Linear Range in Gradient Descent

Angxiu Ni
Department of Mathematics, University of California, Berkeley, CA 94020, USA
niangxiu@math.berkeley.edu
Chaitanya Talnikar
Nvidia Corporation, Santa Clara, CA 95051, USA
chaitukca@gmail.com

Abstract
This paper defines linear range as the range of parameter perturbations which approximately leads to linear perturbations in states. We compute linear range by comparing the actual perturbations in states and the tangent solution of a network. Linear range is a new criterion for gradients to be meaningful, thus having many possible applications. In particular, we propose that the optimal learning rate at the beginning of training can be found automatically, by selecting a stepsize such that all minibatches are within linear range. We demonstrate our algorithm on a network with canonical architecture and a ResNet.

1 Introduction
Machine learning is a popular method for approximating complex functions arising in a variety of fields like computer vision [Krizhevsky et al., 2012], speech recognition [Graves et al., 2013] and many more. Stochastic Gradient Descent (SGD) is a common choice for optimizing the parameters of the deep neural network representing the functions [LeCun et al., 2012]. But, the correct application of the SGD algorithm requires setting an initial learning rate and a schedule that reduces the learning rate, or stepsize, as the algorithm proceeds closer to the optimal parameter values. Choosing a stepsize too small results in no improvement in the cost function, while choosing a stepsize too large causes chaotic oscillations in the value of the cost function.

There have been a number of approaches suggested to solve this problem. Adaptive SGD algorithms like AdaGrad and Adam [Duchi et al., 2011] require the setting of a global stepsize and other hyperparameters. In non-stochastic gradient descent, there are two methods popular for determining the optimal step size for a gradient. The first is the trust region method [Conn et al., 2000, Byrd et al., 1987, Sorensen, 1982], and the second is the line search method, like backtracking line search [Armijo, 1965], Wolfe conditions [Wolfe, 1969], and probabilistic line search [Mahsereci and Hennig, 2017]. However, when applied to SGD, both methods solve an optimization within a prescribed region or along a direction, which could lead to over-optimization for the current minibatch but deterioration of other minibatches. Moreover, trust regions typically uses second-order model, which can be expensive to build; line search methods typically give only upper but not lower bound on the objective change, which might lead to even more over-optimization.

Instead of optimizing for each minibatch, we propose to select initial stepsizes by checking a ‘speed limit’ given by such that all minibatches are within the linear range. The tangent solution describes the linearized perturbation on the entire network, and its difference with actual perturbations tells how nonlinear the entire network is, thus allowing good inference of the largest possible stepsize within the linear range. Our approach will be elaborated in the rest of this paper, organized as follows. First, we define tangent and adjoint solutions and show their relations in sensitivity analysis. Then
we define linear range and describe how to compute and use this criterion. Finally, we demonstrate
linGrad on some networks with different architectures.

2 Preparations

In this section, we first define networks from a dynamical system point of view. Then in subsection 2.1 we show that small perturbations on parameters will lead to roughly linear perturbation in all states of the network, which further leads to a tangent formula for linear perturbation in the objective. Finally, in subsection 2.2, we show this tangent formula is equivalent to the adjoint formula, which is a generalization of the backpropagation. Above discussion leads to the natural conclusion that, gradients are meaningful only when stepsizes lead to roughly linear perturbation in the states.

To start, we define a neural network with \( I \) layers as a discrete dynamical system, with the governing equation at the \( i \)-th layer:

\[
\begin{align*}
u_0 &= x, \\
u_{i+1} &= f_i(u_i, s_i) \quad \text{for } 0 \leq i \leq I - 1, \\
\end{align*}
\]

where \( x \) is the input, \( u_i \in \mathbb{R}^{m_i \times 1} \) is the states of neurons, and \( s_i \in \mathbb{R}^{n_i \times 1} \) is the parameters at the \( i \)-th layer to be trained, to minimize some objective \( J \), defined as:

\[
J\left(\{u_i, s_i\}_{i=0}^I\right) = \sum_{i=0}^I J_i(u_i, s_i).
\]

Typically the objective is the error between the actual output of network and the output provide by the data: in this case, the objective depends only on \( J_I(u_I) \). However, for future development, we allow the objective depend on all layers of network.

