GLOBALLY 1-LIPSCHITZ ATTENTION WITHOUT SEQUENCE-LENGTH DEPENDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-attention powers modern deep learning; however, dot-product attention is not globally Lipschitz, which limits stability and robustness. Prior fixes enforce Lipschitz continuity by changing the geometry of attention, which departs from the standard mechanism and still yields bounds that scale poorly with sequence length or spectral norms. We introduce AttLip, a new attention block that derives coefficients from a convex potential and realizes them through an implicit proximal update, guaranteeing architectural stability. This design ensures the entire block is firmly non-expansive and thus *unconditionally 1-Lipschitz*, independent of sequence length or parameter norms, while remaining structurally close to dot-product attention. Building on monotone operator theory, we establish its contractive properties and develop efficient first-order solvers. Experiments on OpenWebText show that AttLip achieves meaningful token mixing and retains learning capacity.

1 INTRODUCTION

Self-attention (Bahdanau et al., 2015; Vaswani et al., 2017) is the driving force behind modern deep learning, powering large language models (LLMs) (Brown et al., 2020; Bubeck et al., 2023) and vision transformers (ViTs) (Dosovitskiy et al., 2021; Radford et al., 2021). Despite its success, its mathematical behavior is less well understood than that of classical layers such as convolutions, for which operator norms and Lipschitz constants are well characterized (Sedghi et al., 2019; Delattre et al., 2023). In particular, dot-product attention is not globally Lipschitz (Kim et al., 2021), meaning small input perturbations can induce disproportionately large output changes. This raises concerns for optimization stability (Qi et al., 2023) and robustness (Zhou et al., 2022; Cissé et al., 2017). The fragility of neural networks to imperceptible adversarial perturbations (Biggio et al., 2013; Szegedy et al., 2014) has motivated certified defenses that provide provable robustness guarantees (Raghunathan et al., 2018; Wong & Kolter, 2018). A principled route to certification is to enforce Lipschitz continuity: if the global Lipschitz constant of a network is bounded, one can derive certified robustness radii around any input. Two strategies have emerged. Randomized smoothing (Lécuyer et al., 2019; Cohen et al., 2019) scales to large datasets but is constrained by impossibility results (Yang et al., 2020). Alternatively, networks can be built from 1-Lipschitz layers, ensuring tight global control (Cissé et al., 2017; Li et al., 2019; Anil et al., 2019; Trockman & Kolter, 2021b; Singla & Feizi, 2021). While well studied for convolutions and linear layers, Lipschitz control of attention remains largely unresolved. Recent work has addressed this gap along two directions. First, Lipschitz-constrained variants such as ℓ_2-attention (Kim et al., 2021; Dasoulas et al., 2021) enforce global continuity but yield bounds that scale poorly with sequence length. Second, local smoothness analyses of dot-product attention (Castin et al., 2024; Xixu, 2023) provide finer control but again degrade unfavorably with sequence length. As a result, standard attention remains highly expressive but inherently non-Lipschitz, while Lipschitz alternatives offer guarantees that become vacuous at realistic context sizes.

We propose to bridge this gap with an implicit attention block, denoted AttLip. The key idea is to reinterpret attention coefficients as the gradient of a convex potential (Meunier et al., 2022) and to realize them through an implicit proximal update. This construction guarantees that the block is firmly non-expansive and thus globally 1-Lipschitz, independent of sequence length or parameter norms. Intuitively, the convex potential encodes pairwise interactions, while the proximal formulation enforces stability by design. In this way, AttLip provides rigorous Lipschitz control without abandoning the expressive token-mixing behavior that makes attention effective. Our main contributions are:

1. **First global 1-Lipschitz attention block independent of sequence length.** We introduce AttLip, an implicit attention mechanism derived from convex potential theory and realized through a proximal operator, ensuring strict 1-Lipschitzness by construction.
2. **Theoretical guarantees and comparison.** We develop an operator-theoretic analysis of AttLip, establish its contractive properties, and compare its guarantees to those of ℓ_2-attention (Kim et al., 2021), highlighting the independence from sequence length.
3. **Practical trainability and empirical validation.** We show that AttLip can be trained efficiently with first-order solvers and retains meaningful learning capacity. Experiments on Transformer pretraining with OpenWebText validate its effectiveness in practice.

2 Related Work

Global Lipschitz control in deep networks. A function f is L-Lipschitz w.r.t. the ℓ_2 norm if $\|f(x) - f(x+\delta)\|_2 \quad \leq \quad L\|\delta\|_2, \qquad \forall\, x, \delta,$ meaning that its output cannot change faster than a constant multiple of the input perturbation. The smallest such constant is the global Lipschitz constant $\mathrm{L}_2(f)$. Controlling $\mathrm{L}_2(f)$ has been shown to improve generalization, through spectrally-normalized margin bounds (Bartlett et al., 2017), and is also central to certified robustness: the certified radius of a classifier scales proportionally to its margin and inversely with $\mathrm{L}_2(f)$ (Tsuzuku et al., 2018). Because computing L_2 exactly is intractable for deep models, two families of approaches have emerged. *Estimation/certification:* AutoLip and SeqLip yield generic and structured upper bounds, while convex relaxations based on quadratic constraints provide tight certificates at scale (Virmaux & Scaman, 2018; Fazlyab et al., 2019). Operator-theoretic analyses further give a posteriori layerwise certificates by modeling activations as averaged (firmly non-expansive) operators (Combettes & Pesquet, 2020). *Design by construction:* spectral normalization constrains operator norms during training (Miyato et al., 2018; Farnia et al., 2019); Parseval regularization encourages near-orthonormal weights (Cissé et al., 2017); exact singular-value characterizations enable norm control for convolutions (Sedghi et al., 2019); and orthogonal/near-isometric parameterizations (e.g., Cayley) preserve norms in deep convnets (Trockman & Kolter, 2021a; Prach & Lampert, 2022; Boissin et al., 2025). On the function-class side, GroupSort networks with norm-constrained weights are universal Lipschitz approximators with refined approximation guarantees (Anil et al., 2019; Tanielian et al., 2021). Dynamical-systems and control viewpoints motivate *implicit/contractive* layers; in particular, convex potential layers realize transformations as resolvents/proximal maps, making them 1-Lipschitz by design (Meunier et al., 2022; Araujo et al., 2023; Wang & Manchester, 2023). Monotone operator theory underpins these guarantees via firm non-expansiveness of resolvents (Bauschke & Combettes, 2017).

Lipschitzness of self-attention. Standard dot-product self-attention is not globally Lipschitz: unbounded logits can induce sharp changes in softmax weights. Two main directions address this limitation. (i) *Local smoothness analyses* bound the *local* Lipschitz constant (or sensitivity), typically with dependence on input norm and sequence length (e.g., worst-case local constants scaling like $\widetilde{\mathcal{O}}(\sqrt{n})$), which limits certification at realistic context sizes (Castin et al., 2024; Havens et al., 2024a). (ii) *Lipschitz-by-modification* enforces global control by altering the mechanism, most commonly by replacing dot products with negative squared distances (*ℓ_2-attention*) or by score normalization (Kim et al., 2021; Dasoulas et al., 2021). While theoretically sound, these variants change attention geometry/distributional behavior, and their bounds often grow with sequence length. Beyond analysis, certified robustness for Transformer-style models has been advanced by (post-hoc) proximal/projection schemes for ViTs that tighten bounds and can swap dot-product attention for ℓ_2-attention without full retraining (Gupta & Verma, 2023); by contractive-flow formulations integrating attention (Mukherjee et al., 2023); and by 1-Lipschitz Transformer variants under additional constraints (Xu et al., 2023). Vision-focused studies further document robustness/sensitivity patterns (Zhou et al., 2022; Qi et al., 2023). Overall, existing results either provide *local*, length-dependent guarantees for dot-product attention, or achieve *global* control at the cost of changing the attention geometry and/or introducing length-dependent constants.

Adversarial robustness in NLP. Modern NLP systems built on self-attention are highly capable yet *fragile* to small input edits. A large body of work shows that character flips, synonym substitutions, short appended phrases, or crafted prompt suffixes can reliably destabilize model behavior, motivating *stability by design* through architectural control of input–output sensitivity.

Early attacks established that tiny, often imperceptible edits can flip predictions. Ebrahimi et al. (2018) introduced *HotFlip*, using gradients to propose character- and word-level substitutions that efficiently degrade accuracy. In a black-box setting, Jin et al. (2020) showed that meaning-preserving synonym swaps reliably fool strong classifiers, including BERT. Follow-ups refined search constraints (e.g., part-of-speech consistency, semantic similarity thresholds) and explored gradient-free strategies and genetic algorithms for discrete sequences (Alzantot et al., 2018; Liang et al., 2018). Tooling such as Morris et al. (2020) (TextAttack) standardized constraints, transformations, and search procedures, enabling reproducible comparisons across models and tasks.

Beyond instance-specific edits, Wallace et al. (2019) discovered *universal adversarial triggers*: short token sequences, optimized once, that when appended consistently elicit targeted behaviors across tasks and architectures. This foreshadowed *jailbreak* suffixes for instruction-tuned LLMs: Zou et al. (2023) ("GCG") demonstrated automatically optimized suffixes that transfer across open and proprietary models, reliably bypassing guardrails.

On the defense side, adversarial training and robust fine-tuning improve empirical robustness but lack worst-case guarantees (Jiang et al., 2020; Aghajanyan et al., 2021). Randomized smoothing provides probabilistic certificates at scale but faces intrinsic limits on achievable radii and sensitivity to distributional mismatch (Lécuyer et al., 2019; Cohen et al., 2019; Yang et al., 2020). A parallel line of work develops *certified defenses* tailored to word substitutions, including interval-bound and smoothing methods (Jia et al., 2019; Ye et al., 2020; Zhao et al., 2022), which demonstrate non-trivial certified robust accuracy on IMDB and YELP benchmarks.

These threads highlight the need for *architectural* stability mechanisms, specifically block-level non-expansiveness that delivers global (sequence–length–independent) guarantees.

