Exponential Moving Average of Weights in Deep Learning: Dynamics and Benefits

Anonymous authors
Paper under double-blind review

Abstract

Weight averaging of Stochastic Gradient Descent (SGD) iterates is a popular method for training deep learning models. While it is often used as part of complex training pipelines to improve generalization or serve as a 'teacher' model, weight averaging lacks proper evaluation on its own. In this work, we present a systematic study of the Exponential Moving Average (EMA) of weights. We first explore the training dynamics of EMA, give guidelines for hyperparameters tuning, and highlight its good early performance, partly explaining its success as a teacher. We also observe that EMA requires less learning rate decay compared to SGD since averaging naturally reduces noise, introducing a form of implicit regularization. Through extensive experiments, we show that EMA solutions differ from last-iterate solutions. EMA models not only generalize better but also exhibit improved i) robustness to noisy labels, ii) repeatability, iii) calibration and iv) transfer learning. Therefore, we suggest that an EMA of weights is a simple yet effective plug-in to improve the performance of deep learning models.

1 Introduction

The performance of modern deep learning models is tightly linked to their training. In order to converge to a good solution, reducing the noise coming from stochastic updates is eventually required. For example, Stochastic Gradient Descent (SGD) (Robbins & Monro, 1951; Bottou, 2010) needs a carefully tuned learning rate and decay schedule, while adaptive variants such as Adam (Kingma & Ba, 2014) essentially decay the learning rate based on gradients retrieved: while a large learning rate is initially required to obtain fast training and good generalization, only when the learning rate is low enough does the model actually converge to a good solution. In an orthogonal approach to learning rate decay, a standard way to reduce noise in *convex* optimization is (tail) averaging (Polyak & Juditsky, 1992): if SGD stops making progress because of stochastic noise, a more accurate solution can be retrieved despite high learning rates by averaging the last iterates.

While the theoretical gains of averaging are less clear in the non-convex setting, Weight averaging (WA) is also popular in deep learning, and has been explored primarily in two ways: WA inside the training loop as a teacher model, and WA outside the training loop to improve generalization. In the first case, an Exponential Moving Average (EMA) of parameters is used as a teacher in Student-Teacher frameworks, for example in popular representation learning methods (Tarvainen & Valpola, 2017; Grill et al., 2020). The averaged model provides more accurate and consistent predictions during training, which the student uses for training. For the second case, Stochastic Weight Averaging (SWA) (Izmailov et al., 2018) uses an average of multiple checkpoints along the SGD trajectory to improve the generalization of the final model, arguing that it finds flatter solutions than SGD. Note that SWA does not affect optimization since the averaged model is not used in the training loop.

Despite the popularity of EMA teachers, the properties of EMA models have not been studied thoroughly. Previous works (Grill et al., 2020; He et al., 2020) mainly justify its effectiveness by enhanced consistency and stability of predictions during training, and often mix EMA with other mechanisms as part of a complex

framework, making it impossible to disentangle the impact of averaging. EMA has mainly been used as a teacher, leaving aside the potential of EMA itself to improve the final solution in favor of SWA.

In this work, we focus on EMA models outside of the training loop, exploring their training dynamics and benefits in generalization and beyond. By doing so, we unveil new reasons why EMA are such good teachers: we find that the solutions reached when reducing stochastic noise by averaging are different than by learning rate decay. They improve in robustness to label noise, calibration, repeatability and transfer learning. In a nutshell, we ask the following question:

What are the properties of weight averaging when training deep neural networks?

To answer this question, we empirically study weight averaging during a (momentum) SGD trajectory by means of an Exponential Moving Average (EMA). More specifically, if the SGD models form a trajectory $(x_t)_{t\geq 0}$, the corresponding EMA models would be obtained by taking, for some $\alpha \in [0, 1]$:

$$\mathbf{x}_0^{\text{EMA}} = \mathbf{x}_0, \text{ and } \mathbf{x}_{t+1}^{\text{EMA}} = \alpha \mathbf{x}_t^{\text{EMA}} + (1-\alpha)\mathbf{x}_{t+1}. \tag{1}$$

Keeping an EMA model is simple, has minimal overhead, and can easily be plugged into any existing pipeline. We study EMA models outside of the training loop, such that they have no effect on the underlying SGD trajectory. We split our study in two parts, described in the two paragraphs below.