2.1 Tangent solutions

Assume we want to make perturbations to parameters \( \{s_i\}_{i=0}^I \) in the direction of \( \{\sigma_i \in \mathbb{R}^{n_i \times 1}\}_{i=0}^I \), the perturbations will be \( \{\Delta s_i = \sigma_i \psi\}_{i=0}^I \), where \( \psi \in \mathbb{R} \) is the stepsizes, or learning rate. When the stepsizes are infinitesimal \( \delta \psi \), the first order approximation at each layer is: \( \delta u_0 = 0, \delta u_{i+1} = f_{ui} \delta u_i + f_{si} \sigma_i \psi \), where \( \delta \) indicates infinitesimal perturbations, and \( \delta u_i \) includes perturbations propagated from previous layers. Here \( f_{ui} := \partial f_i / \partial u(u_i, s_i) \in \mathbb{R}^{m_i+1 \times m_i}, \) and \( f_{si} := \partial f_i / \partial s(u_i, s_i) \in \mathbb{R}^{m_i+1 \times n_i} \).

There is no perturbation on \( u_0 \), since the input data are accurate. Define \( v_i := \delta u_i / \delta \psi \), the governing equation for \( v_i \) is the so-called conventional inhomogeneous tangent equation, given by:

\[
v_0 = 0, \quad v_{i+1} = f_{ui} v_i + f_{si} \sigma_i .
\]

Later we are interested in computing \( v_i \psi \), which is easier to compute via:

\[
v_0 \psi = 0, \quad v_{i+1} \psi = f_{ui} (v_i \psi) + f_{si} \Delta s_i .
\]

Tangent equations for ResNet and neural ODE are given in appendix A.

Investigating inhomogeneous tangent solutions calls for first defining the homogeneous tangent equation: \( u_{i+1} = f_{ui} u_i \), which describes the evolution of perturbation on states while the parameters are fixed. The propagation operator \( D_i \in \mathbb{R}^{m_i \times m_i} \) is defined as the matrix that maps a homogeneous tangent solution at \( i \)-th layer to \( i \)-th layer. More specifically, \( D_i \) is defined as:

\[
D_i := \begin{cases} I_d \text{ (the identity matrix)}, & \text{when } i = 1; \\
f_{u,i-1} f_{u,i-2} \cdots f_{u,1} f_{u,1}, & \text{when } i > 1. 
\end{cases}
\]

We can use Duhamel’s principle to analytically write out a solution to equation 3. Intuitively, an inhomogeneous solution can be viewed as linearly adding up homogeneous solutions, each starting afresh at a previous layer, with initial condition given by the inhomogeneous term. More specifically,

\[
v_0 = 0, \quad v_i = \sum_{i=0}^{i-1} D_{i+1} f_{si} \sigma_i \quad \text{for } 1 \leq i \leq I.
\]

Now we can write out a formula for the sensitivity \( dJ / d\psi \). More specifically, first differentiate each term in equation 2, apply the definition of \( v_i \), then substitute equation 6, we get:

\[
\frac{dJ}{d\psi} = \sum_{i=0}^I \left( J_{ui} v_i + J_{si} \sigma_i \right) = \sum_{i=1}^{I-1} \sum_{i=0}^{i-1} J_{ui} D_{i+1} f_{si} \sigma_i + \sum_{i=0}^I J_{si} \sigma_i .
\]

Here both \( J_{ui} := \partial J_i / \partial u(u_i, s_i) \in \mathbb{R}^{1 \times m_i} \) and \( J_{si} := \partial J_i / \partial s(u_i, s_i) \in \mathbb{R}^{1 \times n_i} \) are row vectors.
2.2 Adjoint solutions

In this subsection, we first use a technique similar to backpropagation to derive an adjoint sensitivity formula, which we then directly show is equivalent to the tangent sensitivity formula in equation (7).

Assume we perturb the $l$-th layer by $\delta u_l$, and let it propagate through the entire network, then the change in the objective is: $J + \delta J = \sum_{j=1}^{l} J_j (f_{j-1}(\cdots f_1(u_l + \delta u_l, s_l) \cdots, s_j)$. Neglecting higher order terms, we can verify the inductive relation $\delta J/\delta u_l = \delta J/\delta u_{l+1} f_{u_{l+1}} + J_{u_l}$. Define $\tau_l := \delta J/\delta u_l \in \mathbb{R}^{1 \times m}$, it satisfies the conventional inhomogeneous adjoint equation:

$$\tau_{l+1} = 0, \quad \tau_l = \tau_{l+1} f_{u_l} + J_{u_l}.$$  

Notice that the order of layers is reversed. Here the terminal condition is because we can assume there is $(I+1)$-th layer which $J$ does not depend on. Hence the adjoint sensitivity formula is:

$$\frac{dJ}{d\psi} = \sum_{l=0}^{I} \frac{\delta J}{\delta u_l} \frac{\partial u_l}{\partial \psi} + J_{u_I} \sigma_I = \sum_{l=1}^{I} \tau_l f_{u_{l-1}} \sigma_{l-1} + \sum_{l=0}^{I} J_{u_I} \sigma_I. \tag{9}$$

