Second-Order Neural ODE Optimizer

Guan-Horng Liu, Tianrong Chen, Evangelos A. Theodorou
Georgia Institute of Technology, USA
{ghliu, tianrong.chen, evangelos.theodorou}@gatech.edu

Abstract

We propose a novel second-order optimization framework for training the emerging deep continuous-time models, specifically the Neural Ordinary Differential Equations (Neural ODEs). Since their training already involves expensive gradient computation by solving a backward ODE, deriving efficient second-order methods becomes highly nontrivial. Nevertheless, inspired by the recent Optimal Control (OC) interpretation of training deep networks, we show that a specific continuous-time OC methodology, called *Differential Programming*, can be adopted to derive backward ODEs for higher-order derivatives at the same $\mathcal{O}(1)$ memory cost. We further explore a low-rank representation of the second-order derivatives and show that it leads to efficient preconditioned updates with the aid of Kronecker-based factorization. The resulting method – named **SNOpt** – converges much faster than first-order baselines in wall-clock time, and the improvement remains consistent across various applications, *e.g.* image classification, generative flow, and time-series prediction. Our framework also enables direct architecture optimization, such as the integration time of Neural ODEs, with second-order feedback policies, strengthening the OC perspective as a principled tool of analyzing optimization in deep learning. Our code is available at https://github.com/ghliu/snopt.

1 Introduction

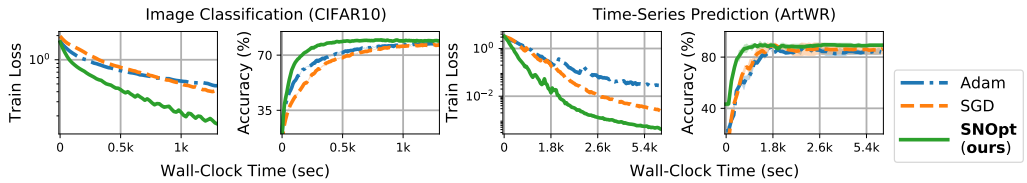

Figure 1: Our second-order method (SNOpt; green solid curves) achieves superior convergence compared to first-order methods (SGD, Adam) on various Neural-ODE applications.

Neural ODEs (Chen et al., 2018) have received tremendous attention over recent years. Inspired by taking the continuous limit of the "discrete" residual transformation, $\mathbf{x}_{k+1} = \mathbf{x}_k + \epsilon F(\mathbf{x}_k, \theta)$, they propose to directly parameterize the vector field of an ODE as a deep neural network (DNN), *i.e.*

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = F(t, \mathbf{x}(t), \theta), \quad \mathbf{x}(t_0) = \boldsymbol{x}_{t_0}, \tag{1}$$

where $\mathbf{x}(t) \in \mathbb{R}^m$ and $F(\cdot, \cdot, \theta)$ is a DNN parameterized by $\theta \in \mathbb{R}^n$. This provides a powerful paradigm connecting modern machine learning to classical differential equations (Weinan, 2017) and has since then achieved promising results on time series analysis (Rubanova et al., 2019; Kidger et al., 2020b), reversible generative flow (Grathwohl et al., 2018; Nguyen et al., 2019), image classification (Zhuang et al., 2020, 2021), and manifold learning (Lou et al., 2020; Mathieu & Nickel, 2020).

Due to the continuous-time representation, Neural ODEs feature a distinct optimization process (see Fig. 2) compared to their discrete-time counterparts, which also poses new challenges. First, the

35th Conference on Neural Information Processing Systems (NeurIPS 2021)

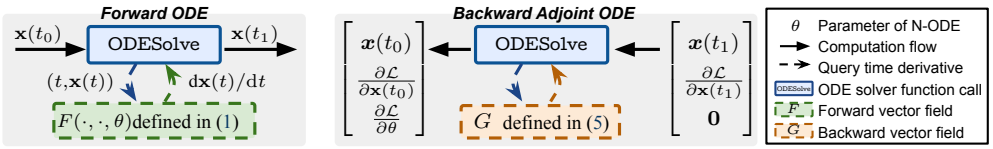

Figure 2: Neural ODE features distinct training process: Both forward and backward passes parameterize vector fields so that any generic ODE solver (which can be non-differentiable) can query time derivatives, *e.g.* $\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t}$, to solve the ODEs (1, 5). In this work, we extend it to second-order training.

forward pass of Neural ODEs involves solving (1) with a black-box ODE solver. Depending on how its numerical integration is set up, the propagation may be refined to arbitrarily small step sizes and become prohibitively expensive to solve without any regularization (Ghosh et al., 2020; Finlay et al., 2020). On the other hand, to prevent Back-propagating through the entire ODE solver, the gradients are typically obtained by solving a *backward adjoint ODE* using the Adjoint Sensitivity Method (ASM; Pontryagin et al. (1962)). While this can be achieved at a favorable $\mathcal{O}(1)$ memory, it further increases the runtime and can suffer from inaccurate integration (Gholami et al., 2019). For these reasons, Neural ODEs often take notoriously longer time to train, limiting their applications to relatively small or synthetic datasets (Massaroli et al., 2020) until very recently (Zhuang et al., 2021).

To improve the convergence rate of training, it is natural to consider higher-order optimization. While efficient second-order methods have been proposed for discrete models (Ba et al., 2016; George et al., 2018), it remains unclear how to extend these successes to Neural ODEs, given their distinct computation processes. Indeed, limited discussions in this regard only note that one may repeat the backward adjoint process recursively to obtain higher-order derivatives (Chen et al., 2018). This is, unfortunately, impractical as the recursion will accumulate the aforementioned integration errors and scale the per-iteration runtime linearly. As such, second-order methods for Neural ODEs are seldom considered in practice, nor have they been rigorously explored from an optimization standpoint.

In this work, we show that efficient second-order optimization is in fact viable for Neural ODEs. Our method is inspired by the emerging Optimal Control perspective (Weinan et al., 2018; Liu & Theodorou, 2019), which treats the parameter θ as a control variable, so that the training process, *i.e.* optimizing θ w.r.t. some objective, can be interpreted as an Optimal Control Programming (OCP). Specifically, we show that a continuous-time OCP methodology, called *Differential Programming*, provides analytic second-order derivatives by solving a set of coupled matrix ODEs. Interestingly, these matrix ODEs can be augmented to the backward adjoint ODE and solved simultaneously. In other words, a single backward pass is sufficient to compute all derivatives, including the original ASM-based gradient, the newly-derived second-order matrices, or even higher-order tensors. Further, these higher-order computations enjoy the same $\mathcal{O}(1)$ memory and a comparable runtime to first-order methods by adopting Kronecker factorization (Martens & Grosse, 2015). The resulting method – called **SNOpt** – admits superior convergence in wall-clock time (Fig. 1), and the improvement remains consistent across image classification, continuous normalizing flow, and time-series prediction.

Our OCP framework also facilitates progressive training of the network architecture. Specifically, we study an example of jointly optimizing the "integration time" of Neural ODEs, in analogy to the "depth" of discrete DNNs. While analytic gradients w.r.t. this architectural parameter have been derived under the ASM framework, they were often evaluated on limited synthetic datasets (Massaroli et al., 2020). In the context of OCP, however, free-horizon optimization is a well-studied problem for practical applications with a priori unknown terminal time (Sun et al., 2015; De Marchi & Gerdts, 2019). In this work, we show that these principles can be applied to Neural ODEs, yielding a novel second-order *feedback policy* that adapts the integration time throughout training. On training CIFAR10, this further leads to a 20% runtime reduction, yet without hindering test-time accuracy.

In summary, we present the following contributions.

- We propose a novel computational framework for computing higher-order derivatives of deep continuous-time models, with a rigorous analysis using continuous-time Optimal Control theory.
- We propose an efficient second-order method, **SNOpt**, that achieves superior convergence (in wall-clock time) over first-order methods in training Neural ODEs, while retaining constant memory complexity. These improvements remain consistent across various applications.
- To show that our framework also enables direct architecture optimization, we derive a second-order feedback policy for adapting the integration horizon and show it further reduces the runtime.

2 Preliminaries

Notation. We use roman and italic type to represent a variable $\mathbf{x}(t)$ and its realization $\boldsymbol{x}(t)$ given an ODE. `ODESolve` denotes a function call that solves an initial value problem given an initial condition, start and end integration time, and vector field, *i.e.* $\texttt{ODESolve}(\mathbf{x}(t_0), t_0, t_1, F)$ where $\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = F$.

Forward and backward computations of Neural ODEs. Given an initial condition $\mathbf{x}(t_0)$ and integration interval $[t_0, t_1]$, Neural ODEs concern the following optimization over an objective $\mathcal{L}$,

$$\min_{\theta} \mathcal{L}(\mathbf{x}(t_1)), \quad \text{where } \mathbf{x}(t_1) = \mathbf{x}(t_0) + \int_{t_0}^{t_1} F(t, \mathbf{x}(t), \theta)\, \mathrm{d}t \tag{2}$$

is the solution of the ODE (1) and can be solved by calling a black-box ODE solver, *i.e.* $\mathbf{x}(t_1) = \texttt{ODESolve}(\mathbf{x}(t_0), t_0, t_1, F)$. The use of `ODESolve` allows us to adopt higher-order numerical methods, *e.g.* adaptive Runge-Kutta (Press et al., 2007), which give more accurate integration compared with *e.g.* vanilla Euler discretization in residual-based discrete models. To obtain the gradient $\frac{\partial \mathcal{L}}{\partial \theta}$ of Neural ODE, one may naively Back-propagate through `ODESolve`. This, even if it could be made possible, leads to unsatisfactory memory complexity since the computation graph can grow arbitrarily large for adaptive ODE solvers. Instead, Chen et al. (2018) proposed to apply the Adjoint Sensitivity Method (ASM), which states that the gradient can be obtained through the following integration.

$$\frac{\partial \mathcal{L}}{\partial \theta} = -\int_{t_1}^{t_0} \mathbf{a}(t)^{\mathsf{T}} \frac{\partial F(t, \mathbf{x}(t), \theta)}{\partial \theta}\, \mathrm{d}t\,, \tag{3}$$

where $\mathbf{a}(t) \in \mathbb{R}^m$ is referred to the *adjoint* state whose dynamics obey a *backward adjoint ODE*,

$$-\frac{\mathrm{d}\mathbf{a}(t)}{\mathrm{d}t} = \mathbf{a}(t)^{\mathsf{T}} \frac{\partial F(t, \mathbf{x}(t), \theta)}{\partial \mathbf{x}(t)}, \quad \mathbf{a}(t_1) = \frac{\partial \mathcal{L}}{\partial \mathbf{x}(t_1)}\,. \tag{4}$$

Equations (3, 4) present two coupled ODEs that can be viewed as the continuous-time expression of the Back-propagation (LeCun et al., 1988). Algorithmically, they can be solved through another call of `ODESolve` (see Fig. 2) with an augmented dynamics G, *i.e.*

$$\begin{bmatrix} \mathbf{x}(t_0) \\ \mathbf{a}(t_0) \\ \partial\mathcal{L}/\partial\theta \end{bmatrix} = \texttt{ODESolve}\left(\begin{bmatrix} \mathbf{x}(t_1) \\ \mathbf{a}(t_1) \\ \mathbf{0} \end{bmatrix}, t_1, t_0, G \right), \text{ where } G\left(t, \begin{bmatrix} \mathbf{x}(t) \\ \mathbf{a}(t) \\ \cdot \end{bmatrix}, \theta \right) := \begin{bmatrix} F(t, \mathbf{x}(t), \theta) \\ -\mathbf{a}(t)^{\mathsf{T}} \frac{\partial F}{\partial \mathbf{x}} \\ -\mathbf{a}(t)^{\mathsf{T}} \frac{\partial F}{\partial \theta} \end{bmatrix} \tag{5}$$

augments the original dynamics F in (1) with the adjoint ODEs (3, 4). Notice that this computation (5) depends only on $(\mathbf{x}(t_1), \mathbf{a}(t_1))$. This differs from naive Back-propagation, which requires storing intermediate states along the entire computation graph of forward `ODESolve`. While the latter requires $\mathcal{O}(\widetilde{T})$ memory cost,[1] the computation in (5) only consumes constant $\mathcal{O}(1)$ memory cost.

Chen et al. (2018) noted that if we further encapsulate (5) by $\frac{\partial}{\partial\theta}\mathcal{L} = \texttt{grad}(\mathcal{L}, \theta)$, one may compute higher-order derivatives by recursively calling $\frac{\partial^n \mathcal{L}}{\partial \theta^n} = \texttt{grad}(\frac{\partial^{n-1}\mathcal{L}}{\partial\theta^{n-1}}, \theta)$, starting from n=1. This can scale unfavorably due to its recursive dependence and accumulated integration errors. Indeed, Table 1 suggests that the errors of second-order derivatives, $\frac{\partial^2\mathcal{L}}{\partial\theta^2}$, obtained from the recursive adjoint procedure can be 2-6 orders of magnitude larger than the ones from the first-order adjoint, $\frac{\partial\mathcal{L}}{\partial\theta}$. In the next section, we will present a novel optimization framework that computes these higher-order derivatives *without* any recursion (Section 3.1) and discuss how it can be implemented efficiently (Section 3.2).

Table 1: Numerical errors between ground-truth and adjoint derivatives using different `ODESolve` on CIFAR10.

	`rk4`	`implicit adams`	`dopri5`
$\frac{\partial\mathcal{L}}{\partial\theta}$	7.63×10^{-5}	2.11×10^{-3}	3.44×10^{-4}
$\frac{\partial^2\mathcal{L}}{\partial\theta^2}$	6.84×10^{-3}	2.50×10^{-1}	41.10

3 Approach

3.1 Dynamics of Higher-order Derivatives using Continuous-time Optimal Control Theory

OCP perspective is a recently emerging methodology for analyzing optimization of discrete DNNs. Central to its interpretation is to treat the layer propagation of a DNN as discrete-time dynamics, so

[1] $\widetilde{T}$ is the number of the adaptive steps used to solve (1), as an analogy of the "depth" of Neural ODEs.

that the training process, *i.e.* finding an *optimal parameter* of *a DNN*, can be understood like an OCP, which searches for an *optimal control* subjected to *a dynamical constraint*. This perspective has provided useful insights on characterizing the optimization process (Hu et al., 2019) and enhancing principled algorithmic design (Liu et al., 2021a). We leave a complete discussion in Appendix A.1.

Lifting this OCP perspective from discrete DNNs to Neural ODEs requires special treatments from continuous-time OCP theory (Todorov, 2016). Nevertheless, we highlight that training Neural ODEs and solving continuous-time OCP are fundamentally intertwined since these models, by construction, represent continuous-time dynamical systems. Indeed, the ASM used for deriving (3, 4) originates from the celebrated Pontryagin's principle (Pontryagin et al., 1962), which is an optimality condition to OCP. Hence, OCP analysis is not only motivated but principled from an optimization standpoint.

We begin by first transforming (2) to a form that is easier to adopt the continuous-time OCP analysis.

$$\min_{\theta}\left[\Phi(\mathbf{x}_{t_1})+\int_{t_0}^{t_1}\ell(t,\mathbf{x}_t,\mathbf{u}_t)\mathrm{d}t\right]\quad \text{subjected to}\ \begin{cases}\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t}=F(t,\mathbf{x}_t,\mathbf{u}_t), & \mathbf{x}_{t_0}=\boldsymbol{x}_{t_0}\\ \frac{\mathrm{d}\mathbf{u}_t}{\mathrm{d}t}=\mathbf{0}, & \mathbf{u}_{t_0}=\theta\end{cases}, \tag{6}$$

where $\mathbf{x}(t)\equiv\mathbf{x}_t$, and etc. It should be clear that (6) describes (2) without loss of generality by having $(\Phi,\ell):=(\mathcal{L},0)$. These functions are known as the terminal and intermediate costs in standard OCP. In training Neural ODEs, ℓ can be used to describe either the weight decay, *i.e.* $\ell\propto\|\mathbf{u}_t\|$, or more complex regularization (Finlay et al., 2020). The time-invariant ODE imposed for $\mathbf{u}_t$ makes the ODE of $\mathbf{x}_t$ equivalent to (1). Problem (6) shall be understood as a particular type of OCP that searches for an optimal initial condition θ of a time-invariant control $\mathbf{u}_t$. Despite seemly superfluous, this is a necessary transformation that enables rigorous OCP analysis for the original training process (2), and it has also appeared in other control-related analyses (Zhong et al., 2020; Chalvidal et al., 2021).

