Escaping Random Teacher Initialization Enhances Signal Propagation and Representation

Felix Sarnthein
Department of Computer Science
ETH Zurich
Zurich, Switzerland
safelix@student.ethz.ch

Sidak Pal Singh
Department of Computer Science
ETH Zurich
Zurich, Switzerland
sidak.singh@inf.ethz.ch

Antonio Orvieto
ELLIS Institute Tübingen and MPI-IS
Tübingen AI Center
Tübingen, Germany
antonio@tue.ellis.eu

Thomas Hofmann
Department of Computer Science
ETH Zurich
Zurich, Switzerland
thomas.hofmann@inf.ethz.ch

Abstract

Recent work shows that by mimicking a random teacher network, student networks learn to produce better feature representations, even if they are initialized at the teacher. In this paper, we characterize how students escape this global optimum and investigate how this process translates into concrete properties of the representations. To that end, we first describe a simplified setup and identify very large step sizes as the main driver of this phenomenon. Then, we investigate key signal propagation and representation separability properties during the escape. Our analysis reveals a two-stage process: the network first undergoes a form of representational collapse, then steers to a parameter region that not only allows for better propagation of input signals but also gives rise to well-conditioned representations. This might relate to the edge of stability and label-independent dynamics.

1 Introduction

The teacher-student setting has been a rich source of phenomena and has been used widely to better understand the behavior of modern deep neural networks (like their dynamics [9] or double descent [16]). But the influence of the teacher-student is not just restricted to theoretical endeavors. As a matter of fact, distillation for model compression is perhaps the most popular instance of this and has seen widespread practical adoption. The surprising properties of self-distillation, i.e. between models of identical architecture, has enabled significant advancements in self-supervised learning (SSL) for computer vision [10, 1].

An investigation on the role of self-distillation for self-supervised representation learning by Sarnthein et al. [19] brought to light an intriguing phenomenon; students distilled from randomly initialized teachers produce superior feature mappings as measured by linear probing accuracy. Additionally, the authors observe that initializing the student networks close to the teacher amplifies the effect. Instead of arriving at the solution corresponding to the teacher network, the student navigates away into a different region of the loss landscape which produces better representations. Switching to a finetuning setup, they observed that students lie on the border of a convex basins in the supervised loss landscape and contain sparse subnetworks, so-called lottery tickets [7, 8]. From this, raise the question whether this can be explained by label-independent training dynamics.

Mathematics of Modern Machine Learning Workshop at NeurIPS 2023.

Our aim is to contribute to the understanding of this phenomenon, to extract the teacher-student formulation from self-supervision and distillation, and pose it as a more general object of study. As opposed to theoretically analyzing the convergence of the student toward an optimum, we propose to investigate a special form of divergence away from the global optimum of the teacher. We believe that the theoretical accessibility of teacher-student formulations in combination with our empirical results yields an interesting question to narrow the gap between theoretical understanding and deep learning practice.

Outline. We first state the teacher-student formulation and introduce the key assumptions. Then, we empirically elucidate the dynamics of the escaping phenomenon, comparing the effect of finite step sizes versus stochasticity in the gradient. Using the derived minimal setting, we investigate the signal propagation properties of a network and show how they improve during the escape from the random teacher initialization. Finally, we investigate how representations are propagated and find that enhanced signal propagation alone does not explain the gain in representational power. Instead, early layers strongly separate the data already at initialization — a property that is preserved more efficiently after escaping the initialization.