Here $\tau_{l-1} \sigma_{l-1} = \partial u_l/\partial \psi \neq \delta u_l/\delta \psi = v_l$; in the term $\partial u_l/\partial \psi$ we do not consider the perturbation on $u_l$ due to the perturbation propagated from previous layers, which has been accounted for in $\tau_l$. We also used $\partial u_0/\partial \psi = 0$ since $u_0$ is fixed.

The advantage of the adjoint sensitivity formula, comparing to the tangent formula, is a clearer view of how the sensitivity depends on $\sigma_i$, which further enables us to select the direction for perturbing parameters, $\{\sigma_i\}_{i=0}^{I}$. Not surprisingly, the inhomogeneous adjoint solution is a generalization of the backpropagation. To illustrate, setting the objective to take the common form, $J = J_I(u_I)$, then $\tau_l = J_{u_I} f_{u_{l-1}} \cdots f_{u_{l+1}}$. The gradient of $J$ to parameters, given by the backpropagation, is:

$$\partial J/\partial s_I = \partial u_I f_{u_{l-I}} \cdots f_{u_{l+1}} = \tau_{l-1} f_{s_{l-I}}. \tag{10}$$

The sensitivity can be given by either a tangent formula in equation (7), or an adjoint formula in equation (9); of course the two formula should be equivalent. Since later development heavily depends on this equivalence, we also prove it directly. To start, first define the homogeneous adjoint equation: $\tau_l = \tau_{l+1} f_{u_I}$, where $\tau_l \in \mathbb{R}^{1 \times m}$ is a row vector. The adjoint propagation operator $\overline{D}_I^l$ is the matrix which, multiplying on the right of a row vector, maps a homogeneous adjoint solution at $l$-th layer to $l$-th layer. A direct computation shows that $D_I^l = \overline{D}_I^l$. Using Duhamel’s principle with reversed order of layers, we can analytically write out the inhomogeneous adjoint solution:

$$\tau_{l+1} = 0, \quad \tau_l = \sum_{i=1}^{I} J_{u_i} \overline{D}_I^{l-i} \text{ for } 0 \leq l \leq I. \tag{11}$$

To directly show the equivalence between tangent and adjoint sensitivity formula, first change the order for the double summation in equation (7), then assemble terms not related to $f_{u_I} \sigma_I$:

$$\frac{dJ}{d\psi} = \sum_{i=1}^{I} J_{u_i} v_i + \sum_{i=1}^{I} J_{u_I} \sigma_I = \sum_{i=1}^{I} \sum_{l=0}^{I} J_{u_i} \overline{D}_I^{l-i} f_{u_I} \sigma_I + \sum_{i=0}^{I} J_{u_I} \sigma_I \tag{12}$$

$$= \sum_{l=0}^{I} \left( \sum_{i=l+1}^{I} J_{u_i} \overline{D}_I^{l+i-1} \right) f_{u_I} \sigma_I + \sum_{i=0}^{I} J_{u_I} \sigma_I = \sum_{l=0}^{I} \tau_{l+1} f_{u_I} \sigma_I + \sum_{i=0}^{I} J_{u_I} \sigma_I.$$

3 Linear range

3.1 Definition

Assume that the direction to perturb parameters, $\{\sigma_i\}_{i=0}^{I}$, has been decided, we still need to specify the stepsize $\psi$ to get the new parameters, the selection of which is the topic of the this section. There are two equivalent methods for computing the sensitivity $dJ/d\psi$: the tangent formula in equation (7) and the adjoint formula in equation (12). The adjoint formula is useful for deciding $\sigma_i$, and the tangent formula is useful for checking the effectiveness of the sensitivity, which is its ability to predict
the linear change in objective after we slightly perturb the parameters. A sufficient condition for the approximate linearity, as suggested by the tangent formula and definition of tangent solutions, is that the perturbation in all of the states are roughly linear to \( \{ s_i \}_{i=0}^l \).

