SELF-TUNING NETWORKS: BILEVEL OPTIMIZATION OF HYPERPARAMETERS USING STRUCTURED BEST-RESPONSE FUNCTIONS

Anonymous authors

Paper under double-blind review

ABSTRACT

Hyperparameter optimization is a bi-level optimization problem, where the optimal parameters on the training set depend on the current hyperparameters. The *best-response function* which maps hyperparameters to these optimal parameters allows gradient-based hyperparameter optimization but is difficult to represent and compute when the parameters are high dimensional, as in neural networks. We develop efficient best-response approximations for neural networks by applying insights from the structure of the optimal response in a Jacobian-regularized two-layer linear network to deep, nonlinear networks. The approximation works by scaling and shifting the hidden units by amounts which depend on the current hyperparameters. We use our approximation for a gradient-based hyperparameter optimization algorithm which alternates between approximating the best-response in a neighborhood around the current hyperparameters and optimizing the hyperparameters using the approximate best-response. We show this method outperforms competing hyperparameter optimization methods on large-scale deep learning problems. We call our networks, which update their own hyperparameters online during training, *Self-Tuning Networks*.

1 INTRODUCTION

Regularization hyperparameters such as weight decay, data augmentation, and dropout (Srivastava et al., 2014) are crucial to the generalization of neural networks (NNs), but are difficult to tune. Popular approaches to hyperparameter optimization (HO) include grid search, random search (Bergstra & Bengio, 2012), and Bayesian optimization (Snoek et al., 2012). These approaches work well with low-dimensional hyperparameters and ample computational resources. However, they pose HO as a black-box optimization problem, ignoring structure which could be exploited for faster convergence.

We can formulate HO as a bilevel optimization problem. Let $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ denote parameters and hyperparameters, and let $\mathcal{L}_T$ and $\mathcal{L}_V$ be functions mapping parameters and hyperparameters to the loss on the training and validation sets. We aim to solve[1] :

$$\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}} \mathcal{L}_V(\boldsymbol{\lambda}, \boldsymbol{\theta}) \quad \text{subject to } \boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_T(\boldsymbol{\lambda}, \boldsymbol{\theta}) \tag{1}$$

Substituting the *best-response function* $\boldsymbol{\theta}^*(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_T(\boldsymbol{\lambda}, \boldsymbol{\theta})$ gives a single-level problem:

$$\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}} \mathcal{L}_V(\boldsymbol{\lambda}, \boldsymbol{\theta}^*(\boldsymbol{\lambda})) \tag{2}$$

Gradient-based HO methods attempt to solve Problem 1 through approximation of the best-response or its Jacobian. It is possible to do so by differentiating through the training procedure (Maclaurin et al., 2015) or by implicitly deriving the best-response Jacobian (Larsen et al., 1996), but these methods struggle with discrete hyperparameters, and are computationally expensive. Recently, Lorraine & Duvenaud (2018) proposed directly approximating the best-response by a parametric function $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}$. They jointly optimized $\boldsymbol{\phi}$ and $\boldsymbol{\lambda}$, optimizing $\boldsymbol{\phi}$ so $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}} \approx \boldsymbol{\theta}^*$ in a neighborhood around the

[1] Uniqueness of the $\arg\min$ is assumed.

current hyperparameters, and optimizing $\boldsymbol{\lambda}$ by using $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}$ as a proxy for $\boldsymbol{\theta}^*$ in Eq. 2:

$$\boldsymbol{\lambda}^* \approx \arg\min_{\boldsymbol{\lambda}} \mathcal{L}_V(\boldsymbol{\lambda}, \hat{\boldsymbol{\theta}}_{\phi}(\boldsymbol{\lambda})) \tag{3}$$

This worked well for small-scale problems on MNIST using L_2 regularization, but it is difficult to represent and compute $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}$ when $\boldsymbol{\theta}$ is high-dimensional.

In this paper, we construct scalable and compact best-response approximations for NNs. We analyze the best-response structure in a Jacobian-regularized two-layer linear network, and show it can be modelled as scaling network's hidden units by amounts which depend on the current hyperparameters. Exploiting this structure, we extend the approach to deep, nonlinear networks, giving a memory efficient best-response approximation which scales to modern architectures.

Using this approximation for HO gives a gradient-based HO algorithm that is (1) computationally inexpensive, (2) can optimize all regularization hyperparameters, including discrete hyperparameters, and (3) scales to large NNs. We call the resulting networks *Self-Tuning Networks* (STNs), since they update their own hyperparameters online during training. We empirically evaluate the performance of STNs for HO on large-scale deep-learning problems on the PTB and CIFAR-10 datasets. We compare their performance to other popular HO approaches and find that STNs achieve lower validation loss in less time.

2 Bilevel Optimization

A bilevel optimization problem consists of two sub-problems called the *upper-level* and *lower-level* problems, where the upper level problem must be solved subject to optimality of the lower-level problem. Minimax problems are a subclass of bilevel programs. Bilevel problems were first studied in economics to model leader/follower firm dynamics (Von Stackelberg, 2010) and have since found uses in many different fields (see Colson et al. (2007) for an overview). In machine learning, many problems can be formulated as bilevel programs, including HO, GAN training (Goodfellow et al., 2014), meta-learning, and neural architecture search (Zoph & Le, 2016).

Solving bilevel problems exactly is difficult. Even if all objectives and constraints are linear, bilevel problems are strongly NP-hard (Hansen et al., 1992; Vicente et al., 1994). Due to the difficulty of obtaining exact solutions, most work has focused on restricted settings, considering linear, quadratic, and convex functions. In contrast, we focus on obtaining local solutions in the nonconvex, differentiable, and unconstrained setting. We let $F, f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ denote the upper- and lower-level objectives and $\boldsymbol{\lambda} \in \mathbb{R}^n, \boldsymbol{\theta} \in \mathbb{R}^m$ denote the upper- and lower-level variables. We aim to solve:

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}^n} F(\boldsymbol{\lambda}, \boldsymbol{\theta}) \tag{4a}$$

$$\text{subject to} \quad \boldsymbol{\theta} \in \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^m} f(\boldsymbol{\lambda}, \boldsymbol{\theta}) \tag{4b}$$

2.1 Gradient Descent via the Best-Response Function

Using gradient information provides drastic speed-ups over black-box optimization methods (Nesterov, 2013). Thus, it is desirable to design a gradient-based algorithm for solving Problem 4. The simplest method is simultaneous gradient descent, updating $\boldsymbol{\lambda}$ using $\frac{\partial F}{\partial \boldsymbol{\lambda}}$ and $\boldsymbol{\theta}$ using $\frac{\partial f}{\partial \boldsymbol{\theta}}$. However, simultaneous gradient descent fails if $\boldsymbol{\lambda}$ does not affect the upper-level objective explicitly, i.e. if $\frac{\partial F}{\partial \boldsymbol{\lambda}} \equiv \mathbf{0}$. In this case, one solution is to use the *best-response function* (Gibbons, 1992). Assume the lower-level Problem 4b has a unique optimum $\boldsymbol{\theta}^*(\boldsymbol{\lambda})$ for each $\boldsymbol{\lambda}$. Substituting $\boldsymbol{\theta}^*$, the best-response function, converts Problem 4 into a single-level problem:

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}^n} F^*(\boldsymbol{\lambda}) := F(\boldsymbol{\lambda}, \boldsymbol{\theta}^*(\boldsymbol{\lambda})) \tag{5}$$

If $\boldsymbol{\theta}^*$ is differentiable, we can minimize Eq. 5 using gradient descent on F^* w.r.t. $\boldsymbol{\lambda}$. This method requires a unique optimum for Problem 4b for each $\boldsymbol{\lambda}$ and differentiability of $\boldsymbol{\theta}^*$. In general, this is difficult to verify. We give sufficient conditions, which almost certainly do not hold in practice, for this to be the case in a neighborhood of a point $(\boldsymbol{\lambda}_0, \boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0$ solves Problem 4b given $\boldsymbol{\lambda}_0$.

Lemma 1. *(Fiacco & Ishizuka, 1990) Let $\boldsymbol{\theta}_0$ solve Problem equation 4b for $\boldsymbol{\lambda}_0$. Suppose f is $\mathcal{C}^2$ in a neighborhood of $(\boldsymbol{\lambda}_0, \boldsymbol{\theta}_0)$ and the Hessian $\frac{\partial^2 f}{\partial \boldsymbol{\theta}^2}(\boldsymbol{\lambda}_0, \boldsymbol{\theta}_0)$ is positive definite. Then for some neighborhood U of $\boldsymbol{\lambda}_0$, there exists a continuously differentiable function $\boldsymbol{\theta}^* : U \to \mathbb{R}^m$ such that $\boldsymbol{\theta}^*(\boldsymbol{\lambda})$ is the unique solution to Problem equation 4b given $\boldsymbol{\lambda}$ for all $\boldsymbol{\lambda} \in U$ and $\boldsymbol{\theta}^*(\boldsymbol{\lambda}_0) = \boldsymbol{\theta}_0$.*

Proof. See Appendix A.1. □

2.2 Approximating the Best-Response Function

Although the solution to Problem 4b is, in general, a set, assuming uniqueness of a solution and differentiability of $\boldsymbol{\theta}^*$, as is often done in the HO literature, can yield fruitful algorithms in practice. Gradient-based HO methods can be interpreted as approximating either the best-response $\boldsymbol{\theta}^*$ or its Jacobian $\frac{\partial \boldsymbol{\theta}^*}{\partial \boldsymbol{\lambda}}$. The most popular approaches, described in Section 5, either approximate $\boldsymbol{\theta}^*$ by differentiating gradient descent or implicitly derive $\frac{\partial \boldsymbol{\theta}^*}{\partial \boldsymbol{\lambda}}$. However, these approaches can be computationally expensive. When applied to HO, they struggle with discrete hyperparameters and some continuous hyperparameters like dropout probabilities. Alternative approaches which directly approximate $\boldsymbol{\theta}^*$ were proposed by Lorraine & Duvenaud (2018). We detail their algorithms below.

