# A Constrained Optimization Perspective of Unrolled Transformers

Javier Porras-Valenzuela, Samar Hadou, Alejandro Ribeiro

University of Pennsylvania

ICML International Conference On Machine Learning

## Transformers interpreted as descent algorithms

- Algorithmic unrolling interprets transformer layers as optimization steps, therefore loss should decrease layerwise.
- Standard ERM training:

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

- **Problem:** ERM training optimizes only for last-layer value of loss function $f$. The resulting unconstrained transformer is non-monotonic across layers.
- Leads to decreased robustness to OOD perturbations.

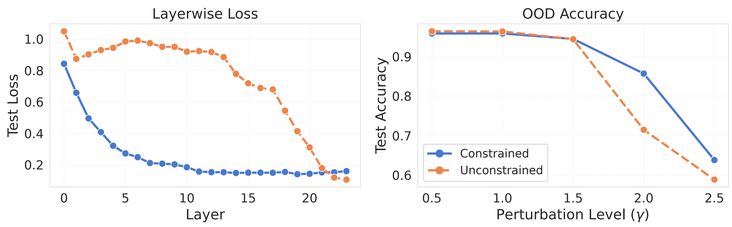


Figure 1. Unconstrained transformers show non-monotonic loss, constrained models descend smoothly with better OOD robustness. (Left: ↓, Right: ↑)

## Training transformers to descend improves OOD robustness

- **Key Idea:** We impose constraints to decrease loss monotonically in expectation, inspired by descent algorithms.
- Consider transformer given by layers

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

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

- Train each layer $\mathbf{\Phi}_l$ to decrease $f$ with step size $0 < \alpha < 1$:

**Constrained Unrolled Transformer**

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

- Lagrangian of (P-CUT):

$$\mathcal{L}(\mathbf{T}, \boldsymbol{\lambda}) = \mathbb{E}\Big[f(\mathbf{X}, \mathbf{\Phi}(\mathbf{X}; \mathbf{T}))\Big] + \sum_{l=1}^{L} \lambda_l \mathbb{E}\Big[f\big(\mathbf{X}, \mathbf{\Phi}_l(\mathbf{X}; \mathbf{T})\big) - (1-\alpha_l) f\big(\mathbf{X}, \mathbf{\Phi}_{l-1}(\mathbf{X}; \mathbf{T})\big)\Big],$$

- Optimize (P-CUT) by alternating primal and dual steps on the empirical dual problem:

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

## Convergence & OOD Generalization guarantees

We bound the optimality gap $\Delta_k^{*} := f\big(\mathbf{X}, \mathbf{\Phi}_k(\mathbf{X}; \mathbf{T}^{*})\big) - f\big(\mathbf{X}, \mathbf{Y}^{*}\big)$.

- **Theorem 1 (Convergence):** Constrained transformers attain 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 2 (OOD Generalization):** Under 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.

- Bounds depend on sample complexity, expressivity, distribution distance and step size $\alpha$.

## Unrolled text classifiers are robust to noisy embeddings

- Leverage robustness to **distribution shifts** to classify with perturbed embeddings $\tilde{\mathbf{X}}$.
- Evaluate layerwise loss by attaching a shared readout layer $\psi$.

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

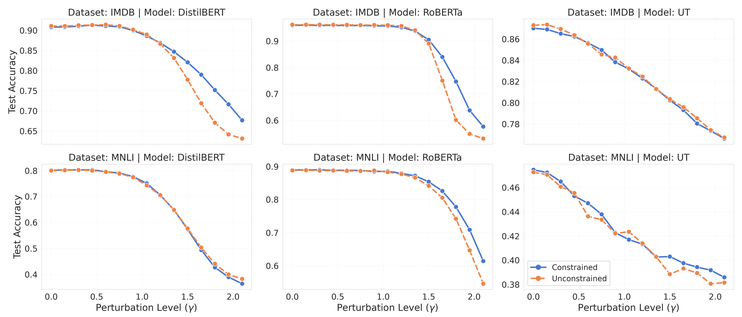


Figure 2. Accuracy vs. test perturbation $\gamma$ (↑ higher is better).

- Similar results with uniform noise and discrete perturbations:

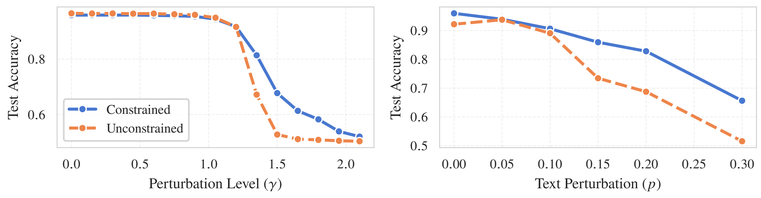


Figure 3. Uniformly distributed perturbations (left) and random text perturbations (character and word level, right).

## LLM supervised fine-tuning: Improves OOD, Maintains ID

- Llama 3.1 8B SFT on Alpaca dataset with perturbed token embeddings.

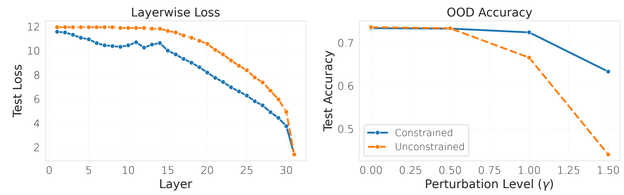


Figure 4. Constraints improve token accuracy (Left: ↓, Right: ↑).

- **AlpacaEval (constrained vs unconstrained):** $\gamma = 0$: 50% win rate → preserves ID. $\gamma = 1.5$: 69.9% win rate → robust OOD.

## Unrolled video denoisers generalize to OOD noise

- **Task:** Reconstruct video $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$ from noisy $\tilde{\mathbf{X}}$,

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

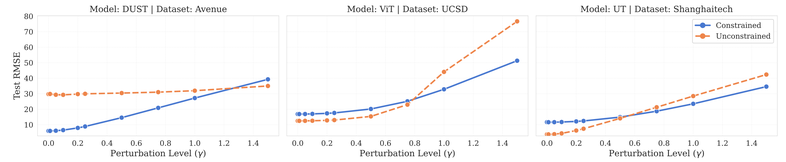


Figure 5. Constraints improve robustness. RMSE vs. test perturbation $\gamma$ (↓ lower is better).

## Ablations: Step size, dual LR, training perturbations

- Low sensitivity to dual LR $\eta_2$; OOD accuracy increases with step size $\alpha$.

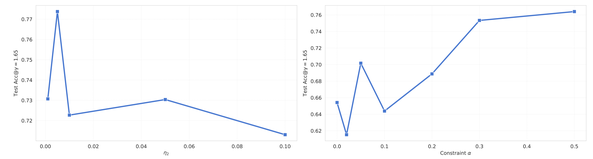


Figure 6. Dual learning rate $\eta_2$ (left) and step size $\alpha$ (right) ablations.

- Constraints compose with standard robustness techniques (noise injection).

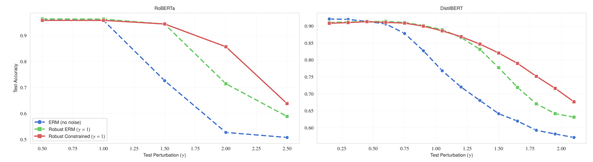


Figure 7. Comparison of baseline, robust ERM ($\gamma = 1$), and robust + constraints.

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