To elaborate, first define a nonlinear measurement for the perturbations in the states of one network, 

\[
\epsilon := \frac{1}{l} \sum_{i=1}^l \frac{\| u_{new,i} - u_{old,i} - v_i \psi \|}{\| v_i \psi \|}, \tag{13}
\]

where \( v \) is the conventional tangent solution, \( u_{old} \) and \( u_{new} \) are the states before and after parameter change, subscript \( i \) indicates the layer, and the norm is \( \| \cdot \|_2 \). Assume that we can use Taylor expansion for \( u_{new} \) around \( u_{old} \), and that \( v = \frac{du_{new}}{d\psi} \) is non-zero, we have \( u_{new} = u_{old} + v \psi + v' \psi^2 + O(\psi^3) \), where \( v' \) is some unknown constant vector. Hence \( \epsilon = C \psi + O(\psi^2) \) for some constant \( C \), and for small \( \psi \) we may regard \( \epsilon \) as linear to \( \psi \). With this preparation of nonlinearity measurement, we can finally define linear range.

**Definition.** Given a batch of samples, the \( \epsilon^*- \)linear range on parameters is the range of parameter perturbations such that \( \epsilon \leq \epsilon^* \). The linear range on objective and on states are the image sets of the linear range on parameters.

### 3.2 Gradient descent by linear range

Linear range is a criterion can be used in many ways. In this subsection, we use it to develop linGrad, a stochastic gradient descent (SGD) method. In linGrad, the stepsize is determined by that, for some \( \epsilon^* \approx 0.3 \) at the start of training, since \( 0.3^2 \approx 0.1 \), meaning the second order terms now are one magnitude smaller, which can be neglected in many engineering problems, resulting in a barely adequate linear approximation. For easy problems where most data achieves minimum for similar parameters, it is unlikely that optimization for one minibatch deteriorates others, and we could use larger, yet still below 1, \( \epsilon^* \).

**Algorithm 1 linGrad: Linear range gradient descent (with fixed \( \epsilon^* \))**

**Precondition:** \( \epsilon^* = 0.3; \) empty list \( L \); \( N_0 \) samples in a minibatch; \( N_b \) minibatches.

1. **for** each epoch \( n \), **do**
2.  **for** each minibatch, **do**
3.   **for** \( n \leftarrow 1, N_s \) **do** \( \triangleright \) We sometimes omit the subscript \( n \).
4.   Compute \( u_{old} \) using parameters \( \{ s_i \}_{i=0}^l \) and input data.
5.   Compute adjoint solution \( \{ \eta_i \}_{i=1}^l \).
6.   Select \( \{ \sigma_i \}_{i=0}^l \) according to some rules.
7.   Compute tangent solution \( \{ v_i \psi \}_{i=1}^l \) by equation (4).
8.   Compute new states \( u_{new} \) using parameters \( \{ s_i + \sigma_i \psi \}_{i=0}^l \).
9.   Compute \( \epsilon_n \) for this sample by equation (13).
10. **end for**
11. Compute \( \epsilon = \frac{\sum_{n=1}^{N_s} \epsilon_n}{N_s} \).
12. Append \( \psi \epsilon^*/\epsilon \) to the list \( L \).
13. \( \psi \leftarrow \min \{ \text{last } N_{hist} \text{ elements in } L \} \) \( \triangleright \) Set \( N_{hist} \approx N_b/N_{lin} \)
14. Update parameters \( s_i \leftarrow s_i + \sigma_i \psi \).
15. **end for** \( \triangleright \) Only need to do steps 7 to 13 once every \( N_{lin} \) minibatches.
16. **end for**

### 3.3 Remarks

Notice that the nonlinear measurement is defined over the entire network rather than just over the objective. Since the objective is only one number, it may not fully convey the information how far
we are within the linear range; even worse, we might just be lucky sometimes to get the objective linear to parameter perturbation, whereas the network has already gone into nonlinear regions. We need the full information on the whole network to better infer the linear range. Moreover, a nonlinear measurement over the objective still requires computing $u_{new}$, which is typically more expensive than computing the tangent solution. Hence, a nonlinear measurement over the objective would have half the computational cost, but produce much less information than our current measurement.