Training dynamics of EMA (Sec. 3). We find that weight averaging reduces noise and can replace the last phase of learning rate decay, while at the same time enabling implicit regularization via stochastic noise. More specifically, **(i)** While last-iterate SGD requires decaying the learning rate to nearly 0 for convergence, averaging reduces noise and yields good solutions under reasonably high learning rates, which we argue promotes learning more general representations. We propose a one-shot tuning of the strength of implicit regularization by using cosine annealing of the learning rate and early stopping. **(ii)** We highlight the impressive performance of EMA in the early stages of training and postulate this observation as a key reason for the success of popular EMA teachers. Combined with early stopping, the EMA model can reduce the compute budget by sparing the last phase of SGD training at low learning rates. **(iii)** We look into Batch Normalization (BN) in EMA models and find that it is the limiting factor when choosing the EMA decay. If BN statistics are recomputed, larger averaging windows can be used and actually may improve further generalization.

Properties of the final EMA model (Sec. 4). We find that the solutions reached with EMA and early stopping are different from the baseline solution obtained by last-iterate momentum SGD, and conjecture that this is due to implicit regularization via stochastic noise. More specifically, **(i)** in extensive experiments on image classification tasks we find a consistent improvement in **generalization** using an EMA model. **(ii)** We also find a great improvement in **robustness to label noise** in training data, as the implicit regularization largely prevents memorization of wrong labels. A simple EMA model proves to be competitive with specialized methods for robust training to label noise. **(iii)** We compare the EMA model to the SGD baseline in a number of other metrics and find that it improves considerably in **calibration**, **prediction disagreement** and **transfer learning**.

2 Related Work

Weight averaging during deep learning training is already a popular method that leads to better practical performances. Yet, it lacks a systematic study, since contributions are mostly scattered in domain-specific literature. Besides, the use of weight averaging is often justified by intuitive alleged properties which are rarely investigated in isolation. In this section, we review different areas in which weight averaging is used, along with the corresponding alleged folklore benefits. We then test them, as well as give alternative explanations for why it is useful in the remaining sections.

Weight averaging to improve generalization. Averaging the iterates during a trajectory has a long history in stochastic approximation (Ruppert, 1988; Polyak, 1990; Polyak & Juditsky, 1992), and its correct use and understanding have been an active area of research since (Bach & Moulines, 2011; Dieuleveut et al., 2017; Lakshminarayanan & Szepesvari, 2018; Mücke et al., 2019; Gadat & Panloup, 2023). Geometric averaging (EMA but with more weights on old iterates) can also be connected with a form of explicit regularization Neu

& Rosasco (2018). Yet, these methods assume quadratic or (strongly) convex objectives, and so do not apply to deep learning training.

In deep learning, a popular averaging method is Stochastic Weight Averaging (SWA) (Izmailov et al., 2018). SWA keeps a uniform average of checkpoints during the final epochs of an SGD trajectory, while holding a reasonably high and constant learning rate. SWA is argued to find flatter solutions than SGD, thus generalizing better to unseen data. A potential explanation is that the loss function near a minimum is often asymmetric, sharp in some directions and flat in others. While SGD tends to land near a sharp ascent, averaging iterates biases solutions towards a flat region (He et al., 2019).

Many extensions of SWA have been proposed for specialized tasks (Gupta et al., 2020; Li et al., 2022; Kaddour, 2022), and in particular semi-supervised learning (Athiwaratkun et al., 2019), low-precision training (Yang et al., 2019) and domain generalization (Cha et al., 2021). The latter introduces enhancements such as dense averaging (*i.e.*, every iteration) and overfit-aware sampling by tracking validation loss. The flatness argument has also been leveraged for robustness: weight averaging on top of adversarial training helps finding flatter minima and boosts adversarial robustness. This has been shown using both SWA (Chen et al., 2021) and EMA (Gowal et al., 2020; Rebuffi et al., 2021). EMA has also been studied in minimax optimization, with applications to GANs (Yaz et al., 2019). Finally, many works use averaging as part of their implementation but do not emphasize it or discuss its effect. Such works are hard to review since they are not explicitly listed as working on averaging, but include for instance Berthelot et al. (2019); Sohn et al. (2020); Oord et al. (2018); Oquab et al. (2023), which rely on EMA or uniform averaging, sometimes replacing the decay of learning rate.