Next, define the accumulated loss from any time $t\in[t_0,t_1]$ to the integration end time t_1 as

$$Q(t,\mathbf{x}_t,\mathbf{u}_t):=\Phi(\mathbf{x}_{t_1})+\int_{t}^{t_1}\ell(\tau,\mathbf{x}_\tau,\mathbf{u}_\tau)\,\mathrm{d}\tau, \tag{7}$$

which is also known in OCP as the *cost-to-go* function. Recall that our goal is to compute higher-order derivatives w.r.t. the parameter θ of Neural ODEs. Under the new OCP representation (6), the first-order derivative $\frac{\partial\mathcal{L}}{\partial\theta}$ is identical to $\frac{\partial Q(t_0,\mathbf{x}_{t_0},\mathbf{u}_{t_0})}{\partial\mathbf{u}_{t_0}}$. This is because $Q(t_0,\mathbf{x}_{t_0},\mathbf{u}_{t_0})$ accumulates all sources of losses between $[t_0,t_1]$ (hence it sufficiently describes $\mathcal{L}$) and $\mathbf{u}_{t_0}=\theta$ by construction. Likewise, the second-order derivatives can be captured by the Hessian $\frac{\partial^2 Q(t_0,\mathbf{x}_{t_0},\mathbf{u}_{t_0})}{\partial\mathbf{u}_{t_0}\partial\mathbf{u}_{t_0}}=\frac{\partial^2\mathcal{L}}{\partial\theta\partial\theta}\equiv\mathcal{L}_{\theta\theta}$. In other words, we are only interested in obtaining the derivatives of Q at the integration start time t_0.

To obtain these derivatives, notice that we can rewrite (7) as

$$0=\ell(t,\mathbf{x}_t,\mathbf{u}_t)+\frac{\mathrm{d}Q(t,\mathbf{x}_t,\mathbf{u}_t)}{\mathrm{d}t},\quad Q(t_1,\mathbf{x}_{t_1})=\Phi(\mathbf{x}_{t_1}), \tag{8}$$

since the definition of Q implies that $Q(t,\mathbf{x}_t,\mathbf{u}_t)=\ell(t,\mathbf{x}_t,\mathbf{u}_t)\mathrm{d}t+Q(t+\mathrm{d}t,\mathbf{x}_{t+\mathrm{d}t},\mathbf{u}_{t+\mathrm{d}t})$. We now state our main result, which provides a local characterization of (8) with a set of coupled ODEs expanded along a solution path. These ODEs can be used to obtain all second-order derivatives at t_0.

Theorem 1 (Second-order Differential Programming)**.** *Consider a solution path $(\bar{\boldsymbol{x}}_t,\bar{\boldsymbol{u}}_t)$ that solves the ODEs in (6). Then the first and second-order derivatives of $Q(t,\mathbf{x}_t,\mathbf{u}_t)$, expanded locally around this solution path, obey the following backward ODEs:*

$$-\frac{\mathrm{d}Q_{\bar{\boldsymbol{x}}}}{\mathrm{d}t}=\ell_{\bar{\boldsymbol{x}}}+F_{\bar{\boldsymbol{x}}}^{\mathsf{T}}Q_{\bar{\boldsymbol{x}}},\qquad -\frac{\mathrm{d}Q_{\bar{\boldsymbol{u}}}}{\mathrm{d}t}=\ell_{\bar{\boldsymbol{u}}}+F_{\bar{\boldsymbol{u}}}^{\mathsf{T}}Q_{\bar{\boldsymbol{x}}}, \tag{9a}$$

$$-\frac{\mathrm{d}Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}}{\mathrm{d}t}=\ell_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}+F_{\bar{\boldsymbol{x}}}^{\mathsf{T}}Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}+Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}F_{\bar{\boldsymbol{x}}},\qquad -\frac{\mathrm{d}Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}}}{\mathrm{d}t}=\ell_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}}+Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}F_{\bar{\boldsymbol{u}}}+F_{\bar{\boldsymbol{x}}}^{\mathsf{T}}Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}}, \tag{9b}$$

$$-\frac{\mathrm{d}Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}}}{\mathrm{d}t}=\ell_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}}+F_{\bar{\boldsymbol{u}}}^{\mathsf{T}}Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}}+Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}}F_{\bar{\boldsymbol{u}}},\qquad -\frac{\mathrm{d}Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}}}{\mathrm{d}t}=\ell_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}}+F_{\bar{\boldsymbol{u}}}^{\mathsf{T}}Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}+Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}}F_{\bar{\boldsymbol{x}}}, \tag{9c}$$

where $F_{\bar{\boldsymbol{x}}}(t)\equiv\frac{\partial F}{\partial\mathbf{x}_t}|_{(\bar{\boldsymbol{x}}_t,\bar{\boldsymbol{u}}_t)}$, $Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}(t)\equiv\frac{\partial^2 Q}{\partial\mathbf{x}_t\partial\mathbf{x}_t}|_{(\bar{\boldsymbol{x}}_t,\bar{\boldsymbol{u}}_t)}$, and etc. All terms in (9) are time-varying vector-valued or matrix-valued functions expanded at $(\bar{\boldsymbol{x}}_t,\bar{\boldsymbol{u}}_t)$. The terminal condition is given by

$$Q_{\bar{\boldsymbol{x}}}(t_1)=\Phi_{\bar{\boldsymbol{x}}},\quad Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}(t_1)=\Phi_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}},\quad \textit{and}\quad Q_{\bar{\boldsymbol{u}}}(t_1)=Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}}(t_1)=Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}}(t_1)=Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}}(t_1)=\mathbf{0}.$$

The proof (see Appendix A.2) relies on rewriting (8) with *differential states*, $\delta \mathbf{x}_t := \mathbf{x}_t - \bar{\boldsymbol{x}}_t$, which view the deviation from $\bar{\boldsymbol{x}}_t$ as an *optimizing variable* (hence the name "*Differential Programming*"). It can be shown that $\delta \mathbf{x}_t$ follows a linear ODE expanded along the solution path. Theorem 1 has several important implications. First, the ODEs in (9a) recover the original ASM computation (3,4), as one can readily verify that $Q_{\bar{\boldsymbol{x}}}(t) \equiv \mathbf{a}(t)$ follows the same backward ODE in (4) and the solution of the second ODE in (9a), $Q_{\bar{\boldsymbol{u}}}(t_0) = -\int_{t_1}^{t_0} F_{\bar{\boldsymbol{u}}}{}^{\mathsf{T}} Q_{\bar{\boldsymbol{x}}} \mathrm{d}t$, gives the exact gradient in (3). Meanwhile, solving the coupled matrix ODEs presented in (9b, 9c) yields the desired second-order matrix, $Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}}(t_0) \equiv \mathcal{L}_{\theta\theta}$, for preconditioning the update. Finally, one can derive the dynamics of other higher-order tensors using the same Differential Programming methodology by simply expanding (8) beyond the second order. We leave some discussions in this regard in Appendix A.2.

3.2 Efficient Second-order Preconditioned Update

Theorem 1 provides an attractive computational framework that does not require recursive computation (as mentioned in Section 2) to obtain higher-order derivatives. It suggests that we can obtain first and second-order derivatives all at once with a single function call of `ODESolve`:

$$\begin{aligned}&[\boldsymbol{x}_{t_0}, Q_{\bar{\boldsymbol{x}}}(t_0), Q_{\bar{\boldsymbol{u}}}(t_0), Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}(t_0), Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}}(t_0), Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}}(t_0), Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}}(t_0)] \\ &\qquad = \texttt{ODESolve}([\boldsymbol{x}_{t_1}, \Phi_{\bar{\boldsymbol{x}}}, \mathbf{0}, \Phi_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}, \mathbf{0}, \mathbf{0}, \mathbf{0}], t_1, t_0, \tilde{G}),\end{aligned} \tag{10}$$

where $\tilde{G}$ augments the original dynamics F in (1) with all 6 ODEs presented in (9). Despite that this OCP-theoretic backward pass (10) retains the same $\mathcal{O}(1)$ memory complexity as in (5), the dimension of the new augmented state, which now carries second-order matrices, can grow to an unfavorable size that dramatically slows down the numerical integration. Hence, we must consider other representations of (9), if any, in order to proceed. In the following proposition, we present one of which that transforms (9) into a set of vector ODEs, so that we can compute them much efficiently.

Proposition 2 (Low-rank representation of (9))**.** *Suppose ℓ:=0 in (6) and let $Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}(t_1) = \sum_{i=1}^{R} \boldsymbol{y}_i \otimes \boldsymbol{y}_i$ be a symmetric matrix of rank $R \leq n$, where $\boldsymbol{y}_i \in \mathbb{R}^m$ and $\otimes$ is the Kronecker product. Then, for all $t \in [t_0, t_1]$, the second-order matrices appeared in (9b, 9c) can be decomposed into*

$$Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}(t) = \sum_{i=1}^{R} \mathbf{q}_i(t) \otimes \mathbf{q}_i(t), \quad Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}}(t) = \sum_{i=1}^{R} \mathbf{q}_i(t) \otimes \mathbf{p}_i(t), \quad Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}}(t) = \sum_{i=1}^{R} \mathbf{p}_i(t) \otimes \mathbf{p}_i(t),$$

where the vectors $\mathbf{q}_i(t) \in \mathbb{R}^m$ and $\mathbf{p}_i(t) \in \mathbb{R}^n$ obey the following backward ODEs:

$$-\frac{\mathrm{d}\mathbf{q}_i(t)}{\mathrm{d}t} = F_{\bar{\boldsymbol{x}}}(t)^{\mathsf{T}} \mathbf{q}_i(t), \quad -\frac{\mathrm{d}\mathbf{p}_i(t)}{\mathrm{d}t} = F_{\bar{\boldsymbol{u}}}(t)^{\mathsf{T}} \mathbf{q}_i(t), \tag{11}$$

with the terminal condition given by $(\mathbf{q}_i(t_1), \mathbf{p}_i(t_1)) := (\boldsymbol{y}_i, \mathbf{0})$.

The proof is left in Appendix A.2. Proposition 2 gives a nontrivial conversion. It indicates that the *coupled matrix* ODEs presented in (9b, 9c) can be disentangled into a set of *independent vector* ODEs where each of them follows its own dynamics (11). As the rank R determines the number of these vector ODEs, this conversion will be particularly useful if the second-order matrices exhibit low-rank structures. Fortunately, this is indeed the case for many Neural-ODE applications which often propagate $\mathbf{x}_t$ in a latent space of higher dimension (Chen et al., 2018; Grathwohl et al., 2018; Kidger et al., 2020b).

Based on Proposition 2, the second-order precondition matrix $\mathcal{L}_{\theta\theta}$ is given by[2]

$$\mathcal{L}_{\theta\theta} \equiv Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}}(t_0) = \sum_{i=1}^{R} \left(\int_{t_1}^{t_0} F_{\bar{\boldsymbol{u}}}{}^{\mathsf{T}} \mathbf{q}_i \, \mathrm{d}t \right) \otimes \left(\int_{t_1}^{t_0} F_{\bar{\boldsymbol{u}}}{}^{\mathsf{T}} \mathbf{q}_i \, \mathrm{d}t \right), \tag{12}$$

where $\mathbf{q}_i \equiv \mathbf{q}_i(t)$ follows (11). Our final step is to facilitate efficient computation of (12) with Kronecker-based factorization, which underlines many popular second-order methods for discrete DNNs (Grosse & Martens, 2016; Martens et al., 2018). Recall that the vector field F is represented

[2] We drop the dependence on t for brevity, yet all terms inside the integrations of (12, 13) are time-varying.

Algorithm 1 SNOpt: Second-order Neural ODE Optimizer

1: **Input:** dataset $\mathcal{D}$, parametrized vector field $F(\cdot,\cdot,\theta)$, integration time $[t_0, t_1]$, black-box ODE solver `ODESolve`, learning rate η, rank R, interval of the time grid Δt
2: **repeat**
3: Solve $\mathbf{x}(t_1) = \texttt{ODESolve}(\mathbf{x}(t_0), t_0, t_1, F)$, where $\mathbf{x}(t_0) \sim \mathcal{D}$. ▷ Forward pass
4: Initialize $(\bar{\boldsymbol{A}}_n, \bar{\boldsymbol{B}}_n) := (\mathbf{0}, \mathbf{0})$ for each layer n and set $\mathbf{q}_i(t_1) := \boldsymbol{y}_i$.
5: **for** t' **in** $\{t_1, t_1 - \Delta t, \cdots, t_0 + \Delta t, t_0\}$ **do**
6: Set $t := t' - \Delta t$ as the small integration step, then call
$[\mathbf{x}(t), Q_{\bar{\boldsymbol{x}}}(t), Q_{\bar{\boldsymbol{u}}}(t), \{\mathbf{q}_i(t)\}_{i=1}^R]$
$= \texttt{ODESolve}([\mathbf{x}(t'), Q_{\bar{\boldsymbol{x}}}(t'), Q_{\bar{\boldsymbol{u}}}(t'), \{\mathbf{q}_i(t')\}_{i=1}^R], t', t, \widehat{G})$, ▷ Backward pass
where $\widehat{G}$ augments the ODEs of state (1), first and second-order derivatives (9a, 11).
7: Evaluate $\mathbf{z}^n(t)$, $\mathbf{h}^n(t)$, $F(t, \mathbf{x}_t, \theta)$, then compute $\boldsymbol{A}_n(t), \boldsymbol{B}_n(t)$ in (13).
8: Update $\bar{\boldsymbol{A}}_n \leftarrow \bar{\boldsymbol{A}}_n + \boldsymbol{A}_n(t) \cdot \Delta t$ and $\bar{\boldsymbol{B}}_n \leftarrow \bar{\boldsymbol{B}}_n + \boldsymbol{B}_n(t) \cdot \Delta t$.
9: **end for**
10: $\forall n$, apply $\theta^n \leftarrow \theta^n - \eta \cdot \mathrm{vec}(\bar{\boldsymbol{B}}_n^{-1} Q_{\bar{\boldsymbol{u}}^n}(t_0) \bar{\boldsymbol{A}}_n^{-\mathsf{T}})$. ▷ Second-order parameter update
11: **until** converges

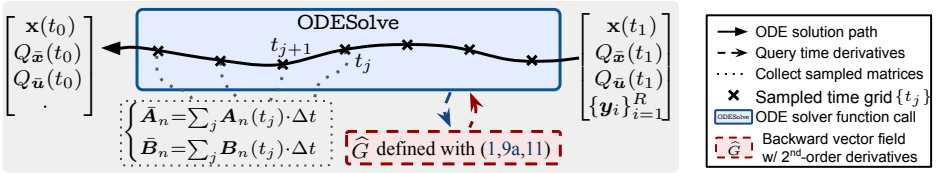

Figure 4: Our second-order method, SNOpt, solves a new backward ODE, *i.e.* the $\widehat{G}$ appeared in line 6 of Alg. 1, which augments second-order derivatives, while simultaneously collecting the matrices $\boldsymbol{A}_n(t_j)$ and $\boldsymbol{B}_n(t_j)$ on a sampled time grid $\{t_j\}$ for computing the preconditioned update in (14).

by a DNN. Let $\mathbf{z}^n(t)$, $\mathbf{h}^n(t)$, and $\mathbf{u}^n(t)$ denote the activation vector, pre-activation vector, and the parameter of layer n when evaluating $\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}$ at time t (see Fig. 3), then the integration in (12) can be broken down into each layer n,

$$\int_{t_1}^{t_0} \left(F_{\bar{\boldsymbol{u}}}^{\mathsf{T}} \mathbf{q}_i\right) \mathrm{d}t = [\cdots, \textstyle\int_{t_1}^{t_0} \left(F_{\bar{\boldsymbol{u}}^n}^{\mathsf{T}} \mathbf{q}_i\right) \mathrm{d}t, \cdots]$$
$$= [\cdots, \textstyle\int_{t_1}^{t_0} \left(\mathbf{z}^n \otimes (\frac{\partial F}{\partial \mathbf{h}^n}^{\mathsf{T}} \mathbf{q}_i)\right) \mathrm{d}t, \cdots],$$

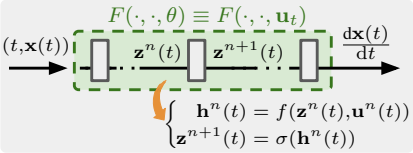

Figure 3: The layer propagation inside the vector field F, where f and σ denote affine and nonlinear activation functions.

where the second equality holds by $F_{\bar{\boldsymbol{u}}^n}^{\mathsf{T}} \mathbf{q}_i = (\frac{\partial F}{\partial \mathbf{h}^n} \frac{\partial \mathbf{h}^n}{\partial \mathbf{u}^n})^{\mathsf{T}} \mathbf{q}_i = \mathbf{z}^n \otimes (\frac{\partial F}{\partial \mathbf{h}^n}^{\mathsf{T}} \mathbf{q}_i)$. This is an essential step towards the Kronecker approximation of the layer-wise precondition matrix:

$$\mathcal{L}_{\theta^n \theta^n} \equiv Q_{\bar{\boldsymbol{u}}^n \bar{\boldsymbol{u}}^n}(t_0) = \sum_{i=1}^{R} \left(\int_{t_1}^{t_0} \left(\mathbf{z}^n \otimes (\tfrac{\partial F}{\partial \mathbf{h}^n}^{\mathsf{T}} \mathbf{q}_i)\right) \mathrm{d}t\right) \otimes \left(\int_{t_1}^{t_0} \left(\mathbf{z}^n \otimes (\tfrac{\partial F}{\partial \mathbf{h}^n}^{\mathsf{T}} \mathbf{q}_i)\right) \mathrm{d}t\right)$$
$$\approx \int_{t_1}^{t_0} \underbrace{(\mathbf{z}^n \otimes \mathbf{z}^n)}_{\boldsymbol{A}_n(t)} \mathrm{d}t \otimes \int_{t_1}^{t_0} \underbrace{\sum_{i=1}^{R} \left((\tfrac{\partial F}{\partial \mathbf{h}^n}^{\mathsf{T}} \mathbf{q}_i) \otimes (\tfrac{\partial F}{\partial \mathbf{h}^n}^{\mathsf{T}} \mathbf{q}_i)\right)}_{\boldsymbol{B}_n(t)} \mathrm{d}t. \tag{13}$$

We discuss the approximation behind (13), and also the one for (14), in Appendix A.2. Note that $\boldsymbol{A}_n(t)$ and $\boldsymbol{B}_n(t)$ are much smaller matrices in $\mathbb{R}^{m \times m}$ compared to the ones in (9), and they can be efficiently computed with automatic differentiation packages (Paszke et al., 2017). Now, let $\{t_j\}$ be a time grid uniformly distributed over $[t_0, t_1]$ so that $\bar{\boldsymbol{A}}_n = \sum_j \boldsymbol{A}_n(t_j)\Delta t$ and $\bar{\boldsymbol{B}}_n = \sum_j \boldsymbol{B}_n(t_j)\Delta t$ approximate the integrations in (13), then our final preconditioned update law is given by

$$\forall n, \quad \mathcal{L}_{\theta^n \theta^n}^{-1} \mathcal{L}_{\theta^n} \approx \mathrm{vec}\left(\bar{\boldsymbol{B}}_n^{-1} Q_{\bar{\boldsymbol{u}}^n}(t_0) \bar{\boldsymbol{A}}_n^{-\mathsf{T}}\right), \tag{14}$$

where vec denotes vectorization. Our second-order method – named **SNOpt** – is summarized in Alg. 1, with the backward computation (*i.e.* line 4-9 in Alg. 1) illustrated in Fig. 4. In practice, we also adopt eigen-based amortization with Tikhonov regularization (George et al. (2018); see Alg. 2 in Appendix A.4), which stabilizes the updates over stochastic training.