The concept of linear range is useful for other scenarios beyond linGrad. One possible application is that it offers a criterion for comparing different descent directions, $\sigma_i$. For shallow to medium-depth networks, there is only one typical adjoint solvers; for deep networks, because of the exploding gradient, there are also multiple choices on adjoint solvers, adding even more complexity for further choosing $\sigma_i$. Now provided with different descent directions, $\{\sigma_i\}_{i=0}$, the linear range can serve as a standard for comparison: the larger the linear range, the larger we may effectively perturb parameters, and the faster the convergence.

As we shall see in section 4 for linGrad with fixed $\epsilon^*$, stepsize remains roughly the same value. Should we desire a stepsize decreasing scheme, we can as well decrease $\epsilon^*$; further, we may use linGrad only to compute an initial stepsize and perform some stepsize decreasing thereafter. However, decreasing $\epsilon^*$ and stepsize may slow down escaping saddle points [Jin et al., 2017, 2018]: for this reason, we do not include $\epsilon^*$ decreasing as part of linGrad. It is also possible to increase batch size instead of decrease stepsize [Byrd et al., 2012; Friedlander and Schmidt, 2012]. There are also many choice in terminating criteria, such as terminate when the ‘signal-to-noise’ ratio, the ratio between the average and RMS of gradients, is too low [De et al., 2017], or when the ratio of counter-directions, the pairs of gradients with negative inner-products, is about half. LinGrad can be added to most existing training schemes, and we suggest readers to experiment.

Although not implemented in this paper, it is possible to code tangent solvers by tracing computation graphs, in a way similar to backpropagation. The tangent propagation is also a local process as existing training schemes, and we suggest readers to experiment.

4 Applications

4.1 Application on an artificial network

We first apply linGrad on a network where each layer is given by $f_i(u_i, W_i) = g(W_i u_i + b_i)$, where $g$ is the vectorized logistic function. Our parameters to be learned are $W_i \in \mathbb{R}^{m_{i+1} \times m_i}$ and $b_i \in \mathbb{R}^{m_{i+1}}$. The perturbations on parameters are $\Sigma_i \psi = \Delta W_i$ and $\beta_i \psi = \Delta b_i$. Our objective is defined only on the last layer as the square difference $J := J_f(u_1) = \frac{1}{2} \sum_{j=1}^{m_1} (u_i^j - y_i^j)^2$, where $y$ is the output data for this sample. To adapt with our previous notations, we regard $(W_i, b_i)$ and $(\sigma_i, \beta_i)$ as one-dimensional vectors of length $n_i = m_i+1 \times m_i + m_{i+1}$, obtained by flattening the matrix and append the vector. Then, for programming convenience, we reshape this long dimension back to a matrix and a vector in the list of results below.

$$f_{ui} = \Lambda_i W_i, \quad f_{ui} \sigma_i = \Lambda_i (\Sigma_i u_i + \beta_i), \quad J_{ui} = (u_i - y_i)^T,$$

where $\Lambda_i = diag[g_i(1-g_i)] \in \mathbb{R}^{m_{i+1} \times m_i+1}$ is a diagonal matrix due to differentiating the component-wise logistic function. By either carefully managing subscripts in partial derivatives or directly deriving from the definition, we get tangent and adjoint equations:

$$v_{ui} \psi = 0, \quad v_{i+1} \psi = \Lambda_i (W_i v_i \psi + \Delta W_i u_i + \Delta b_i);$$
$$\overline{v_I} = J_{ui}, \quad \overline{v_I} = v_{i+1} \Lambda_i W_i.$$  

The feedforward and backpropagation in our implementation are from the code complementing [Nielsen, 2015]. For our particular example, we use two hidden layers. All layers have the same number of neurons, $m_i = 50$. We first fix an network with randomly generated parameters, then generate $5 \times 10^4$ training samples and $1 \times 10^4$ test samples by feeding this fixed network with random inputs. Here all random numbers are from independent standard normal distribution.
For the training, initial parameters are generated randomly, and all samples are randomly shuffled for each epoch. We compute the nonlinear measurement and adjust stepsize every $N_{\text{lin}} = 100$ mini-batches, and take stepsize as the smallest of the last $N_{\text{hist}} = \max(50, 50000/(N_{\text{lin}}N_s))$ candidate values, where $N_s$ varies. We choose $\Sigma_i$ as the fastest descent direction:

$$
\Sigma_i = -\pi_{i+1}f_{si} = -\Lambda_i\pi_{i+1}^T u_i^T, \quad \beta_i = -\Lambda_i\pi_{i+1}^T.
$$

