A Constrained Optimization Perspective of Unrolled Transformers

Javier Porras-Valenzuela, Samar Hadou, Alejandro Ribeiro

University of Pennsylvania

NEURAL INFORMATION PROCESSING SYSTEMS

Transformers as Optimizers
Transformers can be interpreted as descent algorithms

- Neural network unrolling interprets transformers as optimizers. Each layer is an optimization step that decreases an objective.
- **Problem:** In practice, training unconstrained transformers may result in non-monotonic losses along its layers.

$$\mathbf{T}_{\mathsf{U}}^{*} = \underset{\mathbf{T}}{\operatorname{argmin}}\, \mathbb{E}\Big[f\big(\mathbf{X}, \Phi(\mathbf{X};\mathbf{T})\big)\Big] \qquad \text{(ERM)}$$

- This leads to less robustness to OOD perturbations.

Training transformers to descend improves OOD robustness

- **Key Idea:** Constrain layers to decrease loss, like descent algorithms.
- Consider a transformer given by equations

$$\mathbf{Z}_l = \mathbf{V}_l\mathbf{X}_{l-1} \times \text{softmax}\Big[(\mathbf{Q}_l\mathbf{X}_{l-1})^T(\mathbf{K}_l\mathbf{X}_{l-1})\Big]$$

$$\Phi_l(\mathbf{X};\mathbf{T}) = \sigma\Big[\mathbf{W}_l\mathbf{Z}_l + \mathbf{U}_l\mathbf{X}_{l-1}\Big]$$

- Train each layer Φ_l to decrease f with step size $0 < \alpha < 1$:

Constrained Unrolled Transformer

$$\mathbf{T}^{*} = \underset{\mathbf{T}}{\operatorname{argmin}}\, \mathbb{E}\Big[f\big(\mathbf{X}, \Phi(\mathbf{X};\mathbf{T})\big)\Big] \qquad \text{(P-CUT)}$$
$$\text{s.t. } \mathbb{E}\Big[f\big(\mathbf{X}, \Phi_l(\mathbf{X};\mathbf{T})\big)\Big] \leq (1-\alpha_l)\mathbb{E}\Big[f\big(\mathbf{X}, \Phi_{l-1}(\mathbf{X};\mathbf{T})\big)\Big],\ \forall\, l$$

- Models trained to optimize this problem exhibit **enhanced robustness to OOD perturbations.**

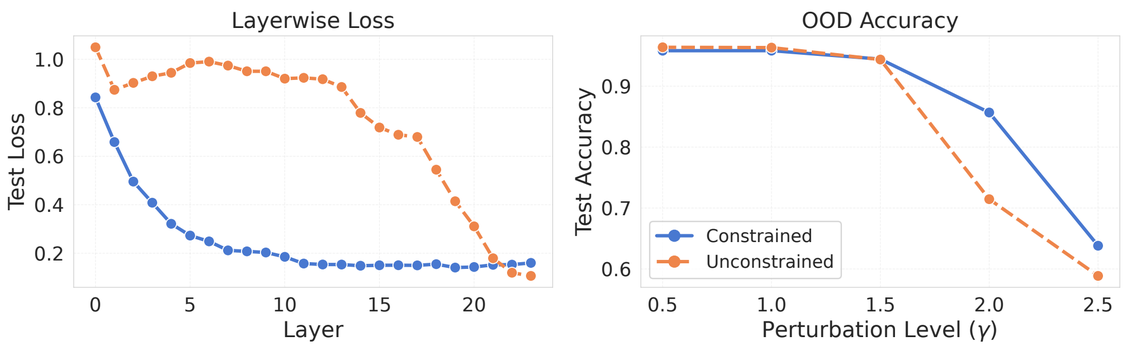

Figure 1. Unconstrained transformers show non-monotonic loss; constrained models descend smoothly with better OOD robustness. (Left: ↓, Right: ↑)

Enforce constraints with primal-dual training

- Optimize (P-CUT) by alternating primal and dual steps.

