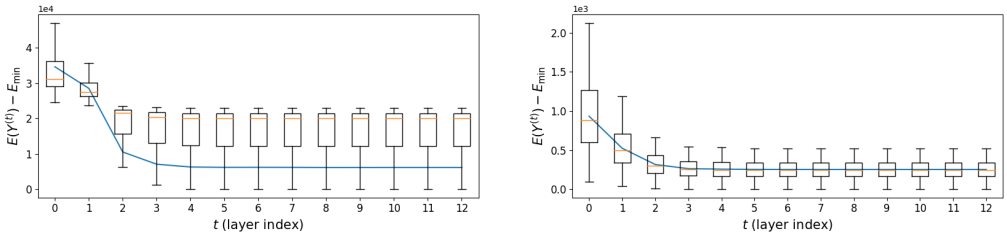

Figure 6: Box plots of the energy curves from trained 12-layer Transformers on IMDB (*left*) and SST2 (*right*) datasets; results are averaged over samples.

analysis. Moreover, with 12 layers, which represents a depth range not uncommon in practice (e.g., BERT), the model has not entered the fluctuation region. Moreover, this observation holds even with trained Transformer models composed from many/most of the typical components used in practice, namely, layers with self-attention, a linear transform followed by nonlinearity, and residual connections, etc. Hence although seemingly complicated, the conditions adopted by Lemma 4.3 and Theorem 5.1 are nonetheless likely hold in many practical settings.

7 Conclusion

While our contributions here have been primarily of a theoretical nature, there are nonetheless take-home messages that may be of practical relevance. First, as our overriding goal was to reproduce Transformer layers as closely as possible using an unfolded optimization perspective, very specific design choices were made in constructing the core underlying energy function. However, in practice we are free to choose alternative energies that can lead to different forms of tailored self-attention that may be be advantageous on an application-specific basis.

As a brief/simple representative example of the latter consider the following: The canonical Transformer uses softmax to normalize the attention coefficients. However, multiple works have questioned the appropriateness of softmax normalization, e.g., [31, 39]. Under our optimization unfolding perspective, we can see that the softmax normalization in the Transformer is derived from a specific choice of ρ function, namely $\rho(z) = -e^{-z}$ (Section 3.1 and Theorem 3.1). But if we choose a different ρ, it would lead to a new normalization method other than softmax with interpretable properties. For instance, if $\rho(z) = \log(z+2)$, then the attention coefficients would behave as (for simplicity we assume all $\mathbf{y}_i$'s have unit norm below):

$$a_{i,j} = \frac{1}{2 - \mathbf{y}_i^\top \mathbf{y}_j} \left(\sum_{k=1}^{n} \frac{1}{2 - \mathbf{y}_i^\top \mathbf{y}_k} \right)^{-1},$$

where $a_{i,j}$ is the attention coefficient between the i-th and j-th token. Since $\log(z+2)$ grows slower than $-e^{-z}$ for $z \in [0,1]$, the associated attention formulation would tend to encourage some more dissimilar representations between tokens. This is because in the new energy, a large value of $\|\mathbf{y}_i - \mathbf{y}_j\|$ (which means dissimilar $\mathbf{y}_i$ and $\mathbf{y}_j$) contributes less to the energy compared to before, and thus the optimization process has less motivation to reduce it. Conversely, if we choose a ρ that grows faster than $-e^{-z}$ in the stated range (e.g. $\rho(z) = \log(z+1)$), then the derived model would likely be encouraging more similar representations between tokens.

There are other ramifications of this overall framework with practical relevance as well. For example, the actual distribution of attention weights can potentially be better understood or influenced by properties of the energy function that produces them. Moreover, especially for regimes with limited data and therefore fewer free model parameters, the unfolding perspective can be used to devise architectures with inductive biases aligned with downstream tasks to help compensate for less model flexibility.

We conclude by noting that the techniques introduced in this paper can in some sense be viewed as universal tools for constructing a broader family of unfolding architectures. In fact, many/most deep learning models, including those with feed-forward structure [29], token-level interactions [49], residual connections [21], and heterogeneous layer types, can be reinterpreted through the lens of unfolded optimization using the framework we have introduced, at least up to some potential constraints like symmetric weights (which may be handled in other ways).

References

[1] Brandon Amos and J. Zico Kolter. OptNet: Differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning, ICML*, Proceedings of Machine Learning Research, 2017.

[2] Brandon Amos, Lei Xu, and J. Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning, ICML*, 2017.

[3] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Annual Conference on Neural Information Processing Systems, NeurIPS*, 2019.

[4] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Rev.*, 60(2):223–311, 2018.

[5] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velickovic. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *CoRR*, abs/2104.13478, 2021.

[6] Giuseppe C Calafiore and Laurent El Ghaoui. *Optimization models*. Cambridge university press, 2014.

[7] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *International Conference on Machine Learning, ICML*, Proceedings of Machine Learning Research, 2020.

[8] Siheng Chen and Yonina C Eldar. Graph signal denoising via unrolling networks. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2021.

[9] Patrick Cheridito. Convex analysis. *Lecture Notes (Princeton University, Princeton, NJ)*, 2013.

[10] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *CoRR*, abs/1904.10509, 2019.

[11] Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Ann. Oper. Res.*, 153(1):235–256, 2007.

[12] Patrick L Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

[13] Olivier Devolder, François Glineur, and Yurii E. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Math. Program.*, 146(1-2):37–75, 2014.

[14] Jordan Frécon, Gilles Gasso, Massimiliano Pontil, and Saverio Salzo. Bregman neural networks. In *International Conference on Machine Learning, ICML*, 2022.

[15] Zhengyang Geng, Meng-Hao Guo, Hongxu Chen, Xia Li, Ke Wei, and Zhouchen Lin. Is attention better than matrix decomposition? In *International Conference on Learning Representations, ICLR*, 2021.

[16] Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *International Conference on Machine Learning, ICML*, 2010.

[17] Bin Gu, De Wang, Zhouyuan Huo, and Heng Huang. Inexact proximal gradient methods for non-convex and non-smooth optimization. In *AAAI Conference on Artificial Intelligence*, 2018.

[18] Qipeng Guo, Xipeng Qiu, Pengfei Liu, Yunfan Shao, Xiangyang Xue, and Zheng Zhang. Star-transformer. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, 2019.

