Δ-STN: Efficient Bilevel Optimization for Neural Networks using Structured Response Jacobians (Neurips 2020)

Juhan Bae, Roger Grosse

University of Toronto, Vector Institute
{*jbae, rgrosse*}*@cs.toronto.edu*

Introduction: Hyperparameter Optimization

Hyperparameter is a bilevel optimization problem.

$$\min_{\boldsymbol{\lambda}} \mathcal{L}_V(\boldsymbol{\lambda}, \mathbf{w}^*) \text{ subject to } \mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}_T(\boldsymbol{\lambda}, \mathbf{w}),$$

- $\mathcal{L}_V, \mathcal{L}_T$,: Validation, Training objectives
- $\boldsymbol{\lambda}, \mathbf{w}$: Hyperparmaeters, Parameters

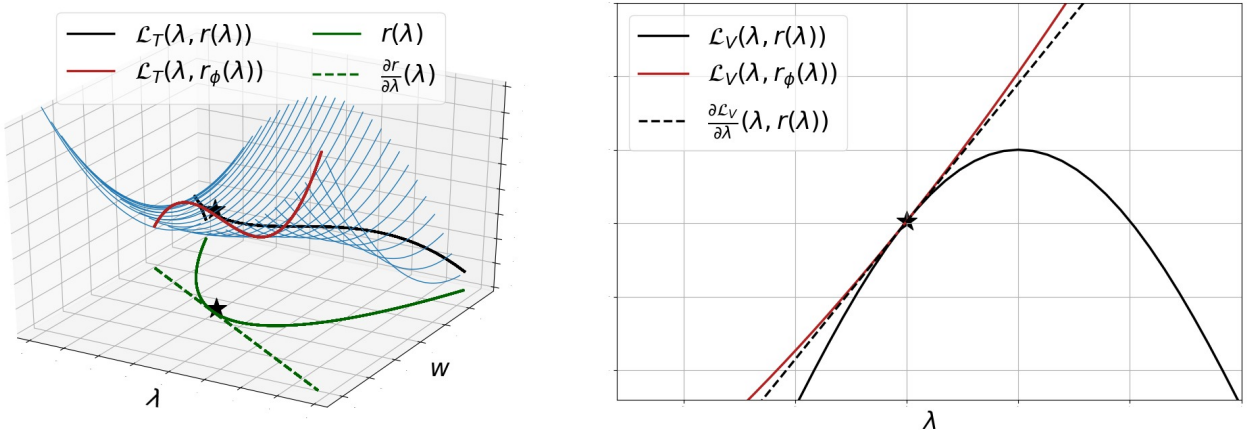
Goal:

Find the optimal hyperparameters $\boldsymbol{\lambda}^*$ that minimizes the validation objective after training.

Introduction: Parametric Best-Response Function

Best-response function is defined as:

$$\mathbf{r}(\boldsymbol{\lambda}) = \arg\min_{\mathbf{w}} \mathcal{L}_T(\boldsymbol{\lambda}, \mathbf{w})$$



Idea: Learn a parametric best-response function with hypernetwork and jointly optimize hyperparameters and parameters.

$$\min_{\phi} \mathbb{E}_{\epsilon \sim p(\epsilon|\sigma)} \left[\mathcal{L}_T(\boldsymbol{\lambda} + \epsilon, \mathbf{r}_\phi(\boldsymbol{\lambda} + \epsilon)) \right], \quad \min_{\boldsymbol{\lambda}} \mathcal{L}_V(\boldsymbol{\lambda}, \mathbf{r}_\phi(\boldsymbol{\lambda}))$$

Δ-STNs

Self-Tuning Networks (STNs) parameterize the best-response functions (with further structure imposed) as:

$$\mathbf{r}_{\phi}(\boldsymbol{\lambda}) = \boldsymbol{\Phi}\boldsymbol{\lambda} + \boldsymbol{\phi}_0,$$