1. Global Approximation. The first algorithm proposed by Lorraine & Duvenaud (2018) approximated $\boldsymbol{\theta}^*$ as a differentiable function $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}$ with parameters $\boldsymbol{\phi}$. If $\boldsymbol{\theta}$ represents NN weights, then the mapping $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}$ is a hypernetwork (Lorraine & Duvenaud, 2018; Schmidhuber, 1992; Ha et al., 2016). A distribution $p(\boldsymbol{\lambda})$ is fixed then gradient descent w.r.t. $\boldsymbol{\phi}$ minimizes:

$$\mathbb{E}_{\boldsymbol{\lambda} \sim p(\boldsymbol{\lambda})} \left[f(\boldsymbol{\lambda}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}(\boldsymbol{\lambda})) \right] \tag{6}$$

If support(p) is broad and $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}$ is sufficiently flexible, then Eq. 6 is minimized exactly and $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}$ is used as a proxy for $\boldsymbol{\theta}^*$ in Problem 5, resulting in the following objective $\hat{F}$:

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}^n} \hat{F}(\boldsymbol{\lambda}, \boldsymbol{\phi}) := F(\boldsymbol{\lambda}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}(\boldsymbol{\lambda})) \tag{7}$$

2. Local Approximation. In practice, $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}$ is usually insufficiently flexible to model $\boldsymbol{\theta}^*$ on support(p). The second algorithm of Lorraine & Duvenaud (2018) locally approximated $\boldsymbol{\theta}^*$ in a neighborhood around the current upper-level parameter $\boldsymbol{\lambda}$. A noise distribution $p(\boldsymbol{\epsilon}|\boldsymbol{\sigma})$ parametrized by a scale parameter $\boldsymbol{\sigma} \in \mathbb{R}^n_+$ is fixed, then $\boldsymbol{\phi}$ is found by minimizing the objective:

$$\hat{f}(\boldsymbol{\lambda}, \boldsymbol{\phi}, \boldsymbol{\sigma}) = \mathbb{E}_{\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}|\boldsymbol{\sigma})} \left[f(\boldsymbol{\lambda} + \boldsymbol{\epsilon}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}(\boldsymbol{\lambda} + \boldsymbol{\epsilon})) \right] \tag{8}$$

Intuitively, the upper-level parameter $\boldsymbol{\lambda}$ is perturbed by a small amount, so the lower-level parameter learns how to respond. An alternating gradient descent scheme is then used, where $\boldsymbol{\phi}$ is updated to minimize equation 8 and $\boldsymbol{\lambda}$ is updated to minimize equation 7. This approach worked well for problems using L_2 weight decay on MNIST (LeCun et al., 1998). However, it is unclear if the approach works with different regularizers and whether it scales to larger problems. It requires $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}$, which is a priori unwieldy for high dimensional $\boldsymbol{\theta}$. It is also unclear how to set the scale $\boldsymbol{\sigma}$, which defines the size of the neighborhood on which $\boldsymbol{\phi}$ is trained, and whether the approach can be adapted to discrete hyperparameters.

3 Self-Tuning Networks

We first construct a compact best-response approximation $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}$ that is memory efficient and scales to large NNs. We then describe a method to automatically adjust the size of the neighborhood on which $\boldsymbol{\phi}$ is trained during training. We formally describe our algorithm and discuss how it easily handles discrete hyperparameters. We call the resulting networks, which update their own hyperparameters online during training, *self-tuning networks* (STNs).

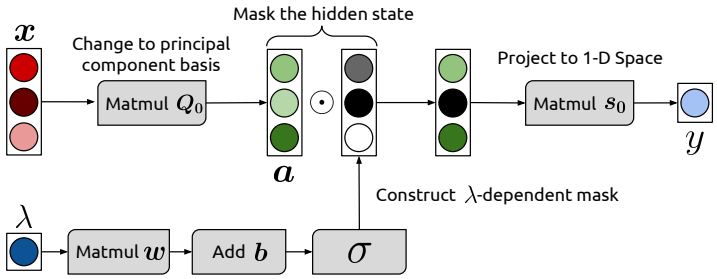


Figure 1: Best-response architecture for an L_2-Jacobian regularized two-layer linear network.

3.1 Compact Best-Response Approximations for Neural Networks

When $\boldsymbol{\theta}$ represents the weights of a NN, $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}$ must be flexible enough to model how the weights should respond, while remaining memory-efficient and tractable for larger networks. We begin by investigating the structure of the best-response in a simpler setting to glean insights which can be applied to more complicated problems. Consider a regression problem where a 2-layer linear network with weights $\boldsymbol{\theta} = (\boldsymbol{Q}, \boldsymbol{s}) \in \mathbb{R}^{D\times D} \times \mathbb{R}^D$ predicts targets t from inputs $\boldsymbol{x} \in \mathbb{R}^D$:

$$\boldsymbol{a}(\boldsymbol{x};\boldsymbol{\theta}) = \boldsymbol{Q}\boldsymbol{x}, \quad y(\boldsymbol{x};\boldsymbol{\theta}) = \langle \boldsymbol{s}, \boldsymbol{a}(\boldsymbol{x};\boldsymbol{\theta})\rangle \tag{9}$$

Suppose we use a squared-error loss regularized with an L_2 penalty on the Jacobian $\frac{\partial y}{\partial \boldsymbol{x}}$, where the penalty weight hyperparameter is reparametrized using $\exp$ to lie in $\mathbb{R}$ rather than $\mathbb{R}_+$:

$$\mathcal{L}_T(\lambda, \boldsymbol{\theta}) = \sum_{(\boldsymbol{x},t)\in\mathcal{D}} (y(\boldsymbol{x};\boldsymbol{\theta}) - t)^2 + \frac{1}{|\mathcal{D}|}\exp(\lambda)\left\|\frac{\partial y}{\partial \boldsymbol{x}}(\boldsymbol{x};\boldsymbol{\theta})\right\|^2 \tag{10}$$

Lemma 2. *Let $\boldsymbol{\theta}_0 = (\boldsymbol{Q}_0, \boldsymbol{s}_0)$, where $\boldsymbol{Q}_0$ is the change-of-basis matrix to the principal components of the data matrix and $\boldsymbol{s}_0$ solves the unregularized version of Problem 10 given $\boldsymbol{Q}_0$. Then there exist $\boldsymbol{w}, \boldsymbol{b} \in \mathbb{R}^D$ such that the best-response function[2] $\boldsymbol{\theta}^*(\lambda) = (\boldsymbol{Q}^*(\lambda), \boldsymbol{s}^*(\lambda))$ for Problem 10 is:*

$$\boldsymbol{Q}^*(\lambda) = \sigma(\lambda\boldsymbol{w} + \boldsymbol{b}) \odot \boldsymbol{Q}_0, \quad \boldsymbol{s}^*(\lambda) = \boldsymbol{s}_0 \tag{11}$$

Here, $\odot$ indicates row-wise rescaling.

Proof. See Appendix A.2. □

$y(\boldsymbol{x};\boldsymbol{\theta}^*(\lambda))$ implements $y(\boldsymbol{x};\boldsymbol{\theta}_0)$ with an additional sigmoidal gating of the hidden units:

$$\boldsymbol{a}(\boldsymbol{x};\boldsymbol{\theta}^*(\lambda)) = \boldsymbol{Q}^*(\lambda)\boldsymbol{x} = \sigma(\lambda\boldsymbol{w} + \boldsymbol{b}) \odot (\boldsymbol{Q}_0\boldsymbol{x}) = \sigma(\lambda\boldsymbol{w} + \boldsymbol{b}) \odot \boldsymbol{a}(\boldsymbol{x};\boldsymbol{\theta}_0) \tag{12}$$

We show this architecture as Figure 1. Inspired by this example, we use a linearized version as a best-response architecture for deep nonlinear networks. Our architecture scales and shifts the rows of the weights and biases (equivalently, the hidden units of the network) by amounts which depend on the hyperparameters. For a layer's weight matrix $\boldsymbol{W}$ and a bias $\boldsymbol{b}$, we approximate the best-response as[3]:

$$\boldsymbol{W}^*(\boldsymbol{\lambda}) = \boldsymbol{W}_{\text{shift}} + \langle \boldsymbol{\lambda}, \boldsymbol{u}\rangle \odot \boldsymbol{W}_{\text{scale}}, \quad \boldsymbol{b}^*(\boldsymbol{\lambda}) = \boldsymbol{b}_{\text{shift}} + \langle \boldsymbol{\lambda}, \boldsymbol{v}\rangle \odot \boldsymbol{b}_{\text{scale}}, \tag{13}$$

This best-response architecture is tractable to compute and memory-efficient: if a normal layer uses p weights, it uses $2p + n$ weights. There is no guarantee this architecture can represent the best-response function in general, but intuitively it makes sense that some hidden units should be (de-)emphasized for stronger regularization. NN generalization is strongly related to the network's Jacobian norm (Novak et al., 2018), suggesting that the structure of the best-response from Problem 10 may carry over to other types of regularization. Louizos & Welling (2017) showed probability distributions induced by noisily scaling activations could be complex and multimodal.

This architecture allows the use of different perturbed hyperparameters for each minibatch element since the activations of each batch element can be scaled and shifted separately, enabling $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}$ to quickly learn how to respond in a neighborhood around the current hyperparameter values.

[2] This is an abuse of notation since there is not a unique solution to Problem 10 for a given λ.

[3] We describe modifications for convolutional filters in Appendix B.

3.2 ADAPTING THE HYPERPARAMETER DISTRIBUTION

The entries of $\boldsymbol{\sigma}$, which define the size of the neighborhood $\boldsymbol{\phi}$ is trained on, are crucial to the STN's performance. If the entries are too large, then $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}$ will not be flexible enough to capture the best-response over the sampled neighborhood. However, its entries must remain sufficiently large so that $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}$ captures the local shape around the current hyperparameter values. As the smoothness of the loss landscape changes during training, it may be beneficial to enlarge or shrink the entries of $\boldsymbol{\sigma}$.