[19] Moritz Hardt. Understanding alternating minimization for matrix completion. In *Annual Symposium on Foundations of Computer Science, FOCS*, 2014.

[20] Hao He, Bo Xin, Satoshi Ikehata, and David Wipf. From bayesian sparsity to gated recurrent nets. In *Annual Conference on Neural Information Processing Systems, NeurIPS*, 2017.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2016.

[22] John Hershey, Jonathan Le Roux, and Felix Weninger. Deep unfolding: Model-based inspiration of novel deep architectures. *CoRR*, abs/1409.2574, 2014.

[23] Shell Xu Hu, Sergey Zagoruyko, and Nikos Komodakis. Exploring weight symmetry in deep neural networks. *Comput. Vis. Image Underst.*, 187, 2019.

[24] Simon Jenni and Paolo Favaro. Deep bilevel learning. In *European Conference on Computer Vision, ECCV*, 2018.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Annual Conference on Neural Information Processing Systems, NeurIPS*, 2012.

[26] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[27] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[28] Jia Li, Mingqing Xiao, Cong Fang, Yue Dai, Chao Xu, and Zhouchen Lin. Training neural networks by lifted proximal operator machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[29] Hanxiao Liu, Zihang Dai, David R. So, and Quoc V. Le. Pay attention to mlps. *CoRR*, abs/2105.08050, 2021.

[30] Xiaorui Liu, Wei Jin, Yao Ma, Yaxin Li, Hua Liu, Yiqi Wang, Ming Yan, and Jiliang Tang. Elastic graph neural networks. In *International Conference on Machine Learning, ICML*, 2021.

[31] Jiachen Lu, Jinghan Yao, Junge Zhang, Xiatian Zhu, Hang Xu, Weiguo Gao, Chunjing Xu, Tao Xiang, and Li Zhang. SOFT: softmax-free transformer with linear complexity. In *Annual Conference on Neural Information Processing Systems, NeurIPS*, 2021.

[32] Songtao Lu, Mingyi Hong, and Zhengdao Wang. PA-GD: on the convergence of perturbed alternating gradient descent to second-order stationary points for structured nonconvex optimization. In *International Conference on Machine Learning, ICML*, 2019.

[33] Yao Ma, Xiaorui Liu, Tong Zhao, Yozen Liu, Jiliang Tang, and Neil Shah. A unified view on graph neural networks as graph signal denoising. *CoRR*, abs/2010.01777, 2020.

[34] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Annual Meeting of the Association for Computational Linguistics, ACL*, 2011.

[35] S. Deepak Narayanan, Aditya Sinha, Prateek Jain, Purushottam Kar, and Sundararajan Sellamanickam. IGLU: efficient GCN training via lazy updates. *CoRR*, abs/2109.13995, 2021.

[36] Xuran Pan, Shiji Song, and Gao Huang. A unified framework for convolution-based graph neural networks, 2021.

[37] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

[38] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing, EMNLP*, 2014.

[39] Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosFormer: Rethinking softmax in attention. In *The Tenth International Conference on Learning Representations, ICLR*, 2022.

[40] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David P. Kreil, Michael K. Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations, ICLR*, 2021.

[41] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[42] Pierre Samuel. *Projective Geometry*. Springer, 1988.

[43] Mark Schmidt, Nicolas Le Roux, and Francis R. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Annual Conference on Neural Information Processing Systems, NeurIPS*, 2011.

[44] Han Shi, Jiahui Gao, Hang Xu, Xiaodan Liang, Zhenguo Li, Lingpeng Kong, Stephen Lee, and James T Kwok. Revisiting over-smoothing in BERT from the perspective of graph. *International Conference on Learning Representations, ICLR*, 2022.

[45] Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. Parsing with compositional vector grammars. In *Annual Meeting of the Association for Computational Linguistics, ACL*, 2013.

[46] Pablo Sprechmann, Alex Bronstein, and Guillermo Sapiro. Learning efficient sparse and low rank models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 37(9), 2015.

[47] Ying Sun, Prabhu Babu, and Daniel P Palomar. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Transactions on Signal Processing*, 65(3):794–816, 2016.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Annual Conference on Neural Information Processing Systems, NeurIPS*, 2017.

[49] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations, ICLR*, 2018.

[50] Zhangyang Wang, Qing Ling, and Thomas Huang. Learning deep ℓ_0 encoders. In *AAAI Conference on Artificial Intelligence*, 2016.

[51] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM J. Optim.*, 24(4):2057–2075, 2014.

[52] Xingyu Xie, Qiuhao Wang, Zenan Ling, Xia Li, Yisen Wang, Guangcan Liu, and Zhouchen Lin. Optimization induced equilibrium networks. *CoRR*, abs/2105.13228, 2021.

[53] Chaoqi Yang, Ruijie Wang, Shuochao Yao, Shengzhong Liu, and Tarek F. Abdelzaher. Revisiting "over-smoothing" in deep GCNs. *CoRR*, abs/2003.13663, 2020.

[54] Yongyi Yang, Tang Liu, Yangkun Wang, Jinjing Zhou, Quan Gan, Zhewei Wei, Zheng Zhang, Zengfeng Huang, and David Wipf. Graph neural networks inspired by classical iterative algorithms. In *International Conference on Machine Learning, ICML*, 2021.

[55] Yongyi Yang, Yangkun Wang, Zengfeng Huang, and David Wipf. Implicit vs unfolded graph neural networks. *CoRR*, abs/2111.06592, 2021.

[56] Alan L. Yuille and Anand Rangarajan. The concave-convex procedure (CCCP). In *Annual Conference on Neural Information Processing Systems, NeurIPS*, 2001.

[57] Hongwei Zhang, Tijin Yan, Zenjun Xie, Yuanqing Xia, and Yuan Zhang. Revisiting graph convolutional network on semi-supervised node classification from an optimization perspective. *CoRR*, abs/2009.11469, 2020.

[58] Kaixiong Zhou, Xiao Huang, Daochen Zha, Rui Chen, Li Li, Soo-Hyun Choi, and Xia Hu. Dirichlet energy constrained learning for deep graph neural networks. *CoRR*, abs/2107.02392, 2021.

[59] Meiqi Zhu, Xiao Wang, Chuan Shi, Houye Ji, and Peng Cui. Interpreting and unifying graph neural networks with an optimization framework. *CoRR*, abs/2101.11859, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work is mostly theoretical and devoted to the better understanding of an existing, well-established modeling framework; we do not see any clear pathway to negative societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] They are in the supplementary.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Theory Background

This section provides background and references regarding some of the concepts and techniques used for proving our main results.