Overall, while weight averaging seems to improve generalization, existing studies mostly focus on SWA. We close this gap, and show similar results for EMA in this work.

Weight averaging in Student-Teacher methods. Consistency training is a popular technique for learning with unlabeled data (Laine & Aila, 2016; Berthelot et al., 2019), based on generating pseudo-labels during training, often through a *teacher* model, which does not receive gradient updates. Mean Teacher (Tarvainen & Valpola, 2017) first proposed to use an EMA of model weights as a teacher, such that $\theta' = \text{EMA}(\theta)$, in a method for semi-supervised image classification. EMA has since become a popular choice for teacher models, used for tasks such as semi-supervised semantic segmentation (French et al., 2019), unsupervised domain adaptation (Hoyer et al., 2022), continual adaptation (Wang et al., 2022), and robustness to label noise Liu et al. (2020); Nguyen et al. (2019). On the other hand, SWA requires recomputing batch norm (BN) statistics for the averaged model with a full pass over the train set, thus making it unfit for its online use (*e.g.*, as teacher). While these works find that EMA teachers are beneficial and provide accurate pseudo-labels, they do not specifically study the properties of EMA models.

In self-supervised learning, EMA plays a central role in a handful of popular frameworks. BYOL (Grill et al., 2020) employs consistency training with an EMA teacher (*a.k.a.* self-distillation) to learn visual representations from unsupervised data. MoCo (He et al., 2020) rebrands the EMA teacher as a momentum encoder and proposes a student-teacher framework with a contrastive learning objective. CURL (Laskin et al., 2020) applies the same idea to learn unsupervised representations for reinforcement learning. These methods attribute the effectiveness of EMA to smoother changes in target representations, maintaining consistency and stability, rather than the quality of the representations. DINO (Caron et al., 2021) explores self-distillation in Transformers and studies, for the first time, the training dynamics of the EMA teacher, including the key observation that the teacher consistently outperforms the student during training.

In summary, weight averaging is a key component of student-teacher methods. EMA is generally preferred over other averaging methods (such as SWA) to avoid recomputing BN stats. In this work, we investigate the relation between averaging window and BN statistics, and show that this is actually only the case for short averaging windows.

3 Insights on Weight Averaging during Training

Although EMA models are built from SGD iterates, their dynamics during training and final solutions are very different. We argue that EMA is a simple, lightweight and effective plug-in to SGD training.

3.1 Training with EMA

Computation overhead. The overhead of using an EMA of weights outside of the training loop is generally very low, as it only requires keeping a running average of parameters and possibly evaluating every epoch. Moreover, the running average can be updated every T steps instead of after every parameter update. We set $T = 16$ by default and find no difference to $T = 1$ in the results. In terms of computation, the optimization step remains the dominating factor. In terms of memory, keeping an additional set of weights is feasible for most deep learning models used in practice, other than foundation models. For example, a ResNet-50 (23.7M parameters) requires 90.43 MB of storage.

Hyperparameters tuning. There are two main sources of potential tuning overhead when training averaging models: 1) deciding on the averaging window and 2) tuning the final learning rate, a sensible hyperparameter crucial for the final performance. The averaging window for an EMA is determined by the decay factor α. An EMA naturally avoids the need for (1), since we can simultaneously keep multiple EMA models with different decays to compare different averaging windows. We prevent (2) by using cosine annealing of the learning rate and finding the best early stopping epoch on a validation set. This allows us to search for (1) and (2) on the go, training only once. Admittedly, keeping multiple EMAs (say, M) to avoid tuning does increase the overhead by a factor M, but it is still a tiny fraction of the computation time for small enough values (*e.g.*, $M = 5$). With this scheme, we need to decide on epoch budget, number and selection of EMA decays, and to search for the best initial learning rate, as usual for regular training of DNNs.