2 Escaping Random Teacher Initializations

Typically, in teacher-student frameworks, a teacher network f_{θ_T} provides labels for an input dataset $D = \{x_i\}_{i=1}^n$. Given a student network f_θ, the aim is to minimize the discrepancy between the teacher and the student outputs:

$$\theta^{(S)} = \underset{\theta}{\operatorname{argmin}} \quad \mathcal{L}\big(f_\theta(D),\ f_{\theta_T}(D)\big)\,.$$

Instead of being pre-trained like in distillation, we take the teacher network to be random, i.e., its parameters θ_T are drawn from an underlying distribution, as in Kaiming or Xavier initialization. The student network is then initialized in close proximity to its teacher

$$\theta^{(0)} = \theta_T + \varepsilon\widetilde{\theta}\,,$$

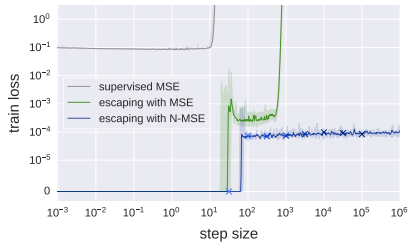
where $\widetilde{\theta}$ is yet another draw from the same initializing distribution and $\varepsilon \sim \mathcal{N}(0; \sigma_\varepsilon I)$. If $\varepsilon = 0$, then there is nothing to optimize since the student is at a global minimum. So, in this procedure, typically one starts close to the teacher, with a small $\sigma_\varepsilon \approx 10^{-10}$, which together with the finite step size η and batch size B, allows for a departure from the teacher and gradient dynamics:

$$\theta^{(i+1)} = \theta^{(i)} - \eta \cdot \nabla_\theta \mathcal{L}\big(f_{\theta^{(i)}}(D_B),\ f_{\theta_T}(D_B)\big)\,.$$

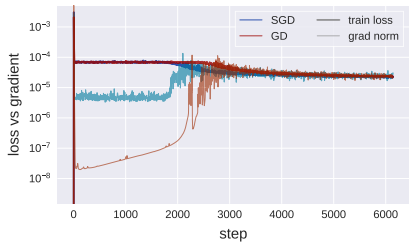
Representations. How does this escaping process act on the student network? As the the teacher network cannot convey any information about the underlying dataset labels, we investigate the representations generated by the student. For now, we reproduce the phenomenon of [19] in Fig. 1c and (d), but in our more concise setting. As is common in self-supervised learning, we measure the feature quality by applying a linear probe, i.e. we train and evaluate a logistic regression on $\varphi(D)$, using the models' encoders as feature map. We observe that the linear probing accuracy collapses initially, then recovers, and even outperforms the teacher. Next, we aim to understand the driver behind this phenomenon and in Sec. 3, we investigate other properties of the feature representations.

Empirical Setup. We simplify the setup of [19]: (1) we append only one linear layer ($d_{out} = 512$) to a *VGG11* [21] encoder, both initialized according to K. He et al. [12], (2) we use plain SGD without momentum or weight decay instead of AdamW and (3) we replace the cross-entropy (CE) with a mean-squared error (MSE) loss function. We use *CIFAR10* [14] and $B = 256$ if not stated otherwise.

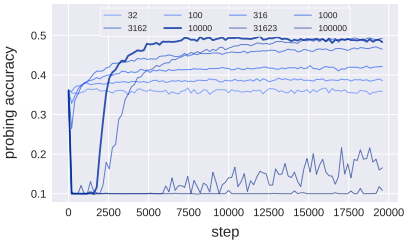
Step size warm-up analysis. As the use of Adam [13] in [1, 19] disguises a suitable step size η, we start by determining it empirically. We follow [22] and perform an LR-Range-Test in Fig. 1a. Exponentially increasing the step size from 10^{-6} to 10^6 for 5 epochs captures interactions between the loss landscape and step size. While the supervised loss diverges at $\eta = 10$, the divergence out of the random teacher initialization occurs in two steps, the second being at a large step size $\eta = 500$. We try different ways of stabilizing dynamics with such large step sizes in App. A and reuse the feature normalized loss N-$\mathcal{L}(x_S, x_T) = \mathcal{L}(x_S/||x_S||_2, x_T/||x_T||_2)$, from the original DINO setup [1]. The resulting N-MSE loss and its gradient are conveniently bounded from above and therefore enable the use of uncommonly large step sizes η such as `1e4`!