A.1 Lipschitz Conditions and Strong Convexity

For a function $f : \mathbb{R}^d \to \mathbb{R}$, if

$$\exists L > 0, \forall \boldsymbol{x}, \mathbf{y} \in \mathbb{R}^d, \|f(\boldsymbol{x}) - f(\mathbf{y})\| \leq L\|\boldsymbol{x} - \mathbf{y}\|, \tag{21}$$

we say f satisfies the Lipschitz condition with Lipschitz constant L or is simply L-Lipschitz. If *the gradient* of a function is L-Lipschitz, we say it is Lipschitz smooth with Lipschitz constant L. Moreover, if f satisfies

$$\exists c > 0, \forall \boldsymbol{x}, \mathbf{y} \in \mathbb{R}^d, \|\nabla f(\boldsymbol{x}) - \nabla f(\mathbf{y})\| \geq c\|\boldsymbol{x} - \mathbf{y}\|, \tag{22}$$

we say it is (strongly) convex with convexity c, or simply c-strongly convex.

There are two important inequalities for Lipschitz smooth and strongly convex functions:

Proposition A.1. *If $f : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz smooth, then*

$$\forall \boldsymbol{x}, \mathbf{y} \in \mathbb{R}^d, f(\mathbf{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top(\mathbf{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{x} - \mathbf{y}\|^2. \tag{23}$$

If f is c-strongly convex, then

$$\forall \boldsymbol{x}, \mathbf{y} \in \mathbb{R}^d, f(\mathbf{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top(\mathbf{y} - \boldsymbol{x}) + \frac{c}{2}\|\boldsymbol{x} - \mathbf{y}\|^2. \tag{24}$$

The proof of Proposition A.1, and more details of these two properties can be found in [4].

A.2 Proximal Operators and Proxinal Gradient Descent

The proximal problem of a function ϕ with parameter λ is defined as

$$P(\boldsymbol{z}; \mathbf{y}) = \frac{1}{2\lambda}\|\boldsymbol{z} - \mathbf{y}\|^2 + \phi(\boldsymbol{z}), \tag{25}$$

and the mapping from $\mathbf{y}$ to the minimal point of $P(\cdot; \mathbf{y})$, is called the proximal operator of ϕ:

$$\mathrm{prox}_\phi : \mathbf{y} \mapsto \arg\min_{\boldsymbol{z}} \frac{1}{2\lambda}\|\boldsymbol{z} - \mathbf{y}\|^2 + \phi(\boldsymbol{z}). \tag{26}$$

It can be shown that, for an objective function $h = f + \phi$, where f is smooth, alternatively performing gradient descent on f and proximal projection of ϕ, i.e.

$$\mathbf{y}^{(t+1)} = \mathrm{prox}_\phi\left(\mathbf{y}^{(t)} - \alpha\nabla f\left(\mathbf{y}^{(t)}\right)\right), \tag{27}$$

is a descent algorithm of h, which is often referred to as proximal gradient descent [37].

A.3 Subgradient

For a (not necessarily smooth) convex function $f : \mathbb{R}^d \to \mathbb{R}$, it can be proven that the following set is always not empty [9]:

$$\partial f(\boldsymbol{x}) = \{\boldsymbol{g} | \forall \mathbf{y} \in \mathbb{R}^d, f(\mathbf{y}) \geq f(\boldsymbol{x}) + \boldsymbol{g}^\top(\mathbf{y} - \boldsymbol{x})\}. \tag{28}$$

We call $\partial f(\mathbf{y})$ the subdifferential of f at point $\mathbf{y}$, and the elements of $\partial f(\mathbf{y})$ the subgradients of ∂f at point $\mathbf{y}$.

Subdifferential satisfies linearity, which means

$$\alpha\partial f(\mathbf{y}) + \beta\partial g(\mathbf{y}) = \partial(\alpha f + \beta g)(\mathbf{y}), \tag{29}$$

where the summation of two sets is defined as the elementary summation

$$\partial f(\mathbf{y}) + \partial g(\mathbf{y}) = \{\boldsymbol{a} + \boldsymbol{b} | \boldsymbol{a} \in \partial f(\mathbf{y}), \boldsymbol{b} \in \partial \in g(\mathbf{y})\}. \tag{30}$$

Although subdifferentials are sets, based on the linearity, it is easier to use the equation symbol $=$ instead of $\in$ to denote subgradient, e.g. $\partial f(\mathbf{y}) = \boldsymbol{a}$ means $\boldsymbol{a} \in \partial f(\mathbf{y})$.

One very useful property of the subgradient is that it can indicate the global minimum of a convex function :

Proposition A.2. *If f is convex, then*

$$0 \in \partial f(\mathbf{y}) \iff \mathbf{y} \in \arg\min_{\boldsymbol{z}} f(\boldsymbol{z}), \tag{31}$$

which is quite straight forward from the definition of subgradient. More detailed introduction and discussion of subgradients can be found in [6, 9].

B Proofs, Discussion, and Extensions of Section 3 Results

We now provide a complement of Section 3 of the main paper, including the proofs and discussion. For Theorem 3.1, we prove a more generalized version, taking into account of the structure of attention and the step size, as mentioned in Remark 3.2.

To achieve this, we extend the techniques used in [54] and show how to construct an energy function whose iterative optimization steps match Transformer-style self-attention on arbitrary graph structures – including fully connected graphs.

Note that this section provides a more general framework that assumes a certain graph structure exists in token-level interactions, which covers some attention sparsification works like [10, 18]. And when the graph is fully connected, we obtain a normal Transformer structure as analyzed in the main paper.