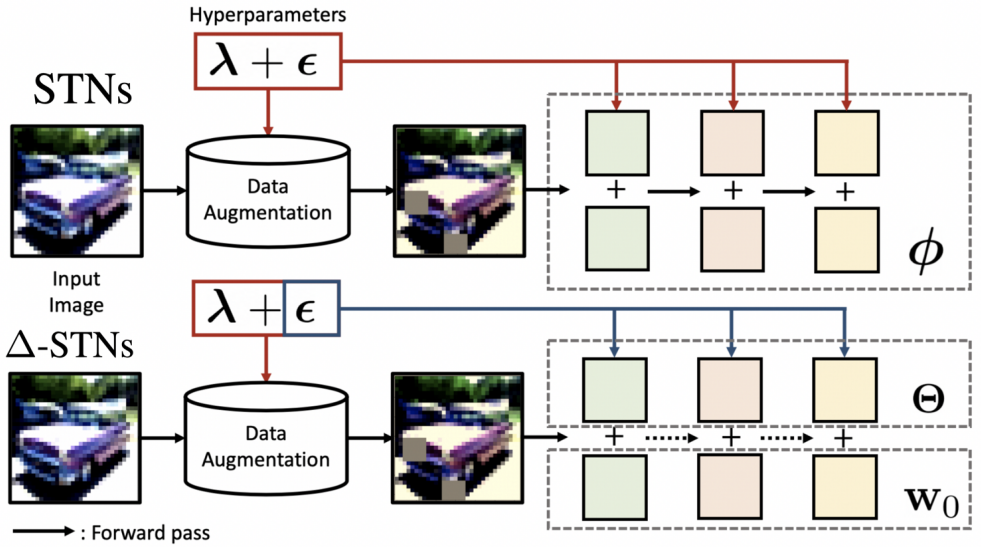
Proposed:

Δ-STNs that fix subtle pathologies in training STNs. It can optimize hyperparameters with:

- Higher accuracy
- Faster convergence
- Improved stability

Δ-STNs: Centered Parameterization

STNs	Δ-STNs
$\mathbf{r}_{\phi}(\boldsymbol{\lambda}) = \boldsymbol{\Phi}\boldsymbol{\lambda} + \boldsymbol{\phi}_0$	$\mathbf{r}_{\boldsymbol{\theta}}(\boldsymbol{\lambda}, \boldsymbol{\lambda}_0) = \boldsymbol{\Theta}(\boldsymbol{\lambda} - \boldsymbol{\lambda}_0) + \mathbf{w}_0$



Advantages

- Improves the conditioning of a lower-level problem
- Fixes undesirable bilevel dynamics

Δ-STNs: Modified Update Rule

Proposed parameterization:

$$\mathbf{r}_{\boldsymbol{\theta}}(\boldsymbol{\lambda}, \boldsymbol{\lambda}_0) = \boldsymbol{\Theta}(\boldsymbol{\lambda} - \boldsymbol{\lambda}_0) + \mathbf{w}_0$$

- $\boldsymbol{\lambda}_0$: "current hyperparameters"
- $\mathbf{w}_0$: "current weights"
- $\boldsymbol{\Theta}$: how the weights are adjusted in response to a perturbation to $\boldsymbol{\lambda}$

Modify the training objectives by separating:

$$\arg\min_{\mathbf{w}_0} \mathcal{L}_T(\boldsymbol{\lambda}, \mathbf{w}_0)$$

$$\arg\min_{\boldsymbol{\Theta}} \mathbb{E}_{\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}|\boldsymbol{\sigma})} \left[\mathcal{L}_T(\boldsymbol{\lambda} + \boldsymbol{\epsilon}, \mathbf{r}_{\boldsymbol{\theta}}(\boldsymbol{\lambda} + \boldsymbol{\epsilon}, \boldsymbol{\lambda}))\right]$$

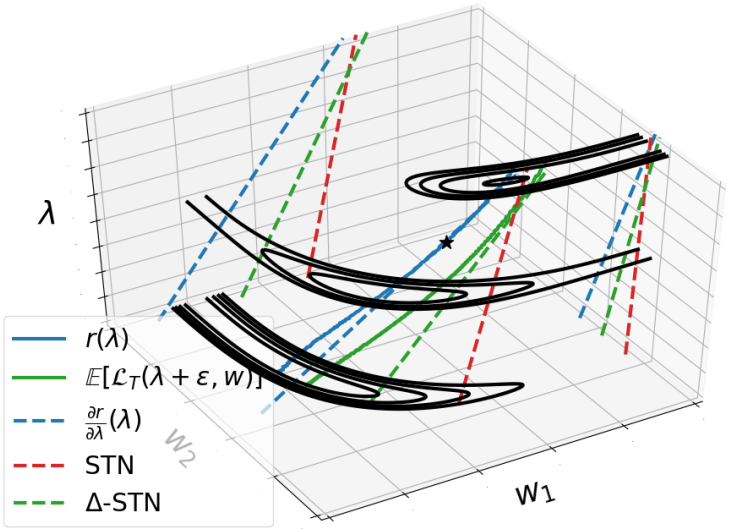
Δ-STNs: Direct Approximation of Best-Response Jacobian

Problem:

If the perturbation is large, it is difficult to approximate best-response function as linear within the region.