3.2 Implicit Regularization with SGD Noise and Learning Rate Schedule

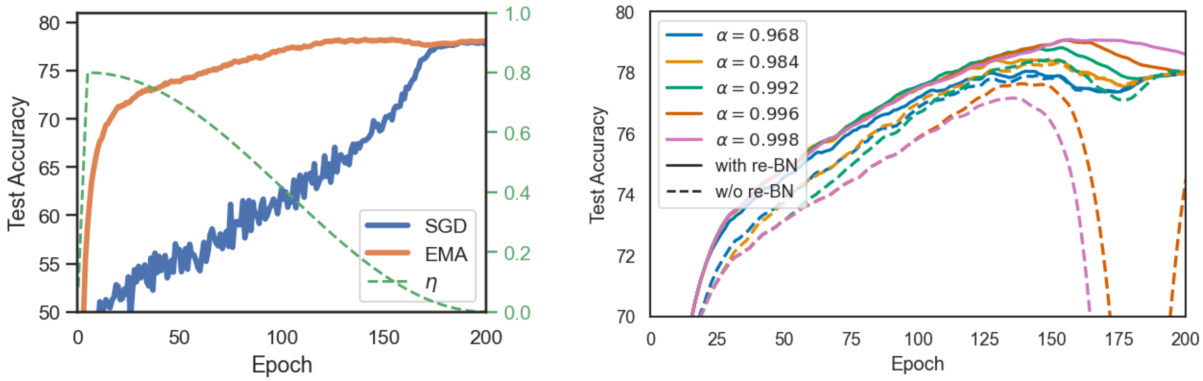
Noisy SGD updates are argued to bias solutions towards flatter regions that are believed to generalize better, partly explaining the success of deep learning (Keskar et al., 2016). This implicit regularization effect makes a case for large learning rates and small mini-batches (Pesme et al., 2021; Even et al., 2023). Nonetheless, standard training of DNNs requires decaying the learning rate to reduce stochastic noise and converge to a good solution. Averaging during training is an alternative way of reducing noise and reach a good solution without too much learning rate decay. This allows to freely tune the final learning rate to control the strength of implicit regularization, while still converging to a good solution within the neighborhood.

We demonstrate the dynamics of EMA models in Fig. 1a, where we use a continuous decaying of the learning rate η with cosine annealing and track the validation accuracy of the EMA model to find the best early stopping epoch (for implementation details see Sec. 4.1). We first highlight that the EMA model outperforms SGD throughout training. We also observe that the best EMA model is obtained when averaging updates with a reasonably high learning rate, benefiting from a stronger implicit regularization. The EMA accuracy rises fast at first, then slowly increases as η is continuously decayed (and so does the strength of regularization), peaks at epoch 150, and finally deteriorates as η is decayed further. Early stopping while sweeping through the learning rate values allows for a one-shot tuning of the regularization strength. Finally, the EMA model matches the SGD sequence when the iterates don't advance ($\eta \to 0$). For more examples of EMA dynamics, see Fig. 2 and App. A. As we will see in Sec. 4, the solutions reached by EMA and SGD are different: the implicit regularization of averaging improves generalization and promotes learning more general representations.

We emphasize that the EMA solutions not only generalize similarly or better than SGD, but also require fewer training epochs. In our experiments with cosine annealing, early stopping for EMA was always at $< 3/4$ of the epochs budget (see App. B). This suggests that the last phase of SGD training is *mostly wasteful*, as the iterates are already around a good solution that cannot be accessed (without averaging) because of stochastic noise. With averaging, there is no need to decay the learning rate that much, and the entire last phase of training can be spared.

3.3 EMA in early training

A key difference in the training dynamics of EMA and SGD models is the early-stage performance: **EMA models are very effective early in training**, as shown in Fig. 1a. Its remarkable performance after just a few epochs partly explains the success of EMA teachers. While popular self-supervised methods (Grill et al.,



(a) EMA vs SGD

(b) EMA by decay α, with vs w/o recomputing BN stats

Figure 1: CIFAR-100 on ResNet-18. **Left:** EMA vs SGD baseline, and learning rate (μ). EMA is the best among the 5 EMA models at any given epoch, without recomputing BN stats. We observe that EMA dominates momentum SGD and has a good performance since early on. EMA peaks at epoch 150, at the *optimal* μ, and then deteriorates. **Right:** Breakdown of the 5 EMA models per decay (with and without BN recomputation after every epoch). EMAs with the largest averaging windows fail unless BN stats are recomputed. Sliding window of 5 used for smoothing. All results are the mean of 3 runs.

2020; He et al., 2020) attribute the benefit of a slow-moving average to an improved consistency between predictions during training, we argue that the improved quality of the EMA representations also plays a crucial role in their frameworks. Thanks to noise reduction from averaging, the EMA model can achieve notable performances while keeping a large learning rate for fast progress. Student-teacher frameworks leverage this fact and distill knowledge from the EMA teacher.