To address these issues we propose adjusting $\boldsymbol{\sigma}$ during training based on the sensitivity of the upper-level objective to sampled hyperparameters. We include an entropy term weighted by $\tau \in [0, 1]$ which enlarges entries of $\boldsymbol{\sigma}$. The resulting objective is:

$$\hat{F}_V(\boldsymbol{\lambda}, \boldsymbol{\phi}, \boldsymbol{\sigma}) = \mathbb{E}_{\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}|\boldsymbol{\sigma})}[F(\boldsymbol{\lambda} + \boldsymbol{\epsilon}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}(\boldsymbol{\lambda} + \boldsymbol{\epsilon}))] - \tau \mathbb{H}[p(\boldsymbol{\epsilon}|\boldsymbol{\sigma})] \tag{14}$$

The resulting objective is similar to a variational inference (VI) objective, where the first term is analogous to the negative-log likelihood, but $\tau \neq 1$. As τ ranges from 0 to 1, our objective interpolates between variational optimization (Staines & Barber, 2012) and VI, as noted by Khan et al. (2018). Similar objectives have been used in the VI literature for better training (Blundell et al., 2015) and representation learning (Higgins et al., 2016).

On its own, minimizing the first term eventually moves all probability mass towards an optimum $\boldsymbol{\lambda}^*$, resulting in $\boldsymbol{\sigma} = \mathbf{0}$. This compels $\boldsymbol{\sigma}$ to achieve a balance between shifting towards $\mathbf{0}$ to achieve better performance on the first term while remaining sufficiently high to avoid a heavy entropy penalty. τ must be set carefully to ensure the $\boldsymbol{\sigma}$ is not pushed too high. When benchmarking our algorithm's performance, we evaluate $\hat{F}$ at the current hyperparameter $\boldsymbol{\lambda}_0$ and do not evaluate over a range as required by $\hat{F}_V$. This is common practice when using stochastic operations during training, e.g., batch-norm or dropout.

3.3 TRAINING ALGORITHM

We now describe the complete STN training algorithm and discuss how it can tune hyperparameters which other gradient-based algorithms cannot, such as discrete hyperparameters. We use an unconstrained parametrization $\boldsymbol{\lambda} \in \mathbb{R}^n$ of the hyperparameters. We let r denote the element-wise function which maps $\boldsymbol{\lambda}$ to the appropriate constrained space. These element-wise transformations will involve a non-differentiable discretization for discrete hyperparameters.

We let $\mathcal{L}_T$ and $\mathcal{L}_V$ denote training and validation functions which are (possibly stochastic, e.g., if using dropout) functions of the hyperparameters and model parameters, and define functions f, F by $f(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \mathcal{L}_T(r(\boldsymbol{\lambda}), \boldsymbol{\theta})$ and $F(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \mathcal{L}_V(r(\boldsymbol{\lambda}), \boldsymbol{\theta})$. STNs are trained by a gradient descent scheme which alternates between updating $\boldsymbol{\phi}$ for T_{train} steps to minimize $\hat{f}(\boldsymbol{\lambda}, \boldsymbol{\phi}, \boldsymbol{\sigma})$ and updating $\boldsymbol{\lambda}$ and $\boldsymbol{\sigma}$ for T_{valid} steps to minimize $\hat{F}_V(\boldsymbol{\lambda}, \boldsymbol{\phi}, \boldsymbol{\sigma})$. We give our complete algorithm as Algorithm 1.

The possible non-differentiability of r due to discrete hyperparameters poses no problem. To differentiate $\hat{f}$ w.r.t. $\boldsymbol{\phi}$, we only require $\frac{\partial f}{\partial \boldsymbol{\theta}}$ and $\frac{\partial \hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}}{\partial \boldsymbol{\phi}}$. To differentiate $\hat{F}_V$ w.r.t. λ_i corresponding to a discrete hyperparameter, there are two cases we must consider:

Case 1: For most regularization schemes, $\mathcal{L}_V$ does not depend on λ_i directly and hence the only gradient is through $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}$. Thus, the reparametrization gradient can be used when differentiating $\hat{F}_V$.

Case 2: If $\mathcal{L}_V$ relies explicitly on λ_i, then we can use the REINFORCE gradient estimator (Williams, 1992) to estimate $\frac{\partial \hat{F}_V}{\partial \lambda_i}$. An example hyperparameter requiring this approach is the number of hidden units, which directly affects the validation loss.

3.4 SUFFICIENCY OF LINEAR BEST-RESPONSE APPROXIMATIONS

To assess how flexible the approximation must be to model the best-response, we provide theoretical results for the behavior of linear approximations in the quadratic setting. We show that minimizing equation 8 yields the correct optimal-response gradient, hence linear approximations are sufficient in this restricted setting.

Algorithm 1 STNN Training Algorithm

Initialize: Model parameters $\boldsymbol{\theta}$, hyperparameters $\boldsymbol{\lambda}$
while not converged **do**
  for $t = 1, \ldots, T_{train}$ **do**
    $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}|\boldsymbol{\sigma})$
    $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} - \alpha \frac{\partial}{\partial \boldsymbol{\phi}} \left[f(\boldsymbol{\lambda} + \boldsymbol{\epsilon}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}(\boldsymbol{\lambda} + \boldsymbol{\epsilon})) \right]$
  for $t = 1, \ldots, T_{valid}$ **do**
    $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}|\boldsymbol{\sigma})$
    $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} - \beta \frac{\partial}{\partial \boldsymbol{\lambda}} \left[F(\boldsymbol{\lambda} + \boldsymbol{\epsilon}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}(\boldsymbol{\lambda} + \boldsymbol{\epsilon})) - \tau \mathbb{H}[p(\boldsymbol{\epsilon}|\boldsymbol{\sigma})] \right]$
    $\boldsymbol{\sigma} \leftarrow \boldsymbol{\sigma} - \gamma \frac{\partial}{\partial \boldsymbol{\sigma}} \left[F(\boldsymbol{\lambda} + \boldsymbol{\epsilon}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}(\boldsymbol{\lambda} + \boldsymbol{\epsilon})) - \tau \mathbb{H}[p(\boldsymbol{\epsilon}|\boldsymbol{\sigma})] \right]$

Theorem 3. *Suppose f is quadratic with $\frac{\partial^2 f}{\partial \boldsymbol{\theta}^2} \succ \mathbf{0}$ and $p(\boldsymbol{\epsilon}|\sigma)$ is Gaussian with mean $\mathbf{0}$ and variance $\sigma^2 \boldsymbol{I}$. Suppose $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}$ is linear and $\boldsymbol{\phi}^* = \arg\min_{\boldsymbol{\phi}} \hat{f}(\boldsymbol{\lambda}, \boldsymbol{\phi}, \sigma)$. Then we have $\frac{\partial \hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}_0) = \frac{\partial \boldsymbol{\theta}^*}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda})$.*

Proof. See Appendix A.3. □

4 Experiments

We empirically investigated the behavior of STNs when tuning a single hyperparameter, and found STNs greedily adapt hyperparameters and discover *hyperparameter schedules* which outperform any single, fixed hyperparameter. We provide experiments showing that better generalization is a consequence of the schedule itself, and not due to increased regularization from stochasticity of sampling hyperparameters or limited capacity of the approximation $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}$. Next, we compared the performance of STNs to commonly used HO methods on the CIFAR-10 and PTB datasets. We show that STNs reach a lower validation loss more quickly than other popular approaches. We do not compare with other gradient-based approaches since they cannot tune dropout probabilities.

4.1 Hyperparameter Schedules

By tuning just a single hyperparameter, we found that STNs do not converge to a single fixed hyperparameter, but rather discover schedules for adapting the hyperparameters online. We examined this in detail on the PTB corpus (Marcus et al., 1993) using a Self-Tuning LSTM (Hochreiter & Schmidhuber, 1997) (ST-LSTM) and tuning the output dropout rate applied to the hidden units.

First, we found the schedule discovered by an ST-LSTM for output dropout, shown in Figure 2, outperforms the best, fixed output dropout rate (0.68) found by a fine-grained grid search, achieving 83.25 vs 85.33 validation perplexity. We argue this is a consequence of the schedule itself, and not due to regularizing effects from sampling hyperparameters or the limited capacity of $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}$.

To determine whether hyperparameter stochasticity is responsible for improved generalization, we trained a standard LSTM while perturbing its dropout rate around the best value found by grid search. We perturbed using (1) random Gaussian perturbations, and (2) sinusoid perturbations, which implement a cyclic regularization schedule. STNs outperform both perturbation methods.

To determine whether limited capacity of $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}$ acts as an additional regularizer, we trained a standard LSTM from scratch using the schedule for output dropout discovered by the ST-LSTM, and again find that it outperforms the best fixed value, proving that the schedule itself provides some benefit. The final validation and test perplexities of each variant are shown in Table 1.

Next, we show that the schedules discovered by STNs are *greedy*, i.e., the hyperparameters are adapted to be well-suited for the current training phase. As shown empirically in Appendix E, lower regularization performs best early in training, while higher regularization performs better later. We found the STN schedule implements a type of curriculum by using a low dropout rate early in training, which aids optimization, and gradually increases the dropout rate as training proceeds,

	PTB		CIFAR-10	
Method	**Val Perplexity**	**Test Perplexity**	**Val Loss**	**Test Loss**
Grid Search	97.32	94.58	.1686	.1701
Random Search	84.81	81.46	.1518	.1517
Bayesian Optimization	72.13	69.29	.1366	**.1403**
STN	**70.93**	**68.12**	**.1339**	.1408

Table 2: Final validation and test performance of each method on the PTB word-level language modeling task, and the CIFAR-10 image-classification task.

leading to better generalization. We show the schedules the STN discovers for different initial dropout rates in Figure 2. The same schedule is followed no matter what initial rate is used.