B.1 General Unfolded Optimization Results for Creating Self-Attention on Graphs

Suppose $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is an undirected graph, whose adjacency matrix is $A \in \mathbb{R}^{n\times n}$, incidence-matrix is $B \in \mathbb{R}^{m\times n}$, degree matrix is $D = \operatorname{diag}(A)\,\mathbf{1}$, and Laplacian matrix is $\mathcal{L} = D - A = B^\top B$, where $n = |\mathcal{V}|$ and $m = |\mathcal{E}|$. We use a tuple (u, v) to represent an edge in the graph that connects node u and v. Furthermore, let there be an index for each edge, suppose the ith edge connects $\mathrm{fr}(i)$ and $\mathrm{to}(i)$ and the index of edge that connecting (u, v) is $\mathrm{ind}(u, v)$. Then, the incidence matrix B can be defined as

$$B_{i,j} = \begin{cases} 1 & \text{if } j = \mathrm{fr}(i) \\ -1 & \text{if } j = \mathrm{to}(i) \\ 0 & \text{others.} \end{cases} \tag{32}$$

Consider a generalized version of the energy function (9):

$$E_1(Y) = \sum_{u,v\in\mathcal{E}} \rho\left(\frac{1}{2}\|\mathbf{y}_u - \mathbf{y}_v\|^2\right) + R(Y), \tag{33}$$

where $\mathbf{y}_u \in \mathbb{R}^{d\times 1}$ is the u-th row of matrix $Y \in \mathbb{R}^{n\times d}$, $\rho : \mathbb{R}^+ \to \mathbb{R}$ is a concave non-decreasing function and $R : \mathbb{R}^{n\times d} \to \mathbb{R}$ is a convex function.

Algorithm 3

Execute the following two assignment operations in arbitrary order:

1. $\gamma_{u,v}^{(t)} = \left.\frac{\partial\rho(z^2)}{\partial z^2}\right|_{z^2=\frac{1}{2}\left\|\mathbf{y}_u^{(t)}-\mathbf{y}_j^{(t)}\right\|^2}$;
2. $Y^{(t+1)} = Y'^{(t)}$ such that $\tilde{E}_1\left(Y'^{(t)}, \Gamma^{(t)}\right) \le \tilde{E}_1\left(Y^{(t+1)}, \Gamma^{(t)}\right)$,

where $\Gamma^{(t)} \in \mathbb{R}^{n\times n}, \Gamma_{i,j}^{(t)} = \gamma_{i,j}^{(t)}$, and $\tilde{E}_1(Y, \Gamma) = \sum_{u,v\in\mathcal{E}} \frac{1}{2}\gamma_{u,v}\|\mathbf{y}_u - \mathbf{y}_v\|^2 + R(Y)$.

We refer to [54] that Algorithm 3 optimizes E_1:

Proposition B.1 (Lemma 3.2 in [54])**.** *Let $Y^{(t)}$ be the input of Algorithm 3 and $Y^{(t+1)}$ be the output, then $E_1\left(Y^{(t+1)}\right) \le E_1\left(Y^{(t)}\right)$.*

In Algorithm 3, Step 1 converts the non-linear function ρ to edge weights in graphs, which correspond to attention weight matrices. If in Step 2 we use gradient descent, then the update rule is

$$Y^{(t+1)} = \alpha(I - B^\top\Gamma B)Y^{(t)} + \left[(1-\alpha)I - \frac{\partial R\left(Y^{(t)}\right)}{\partial Y^{(t)}}\right] Y^{(t)}, \tag{34}$$

where the attention weights are injected into coefficients $1 - B^\top \Gamma B$. However, this naive framework does not include softmax used in Transformers. In the following subsections we show that with a special choice of ρ and a reweighting of each row of Y, we get exactly self-attention updates with softmax.

Now, consider $\rho\left(z^2\right) = -\exp\left\{-z^2\right\}$, $R(Y) = \frac{1}{2}\|Y\|_{\mathcal{F}}^2$. Then Step 1 gives

$$\begin{aligned} \gamma_{u,v}^{(t)} &= \exp\left\{-\frac{1}{2}\left\|\mathbf{y}_u^{(t)} - \mathbf{y}_v^{(t)}\right\|^2\right\} \\ &= \exp\left\{\mathbf{y}_u^{(t)\top}\mathbf{y}_v^{(t)}\right\}\boldsymbol{\beta}_u\boldsymbol{\beta}_v, \end{aligned} \tag{35}$$

where $\boldsymbol{\beta}_u = \exp\left\{-\frac{1}{2}\left\|\mathbf{y}_u^{(t)}\right\|^2\right\}$. Note that $\gamma_{u,u} = e^0 = 1$.

For Step 2, consider using gradient descent with Jacobi preconditioning, from which we arrive at the following Theorem:

Theorem B.2. *Consider updating $\tilde{E}_1$ using a gradient step with step size α and Jacobi preconditioner $\mathcal{D}^{-(t)}$:*

$$Y^{(t+1)} = Y^{(t)} - \alpha\mathcal{D}^{-(t)}\left.\frac{\partial\tilde{E}_1\left(Y,\Gamma^{(t)}\right)}{\partial Y}\right|_{Y=Y^{(t)}}, \tag{36}$$

where

$$\mathcal{D}^{(t)} = \left.\frac{\partial^2\tilde{E}_1\left(Y,\Gamma^{(t)}\right)}{\partial Y^2}\right|_{Y=Y^{(t)}}, \tag{37}$$

and $\alpha \leq 1$, it follows that $\tilde{E}_1\left(Y^{(t+1)}\right) \leq \tilde{E}_1\left(Y^{(t)}\right)$. And the update rule in (36) can be written as

$$\mathbf{y}_u^{(t+1)} = (1-\alpha)\mathbf{y}_u^{(t)} + \alpha\frac{\sum\limits_{v\in\tilde{\mathcal{N}}(u)}\boldsymbol{\beta}_v\exp\left\{\mathbf{y}_u^{(t)\top}\mathbf{y}_v^{(t)}\right\}\mathbf{y}_v^{(t)}}{\sum\limits_{v\in\tilde{\mathcal{N}}(u)}\boldsymbol{\beta}_v\exp\left\{\mathbf{y}_u^{(t)\top}\mathbf{y}_v^{(t)}\right\}}, \quad \forall u. \tag{38}$$

Notice that it requires $\alpha \leq L^{-(t)}$ to guarantee (36) is a descent step of $\tilde{E}_1$, where $L^{(t)}$ is the Lipschitz constant of the gradient of $\mathcal{D}^{-(t)}\tilde{E}_1\left(Y,\Gamma^{(t)}\right)$ [4]. And by the definition of $\mathcal{D}^{(t)}$, $L^{(t)}$ is always 1. Therefore, $\alpha \leq 1$ always satisfies the condition.