Remark. The fact that Proposition 2 holds only for degenerate ℓ can be easily circumvented in practice. As ℓ typically represents weight decay, $\ell := \frac{1}{t_1 - t_0} \|\theta\|_2$, which is time-independent, it can be separated from the backward ODEs (9) and added after solving the backward integration, *i.e.*

$$Q_{\bar{\boldsymbol{u}}}(t_0) \leftarrow \gamma\theta + Q_{\bar{\boldsymbol{u}}}(t_0), \quad Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}}(t_0) \leftarrow \gamma \boldsymbol{I} + Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}}(t_0),$$

where γ is the regularization factor. Finally, we find that using the scaled *Gaussian-Newton* matrix, *i.e.* $Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}(t_1) \approx \frac{1}{t_1 - t_0} \Phi_{\bar{\boldsymbol{x}}} \otimes \Phi_{\bar{\boldsymbol{x}}}$, generally provides a good trade-off between the performance and runtime complexity. As such, we adopt this approximation to Proposition 2 for all experiments.

3.3 Memory Complexity Analysis

Table 2: Memory complexity at different stages of our derivation in terms of $\mathbf{x}_t \in \mathbb{R}^m$, $\theta \in \mathbb{R}^n$, and the rank R. Note that *all* methods have $\mathcal{O}(1)$ in terms of depth.

	Theorem 1 Eqs. (9,10)	Proposition 2 Eqs. (11,12)	**SNOpt** (Alg. 1) Eqs. (13,14)	first-order adjoint Eqs. (3,4)
backward storage	$\mathcal{O}((m+n)^2)$	$\mathcal{O}(Rm + Rn)$	$\mathcal{O}(Rm + 2n)$	$\mathcal{O}(m+n)$
parameter update	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$	$\mathcal{O}(2n)$	$\mathcal{O}(n)$

Table 2 summarizes the memory complexity of different computational methods that appeared along our derivation in Section 3.1 and 3.2. Despite that *all* methods retain $\mathcal{O}(1)$ memory as with the first-order adjoint method, their complexity differs in terms of the state and parameter dimension. Starting from our encouraging result in Theorem 1, which allows one to compute all derivatives with a single backward pass, we first exploit their low-rank representation in Proposition 2. This reduces the storage to $\mathcal{O}(Rm + Rn)$ and paves a way toward adopting Kronecker factorization, which further facilitates efficient preconditioning. With all these, our SNOpt is capable of performing efficient second-order updates while enjoying similar memory complexity (up to some constant) compared to first-order adjoint methods. Lastly, for image applications where Neural ODEs often consist of convolution layers, we adopt convolution-based Kronecker factorization (Grosse & Martens, 2016; Gao et al., 2020), which effectively makes the complexity to scale w.r.t. the number of feature maps (*i.e.* number of channels) rather than the full size of feature maps.

3.4 Extension to Architecture Optimization

Let us discuss an intriguing extension of our OCP framework to optimizing the architecture of Neural ODEs, specifically the integration bound t_1. In practice, when problems contain no prior information on the integration, $[t_0, t_1]$ is typically set to some trivial values (usually $[0, 1]$) without further justification. However, these values can greatly affect both the performance and runtime. Take CIFAR10 for instance (see Fig. 5), the required training time decreases linearly as we drop t_1 from 1, yet the accuracy retains mostly the same unless t_1 becomes too small. Similar results also appear on MNIST (see Fig. 12 in Appendix A.5). In other words, we may interpret the integration bound t_1 as an *architectural parameter* that needs to be jointly optimized during training.

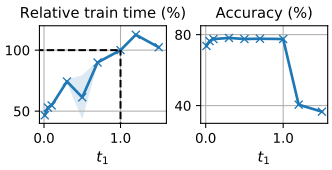

Figure 5: Training performance of CIFAR10 with Adam when using different t_1, which motivates joint optimization of t_1. Experiment setup is left in Appendix A.4.

The aforementioned interpretation fits naturally into our OCP framework. Specifically, we can consider the following extension of Q, which introduces the terminal time T as a new variable:

$$\widetilde{Q}(t, \mathbf{x}_t, \mathbf{u}_t, \mathrm{T}) := \widetilde{\Phi}(\mathrm{T}, \mathbf{x}(\mathrm{T})) + \int_t^{\mathrm{T}} \ell(\tau, \mathbf{x}_\tau, \mathbf{u}_\tau)\, \mathrm{d}\tau, \tag{15}$$

where $\widetilde{\Phi}(\mathrm{T}, \mathbf{x}(\mathrm{T}))$ explicitly imposes the penalty for longer integration time, *e.g.* $\widetilde{\Phi} := \Phi(\mathbf{x}(\mathrm{T})) + \frac{c}{2}\mathrm{T}^2$. Following a similar procedure presented in Section 3.1, we can transform (15) into its ODE form (as in (8)) then characterize its local behavior (as in (9)) along a solution path $(\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t, \bar{T})$. After some tedious derivations, which are left in Appendix A.3, we will arrive at the update rule below,

$$\mathrm{T} \leftarrow \bar{T} - \eta \cdot \delta\mathrm{T}(\delta\theta), \quad \text{where} \quad \delta\mathrm{T}(\delta\theta) = [\widetilde{Q}_{\bar{T}\bar{T}}(t_0)]^{-1} \left(\widetilde{Q}_{\bar{T}}(t_0) + \widetilde{Q}_{\bar{T}\bar{\boldsymbol{u}}}(t_0)\delta\theta \right). \tag{16}$$

Similar to what we have discussed in Section 3.1, one shall view $\widetilde{Q}_{\bar{T}}(t_0) \equiv \frac{\partial \mathcal{L}}{\partial \mathrm{T}}$ as the first-order derivative w.r.t. the terminal time T. Likewise, $\widetilde{Q}_{\bar{T}\bar{T}}(t_0) \equiv \frac{\partial^2 \mathcal{L}}{\partial \mathrm{T} \partial \mathrm{T}}$, and etc. Equation (16) is a second-order *feedback* policy that adjusts its updates based on the change of the parameter θ. Intuitively, it moves in the descending direction of the preconditioned gradient (*i.e.* $\widetilde{Q}_{\bar{T}\bar{T}}^{-1}\widetilde{Q}_{\bar{T}}$), while accounting for the fact that θ *is also progressing during training* (via the feedback $\widetilde{Q}_{\bar{T}\bar{\boldsymbol{u}}}\delta\theta$). The latter is a distinct feature arising from the OCP principle. As we will show later, this update (16) leads to distinct behavior with superior convergence compared to first-order baselines (Massaroli et al., 2020).

4 Experiments

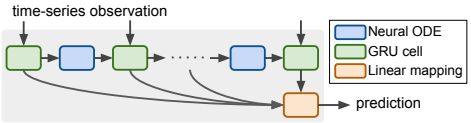

Figure 6: Hybrid model for time-series prediction.

Table 3: Sample size of time-series datasets (input dimension, class label, series length)

SpoAD	ArtWR	CharT
(27, 10, 93)	(19, 25, 144)	(7, 20, 187)

Dataset. We select 9 datasets from 3 distinct applications where N-ODEs have been applied, including image classification (•), time-series prediction (•), and continuous normalizing flow (•; CNF):

- **MNIST**, **SVHN**, **CIFAR10**: MNIST consists of 28×28 gray-scale images, while SVHN and CIFAR10 consist of 3×32×32 colour images. All 3 image datasets have 10 label classes.
- **SpoAD**, **ArtWR**, **CharT**: We consider UEA time series archive (Bagnall et al., 2018). SpokenArabicDigits (SpoAD) is a speech dataset, whereas ArticularyWordRecognition (ArtWR) and CharacterTrajectories (CharT) are motion-related datasets. Table 3 details their sample sizes.
- **Circle**, **Gas**, **Miniboone**: Circle is a 2-dim synthetic dataset adopted from Chen et al. (2018). Gas and Miniboone are 8 and 43-dim tabular datasets commonly used in CNF (Grathwohl et al., 2018; Onken et al., 2020). All 3 datasets transform a multivariate Gaussian to the target distributions.

Models. The models for image datasets and CNF resemble standard feedforward networks, except now consisting of Neural ODEs as continuous transformation layers. Specifically, the models for image classification consist of convolution-based feature extraction, followed by a Neural ODE and linear mapping. Meanwhile, the CNF models are identical to the ones in Grathwohl et al. (2018), which consist of 1-5 Neural ODEs, depending on the size of the dataset. As for the time-series models, we adopt the hybrid models from Rubanova et al. (2019), which consist of a Neural ODE for hidden state propagation, standard recurrent cell (*e.g.* GRU (Cho et al., 2014)) to incorporate incoming time-series observation, and a linear prediction layer. Figure 6 illustrates this process. We detail other configurations in Appendix A.4.

ODE solver. We use standard Runge-Kutta 4(5) adaptive solver (`dopri5`; Dormand & Prince (1980)) implemented by the `torchdiffeq` package. The numerical tolerance is set to 1e-6 for CNF and 1e-3 for the rest. We fix the integration time to $[0, 1]$ whenever it appears as a hyper-parameter (*e.g.* for image and CNF datasets[3]); otherwise we adopt the problem-specific setup (*e.g.* for time series).

Training setup. We consider Adam and SGD (with momentum) as the first-order baselines since they are default training methods for most Neural-ODE applications. As for our second-order SNOpt, we set up the time grid $\{t_j\}$ such that it collects roughly 100 samples along the backward integration to estimate the precondition matrices (see Fig. 4). The hyper-parameters (*e.g.* learning rate) are tuned for each method on each dataset, and we detail the tuning process in Appendix A.4. We also employ practical acceleration techniques, including the $\mathrm{semi\text{-}norm}$ (Kidger et al., 2020a) for speeding up `ODESolve`, and the Jacobian-free estimator (FFJORD; Grathwohl et al. (2018)) for accelerating CNF models. The batch size is set to 256, 512, and 1000 respectively for ArtWord, CharTraj, and Gas. The rest of the datasets use 128 as the batch size. All experiments are conducted on a TITAN RTX.

4.1 Results

Convergence and computation efficiency. Figures 1 and 7 report the training curves of each method measured by wall-clock time. It is obvious that our SNOpt admits a superior convergence

[3] except for Circle where we set $[t_0, t_1]$:=$[0, 10]$ in order to match the original setup in Chen et al. (2018).

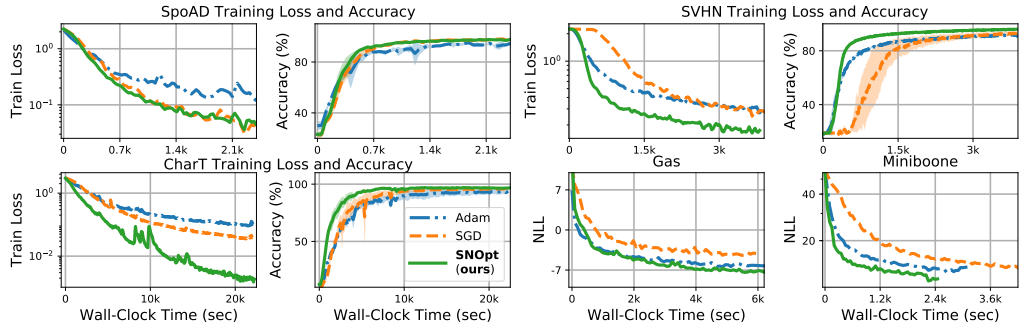

Figure 7: Training performance in *wall-clock* runtime, averaged over 3 trials. Our SNOpt achieves faster convergence against first-order baselines. See Fig. 14 in Appendix A.5 for MNIST and Circle.

Table 4: Test-time performance: accuracies for image and time-series datasets; NLL for CNF datasets

	MNIST	SVHN	CIFAR10	SpoAD	ArtWR	CharT	Circle	Gas	Miniboone
Adam	98.83	91.92	77.41	94.64	84.14	93.29	0.90	-6.42	13.10
SGD	98.68	93.34	76.42	**97.70**	85.82	95.93	0.94	-4.58	13.75
SNOpt	**98.99**	**95.77**	**79.11**	97.41	**90.23**	**96.63**	**0.86**	**-7.55**	**12.50**

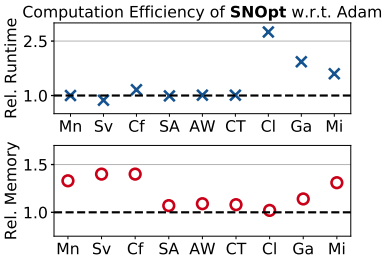

Figure 8: Relative runtime and memory of our SNOpt compared to Adam (denoted by the dashed black lines) on all 9 datasets, where 'Mn' is the shorthand for MNIST, and *etc*.

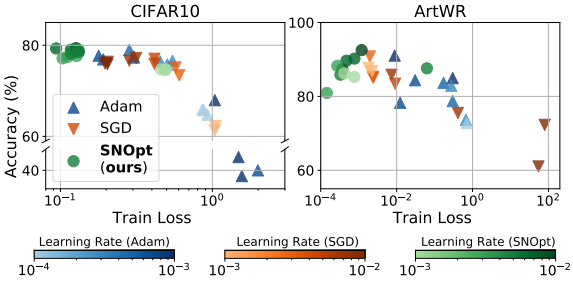

Figure 9: Sensitivity analysis where each sample represents a training result using different optimizer and learning rate (annotated by different symbol and color). Our SNOpt achieves higher accuracies and is insensitive to hyper-parameter changes. Note that x-axes are in log scale.

rate compared to the first-order baselines, and in many cases exceeds their performances by a large margin. In Fig. 8, we report the computation efficiency of our SNOpt compared to Adam on each dataset, and leave their numerical values in Appendix A.4 (Table 9 and 10). For image and time-series datasets (*i.e.* Mn~CT), our SNOpt runs nearly as fast as first-order methods. This is made possible through a rigorous OCP analysis in Section 3, where we showed that second-order matrices can be constructed along with the *same* backward integration when we compute the gradient. Hence, only a minimal overhead is introduced. As for CNF, which propagates the probability density additional to the vanilla state dynamics, our SNOpt is roughly 1.5 to 2.5 times slower, yet it still converges faster in the overall wall-clock time (see Fig. 7). On the other hand, the use of second-order matrices increases the memory consumption of SNOpt by 10-40%, depending on the model and dataset. However, the actual increase in memory (less than 1GB for all datasets; see Table 10) remains affordable on standard GPU machines. More importantly, our SNOpt retains the $\mathcal{O}(1)$ memory throughout training.

Test-time performance and hyper-parameter sensitivity. Table 4 reports the test-time performance, including the accuracies (%) for image and time-series classification, and the negative log-likelihood (NLL) for CNF. On most datasets, our method achieves competitive results against standard baselines. In practice, we also find that using the preconditioned updates greatly reduce the sensitivity to hyper-parameters (*e.g.* learning rate). This is demonstrated in Fig. 9, where we sample distinct learning rates from a proper interval for each method (shown with different color bars) and record their training results after convergence. It is clear that our method not only converges to higher

Table 5: Performance of jointly optimizing the integration bound t_1 on CIFAR10

Method	Train time (%) w.r.t. t_1=1.0	Accuracy (%)
ASM baseline	96	76.61
SNOpt (ours)	**81**	**77.82**

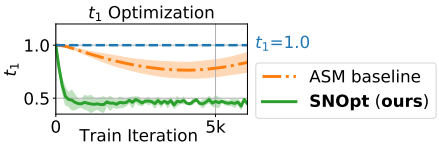

Figure 10: Dynamics of t_1 over CIFAR10 training using different methods.