Method	**Val**	**Test**
$p = 0.68$ (Fixed)	85.33	82.82
$p \sim \mathcal{N}(0.68, \sigma = 0.05)$	84.39	81.58
$p = 0.68 + 0.1 \sin(k\pi)$	85.05	82.42
$p = 0.78$ (Converged STN)	88.12	84.94
STN (Ours)	**83.25**	**80.54**
Following STN Schedule	83.62	80.58

Table 1: Comparing fixed and time-varying methods for setting the output dropout hyperparameter for an LSTM.

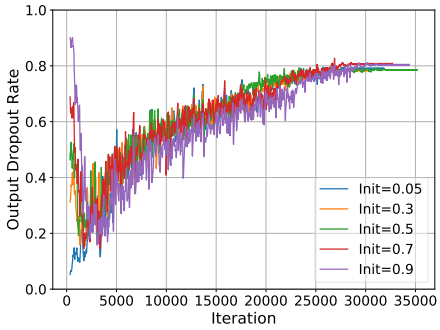
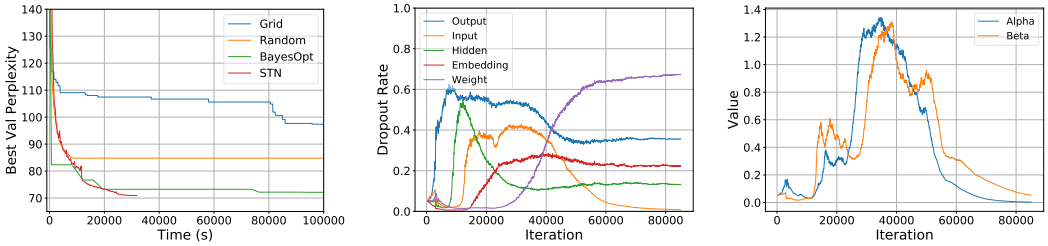


Figure 2: Output dropout schedules found by the ST-LSTM for different initial dropout rates

4.2 Language Modelling

We evaluated a ST-LSTM on the PTB corpus (Marcus et al., 1993), which is widely used as a benchmark for RNN regularization due to its small size (Gal & Ghahramani, 2016; Merity et al., 2018; Wen et al., 2018). We used a 2-layer LSTM with 650 hidden units per layer, and 650-dimensional word embeddings. We tuned 7 hyperparameters: variational dropout rates for the input, hidden state, and output; embedding dropout (i.e., setting rows of the embedding matrix to $\mathbf{0}$); DropConnect on the hidden-to-hidden weight matrix; and coefficients α and β that control the strength of activation regularization and temporal activation regularization, respectively. For LSTM tuning, we obtained the best results when using a fixed perturbation scale of 1 for the hyperparameters. Additional details about the experimental setup and the role of these hyperparameters can be found in Appendix C.

We compare STNs to grid search, random search, and Bayesian optimization. [4]. Figure 3a shows the best validation perplexity achieved by each method over time. STNs outperform the other methods, achieving lower validation perplexity more quickly. We show the schedules the STN finds for each hyperparameter in Figures 3b and 3c.



(a) Time comparison (b) STN schedule for dropout (c) STN schedule for α and β

Figure 3: A comparison of the best validation perplexity achieved on PTB over time, by grid search, random search, Bayesian optimization, and STNs. STNs achieve better (lower) validation perplexity in less time.

[4]For grid search and random search, we used the Ray Tune libraries (https://github.com/ray-project/ray/tree/master/python/ray/tune). For Bayesian optimization, we used Spearmint (https://github.com/HIPS/Spearmint)

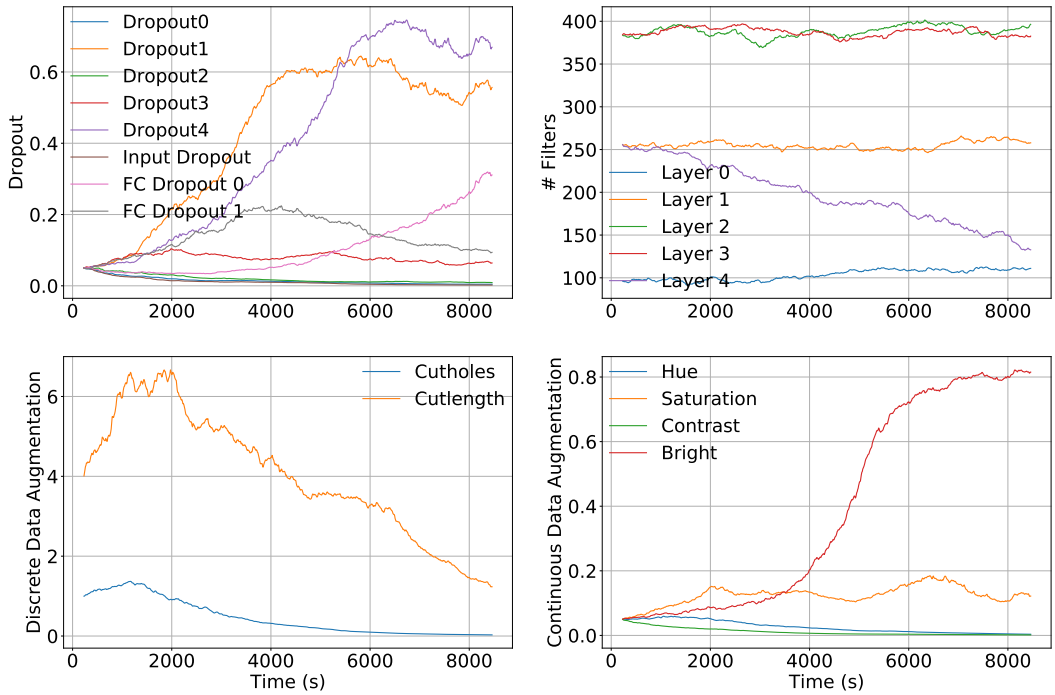


Figure 5: The hyperparameter schedule prescribed by the STN while training for image classification. The dropouts are indexed by the convolutional layer their applied to. FC dropout is for the fully-connected layers.

4.3 Image Classification

We evaluated Self-Tuning CNNs (ST-CNNs) on the CIFAR-10 (Krizhevsky, 2009) dataset, where it is easy to overfit with high-capacity networks. We used the AlexNet architecture (Krizhevsky et al., 2012). We optimized: (1) discrete architectural hyperparameters of the number of filters in each convolutional layer of AlexNet; (2) continuous hyperparameters of per-layer activation dropout and input dropout. (3) discrete data augmentation hyperparameters of the length and number of cut-out holes (DeVries & Taylor, 2017), and (4) continuous data augmentation hyperparameters controlling the amount of noise to apply to the hue, saturation, brightness, and contrast of an image.

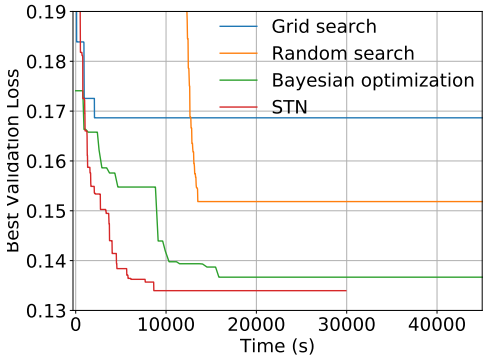


Figure 4: A comparison of the best validation loss achieved on CIFAR-10 over time, by grid search, random search, Bayesian optimization, and STNs. STNs outperform other methods for many computational budgets by finding the best validation loss.

We compared STNs to grid search, random search, and Bayesian optimization. Figure 4 shows the lowest validation loss achieved by each method over time. Details of the optimization procedures are in Appendix D. The baselines methods decay the learning rate at most 3 times by a factor of 10 every time we complete an epoch without the validation loss improving, after which we halt optimization early. Again, we find STNs find better hyperparameter configurations in less time than other methods.

5 Related Work

Bi-level Optimization Colson et al. (2007) is an overview of bilevel problems and Bard (2013) is a comprehensive textbook. When the objectives/constraints are restricted to be linear, quadratic or convex, a common approach replaces the lower-level problem with its KKT conditions added as constraints for the upper-level problem (Hansen et al., 1992; Vicente et al., 1994). In the unrestricted setting, our work loosely resembles trust-region methods (Colson et al., 2005), which repeatedly approximate the problem locally using a simpler bilevel program. In closely related work, Sinha et al. (2013) used evolutionary techniques to estimate the best-response function iteratively.

Hypernetworks Hypernetworks, first considered by Schmidhuber (1993; 1992), are functions mapping to the weights of a NN . Predicting certain weights in CNNs has been developed in various forms (Denil et al., 2013; Yang et al., 2015). Ha et al. (2016) used hypernetworks to generate weights for modern CNNs and RNNs. Brock et al. (2017) used hypernetworks to globally approximate an best-response for architecture search. Because the architecture is not optimized during training, this required a large hypernetwork, unlike ours which locally approximates the best-response.

Gradient-Based HO There are two main approaches. The first approximates $\boldsymbol{\theta}^*(\boldsymbol{\lambda}_0)$ using $\boldsymbol{\theta}_T(\boldsymbol{\lambda}_0, \boldsymbol{\theta}_0)$, the value of $\boldsymbol{\theta}$ after T steps of gradient descent on f w.r.t. $\boldsymbol{\theta}$ starting at $(\boldsymbol{\lambda}_0, \boldsymbol{\theta}_0)$. The descent steps are differentiated through to approximate $\frac{\partial \boldsymbol{\theta}^*}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}_0) \approx \frac{\partial \boldsymbol{\theta}_T}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}_0, \boldsymbol{\theta}_0)$. This approach was proposed by Domke (2012) and used by Maclaurin et al. (2015), Luketina et al. (2016) and Franceschi et al. (2018). The second approach uses the Implicit Function Theorem to derive $\frac{\partial \boldsymbol{\theta}^*}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}_0)$ under certain conditions (see Appendix A.4). This was first used for HO in NNs (Larsen et al., 1996) and developed further by Pedregosa (2016). Similar approaches have been used for HO in log-linear models (Foo et al., 2008), kernel selection (Chapelle et al., 2002; Seeger, 2007), and image reconstruction (Kunisch & Pock, 2013; Calatroni et al., 2016). Both approaches struggle with certain hyperparameters, since they differentiate gradient descent or the training loss w.r.t. the hyperparameters. In addition, differentiating gradient descent becomes prohibitively expensive as the number of descent steps increases, while implicitly deriving $\frac{\partial \boldsymbol{\theta}^*}{\partial \boldsymbol{\lambda}}$ requires using Hessian-vector products with conjugate gradient solvers to avoid directly computing the Hessian.