B.2 The Proof of Theorem 3.1

In Theorem B.2, if we take $\mathcal{G}$ to be a complete graph, then $\tilde{\mathcal{N}}(u) = \{1, 2, \cdots, n\}$. Besides, if $\alpha = 1$, (38) becomes

$$\mathbf{y}_u^{(t+1)} = \frac{\sum_{v=1}^n \exp\left\{\mathbf{y}_u^{(t)\top}\mathbf{y}_v^{(t)}\right\}\boldsymbol{\beta}_v\mathbf{y}_v^{(t)}}{\sum_{v=1}^n \exp\left\{\mathbf{y}_u^{(t)\top}\mathbf{y}_v^{(t)}\right\}\boldsymbol{\beta}_v}, \tag{39}$$

which guarantees the descent of $\tilde{E}_1\left(Y\right)$ by Theorem B.2. And further by Proposition B.1, Theorem 3.1 is proved.

B.3 Discussion

From the analysis above, we have the following observations. Firstly, if we set α to other values, say $\alpha = \frac{1}{2}$, then the residual term in (10) is maintained, which corresponds with the common use of skip connections in the self-attention. Secondly, by assuming arbitrary graphs $\mathcal{G}$ (as opposed to merely the fully connected case described in the main paper), this framework can naturally handle attention mechanisms with special structure as in [10, 18].

C Proofs of Section 4 Results

In this section, we prove the propositions in Section 4.

C.1 The Proof of Theorem 4.1

Firstly, from the Lipschitz smooth assumption of the objective functions, we can bound the norm of the noise term:

$$\begin{aligned}
\|\Delta_t\| &= \left\|\nabla f\left(\mathbf{y}^{(t)}\right) + \nabla g\left(\mathbf{y}^{(t)}\right) - \frac{\alpha_1}{\alpha_2}\nabla f\left(\mathbf{y}^{(t)}\right) - \nabla g\left(\mathbf{y}^{(t)} - \alpha_1 \nabla f\left(\mathbf{y}^{(t)}\right)\right)\right\| \\
&\leq \left\|\left(1 - \frac{\alpha_1}{\alpha_2}\right)\nabla f\left(\mathbf{y}^{(t)}\right)\right\| + \left\|\nabla g\left(\mathbf{y}^{(t)}\right) - \nabla g\left(\mathbf{y}^{(t)} - \alpha_1 \nabla f\left(\mathbf{y}^{(t)}\right)\right)\right\| \\
&\leq \left\|\left(1 - \frac{\alpha_1}{\alpha_2}\right)\nabla f\left(\mathbf{y}^{(t)}\right)\right\| + \alpha_1 L_g \left\|\nabla f\left(\mathbf{y}^{(t)}\right)\right\| \\
&= \left(1 - \frac{\alpha_1}{\alpha_2} + \alpha_1 L_g\right)\left\|\nabla f\left(\mathbf{y}^{(t)}\right)\right\| \\
&\leq \left(1 - \frac{\alpha_1}{\alpha_2} + \alpha_1 L_g\right)\delta\left(\mathbf{y}^{(t)}\right)\left\|\nabla h\left(\mathbf{y}^{(t)}\right)\right\|.
\end{aligned} \tag{40}$$

In the following we write ∇h for short to refer to $\nabla h\left(\mathbf{y}^{(t)}\right)$; similarly δ for $\delta\left(\mathbf{y}^{(t)}\right)$ and Δ for Δ_t.

From (40) we have

$$\|\Delta\|^2 \leq \left(1 - \frac{\alpha_1}{\alpha_2} + \alpha_1 L_g\right)^2 \delta^2 \|\nabla h\|^2, \tag{41}$$

and from Cauchy-Schwarz

$$\nabla h^\top \Delta \geq -\|\nabla h\|\|\Delta\| \geq -\left(1 - \frac{\alpha_1}{\alpha_2} + \alpha_1 L_g\right)\delta\|\nabla h\|^2. \tag{42}$$

Therefore, by Lipschitz smoothness and convexity assumptions, and the inequalities (40), (41) and (42), and notice that $\alpha_2 L_h \leq 1$, we have

$$\begin{aligned}
&h\left(\mathbf{y}^{(t+1)}\right) - h\left(\mathbf{y}^{(t)}\right) \\
&\quad\leq -\alpha_2 \nabla h^\top (\nabla h + \Delta) + \frac{L_h}{2}\alpha_2^2 \|\nabla h + \Delta\|^2 \\
&\quad= -\alpha_2\|\nabla h\|^2 - \alpha_2 \nabla h^\top \Delta + \frac{L_h \alpha_2^2}{2}\left(\|\nabla h\|^2 + \|\Delta\|^2 + 2\nabla h^\top \Delta\right) \\
&\quad=\alpha_2\left[\left(\frac{L_h\alpha_2}{2} - 1\right)\|\nabla h\|^2 + \frac{L_h\alpha_2}{2}\|\Delta\|^2 + (L_h\alpha_2 - 1)\nabla h^\top \Delta\right] \\
&\quad\leq\alpha_2\|\nabla h\|^2\left[\frac{1}{2}L_h\alpha_2\left(1 - \frac{\alpha_1}{\alpha_2} + \alpha_1 L_g\right)^2\delta^2\right. \\
&\qquad\left. +(1 - L_h\alpha_2)\left(1 - \frac{\alpha_1}{\alpha_2} + \alpha_1 L_g\right)\delta + \left(\frac{L_h\alpha_2}{2} - 1\right)\right].
\end{aligned} \tag{43}$$

While seemingly complex, if we define $a = \frac{1}{2}L_h\alpha_2 \in \left[0, \frac{1}{2}\right]$ and $b = 1 - \frac{\alpha_1}{\alpha_2} + \alpha_1 L_g \geq 0$, (43) can be rewritten as

$$\frac{1}{\alpha_2\|\nabla h\|^2}\left[h\left(\mathbf{y}^{(t+1)}\right) - h\left(\mathbf{y}^{(t)}\right)\right] \leq ab^2\delta^2 + (1 - 2a)b\delta + (a - 1). \tag{44}$$

Clearly, when $\delta \in \left[\frac{a-1}{ab}, \frac{1}{b}\right]$, we have

$$ab^2\delta^2 + (1 - 2a)b\delta + (a - 1) \leq 0, \tag{45}$$

and $\delta \geq 0 \geq \frac{a-1}{ab}$ by definition. Therefore, we conclude that $h\left(\mathbf{y}^{(t+1)}\right) - h\left(\mathbf{y}^{(t)}\right) \leq 0$ is guaranteed when

$$\delta \leq \frac{1}{b} = \frac{\alpha_2}{\alpha_2 - \alpha_2 + \alpha_1\alpha_2 L_g}. \tag{46}$$