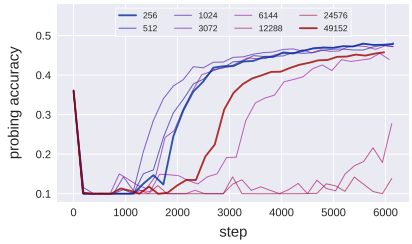
(a) Step size divergence analysis.



(b) Escaping dynamics of SGD and GD.



(c) Step size vs linear probing.



(d) Batch size vs linear probing.

Figure 1: Escaping random teacher initializations.

Role of step size. Using the insight from Fig. 1a, we select a few exponentially spaced step size candidates, attempt to escape under N-MSE and measure linear probing accuracy. In Fig. 1c, we observe no progress for small $\eta = 32$, as predicted by Fig. 1a. For increasing η (light to dark), we reproduce the phenomenon from [19] with our simplified model and additionally observe correlation between step size and probing performance. Note that for $\eta \in \{3162, \texttt{1e5}\}$, our model undergoes a collapse during training from which it can recover, whereas $\eta > \texttt{1e4}$ results in unrecoverable collapse during 20k steps.

Optimization stochasticity: integral or superfluous? Given the necessity of noise to escape, we investigate if there is something crucial about the gradient noise, which for SGD has been shown to exhibit interesting properties [23, 2]. In Fig. 1d, we experiment with increasing batch sizes up to full-batch ($B = 49152$, reduced CIFAR10 set) and fixed $\eta = \texttt{1e4}$. We observe that with fewer noise, the model takes longer to uncollapse in 6k steps. Most importantly, we observe a successful escape and better probing for full-batch gradient descent with no gradient noise and only large learning rates.

Escaping Dynamics. To understand the full-batch ($B = 49152$) vs mini-batch ($B = 256$) escaping dynamics better, we plot training loss and gradient norm in Fig. 1b. Within the first three steps, the very large step size causes a loss spike and overshoots into a plateau, where loss and gradient magnitude are again relatively small. Interestingly, this plateau corresponds to low probing performance in Fig. 1d and rank-collapsed representations, as we shall see in Sec. 3. Focusing on GD nicely uncovers how the gradient magnitude progressively increases until it reaches an oscillatory state. Then, the probing performance starts to rise and significantly exceed its value at initialization.

3 Signal Propagation & Representation

Motivated by the observed representational collapse in Fig. 1, we investigate how the student network incurs a transition in the quality of representations as it proceeds from the teacher θ_T, via the point of collapse $\theta^{(C)}$, to the student after 100 epochs $\theta^{(S)}$. To that end, we focus on signal propagation and representation properties, following [5, 2].

Approach. We will denote the matrix of representations over the dataset D of size n at some layer l of a given network, with width d_l as the matrix $\varphi_{\theta,l}(D) = \mathbf{X}^{(l)} \in \mathrm{R}^{d_l \times n}$. Investigating the three checkpoints θ_T, $\theta^{(C)}$, and $\theta^{(S)}$, we denote the corresponding representations as $\mathbf{X}_T^{(l)}$, $\mathbf{X}_C^{(l)}$, and $\mathbf{X}_S^{(l)}$.

Effective Rank. We compute the effective rank $\mathrm{erank}(\mathbf{X}) = \exp(H(p(\mathbf{X})) \leq \mathrm{rank}(\mathbf{X})$, where $H(p) = \sum p_k \log p_k$ is the entropy of $p_k(\mathbf{X}) = \frac{\sigma_k(\mathbf{X})}{\sum \sigma_j(\mathbf{X})}$, the normalized singular value distribution of X [18]. This is a continuous approximation of the rank, which reaches the exact rank if the rows of $\mathbf{X}^{(l)}$ are orthogonal. In Fig. 2a, we display the layerwise $\mathrm{erank}(\mathbf{X}^{(l)})$ for these three points in the

landscape. Starting with the teacher, we observe that the representations are rank deficient across the layers. This is well known and can be attributed to a combination of (lower) input rank and improper signal propagation [20, 17]. Next, we consider the effective rank behavior of $\theta^{(C)}$ and observe that the rank decreases to single digits. Such a rank deficiency, means that multiple samples are mapped to one common, or a few linearly dependent representations. Finally, $\theta^{(S)}$ consistently retains more signal per propagation step than the teacher network. Crucially, this higher rank implies that more useful information to distinguish between data points is being retained.