3 Background on Self-Attention and Convex Potential Layers

We focus on *self*-attention with *global* context (not cross- or causal attention), which is the regime compatible with our convex/PSD construction.

Standard attention. Consider an input matrix $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ with N tokens of dimension d. For each head h, define the score matrix

$$\boldsymbol{S}_{ij}^{\text{std},h} \;=\; \frac{(\boldsymbol{x}_i \boldsymbol{W}_Q^h)(\boldsymbol{x}_j \boldsymbol{W}_K^h)^\top}{\sqrt{d_h}}, \qquad i, j = 1, \ldots, N. \tag{1}$$

Standard multi-head dot-product self-attention (Vaswani et al., 2017) is then

$$\mathrm{Att}(\boldsymbol{X}) \;=\; \mathrm{concat}_{h=1}^{H}\left(\mathrm{softmax}_{\text{row}}\big(\boldsymbol{S}^{\text{std},h}\big)\, \boldsymbol{X} \boldsymbol{W}_V^h\right) \boldsymbol{W}^O, \tag{2}$$

where $\mathrm{softmax}_{\text{row}}$ applies the softmax function independently to each row, $\boldsymbol{W}_Q^h, \boldsymbol{W}_K^h, \boldsymbol{W}_V^h \in \mathbb{R}^{d \times d_h}$ are the query, key, and value matrices for head h, $\boldsymbol{W}^O \in \mathbb{R}^{H d_h \times d}$ is the output projection, and concat denotes concatenation along the feature dimension.

While extremely effective in practice, this operator is not globally Lipschitz in the ℓ_2 norm. The attention weights depend exponentially on pairwise dot products, which scale quadratically with $\|\boldsymbol{X}\|_2$. Small perturbations of $\boldsymbol{X}$ near score ties can cause abrupt changes in the attention distribution, and this effect grows without bound as $\|\boldsymbol{X}\|$ increases. However, when input $\boldsymbol{X}$ is bounded, one can derive a local Lipschitzness property for self-attention.

Theorem 1 (Multi-head adaptation of the local Lipschitz bound; adapted from Xixu (2023))**.** *Consider a multi-head self-attention layer with H heads concatenated and a final output projection $\boldsymbol{W}_O$. For $\boldsymbol{X}$ in the Frobenius ball $B_2(\boldsymbol{X}_0, \delta_0)$, define for each head*

$$L_h \;=\; N(N+1)\,(\|\boldsymbol{X}_0\|_F + \delta_0)^2 \Big[\|\boldsymbol{W}_V^h\|_2\, \|\boldsymbol{W}_Q^h\|_2\, \|(\boldsymbol{W}_K^h)^\top\|_2 \;+\; \|\boldsymbol{W}_V^h\|_2\Big].$$

Then the multi-head attention mapping Att *is locally Lipschitz on* $B_2(\boldsymbol{X}_0, \delta_0)$ *with*

$$\mathrm{L}_2\big(\mathrm{Att}, \boldsymbol{X}_0\big) \;\le\; \|\boldsymbol{W}_O\|_2 \left(\sum_{h=1}^{H} L_h^2\right)^{1/2}. \tag{3}$$

In parallel, Castin et al. (2024) also derive worst-case bounds on the local Lipschitz constant of self-attention, Havens et al. (2024b) develop a fine-grained local sensitivity analysis of dot-product self-attention, showing that while the layer is not globally Lipschitz, its local sensitivity to small ℓ_2 perturbations can be bounded and used to obtain certified robustness guarantees on vision transformers. However, these guarantees are only valid for small perturbation budgets. However, local Lipschitz analyses of standard dot-product attention remain limited, as the constants depend on the input norm.

Modified Lipschitz attention. This phenomenon was formalized by Kim et al. (2021), who proposed an alternative ℓ_2-attention mechanism that is provably Lipschitz continuous. Specifically, they replace dot-product similarity with negative squared ℓ_2 distances between transformed tokens.

For each head h, define the ℓ_2-attention score matrix

$$\boldsymbol{S}_{ij}^{\mathrm{Kim},h} \;=\; -\frac{1}{\sqrt{d_h}}\,\|\boldsymbol{x}_i\boldsymbol{W}_Q^h - \boldsymbol{x}_j\boldsymbol{W}_K^h\|_2^2, \qquad i,j = 1,\ldots,N, \tag{4}$$

and the multi-head ℓ_2-attention layer

$$\mathrm{Att}_{\ell_2}(\boldsymbol{X}) \;=\; \mathrm{concat}_{h=1}^{H}\left(\mathrm{softmax}_{\text{row}}\big(\boldsymbol{S}^{\mathrm{Kim},h}\big)\,\boldsymbol{X}\boldsymbol{W}_V^h\right)\boldsymbol{W}^O. \tag{5}$$

Note that for this layer to be Lipschitz continuous, weight tying is necessary: $\boldsymbol{W}_Q^h = \boldsymbol{W}_K^h$.

Theorem 2 (Lipschitz bound for ℓ_2-attention (Kim et al., 2021)). *The operator* Att_{ℓ_2} *is Lipschitz continuous in the* ℓ_2 *norm, with Lipschitz constant bounded as*

$$\mathrm{L}_2(\mathrm{Att}_{\ell_2}) \;\le\; \frac{\sqrt{N}}{\sqrt{d_h}}\Big(4\,\Phi^{-1}(N-1)+1\Big)\left(\sqrt{\sum_{h=1}^{H}\|\boldsymbol{W}_Q^h\|_2^2\,\|\boldsymbol{W}_V^h\|_2^2}\right)\|\boldsymbol{W}^O\|_2, \tag{6}$$

where $\Phi(x) = x\exp(x+1), \quad x > 0$, *and* $\Phi^{-1}(N-1) = W_0\big(\frac{N-1}{e}\big)$, *where* W_0 *is the principal branch of the Lambert W function.*

While theoretically sound, the bound scales in a nontrivial way with N, and the operator departs from dot-product attention both geometrically (distance-based vs. inner-product) and in its distributional behaviour. Even in this Lipschitz formulation, the bound remains intricate and explicitly tied to the sequence length N, which is problematic in practice, especially when contrasted with simpler constructions where the Lipschitz constant can be exactly 1 (Trockman & Kolter, 2021a; Meunier et al., 2022; Miyato et al., 2018).

Scalability of Lipschitz bounds. The analytic bound for ℓ_2-attention proposed by Kim et al. (2021) grows with the sequence length N as $\mathcal{O}\Big(\frac{\sqrt{N}}{\sqrt{d_h}}\big(4\phi^{-1}(N-1)+1\big)\Big) \;\sim\; \mathcal{O}(N\log N)$, whereas local Lipschitz constants obtained from the Jacobian typically exhibit quadratic growth, $\mathcal{O}(N^2)$. This scaling becomes problematic given the sequence lengths used in practice: BERT is trained with up to 512 tokens (Devlin et al., 2019), GPT-2 with 1024 (Radford et al., 2019), GPT-3 with 2048 (Brown et al., 2020), and the LLaMA family ranges from 2048 in LLaMA-1 to 4096 in LLaMA-2 and 16k in LLaMA-3 (Touvron et al., 2023). More recent proprietary models push these limits much further, with GPT-4 and Claude 2 supporting context windows of 32k tokens, and specialized versions extending to over 100k. Although the Kim et al. (2021) bound is formally subquadratic, inserting these practical sequence lengths produces Lipschitz constants of astronomical scale, far beyond any interpretable range. These results highlight that such bounds should be viewed primarily as formal guarantees rather than tight measures of effective sensitivity.

This reveals a fundamental tension: standard attention is powerful but non-Lipschitz, while Lipschitz alternatives exist but induce different behaviour. Our objective is therefore to design an attention-like operator that is provably Lipschitz yet retains token-mixing behavior.

3.1 Contractivity from convex potential flow

The input matrix $\boldsymbol{X} \in \mathbb{R}^{N\times d}$ is written in vectorized form as $\boldsymbol{x} = [\boldsymbol{x}_1^\top \dots \boldsymbol{x}_N^\top]^\top \in \mathbb{R}^{Nd}$ with $\boldsymbol{x}_i \in \mathbb{R}^d$. Convex gradient flows are known to generate contractive dynamics in Hilbert spaces, which provides a dimension-free stability guarantee.

Theorem 3 (Contractivity of convex gradient flows (Brézis, 1973))**.** *Let $f : \mathcal{H} \to (-\infty, +\infty]$ be proper, lower semicontinuous, and convex on a Hilbert space $\mathcal{H}$. Consider the (sub)gradient flow*

$$\frac{d\boldsymbol{x}(t)}{dt} + \partial f(\boldsymbol{x}(t)) \ni 0.$$

Then the associated semigroup is nonexpansive:

$$\|\boldsymbol{x}(t) - \boldsymbol{y}(t)\| \le \|\boldsymbol{x}(0) - \boldsymbol{y}(0)\| \qquad \forall\, t \ge 0,$$

for any two solutions $\boldsymbol{x}(\cdot), \boldsymbol{y}(\cdot)$. In particular, if f is differentiable, the gradient flow $\frac{d\boldsymbol{x}(t)}{dt} = -\nabla f(\boldsymbol{x}(t))$ is 1-contractive.

This theorem formalizes the fact that the gradient of a convex potential defines a contractive dynamical system. A more general operator-theoretic result due to Rockafellar (Rockafellar, 1976) shows that the resolvent of any maximal monotone operator is firmly nonexpansive, and therefore 1-Lipschitz. Together, these results provide both the continuous-time and discrete-time foundations for constructing contractive mappings from convex potentials.

Application to neural networks. Building on these principles, Meunier et al. (2022) proposed convex-potential layers where the vector field is taken as $F = -\nabla f$ with f convex. By Theorem 3, the resulting continuous-time flow is contractive, and discretization yields a 1-Lipschitz residual mapping. Their implementation uses the *explicit Euler scheme*,

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \eta F(\boldsymbol{x}_t),$$

which ensures non-expansiveness under smoothness assumptions on f. This line of work is conceptually related to *Input Convex Neural Networks* (ICNNs) (Amos & Kolter, 2017), which enforce convexity through architectural constraints, though here convexity is leveraged to design contractive flows.