C.2 The Proof of Lemma 4.3

We only need to notice that $\nabla f\left(\mathbf{y}_f^*\right) = \nabla h\left(\mathbf{y}_h^*\right) = 0$, then by Lipschitz smoothness and convexity of the objectives,

$$\|\nabla f(\mathbf{y})\| \leq L_f \left\|\mathbf{y} - \mathbf{y}_f^*\right\| \text{ and } \|\nabla h(\mathbf{y})\| \geq c_h \left\|\mathbf{y} - \mathbf{y}_h^*\right\|. \tag{47}$$

Therefore, when $\mathbf{y} \in \mathcal{S}(\mathscr{C})$,

$$\delta(\mathbf{y}) = \frac{\|\nabla f(\mathbf{y})\|}{\|\nabla h(\mathbf{y})\|} \leq \frac{L_f\|\mathbf{y} - \mathbf{y}_f^*\|}{c_h\|\mathbf{y} - \mathbf{y}_h^*\|} \leq \mathscr{C}. \tag{48}$$

D Proofs of Section 5 Results

For simplicity, in the following we use $\mathbf{y}$ to denote $\mathbf{y}^{(t+1)}$ and $\boldsymbol{x}$ to denote $\mathbf{y}^{(t)}$ this section. We also use $\boldsymbol{x}^*$ to denote the optimal solution of the proximal problem (19):

$$\boldsymbol{x}^* = \arg\min_{z} P(\boldsymbol{z}; \boldsymbol{x}) = \mathrm{ReLU}(\boldsymbol{x}). \tag{49}$$

Recall that

$$\mathbf{y} = \arg\min_{\boldsymbol{z}} \frac{1}{2\lambda} \|\boldsymbol{z} - \boldsymbol{x} + \alpha_2 \nabla h(\boldsymbol{x}) - \alpha_2 \Delta_t\|^2 + \phi(\boldsymbol{z}). \tag{50}$$

First we shall prove the bound of the subgradient of proximal problem (19) at the point of $\mathbf{y}$:

Lemma D.1.

$$\partial P(\mathbf{y}; \boldsymbol{x}) = -\frac{\alpha_2}{\lambda}\Delta_t \tag{51}$$

Proof. Since $\mathbf{y}$ is the optimal point of (50), we have

$$\frac{\partial}{\partial \mathbf{y}}\left(\frac{1}{2\lambda}\|\boldsymbol{z} - \boldsymbol{x} + \alpha_2 \nabla h(\boldsymbol{x}) - \alpha_2 \Delta_t\|^2 + \phi(\boldsymbol{z})\right) = 0, \tag{52}$$

which gives

$$\partial \phi(\mathbf{y}) = \frac{1}{\lambda}\left(\boldsymbol{x} - \mathbf{y} - \alpha_2 \nabla h(\boldsymbol{x}) + \alpha_2 \Delta(t)\right). \tag{53}$$

Therefore, the subgradient of (19) at $\mathbf{y}$ is

$$\begin{aligned} \partial P(\mathbf{y}; \boldsymbol{x}) &= \frac{1}{\lambda}\left(\mathbf{y} - \boldsymbol{x} + \alpha_2 \nabla h(\boldsymbol{x})\right) + \partial \phi(\mathbf{y}) \\ &= -\frac{\alpha_2}{\lambda}\Delta_t. \end{aligned} \tag{54}$$

□

Then, we shall show the descent of $P(\mathbf{y}; \boldsymbol{x})$ to $P(\boldsymbol{x}; \boldsymbol{x})$ can be bounded by the distance between $\boldsymbol{x}$ and $\boldsymbol{x}^*$.

Lemma D.2.

$$P(\boldsymbol{x}; \boldsymbol{x}) - P(\mathbf{y}; \boldsymbol{x}) \geq \frac{c_P}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 - \frac{\alpha_2^2(1 + \alpha_1 L_g)^2 \delta(\boldsymbol{x})^2}{c_P \lambda^2}\|\nabla h(\boldsymbol{x})\|^2. \tag{55}$$

Proof. Let c_P denote the convexity of proximal problem $P(\boldsymbol{z}; \boldsymbol{x})$. And note since $\boldsymbol{x}^*$ is the optimal point of $P(\boldsymbol{z}; \boldsymbol{x})$, we have $\partial P\left(\boldsymbol{x}^*\right) = 0$. Then by convexity we have

$$P(\boldsymbol{x}; \boldsymbol{x}) \geq P\left(\boldsymbol{x}^*\right) + \frac{c_P}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \tag{56}$$

and

$$
\begin{aligned}
P\left(\boldsymbol{x}^{*}\right) &\geq P(\mathbf{y};\boldsymbol{x})+\partial P(\mathbf{y};\boldsymbol{x})^{\top}\left(\boldsymbol{x}^{*}-\mathbf{y}\right)+\frac{c_P}{2}\left\|\boldsymbol{x}^{*}-\mathbf{y}\right\|^{2} \\
&\geq P(\mathbf{y};\boldsymbol{x})-\|\partial P(\mathbf{y};\boldsymbol{x})\|\left\|\boldsymbol{x}^{*}-\mathbf{y}\right\| \\
&\geq P(\mathbf{y};\boldsymbol{x})-\frac{1}{c_P}\|\partial P(\mathbf{y};\boldsymbol{x})\|^{2} \\
&= P(\mathbf{y};\boldsymbol{x})-\frac{\alpha_2^2}{c_P\lambda^2}\|\Delta_t\|^2 && \text{(By Lemma D.1)} \\
&\geq P(\mathbf{y};\boldsymbol{x})-\frac{\alpha_2^2}{c_P\lambda^2}(1+\alpha_1 L_g+\alpha_1\alpha_2^{-1})^2\delta(\boldsymbol{x})^2\left\|\nabla h(\boldsymbol{x})\right\|^2. && \text{(By (40)}
\end{aligned}
\tag{57}
$$

Combining (56) and (57) we prove the lemma.