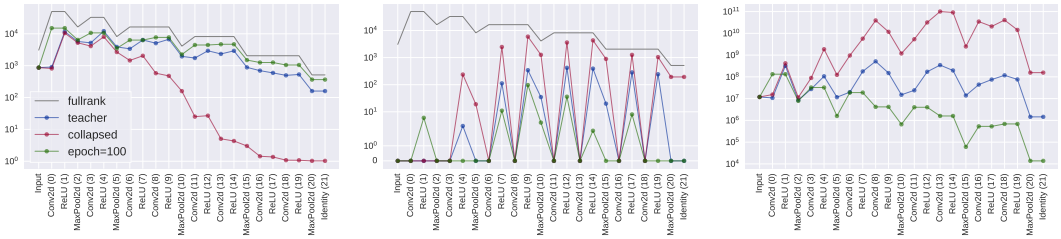
Dead Neurons. If we go a step further, we can ask how the representations are transformed to low rank. Is it due to some of the weight matrices exhibiting rank deficiencies? Or do the representations across layers get organized such that less number of neurons are fired? We find that the latter case of rank collapse is more prominent, where more neurons are inactive across the entire dataset D, which means that rows of post-activation $\mathbf{X}^{(l)}$ are zero. This is observable quite clearly in Fig. 2b, where we count the number of rows $= \vec{0}$ in $\mathbf{X}^{(l)}$, revealing that some neurons are dead already at initialization $\Theta^{(T)}$. Furthermore, up to 60% of the neurons are inactive in the collapsed state $\theta^{(C)}$. But remarkably, the model manages to recover from this through its large learning gradient dynamics.

Condition Number. Finally, we estimate the condition number of $\mathbf{X}^{(l)} = \varphi_l(D)$, as a metric of label-independent feature quality. We approximate $\mathrm{cond}(\mathbf{X}) = \sigma_{\max}/\sigma'_{\min}$, where $\sigma'_{\min} = \min_k \sigma_k$, s.t. $\sum_{j=0}^{k} p_j > 0.999$. This means that we select the k-th singular value that captures 99.9% of the variance in the spectrum as $\sigma'_{\min}$. We observe the common trend that $\theta^{(S)}$ provides better conditioned features than at collapse $\theta^{(C)}$ and initialization $\theta^{(T)}$. Better conditioning of the data means that optimization on a downstream task such as finetuning or linear probing is is well behaved. Therefore, this could explain the improved linear probing.

Separability of Representations. We have seen in the above sections that the escaping process leads to more well-behaved representations, namely in the sense of higher effective rank and better conditioning. While this gives an explanation towards better linear probing performance, we are going to investigate the whether the organization of the representations enables better discriminability and thus higher probing performance in this section.

We now make use of the underlying label information, which, notably, was *not utilized in the escaping process*[1]. To that end, we investigate the separability of representations using classical metrics arising in Fisher discriminant analysis such as the (trace of) between/within cluster and total covariance Σ_b, Σ_w, Σ_t [6, Chapter 4.11]. In particular, the separability discriminant $\mathrm{tr}(\Sigma_w^{\dagger}\Sigma_b)$ indicates the variance of the representations within classes (or clusters) relative to the variance of the cluster center representations. This has lately been used in understanding representations of deep networks [11].