Dual problem of (P-CUT)

$$\widehat{D}^{*} = \max_{\boldsymbol{\lambda}\geq 0}\min_{\mathbf{T}} \widehat{\mathcal{L}}(\mathbf{T},\boldsymbol{\lambda})$$

Theoretical Guarantees

- We bound the optimality gap $\Delta_k^{*} := f\big(\mathbf{X}, \Phi_k(\mathbf{X};\mathbf{T}^{*})\big) - f\big(\mathbf{X},\mathbf{Y}^{*}\big)$.
- **Theorem 2 (Convergence):** Constrained transformers converge to near-optimal loss:

$$\lim_{l\to\infty}\min_{k\leq l}\mathbb{E}\Big[\Delta_k^{*}\Big] \leq \frac{1}{\alpha}\left(\zeta(M,\delta) + \frac{C\delta\nu}{1-\delta}\right)$$

- **Theorem 3 (OOD Generalization):** For shifted distribution $D_{x'}$

$$\lim_{l\to\infty}\min_{k\leq l}\mathbb{E}_{D_{x'}}\Big[\Delta_k^{*}\Big] \leq \frac{1}{\alpha}\left(\zeta(M,\delta) + \frac{C\delta\nu}{1-\delta} + C\tau\right)$$

where $\tau = d(D_x, D_{x'}) + d(D_{x'}, D_x)$ measures distribution shift.

Unrolled video denoisers generalize to OOD noise

- **Task:** Reconstruct video $\mathbf{X} = [\mathbf{x}_1,\dots,\mathbf{x}_T]$ from noisy $\tilde{\mathbf{x}}_t = \mathbf{x}_t + \boldsymbol{\epsilon}_t$, with $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mu, \gamma\sigma_x\mathbf{I})$

$$f(\mathbf{X}, \Phi(\tilde{\mathbf{X}};\mathbf{T})) = \frac{1}{2}\sum_{t=1}^{T}\|\mathbf{x}_t - [\Phi(\tilde{\mathbf{X}};\mathbf{T})]_t\|_2$$

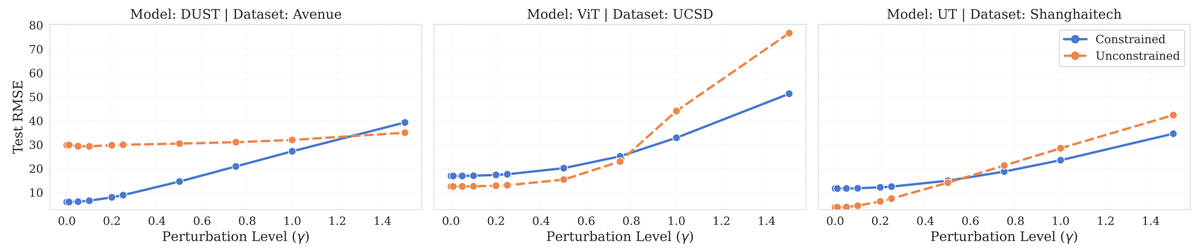

Figure 2. Constraints improve robustness. RMSE vs. test perturbation γ (↓ lower is better).

Unrolled text classifiers are robust to noisy embeddings

- **Task:** Classify text with perturbed embeddings.
- Evaluate layerwise loss by attaching a shared readout layer ψ.

$$f(\tilde{\mathbf{X}}, \mathbf{q}, \Phi(\tilde{\mathbf{X}};\mathbf{T})) = -\sum_{c=1}^{C} q_c \log \psi(\Phi(\tilde{\mathbf{X}};\mathbf{T}))_c$$