4 Designing Lipschitz Attention Mechanism

We now present our construction of a globally 1-Lipschitz attention block. The key idea is to reinterpret attention scores as the gradient of a convex potential, and to realize the update through an *implicit proximal step*. This ensures contractivity by design, independent of sequence length or parameter norms.

4.1 From residual updates to convex potentials

Residual formulation. In the *post-norm* Transformer (Vaswani et al., 2017), ignoring LayerNorm for clarity, the self-attention block reduces to a residual update

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t + \mathrm{Att}(\boldsymbol{X}_t)\boldsymbol{W}_O,$$

with output projection $\boldsymbol{W}_O \in \mathbb{R}^{d_h \times d}$. This is exactly an explicit Euler step of the ODE

$$\frac{d\boldsymbol{X}(t)}{dt} = F(\boldsymbol{X}(t)), \qquad F(\boldsymbol{X}) = \mathrm{Att}(\boldsymbol{X})\,\boldsymbol{W}_O,$$

with step size $\eta = 1$. Similar dynamical-systems interpretations have been explored in neural ODEs, equilibrium models, and convex potential layers (Chen et al., 2018; Bai et al., 2019; Meunier et al., 2022).

Convex potential formulation. We now instantiate this principle for self-attention by defining a convex potential whose gradient has an attention-like structure. Using Theorem 3 we have:

Proposition 1 (Convex Potential for Attention Flow)**.** *Define the convex attention potential*

$$f(\boldsymbol{x}) \;=\; \tfrac{1}{2}\sum_{i=1}^{N} \mathrm{LogSumExp}\big(S^{\text{convex}}_{i,:}(\boldsymbol{x})\big),$$

where the score matrix is given by $S^{\text{convex}}_{ij}(\boldsymbol{x}) = (\boldsymbol{x}_i + \boldsymbol{x}_j)^\top \boldsymbol{W}^\top \boldsymbol{W}(\boldsymbol{x}_i + \boldsymbol{x}_j)$.

Letting $\alpha_{i,j}(\boldsymbol{x}) = \mathrm{softmax}_j\big(S^{\text{convex}}_{i,:}(\boldsymbol{x})\big)$ *denote the row-wise softmax weights, the gradient of* f *takes the attention-like form*

$$\nabla f(\boldsymbol{x})_k = \Big(1 + \sum_{i=1}^{N} \alpha_{i,k}(\boldsymbol{x})\Big)\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{x}_k \;+\; \sum_{i=1}^{N} \big(\alpha_{i,k}(\boldsymbol{x}) + \alpha_{k,i}(\boldsymbol{x})\big)\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{x}_i.$$

Therefore, the induced gradient flow $\frac{d\boldsymbol{x}}{dt} = -\nabla f(\boldsymbol{x})$ *is contractive.*

Our construction enforces $\boldsymbol{W}_Q = \boldsymbol{W}_K = \boldsymbol{W}$ as in Kim et al. (2021) and $\boldsymbol{W}_V = \boldsymbol{W}^\top \boldsymbol{W}$, yielding a symmetric score and attention matrix. This parameter sharing departs from standard attention but is what guarantees convexity. See Algorithm 4 for efficient way to compute $\nabla_f(\boldsymbol{x})$.

Choosing a discretization scheme. In our setting, the convex potentials of interest involve quadratic forms and log-sum-exp terms, which are convex but generally non-smooth. For such functions, explicit Euler no longer guarantees stability. We therefore adopt the *implicit Euler scheme*,

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \nabla f(\boldsymbol{x}_{t+1}), \tag{IE}$$

which remains contractive under convexity alone and can be interpreted as an implicit layer (Bai et al., 2019). Although this step requires solving a nonlinear system, we reinterpret it as a proximal operator, preserving non-expansiveness while allowing efficient first-order solvers in practice.

4.2 Casting as a proximal problem

The implicit Euler step (Eq. IE) involves the unknown $\boldsymbol{x}_{t+1}$ on both sides. It is equivalent to applying the *proximal operator* of f:

$$\boldsymbol{x}_{t+1} \;=\; \operatorname*{prox}_{\eta f}(\boldsymbol{x}_t) \;=\; \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^{Nd}} \Big\{ f(\boldsymbol{x}) \;+\; \tfrac{1}{2\eta}\,\|\boldsymbol{x} - \boldsymbol{x}_t\|_2^2 \Big\}. \tag{P$_t$}$$

This convex reformulation ensures that the update is well posed, admits a unique minimizer, and by maximal monotone operator theory the map $\mathrm{prox}_{\eta f} = (I + \eta\nabla f)^{-1}$ is firmly non-expansive and thus 1-Lipschitz (Bauschke & Combettes, 2017). Hence the implicit attention layer is contractive by design. This reformulation shows that the implicit step is a proximal operator of f, which not only guarantees well-posedness and contractivity, but also makes the update computationally tractable: instead of inverting a large Jacobian, one can solve the convex subproblem (P$_t$) efficiently with first-order methods.

Non-smooth convex optimization. The potential f is convex but not globally smooth. Nevertheless, on any bounded sublevel set, its gradient is Lipschitz, which enables local convergence guarantees. We therefore minimize $\phi_t(\boldsymbol{x}) := f(\boldsymbol{x}) + \frac{1}{2\eta}\|\boldsymbol{x} - \boldsymbol{x}_t\|_2^2$ by gradient descent with updates

$$y^{(t+1)} \leftarrow y^{(t)} - \lambda_t\big(\nabla f(y^{(t)}) + \tfrac{1}{\eta}(y^{(t)} - \boldsymbol{x}_t)\big).$$

When a local Lipschitz constant $L_2(f, R)$ on the relevant sublevel set is available, any fixed step $\lambda \in \big(0,\, 2/(L_2(f,R){+}1/\eta)\big)$ ensures linear convergence. Otherwise, Armijo backtracking automatically selects a valid λ_t at each iteration. In both cases, the iterates converge linearly to $\mathrm{prox}_{\eta f}(\boldsymbol{x}_t)$. The following is a direct application of standard results from convex optimization (e.g., Nesterov (2018), Bauschke & Combettes (2017), Bertsekas (1999)). We restate this standard result for completeness: it ensures the proximal subproblem can be solved to arbitrary precision, so the 1-Lipschitz property of $\mathrm{prox}_{\eta f}$ carries over to the implemented layer (up to solver tolerance).

Proposition 2 (Well-posedness and convergence (informal))**.** *For convex f, the proximal objective*

$$\phi_t(\boldsymbol{x}) = f(x) + \tfrac{1}{2\eta}\|\boldsymbol{x} - \boldsymbol{x}_t\|_2^2$$

is $(1/\eta)$-strongly convex and admits a unique minimizer $\boldsymbol{x}_{t+1} = \mathrm{prox}_{\eta f}(\boldsymbol{x}_t)$.

On a bounded sublevel set around x_t, say contained in a ball $\overline{B}(\boldsymbol{x}_t, R)$ of radius R, the gradient ∇f has a local Lipschitz constant $L_2(f, R)$. In this region, $\nabla \phi_t$ is Lipschitz with constant $L_2(\phi, R) = L_2(f, R) + 1/\eta$, and gradient descent with either fixed stepsizes $\lambda \leq 1/L_2(\phi, R)$ or Armijo backtracking (Armijo, 1966) converges linearly. The rate is controlled by the local condition number $\kappa(R) = 1 + \eta L_2(f, R)$.

Practical solver. In practice, we approximate $\mathrm{prox}_{\eta\, f}(\boldsymbol{x}_t)$ via gradient descent with Armijo backtracking, summarized in Algorithm 1. This ensures convergence to arbitrary precision, so the 1-Lipschitz property of $\mathrm{prox}_{\eta\, f}$ carries over to the implemented layer (up to solver tolerance).

Algorithm 1 Gradient descent with Armijo backtracking for (P$_t$)

```
Require: $\boldsymbol{x}_t \in \mathbb{R}^{Nd}$, tolerance $\varepsilon > 0$
$\boldsymbol{y}^{(0)} \leftarrow \boldsymbol{x}_t$
for $k = 0, 1, 2, \dots$ do
    $\boldsymbol{g}^{(k)} \leftarrow \nabla f(\boldsymbol{y}^{(k)}) + \frac{1}{\eta}(\boldsymbol{y}^{(k)} - \boldsymbol{x}_t)$
    if $\|\boldsymbol{g}^{(k)}\|_2 \leq \varepsilon$ then
        break
    end if
    Choose $\lambda_k$ by Armijo backtracking
    $\boldsymbol{y}^{(k+1)} \leftarrow \boldsymbol{y}^{(k)} - \lambda_k \boldsymbol{g}^{(k)}$
end for
$\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{y}^{(k)}$
```

Armijo backtracking (see Algorithm 2) requires an initial step guess. In practice, setting λ_{init} from an estimate of the local Lipschitz constant L provides a good starting point and often accelerates convergence. Note that the inner iteration of Armijo does not require gradient accumulation.

Role of the step parameter. A central advantage of the implicit (proximal) scheme is that $\eta > 0$ can be chosen without restriction: for any value, $\mathrm{prox}_{\eta\, f}$ remains firmly non-expansive and therefore 1-Lipschitz. In practice, η serves as a tunable hyperparameter: larger values increase the influence of f, while smaller values improve the conditioning of the subproblem. The choice of η directly impacts the local condition number $\kappa(R) = 1 + \eta L_2(f, R)$, and can thus be guided by local Lipschitz estimates of ∇f (see Appendix A.3.2).

From dynamical flow to block. Our construction starts from a continuous-time gradient flow derived from a convex potential and stabilizes it via an implicit Euler step. Reformulating this step as a proximal operator yields a tractable update rule that is contractive by design. At the discrete level, this update defines the actual attention block, taking an input $\boldsymbol{x}_t$ and producing an output $\boldsymbol{x}_{t+1}$.