□

Given Lemma D.2, we only need to bound $\|\boldsymbol{x}-\boldsymbol{x}^*\|$. The following lemma shows that, $\|\boldsymbol{x}-\boldsymbol{x}^*\|$ is related to

$$
\mathfrak{D}:\left(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2\right)\mapsto\sum_{i=1}^{d}\min\left(\boldsymbol{\xi}_{2,i}^2-\boldsymbol{\xi}_{1,i}^2,0\right), \tag{58}
$$

as stated in the main paper.

Lemma D.3. *Given $\mathfrak{D}\left(\alpha_2\nabla h\left(\boldsymbol{x}\right);\boldsymbol{x}\right)\geq-\kappa$, we have*

$$
\|\boldsymbol{x}-\boldsymbol{x}^*\|^2\geq\left(1-\kappa\right)\alpha_2^2\|\nabla h(\boldsymbol{x})\|^2. \tag{59}
$$

Proof. Note $\boldsymbol{x}_i^*=\sigma(\boldsymbol{x}_i-\alpha_2\nabla h(\boldsymbol{x})_i)=\max(\boldsymbol{x}_i-\alpha_2\nabla h(\boldsymbol{x})_i,0)$. We have

$$
\boldsymbol{x}_i^*-\boldsymbol{x}_i=\max(-\alpha_2\nabla h(\boldsymbol{x})_i,-\boldsymbol{x}_i)=-\min(\alpha_2\nabla h(\boldsymbol{x})_i,\boldsymbol{x}_i). \tag{60}
$$

Therefore

$$
\begin{aligned}
\|\boldsymbol{x}-\boldsymbol{x}^*\|^2 &= \sum_{i=1}^{d}\min(\alpha_2\nabla h(\boldsymbol{x})_i,\boldsymbol{x}_i)^2 \\
&\geq \sum_{i=1}^{d}\min(\alpha_2^2\nabla h(\boldsymbol{x})_i^2,\boldsymbol{x}_i^2) \\
&= \sum_{i=1}^{d}\alpha_2^2\nabla h(\boldsymbol{x})_i^2+\sum_{i=1}^{d}\min\left(\boldsymbol{x}_i^2-\alpha_2^2\nabla h(\boldsymbol{x})_i^2,0\right) \\
&= \alpha_2^2\left\|\nabla h(\boldsymbol{x})\right\|^2+\alpha_2^2\|\nabla h(\boldsymbol{x})\|^2\mathfrak{D}(\alpha_2\nabla h(\boldsymbol{x});\boldsymbol{x}) \\
&= \left(1+\mathfrak{D}(\alpha_2\nabla h(\boldsymbol{x}),\boldsymbol{x})\right)\alpha_2^2\|\nabla h(\boldsymbol{x})\|^2 \\
&\geq \left(1-\kappa\right)\alpha_2^2\|\nabla h(\boldsymbol{x})\|^2
\end{aligned}
\tag{61}
$$

□

Next, we shall show that $P(\boldsymbol{z};\boldsymbol{x})$ is actually an upper bound of $f(\boldsymbol{z})+\phi(\boldsymbol{z})$:

Lemma D.4. *If $\lambda\leq\alpha_2\leq\frac{1}{L_h}$, there exists a function $\eta(\boldsymbol{x})$, which is only dependent on $\boldsymbol{x}$, which satisfies*

$$
P(\boldsymbol{z};\boldsymbol{x})+\eta(\boldsymbol{x})\geq h(\boldsymbol{z})+\phi(\boldsymbol{z}) \tag{62}
$$

and

$$
P(\boldsymbol{x};\boldsymbol{x})+\eta(\boldsymbol{x})=h(\boldsymbol{x})+\phi(\boldsymbol{x}). \tag{63}
$$

Proof. Let $\eta(\boldsymbol{x}) = h(\boldsymbol{x}) - \frac{\alpha_2^2}{2\lambda}\|\nabla h(\boldsymbol{x})\|^2$, then by the assumed Lipschitz condition we have

$$\begin{aligned}
h(\boldsymbol{z}) + \phi(\boldsymbol{z}) &\leq h(\boldsymbol{x}) + \nabla h(\boldsymbol{x})^\top(\boldsymbol{z}-\boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{z}-\boldsymbol{x}\|^2 + \phi(\boldsymbol{x}) \\
&\leq h(\boldsymbol{x}) + \nabla h(\boldsymbol{x})^\top(\boldsymbol{z}-\boldsymbol{x}) + \frac{1}{2\alpha_2}\|\boldsymbol{z}-\boldsymbol{x}\|^2 + \phi(\boldsymbol{z}) \\
&\leq \frac{1}{2\alpha_2}\left[\|\boldsymbol{z}-\boldsymbol{x}\|^2 + 2\alpha_2\nabla h(\boldsymbol{x})^\top(\boldsymbol{z}-\boldsymbol{x}) + \alpha_2^2\|\nabla h(\boldsymbol{x})\|^2\right] \\
&\quad + \phi(\boldsymbol{z}) - \frac{\alpha_2}{2}\|\nabla h(\boldsymbol{x})\|^2 + h(\boldsymbol{x}) \\
&\leq \frac{1}{2\lambda}\|\boldsymbol{z}-\boldsymbol{x}+\alpha_2\nabla h(\boldsymbol{x})\|^2 + \phi(\boldsymbol{z}) - \frac{\lambda}{2}\|\nabla h(\boldsymbol{x})\|^2 + h(\boldsymbol{x}) \\
&= P(\boldsymbol{z};\boldsymbol{x}) + \eta(\boldsymbol{x}).
\end{aligned} \tag{64}$$

And it is also straightforward to verify that

$$P(\boldsymbol{x};\boldsymbol{x}) + \beta(\boldsymbol{x}) = \frac{\alpha_2^2}{2\lambda}\|\nabla h(\boldsymbol{x})\|^2 + \phi(\boldsymbol{x}) - \frac{\alpha_2^2}{2\lambda}\|\nabla h(\boldsymbol{x})\|^2 + h(\boldsymbol{x}) = h(\boldsymbol{x}) + \phi(\boldsymbol{x}). \tag{65}$$

□

It is worth noting that, although there's a parameter λ in the definition of P, when ϕ is defined as $\phi(z) = \begin{cases} +\infty & \text{if } z < 0 \\ 0 & \text{if } z \geq 0 \end{cases}$, which we used in the main paper to derive ReLU, the proximal operator prox_ϕ is independent of λ, which means we can always select a λ small enough to satisfy $\lambda \leq \alpha_2$, which is required by Lemma D.4.