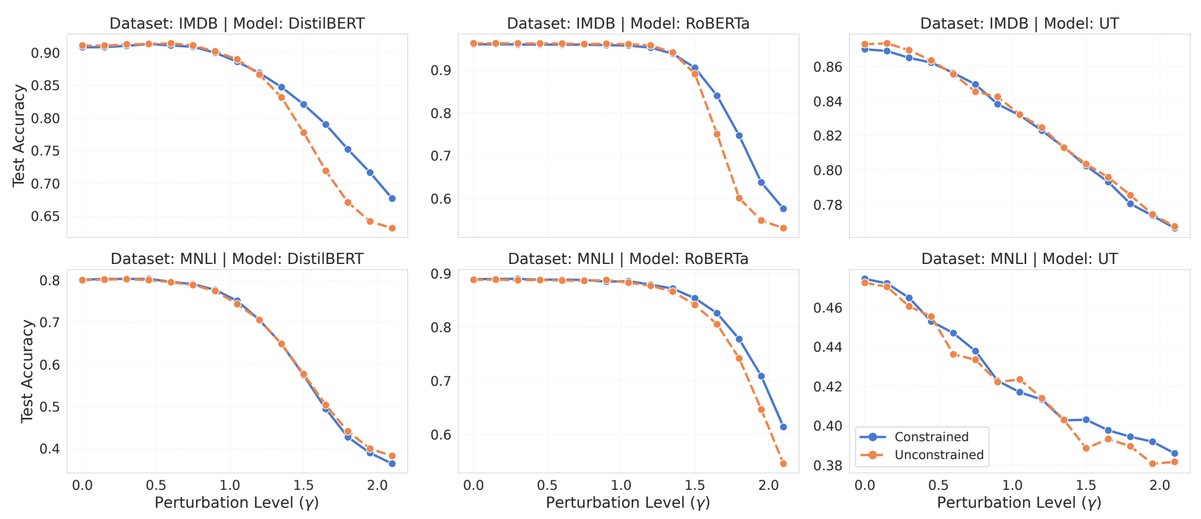

Figure 3. Accuracy vs. test perturbation γ (↑ higher is better).

LLM supervised finetuning: Improves OOD, Maintains ID

- **Setup:** Llama 3.1 8B SFT on Alpaca instruction dataset with perturbed token embeddings.

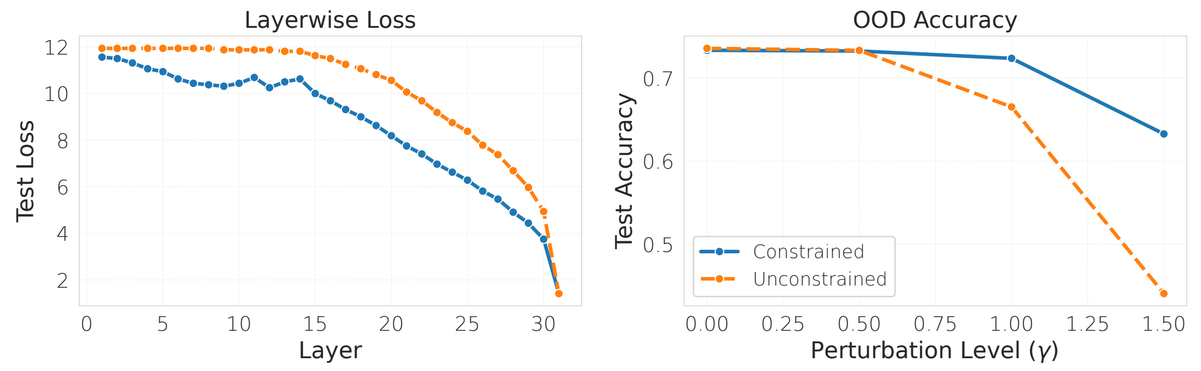

Figure 4. Constraints improve eval token accuracy (Left: ↓, Right: ↑).

- **AlpacaEval** win rate vs unconstrained: 50% at $\gamma = 0.0$, 69.83% at $\gamma = 1.5$ → preserved ID while robust OOD.

Larger step size α increases robustness

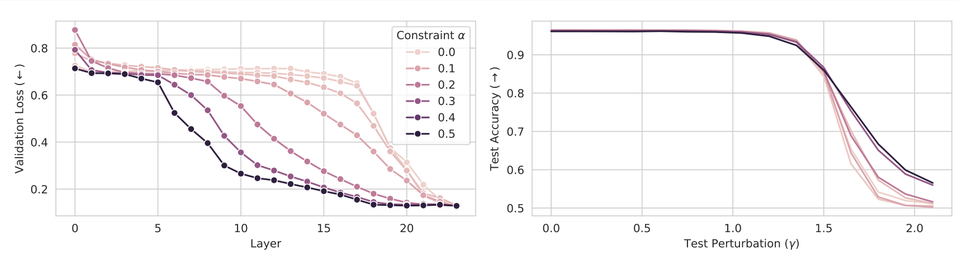

Figure 5. Increasing α improves monotonic descent and OOD robustness.

jporras@seas.upenn.edu, selaraby@seas.upenn.edu