Model-Based HO A common model-based approach is Bayesian optimization, which models $p(r|\boldsymbol{\lambda}, \mathcal{D})$, the conditional probability of the performance on some metric r given hyperparameters $\boldsymbol{\lambda}$ and a dataset $\mathcal{D} = \{(\boldsymbol{\lambda}_i, r_i)\}$. $p(r|\boldsymbol{h}, \mathcal{D})$ can be modelled using various methods (Hutter et al., 2011; Bergstra et al., 2011; Snoek et al., 2012; 2015). $\mathcal{D}$ is constructed iteratively, where the next $\boldsymbol{\lambda}$ to train on is chosen by maximizing an acquisition function $C(\boldsymbol{h}; p(r|\boldsymbol{\lambda}, \mathcal{D}))$ which balances exploration and exploitation. Training each model to completion can be avoided if assumptions are made on learning curve behavior (Swersky et al., 2014; Klein et al., 2016). These approaches require building inductive biases into $p(r|\boldsymbol{h}, \mathcal{D})$ which may not hold in practice, do not take advantage of NN structure when used for HO, and do not scale well with the number of hyperparameters. However, these approaches have consistency guarantees in the limit, unlike ours.

Model-Free HO Model-free approaches include grid-search and random search. Bergstra & Bengio (2012) advocated using random search over grid search. Successive Halving (Jamieson & Talwalkar, 2016) and Hyperband (Li et al., 2016) extend random search by adaptively allocating resources to promising configurations using multi-armed bandit techniques. These methods ignore any structure in the problem, unlike ours which makes use of rich gradient information. However, unlike our method, it is possible to parallelize model-free methods over computing resources.

Hyperparameter Scheduling To our knowledge, the only other work that has considered schedules for hyperparameters is Population Based Training (PBT) (Jaderberg et al., 2017). In PBT, a population of NNs is trained in parallel, where the hyperparameters and weights of under-performing NNs are periodically replaced by their better-performing counterparts. The hyperparameters are randomly perturbed before copying. Thus, a single model can experience different hyperparameters over the course of training, implementing a schedule. STNs replace the population of NNs by a single hypernetwork and use gradients to tune hyperparameters during a single training run. The schedule discovered by the STN can be seen as an instance of curriculum learning (Bengio et al., 2009), a strategy for training machine learning models, where the model is first trained on simple examples and then on gradually more difficult ones.

6 Conclusion

We proposed Self-Tuning Networks (STNs), which efficiently approximate the best-response to hyperparameters by scaling and shifting their hidden units. This allowed us to use gradient-based optimization to tune all kinds of regularization hyperparameters, including discrete hyperparameters. We validated this approach on large-scale problems by showing STNs can achieve better generalization performance than competing approaches in less time. We believe STNs offer a compelling path towards large-scale, automated hyperparameter tuning for neural networks.

REFERENCES

Eugene L Allgower and Kurt Georg. *Numerical continuation methods: An introduction*, volume 13. Springer Science & Business Media, 2012.

Jonathan F Bard. *Practical bilevel optimization: Algorithms and applications*, volume 30. Springer Science & Business Media, 2013.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 41–48. ACM, 2009.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

James S. Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, pp. 2546–2554. 2011.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.

Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. SMASH: One-shot model architecture search through hypernetworks. *arXiv preprint arXiv:1708.05344*, 2017.

Luca Calatroni, Chung Cao, Juan Carlos De Los Reyes, Carola-Bibiane Schönlieb, and Tuomo Valkonen. Bilevel approaches for learning of variational imaging models. 2016.

Olivier Chapelle, Vladimir Vapnik, Olivier Bousquet, and Sayan Mukherjee. Choosing multiple parameters for Support Vector Machines. *Machine Learning*, 46(1-3):131–159, 2002.

Benoît Colson, Patrice Marcotte, and Gilles Savard. A trust-region method for nonlinear bilevel programming: Algorithm and computational experience. *Computational Optimization and Applications*, 30(3):211–227, 2005.

Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Annals of Operations Research*, 153(1):235–256, 2007.

Misha Denil, Babak Shakibi, Laurent Dinh, Nando De Freitas, et al. Predicting parameters in deep learning. In *Advances in Neural Information Processing Systems*, pp. 2148–2156, 2013.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Justin Domke. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pp. 318–326, 2012.

John Duchi. Properties of the trace and matrix derivatives. 2007.

Anthony V Fiacco and Yo Ishizuka. Sensitivity and stability analysis for nonlinear programming. *Annals of Operations Research*, 27(1):215–235, 1990.

Chuan-sheng Foo, Chuong B Do, and Andrew Y Ng. Efficient multiple hyperparameter learning for log-linear models. In *Advances in Neural Information Processing Systems*, pp. 377–384, 2008.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, and Massimilano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. *arXiv preprint arXiv:1806.04910*, 2018.

Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pp. 1019–1027, 2016.

Robert Gibbons. *A primer in game theory*. Harvester Wheatsheaf, 1992.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

Pierre Hansen, Brigitte Jaumard, and Gilles Savard. New branch-and-bound rules for linear bilevel programming. *SIAM Journal on Scientific and Statistical Computing*, 13(5):1194–1217, 1992.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *International Conference on Learning and Intelligent Optimization*, pp. 507–523. Springer, 2011.

Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, et al. Population-based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017.

Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Artificial Intelligence and Statistics*, pp. 240–248, 2016.

Mohammad Emtiyaz Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable Bayesian deep learning by weight-perturbation in Adam. *arXiv preprint arXiv:1806.04854*, 2018.

Aaron Klein, Stefan Falkner, Jost Tobias Springenberg, and Frank Hutter. Learning curve prediction with Bayesian neural networks. 2016.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

Karl Kunisch and Thomas Pock. A bilevel optimization approach for parameter learning in variational models. *SIAM Journal on Imaging Sciences*, 6(2):938–983, 2013.

Jan Larsen, Lars Kai Hansen, Claus Svarer, and M Ohlsson. Design and regularization of neural networks: The optimal use of a validation set. In *Neural Networks for Signal Processing [1996] VI. Proceedings of the 1996 IEEE Signal Processing Society Workshop*, pp. 62–71. IEEE, 1996.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: Bandit-based configuration evaluation for hyperparameter optimization. 2016.

Jonathan Lorraine and David Duvenaud. Stochastic hyperparameter optimization through hypernetworks. *arXiv preprint arXiv:1802.09419*, 2018.

Christos Louizos and Max Welling. Multiplicative normalizing flows for variational Bayesian neural networks. *arXiv preprint arXiv:1703.01961*, 2017.

Jelena Luketina, Mathias Berglund, Klaus Greff, and Tapani Raiko. Scalable gradient-based tuning of continuous regularization hyperparameters. In *International Conference on Machine Learning*, pp. 2952–2960, 2016.

Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *International Conference on Machine Learning*, pp. 2113–2122, 2015.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing LSTM language models. *International Conference on Learning Representations*, 2018.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: An empirical study. *arXiv preprint arXiv:1802.08760*, 2018.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *International Conference on Machine Learning*, pp. 737–746, 2016.

Jürgen Schmidhuber. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139, 1992.

Jürgen Schmidhuber. A 'self-referential' weight matrix. In *ICANN'93*, pp. 446–450. Springer, 1993.

Matthias Seeger. Cross-validation optimization for large scale hierarchical classification kernel methods. In *Advances in Neural Information Processing Systems*, pp. 1233–1240, 2007.

Ankur Sinha, Pekka Malo, and Kalyanmoy Deb. Efficient evolutionary algorithm for single-objective bilevel optimization. *arXiv preprint arXiv:1303.3901*, 2013.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pp. 2951–2959, 2012.

Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, M Prabhat, and Ryan Adams. Scalable Bayesian optimization using deep neural networks. In *International Conference on Machine Learning*, pp. 2171–2180, 2015.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Joe Staines and David Barber. Variational optimization. *arXiv preprint arXiv:1212.4507*, 2012.

Kevin Swersky, Jasper Snoek, and Ryan Prescott Adams. Freeze-thaw Bayesian optimization. *arXiv preprint arXiv:1406.3896*, 2014.

Luis Vicente, Gilles Savard, and Joaquim Júdice. Descent approaches for quadratic bilevel programming. *Journal of Optimization Theory and Applications*, 81(2):379–399, 1994.

Heinrich Von Stackelberg. *Market Structure and Equilibrium*. Springer Science & Business Media, 2010.

Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *International Conference on Learning Representations*, 2018.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Zichao Yang, Marcin Moczulski, Misha Denil, Nando de Freitas, Alex Smola, Le Song, and Ziyu Wang. Deep fried convnets. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1476–1483, 2015.

Wojciech Zaremba and Ilya Sutskever. Reinforcement learning neural Turing machines-revised. *arXiv preprint arXiv:1505.00521*, 2015.

Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

A PROOFS

A.1 LEMMA 1

Because $\boldsymbol{\theta}_0$ solves Problem 4b given $\boldsymbol{\lambda}_0$, by the first-order optimality condition, we must have:

$$\frac{\partial f}{\partial \boldsymbol{\theta}}(\boldsymbol{\lambda}_0, \boldsymbol{\theta}_0) = 0 \tag{15}$$

The Jacobian of $\frac{\partial f}{\partial \boldsymbol{\theta}}$ decomposes as a block matrix with subblocks given by:

$$\left[\begin{array}{c|c} \frac{\partial^2 f}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\theta}} & \frac{\partial^2 f}{\partial \boldsymbol{\theta}^2} \end{array}\right] \tag{16}$$

We know f is $\mathcal{C}^2$ in some neighborhood of $(\boldsymbol{\lambda}_0, \boldsymbol{\theta}_0)$, so $\frac{\partial f}{\partial \boldsymbol{\theta}}$ is continuously differentiable in this neighborhood. By assumption, the Hessian $\frac{\partial^2 f}{\partial \boldsymbol{\theta}^2}$ is positive definite and hence invertible at $(\boldsymbol{\lambda}_0, \boldsymbol{\theta}_0)$. By the Implicit Function Theorem, there exists a neighborhood V of $\boldsymbol{\lambda}_0$ and a unique continuously differentiable function $\boldsymbol{\theta}^* : V \to \mathbb{R}^m$ such that $\frac{\partial f}{\partial \boldsymbol{\theta}}(\boldsymbol{\lambda}, \boldsymbol{\theta}^*(\boldsymbol{\lambda})) = 0$ for $\boldsymbol{\lambda} \in V$ and $\boldsymbol{\theta}^*(\boldsymbol{\lambda}_0) = \boldsymbol{\theta}_0$.

Furthermore, by continuity we know there is a neighborhood $W_1 \times W_2$ of $(\boldsymbol{\lambda}_0, \boldsymbol{\theta}_0)$ such that $\frac{\partial^2 f}{\partial \boldsymbol{\theta}^2}$ is positive definite on this neighborhood. Setting $U = V \cap W_1 \cap (\theta^*)^{-1}(W_2)$, we can conclude that $\frac{\partial^2 f}{\partial \boldsymbol{\theta}^2}(\boldsymbol{\lambda}, \boldsymbol{\theta}^*(\boldsymbol{\lambda})) \succ 0$ for all $\boldsymbol{\lambda} \in U$. Combining this with $\frac{\partial f}{\partial \boldsymbol{\theta}}(\boldsymbol{\lambda}, \boldsymbol{\theta}^*(\boldsymbol{\lambda})) = 0$ and using second-order sufficient optimality conditions, we conclude that $\boldsymbol{\theta}^*(\boldsymbol{\lambda})$ is the unique solution to Problem 4b for all $\boldsymbol{\lambda} \in U$.

A.2 LEMMA 2

This discussion mostly follows from Hastie et al. (2001). We let $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ denote the data matrix where N is the number of training examples, and D is the dimensionality of the data. We let $\boldsymbol{t} \in \mathbb{R}^N$ denote the associated targets. We can write the SVD decomposition of $\boldsymbol{X}$ as:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T \tag{17}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are $N \times D$ and $D \times D$ orthogonal matrices and $\boldsymbol{D}$ is a diagonal matrix with entries $d_1 \geq d_2 \geq \cdots \geq d_D \geq 0$. We next simplify the function $y(\boldsymbol{x}; \boldsymbol{\theta})$ by setting $\boldsymbol{v} = \boldsymbol{s}^T\boldsymbol{Q}$, so that $y(\boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{s}^T\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{v}^T\boldsymbol{x}$. We see that the Jacobian $\frac{\partial y}{\partial \boldsymbol{x}} \equiv \boldsymbol{v}$ is constant, and Problem 10 simplifies to standard L_2-regularized least-squares linear regression with the following loss function:

$$\sum_{(\boldsymbol{x}, t) \in \mathcal{D}} (\boldsymbol{v}^T\boldsymbol{x} - t)^2 + \frac{1}{|\mathcal{D}|} \exp(\lambda) \|\boldsymbol{v}\|^2 \tag{18}$$

It is well-known (see Hastie et al. (2001), Chapter 3) that the optimal solution $\boldsymbol{v}^*(\lambda)$ minimizing Equation 18 is given by:

$$\boldsymbol{v}^*(\lambda) = (\boldsymbol{X}^T\boldsymbol{X} + \exp(\lambda)\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{t} = \boldsymbol{V}(\boldsymbol{D}^2 + \exp(\lambda)\boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{t} \tag{19}$$

Furthermore, the optimal solution $\boldsymbol{v}^*$ to the unregularized version of Problem 18 is given by:

$$\boldsymbol{v}^* = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^T\boldsymbol{t} \tag{20}$$

Recall that we defined $\boldsymbol{Q}_0 = \boldsymbol{V}^T$, i.e. the change-of-basis matrix from the standard basis to the principal components of the data matrix, and we defined $\boldsymbol{s}_0$ to solve the unregularized regression problem given $\boldsymbol{Q}_0$. Thus, we require that $\boldsymbol{Q}_0^T\boldsymbol{s}_0 = \boldsymbol{v}^*$ which implies $\boldsymbol{s}_0 = \boldsymbol{D}^{-1}\boldsymbol{U}^T\boldsymbol{t}$.

There are not unique solutions to Problem 10, so we take any functions $\boldsymbol{Q}(\lambda), \boldsymbol{s}(\lambda)$ which satisfy $\boldsymbol{Q}(\lambda)^T\boldsymbol{s}(\lambda) = \boldsymbol{v}^*(\lambda)$ as "best-response functions". We will show that our chosen functions $\boldsymbol{Q}^*(\lambda) = \sigma(\lambda\boldsymbol{w} + \boldsymbol{b}) \odot \boldsymbol{Q}_0$ and $\boldsymbol{s}^*(\lambda) = \boldsymbol{s}_0$, where $\boldsymbol{w} = \boldsymbol{0}$ and $b_i = 2\log(d_i)$ for $i = 1, \ldots, D$, meet this criteria. We start by noticing that for any $d \in \mathbb{R}_+$, we have:

$$\sigma(\lambda + 2\log(d)) = \frac{1}{1 + \exp(-\lambda - 2\log(d))} = \frac{1}{1 + d^{-2}\exp(\lambda)} = \frac{d^2}{d^2 + \exp(\lambda)} \tag{21}$$

It follows that:

$$\boldsymbol{Q}^*(\lambda)^T \boldsymbol{s}^*(\lambda) = [\sigma(\lambda \boldsymbol{w} + \boldsymbol{b}) \odot \boldsymbol{Q}_0]^T \boldsymbol{s}_0 \tag{22}$$

$$= \left[\operatorname{diag} \begin{pmatrix} \sigma(\lambda + 2\log(d_1)) \\ \vdots \\ \sigma(\lambda + 2\log(d_m)) \end{pmatrix} \boldsymbol{Q}_0 \right]^T \boldsymbol{s}_0 \tag{23}$$

$$= \boldsymbol{Q}_0^T \left[\operatorname{diag} \begin{pmatrix} \frac{d_1^2}{d_1^2 + \exp(\lambda)} \\ \vdots \\ \frac{d_D^2}{d_D^2 + \exp(\lambda)} \end{pmatrix} \right] \boldsymbol{s}_0 \tag{24}$$

$$= \boldsymbol{V} \left[\operatorname{diag} \begin{pmatrix} \frac{d_1^2}{d_1^2 + \exp(\lambda)} \\ \vdots \\ \frac{d_D^2}{d_D^2 + \exp(\lambda)} \end{pmatrix} \right] \boldsymbol{D}^{-1} \boldsymbol{U}^T \boldsymbol{t} \tag{25}$$

$$= \boldsymbol{V} \left[\operatorname{diag} \begin{pmatrix} \frac{d_1}{d_1^2 + \exp(\lambda)} \\ \vdots \\ \frac{d_D}{d_D^2 + \exp(\lambda)} \end{pmatrix} \right] \boldsymbol{U}^T \boldsymbol{t} \tag{26}$$

$$= \boldsymbol{V} \left[(\boldsymbol{D}^2 + \exp(\lambda)\boldsymbol{I})^{-1} \boldsymbol{D} \right] \boldsymbol{U}^T \boldsymbol{t} \tag{27}$$

$$= \boldsymbol{v}^*(\lambda) \tag{28}$$

A.3 THEOREM 3

By assumption, f is quadratic so there exist $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, $\boldsymbol{B} \in \mathbb{R}^{n \times m}$, $\boldsymbol{C} \in \mathbb{R}^{m \times m}$ and $\boldsymbol{d} \in \mathbb{R}^n, \boldsymbol{e} \in \mathbb{R}^m$ such that:

$$f(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \frac{1}{2} \begin{pmatrix} \boldsymbol{\lambda}^T & \boldsymbol{\theta}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^T & \boldsymbol{C} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\theta} \end{pmatrix} + \boldsymbol{d}^T \boldsymbol{\lambda} + \boldsymbol{e}^T \boldsymbol{\theta} \tag{29}$$

Furthermore, since we assume $\frac{\partial^2 f}{\partial \boldsymbol{\theta}^2} \succ \boldsymbol{0}$, we must have $\boldsymbol{C} \succ \boldsymbol{0}$. One can easily compute that:

$$\frac{\partial f}{\partial \boldsymbol{\theta}}(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \boldsymbol{B}^T \boldsymbol{\lambda} + \boldsymbol{C}\boldsymbol{\theta} + \boldsymbol{e} \tag{30}$$

$$\frac{\partial^2 f}{\partial \boldsymbol{\theta}^2}(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \boldsymbol{C} \tag{31}$$

Setting the derivative equal to $\boldsymbol{0}$ and using second-order sufficient conditions, we have:

$$\boldsymbol{\theta}^*(\boldsymbol{\lambda}) = -\boldsymbol{C}^{-1}(\boldsymbol{e} - \boldsymbol{B}^T \boldsymbol{\lambda}) \tag{32}$$

Hence, we find:

$$\frac{\partial \boldsymbol{\theta}^*}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}) = -\boldsymbol{C}^{-1} \boldsymbol{B}^T \tag{33}$$