As we can see from the left of figure 1, for batch size $N_s = 10$, comparing to SGD with fixed stepsizes, linGrad with $\epsilon^* = 0.3$ descent the fastest, especially in the first 50 epochs, confirming that the 'speed limit' during the first phase of training neural networks is given by the criterion of linear range. In fact, if the objective function is defined as 10 times the current, SGD would have 10 times larger parameter perturbations and totally different convergence behavior, whereas the linear range and hence linGrad would remain unaffected. Moreover, from the right of figure 1 we can see that $\epsilon^* = 0.3$ persists to be the optimal for linGrad with different batch sizes, as suggested in section 3.2.

![Figure 1: LinGrad applied on data generated by an artificial network. Each objective history is averaged over 5 runs. The vertical axis is average of normalized distance $\left(\sum_{j=1}^{m_I}(u_j^I - y_j^I)^2\right)^{0.5}/\sqrt{m_I}$. Left: linGrad with minibatch size $N_s = 10$ and different $\epsilon^*$ versus SGD with different fixed stepsizes. Right: linGrad with different $N_s$ and $\epsilon^*$.](image)

Histories of stepsize $\psi$ and nonlinear measurement $\epsilon$ of linGrad are shown in figure 2. We run linGrad with different initial stepsizes $\psi_0 = 0.01$ and $\psi_0 = 1$. As shown, $\psi_0$ does not affect much: this is expected, since $\psi_0$ is used only to infer the first $\epsilon^*$-linear range. This confirms that linGrad relieves us from the headache of choosing initial stepsize. Also we can see that the stepsize remains roughly the same value, indicating that we can adjust stepsize by nonlinear measurement even less frequently. Finally, the nonlinear measurement remains below 0.3, confirming that our implementation correctly keeps the stepsize within the linear range.

![Figure 2: History of stepsize and nonlinear measurement for linGrad with $\epsilon^* = 0.3$ and $N_s = 10$.](image)

### 4.2 Application on MNIST

We then apply linGrad on MNIST with $6 \times 10^4$ data. The network has three layers, where the input layer has 764 neurons, the hidden layer 30 neurons, and the output layer 10 neurons. The classification is done by selecting the largest component in the output layer. Other aspects of the architecture and settings are the same as we used in section 4.1. We compare linGrad to SGD with constant stepsizes and compare linGrad with different minibatch sizes in figure 3. For this problem...
linGrad converges fastest for either $\epsilon^* = 0.3, 0.5$ or $0.8$, both comparable to SGD with optimal stepsize. Again, we can see the selection of $\epsilon^*$ is robust to $N_s$.

Figure 3: LinGrad applied on MNIST. Same settings as figure 1. The history for SGD with $\psi = 100$ remains similar as the initial value, 0.9, and hence out of the picture.

5 Conclusion

This paper defines linear range and states how to compute it via comparing tangent solutions with finite differences. Linear range is a new criteria which can be used for evaluating the quality of stepsizes and descent directions, and it could have many theoretical and practical applications. In particular, we develop a stochastic gradient descent algorithm, linGrad, where the stepsize is given by such that all minibatches are within linear range. By applying linGrad on several test cases, we suggest that setting the target nonlinear measurement as $\epsilon^* \approx 0.3$ yields the optimal convergence rate, at least at the beginning of training. LinGrad can be integrated with existing gradient descent algorithms to improve the selection of initial stepsizes.

A Tangent equations for other architectures

To compute nonlinear measurement defined by equation (13) for other architectures, we further provide corresponding tangent equations.

For ResNet [He et al., 2016], the dynamical system corresponding to equation (1) and the inhomogeneous tangent equation corresponding to equation (4) are:

\begin{align*}
  u_0 &= x, \quad u_{i+1} = g \left( \sum_{j \leq i} W_{j,i+1} u_j + b_i \right); \\
  v_0 \psi &= 0, \quad v_{i+1} \psi = \Lambda_i \left( \sum_{j \leq i} (W_{j,i+1} v_j \psi + \Delta W_{j,i+1} u_j) + \Delta b_i \right). \tag{17}
\end{align*}

Further, we suggest that the objective function should not depend on the intermediate layers in a residual block, since they have different interpretations as the input/output layers. In fact, if we do not regard intermediate layers in residual blocks as states, then we can recover the basic form of dynamical system in equation (1) with complicated $f_i$’s.