Theorem 4 (Multi-head convex attention potential)**.** *For each head $h = 1, \dots, H$, define the convex potential*

$$f^h(\boldsymbol{x}) \;=\; \tfrac{1}{2}\sum_{i=1}^{N} \mathrm{LogSumExp}\big(\boldsymbol{z}_i^h(\boldsymbol{x})\big), \qquad \boldsymbol{z}_i^h(\boldsymbol{x}) = \Big\{ S_{ij}^{convex,h}(\boldsymbol{x}) \Big\}_{j=1}^{N},$$

and let $f(\boldsymbol{x}) \;=\; \sum_{h=1}^{H} f^h(\boldsymbol{x})$. Then the implicit update

$$\mathrm{AttLip}_\eta(\boldsymbol{x}_t) \;=\; \operatorname*{prox}_{\eta\, f}(\boldsymbol{x}_t)$$

defines the multi-head convex attention block. The resulting mapping $x_t \mapsto \mathrm{AttLip}_\eta(x_t) = x_{t+1}$ is firmly non-expansive and therefore 1-Lipschitz.

Thus, the block we implement is both faithful to attention structure and globally 1-Lipschitz—unlike ad-hoc clipping or normalization, which retain sequence-length dependence, while remaining independent of sequence length or parameter norms.

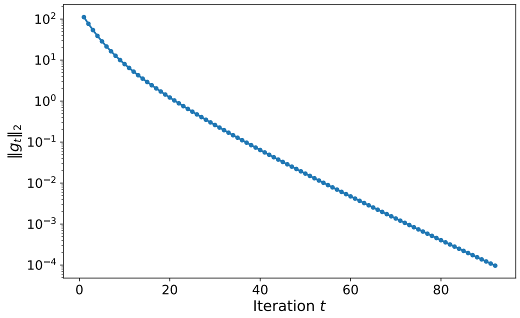

Figure 1: Convergence speed of the inner gradient descent solver for the proximal subproblem (P_t). We report $\|g^{(t)}\|_2$ as a function of inner iteration t for $N = 252$, embedding dimension $d = 512$, and a single head ($H = 1$). Although the proximal objective is convex but not smooth, the residual decreases rapidly within a few iterations.

Layer	Metric	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$	$N = 1024$	$N = 2048$
AttLip (ours)	PGD	0.991	0.991	0.990	0.990	0.989	0.987	0.987	0.987
	Theoretical	1	1	1	1	1	1	1	1
Att_{ℓ_2}	PGD	2.521	2.480	2.558	2.551	2.513	2.622	2.455	2.518
	Theoretical	28.7	52.6	90.9	153.7	253.1	407.7	662.5	1055.5

Table 1: Comparison between our AttLip (ours) block and the Att_{ℓ_2} layer of Kim et al. (2021), with embedding dimension $d = 512$ and number of heads $H = 8$. We report the empirical Lipschitz constant estimated by projected gradient ascent (PGD) and the corresponding theoretical upper bound. Here N denotes the input sequence length. We run the solver for $K = 20$ iterations in our layer to ensure convergence. Our bound is fixed to 1, independent of N, while Att_{ℓ_2} grows as $O(N \log N)$ and becomes rapidly vacuous.

5 Experiments

In all experiments, we arbitrarily fix $\eta = 1$, and for the Armijo rule we use standard $\beta = 0.5$, $c = 10^{-4}$, with at most 20 backtracking steps.

5.1 Theoretical validations

Speed of convergence of proximal problem Figure 1 shows that the solver (Algorithm 1 used) converges quickly even without smoothness assumptions. While general convex non-smooth optimization admits only $O(1/t)$ rates, the structure of ϕ_k leads to much faster empirical decay, keeping the overhead of the proximal formulation modest in practice. Empirically, the convergence rate seems independent of N, so the cost of computing the proximal update remains stable as sequence length grows.

Comparison with previous Lipschitz attention Table 1 highlights the difference between our AttLip and the Att_{ℓ_2} layer of Kim et al. (2021). All layers are initialized with the standard initialization used in Transformer architectures (Gaussian weights with variance scaling, as in Vaswani et al. (2017)). A key advantage of our construction is that the theoretical Lipschitz bound is *independent of the sequence length* N and remains fixed at 1 by design. This contrasts with Att_{ℓ_2}, whose theoretical bound grows linearly with N and quickly becomes vacuous. Moreover, our theoretical bound is much tighter and consistently matches the empirical estimates obtained with adversarial search ("PGD max"), where we maximize the local Lipschitz ratio by projected gradient ascent within an ℓ_2-ball around the input (Shi et al., 2022), while Att_{ℓ_2} exhibits a large gap between its loose theoretical guarantee and its empirical Lipschitz behavior. In large-scale foundation models such as LLaMA–65B (Touvron et al., 2023), the Transformer stack contains up to 80 layers. In such regimes, estimating the global Lipschitz constant quickly becomes computationally prohibitive, and the resulting bound can be extremely large and thus uninformative.

These results confirm that our bound is not only formally tight but also practically verifiable, which is a prerequisite before testing whether such stability constraints still permit non-trivial learning capacity.

5.2 One-layer experiments on OpenWebText

We train on the OpenWebText corpus in the non-causal next-token prediction setting, following the preprocessing and hyperparameters of the original RoBERTa paper (Liu et al., 2019). As a first step, we restrict to small one-layer networks in order to isolate the role of attention. Specifically, we compare four variants: (i) a baseline with standard dot-product attention, (ii) Att_{ℓ_2} from Kim et al. (2021), (iii) a one-layer RoBERTa model without attention (MLP head only), and (iv) our proposed block AttLip_η, which enforces global 1-Lipschitzness. This setup allows us to test whether token-mixing through different attention mechanisms improves learning beyond feedforward-only architectures.

For our method, we denote by K_{prox} the number of inner gradient steps used to approximately solve the proximal subproblem in each layer, and by K_{armijo} the iterations of Armijo backtracking. Since Armijo steps do not accumulate gradients, their cost is negligible compared to K_{prox}. In practice, we train with very few proximal steps (typically $K_{\text{prox}} = 1$–3) and use larger values only at inference. To further reduce cost, we adopt a simple schedule where training begins with a single step and gradually increases to two or more. This keeps the overhead modest while ensuring strict 1-Lipschitzness at inference.

Model	**Lipschitz constant**	**Val. PPL**	**PPL × Lip**
MLP (no attention)	—	988	—
Standard attention	unbounded	22	—
Att_{ℓ_2} (Kim et al., 2021)	2117	28	5.9×10^4
Our block AttLip	1	141	141

Table 2: Validation perplexity and Lipschitz constants on OpenWebText (1-layer models). Both metrics should be minimized. The product (PPL × Lip) provides a simple joint measure of efficiency and robustness: **lower values are better**.

The results in Table 2 highlight the trade-off between expressivity and contractivity. Our block enforces global 1-Lipschitzness by design, a qualitatively different guarantee than existing mechanisms. Although its validation perplexity (141) is higher than that of Att_{ℓ_2} (28) or standard attention (22), the combined score (PPL × Lip) provides a simple proxy measure for robust learning and shows that AttLip is the only mechanism achieving non-trivial modeling ability under exact Lipschitz control (141 vs. 5.9×10^4 for Att_{ℓ_2}). This demonstrates that non-trivial token interactions can be captured while maintaining strict robustness guarantees. Standard attention attains the best perplexity but at the cost of being non-Lipschitz, whereas Att_{ℓ_2} sacrifices robustness entirely with a very large constant.

6 Conclusion

We introduced a novel self-attention mechanism based on convex proximal flows, which guarantees global 1-Lipschitz continuity by construction. Our approach provides a principled alternative to previous Lipschitz attention variants, combining operator-theoretic contractivity with efficient approximate solvers. The experiment on OpenWebText demonstrates that, despite operating under strict Lipschitz constraints, the proposed block can capture non-trivial token dependences.

The present work establishes a theoretical and algorithmic foundation for Lipschitz attention in large-scale architectures. Future work should investigate scaling to deeper models and applying the framework to downstream tasks where robustness guarantees are critical. In particular, extending the analysis to adversarial perturbations in the embedding space for NLP remains an open direction.

REFERENCES

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. Better fine-tuning by reducing representational collapse. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=OQ08SN70M1V.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2890–2896, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1316. URL https://aclanthology.org/D18-1316/.

Brandon Amos and J. Zico Kolter. Input convex neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Cem Anil, James Lucas, and Roger B. Grosse. Sorting out lipschitz function approximation. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Alexandre Araujo, Aaron Havens, Blaise Delattre, Alexandre Allauzen, and Bin Hu. A unified algebraic perspective on Lipschitz neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Larry Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16, 1966.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, 2015.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2017.

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, 1999.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2013.

Thibaut Boissin, Franck Mamalet, Thomas Fel, Agustin Martin Picard, Thomas Massena, and Mathieu Serrurier. An adaptive orthogonal convolution scheme for efficient and flexible CNN architectures. In *Forty-second International Conference on Machine Learning*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Haim Brézis. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*. North-Holland, 1973.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Yin Tat Lee, Yi Lee, Guido Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv*, 2023.

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Moustapha Cissé, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Patrick L. Combettes and Jean-Christophe Pesquet. Lipschitz certificates for layered network structures driven by averaged activation operators. *SIAM Journal on Mathematics of Data Science*, 2020.

George Dasoulas, Kevin Scaman, and Aladin Virmaux. Lipschitz normalization for self-attention layers with application to graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Blaise Delattre, Quentin Barthélemy, Alexandre Araujo, and Alexandre Allauzen. Efficient bound of Lipschitz constant for convolutional layers by gram iteration. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. HotFlip: White-box adversarial examples for text classification. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 31–36, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2006. URL `https://aclanthology.org/P18-2006/`.

Farzan Farnia, Jesse Zhang, and David Tse. Generalizable adversarial training via spectral normalization. In *International Conference on Learning Representations*, 2019.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George J. Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Kavya Gupta and Sagar Verma. Certvit: Certified robustness of pre-trained vision transformers, 2023. URL `https://arxiv.org/abs/2302.10287`.