Fix:

Linearize the network around the current parameters $r(\boldsymbol{\lambda}_0) = \mathbf{w}_0$. Directly approximates the best-response Jacobian instead of learning the full best-response function.



Experiments

Linear Regression: UCI Regression dataset

Tune L_2 regularization penalty & input dropout rate.

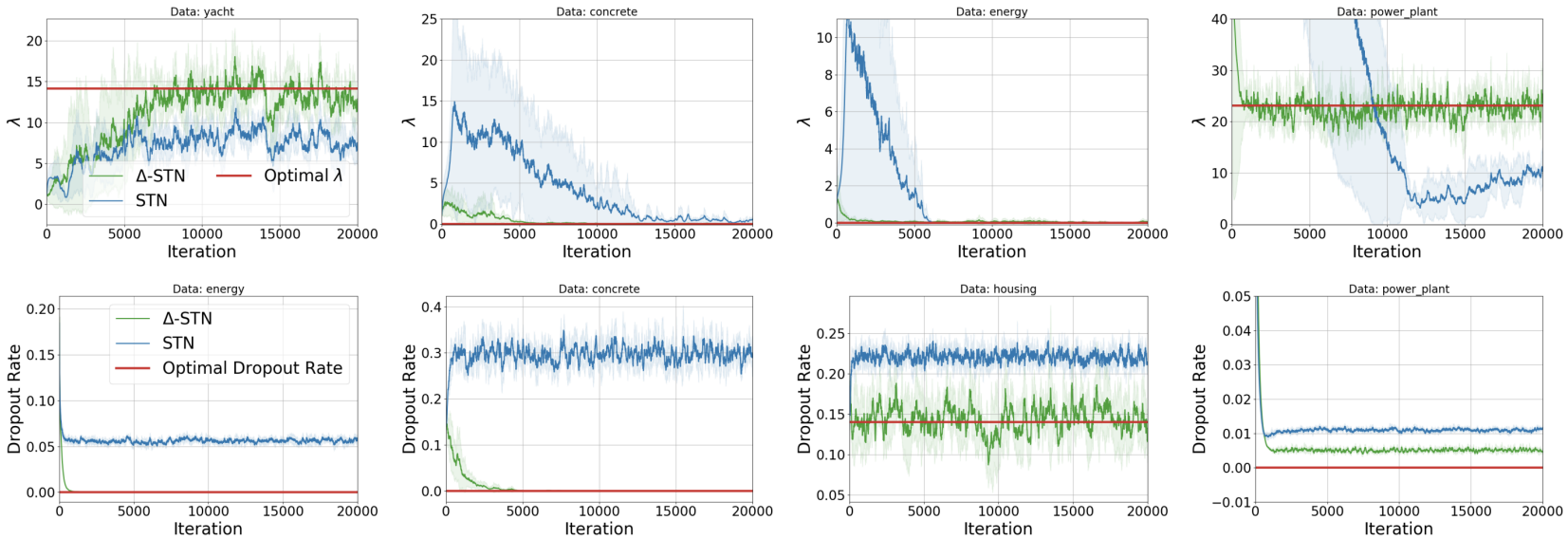
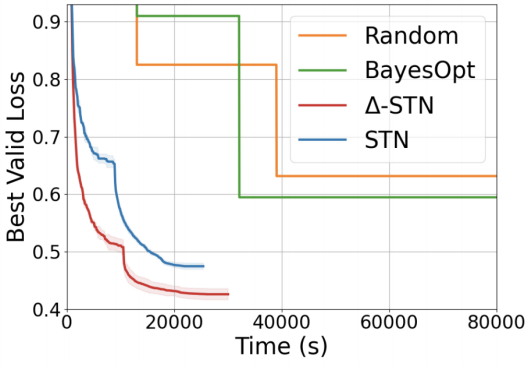
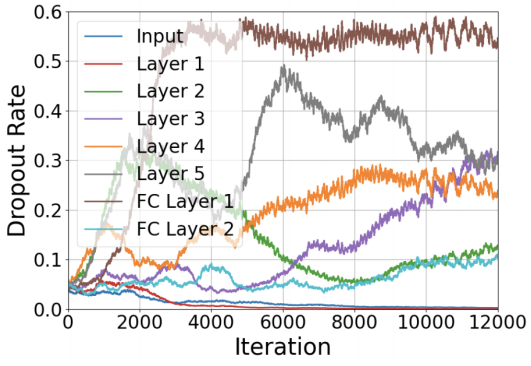