Table 6: Measure of implicit regularization on SVHN

	# of function evaluation (NFE)	Regularization ($\int \|\nabla_{\mathbf{x}} F\|^2 + \int \|F\|^2$)
Adam	42.1	323.88
SNOpt	**32.6**	**199.1**

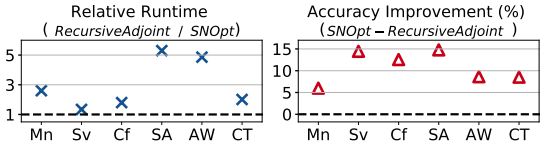

Figure 11: Comparison between SNOpt and second-order recursive adjoint. SNOpt is at least 2 times faster and improves the accuracies of baselines by 5-15%.

accuracies with lower losses, these values are also more concentrated on the plots. In other words, our method achieves better convergence in a more consistent manner across different hyper-parameters.

Joint optimization of the integration bound t_1. Table 5 and Fig. 10 report the performance of optimizing t_1 along with its convergence dynamics. Specifically, we compare our second-order feedback policy (16) derived in Section 3.4 to the first-order ASM baseline proposed in Massaroli et al. (2020). It is clear that our OCP-theoretic method leads to substantially faster convergence, and the optimized t_1 stably hovers around 0.5 without deviation (as appeared for the baseline). This drops the training time by nearly 20% compared to the vanilla training, where we fix t_1 to 1.0, yet without sacrificing the test-time accuracy. A similar experiment for MNIST (see Fig. 13 in Appendix A.5) shows a consistent result. We highlight these improvements as the benefit gained from introducing the well-established OCP principle to these emerging deep continuous-time models.

Comparison with recursive adjoint. Finally, Fig. 11 reports the comparison between our SNOpt and the recursive adjoint baseline (see Section 2 and Table 1). It is clear that our method outperforms this second-order baseline by a large margin in both runtime efficiency and test-time performance. Note that we omit the comparison on CNF datasets since the recursive adjoint simply fails to converge.

Remark (Implicit regularization)**.** In some cases (*e.g.* SVHN in Fig. 8), our method may run slightly faster than first-order methods. This is a distinct phenomenon arising exclusively from training these continuous-time models. Since their forward and backward passes involve solving *parameterized* ODEs (see Fig. 2), the computation graphs are *parameter-dependent*; hence adaptive throughout training. In this vein, we conjecture that the preconditioned updates in these cases may have guided the parameter to regions that are numerically stabler (hence faster) for integration.[4] With this in mind, we report in Table 6 the value of Jacobian, $\int \|\nabla_{\mathbf{x}} F\|^2$, and Kinetic, $\int \|F\|^2$, regularization (Finlay et al., 2020) in SVHN training. Interestingly, the parameter found by our SNOpt indeed has a substantially lower value (hence stronger regularization and better-conditioned ODE dynamics) compared to the one found by Adam. This provides a plausible explanation of the reduction in the NFE when using our method, yet without hindering the test-time performance (see Table 4).

5 Conclusion

We present an efficient higher-order optimization framework for training Neural ODEs. Our method – named **SNOpt** – differs from existing second-order methods in various aspects. While it leverages similar factorization inherited in Kronecker-based methods (Martens & Grosse, 2015), the two methodologies differ fundamentally in that we construct analytic ODE expressions for higher-order derivatives (Theorem 1) and compute them through `ODESolve`. This retains the favorable $\mathcal{O}(1)$ memory as opposed to their $\mathcal{O}(T)$. It also enables a flexible rank-based factorization in Proposition 2. Meanwhile, our method extends the recent trend of OCP-inspired methods (Li et al., 2017; Liu et al., 2021b) to deep continuous-time models, yet using a rather straightforward framework without imposing additional assumptions, such as Markovian or game transformation. To summarize, our work advances several methodologies to the emerging deep continuous-time models, achieving strong empirical results and opening up new opportunities for analyzing models such as Neural SDEs/PDEs.

[4] In Appendix A.4, we provide some theoretical discussions (see Corollary 9) in this regard.

Acknowledgments and Disclosure of Funding

The authors would like to thank Chia-Wen Kuo and Chen-Hsuan Lin for the meticulous proofreading. Guan-Horng Liu was supported by CPS NSF Award #1932068, and Tianrong Chen was supported by ARO Award #W911NF2010151.

References

Almubarak, H., Sadegh, N., and Taylor, D. G. Infinite horizon nonlinear quadratic cost regulator. In *2019 American Control Conference (ACC)*, pp. 5570–5575. IEEE, 2019.

Amari, S.-i. and Nagaoka, H. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

Ba, J., Grosse, R., and Martens, J. Distributed second-order optimization using kronecker-factored approximations. 2016.

Bagnall, A., Dau, H. A., Lines, J., Flynn, M., Large, J., Bostrom, A., Southam, P., and Keogh, E. The uea multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018.

Botev, A., Ritter, H., and Barber, D. Practical gauss-newton optimisation for deep learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 557–565. JMLR. org, 2017.

Chalvidal, M., Ricci, M., VanRullen, R., and Serre, T. Go with the flow: Adaptive control for neural odes. 2021.

Chen, T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, pp. 6572–6583, 2018.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

De Marchi, A. and Gerdts, M. Free finite horizon lqr: a bilevel perspective and its application to model predictive control. *Automatica*, 100:299–311, 2019.

Desjardins, G., Simonyan, K., Pascanu, R., and Kavukcuoglu, K. Natural neural networks. *arXiv preprint arXiv:1507.00210*, 2015.

Dormand, J. R. and Prince, P. J. A family of embedded runge-kutta formulae. *Journal of computational and applied mathematics*, 6(1):19–26, 1980.

Finlay, C., Jacobsen, J.-H., Nurbekyan, L., and Oberman, A. How to train your neural ode: the world of jacobian and kinetic regularization. In *International Conference on Machine Learning*, pp. 3154–3164. PMLR, 2020.

Gao, K.-X., Liu, X.-L., Huang, Z.-H., Wang, M., Wang, Z., Xu, D., and Yu, F. A trace-restricted kronecker-factored approximation to natural gradient. *arXiv preprint arXiv:2011.10741*, 2020.

George, T., Laurent, C., Bouthillier, X., Ballas, N., and Vincent, P. Fast approximate natural gradient descent in a kronecker factored eigenbasis. In *Advances in Neural Information Processing Systems*, pp. 9550–9560, 2018.

Gholami, A., Keutzer, K., and Biros, G. Anode: Unconditionally accurate memory-efficient gradients for neural odes. *arXiv preprint arXiv:1902.10298*, 2019.

Ghosh, A., Behl, H. S., Dupont, E., Torr, P. H., and Namboodiri, V. Steer: Simple temporal regularization for neural odes. *arXiv preprint arXiv:2006.10711*, 2020.

Grathwohl, W., Chen, R. T., Betterncourt, J., Sutskever, I., and Duvenaud, D. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.

Grosse, R. and Martens, J. A kronecker-factored approximate fisher matrix for convolution layers. In *International Conference on Machine Learning*, pp. 573–582, 2016.

Gupta, V., Koren, T., and Singer, Y. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pp. 1842–1850. PMLR, 2018.

Hu, K., Kazeykina, A., and Ren, Z. Mean-field langevin system, optimal control and deep neural networks. *arXiv preprint arXiv:1909.07278*, 2019.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Kelly, J., Bettencourt, J., Johnson, M. J., and Duvenaud, D. Learning differential equations that are easy to solve. *arXiv preprint arXiv:2007.04504*, 2020.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Kidger, P., Chen, R. T., and Lyons, T. " hey, that's not an ode": Faster ode adjoints with 12 lines of code. *arXiv preprint arXiv:2009.09457*, 2020a.

Kidger, P., Morrill, J., Foster, J., and Lyons, T. Neural controlled differential equations for irregular time series. *arXiv preprint arXiv:2005.08926*, 2020b.

Laurent, C., George, T., Bouthillier, X., Ballas, N., and Vincent, P. An evaluation of fisher approximations beyond kronecker factorization. 2018.

LeCun, Y., Touresky, D., Hinton, G., and Sejnowski, T. A theoretical framework for back-propagation. In *Proceedings of the 1988 connectionist models summer school*, volume 1, pp. 21–28. CMU, Pittsburgh, Pa: Morgan Kaufmann, 1988.

Li, Q., Chen, L., Tai, C., and Weinan, E. Maximum principle based algorithms for deep learning. *The Journal of Machine Learning Research*, 18(1):5998–6026, 2017.

Liu, G.-H. and Theodorou, E. A. Deep learning theory review: An optimal control and dynamical systems perspective. *arXiv preprint arXiv:1908.10920*, 2019.

Liu, G.-H., Chen, T., and Theodorou, E. A. Ddpnopt: Differential dynamic programming neural optimizer. In *International Conference on Learning Representations*, 2021a.

Liu, G.-H., Chen, T., and Theodorou, E. A. Dynamic game theoretic neural optimizer. In *International Conference on Machine Learning*, 2021b.

Lou, A., Lim, D., Katsman, I., Huang, L., Jiang, Q., Lim, S.-N., and De Sa, C. Neural manifold ordinary differential equations. *arXiv preprint arXiv:2006.10254*, 2020.

Ma, L., Montague, G., Ye, J., Yao, Z., Gholami, A., Keutzer, K., and Mahoney, M. W. Inefficiency of k-fac for large batch size training. *arXiv preprint arXiv:1903.06237*, 2019.

Martens, J. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*, 2014.

Martens, J. and Grosse, R. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417, 2015.

Martens, J., Ba, J., and Johnson, M. Kronecker-factored curvature approximations for recurrent neural networks. In *International Conference on Learning Representations*, 2018.

Massaroli, S., Poli, M., Park, J., Yamashita, A., and Asama, H. Dissecting neural odes. *arXiv preprint arXiv:2002.08071*, 2020.

Mathieu, E. and Nickel, M. Riemannian continuous normalizing flows. *arXiv preprint arXiv:2006.10605*, 2020.

Nguyen, T. M., Garg, A., Baraniuk, R. G., and Anandkumar, A. Infocnf: An efficient conditional continuous normalizing flow with adaptive solvers. *arXiv preprint arXiv:1912.03978*, 2019.

Onken, D., Fung, S. W., Li, X., and Ruthotto, L. Ot-flow: Fast and accurate continuous normalizing flows via optimal transport. *arXiv preprint arXiv:2006.00104*, 2020.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Pontryagin, L. S., Mishchenko, E., Boltyanskii, V., and Gamkrelidze, R. The mathematical theory of optimal processes. 1962.

Press, W. H., William, H., Teukolsky, S. A., Vetterling, W. T., Saul, A., and Flannery, B. P. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

Rubanova, Y., Chen, R. T., and Duvenaud, D. Latent odes for irregularly-sampled time series. *arXiv preprint arXiv:1907.03907*, 2019.

Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A. How does batch normalization help optimization? *arXiv preprint arXiv:1805.11604*, 2018.

Schacke, K. On the kronecker product. *Master's thesis, University of Waterloo*, 2004.

Sun, W., Theodorou, E., and Tsiotras, P. Model based reinforcement learning with final time horizon optimization. *arXiv preprint arXiv:1509.01186*, 2015.

Tassa, Y., Mansard, N., and Todorov, E. Control-limited differential dynamic programming. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1168–1175. IEEE, 2014.

Theodorou, E., Tassa, Y., and Todorov, E. Stochastic differential dynamic programming. In *Proceedings of the 2010 American Control Conference*, pp. 1125–1132. IEEE, 2010.

Todorov, E. Optimal control theory. *Bayesian brain: probabilistic approaches to neural coding*, pp. 269–298, 2016.

Weinan, E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017.

Weinan, E., Han, J., and Li, Q. A mean-field optimal control formulation of deep learning. *arXiv preprint arXiv:1807.01083*, 2018.

Wu, Y., Zhu, X., Wu, C., Wang, A., and Ge, R. Dissecting hessian: Understanding common structure of hessian in neural networks. *arXiv preprint arXiv:2010.04261*, 2020.

Zhang, G., Martens, J., and Grosse, R. Fast convergence of natural gradient descent for overparameterized neural networks. *arXiv preprint arXiv:1905.10961*, 2019.

Zhong, Y. D., Dey, B., and Chakraborty, A. Symplectic ode-net: Learning hamiltonian dynamics with control. 2020.

Zhuang, J., Dvornek, N., Li, X., Tatikonda, S., Papademetris, X., and Duncan, J. Adaptive checkpoint adjoint method for gradient estimation in neural ode. In *International Conference on Machine Learning*, pp. 11639–11649. PMLR, 2020.

Zhuang, J., Dvornek, N. C., Tatikonda, S., and Duncan, J. S. Mali: A memory efficient and reverse accurate integrator for neural odes. *arXiv preprint arXiv:2102.04668*, 2021.

A Appendix

A.1 Review of Optimal Control Programming (OCP) Perspective of Training Discrete DNNs and Continuous-time OCP

Here, we review the OCP perspective of training discrete DNNs and discuss how the continuous-time OCP can be connected to the training process of Neural ODEs. For a complete treatment, we refer readers to *e.g.* Weinan (2017); Li et al. (2017); Weinan et al. (2018); Liu & Theodorou (2019); Liu et al. (2021a), and their references therein.

Abuse the notation and let the layer propagation rule in standard feedforward DNNs with depth T be

$$\boldsymbol{z}_{t+1} = f(\boldsymbol{z}_t, \boldsymbol{u}_t), \quad t \in \{0, 1, \cdots, T\}. \tag{17}$$

Here, $\boldsymbol{z}_t$ and $\boldsymbol{u}_t$ represent the (vectorized) hidden state and parameter of layer t. For instance, consider the propagation of a fully-connected layer, *i.e.* $\boldsymbol{z}_{t+1} = \sigma(\boldsymbol{W}_t \boldsymbol{z}_t + \boldsymbol{b}_t)$, where $\boldsymbol{W}_t$, $\boldsymbol{b}_t$, and $\sigma(\cdot)$ are respectively the weight, bias, and nonlinear activation function. Then, (17) treats $\boldsymbol{u}_t := \mathrm{vec}([\boldsymbol{W}_t, \boldsymbol{b}_t])$ as the vectorized parameter and f as the composition of $\sigma(\cdot)$ and the affine transformation (Do not confuse with Fig. 3 which denotes f as the affine transformation).

The OCP perspective notices that (17) can also be interpreted as a discrete-time dynamical system that propagates the state $\boldsymbol{z}_t$ with the control variable $\boldsymbol{u}_t$. In this vein, computing the forward pass of a DNN can be seen as propagating a nonlinear dynamical system from time $t = 0$ to T. Furthermore, the training process, *i.e.* finding optimal parameters $\{\boldsymbol{u}_t : \forall t\}$ for all layers, can be seen as a discrete-time Optimal Control Programming (OCP), which searches for an optimal control sequence $\{\boldsymbol{u}_t : \forall t\}$ that minimizes some objective.

In the case of Neural ODEs, the discrete-time layer propagation rule in (17) is replaced with the ODE in (1). However, as we have shown in Section 3.1, the interpretation between the trainable parameter θ and control variable (hence the connection between the training process and OCP) remains valid. In fact, consider the vanilla form of continuous-time OCP,

$$\min_{\mathbf{u}(t):t\in[t_0,t_1]} \left[\Phi(\mathbf{x}_{t_1}) + \int_{t_0}^{t_1} \ell(t, \mathbf{x}_t, \mathbf{u}_t) \mathrm{d}t \right], \quad \dot{\mathbf{x}}_t = F(t, \mathbf{x}_t, \mathbf{u}_t), \quad \mathbf{x}_{t_0} = \boldsymbol{x}_{t_0}, \tag{18}$$

which resembles the one we used in (6) except considering a time-varying control process $\mathbf{u}(t)$. The necessary condition to the programming (18) can be characterized by the celebrated Pontryagin's maximum principle (Pontryagin et al., 1962).