For neural ODE [Chen et al., 2018], the dynamical system and inhomogeneous tangent equation are:

\begin{align*}
  u_0 &= x, \quad \frac{du}{dt} = f(u, s); \\
  v_0 &= 0, \quad \frac{dv \psi}{dt} = f_u v \psi + f_s \Delta s. \tag{18}
\end{align*}

Here $f_u := \partial f / \partial u$, $f_s := \partial f / \partial s$. Moreover, the norm in the definition of nonlinear measurement in equation (13) should change to function norms.
References

Larry Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 2012. ISSN 0030-8730. doi: 10.2140/pjm.1966.16.1.

R. H. Byrd, G. M. Chin, J. Nocedal, and Y. Wu. Sample Size Selection in Optimization Methods for Machine Learning. *Mathematical Programming*, 134(1):127–155, 2012.

Richard H. Byrd, Robert B. Schnabel, and Gerald A. Shultz. A Trust Region Algorithm for Nonlinearly Constrained Optimization. *SIAM Journal on Numerical Analysis*, 24(5):1152–1170, 1987. ISSN 0036-1429. doi: 10.1137/0724076.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems*, pages 6571–6583, Montreal, Canada, 2018. ISBN 9781139108188. doi: arXiv:1806.07366v3. URL [http://arxiv.org/abs/1806.07366](http://arxiv.org/abs/1806.07366).

A R Conn, N I M Gould, and P L Toint. *Trust Region Methods*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, 2000.

Soham De, Abhay Yadav, David Jacobs, and Tom Goldstein. Big Batch SGD: Automated Inference using Adaptive Batch Sizes. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1–20, Fort Lauderdale, Florida, USA, 2017. URL [http://arxiv.org/abs/1610.05792](http://arxiv.org/abs/1610.05792).

John Duchi, Elad Hazan, and Yoram Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

Michael P. Friedlander and Mark Schmidt. Hybrid Deterministic-Stochastic Methods for Data Fitting. *SIAM Journal on Scientific Computing*, 34(3):A1380–A1405, 2012. ISSN 1064-8275. doi: 10.1137/130908257.

Alex Graves, Abdel Rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 6645–6649, 2013. ISSN 15206149. doi: 10.1109/ICASSP.2013.6638947.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE conference on computer vision and pattern recognition*, 2016. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90. URL [http://arxiv.org/abs/1512.03385](http://arxiv.org/abs/1512.03385).

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to Escape Saddle Points Efficiently. In *Proceedings of the 34 th International Conference on Machine Learning*, volume 70, pages 1724–1732, Sydney, Australia, 2017. URL [http://arxiv.org/abs/1703.00887](http://arxiv.org/abs/1703.00887).

Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. Accelerated Gradient Descent Escapes Saddle Points Faster than Gradient Descent. In *Proceedings of the 31st Conference On Learning Theory*, number 75, pages 1042–1085, Stockholm, Sweden, 2018. URL [http://arxiv.org/abs/1711.10456](http://arxiv.org/abs/1711.10456).

Alex Krizhevsky, Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, Lake Tahoe, Nevada, USA, 2012. URL [http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.299.205](http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.299.205).

Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus Robert Müller. Efficient backprop. In Montavon G., Orr G.B., and Müller KR., editors, *Neural networks: Tricks of the trade*, volume 7700 of *Lecture Notes in Computer Science*, pages 9–48. Springer, Berlin, Heidelberg, 2012. ISBN 9783642352881. doi: 10.1007/978-3-642-35289-8-3.

Maren Mahsereci and Philipp Hennig. Probabilistic Line Searches for Stochastic Optimization. *Journal of Machine Learning Research*, 18:1–59, 2017. URL [http://arxiv.org/abs/1703.10034](http://arxiv.org/abs/1703.10034).
Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015.

D. C. Sorensen. Newton’s Method with a Model Trust Region Modification. *SIAM Journal on Numerical Analysis*, 19(2):409–426, 1982. ISSN 0036-1429. doi: 10.1137/0719026.

Philip Wolfe. Convergence Conditions for Ascent Methods. *SIAM Review*, 11(2):226–235, 2005. ISSN 0036-1445. doi: 10.1137/1011036.