Image Classification & Language Modelling

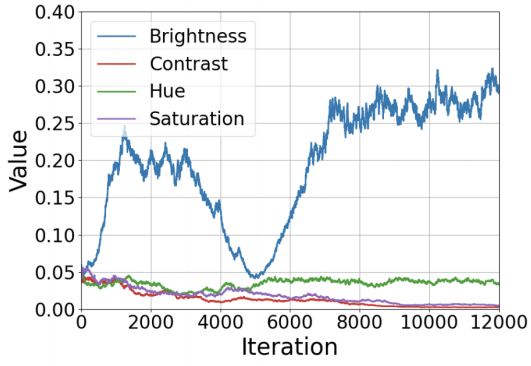
Tune per-layer dropouts & data augmentation hyperparameters.



(a) Time comparison



(b) Schedule for dropouts

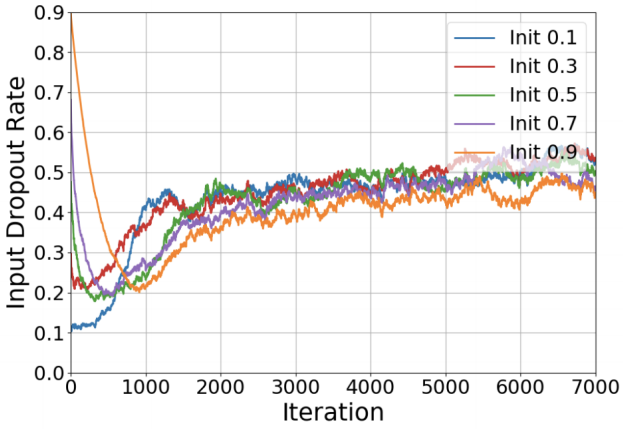


(c) Schedule for data augmentation

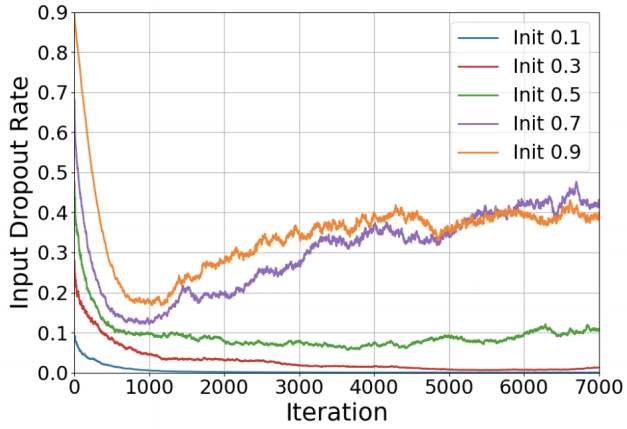
Dataset	Network	RS	BO	STN	Centered STN	Δ-STN
MNIST	MLP	0.043 (0.042)	0.042 (0.043)	0.043 (0.041)	0.041 (0.039)	**0.040** (**0.038**)
FMNIST	SimpleCNN	0.206 (0.214)	0.217 (0.215)	0.196 (0.218)	0.191 (0.212)	**0.189** (**0.209**)
CIFAR10	AlexNet	0.631 (0.671)	0.594 (0.598)	0.474 (0.488)	0.431 (0.450)	**0.425** (**0.446**)
	VGG16	0.566 (0.595)	0.421 (0.446)	0.330 (0.354)	0.286 (0.321)	**0.272** (**0.296**)
	ResNet18	0.264 (0.298)	0.230 (0.267)	0.266 (0.312)	0.222 (0.258)	**0.204** (**0.238**)
PTB	LSTM	84.81 (81.46)	72.13 (69.29)	70.67 (67.78)	69.40 (66.67)	**68.63** (**66.26**)

Sensitivity Analysis

Δ-STNs are more robust to hyperparameter initialization & perturbation scale.



(a) Schedule found by Δ-STNs



(b) Schedule found by STNs