Theorem 3 (Pontryagin's maximum principle)**.** *Let $\mathbf{u}_t^* \equiv \mathbf{u}^*(t)$ be a solution that achieved the minimum of (18). Then, there exists continuous processes, $\mathbf{x}_t^*$ and $\mathbf{a}_t^*$, such that*

$$\dot{\mathbf{x}}_t^* = \nabla_{\mathbf{a}} H\left(t, \mathbf{x}_t^*, \mathbf{a}_t^*, \mathbf{u}_t^*\right) \qquad \mathbf{x}_0^* = \mathbf{x}_0\,, \tag{19a}$$

$$\dot{\mathbf{a}}_t^* = -\nabla_{\mathbf{x}} H\left(t, \mathbf{x}_t^*, \mathbf{a}_t^*, \mathbf{u}_t^*\right), \qquad \mathbf{a}_{t_1}^* = \nabla_{\mathbf{x}} \Phi\left(\mathbf{x}_{t_1}^*\right)\,, \tag{19b}$$

$$H\left(t, \mathbf{x}_t^*, \mathbf{a}_t^*, \mathbf{u}_t^*\right) \leq H\left(t, \mathbf{x}_t^*, \mathbf{a}_t^*, \mathbf{u}_t\right), \qquad \forall \mathbf{u}_t \in \mathbb{R}^m, \quad t \in [t_0, t_1]\,, \tag{19c}$$

where the Hamiltonian function is defined as

$$H\left(t, \mathbf{x}_t, \mathbf{a}_t, \mathbf{u}_t\right) := \mathbf{a}_t \cdot F(t, \mathbf{x}_t, \mathbf{u}_t) + \ell(t, \mathbf{x}_t, \mathbf{u}_t).$$

It can be readily verified that (19b) gives the same backward ODE in (4). In other words, the Adjoint Sensitivity Method used for deriving (3, 4) is a direct consequence arising from the OCP optimization theory. In this work, we provide a full treatment of continuous-time OCP theory and show that it opens up new algorithmic opportunities to higher-order training methods for Neural ODEs.

A.2 Missing Derivations and Discussions in Section 3.1 and 3.2

Proof of Theorem 1. Rewrite the backward ODE of the accumulated loss Q in (8) below

$$0 = \ell(t, \mathbf{x}_t, \mathbf{u}_t) + \frac{\mathrm{d}Q(t, \mathbf{x}_t, \mathbf{u}_t)}{\mathrm{d}t}, \quad Q(t_1, \mathbf{x}_{t_1}) = \Phi(\mathbf{x}_{t_1}).$$

Given a solution path $(\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t)$ of the ODEs in (6), define the differential state and control variables $(\delta\mathbf{x}_t, \delta\mathbf{u}_t)$ by

$$\delta\mathbf{x}_t := \mathbf{x}_t - \bar{\boldsymbol{x}}_t \quad \text{and} \quad \delta\mathbf{u}_t := \mathbf{u}_t - \bar{\boldsymbol{u}}_t.$$

We first perform second-order expansions for ℓ and Q along the solution path, which are given by

$$\ell \approx \ell(t, \bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t) + \ell_{\bar{\boldsymbol{x}}}{}^{\mathsf{T}} \delta \mathbf{x}_t + \ell_{\bar{\boldsymbol{u}}}{}^{\mathsf{T}} \delta \mathbf{u}_t + \frac{1}{2} \begin{bmatrix} \delta \mathbf{x}_t \\ \delta \mathbf{u}_t \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \ell_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} & \ell_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}} \\ \ell_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}} & \ell_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}} \end{bmatrix} \begin{bmatrix} \delta \mathbf{x}_t \\ \delta \mathbf{u}_t \end{bmatrix}, \tag{20a}$$

$$Q \approx Q(t, \bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t) + Q_{\bar{\boldsymbol{x}}}{}^{\mathsf{T}} \delta \mathbf{x}_t + Q_{\bar{\boldsymbol{u}}}{}^{\mathsf{T}} \delta \mathbf{u}_t + \frac{1}{2} \begin{bmatrix} \delta \mathbf{x}_t \\ \delta \mathbf{u}_t \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} & Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}} \\ Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}} & Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}} \end{bmatrix} \begin{bmatrix} \delta \mathbf{x}_t \\ \delta \mathbf{u}_t \end{bmatrix}, \tag{20b}$$

where all derivatives, *i.e.* $\ell_{\bar{\boldsymbol{x}}}, \ell_{\bar{\boldsymbol{u}}}, Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}, Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}}$, and etc, are time-varying. We can thereby obtain the time derivative of the second-order approximated Q in (20b) following standard ordinary calculus.

$$\begin{aligned} \frac{\mathrm{d}Q}{\mathrm{d}t} \approx & \frac{\mathrm{d}Q(t, \bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t)}{\mathrm{d}t} + \left(\frac{\mathrm{d}Q_{\bar{\boldsymbol{x}}}}{\mathrm{d}t}^{\mathsf{T}} \delta \mathbf{x}_t + Q_{\bar{\boldsymbol{x}}}{}^{\mathsf{T}} \frac{\mathrm{d}\delta \mathbf{x}_t}{\mathrm{d}t} \right) + \left(\frac{\mathrm{d}Q_{\bar{\boldsymbol{u}}}}{\mathrm{d}t}^{\mathsf{T}} \delta \mathbf{u}_t + Q_{\bar{\boldsymbol{u}}}{}^{\mathsf{T}} \frac{\mathrm{d}\delta \mathbf{u}_t}{\mathrm{d}t} \right) \\ & + \frac{1}{2} \left(\delta \mathbf{x}_t{}^{\mathsf{T}} \frac{\mathrm{d}Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}}{\mathrm{d}t} \delta \mathbf{x}_t + \frac{\mathrm{d}\delta \mathbf{x}_t}{\mathrm{d}t}^{\mathsf{T}} Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} \delta \mathbf{x}_t + \delta \mathbf{x}_t{}^{\mathsf{T}} Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} \frac{\mathrm{d}\delta \mathbf{x}_t}{\mathrm{d}t} \right) \\ & + \frac{1}{2} \left(\delta \mathbf{u}_t{}^{\mathsf{T}} \frac{\mathrm{d}Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}}}{\mathrm{d}t} \delta \mathbf{u}_t + \frac{\mathrm{d}\delta \mathbf{u}_t}{\mathrm{d}t}^{\mathsf{T}} Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}} \delta \mathbf{u}_t + \delta \mathbf{u}_t{}^{\mathsf{T}} Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}} \frac{\mathrm{d}\delta \mathbf{u}_t}{\mathrm{d}t} \right) \\ & + \frac{1}{2} \left(\delta \mathbf{x}_t{}^{\mathsf{T}} \frac{\mathrm{d}Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}}}{\mathrm{d}t} \delta \mathbf{u}_t + \frac{\mathrm{d}\delta \mathbf{x}_t}{\mathrm{d}t}^{\mathsf{T}} Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}} \delta \mathbf{u}_t + \delta \mathbf{x}_t{}^{\mathsf{T}} Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}} \frac{\mathrm{d}\delta \mathbf{u}_t}{\mathrm{d}t} \right) \\ & + \frac{1}{2} \left(\delta \mathbf{u}_t{}^{\mathsf{T}} \frac{\mathrm{d}Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}}}{\mathrm{d}t} \delta \mathbf{x}_t + \frac{\mathrm{d}\delta \mathbf{u}_t}{\mathrm{d}t}^{\mathsf{T}} Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}} \delta \mathbf{x}_t + \delta \mathbf{u}_t{}^{\mathsf{T}} Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}} \frac{\mathrm{d}\delta \mathbf{x}_t}{\mathrm{d}t} \right). \end{aligned} \tag{21}$$

Next, we need to compute $\frac{\mathrm{d}\delta \mathbf{x}_t}{\mathrm{d}t}$ and $\frac{\mathrm{d}\delta \mathbf{u}_t}{\mathrm{d}t}$, *i.e.* the dynamics of the differential state and control. This can be achieved by linearizing the ODE dynamics along $(\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t)$.

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}(\bar{\boldsymbol{x}}_t + \delta \mathbf{x}_t) = F(t, \bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t) + F_{\bar{\boldsymbol{x}}}(t)^{\mathsf{T}} \delta \mathbf{x}_t + F_{\bar{\boldsymbol{u}}}(t)^{\mathsf{T}} \delta \mathbf{u}_t \ \Rightarrow \frac{\mathrm{d}\delta \mathbf{x}_t}{\mathrm{d}t} = F_{\bar{\boldsymbol{x}}}(t)^{\mathsf{T}} \delta \mathbf{x}_t + F_{\bar{\boldsymbol{u}}}(t)^{\mathsf{T}} \delta \mathbf{u}_t, \\ \frac{\mathrm{d}}{\mathrm{d}t}(\bar{\boldsymbol{u}}_t + \delta \mathbf{u}_t) = \mathbf{0} \ \Rightarrow \frac{\mathrm{d}\delta \mathbf{u}_t}{\mathrm{d}t} = \mathbf{0}, \end{aligned} \tag{22}$$

since $\frac{\mathrm{d}\bar{\boldsymbol{x}}_t}{\mathrm{d}t} = F(t, \bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t)$. Finally, substituting (20a) and (21) back to (8) and replacing all $(\frac{\mathrm{d}\delta \mathbf{x}_t}{\mathrm{d}t}, \frac{\mathrm{d}\delta \mathbf{u}_t}{\mathrm{d}t})$ with (22) yield the following set of backward ODEs.

$$\begin{aligned} -\frac{\mathrm{d}Q_{\bar{\boldsymbol{x}}}}{\mathrm{d}t} &= \ell_{\bar{\boldsymbol{x}}} + F_{\bar{\boldsymbol{x}}}^{\mathsf{T}} Q_{\bar{\boldsymbol{x}}}, & -\frac{\mathrm{d}Q_{\bar{\boldsymbol{u}}}}{\mathrm{d}t} &= \ell_{\bar{\boldsymbol{u}}} + F_{\bar{\boldsymbol{u}}}^{\mathsf{T}} Q_{\bar{\boldsymbol{x}}}, \\ -\frac{\mathrm{d}Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}}{\mathrm{d}t} &= \ell_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} + F_{\bar{\boldsymbol{x}}}^{\mathsf{T}} Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} + Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} F_{\bar{\boldsymbol{x}}}, & -\frac{\mathrm{d}Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}}}{\mathrm{d}t} &= \ell_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}} + Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} F_{\bar{\boldsymbol{u}}} + F_{\bar{\boldsymbol{x}}}^{\mathsf{T}} Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}}, \\ -\frac{\mathrm{d}Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}}}{\mathrm{d}t} &= \ell_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}} + F_{\bar{\boldsymbol{u}}}^{\mathsf{T}} Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}} + Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}} F_{\bar{\boldsymbol{u}}}, & -\frac{\mathrm{d}Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}}}{\mathrm{d}t} &= \ell_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}} + F_{\bar{\boldsymbol{u}}}^{\mathsf{T}} Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} + Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}} F_{\bar{\boldsymbol{x}}}. \end{aligned}$$

□

Remark 4 (Relation to continuous-time OCP algorithm)**.** The proof of Theorem 1 resembles standard derivation of continuous-time Differential Dynamic Programming (DDP), a second-order OCP method that has shown great successes in modern autonomous systems (Tassa et al., 2014). However, our derivation was modified accordingly to account for the particular OCP proposed in (6), which concerns only the initial condition of the time-invariant control. As this equivalently leaves out the "dynamic" aspect of DDP, we shorthand our methodology by *Differential Programming*.

Remark 5 (Computing higher-order derivatives)**.** The proof of Theorem 1 can be summarized by

Step 1. Expand Q and ℓ up to second-order, *i.e.* (20).

Step 2. Derive the dynamics of differential variables. In our case, we consider the linear ODE presented in (22).

Step 3. Substitute the approximations from Step 1 and 2 back to (8), expand all terms using ordinary calculus (21), then collect the dynamics of each derivative.

For higher-order derivatives, we simply need to consider a higher-order expansion of Q and ℓ in Step 1 (see *e.g.* Almubarak et al. (2019) and their reference therein). It is also possible to consider higher-order expression of the linear differential ODEs in Step 2, which may further improve the convergence at the cost of extra overhead (Theodorou et al., 2010).

Remark 6 (Complexity of Remark 5)**.** Let k be the optimization order. Development of higher-order ($k \geq 3$) optimization based on Theorem 1 certainly has few computational obstacles, just like what we have identified and resolved in the case of $k = 2$ (see Section 3.2). In terms of memory, while the number of backward ODEs suggested by Theorem 1 can grow exponentially w.r.t. k, Kelly et al. (2020) has developed an efficient truncated method that reduces the number to $\mathcal{O}(k^2)$ or $\mathcal{O}(k \log k)$. In terms of runtime, analogous to the Kronecker approximation that we use to factorize second-order matrices, Gupta et al. (2018) provided an extension to generic higher-order tensor programming. Hence, it may still be plausible to avoid impractical training.

Proof of Proposition 2. We will proceed the proof by induction. Recall that when ℓ degenerates, the matrix ODEs presented in (9b, 9c) from Theorem 1 take the form,

$$-\frac{\mathrm{d}Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}}{\mathrm{d}t} = F_{\bar{\boldsymbol{x}}}^{\mathsf{T}} Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} + Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} F_{\bar{\boldsymbol{x}}}, \qquad Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}(t_1) = \Phi_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}, \tag{24a}$$

$$-\frac{\mathrm{d}Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}}}{\mathrm{d}t} = F_{\bar{\boldsymbol{u}}}{}^{\mathsf{T}} Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}} + Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}} F_{\bar{\boldsymbol{u}}}, \qquad Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}}(t_1) = \mathbf{0}, \tag{24b}$$

$$-\frac{\mathrm{d}Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}}}{\mathrm{d}t} = Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} F_{\bar{\boldsymbol{u}}} + F_{\bar{\boldsymbol{x}}}{}^{\mathsf{T}} Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}}, \qquad Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}}(t_1) = \mathbf{0}, \tag{24c}$$

where we leave out the ODE of $Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}}$ since $Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}}(t) = Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}}^{\mathsf{T}}(t)$ for all $t \in [t_0, t_1]$.

From (24), it is obvious that the decomposition given in Proposition 2 holds at the terminal stage t_1. Now, suppose it also holds at $t \in (t_0, t_1)$, then the backward dynamics of second-order matrices at this specific time step t, take $\frac{\mathrm{d}Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}(t)}{\mathrm{d}t}$ for instance, become

$$\begin{aligned}
-\frac{\mathrm{d}Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}}{\mathrm{d}t} &= F_{\bar{\boldsymbol{x}}}^{\mathsf{T}} Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} + Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} F_{\bar{\boldsymbol{x}}} \\
&= F_{\bar{\boldsymbol{x}}}^{\mathsf{T}} \left(\sum_{i=1}^{R} \mathbf{q}_i \otimes \mathbf{q}_i \right) + \left(\sum_{i=1}^{R} \mathbf{q}_i \otimes \mathbf{q}_i \right) F_{\bar{\boldsymbol{x}}} \\
&= \sum_{i=1}^{R} \left[\left(F_{\bar{\boldsymbol{x}}}^{\mathsf{T}} \mathbf{q}_i \right) \otimes \mathbf{q}_i + \mathbf{q}_i \otimes \left(F_{\bar{\boldsymbol{x}}}^{\mathsf{T}} \mathbf{q}_i \right) \right],
\end{aligned} \tag{25}$$

where $\mathbf{q}_i \equiv \mathbf{q}_i(t)$ for brevity. On the other hand, the LHS of (25) can be expanded as

$$-\frac{\mathrm{d}Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}}{\mathrm{d}t} = -\frac{\mathrm{d}}{\mathrm{d}t} \left(\sum_{i=1}^{R} \mathbf{q}_i \otimes \mathbf{q}_i \right) = -\sum_{i=1}^{R} \left[\frac{\mathrm{d}\mathbf{q}_i}{\mathrm{d}t} \otimes \mathbf{q}_i + \mathbf{q}_i \otimes \frac{\mathrm{d}\mathbf{q}_i}{\mathrm{d}t} \right], \tag{26}$$

which follows by standard ordinary calculus. Equating (25) and (26) implies that following relation should hold at time t,

$$-\frac{\mathrm{d}\mathbf{q}_i}{\mathrm{d}t} = F_{\bar{\boldsymbol{x}}}^{\mathsf{T}} \mathbf{q}_i,$$

which yields the first ODE appeared in (11). Similarly, we can repeat the same process (25, 26) for the matrices $Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}}$ and $Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}}$. This will give us

$$\begin{aligned}
-\frac{\mathrm{d}Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}}}{\mathrm{d}t} &= F_{\bar{\boldsymbol{u}}}{}^{\mathsf{T}} Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}} + Q_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}} F_{\bar{\boldsymbol{u}}} \\
&\Rightarrow -\sum_{i=1}^{R} \left[\frac{\mathrm{d}\mathbf{p}_i}{\mathrm{d}t} \otimes \mathbf{p}_i + \mathbf{p}_i \otimes \frac{\mathrm{d}\mathbf{p}_i}{\mathrm{d}t} \right] = \sum_{i=1}^{R} \left[\left(F_{\bar{\boldsymbol{u}}}^{\mathsf{T}} \mathbf{q}_i \right) \otimes \mathbf{p}_i + \mathbf{p}_i \otimes \left(F_{\bar{\boldsymbol{u}}}^{\mathsf{T}} \mathbf{q}_i \right) \right] \\
-\frac{\mathrm{d}Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}}}{\mathrm{d}t} &= Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} F_{\bar{\boldsymbol{u}}} + F_{\bar{\boldsymbol{x}}}{}^{\mathsf{T}} Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}} \\
&\Rightarrow -\sum_{i=1}^{R} \left[\frac{\mathrm{d}\mathbf{q}_i}{\mathrm{d}t} \otimes \mathbf{p}_i + \mathbf{q}_i \otimes \frac{\mathrm{d}\mathbf{p}_i}{\mathrm{d}t} \right] = \sum_{i=1}^{R} \left[\left(F_{\bar{\boldsymbol{x}}}^{\mathsf{T}} \mathbf{q}_i \right) \otimes \mathbf{p}_i + \mathbf{q}_i \otimes \left(F_{\bar{\boldsymbol{u}}}^{\mathsf{T}} \mathbf{q}_i \right) \right],
\end{aligned}$$

which implies that following relation should also hold at time t,

$$-\frac{\mathrm{d}\mathbf{p}_i}{\mathrm{d}t} = F_{\bar{\boldsymbol{u}}}^{\mathsf{T}} \mathbf{q}_i.$$

Hence, we conclude the proof. □

Derivation and approximation in (13, 14). We first recall two formulas related to the Kronecker product that will be shown useful in deriving (13, 14).

$$(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D})^{\mathsf{T}} = \boldsymbol{A}\boldsymbol{C}^{\mathsf{T}} \otimes \boldsymbol{B}\boldsymbol{D}^{\mathsf{T}}, \tag{27}$$

$$(\boldsymbol{A} \otimes \boldsymbol{B})^{-1}\mathrm{vec}(\boldsymbol{W}) = \mathrm{vec}(\boldsymbol{B}^{-1}\boldsymbol{W}\boldsymbol{A}^{-\mathsf{T}}), \tag{28}$$

where $\boldsymbol{W} \in \mathbb{R}^{l \times p}$, $\boldsymbol{A}, \boldsymbol{C} \in \mathbb{R}^{p \times p}$, and $\boldsymbol{B}, \boldsymbol{D} \in \mathbb{R}^{l \times l}$. Further, $\boldsymbol{A}, \boldsymbol{B}$ are invertible.