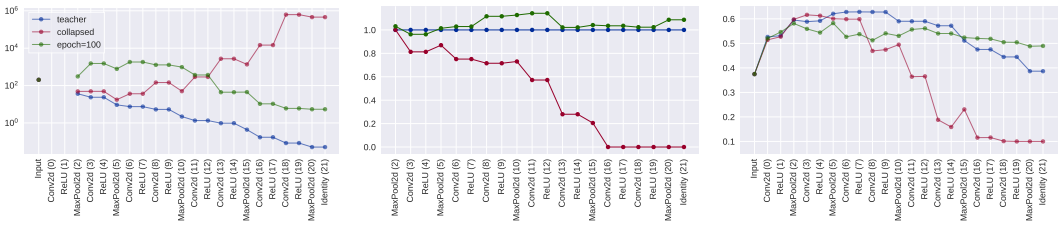
In Fig. 3b, we see how the student produces better separable features than the teacher network, by considering its discriminant relative to the teacher network. Furthermore, the collapsed parameterization $\theta^{(C)}$ can not discrimination between representations of different classes. Likewise, in Fig. 3a, we see that the between-cluster variance of $\theta^{(C)}$ is much higher than the student or the teacher network. Fig. 3c shows that the teacher encodes good features in its early layers, but deteriorates in later layers, presumably due to a loss of signal. Overall, these metrics precisely show the benefit wrought upon by the escaping mechanism that result in higher probing performance.



(a) Effective rank. (b) Dead neurons. (c) Condition number.

Figure 2: Signal Propagation

[1]While here we make use of labeled information, one could also imagine first clustering the representations and then using the cluster membership as a proxy of the labels.



(a) Between-cluster variance. (b) Separability discriminant. (c) Linear probe.

Figure 3: Feature Representation

4 Discussion

We described the inner workings of the random distillation procedure [19] from the same initialization. First, we found that a very large stepsize of `1e4` is the main driver behind this phenomenon, even under full-batch gradient descent. Then we showed how such a procedure inherently results in a student network that has better signal propagation qualities, in terms of higher effective rank and well-conditioning of the representations. Finally, we highlighted how this procedure leads to better organization of the representations, by enhancing their separability and ultimately linear probing accuracy.

Besides, now that we understand phenomenology, carrying out a theoretical characterization would be the next natural step. Our description of the training dynamics might allow to draw a connection to the catapult mechanism [15] or edge of stability phenomenon [3, 4]. Treating the escaping process as a data-dependent initialization, we propose to further investigate the finetuning behavior as has been started in [19]. We hope that our work fosters further research into the mysterious mechanisms that underlie the loss landscapes of deep networks and ultimately contributes to bridging the gap between theoretical understanding and deep learning practice.

Acknowledgments and Disclosure of Funding

Antonio Orvieto acknowledges the financial support of the Hector Foundation. Sidak Pal Singh would also like to acknowledge the financial support from Max Planck ETH Center for Learning Systems.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. “Emerging Properties in Self-Supervised Vision Transformers”. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021. URL: `https://arxiv.org/abs/2104.14294`.

[2] Feng Chen, Daniel Kunin, Atsushi Yamamura, and Surya Ganguli. “Stochastic Collapse: How Gradient Noise Attracts SGD Dynamics Towards Simpler Subnetworks”. In: *arXiv preprint arXiv:2306.04251* (2023).

[3] Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. “Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability”. In: *9th International Conference on Learning Representations (ICLR)*. 2021. URL: `https://arxiv.org/abs/2103.00065`.

[4] Alex Damian, Eshaan Nichani, and Jason D. Lee. “Self-Stabilization: The Implicit Bias of Gradient Descent at the Edge of Stability”. In: (2022). URL: `https://arxiv.org/abs/2209.15594v1`.

[5] Shibhansh Dohare, J. Fernando Hernandez-Garcia, Parash Rahman, Richard S. Sutton, and A. Rupam Mahmood. *Loss of Plasticity in Deep Continual Learning*. 2023. arXiv: `2306.13812 [cs.LG]`.

[6] Richard Duda, Peter Hart, and David G.Stork. “Pattern Classification”. In: vol. xx. 2001.

[7] Jonathan Frankle and Michael Carbin. "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks". In: *7th International Conference on Learning Representations (ICLR)*. 2018. URL: https://arxiv.org/abs/1803.03635.