Lemma D.4 provides us with a majorization minimization perspective, which allow us to bound $h(\boldsymbol{z}) + \phi(\boldsymbol{z})$ by bounding $P(\mathbf{y};\boldsymbol{x})$. And Lemma D.2 provides a way to bound the descent of $P(\mathbf{y};\boldsymbol{x})$. Now we are ready to prove Theorem 5.1.

The proof of Theorem 5.1 Combining Lemmas D.2 and D.3 and recalling the bound of $\delta(\boldsymbol{x})$ we have

$$P(\boldsymbol{x};\boldsymbol{x}) - P(\mathbf{y};\boldsymbol{x}) \geq \alpha_2^2\|\nabla h(\boldsymbol{x})\|^2\left[\frac{c_P(1-\kappa)}{2} - \frac{(1+\alpha_1 L_g + \alpha_1\alpha_2)^2\delta(\boldsymbol{x})^2}{c_P\lambda^2}\right]. \tag{66}$$

Given $\mathfrak{D}\left(\alpha_2\nabla h\left(\mathbf{y}^{(t)}\right);\mathbf{y}^{(t)}\right) \geq -\kappa$ and (66), it follows that $P(\boldsymbol{x};\boldsymbol{x}) - P(\mathbf{y};\boldsymbol{x}) \geq 0$, i.e. $P(\boldsymbol{x};\boldsymbol{x}) \geq P(\mathbf{y};\boldsymbol{x})$.

By Lemma D.4,

$$h(\mathbf{y}) + \phi(\mathbf{y}) \leq P(\mathbf{y};\boldsymbol{x}) + \eta(\boldsymbol{x}) \leq P(\boldsymbol{x};\boldsymbol{x}) + \eta(\boldsymbol{x}) = h(\boldsymbol{x}) + \phi(\boldsymbol{x}), \tag{67}$$

which proves the theorem.

E Experiment Details

In the experiments of Section 6, the dataset and pre-processing scripts are provided by fastNLP[5]. The code to reproduce all experiments in the main paper is available[6] for the reference of the complete implementation of the experiments.

F Further Discussion

In this section, we provide further discussion related to symmetric weights, layer-dependent weights, layer normalization, and initial residual connections [7] as mentioned in the main paper.

[5] https://github.com/fastnlp/fastNLP

[6] https://github.com/FFTYYY/Transformers-From-Optimization

F.1 Asymmetric Weights

It initially seems hard or impossible to directly interpret an asymmetric transformation (i.e. $f(\mathbf{y}) = W\mathbf{y}$ where W is asymmetric) as a gradient (See Lemma 5.3 in [55]). However, non-PSD weights are possible to be viewed as a gradient since

$$\frac{\partial \mathbf{y}^\top W \mathbf{y}}{\partial \mathbf{y}} = \left(W + W^\top\right)\mathbf{y}, \tag{68}$$

where $W + W^\top$ is symmetric but not necessarily PSD. Moreover, if the activation function is linear or meets certain criteria, it is also possible to use symmetric weights [55].

Furthermore, there is other work showing that models with symmetric weights are also universal approximators [23], which suggests that using symmetric weights might not adversely affect the expressivity of the model provided the hidden dimension can be increased sufficiently.

F.2 Layer-dependent Weights

Note that since in the unfolding framework every step is the optimization process of a certain energy function, the model must exhibit a characteristic of recursive models like in (4), where the same weight matrix is shared at each layer. However, as has been discussed in various contexts, if the number of model layers is finite, it is possible to use shared weights to "simulate" layer-dependent weights [3, 52, 55]. Moreover, from the construction in [55], when the number of model layers is finite, the weights can even be asymmetric.

F.3 LayerNorm

In the main paper, the model is actually a simplified version of the Transformer since LayerNorm is missing. Here we note that it is possible to combine LayerNorm in our proposed unfolding model. Apart from parameters, LayerNorm can be viewed as following process:

1. Translation: $\boldsymbol{u}^{(t)} = \mathbf{y}^{(t)} - \mathbf{1}\sum_{i=1}^{d}\mathbf{y}_i^{(t)}$,
2. Rescaling: $\mathbf{y}^{(t+1)} = \boldsymbol{u}^{(t)} / \left\|\boldsymbol{u}^{(t)}\right\|$;

which can both be interpreted as gradient steps:

1. Translation: $\boldsymbol{u}^{(t)} = \mathbf{y}^{(t)} - \frac{\partial}{\partial \mathbf{y}}\left(\sum_{i=1}^{d}\mathbf{y}_i^{(t)}\right)^2$,
2. Rescaling: $\mathbf{y}^{(t+1)} = \boldsymbol{u}^{(t)} - \frac{\partial}{\partial \boldsymbol{u}^{(t)}}\left(\frac{1}{2}\left\|\boldsymbol{u}^{(t)}\right\|^2 - \left\|\boldsymbol{u}^{(t)}\right\|\right)$.

Therefore, using the techniques in Section 4, we can insert LayerNorm in an unfolding model by adding extra energy terms. Moreover, with LayerNorm applied, the reweighted softmax in (39) becomes normal softmax.

F.4 Other forms of Residual Connections

Apart from the residual connections mentioned in Section B.3, it is also possible to derive other forms like initial residual connections that directly connect the input and current layer [7]. In order to achieve this, we can change the $R(Y)$ in (9) and Theorem 3.1 from $\frac{1}{2}\|Y\|_{\mathcal{F}}^2$ to $\frac{1}{2}\|Y - B\|_{\mathcal{F}}^2$, then the basic update in (11) becomes

$$Y^{(t+1)} = \mathrm{softmax}_{\boldsymbol{\beta}}\left(Y^{(t)}Y^{(t)\top}\right) + B, \tag{69}$$

where $B \in \mathbb{R}^{n\times d}$ can be any bias term that is not depended on Y. For example if the entries of B are learnable parameters, then it just changes the linear transformations to affine transformations, and if $B = f(X)$ where $f(X) \in \mathbb{R}^{n\times d}$ is some transformation of input features X, then B serves as an initial residual connection just like the one used in [7]. The same can also be done with the $\|Y\|_{\mathcal{F}}^2$ term in E_2 defined in Section 4.2.