Aaron Havens, Alexandre Araujo, Huan Zhang, and Bin Hu. Fine-grained local sensitivity analysis of standard dot-product self-attention. In *Proceedings of the 41st International Conference on Machine Learning*, 2024a.

Aaron J Havens, Alexandre Araujo, Huan Zhang, and Bin Hu. Fine-grained local sensitivity analysis of standard dot-product self-attention, 2024b.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. Certified robustness to adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4129–4143, 2019.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2177–2190, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.197. URL `https://aclanthology.org/2020.acl-main.197/`.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 8018–8025, 2020.

Hyunjik Kim, George Deligiannidis, Michael Bronstein, and Yee Whye Teh. The lipschitz constant of self-attention. In *International Conference on Machine Learning (ICML)*, 2021.

Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, 2019.

Qiyang Li, Saminul Haque, Cem Anil, James R Lucas, Roger B Grosse, and Jörn-Henrik Jacobsen. Preventing gradient attenuation in lipschitz constrained convolutional networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. Deep text classification can be fooled. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pp. 4208–4215. AAAI Press, 2018. ISBN 9780999241127.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Laurent Meunier, Blaise J. Delattre, Alexandre Araujo, and Alexandre Allauzen. A dynamical system perspective for lipschitz neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 119–126, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.16. URL `https://aclanthology.org/2020.emnlp-demos.16/`.

Avideep Mukherjee, Badri Narayan Patro, and Vinay P. Namboodiri. Attentive contractive flow with lipschitz constrained self-attention. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2023.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer, 2018.

Bernd Prach and Christoph H. Lampert. Almost-orthogonal layers for efficient general-purpose Lipschitz networks. In *Computer Vision – ECCV 2022*, 2022.

Xianbiao Qi, Jianan Wang, Yihao Chen, Yukai Shi, and Lei Zhang. Lipsformer: Introducing lipschitz continuity to vision transformers. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 1976.

Hanie Sedghi, Vineet Gupta, and Philip M. Long. The singular values of convolutional layers. In *International Conference on Learning Representations*, 2019.

Zhouxing Shi, Yihan Wang, Huan Zhang, J. Zico Kolter, and Cho-Jui Hsieh. Efficiently computing local lipschitz constants of neural networks via bound propagation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Sahil Singla and Soheil Feizi. Skew orthogonal convolutions. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. URL `https://proceedings.mlr.press/v139/singla21a.html`.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Ugo Tanielian, Maxime Sangnier, and Gérard Biau. Approximating lipschitz continuous functions with groupsort neural networks. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, 2021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Asher Trockman and J. Zico Kolter. Orthogonalizing convolutional layers with the cayley transform. In *International Conference on Learning Representations*, 2021a.

Asher Trockman and J Zico Kolter. Orthogonalizing convolutional layers with the cayley transform. In *International Conference on Machine Learning (ICML)*, 2021b.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: Analysis and efficient estimation. In *Advances in Neural Information Processing Systems*, 2018.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2153–2162, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1221. URL `https://aclanthology.org/D19-1221/`.

Ruigang Wang and Ian R. Manchester. Direct parameterization of Lipschitz-bounded deep networks. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Eric Wong and J. Zico Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

H. Xixu. Specformer: Guarding vision transformer robustness via maximum singular value penalization. In *Conference on Parsimony and Learning (Recent Spotlight Track)*, 2023.

Xiaojun Xu, Linyi Li, Yu Cheng, Subhabrata Mukherjee, Ahmed Hassan Awadallah, and Bo Li. Certifiably robust transformers with 1-lipschitz self-attention, 2023. URL `https://openreview.net/forum?id=hzG72qB0XQ`.

Greg Yang, Tony Duan, J. Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020.

Mao Ye, Chengyue Gong, and Qiang Liu. SAFER: A structure-free approach for certified robustness to adversarial word substitutions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3465–3475, 2020.

Haiteng Zhao, Chang Ma, Xinshuai Dong, Anh Tuan Luu, Zhi-Hong Deng, and Hanwang Zhang. Certified robustness against natural language attacks by causal intervention. In *International Conference on Machine Learning (ICML)*, pp. 26911–26929. PMLR, 2022.

Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M. Alvarez. Understanding the robustness in vision transformers. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A APPENDIX

A.1 FINITE-K APPROXIMATION AND THE ROLE OF η

In practice, the implicit convex-attention update is computed by running only a finite number K of iterations of the inner solver rather than solving the proximal subproblem exactly. The behaviour of this truncated update is governed by the proximal parameter η, which induces a natural tradeoff between convergence speed and expressivity. When η is small, the quadratic term $(2\eta)^{-1}\|\boldsymbol{x} - \boldsymbol{x}_t\|^2$ strongly penalizes deviations from the input, so the update remains close to the identity map; the subproblem is then strongly convex and the inner iterations converge quickly. Increasing η weakens this quadratic regularization, allowing the update to move further from $\boldsymbol{x}_t$ and thus yielding a more expressive transformation, but at the cost of slower inner convergence due to reduced strong convexity. The next proposition shows that the resulting finite-K update remains close to the exact 1-Lipschitz proximal map, with an explicit error term controlled by the solver's residual.

In the following, let $f : \mathbb{R}^d \to \mathbb{R}$ be the convex potential corresponding to the multi-head convex attention block of Theorem 4, after reshaping $\boldsymbol{x} \in \mathbb{R}^{N\times d_{\text{model}}}$ into a vector in $\mathbb{R}^d$ with $d = Nd_{\text{model}}$.

Proposition 3 (Approximate convex attention)**.** *Fix $\eta > 0$. For each input $\boldsymbol{x}_t \in \mathbb{R}^d$, define*

$$\varphi_{\boldsymbol{x}_t}(\boldsymbol{x}) \;:=\; f(\boldsymbol{x}) \;+\; \frac{1}{2\eta}\,\|\boldsymbol{x} - \boldsymbol{x}_t\|_2^2,$$

and denote the exact proximal update by

$$\operatorname{AttLip}_\eta(\boldsymbol{x}_t) \;:=\; \operatorname*{prox}_{\eta f}(\boldsymbol{x}_t) \;=\; \arg\min_{\boldsymbol{x}\in\mathbb{R}^d}\, \varphi_{\boldsymbol{x}_t}(\boldsymbol{x}).$$

Assume that, instead of $\operatorname{AttLip}_\eta(\boldsymbol{x}_t)$, the block implements an approximate solution $\operatorname{AttLip}_\eta^{(K)}(\boldsymbol{x}_t)$ obtained after K inner iterations of a solver, and define the gradient residual

$$\boldsymbol{g}_K(\boldsymbol{x}_t) \;:=\; \nabla_{\boldsymbol{x}}\varphi_{\boldsymbol{x}_t}\big(\operatorname{AttLip}_\eta^{(K)}(\boldsymbol{x}_t)\big) \;\in\; \mathbb{R}^d.$$

Suppose there exists a function $\varepsilon : \mathbb{R}^d \to [0, +\infty)$ such that

$$\|\boldsymbol{g}_K(\boldsymbol{x}_t)\|_2 \;\le\; \varepsilon(\boldsymbol{x}_t) \qquad \textit{for all } \boldsymbol{x}_t \in \mathbb{R}^d.$$

Then:

1. *(Pointwise approximation of the exact convex attention block) for all $\boldsymbol{x}_t \in \mathbb{R}^d$,*

$$\|\operatorname{AttLip}_\eta^{(K)}(\boldsymbol{x}_t) - \operatorname{AttLip}_\eta(\boldsymbol{x}_t)\|_2 \;\le\; \eta\,\varepsilon(\boldsymbol{x}_t).$$

2. *(Almost 1-Lipschitz behaviour) for all $\boldsymbol{x}_t, \boldsymbol{y}_t \in \mathbb{R}^d$,*

$$\|\operatorname{AttLip}_\eta^{(K)}(\boldsymbol{x}_t) - \operatorname{AttLip}_\eta^{(K)}(\boldsymbol{y}_t)\|_2 \;\le\; \|\boldsymbol{x}_t - \boldsymbol{y}_t\|_2 \;+\; \eta\big(\varepsilon(\boldsymbol{x}_t) + \varepsilon(\boldsymbol{y}_t)\big).$$

   In particular, if $\varepsilon(\boldsymbol{x}_t) \le \varepsilon$ for all $\boldsymbol{x}_t$, then

$$\|\operatorname{AttLip}_\eta^{(K)}(\boldsymbol{x}_t) - \operatorname{AttLip}_\eta^{(K)}(\boldsymbol{y}_t)\|_2 \;\le\; \|\boldsymbol{x}_t - \boldsymbol{y}_t\|_2 + 2\,\eta\,\varepsilon \qquad \forall\, \boldsymbol{x}_t, \boldsymbol{y}_t \in \mathbb{R}^d.$$

Proof. Fix $\boldsymbol{x}_t \in \mathbb{R}^d$. The function

$$\varphi_{\boldsymbol{x}_t}(\boldsymbol{x}) = f(\boldsymbol{x}) + \tfrac{1}{2\eta}\|\boldsymbol{x} - \boldsymbol{x}_t\|_2^2$$

is μ-strongly convex with $\mu = 1/\eta$, hence it has a unique minimizer, denoted $\boldsymbol{x}^\star(\boldsymbol{x}_t) := \operatorname{AttLip}_\eta(\boldsymbol{x}_t)$.