[8] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. "Linear Mode Connectivity and the Lottery Ticket Hypothesis". In: *37th International Conference on Machine Learning (ICML)*. 2019. URL: https://arxiv.org/abs/1912.05671.

[9] Sebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. "Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup". In: *Advances in neural information processing systems* 32 (2019).

[10] Jean Bastien Grill et al. "Bootstrap your own latent: A new approach to self-supervised Learning". In: *34th Conference on Neural Information Processing Systems (NeurIPS)*. 2020. URL: https://arxiv.org/abs/2006.07733.

[11] Hangfeng He and Weijie J Su. "A law of data separation in deep learning". In: *Proceedings of the National Academy of Sciences* 120.36 (2023), e2221704120.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *IEEE International Conference on Computer Vision (ICCV)*. 2015. URL: https://arxiv.org/abs/1502.01852.

[13] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations (ICLR)*. 2014. URL: https://arxiv.org/abs/1412.6980.

[14] Alex Krizhevsky and Geoffrey Hinton. *Learning multiple layers of features from tiny images*. Tech. rep. University of Toronto, 2009. URL: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[15] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. "The large learning rate phase of deep learning: the catapult mechanism". In: (2020). URL: https://arxiv.org/abs/2003.02218v1.

[16] Mohammad Pezeshki, Amartya Mitra, Yoshua Bengio, and Guillaume Lajoie. "Multi-scale feature learning dynamics: Insights for double descent". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 17669–17690.

[17] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. *Exponential expressivity in deep neural networks through transient chaos*. 2016. arXiv: 1606.05340 [stat.ML].

[18] Olivier Roy and Martin Vetterli. "The effective rank: A measure of effective dimensionality". In: *15th European Signal Processing Conference*. 2007. URL: https://ieeexplore.ieee.org/abstract/document/7098875.

[19] Felix Sarnthein, Gregor Bachmann, Sotiris Anagnostidis, and Thomas Hofmann. "Random Teachers are Good Teachers". In: *40th International Conference on Machine Learning (ICML)*. 2023. URL: https://arxiv.org/abs/2302.12091.

[20] Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. "Deep information propagation". In: *arXiv preprint arXiv:1611.01232* (2016).

[21] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *3rd International Conference on Learning Representations (ICLR)*. 2014. URL: https://arxiv.org/abs/1409.1556.

[22] Leslie N. Smith. "Cyclical Learning Rates for Training Neural Networks". In: *Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017* (2015), pp. 464–472. URL: https://arxiv.org/abs/1506.01186v6.

[23] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. "The Anisotropic Noise in Stochastic Gradient Descent: Its Behavior of Escaping from Sharp Minima and Regularization Effects". In: *36th International Conference on Machine Learning, ICML 2019* 2019-June (2018), pp. 13199–13214. URL: https://arxiv.org/abs/1803.00195v5.

A Stabilizing

(a) linear probe vs training epoch and step size. (b) step size schedules.

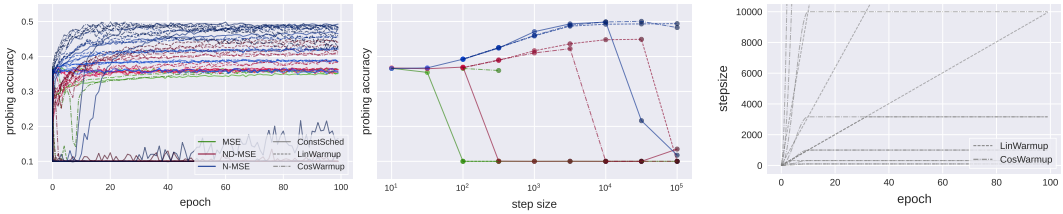


Figure 4: Stabilizing the large step size dynamics using MSE, ND-MSE and N-MSE loss functions, as well as different warm-up schedules. The dynamics can be stabilized under a detached, normalized MSE (ND-MSE) and slow linear warmp of the step size. The detaching the norm from the backpropagation path is a common pattern in BatchNorm and LayerNorm.