Now, we provide a step-by-step derivation of (13). For brevity, we will denote $\mathbf{g}_i^n \equiv \frac{\partial F}{\partial \mathbf{h}^n}^{\mathsf{T}} \mathbf{q}_i$.

$$\begin{aligned}
\mathcal{L}_{\theta^n\theta^n} \equiv Q_{\bar{\boldsymbol{u}}^n\bar{\boldsymbol{u}}^n}(t_0) &= \sum_{i=1}^{R} \left(\int_{t_1}^{t_0} (\mathbf{z}^n \otimes \mathbf{g}_i^n)\, \mathrm{d}t \right) \left(\int_{t_1}^{t_0} (\mathbf{z}^n \otimes \mathbf{g}_i^n)\, \mathrm{d}t \right)^{\mathsf{T}} \\
&\approx \sum_{i=1}^{R} \int_{t_1}^{t_0} (\mathbf{z}^n \otimes \mathbf{g}_i^n)\, (\mathbf{z}^n \otimes \mathbf{g}_i^n)^{\mathsf{T}}\, \mathrm{d}t \\
&= \sum_{i=1}^{R} \int_{t_1}^{t_0} \left(\mathbf{z}^n \mathbf{z}^{n\mathsf{T}}\right) \otimes \left(\mathbf{g}_i^n \mathbf{g}_i^{n\mathsf{T}}\right) \mathrm{d}t && \text{by (27)} \\
&\approx \sum_{i=1}^{R} \int_{t_1}^{t_0} \left(\mathbf{z}^n \mathbf{z}^{n\mathsf{T}}\right) \mathrm{d}t \otimes \int_{t_1}^{t_0} \left(\mathbf{g}_i^n \mathbf{g}_i^{n\mathsf{T}}\right) \mathrm{d}t \\
&= \int_{t_1}^{t_0} (\mathbf{z}^n \otimes \mathbf{z}^n)\, \mathrm{d}t \otimes \int_{t_1}^{t_0} \sum_{i=1}^{R} \mathbf{g}_i^n \otimes \mathbf{g}_i^n \mathrm{d}t. && \text{by Fubini's Theorem}
\end{aligned}$$

There are two approximations in the above derivation. The first one assumes that the contributions of the quantity "$\mathbf{z}^n(t) \otimes \mathbf{g}_i^n(t)$" are uncorrelated across time, whereas the second one assumes $\mathbf{z}^n$ and $\mathbf{g}_i^n$ are pair-wise independents. We stress that both are widely adopted assumptions for deriving practical Kronecker-based methods (Grosse & Martens, 2016; Martens et al., 2018). While the first assumption can be rather strong, the second approximation has been verified in some empirical study (Wu et al., 2020) and can be made exact under certain conditions (Martens & Grosse, 2015). Finally, (14) follows readily from (28) by noticing that $\mathcal{L}_{\theta^n\theta^n} = \bar{\boldsymbol{A}}_n \otimes \bar{\boldsymbol{B}}_n$ under our computation.

Remark 7 (Uncorrelated assumption of $\mathbf{z}^n \otimes \mathbf{g}_i^n$)**.** This assumption is indeed strong yet almost necessary to yield tractable Kronecker matrices for efficient second-order operation. Tracing back to the development of Kronecker-based methods, similar assumptions also appear in convolution layers (*e.g.* uncorrelated between spatial-wise derivatives (Grosse & Martens, 2016)) and recurrent units (*e.g.* uncorrelated between temporal-wise derivatives (Martens et al., 2018)). The latter may be thought of as the discretization of Neural ODEs. We note, however, that it is possible to relax this assumption by considering tractable graphical models (*e.g.* linear Gaussian (Martens et al., 2018)) at the cost of 2-3 times more operations per iteration. In terms of the performance difference, perhaps surprisingly, adopting tractable temporal models provides only minor improvement in test-time performance (see Fig. 4 in Martens et al. (2018)). In some cases, it has been empirically observed that methods adopting the uncorrelated assumption yields better performance (Laurent et al., 2018).

Remark 8 (Relation to Fisher Information Matrix)**.** Recall that for all experiments we apply Gaussian-Newton approximation to the terminal Hessian $Q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}(t_1)$. This specific choice is partially based on empirical performance and computational purpose, yet it turns out that the resulting precondition matrices (12, 13) can be interpreted as Fisher information matrix (FIM). In other words, under this specific setup, (12, 13) can be equivalently viewed as the FIM of Neural ODEs. This implies SNOpt may be thought of as following Natural Gradient Descent (NGD), which is well-known for taking the steepest descent direction in the space of model distributions (Amari & Nagaoka, 2000; Martens, 2014). Indeed, it has been observed that NGD-based methods converge to equally good accuracies, even though its learning rate varies across 1-2 orders (see Fig 10 in Ma et al. (2019) and Fig 4 in George et al. (2018)). These observations coincide with our results (Fig. 9) for Neural ODEs.

A.3 Discussion on the Free-Horizon Optimization in Section 3.4

Derivation of (16). Here we present an extension of our OCP framework to jointly optimizing the architecture of Neural ODEs, specifically the integration bound t_1. The proceeding derivation, despite being rather tedious, follows a similar procedure in Section 3.1 and the proof of Theorem 1.

Recall the modified cost-to-go function that we consider for free-horizon optimization,

$$\widetilde{Q}(t, \mathbf{x}_t, \mathbf{u}_t, \mathrm{T}) := \widetilde{\Phi}(\mathrm{T}, \mathbf{x}(\mathrm{T})) + \int_t^{\mathrm{T}} \ell(\tau, \mathbf{x}_\tau, \mathbf{u}_\tau)\, \mathrm{d}\tau,$$

where we introduce a new variable, *i.e.* the terminal horizon T, that shall be jointly optimized. We use the expression $\mathbf{x}(\mathrm{T})$ to highlight the fact that the terminal state is now a function of T.

Similar to what we have explored in Section 3.1, our goal is to derive an analytic expression for the derivatives of $\widetilde{Q}$ at the integration start time t_0 w.r.t. this new variable T. This can be achieved by characterizing the local behavior of the following ODE,

$$0 = \ell(t, \mathbf{x}_t, \mathbf{u}_t) + \frac{\mathrm{d}\widetilde{Q}(t, \mathbf{x}_t, \mathbf{u}_t, \mathrm{T})}{\mathrm{d}t}, \quad \widetilde{Q}(\mathrm{T}, \mathbf{x}_\mathrm{T}) = \widetilde{\Phi}(\mathrm{T}, \mathbf{x}(\mathrm{T})), \tag{29}$$

expanded on some nominal solution path $(\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t, \bar{T})$.

Let us start from the terminal condition in (29). Given $\widetilde{Q}(\bar{T}, \bar{\boldsymbol{x}}_{\bar{T}}) = \widetilde{\Phi}(\bar{T}, \bar{\boldsymbol{x}}(\bar{T}))$, perturbing the terminal horizon $\bar{T}$ by an infinitesimal amount $\delta\mathrm{T}$ yields

$$\widetilde{Q}(\bar{T} + \delta\mathrm{T}, \bar{\boldsymbol{x}}_{\bar{T}+\delta\mathrm{T}}) = \ell(\bar{\boldsymbol{x}}_{\bar{T}}, \bar{\boldsymbol{u}}_{\bar{T}})\delta\mathrm{T} + \widetilde{\Phi}(\bar{T} + \delta\mathrm{T}, \bar{\boldsymbol{x}}(\bar{T} + \delta\mathrm{T})). \tag{30}$$

It can be shown that the second-order expansion of the last term in (30) takes the form,

$$\begin{aligned}\widetilde{\Phi}\left(\bar{T} + \delta\mathrm{T}, \bar{\boldsymbol{x}}(\bar{T} + \delta\mathrm{T})\right) \approx \widetilde{\Phi}\left(\bar{T}, \bar{\boldsymbol{x}}(\bar{T})\right) &+ \widetilde{\Phi}_{\bar{\boldsymbol{x}}}^\mathsf{T}\delta\mathbf{x}_{\bar{T}} + \left(\widetilde{\Phi}_{\bar{T}} + \widetilde{\Phi}_{\bar{\boldsymbol{x}}}^\mathsf{T}\bar{F}\right)\delta\mathrm{T} + \frac{1}{2}\delta\mathbf{x}_{\bar{T}}^\mathsf{T}\widetilde{\Phi}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}\delta\mathbf{x}_{\bar{T}} \\ &+ \frac{1}{2}\delta\mathbf{x}_{\bar{T}}^\mathsf{T}\left(\widetilde{\Phi}_{\bar{\boldsymbol{x}}\bar{T}} + \widetilde{\Phi}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}\bar{F}\right)\delta\mathrm{T} + \frac{1}{2}\delta\mathrm{T}\left(\widetilde{\Phi}_{\bar{T}\bar{\boldsymbol{x}}} + \bar{F}^\mathsf{T}\widetilde{\Phi}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}\right)\delta\mathbf{x}_{\bar{T}} \\ &+ \frac{1}{2}\delta\mathrm{T}\left(\widetilde{\Phi}_{\bar{T}\bar{T}} + \widetilde{\Phi}_{\bar{T}\bar{\boldsymbol{x}}}\bar{F} + \bar{F}^\mathsf{T}\widetilde{\Phi}_{\bar{\boldsymbol{x}}\bar{T}} + \bar{F}^\mathsf{T}\widetilde{\Phi}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}\bar{F}\right)\delta\mathrm{T},\end{aligned} \tag{31}$$

which relies on the fact that the following formula holds for any generic function that takes t and $\mathbf{x}(t)$ as its arguments:

$$\frac{\mathrm{d}}{\mathrm{d}t}(\cdot) = \frac{\partial}{\partial t}(\cdot) + \frac{\partial}{\partial \mathbf{x}}(\cdot)^\mathsf{T}\bar{F}, \quad \text{where } \bar{F} = F(t, \bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t).$$

Substituting (31) to (30) gives us the local expressions of the terminal condition up to second-order,

$$\widetilde{Q}_{\bar{\boldsymbol{x}}}(\bar{T}) = \widetilde{\Phi}_{\bar{\boldsymbol{x}}}, \qquad \widetilde{Q}_{\bar{T}}(\bar{T}) = \ell(\bar{\boldsymbol{x}}_{\bar{T}}, \bar{\boldsymbol{u}}_{\bar{T}}) + \widetilde{\Phi}_{\bar{T}} + \widetilde{\Phi}_{\bar{\boldsymbol{x}}}^\mathsf{T}\bar{F}, \tag{32a}$$

$$\widetilde{Q}_{\bar{T}\bar{\boldsymbol{x}}}(\bar{T}) = \widetilde{\Phi}_{\bar{T}\bar{\boldsymbol{x}}} + \bar{F}^\mathsf{T}\widetilde{\Phi}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}, \quad \widetilde{Q}_{\bar{\boldsymbol{x}}\bar{T}}(\bar{T}) = \widetilde{\Phi}_{\bar{\boldsymbol{x}}\bar{T}} + \widetilde{\Phi}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}\bar{F}, \tag{32b}$$

$$\widetilde{Q}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}(\bar{T}) = \widetilde{\Phi}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}, \qquad \widetilde{Q}_{\bar{T}\bar{T}}(\bar{T}) = \widetilde{\Phi}_{\bar{T}\bar{T}} + \widetilde{\Phi}_{\bar{T}\bar{\boldsymbol{x}}}^\mathsf{T}\bar{F} + \bar{F}^\mathsf{T}\widetilde{\Phi}_{\bar{\boldsymbol{x}}\bar{T}} + \bar{F}^\mathsf{T}\widetilde{\Phi}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}\bar{F}, \tag{32c}$$

where $\widetilde{Q}_{\bar{\boldsymbol{x}}}(\bar{T}) \equiv \frac{\delta\widetilde{Q}}{\delta\mathbf{x}_{\bar{T}}} = \frac{\widetilde{Q}(\bar{T}+\delta\mathrm{T}, \bar{\boldsymbol{x}}_{\bar{T}+\delta\mathrm{T}}) - \widetilde{Q}(\bar{T}, \bar{\boldsymbol{x}}_{\bar{T}})}{\delta\mathbf{x}_{\bar{T}}}$, and etc.

Next, consider the ODE dynamics in (29). Similar to (20b), we can expand $\widetilde{Q}$ w.r.t. all optimizing variables, *i.e.* $(\mathbf{x}_t, \mathbf{u}_t, \mathrm{T})$, up to second-order. In this case, the approximation is given by

$$\widetilde{Q}(t, \bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t, \bar{T}) + \widetilde{Q}_{\bar{\boldsymbol{x}}}^\mathsf{T}\delta\mathbf{x}_t + \widetilde{Q}_{\bar{\boldsymbol{u}}}^\mathsf{T}\delta\mathbf{u}_t + \widetilde{Q}_{\bar{T}}\delta\mathrm{T} + \frac{1}{2}\begin{bmatrix}\delta\mathbf{x}_t \\ \delta\mathbf{u}_t \\ \delta\mathrm{T}\end{bmatrix}^\mathsf{T}\begin{bmatrix}\widetilde{Q}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} & \widetilde{Q}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}} & \widetilde{Q}_{\bar{\boldsymbol{x}}\bar{T}} \\ \widetilde{Q}_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}} & \widetilde{Q}_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}} & \widetilde{Q}_{\bar{\boldsymbol{u}}\bar{T}} \\ \widetilde{Q}_{\bar{T}\bar{\boldsymbol{x}}} & \widetilde{Q}_{\bar{T}\bar{\boldsymbol{u}}} & \widetilde{Q}_{\bar{T}\bar{T}}\end{bmatrix}\begin{bmatrix}\delta\mathbf{x}_t \\ \delta\mathbf{u}_t \\ \delta\mathrm{T}\end{bmatrix}, \tag{33}$$

which shares the same form as (20b) except having additional terms that account for the derivatives related to T (marked as green). Substitute (33) to the ODE dynamics in (29), then expand the time derivatives $\frac{\mathrm{d}}{\mathrm{d}t}$ as in (21), and finally replace $\frac{\mathrm{d}\delta\mathbf{x}_t}{\mathrm{d}t}$, $\frac{\mathrm{d}\delta\mathbf{u}_t}{\mathrm{d}t}$, and $\frac{\mathrm{d}\delta\mathrm{T}}{\mathrm{d}t}$ with

$$\frac{\mathrm{d}\delta\mathbf{x}_t}{\mathrm{d}t} = F_{\bar{\boldsymbol{x}}}^\mathsf{T}\delta\mathbf{x}_t + F_{\bar{\boldsymbol{u}}}^\mathsf{T}\delta\mathbf{u}_t, \quad \frac{\mathrm{d}\delta\mathbf{u}_t}{\mathrm{d}t} = \mathbf{0}, \quad \text{and } \frac{\mathrm{d}\delta\mathrm{T}}{\mathrm{d}t} = 0.$$

Then, it can be shown that the first and second-order derivatives of $\widetilde{Q}$ w.r.t. T obey the following backward ODEs:

$$-\frac{\mathrm{d}\widetilde{Q}_{\bar{T}}}{\mathrm{d}t} = 0, \quad -\frac{\mathrm{d}\widetilde{Q}_{\bar{T}\bar{T}}}{\mathrm{d}t} = 0, \quad -\frac{\mathrm{d}\widetilde{Q}_{\bar{T}\bar{\boldsymbol{x}}}}{\mathrm{d}t} = \widetilde{Q}_{\bar{T}\bar{\boldsymbol{x}}} F_{\bar{\boldsymbol{x}}}, \quad -\frac{\mathrm{d}\widetilde{Q}_{\bar{T}\bar{\boldsymbol{u}}}}{\mathrm{d}t} = \widetilde{Q}_{\bar{T}\bar{\boldsymbol{x}}} F_{\bar{\boldsymbol{u}}},$$

with the terminal condition given by (32). As for the derivatives that do not involve T, *e.g.* $\widetilde{Q}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}$ and $\widetilde{Q}_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}}$, one can verify that they follow the same backward structures given in (9) except changing the terminal condition from Φ to $\widetilde{\Phi}$.