Strong convexity yields, for all $\boldsymbol{x} \in \mathbb{R}^d$,

$$\big\|\nabla\varphi_{\boldsymbol{x}_t}(\boldsymbol{x}) - \nabla\varphi_{\boldsymbol{x}_t}(\boldsymbol{x}^\star(\boldsymbol{x}_t))\big\|_2 \;\ge\; \mu\,\|\boldsymbol{x} - \boldsymbol{x}^\star(\boldsymbol{x}_t)\|_2.$$

Since $\nabla\varphi_{\boldsymbol{x}_t}(\boldsymbol{x}^\star(\boldsymbol{x}_t)) = \boldsymbol{0}$, we obtain

$$\|\nabla\varphi_{\boldsymbol{x}_t}(\boldsymbol{x})\|_2 \;\ge\; \mu\,\|\boldsymbol{x} - \boldsymbol{x}^\star(\boldsymbol{x}_t)\|_2.$$

Applying this to $\boldsymbol{x} = \mathrm{AttLip}_\eta^{(K)}(\boldsymbol{x}_t)$ gives

$$\|\mathrm{AttLip}_\eta^{(K)}(\boldsymbol{x}_t) - \mathrm{AttLip}_\eta(\boldsymbol{x}_t)\|_2 \;\le\; \frac{1}{\mu}\,\|\nabla\varphi_{\boldsymbol{x}_t}(\mathrm{AttLip}_\eta^{(K)}(\boldsymbol{x}_t))\|_2 = \eta\,\|\boldsymbol{g}_K(\boldsymbol{x}_t)\|_2 \;\le\; \eta\,\varepsilon(\boldsymbol{x}_t),$$

which proves item (1).

For item (2), recall that $\mathrm{AttLip}_\eta = \mathrm{prox}_{\eta f}$ is firmly non-expansive in $(\mathbb{R}^d, \|\cdot\|_2)$, hence

$$\|\mathrm{AttLip}_\eta(\boldsymbol{x}_t) - \mathrm{AttLip}_\eta(\boldsymbol{y}_t)\|_2 \;\le\; \|\boldsymbol{x}_t - \boldsymbol{y}_t\|_2.$$

For arbitrary $\boldsymbol{x}_t, \boldsymbol{y}_t$,

$$\begin{aligned}\|\mathrm{AttLip}_\eta^{(K)}(\boldsymbol{x}_t) - \mathrm{AttLip}_\eta^{(K)}(\boldsymbol{y}_t)\|_2 &\le \|\mathrm{AttLip}_\eta(\boldsymbol{x}_t) - \mathrm{AttLip}_\eta(\boldsymbol{y}_t)\|_2 + \|\mathrm{AttLip}_\eta^{(K)}(\boldsymbol{x}_t) - \mathrm{AttLip}_\eta(\boldsymbol{x}_t)\|_2 \\ &\quad + \|\mathrm{AttLip}_\eta^{(K)}(\boldsymbol{y}_t) - \mathrm{AttLip}_\eta(\boldsymbol{y}_t)\|_2 \\ &\le \|\boldsymbol{x}_t - \boldsymbol{y}_t\|_2 + \eta\,\varepsilon(\boldsymbol{x}_t) + \eta\,\varepsilon(\boldsymbol{y}_t),\end{aligned}$$

using item (1) for both $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$. If $\varepsilon(\boldsymbol{x}_t) \le \varepsilon$ for all $\boldsymbol{x}_t$, then

$$\|\mathrm{AttLip}_\eta^{(K)}(\boldsymbol{x}_t) - \mathrm{AttLip}_\eta^{(K)}(\boldsymbol{y}_t)\|_2 \;\le\; \|\boldsymbol{x}_t - \boldsymbol{y}_t\|_2 + 2\,\eta\,\varepsilon,$$

completing the proof. □

Composition of approximate proximal layers. If several finite-iteration proximal blocks are stacked, each satisfying an almost non-expansive bound of the form $\|F_\ell(u) - F_\ell(v)\|_2 \le \|u - v\|_2 + 2\eta_\ell\varepsilon_\ell$, then their composition also remains almost 1-Lipschitz, with a combined additive defect equal to the sum of the individual ones:

$$\|F_L \circ \cdots \circ F_1(u) - F_L \circ \cdots \circ F_1(v)\|_2 \;\le\; \|u - v\|_2 + 2\sum_{\ell=1}^{L} \eta_\ell\varepsilon_\ell.$$

Thus, a deep stack of such implicit layers preserves stability, and the Lipschitz degradation accumulates in a transparent and explicit controlled way.

A.2 Algorithms

Algorithm 2 Backtracking Armijo rule for ϕ_t

```
Require: $0 < \beta < 1,\ c \in (0,1)$
$\lambda \leftarrow \lambda_{\text{init}}$                                   (e.g. from local $L$ estimate)
while $\phi_t(\boldsymbol{y}^{(k)} - \lambda\boldsymbol{g}^{(k)}) \;>\; \phi_t(\boldsymbol{y}^{(k)}) - c\lambda\|\boldsymbol{g}^{(k)}\|_2^2$ do
    $\lambda \leftarrow \beta\lambda$
end while
return $\lambda$
```

In practice, we use standard Armijo (Algorithm 2) parameters with backtracking factor $\beta = 0.5$, slope constant $c = 10^{-4}$, and at most 20 backtracking steps.

Algorithm 3 Standard multi-head attention

```
Require: Input $\boldsymbol{X} \in \mathbb{R}^{N\times d}$
for $h = 1, \dots, H$ do
    $\boldsymbol{Q}^h \leftarrow \boldsymbol{X}\boldsymbol{W}_{Q,h}^\top, \quad \boldsymbol{K}^h \leftarrow \boldsymbol{X}\boldsymbol{W}_{K,h}^\top, \quad \boldsymbol{V}^h \leftarrow \boldsymbol{X}\boldsymbol{W}_{V,h}^\top$
    $\boldsymbol{S}^h \leftarrow \frac{1}{\sqrt{d_h}}\boldsymbol{Q}^h{\boldsymbol{K}^h}^\top$
    $\boldsymbol{A}^h \leftarrow \mathrm{softmax}_{\text{row}}(\boldsymbol{S}^h)$
    $\boldsymbol{Y}^h \leftarrow \boldsymbol{A}^h\boldsymbol{V}^h$
end for
return $\mathrm{concat}_{h=1}^{H}\boldsymbol{Y}_h\,\boldsymbol{W}^O$
```

Algorithm 4 Gradient of convex attention potential $\nabla f(\boldsymbol{X})$

```
Require: Input $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, head projections $\{\boldsymbol{W}^h\}_{h=1}^{H}$
for $h = 1, \dots, H$ do
    $\boldsymbol{V}^h \leftarrow \boldsymbol{X}(\boldsymbol{W}^h)^\top \quad \in \mathbb{R}^{N\times d_h}$
    $\mathbf{u}^h \leftarrow \|\boldsymbol{V}^h\|_2^2 \quad \in \mathbb{R}^N$    (rowwise squared norms: $(\mathbf{u}^h)_i = \|\boldsymbol{V}_i^h\|_2^2$)
    $\boldsymbol{S}^h \leftarrow \frac{2}{\sqrt{d_h}}\boldsymbol{V}^h(\boldsymbol{V}^h)^\top + \frac{1}{\sqrt{d_h}}\big(\mathbf{u}^h\mathbf{1}^\top + \mathbf{1}(\mathbf{u}^h)^\top\big)$    ($S_{ij}^h = \frac{1}{\sqrt{d_h}}(\|V_i^h\|^2 + \|V_j^h\|^2 + 2(V_i^h)^\top V_j^h)$)
    $\boldsymbol{S}^h \leftarrow \boldsymbol{S}^h - \mathrm{rowmax}(\boldsymbol{S}^h)$    (stability, row-wise)
    $\boldsymbol{A}^h \leftarrow \mathrm{softmax}_{\mathrm{row}}(\boldsymbol{S}^h)$
    $\boldsymbol{A}^h \leftarrow \boldsymbol{A}^h + (\boldsymbol{A}^h)^\top$    (adds column contributions)
    $\boldsymbol{D}^h \leftarrow \mathrm{diag}\big((\boldsymbol{A}^h)^\top\mathbf{1}\big) + \boldsymbol{I}_N$    (self term $1 + \sum_i \alpha_{i,k}^h$)
    $\boldsymbol{Y}^h \leftarrow (\boldsymbol{A}^h + \boldsymbol{D}^h)\,\boldsymbol{V}^h \quad \in \mathbb{R}^{N\times d_h}$
end for
return $\sum_{h=1}^{H} \boldsymbol{Y}^h\boldsymbol{W}^h \ \in \mathbb{R}^{N\times d}$
```

A.3 Proofs

A.3.1 Proof of Proposition 4

Lemma 1 (Convexity of the attention potential)**.** *Let*

$$f(\boldsymbol{x}) \;=\; \frac{1}{2}\sum_{i=1}^{N} \mathrm{LogSumExp}\big(\boldsymbol{z}_i(\boldsymbol{x})\big), \qquad \boldsymbol{z}_i(\boldsymbol{x}) = \big\{(\boldsymbol{x}_i + \boldsymbol{x}_j)^\top \boldsymbol{W}^\top\boldsymbol{W}(\boldsymbol{x}_i + \boldsymbol{x}_j)\big\}_{j=1}^{N},$$

with $\boldsymbol{x} = [\boldsymbol{x}_1^\top \dots \boldsymbol{x}_N^\top]^\top \in \mathbb{R}^{Nd}$, $\boldsymbol{W} \in \mathbb{R}^{d_{\mathrm{head}}\times d}$ and $\boldsymbol{W}^\top\boldsymbol{W} \succeq 0$. Then f is convex over $\mathbb{R}^{Nd}$.

Proof. Fix i and j. Write

$$z_{i,j}(\boldsymbol{x}) = (\boldsymbol{x}_i + \boldsymbol{x}_j)^\top\boldsymbol{W}^\top\boldsymbol{W}(\boldsymbol{x}_i + \boldsymbol{x}_j) = \|\,\boldsymbol{W}(\boldsymbol{x}_i + \boldsymbol{x}_j)\,\|_2^2.$$

Let $\boldsymbol{e}_i, \boldsymbol{e}_j \in \mathbb{R}^N$ be the canonical vectors and define the linear map $T_{i,j} : \mathbb{R}^{Nd} \to \mathbb{R}^d$ by $T_{i,j}\boldsymbol{x} = \boldsymbol{W}\big((\boldsymbol{e}_i^\top \otimes \boldsymbol{I}_d)\boldsymbol{x} + (\boldsymbol{e}_j^\top \otimes \boldsymbol{I}_d)\boldsymbol{x}\big)$. Then $z_{i,j}(\boldsymbol{x}) = \|T_{i,j}\boldsymbol{x}\|_2^2 = (T_{i,j}\boldsymbol{x})^\top(T_{i,j}\boldsymbol{x}) = \boldsymbol{x}^\top T_{i,j}^\top T_{i,j}\boldsymbol{x}$, a quadratic form with PSD matrix $T_{i,j}^\top T_{i,j} \succeq 0$, hence $z_{i,j}$ is convex. Therefore the vector $\boldsymbol{z}_i(\boldsymbol{x}) = (z_{i,1}(\boldsymbol{x}), \dots, z_{i,N}(\boldsymbol{x}))$ has convex components.