We can substitute $\hat{\boldsymbol{\theta}}_{\boldsymbol{\phi}}(\boldsymbol{\lambda}) = \boldsymbol{U}\boldsymbol{\lambda} + \boldsymbol{b}$ and the definition of f into equation 8 to obtain:

$$\begin{aligned} e(\boldsymbol{U}, \boldsymbol{b}, \sigma) := \mathbb{E}_{\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}|\boldsymbol{\sigma})} [& (\boldsymbol{\lambda}_0 + \boldsymbol{\epsilon})^T \boldsymbol{A} (\boldsymbol{\lambda}_0 + \boldsymbol{\epsilon}) + 2(\boldsymbol{\lambda}_0 + \boldsymbol{\epsilon})^T \boldsymbol{B} (\boldsymbol{U}(\boldsymbol{\lambda}_0 + \boldsymbol{\epsilon}) + \boldsymbol{b}) \\ & + (\boldsymbol{U}(\boldsymbol{\lambda}_0 + \boldsymbol{\epsilon}) + \boldsymbol{b})^T \boldsymbol{C} (\boldsymbol{U}(\boldsymbol{\lambda}_0 + \boldsymbol{\epsilon}) + \boldsymbol{b}) \\ & + \boldsymbol{d}^T (\boldsymbol{\lambda}_0 + \boldsymbol{\epsilon}) + \boldsymbol{e}^T (\boldsymbol{U}(\boldsymbol{\lambda}_0 + \boldsymbol{\epsilon}) + \boldsymbol{b})] \end{aligned} \tag{34}$$

Expanding above, we can find that equation 34 is equal to:

$$\mathbb{E}_{\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}|\boldsymbol{\sigma})} \left[\textcircled{1} + \textcircled{2} + \textcircled{3} + \textcircled{4} \right] \tag{35}$$

where we have:

$$\textcircled{1} = \frac{1}{2}\left(\boldsymbol{\lambda}_0^T \boldsymbol{A}\boldsymbol{\lambda}_0 + 2\boldsymbol{\epsilon}^T \boldsymbol{A}\boldsymbol{\lambda}_0 + \boldsymbol{\epsilon}^T \boldsymbol{A}\boldsymbol{\epsilon}\right) \tag{36}$$

$$\textcircled{2} = \boldsymbol{\lambda}_0^T \boldsymbol{B}\boldsymbol{U}\boldsymbol{\lambda}_0 + \boldsymbol{\lambda}_0^T \boldsymbol{B}\boldsymbol{U}\boldsymbol{\epsilon} + \boldsymbol{\lambda}_0^T \boldsymbol{B}\boldsymbol{b} + \boldsymbol{\epsilon}^T \boldsymbol{B}\boldsymbol{U}\boldsymbol{\lambda}_0 + \boldsymbol{\epsilon}^T \boldsymbol{B}\boldsymbol{U}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T \boldsymbol{B}\boldsymbol{b} \tag{37}$$

$$\begin{aligned}\textcircled{3} = \frac{1}{2}(&\boldsymbol{\lambda}_0^T \boldsymbol{U}^T \boldsymbol{C}\boldsymbol{U}\boldsymbol{\lambda}_0 + \boldsymbol{\lambda}_0 \boldsymbol{U}^T \boldsymbol{C}\boldsymbol{U}\boldsymbol{\epsilon} + \boldsymbol{\lambda}_0 \boldsymbol{U}^T \boldsymbol{C}\boldsymbol{b} + \boldsymbol{\epsilon}^T \boldsymbol{U}^T \boldsymbol{C}\boldsymbol{U}\boldsymbol{\lambda}_0 \\ &+ \boldsymbol{\epsilon}^T \boldsymbol{U}^T \boldsymbol{C}\boldsymbol{U}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T \boldsymbol{U}^T \boldsymbol{C}\boldsymbol{b} + \boldsymbol{b}^T \boldsymbol{C}\boldsymbol{U}\boldsymbol{\lambda}_0 + \boldsymbol{b}^T \boldsymbol{C}\boldsymbol{U}\boldsymbol{\epsilon} + \boldsymbol{b}^T \boldsymbol{C})\end{aligned} \tag{38}$$

$$\textcircled{4} = \boldsymbol{d}^T \boldsymbol{\lambda}_0 + \boldsymbol{d}^T \boldsymbol{\epsilon} + \boldsymbol{e}^T \boldsymbol{U}\boldsymbol{\lambda}_0 + \boldsymbol{e}^T \boldsymbol{U}\boldsymbol{\epsilon} + \boldsymbol{e}^T \boldsymbol{b} \tag{39}$$

We can simplify these expressions considerably by using linearity of expectation and that $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}|\sigma)$ has mean $\mathbf{0}$. Doing so, then differentiating and making use of various matrix-derivative equalities (Duchi, 2007), we find:

$$\frac{\partial e}{\partial \boldsymbol{b}}(\boldsymbol{U}, \boldsymbol{b}, \sigma) = \frac{1}{2}\boldsymbol{C}^T \boldsymbol{U}\boldsymbol{\lambda}_0 + \frac{1}{2}\boldsymbol{C}\boldsymbol{U}\boldsymbol{\lambda}_0 + \boldsymbol{B}^T \boldsymbol{\lambda}_0 + \boldsymbol{e} + \boldsymbol{C}\boldsymbol{b} \tag{40}$$

$$\frac{\partial e}{\partial \boldsymbol{U}}(\boldsymbol{U}, \boldsymbol{b}, \sigma) = \boldsymbol{B}^T \boldsymbol{\lambda}_0 \boldsymbol{\lambda}_0^T + \sigma^2 \boldsymbol{B}^T + \boldsymbol{C}\boldsymbol{b}\boldsymbol{\lambda}_0^T + \boldsymbol{e}\boldsymbol{\lambda}_0^T + \boldsymbol{C}\boldsymbol{U}\boldsymbol{\lambda}_0 \boldsymbol{\lambda}_0^T + \sigma^2 \boldsymbol{C}\boldsymbol{U} \tag{41}$$

Setting the derivative $\frac{\partial e}{\partial \boldsymbol{b}}(\boldsymbol{U}, \boldsymbol{b}, \sigma)$ equal to $\mathbf{0}$, we have:

$$\boldsymbol{b} = -\boldsymbol{C}^{-1}(\boldsymbol{C}^T \boldsymbol{U}\boldsymbol{\lambda}_0 + \boldsymbol{B}^T \boldsymbol{\lambda}_0 + \boldsymbol{e}) \tag{42}$$

Setting the derivative for $\frac{\partial e}{\partial \boldsymbol{U}}(\boldsymbol{U}, \boldsymbol{b}, \sigma)$ equal to $\mathbf{0}$, we have:

$$\boldsymbol{C}\boldsymbol{U}(\boldsymbol{\lambda}_0 \boldsymbol{\lambda}_0^T + \sigma^2 \boldsymbol{C}) = -\boldsymbol{B}^T \boldsymbol{\lambda}_0 \boldsymbol{\lambda}_0^T - \sigma^2 \boldsymbol{B}^T - \boldsymbol{C}\boldsymbol{b}\boldsymbol{\lambda}_0^T - \boldsymbol{e}\boldsymbol{\lambda}_0^T \tag{43}$$

Substituting the expression for $\boldsymbol{b}$ given by equation 42 into equation 43 and simplifying gives:

$$\boldsymbol{C}\boldsymbol{U}(\boldsymbol{\lambda}_0 \boldsymbol{\lambda}_0^T + \sigma^2 \boldsymbol{I}) = -\sigma^2 \boldsymbol{B}^T + \boldsymbol{C}\boldsymbol{U}\boldsymbol{\lambda}_0 \boldsymbol{\lambda}_0^T \tag{44}$$

$$\implies \sigma^2 \boldsymbol{C}\boldsymbol{U} = -\sigma^2 \boldsymbol{B}^T \tag{45}$$

$$\implies \boldsymbol{U} = \boldsymbol{C}^{-1}\boldsymbol{B} \tag{46}$$

This is exactly the best-response Jacobian $\frac{\partial \boldsymbol{\theta}^*}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda})$ as given by equation 33. Substituting $\boldsymbol{U} = \boldsymbol{C}^{-1}\boldsymbol{B}$ into the equation 42 gives:

$$\boldsymbol{b} = \boldsymbol{C}^{-1}\boldsymbol{B}^T \boldsymbol{\lambda}_0 - \boldsymbol{C}^{-1}\boldsymbol{B}^T \boldsymbol{\lambda}_0 - \boldsymbol{C}^{-1}\boldsymbol{e} \tag{47}$$

This is $\boldsymbol{\theta}^*(\boldsymbol{\lambda}) - \frac{\partial \boldsymbol{\theta}^*}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda})$, as claimed.

A.4 Best-Response Gradient Lemma

Lemma 4. *Under the same conditions as Lemma 1 and using the same notation, for all $\boldsymbol{\lambda} \in U$, we have that:*

$$\frac{\partial \boldsymbol{\theta}^*}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}) = -\left[\frac{\partial^2 f}{\partial \boldsymbol{\theta}^2}(\boldsymbol{\lambda}, \boldsymbol{\theta}^*(\boldsymbol{\lambda}))\right]^{-1} \frac{\partial^2 f}{\partial \boldsymbol{\lambda}\partial \boldsymbol{\theta}}(\boldsymbol{\lambda}, \boldsymbol{\theta}^*(\boldsymbol{\lambda})) \tag{48}$$

Proof. Define $\iota^* : U \to \mathbb{R}^n \times \mathbb{R}^m$ by $\iota^*(\boldsymbol{\lambda}) = (\boldsymbol{\lambda}, \boldsymbol{\theta}^*(\boldsymbol{\lambda}))$. By first-order optimality conditions, we know:

$$\left(\frac{\partial f}{\partial \boldsymbol{\theta}} \circ \iota^*\right)(\boldsymbol{\lambda}) = 0 \quad \forall \boldsymbol{\lambda} \in U \tag{49}$$