To summarize, solving the following ODEs gives us the derivatives of $\widetilde{Q}$ related to T at t_0:

$$\boxed{\begin{aligned}
&-\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{Q}_{\bar{T}}(t) = 0, && \widetilde{Q}_{\bar{T}}(\bar{T}) = \ell(\bar{\boldsymbol{x}}_{\bar{T}}, \bar{\boldsymbol{u}}_{\bar{T}}) + \widetilde{\Phi}_{\bar{T}} + \widetilde{\Phi}_{\bar{\boldsymbol{x}}}^{\mathsf{T}} \bar{F} && \text{(34a)}\\
&-\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{Q}_{\bar{T}\bar{T}}(t) = 0, && \widetilde{Q}_{\bar{T}\bar{T}}(\bar{T}) = \widetilde{\Phi}_{\bar{T}\bar{T}} + \widetilde{\Phi}_{\bar{T}\bar{\boldsymbol{x}}}^{\mathsf{T}} \bar{F} + \bar{F}^{\mathsf{T}} \widetilde{\Phi}_{\bar{\boldsymbol{x}}\bar{T}} + \bar{F}^{\mathsf{T}} \widetilde{\Phi}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} \bar{F} && \text{(34b)}\\
&-\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{Q}_{\bar{T}\bar{\boldsymbol{x}}}(t) = \widetilde{Q}_{\bar{T}\bar{\boldsymbol{x}}} F_{\bar{\boldsymbol{x}}}, && \widetilde{Q}_{\bar{T}\bar{\boldsymbol{x}}}(\bar{T}) = \widetilde{\Phi}_{\bar{T}\bar{\boldsymbol{x}}} + \bar{F}^{\mathsf{T}} \widetilde{\Phi}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} && \text{(34c)}\\
&-\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{Q}_{\bar{T}\bar{\boldsymbol{u}}}(t) = \widetilde{Q}_{\bar{T}\bar{\boldsymbol{x}}} F_{\bar{\boldsymbol{u}}}, && \widetilde{Q}_{\bar{T}\bar{\boldsymbol{u}}}(\bar{T}) = \mathbf{0} && \text{(34d)}
\end{aligned}}$$

Then, we can consider the following quadratic programming for the optimal perturbation $\delta\mathrm{T}^*$,

$$\begin{aligned}
\min_{\delta\mathrm{T}} \; & Q_{\bar{\boldsymbol{x}}}(t_0)^{\mathsf{T}} \delta\mathbf{x}_{t_0} + Q_{\bar{\boldsymbol{u}}}(t_0)^{\mathsf{T}} \delta\mathbf{u}_{t_0} + \widetilde{Q}_{\bar{T}}(t_0)\delta\mathrm{T} \\
& + \frac{1}{2} \begin{bmatrix} \delta\mathbf{x}_{t_0} \\ \delta\mathbf{u}_{t_0} \\ \delta\mathrm{T} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \widetilde{Q}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}}(t_0) & \widetilde{Q}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{u}}}(t_0) & \widetilde{Q}_{\bar{\boldsymbol{x}}\bar{T}}(t_0) \\ \widetilde{Q}_{\bar{\boldsymbol{u}}\bar{\boldsymbol{x}}}(t_0) & \widetilde{Q}_{\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}}(t_0) & \widetilde{Q}_{\bar{\boldsymbol{u}}\bar{T}}(t_0) \\ \widetilde{Q}_{\bar{T}\bar{\boldsymbol{x}}}(t_0) & \widetilde{Q}_{\bar{T}\bar{\boldsymbol{u}}}(t_0) & \widetilde{Q}_{\bar{T}\bar{T}}(t_0) \end{bmatrix} \begin{bmatrix} \delta\mathbf{x}_{t_0} \\ \delta\mathbf{u}_{t_0} \\ \delta\mathrm{T} \end{bmatrix},
\end{aligned}$$

which has an analytic feedback solution given by

$$\boxed{\delta\mathrm{T}^*(\delta\mathbf{x}_{t_0}, \delta\mathbf{u}_{t_0}) = [\widetilde{Q}_{\bar{T}\bar{T}}(t_0)]^{-1} \left(\widetilde{Q}_{\bar{T}}(t_0) + \widetilde{Q}_{\bar{T}\bar{\boldsymbol{x}}}(t_0)\delta\mathbf{x}_{t_0} + \widetilde{Q}_{\bar{T}\bar{\boldsymbol{u}}}(t_0)\delta\mathbf{u}_{t_0} \right).}$$

In practice, we drop the state differential $\delta\mathbf{x}_{t_0}$ and only keep the control differential $\delta\mathbf{u}_{t_0}$, which can be viewed as the parameter update $\delta\theta$ by recalling (6). With these, we arrive at the second-order feedback policy presented in (16).

Practical implementation. We consider a vanilla quadratic cost, $\widetilde{\Phi}(\mathrm{T}, \mathbf{x}(\mathrm{T})) := \Phi(\mathbf{x}(\mathrm{T})) + \frac{c}{2}\mathrm{T}^2$, which penalizes longer integration time, and impose Gaussian-Newton approximation for the terminal cost, *i.e.* $\Phi_{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}} \approx \Phi_{\bar{\boldsymbol{x}}}\Phi_{\bar{\boldsymbol{x}}}^{\mathsf{T}}$. With these, the terminal conditions in (34) can be simplified to

$$\widetilde{Q}_{\bar{T}}(\bar{T}) = c\bar{T} + \Phi_{\bar{\boldsymbol{x}}}^{\mathsf{T}}\bar{F}, \quad \widetilde{Q}_{\bar{T}\bar{T}}(\bar{T}) = c + \left(\Phi_{\bar{\boldsymbol{x}}}^{\mathsf{T}}\bar{F}\right)^2, \quad \widetilde{Q}_{\bar{T}\bar{\boldsymbol{x}}}(\bar{T}) = \left(\Phi_{\bar{\boldsymbol{x}}}^{\mathsf{T}}\bar{F}\right)\Phi_{\bar{\boldsymbol{x}}}^{\mathsf{T}}.$$

Since $\widetilde{Q}_{\bar{T}}(t)$ and $\widetilde{Q}_{\bar{T}\bar{T}}(t)$ are time-invariant (see (34a, 34b)), we know the values of $\widetilde{Q}_{\bar{T}}(t_0)$ and $\widetilde{Q}_{\bar{T}\bar{T}}(t_0)$ at the terminal stage. Further, one can verify that $\forall t \in [t_0, \bar{T}], \quad \widetilde{Q}_{\bar{T}\bar{\boldsymbol{u}}}(t) = \left(\Phi_{\bar{\boldsymbol{x}}}^{\mathsf{T}}\bar{F}\right) Q_{\bar{\boldsymbol{u}}}(t)^{\mathsf{T}}$. In other words, the feedback term $\widetilde{Q}_{\bar{T}\bar{\boldsymbol{u}}}$ simply rescales the first-order derivative $\widetilde{Q}_{\bar{\boldsymbol{u}}}$ by $\Phi_{\bar{\boldsymbol{x}}}^{\mathsf{T}}\bar{F}$. These reasonings suggest that we can evaluate the second-order feedback policy (16) almost at no cost without augmenting any additional state to `ODESolve`. Finally, to adopt the stochastic training, we keep the moving averages of all terms and update T with (16) every 50-100 training iterations.

A.4 Experiment Details

All experiments are conducted on the same GPU machine (TITAN RTX) and implemented with `pytorch`. Below we provide full discussions on topics that are deferred from Section 4.

Model configuration. Here, we specify the model for each dataset. We will adopt the following syntax to describe the layer configuration.

- `Linear(input_dim, output_dim)`
- `Conv(output_channel, kernel, stride)`

Table 7: Configuration of the vector field $F(t, \mathbf{x}_t, \theta)$ of Neural ODEs used for each dataset ([‡]MIT License; [§]Apache License)

Dataset	DNN architecture as $F(t, \mathbf{x}_t, \theta)$	Model reference
MNIST SVHN CIFAR10	`Conv(64,3,1)` → `ReLU` → `Conv(64,3,1)`	Chen et al. (2018)[‡]
SpoAD CharT	`Linear(32,32)` → `Tanh` → `Linear(32,32)` → `Tanh` → `Linear(32,32)` → `Tanh` → `Linear(32,32)`	Kidger et al. (2020b)[§]
ArtWR	`Linear(64,64)` → `Tanh` → `Linear(64,64)` → `Tanh` → `Linear(64,64)` → `Tanh` → `Linear(64,64)`	Kidger et al. (2020b)[§]
Circle	`Linear(2,64)`[??] → `Tanh` → `Linear(64,2)`[6]	Chen et al. (2018)[‡]
Gas	`ConcatSquashLinear(8,160)` → `Tanh` → `ConcatSquashLinear(160,160)` → `Tanh` → `ConcatSquashLinear(160,160)` → `Tanh` → `ConcatSquashLinear(160,8)`	Grathwohl et al. (2018)[‡]
Miniboone	`ConcatSquashLinear(43,860)` → `SoftPlus` → `ConcatSquashLinear(860,860)` → `SoftPlus` → `ConcatSquashLinear(860,43)`	Grathwohl et al. (2018)[‡]

Table 8: Hyper-parameter grid search considered for each method

Method	Learning rate	Weight decay
Adam	{ 1e-4, 3e-4, 5e-4, 7e-4, 1e-3, 3e-3, 5e-3, 7e-3, 1e-2, 3e-2, 5e-2 }	{0.0, 1e-4, 1e-3 }
SGD	{ 1e-3, 3e-3, 5e-3, 7e-3, 1e-2, 3e-2, 5e-2, 7e-2, 1e-1, 3e-1, 5e-1 }	{0.0, 1e-4, 1e-3 }
Ours	{ 1e-3, 3e-3, 5e-3, 7e-3, 1e-2, 3e-2, 5e-2, 7e-2, 1e-1, 3e-1, 5e-1 }	{0.0, 1e-4, 1e-3 }

- `ConcatSquashLinear(input_dim, output_dim)`[5]
- `GRUCell(input_dim, hidden_dim)`

Table 7 details the vector field $F(t, \mathbf{x}_t, \theta)$ of Neural ODEs for each dataset. All vector fields are represented by some DNNs, and their architectures are adopted from previous references as listed. The convolution-based feature extraction of image-classification models consists of 3 convolution layers connected through `ReLU`, *i.e.* `Conv(64,3,1)` → `ReLU` → `Conv(64,4,2)` → `ReLU` → `Conv(64,4,2)`. For time-series models, We set the dimension of the hidden space to 32, 64, and 32 respectively for SpoAD, ArtWR, and CharT. Hence, their GRU cells are configured by `GRUCell(27,32)`, `GRUCell(19,64)`, and `GRUCell(7,32)`. Since these models take regular time-series with the interval of 1 second, the integration intervals of their Neural ODEs are set to $\{0, 1, \cdots, K\}$, where K is the series length listed in Table 3. Finally, we find that using 1 Neural ODE is sufficient to achieve good performance on Circle and Miniboone, whereas for Gas, we use 5 Neural ODEs stacked in sequence.

Tuning process. We perform a grid search on tuning the hyper-parameters (*e.g.* learning rate, weight decay) for each method on each dataset. The search grid for each method is detailed in Table 8. All figures and tables mentioned in Section 4 report the best-tuned results. For time-series models, we employ standard learning rate decay and note that without this annealing mechanism, we are unable to have first-order baselines converge stably. We also observe that the magnitude of the gradients of the GRU cells is typically 10-50 larger than the one of the Neural ODEs. This can make training unstable when the same configured optimizer is used to train all modules. Hence, in practice we fix Adam to train the GRUs while varying the optimizer for training Neural ODEs. Lastly, for image classification models, we deploy our method together with the standard Kronecker-based method (Grosse & Martens, 2016) for training the convolution layers. This enables full second-order training

[5] https://github.com/rtqichen/ffjord/blob/master/lib/layers/diffeq_layers/basic.py#L76

for the entire model, where the Neural ODE, as a continuous-time layer, is trained using our method proposed in Alg. 1. Finally, the momentum value for SGD is set to 0.9.

Dataset. All image datasets are preprocessed with standardization. To accelerate training, we utilize 10% of the samples in Gas, which still contains 85,217 training samples and 10,520 test samples. In general, the relative performance among training methods remains consistent for larger dataset ratios.

Setup and motivation of Fig. 5. We initialize all Neural ODEs with the *same* parameters while only varying the integration bound t_1. By manually grid-searching over t_1, Fig. 5 implies that despite initializing from the same parameter, different t_1 can yield distinct training time and accuracy; in other words, different t_1 can lead to distinct ODE solution. As an ideal Neural ODE model should keep the training time as small as possible without sacrificing the accuracy, there is a clear motivation to adaptive/optimize t_1 throughout training. Additional comparison w.r.t. standard (*i.e.* static) residual models can be founded in Appendix A.5.

Generating Fig. 8. The numerical values of the per-iteration runtime are reported in Table 9, whereas the ones for the memory consumption are given in Table 10. We use the last rows (*i.e.* $\frac{\text{SNOpt}}{\text{Adam}}$) of these two tables to generate Fig. 8.

Table 9: Per-iteration runtime (seconds) of different optimizers on each dataset

	Image Classification			Time-series Prediction			Continuous NF		
	MNIST	SVHN	CIFAR10	SpoAD	ArtWR	CharT	Circle	Gas	Minib.
Adam	0.15	0.78	0.17	5.24	9.95	14.79	0.34	2.25	0.65
SGD	0.15	0.81	0.17	5.23	10.00	14.77	0.33	2.28	0.74
SNOpt	0.15	0.68	0.20	5.18	10.05	14.89	0.94	4.34	1.04
$\frac{\text{SNOpt}}{\text{Adam}}$	1.00	0.87	1.16	0.99	1.01	1.01	2.75	1.93	1.60

Table 10: Memory Consumption (GBs) of different optimizers on each dataset

	Image Classification			Time-series Prediction			Continuous NF		
	MNIST	SVHN	CIFAR10	SpoAD	ArtWR	CharT	Circle	Gas	Minib.
Adam	1.23	1.29	1.29	1.39	1.18	1.24	1.13	1.17	1.28
SGD	1.23	1.28	1.28	1.39	1.18	1.24	1.13	1.17	1.28
SNOpt	1.64	1.81	1.81	1.49	1.28	1.34	1.15	1.34	1.68
$\frac{\text{SNOpt}}{\text{Adam}}$	1.33	1.40	1.40	1.07	1.09	1.08	1.02	1.14	1.31

Tikhonov regularization in line 10 of Alg. 1. In practice, we apply Tikhonov regularization to the precondition matrix, *i.e.* $\mathcal{L}_{\theta^n\theta^n} + \epsilon \boldsymbol{I}$, where θ^n is the parameter of layer n (see Fig. 3 and (13)) and ϵ is the Tikhonov regularization widely used for stabilizing second-order training (Botev et al., 2017; Zhang et al., 2019). To efficiently compute this ϵ-regularized Kronecker precondition matrix without additional factorization or approximation (*e.g.* Section 6.3 in Martens & Grosse (2015)), we instead follow George et al. (2018) and perform eigen-decompositions, *i.e.* $\bar{\boldsymbol{A}}_n = \boldsymbol{U}_{\boldsymbol{A}} \Sigma_{\boldsymbol{A}} \boldsymbol{U}_{\boldsymbol{A}}^\mathsf{T}$ and $\bar{\boldsymbol{B}}_n = \boldsymbol{U}_{\boldsymbol{B}} \Sigma_{\boldsymbol{B}} \boldsymbol{U}_{\boldsymbol{B}}^\mathsf{T}$, so that we can utilize the property of Kronecker product (Schacke, 2004) to obtain

Algorithm 2 ϵ-regularized Kronecker Update

1: **Input:** Tikhonov regularization ϵ, amortization α, Kronecker matrices $\bar{\boldsymbol{A}}_n$ $\bar{\boldsymbol{B}}_n$
2: $\boldsymbol{U}_{\boldsymbol{A}}, \Sigma_{\boldsymbol{A}} = \texttt{EigenDecomposition}(\bar{\boldsymbol{A}}_n)$
3: $\boldsymbol{U}_{\boldsymbol{B}}, \Sigma_{\boldsymbol{B}} = \texttt{EigenDecomposition}(\bar{\boldsymbol{B}}_n)$
4: $\boldsymbol{X} := \texttt{vec}^{-1}((\boldsymbol{U}_{\boldsymbol{A}} \otimes \boldsymbol{U}_{\boldsymbol{B}})^\mathsf{T} \mathcal{L}_{\theta^n}) = \boldsymbol{U}_{\boldsymbol{B}}^\mathsf{T} \widetilde{\mathcal{L}}_{\theta^n} \boldsymbol{U}_{\boldsymbol{A}}$
5: $\boldsymbol{S}^* := \alpha \boldsymbol{S}^* + (1-\alpha)\boldsymbol{X}^2$
6: $\boldsymbol{X} := \boldsymbol{X}/(\boldsymbol{S}^* + \epsilon)$
7: $\delta\theta := (\boldsymbol{U}_{\boldsymbol{A}} \otimes \boldsymbol{U}_{\boldsymbol{B}})\texttt{vec}(\boldsymbol{X}) = \texttt{vec}(\boldsymbol{U}_{\boldsymbol{B}} \boldsymbol{X} \boldsymbol{U}_{\boldsymbol{A}}^\mathsf{T})$
8: $\theta \leftarrow \theta - \eta \delta\theta$

$$(\bar{\boldsymbol{A}}_n + \bar{\boldsymbol{B}}_n + \epsilon \boldsymbol{I})^{-1} = (\boldsymbol{U}_{\boldsymbol{A}} \otimes \boldsymbol{U}_{\boldsymbol{B}})(\Sigma_{\boldsymbol{A}} \otimes \Sigma_{\boldsymbol{B}} + \epsilon)^{-1}(\boldsymbol{U}_{\boldsymbol{A}} \otimes \boldsymbol{U}_{\boldsymbol{B}})^\mathsf{T}. \tag{35}$$

This, together with the eigen-based amortization which substitutes the original diagonal matrix $\boldsymbol{S} := \Sigma_{\boldsymbol{A}} \otimes \Sigma_{\boldsymbol{B}}$ in (35) with $\boldsymbol{S}^* := ((\boldsymbol{U}_{\boldsymbol{A}} \otimes \boldsymbol{U}_{\boldsymbol{B}})^\mathsf{T} \mathcal{L}_{\theta^n})^2$, leads to the computation in Alg. 2. Note that `vec` is the shorthand for vectorization, and we denote $\mathcal{L}_{\theta^n} = \texttt{vec}(\widetilde{\mathcal{L}}_{\theta^n})$. Finally, α is the

amortizing coefficient, which we set to 0.75 for all experiments. As for ϵ, we test 3 different values from $\{0.1, 0.05, 0.03\}$ and report the best result.