The map $\mathrm{LogSumExp} : \mathbb{R}^N \to \mathbb{R}$, defined by $\mathrm{LogSumExp}(\boldsymbol{u}) = \log\sum_{k=1}^{N} e^{u_k}$, is convex and (componentwise) nondecreasing. By the monotone composition rule, $\mathrm{LogSumExp}(\boldsymbol{z}_i(\boldsymbol{x}))$ is convex as a composition of a convex, nondecreasing function with a vector of convex functions. Summing over i and multiplying by $1/2$ preserves convexity, hence f is convex. □

Proof. By Lemma 1, f is convex and differentiable, so ∇f is a monotone operator. Applying the contractivity result for gradient systems driven by convex potentials (Theorem 3) to $F = -\nabla f$ yields the claimed nonexpansiveness in the Euclidean norm. □

A.3.2 Proof of Proposition 2

First, let us restate Proposition 2 more formally:

Proposition 4 (Well-posedness and conditioning of the proximal subproblem (formal))**.** *Let $N, d, H \in \mathbb{N}_{\geq 1}$, identify $(\mathbb{R}^d)^N \simeq \mathbb{R}^{Nd}$, and fix PSD matrices $\boldsymbol{A}^h \succeq 0$ for $h = 1, \dots, H$. For $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_N) \in \mathbb{R}^{Nd}$ with $\boldsymbol{x}_i \in \mathbb{R}^d$, define the scores*

$$S_{i,j}^h(\boldsymbol{x}) := (\boldsymbol{x}_i + \boldsymbol{x}_j)^\top\boldsymbol{A}^h(\boldsymbol{x}_i + \boldsymbol{x}_j), \qquad f(\boldsymbol{x}) := \tfrac{1}{2}\sum_{h=1}^{H}\sum_{i=1}^{N} \mathrm{LogSumExp}\big(S_{i,1}^h(\boldsymbol{x}), \dots, S_{i,N}^h(\boldsymbol{x})\big).$$

Fix $\eta > 0$ and $\boldsymbol{x}_t \in \mathbb{R}^{Nd}$, and consider

$$\phi_t(\boldsymbol{x}) := f(\boldsymbol{x}) + \frac{1}{2\eta}\|\boldsymbol{x} - \boldsymbol{x}_t\|^2, \qquad \boldsymbol{x}_{t+1} := \arg\min_{\boldsymbol{x}\in\mathbb{R}^{Nd}} \phi_t(\boldsymbol{x}).$$

Let the sublevel set $D := \{\boldsymbol{x} : \phi_t(\boldsymbol{x}) \le \phi_t(\boldsymbol{x}_t)\}$ *and choose* $R > 0$ *with* $D \subset \overline{B}(\boldsymbol{x}_t, R)$*. Define the local smoothness constant*

$$L_2(f, R) := \sup_{\boldsymbol{x} \in \overline{B}(\boldsymbol{x}_t, R)} \|\nabla^2 f(\boldsymbol{x})\| < \infty,$$

then set $L_2(\phi, R) := L_2(f, R) + 1/\eta$ *and the local condition number* $\kappa(R) := 1 + \eta L_2(f, R)$*. Then:*

1. ***Strong convexity, uniqueness, proximal characterization.*** ϕ_t *is* $(1/\eta)$*-strongly convex on* $\mathbb{R}^{Nd}$*, hence admits a unique minimizer* $\boldsymbol{x}_{t+1}$*. Moreover* $\boldsymbol{x}_{t+1} = \operatorname{prox}_{\eta f}(\boldsymbol{x}_t) = (I + \eta\,\partial f)^{-1}(\boldsymbol{x}_t) = (I + \eta\,\nabla f)^{-1}(\boldsymbol{x}_t)$*.*

2. ***Bounded sublevel set and local smoothness.*** D *is nonempty, closed, and bounded. Furthermore,* $f \in C^2$*, so* $\nabla^2 f$ *is continuous and* $L_2(f, R) < \infty$*. Consequently,* ∇f *is* $L_2(f, R)$*-Lipschitz on* $\overline{B}(\boldsymbol{x}_t, R)$ *and* $\nabla \phi_t$ *is* $L_2(\phi, R)$*-Lipschitz on* $\overline{B}(\boldsymbol{x}_t, R)$*.*

3. ***Gradient descent with guaranteed containment and linear rate.*** *Let* $\boldsymbol{y}^{(0)} := \boldsymbol{x}_t$ *and, for any* $\lambda \in \big(0,\, 1/L_2(\phi, R)\big]$*, define* $\boldsymbol{y}^{(k+1)} := \boldsymbol{y}^{(k)} - \lambda\,\nabla\phi_t(\boldsymbol{y}^{(k)})$*. Then* $\boldsymbol{y}^{(k)} \in D \subset \overline{B}(\boldsymbol{x}_t, R)$ *for all* k*, and*

$$\|\boldsymbol{y}^{(k)} - \boldsymbol{x}_{t+1}\| \;\le\; \Big(1 - \frac{1}{\kappa(R)}\Big)^k \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|, \qquad k = 0, 1, 2, \ldots.$$

4. ***Armijo backtracking: step lower bound and linear convergence.*** *Fix* $\sigma \in (0, 1/2)$ *and* $\beta \in (0, 1)$*. At iteration* k*, choose* λ_k *by backtracking from any trial step, shrinking by* β *until*

$$\phi_t\big(\boldsymbol{y}^{(k)} - \lambda_k \nabla\phi_t(\boldsymbol{y}^{(k)})\big) \;\le\; \phi_t(\boldsymbol{y}^{(k)}) - \sigma\,\lambda_k\,\|\nabla\phi_t(\boldsymbol{y}^{(k)})\|^2.$$

Then $\lambda_k \ge \beta(1-\sigma)/L_2(\phi, R)$ *for all* k*,* $\boldsymbol{y}^{(k)} \in D$ *for all* k*, and there exists* $c = c(\beta, \sigma) \in (0, 1)$ *such that*

$$\|\boldsymbol{y}^{(k)} - \boldsymbol{x}_{t+1}\| \;\le\; \Big(1 - \frac{c}{\kappa(R)}\Big)^k \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\| \qquad \text{for all } k \in \mathbb{N}.$$

Proof. **(1) Strong convexity and proximal map.** Each $S^h_{i,j}$ is convex quadratic since $\boldsymbol{A}^h \succeq 0$. The map LogSumExp is convex and coordinatewise nondecreasing; thus $\operatorname{LogSumExp} \circ (S^h_{i,1}, \ldots, S^h_{i,N})$ is convex, and so is f (finite everywhere). The map $\boldsymbol{x} \mapsto \frac{1}{2\eta}\|\boldsymbol{x} - \boldsymbol{x}_k\|^2$ is $(1/\eta)$-strongly convex, hence $\phi_t = f + \frac{1}{2\eta}\|\cdot - \boldsymbol{x}_t\|^2$ is $(1/\eta)$-strongly convex and admits a unique minimizer $\boldsymbol{x}_{t+1}$. The first-order optimality condition $0 \in \partial f(\boldsymbol{x}_{t+1}) + \frac{1}{\eta}(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t)$ reads $\boldsymbol{x}_{t+1} = (\mathrm{Id} + \eta\,\partial f)^{-1}(\boldsymbol{x}_t)$; since f is C^1 convex, $\partial f = \{\nabla f\}$ and $\boldsymbol{x}_{t+1} = (I + \eta\,\nabla f)^{-1}(\boldsymbol{x}_t) = \operatorname{prox}_{\eta f}(\boldsymbol{x}_t)$.

(2) Bounded sublevel set; local smoothness. Nonemptiness: $\boldsymbol{x}_t \in D$ by definition. Closedness: ϕ_t is continuous. To show boundedness, note $S^{(h)}_{i,j}(x) \ge 0$ for all x and (h, i, j), hence $\operatorname{LogSumExp}(S^h_{i,1}(\boldsymbol{x}), \ldots, S^h_{i,N}(\boldsymbol{x})) \ge \log\big(\sum_{j=1}^N e^0\big) = \log N$. Therefore $f(\boldsymbol{x}) \ge \frac{1}{2} HN \log N$, and

$$\phi_t(\boldsymbol{x}) \;\ge\; \frac{1}{2\eta}\|\boldsymbol{x} - \boldsymbol{x}_t\|^2 + \frac{1}{2} HN \log N \;\xrightarrow[\|\boldsymbol{x}\| \to \infty]{}\; \infty,$$

so ϕ_t is coercive and D is bounded. Since f is a finite sum of C^∞ functions, $f \in C^2$ and $\nabla^2 f$ is continuous. Hence $L_2(f, R) := \sup_{\boldsymbol{x} \in \overline{B}(\boldsymbol{x}_t, R)} \|\nabla^2 f(\boldsymbol{x})\| < \infty$ and ∇f is $L_2(f, R)$-Lipschitz on $\overline{B}(\boldsymbol{x}_t, R)$. Adding the quadratic gives that $\nabla\phi_t$ is $L_2(\phi, R) = L_2(f, R) + 1/\eta$-Lipschitz on $\overline{B}(\boldsymbol{x}_r, R)$.

