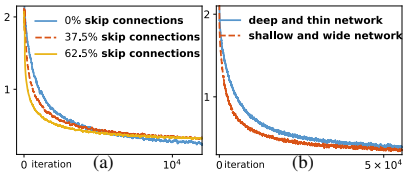
We thank all the reviewers for their insightful and encouraging comments, and will update revision to solve the issues.

To Reviewer #1. Consider channel number $m=\mathcal{O}(n^2)$ and sample number n is much larger than depth h in NAS, our learning rate (LR) is $\eta=\mathcal{O}(\lambda/\sqrt{m}/h^3)=\mathcal{O}(\lambda/n)$. It indeed improves LR requirement in [18-20] which analyze convergence of ResNet, e.g. $\eta=\mathcal{O}(\lambda/n^2)$ in [18,19] and $\eta=\mathcal{O}(\lambda/\text{poly}(n))$ in [20]. As NAS has much dense connections than ResNet, it allows larger LR. So our work makes towards the practice setting, and will continue to improve it later.

To Reviewer #2. 1) We empirically investigate i) more skip connections gives faster convergence and ii) shallow cells have faster convergence rate than deep cells. We first set all operations in NAS cell (normal and reduction cells) as convolution (3×3), and randomly select 0%, 37.5% and 62.5% operations as skip connections. Next, we stack 8 NAS cells to build a network and train on CIFAR10 with same settings. Fig. (a) demonstrates our result i). Moreover, Fig. 3 in [9] also testifies our result i). For result ii), to simply construction, we let each node in NAS cell only has one connection. To construct deep network, we use convolution to connect the current node with its previous node, i.e. $0\to1\to2\cdots\to5$. For shallow network, we connect the i-th node ($i=1,\cdots,5$) to the 0-th node with the same convolution. We also stack 8 cells and train them with same setting. Fig. (b) demonstrates our result ii). We will update it into revision.



2) Pooling operations also converge more slowly than skip connections. We consider function $h(p(g(x)))$, where g are layers before pooling p, h are subsequent layers and loss. Then we can prove ① $\|\frac{\partial h}{\partial p(g(x))}\frac{\partial p(g(x))}{\partial g(x)}\|_F^2<\|\frac{\partial h}{\partial g(x)}\|_F^2$, where the later denotes network h using skip connection. So pooling p reduces gradient and gives slower convergence. For max pooling, it only considers the maximum pixels and ignores others, directly giving ①. For average pooling, by derivation we have $\frac{\partial h}{\partial p(g(x))}\frac{\partial p(g(x))}{\partial g_{ij}(x)}=\frac{s_{ij}}{o^2}\frac{\partial h}{\partial g_{ij}(x)}$ with pooling size $o\times o$, where $s_{ij}(\le o^2)$ denotes how many times $g_{ij}(x)$ attends convolution. If pooling stride $s>1$, then $s_{ij}<o^2$. If $s=1$, for pixels near the edges, their $s_{ij}<o^2$. So ① always holds. Besides, we are sure that with pooling, Theorem 1 still holds for two-layered network and shows that convergence rate depends on skip connection heavier. For deeper networks, more efforts and time are needed for further derivation.

3) For Gram matrix singularity, we set all operations in NAS cell as convolution (3×3), and randomly select 0%, 40%, 80% operations in the shared path as skip connections to obtain cells A, B and C. Due to memory limitation, we use one NAS cell and find that smallest eigenvalues of Gram matrix in A, B and C on CIFAR10 are respectively 1.1×10^{-4}, 3.4×10^{-4} and 5.9×10^{-4}, showing benefits of skip connect to singularity. Then we fix the shared path with 40% skip connections, and randomly replace 0%, 40% and 80% convolutions in private path with zero operations. Then smallest eigenvalues become 3.4×10^{-4}, 1.3×10^{-4} and 8.7×10^{-5}, showing important of convolution in private path to singularity.

4) For depth-wise separable convolution (DSC), we can expect the same convergence rate as standard convolution (SC). Similar to SC, we formulate DSC as $\text{D}(W,X)=\sigma(W_p\Phi_p(\Phi_d(X)W_d))$. Similar to Φ in manuscript, $\Phi_d(X)$ ($\Phi_p(X)$) rearranges features in X along channel (feature) direction for depthwise convolution $\Phi_d(X)W_d$ (pointwise convolution $W_p\Phi_p(X)$). Then we replace convolution $\text{Conv}(W,X)$ in manuscript with $\text{D}(W,X)$, and follow our proof framework to prove same results: the convergence rate replies on skip connections heavier than other types of operations.

5) ReLU is not smooth at only one point, i.e. zero. But the measure of one point is zero. So almost sure, our smoothness assumption holds [arXiv:1706.03175, ICML'17]. Error (%) on CIFAR10 (ImageNet) of the mentioned references [1-3] are respectively 2.62 (24.8), 2.6 (24.6) and 2.7 (25.6). Ours is 2.31 (24.3) and thus is better. We will cite them.

To Reviewer #4. 1) Per your suggestion, we will use $\boldsymbol{X}^{(t)}\to\boldsymbol{X}^{(s)}$ $(0\le t\le s-1)$ to better illustrate the path of λ_s. As $\boldsymbol{X}^{(t)}(1\le t\le h-2)$ are shared by $\boldsymbol{X}^{(s)}(s\ge t)$, our subsequent explanation to Theorem 1 does not need to change.

Skip connection (SC) has formulation $X_{s+1}=X_s+F(X_s)$ where F is a function, e.g. convolution. So to fit ground truth Y of X_s, F only fits the residual $Y-X_s$ instead of Y. Recursively, we have $X_l=X_s+\sum_{t=s}^{l-1}F(X_t)$. Then the gradient of X_s is $\nabla_{X_s}\mathcal{E}=\nabla_{X_l}\mathcal{E}\cdot(1+\nabla_{X_s}\sum_{t=s}^{l-1}F(X_t))$ where $\mathcal{E}$ is loss. In most cases, $\nabla_{X_s}\sum_{t=s}^{l-1}F(X_t)$ is much smaller than 1, especially for along more training iterations, which means that SC propagates the main gradient flow and thus information flow.

2) Our theory could be extended to other losses, e.g. cross entropy. Here we choose square error loss because of its much simpler gradient computation compared with cross entropy. But with more extra efforts, we can follow our framework to establish similar results. Indeed, this is also one reason why recent works [18-21] on network convergence analysis focus on square error loss, as different losses reveal similar results but square error loss gives simpler derivation.

3) Appendix A.1 investigates the effects of $\lambda_1\sim\lambda_3$ to the performance of PR-DARTS. The results show the stable performance of PR-DARTS on CIAFR10 when tuning $\lambda_1\sim\lambda_3$ in relatively large ranges, e.g. $\lambda_1\in[10^{-2},1]$, $\lambda_2\in[10^{-4.5},10^{-2.5}]$ and $\lambda_3\in[10^{-4},10^{-1.5}]$. This is mentioned in line 341. So one can choose $\lambda_1\sim\lambda_3$ from the above ranges.

4) We divide operations in DARTS into skip-connection group and non-skip-connection group, and penalty their average active probabilities/weights. The error on CIFAR10 is 2.69% which is slightly better than 2.76% of vanilla DARTS but is worse than 2.58% of ours without path-depth-wise regularizer, showing the importance of independent stochastic gates.