Error bar in Table 4. Table 11 reports the standard derivations of Table 4, indicating that our result remains statistically sound with comparatively lower variance.

Table 11: Test-time performance: accuracies for image and time-series datasets; NLL for CNF datasets

	MNIST	SVHN	CIFAR10	SpoAD	ArtWR	CharT	Circle	Gas	Minib.
Adam	98.83±0.18	91.92±0.33	77.41±0.51	94.64±1.12	84.14±2.53	93.29±1.59	0.90±**0.02**	-6.42±**0.18**	13.10±0.33
SGD	98.68±0.22	93.34±1.17	76.42±0.51	**97.70**±0.69	85.82±3.83	95.93±0.22	0.94±0.03	-4.58±0.23	13.75±0.19
SNOpt	**98.99±0.15**	**95.77±0.18**	**79.11±0.48**	97.41±**0.46**	**90.23±1.49**	**96.63±0.19**	**0.86**±0.04	**-7.55**±0.46	**12.50±0.12**

Discussion on Footnote 4. Here, we provide some reasoning on why the preconditioned updates may lead the parameter to regions that are stabler for integration. We first adopt the theoretical results in Martens & Grosse (2015), particularly their Theorem 1 and Corollary 3, to our setup.

Corollary 9 (Preconditioned Neural ODEs)**.** *Updating the parameter of a Neural ODE, $F(\cdot, \cdot, \theta)$, with the preconditioned updates in (14) is equivalent to updating the parameter $\theta^\dagger \in \mathbb{R}^n$ of a "preconditioned" Neural ODE, $F^\dagger(\cdot, \cdot, \theta^\dagger)$, with gradient descent. This preconditioned Neural ODE has all the activations $\mathbf{z}^n$ and derivatives $F_{\mathbf{h}^n}^\mathsf{T}\mathbf{q}_i$ (see Fig. 3) centered and whitened.*

These centering and whitening mechanisms are known to enhance convergence (Desjardins et al., 2015) and closely relate to Batch Normalization (Ioffe & Szegedy, 2015), which effectively smoothens the optimization landscape (Santurkar et al., 2018). Hence, one shall expect it also smoothens the diffeomorphism of both the forward and backward ODEs (1, 5) of Neural ODEs.

A.5 Additional Experiments

t_1 optimization. Fig. 12 shows that a similar behavior (as in Fig. 5) can be found when training MNIST: while the accuracy remains almost stationary as we decrease t_1 from 1.0, the required training time can drop by 20-35%. Finally, we provide additional experiments for t_1 optimization in Fig. 13. Specifically, Fig. 13a repeats the same experiment (as in Fig. 10) on training MNIST, showing that our method (green curve) converges faster than the baseline. Meanwhile, Fig. 13b and 13c suggest that our approach is also more effective in recovering from an unstable initialization of t_1. Note that both Fig. 10 and 13 use Adam to optimize the parameter θ.

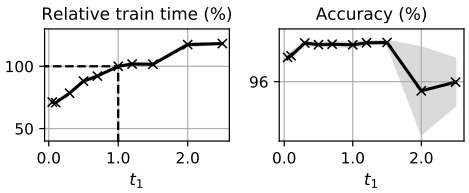

Figure 12: Training performance of MNIST with Adam when using different t_1.

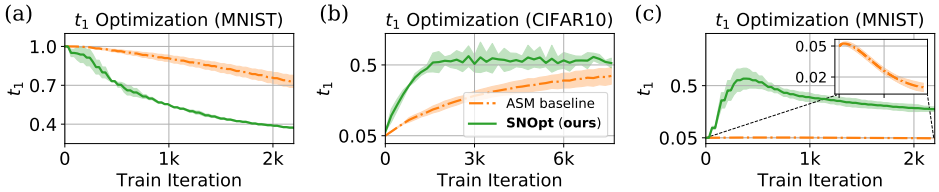

Figure 13: Dynamics of t_1 over training using different methods, where we consider (a) MNIST training with t_1 initialized to 1.0, and (b, c) CIFAR10 and MNIST training with t_1 initialized to some unstable small values (*e.g.* 0.05).

Convergence on all datasets. Figures 14 and 15 report the training curves of all datasets measured either by the wall-clock time or training iteration.

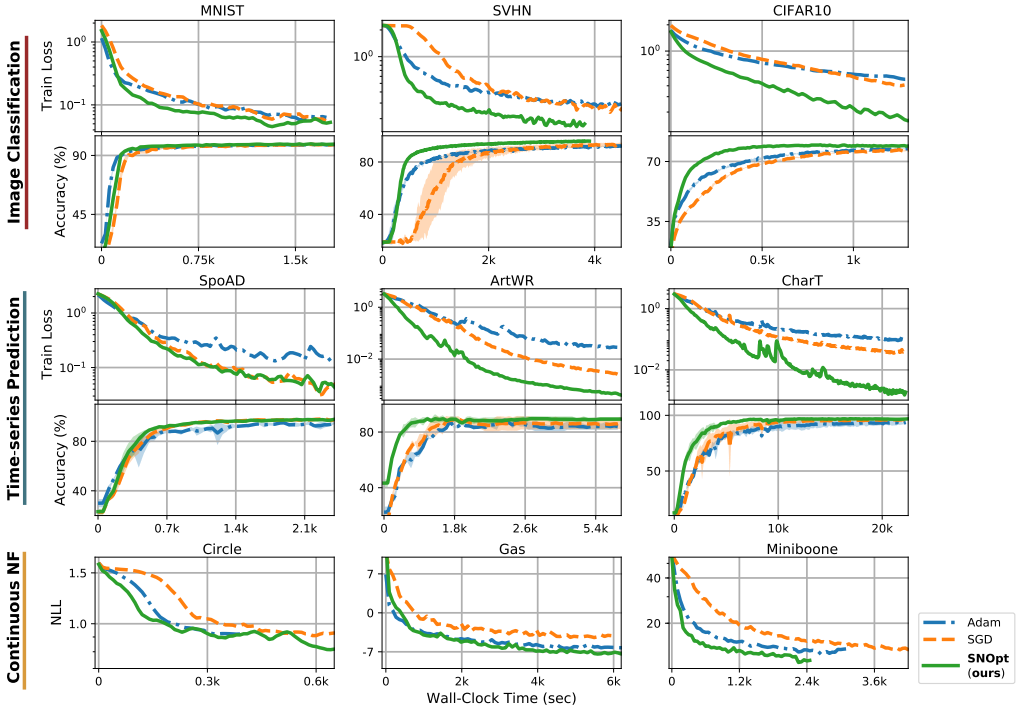

Figure 14: Optimization performance measured by *wall-clock* time across 9 datasets, including image (1st-2nd rows) and time-series (3rd-4th rows) classification, and continuous NF (5th row). We repeat the same figure with update iterations as x-axes in Fig 15. Our method (green) achieves faster convergence rate compared to first-order baselines. Each curve is averaged over 3 random trials.

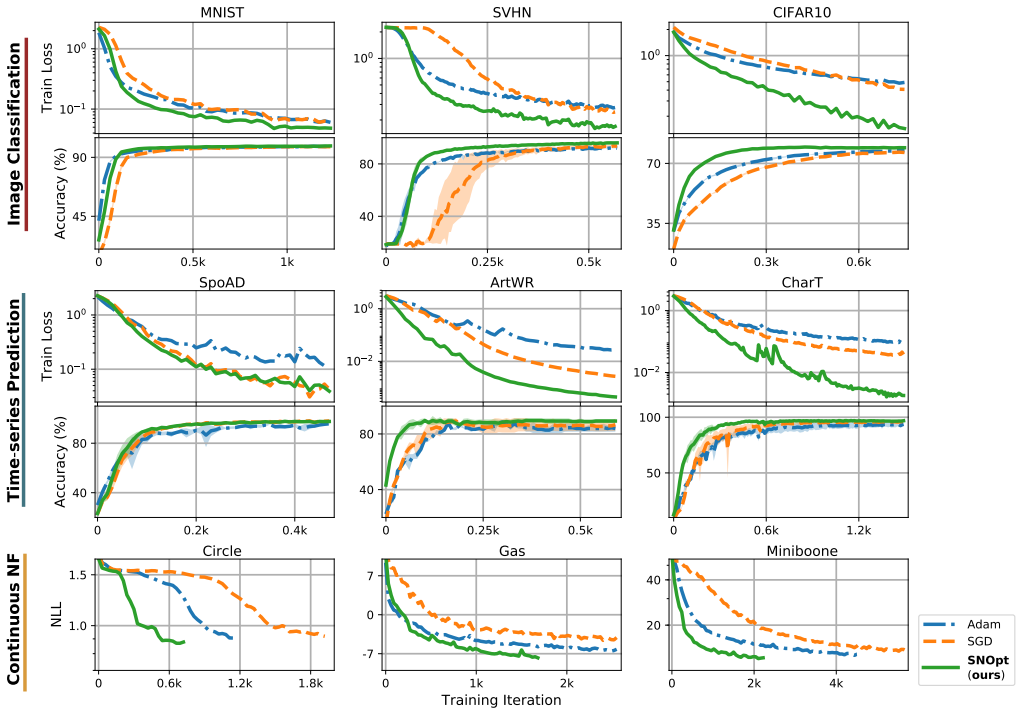

Figure 15: Optimization performance measured by *iteration updates* across 9 datasets, including image (1st-2nd rows) and time-series (3rd-4th rows) classification, and continuous NF (5th row). Each curve is averaged over 3 random trials.

Comparison with first-order methods that handle numerical errors. Table 12 and 13 report the performance difference between vanilla first-order methods (*e.g.* Adam, SGD), first-order methods equipped with error-handling modules (specifically MALI (Zhuang et al., 2021)), and our SNOpt. While MALI does improve the accuracies of vanilla first-order methods at the cost of extra per-iteration runtime (roughly 3 times longer), our method achieves highest accuracy among all optimization methods and retains a comparable runtime compared to *e.g.* vanilla Adam.

Table 12: Test-time performance (accuracies %) w.r.t. different optimization methods

	Adam	Adam + MALI	SGD	SGD + MALI	**SNOpt**
SVHN	91.92	91.98	93.34	94.33	**95.77**
CIFAR10	77.41	77.70	76.42	76.41	**79.11**

Table 13: Per-iteration runtime (seconds) w.r.t. different optimization methods

	Adam	Adam + MALI	SGD	SGD + MALI	**SNOpt**
SVHN	0.78	2.31	0.81	1.28	**0.68**
CIFAR10	**0.17**	0.55	**0.17**	0.23	0.20

Comparison with LBFGS. Table 14 reports various evaluational metrics between LBFGS and our SNOpt on training MNIST. First, notice that our method achieves superior final accuracy compared to LBFGS. Secondly, while both methods are able to converge to a reasonable accuracy (90%) within similar iterations, our method runs 5 times faster than LBFGS per iteration; hence converges much faster in wall-clock time. In practice, we observe that LBFGS can exhibit unstable training without careful tuning on the hyper-parameter of Neural ODEs, *e.g.* the type of ODE solver and tolerance.

Table 14: Comparison between LBFGS and our SNOpt on training MNIST

	Accuracy (%)	Runtime (sec/itr)	Iterations to Accu. 90%	Time to Accu. 90%
LBFGS	92.76	0.75	111 steps	2 min 57 s
SNOpt	98.99	0.15	105 steps	18 s

Results with different ODE solver (`implicit adams`). Table 15 reports the test-time performance when we switch the ODE solver from `dopri5` to `implicit adams`. The result shows that our method retains the same leading position as appeared in Table 4, and the relative performance between optimizers also remains unchanged.

Table 15: Test-time performance using "`implicit adams`" ODE solver: accuracies for image and time-series datasets; NLL for CNF datasets

	MNIST	SVHN	CIFAR10	SpoAD	ArtWR	CharT	Circle	Gas	Miniboone
Adam	98.86	91.76	77.22	95.33	86.28	88.83	0.90	-6.51	13.29
SGD	98.71	94.19	76.48	**97.80**	87.05	95.38	0.93	-4.69	13.77
SNOpt	**98.95**	**95.76**	**79.00**	97.45	**89.50**	**97.17**	**0.86**	**-7.41**	**12.37**

Comparison with discrete-time residual networks. Table 16 reports the training results where we replace the Neural ODEs with standard (*i.e.* discrete-time) residual layers, $\mathbf{x}_{k+1} = \mathbf{x}_k + F(\mathbf{x}_k, \theta)$. Since ODE systems can be made invariant w.r.t. time rescaling (*e.g.* consider $\frac{\mathrm{d}x}{\mathrm{d}t} = F(t, x, \theta)$ and $\tau = ct$, then $\frac{\mathrm{d}x}{\mathrm{d}\tau} = \frac{1}{c}F(\frac{\tau}{c}, x, \theta)$ will give the same trajectory $x(t) = x(\frac{\tau}{c})$), the results of these residual networks provide a performance validation for our joint optimization of t_1 and θ. Comparing Table 16 and 5 on training CIFAR10, we indeed find that SNOpt is able to reach the similar performance (77.82% *vs.* 77.87%) of the residual network, whereas the ASM baseline gives only 76.61%, which is 1% lower.

Table 16: Accuracies (%) of residual networks trained with Adam or SGD

	MNIST	SVHN	CIFAR10
resnet + Adam	98.75 ± 0.21	97.28 ± 0.37	77.87 ± 0.44

Batch size analysis. Table 17 provides results on image classification when we enlarge the batch size by the factor of 4 (*i.e.* $128 \to 512$). It is clear that our method retains the same leading position with a comparatively smaller variance. We also note that while enlarging batch size increases the memory for all methods, the ratio between our method and first-order baselines does not scale w.r.t. this hyper-parameter. Hence, just as enlarging batch size may accelerate first-order training, it can equally improve our second-order training. In fact, a (reasonably) larger batch size has a side benefit for second-order methods as it helps stabilize the preconditioned matrices, *i.e.* $\bar{\boldsymbol{A}}_n$ and $\bar{\boldsymbol{B}}_n$ in (14), throughout the stochastic training (note that too large batch size can still hinder training (Keskar et al., 2016)).

Table 17: Accuracies (%) when using larger ($128 \to 512$) batch sizes

	MNIST	SVHN	CIFAR10
Adam	99.14 ± 0.12	94.19 ± 0.18	77.57 ± 0.30
SGD	98.92 ± 0.08	95.67 ± 0.48	76.66 ± 0.29
SNOpt	$\mathbf{99.18 \pm 0.07}$	$\mathbf{98.00 \pm 0.12}$	$\mathbf{80.03 \pm 0.10}$