(3) Gradient descent: containment and rate for $\lambda \le 1/L_2(\phi, R)$. Let $\mu := 1/\eta$ and $L := L_2(\phi, R)$. Because ϕ_t is convex, each sublevel set is convex; in particular, D is convex. On $\overline{B}(x_k, R)$ the *descent lemma* holds: for any $\boldsymbol{u}, \boldsymbol{v}$ with the segment $[\boldsymbol{u}, \boldsymbol{v}] \subset \overline{B}(\boldsymbol{x}_t, R)$,

$$\phi_t(\boldsymbol{v}) \;\le\; \phi_t(\boldsymbol{u}) + \langle \nabla\phi_t(\boldsymbol{u}), \boldsymbol{v} - \boldsymbol{u} \rangle + \frac{L}{2}\|\boldsymbol{v} - \boldsymbol{u}\|^2.$$

Apply it with $\boldsymbol{u} = \boldsymbol{y}^{(k)}$ and $v = y^{(k)} - \lambda\nabla\phi_k(y^{(k)})$. For any $\lambda \in (0, 1/L]$ we obtain

$$\phi_t(\boldsymbol{y}^{(k+1)}) \;\le\; \phi_t(\boldsymbol{y}^{(k)}) - \Big(\lambda - \tfrac{L}{2}\lambda^2\Big)\|\nabla\phi_t(\boldsymbol{y}^{(k)})\|^2 \;\le\; \phi_t(\boldsymbol{y}^{(k)}) - \tfrac{\lambda}{2}\|\nabla\phi_t(\boldsymbol{y}^{(k)})\|^2 \;\le\; \phi_t(\boldsymbol{y}^{(k)}).$$

By induction, $\phi_t(\boldsymbol{y}^{(k)}) \le \phi_t(x_t)$ for all k, so $y^{(k)} \in D \subset \overline{B}(\boldsymbol{x}_t, R)$. Since D is convex, the segment between $\boldsymbol{y}^{(k)}$ and $\boldsymbol{y}^{(k+1)}$ lies in D and thus in $\overline{B}(\boldsymbol{x}_t, R)$, justifying the use of the descent lemma at every step.

For the linear rate in distance, observe that $\phi_t \in C^2$ and, for all $\boldsymbol{x} \in \overline{B}(x_t, R)$,

$$\mu\mathrm{Id} \;\preceq\; \nabla^2\phi_t(\boldsymbol{x}) \;=\; \nabla^2 f(\boldsymbol{x}) + \tfrac{1}{\eta}\mathrm{Id} \;\preceq\; LI.$$

Let $\boldsymbol{x}_\star := x_{t+1}$ and write $\boldsymbol{e}^{(k)} := \boldsymbol{y}^{(k)} - \boldsymbol{x}_\star$. By the mean value form of the gradient,

$$\boldsymbol{e}^{(k+1)} = \boldsymbol{e}^{(k)} - \lambda\big(\nabla\phi_t(\boldsymbol{y}^{(k)}) - \nabla\phi_t(\boldsymbol{x}_\star)\big) = \Big(\mathrm{Id} - \lambda\,\overline{H}_k\Big)\boldsymbol{e}^{(k)},$$

where $\overline{H}_k := \int_0^1 \nabla^2\phi_t\big(\boldsymbol{x}_\star + \theta\boldsymbol{e}^{(k)}\big)\,d\theta$. Each $\overline{H}_k$ is symmetric with spectrum in $[\mu, L]$, hence

$$\|\boldsymbol{e}^{(k+1)}\| \;\le\; \|I - \lambda\overline{H}_k\|\,\|\boldsymbol{e}^{(k)}\| \;\le\; \rho\,\|\boldsymbol{e}^{(k)}\|, \qquad \rho := \max\{|1-\lambda\mu|,\; |1-\lambda L|\}.$$

For $\lambda \in (0, 1/L]$ one has $\rho = 1-\lambda\mu \le 1-\mu/L = 1-1/\kappa(R)$, so $\|\boldsymbol{e}^{(k)}\| \le (1-1/\kappa(R))^k\|\boldsymbol{e}^{(0)}\|$, as claimed.

(4) Armijo backtracking. Fix $\sigma \in (0, 1/2)$ and $\beta \in (0, 1)$. On $\overline{B}(\boldsymbol{x}_t, R)$ the descent lemma implies that, for any $\boldsymbol{y}$ and any $\lambda \le (1-\sigma)/L$,

$$\phi_t\big(\boldsymbol{y} - \lambda\nabla\phi_t(\boldsymbol{y})\big) \;\le\; \phi_t(\boldsymbol{y}) - \lambda\Big(1 - \frac{L\lambda}{2}\Big)\|\nabla\phi_t(\boldsymbol{y})\|^2 \;\le\; \phi_t(\boldsymbol{y}) - \sigma\lambda\|\nabla\phi_t(\boldsymbol{y})\|^2.$$

Therefore, the backtracking procedure will accept some step $\lambda_k \ge \beta(1-\sigma)/L$ at each iteration. Monotonicity then yields $\boldsymbol{y}^{(k)} \in L$ for all t and (as above) all segments $[\boldsymbol{y}^{(k)}, \boldsymbol{y}^{(k+1)}] \subset L \subset \overline{B}(\boldsymbol{x}_t, R)$, so the local smoothness bound is valid throughout.

To obtain a linear rate, combine the accepted Armijo decrease with the Polyak–Łojasiewicz inequality valid for μ-strongly convex functions:

$$\phi_t(\boldsymbol{y}^{(k+1)}) - \phi_t(\boldsymbol{x}_\star) \;\le\; \phi_t(\boldsymbol{y}^{(k)}) - \phi_t(\boldsymbol{x}_\star) - \sigma\lambda_k\|\nabla\phi_t(\boldsymbol{y}^{(k)})\|^2 \;\le\; \Big(1 - 2\sigma\lambda_k\,\mu\Big)\Big(\phi_t(\boldsymbol{y}^{(k)}) - \phi_t(\boldsymbol{x}_\star)\Big).$$

Using $\lambda_k \ge \beta(1-\sigma)/L$ gives $\phi_t(\boldsymbol{y}^{(k)}) - \phi_t(\boldsymbol{x}_\star) \le \big(1 - c\,\mu/L\big)^t\big(\phi_t(\boldsymbol{x}_t) - \phi_t(\boldsymbol{x}_\star)\big)$ with $c := 2\beta\sigma(1-\sigma) \in (0, 1)$. By strong convexity, $\|\boldsymbol{y}^{(k)} - \boldsymbol{x}_\star\|^2 \le \frac{2}{\mu}\big(\phi_t(\boldsymbol{y}^{(k)}) - \phi_t(\boldsymbol{x}_\star)\big)$, which yields the stated distance contraction with factor $1 - c/\kappa(R)$(after adjusting c if desired). □

Remark On steps larger than $1/L_2(\phi, R)]$ The classical bound $\rho = \max\{|1-\lambda\mu|, |1-\lambda L|\} < 1$ also holds for any $\lambda \in (0, 2/L_2(\phi, R))$ *provided* all iterates remain in $\overline{B}(\boldsymbol{x}_t, R)$ (so that L-smoothness applies along the segments used in the mean-value argument). Our statement enforces containment by requiring $\lambda \le 1/L_2(\phi, R)$ or by using Armijo backtracking.

Heuristic scaling of the local smoothness. Write $\boldsymbol{A}^h = \boldsymbol{W}^{h\top}\boldsymbol{W}^h \succeq 0$ and

$$f(\boldsymbol{x}) = \tfrac{1}{2}\sum_{h=1}^{H}\sum_{i=1}^{N}\mathrm{LogSumExp}\Big(\,S^h_{i,j}(\boldsymbol{x}) \;:\; j = 1, \ldots, N\Big), \qquad S^h_{i,j}(\boldsymbol{x}) = (\boldsymbol{x}_i + \boldsymbol{x}_j)^\top\boldsymbol{A}^h(\boldsymbol{x}_i + \boldsymbol{x}_j).$$

For each fixed (h, i), $\nabla_{\boldsymbol{x}} S^{(h)}_{i,j}$ is *linear* in $\boldsymbol{x}$ and supported only on blocks $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, with $\|\nabla S^h_{i,j}(x)\| \lesssim \|\boldsymbol{A}^h\|\,(\|\boldsymbol{x}_i\| + \|\boldsymbol{x}_j\|)$. On a ball of radius R, this gives $\|\nabla S^h_{i,j}(\boldsymbol{x})\| \lesssim \|\boldsymbol{A}^h\|\,R$. The Hessian of LogSumExp composed with $\{S^h_{i,j}\}_j$ splits as a convex combination of the constant Hessians $\nabla^2 S^h_{i,j}$ (each bounded by $\|\boldsymbol{A}^h\|$ in operator norm) plus a covariance term of the gradients, bounded by $\lesssim \|\boldsymbol{A}^h\|^2R^2$. Summing over $i = 1, \ldots, N$ contributes a linear factor N for each head. Therefore, on $\{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{x}_k\| \le R\}$ one obtains the local bound

$$\|\nabla^2 f(\boldsymbol{x})\| \;\lesssim\; \sum_{h=1}^{H} N\Big(\|\boldsymbol{A}^h\| + \|\boldsymbol{A}^h\|^2R^2\Big), \qquad \Rightarrow \qquad L_2(f, R) \;\lesssim\; \sum_{h=1}^{H} N\Big(\|\boldsymbol{A}^h\| + \|\boldsymbol{A}^h\|^2R^2\Big).$$

If the weights are spectrally normalized so that $\|\boldsymbol{W}^h\| \leq 1$ (hence $\|\boldsymbol{A}_h\| \leq 1$), this simplifies to

$$L_2(f, R) \;\lesssim\; H\,N\,(1 + R^2).$$

Consequently, the local condition number in Proposition **??** behaves as

$$\kappa(R) = 1 + \eta L(f, R) \;\lesssim\; 1 + \eta \sum_{h=1}^{H} N\Big(\|\boldsymbol{A}^h\| + \|\boldsymbol{A}^h\|^2 R^2\Big),$$

which is linear in N under spectral normalization.