Hence, for all $\boldsymbol{\lambda} \in U$:

$$0 = \frac{\partial}{\partial \boldsymbol{\lambda}}\left(\frac{\partial f}{\partial \boldsymbol{\theta}} \circ \iota^*\right)(\boldsymbol{\lambda}) \tag{50}$$

$$= \frac{\partial^2 f}{\partial \boldsymbol{\theta}^2}(\iota^*(\boldsymbol{\lambda}))\frac{\partial \boldsymbol{\theta}^*}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}) + \frac{\partial^2 f}{\partial \boldsymbol{\lambda}\partial \boldsymbol{\theta}}(\iota^*(\boldsymbol{\lambda})) \tag{51}$$

$$= \frac{\partial^2 f}{\partial \boldsymbol{\theta}^2}(\boldsymbol{\lambda}, \boldsymbol{\theta}^*(\boldsymbol{\lambda}))\frac{\partial \boldsymbol{\theta}^*}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}) + \frac{\partial^2 f}{\partial \boldsymbol{\lambda}\partial \boldsymbol{\theta}}(\boldsymbol{\lambda}, \boldsymbol{\theta}^*(\lambda)) \tag{52}$$

Rearranging gives Equation 48. □

B BEST-RESPONSE APPROXIMATIONS FOR CONVOLUTIONAL FILTERS

We let L denote the number of layers, C_l the number of channels in layer l's feature map, and K_l the size of the kernel in layer l. We let $\boldsymbol{W}^{l,c} \in \mathbb{R}^{C_{l-1} \times K_l \times K_l}$ and $\boldsymbol{b}^{l,c} \in \mathbb{R}$ denote the weight and bias respectively of the cth convolution kernel in layer l (so $c \in \{1, \ldots, C_l\}$). For $\boldsymbol{u}^{l,c}, \boldsymbol{a}^{l,c} \in \mathbb{R}^n$, we define best-response approximations $\hat{\boldsymbol{W}}^{l,c}$ and $\hat{\boldsymbol{b}}^{l,c}$ by:

$$\hat{\boldsymbol{W}}^{l,c}(\boldsymbol{\lambda}) = \langle \boldsymbol{\lambda}, \boldsymbol{u}^{l,c} \rangle \boldsymbol{W}^{l,c}_{\text{scale}} + \boldsymbol{W}^{l,c}_{\text{shift}} \tag{53}$$

$$\hat{\boldsymbol{b}}^{l,c}(\boldsymbol{\lambda}) = \langle \boldsymbol{\lambda}, \boldsymbol{a}^{l,c} \rangle \boldsymbol{b}^{l,c}_{\text{scale}} + \boldsymbol{b}^{l,c}_{\text{shift}} \tag{54}$$

Thus, the best-response parameters used for modeling $\boldsymbol{W}^{l,c}$, $\boldsymbol{b}^l$ are $\boldsymbol{\phi}^{l,c} = \{\boldsymbol{u}^{l,c}, \boldsymbol{a}^{l,c}, \boldsymbol{W}^{l,c}_{\text{scale}}, \boldsymbol{W}^{l,c}_{\text{shift}}, \boldsymbol{b}^{l,c}_{\text{scale}}, \boldsymbol{b}^{l,c}_{\text{shift}}\}$. We can compute the number of parameters used as $|\boldsymbol{\phi}^{l,c}| = 2n + 2(|\boldsymbol{W}^{l,c}| + |\boldsymbol{b}^{l,c}|)$. Summing over channels c, we find the total number of parameters is $2nC_l + 2p$, where p is the total number of parameters in the normal CNN layer. Hence, we use twice the number of parameters in a normal CNN, plus an additional overhead which depends on the number of hyperparameters.

C LANGUAGE MODELLING EXPERIMENT DETAILS

Here we present additional details on the setup of our LSTM experiments on PTB, and on the role of each hyperparameter we tune.

We tune variational dropout (re-using the same dropout mask for each step in a sequence) on the input to the LSTM, the hidden state between the LSTM layers, and the output of the LSTM. We also tune embedding dropout, which sets entire rows of the word embedding matrix to 0, effectively removing certain words from all sequences. We regularize the hidden-to-hidden weight matrix using DropConnect (zeroing out weights rather than activations). Finally, we use activation regularization (AR) and temporal activation regularization (TAR). AR penalizes large activations, and is defined as:

$$\alpha ||m \odot h_t||_2 \tag{55}$$

where m is a dropout mask and h_t is the output of the LSTM at time t. TAR is a slowness regularizer, defined as:

$$\beta ||h_t - t_{t+1}||_2 \tag{56}$$

We tune the scaling coefficients α and β.

D IMAGE CLASSIFICATION EXPERIMENT DETAILS

Here we present additional details on the CNN experiments. For all results we hold out 20% of the data for validation.

Our optimization algorithm for the elementary model in the baselines is as follows: We use SGD with a learning rate of 0.01, momentum of 0.9, mini-batches of size 128, and run for at most 200 epochs.

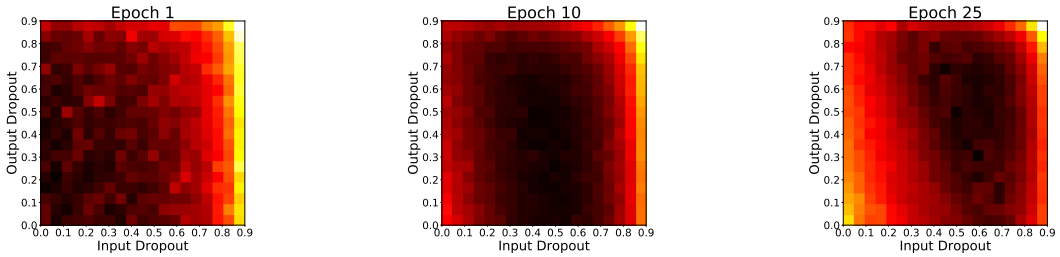
Our optimization algorithm for the ST-CNN is as follows: The ST-CNN is optimized with the same algorithm as the elementary model in the baselines. In other words, we use SGD with learning rate 0.01, momentum of 0.9, mini-batches of size 128. We use the same optimization algorithm as the ST-LSTM for the hyperparameters. This means the hyperparameters are optimized with Adam using a learning rate of 0.003. We alternate between training the hypernet and hyperparameters with the same schedule as the ST-LSTM - i.e. $T_{train} = 15$ steps on the training step and $T_{valid} = 10$ steps on the validation set.

Additionally, we use an entropy weight of 0 during training when optimizing the scale of the hyperparameter distributions.

E ADDITIONAL DETAILS ON HYPERPARAMETER SCHEDULES

Here, we draw connections between the hyperparameter schedules learned by STNs and curriculum learning. Curriculum learning is an instance of a family of *continuation methods* (Allgower & Georg, 2012), which optimize non-convex functions by solving a sequence of functions that are ordered by increasing difficulty. In a continuation method, one considers a family of training criteria $C_\lambda(\theta)$ with parameter λ, where $C_1(\theta)$ is the true/final objective we wish to minimize, and $C_0(\theta)$ represents the training criterion for a simpler version of the problem; one starts by optimizing $C_0(\theta)$ and then gradually increases λ from 0 to 1, while keeping θ at a local minimum of $C_\lambda(\theta)$. When training an RL-NTM (Zaremba & Sutskever, 2015), the authors maintain a distribution over the complexity c of a problem instance (defined as the maximum length of the target output for a given input), which is shifted towards harder problems as the performance of the model improves.

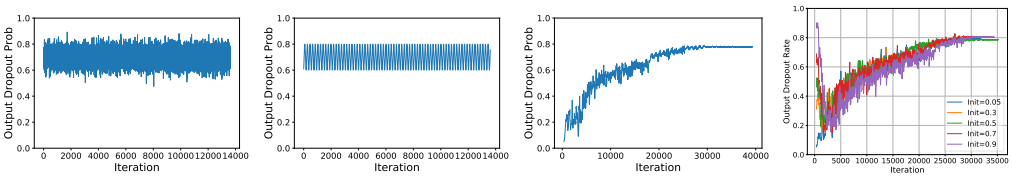
Figure 6 shows the validation perplexity for different sets of input and output dropout, at various points during training. The x- and y-axes represent the values of input and output dropout, respectively. We see that at the start of training, the best validation loss is achieved with small values of both input and output dropout. As we train for more epochs, the best validation performance is achieved with larger dropout rates.



(a) (b) (c)

Figure 6: Validation performance of a baseline LSTM given different settings of input and output dropout over time.

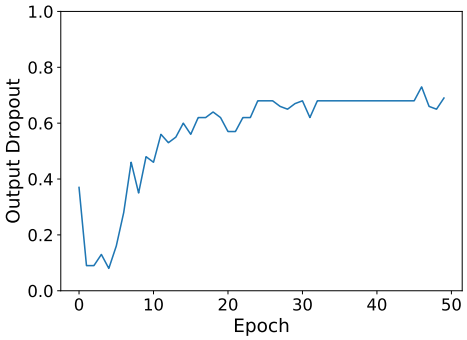
Figure 7 shows the perturbed values for output dropout we use to investigate whether the improved performance yielded by STNs is due to the regularization effect, and not the schedule itself.



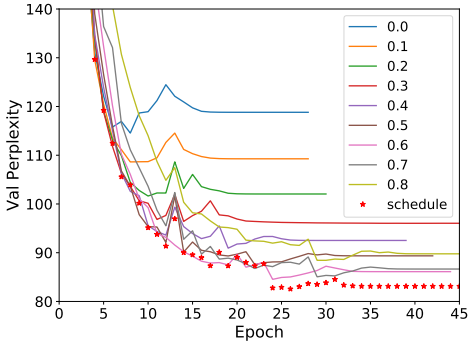
(a) $p \sim \mathcal{N}(0.68, 0.05)$ (b) $p = 0.68$ with sin noise (c) ST-LSTM (d) ST-LSTM Initial Hyperparameters

Figure 7: Output dropout rates

We can observe the benefits of greedy hyperparameter schedules by forming a schedule based on the performance in time of models from a grid search over output dropout. As shown in Figure 8, the schedule formed by taking the best output dropout value in each epoch yields better generalization than any of the fixed hyperparameter values from the initial grid search.



(a) Greedy schedule for output dropout, derived from grid search



(b) Comparison of fixed output dropout values and the dropout schedule derived from grid searches

Figure 8: Grid-search derived output dropout schedules. All experiments here use learning rate decay.