Instead of knowledge distillation, a tempting idea is to regularly bootstrap the SGD model with EMA parameters. We investigate this (see App. C), but conclude that *bootstrapping with the EMA does not offer any benefit*. The EMA model is simply a good point within the local neighborhood of the latest iterates. After bootstrapping, noisy SGD updates quickly take over and deteriorate the model performance. Therefore, *distillation methods are a more effective way of leveraging EMA's early performance to improve training.*

3.4 Batch Norm Statistics and EMA decay

Batch Normalization (BN) presents a challenge for weight averaging. By default, the EMA model uses BN statistics (mean and standard deviation of each activation) from the current batch, but the cross-sample dependency may harm generalization. Cai et al. (2021) improve EMA teachers by keeping a moving average of the BN stats of the student, which we use in our implementation. In another approach, SWA (Izmailov et al., 2018) recomputes BN statistics for the final averaged model with an additional full pass over the train data after training. For the online use of an EMA model, however, recomputing BN stats during training would imply a significant overhead.

We investigate the optimal averaging window size (*i.e.*, decay α) for EMA and find different behaviors for the running average of model parameters and of BN stats. In particular, **model parameters tolerate larger averaging window sizes than BN statistics**. As shown in Fig. 1b, the EMA model can *diverge* when a very slow decay (*i.e.*, large α) is used. Interestingly, if we recompute the BN stats of that same EMA model (once after every epoch) we recover full performance, indicating that it is BN that breaks an EMA model as we increase the size of the averaging window. In fact, if recomputing BN stats, averaging weights tends to benefit from using very large averaging windows. We also observe that recomputing BN stats always improves generalization.

The EMA decay is usually set to $\alpha \in [0.9, 0.999]$, and the optimal value will be task dependent. For an online use of EMA, when recomputing BN stats periodically may be undesirable, a faster decay may be used to avoid divergence. On the other hand, when averaging to improve final performance, it is preferable to use a

slower decay and recompute BN stats after training. Models that do not use BN naturally avoid this problem (*e.g.*, VGG-16). While we focus on CNNs, future work could investigate EMA in Transformer models, which also avoid BN.

4 Results

4.1 Experimental Setup

We perform experiments on several image classification datasets (CIFAR-10, CIFAR-100, Tiny-ImageNet (Le & Yang, 2015)) with various network architectures (ResNet-18 (He et al., 2016), WideResNet-28-10 (Zagoruyko & Komodakis, 2016), VGG-16 (Simonyan & Zisserman, 2014)). We always use SGD with Nesterov Momentum of 0.9 for training. The epochs, batch size and weight decay are fixed (see details in App. F). For the learning rate schedule, we use a linear warmup during the first 5 epochs and then decay with cosine annealing, and search for the best initial learning rate. We always report the mean of 3 independent runs.

To perform a rigorous study, we stress the importance of using a hold-out set for hyperparameter selection. Unfortunately, most image classification benchmarks do not include a standard validation set. We define random 80/20 splits of the training set for train and validation respectively and perform hyperparameter optimization on the validation set, including the early stopping epoch for EMA (without BN stats recomputation). Finally, we train on the full training data using the selected hyperparameters and evaluate on the test set. Note that early stopping is not tuned again on the test set. Also note that most methods would technically require this train/evaluation split: the best step-size for SGD should be selected on a validation set for instance. We explicitly use one here instead of choosing the best performance on the test set (as is often done) since our EMA training pipeline also relies on early stopping, which we do not think would be relevant to choose directly on the test set.

The EMA introduces one hyperparameter, the decay rate α, which governs how fast the moving average forgets past iterations. Since the EMA is outside of the training loop, we can optimize α in a single training run by keeping 5 parallel EMA models. We fix the decays to $\alpha \in [0.968, 0.984, 0.992, 0.996, 0.998]$, and use an EMA sampling period of $T = 16$ steps, to reduce the overhead at no cost in performance. Note that using $T > 1$ affects the effective decay, which becomes $\alpha^{1/T}$ (*e.g.*, $0.984^{1/16} \approx 0.999$). We also use a decay warm-up for a faster EMA update in the first epochs, as $\min(\alpha, \frac{t+1}{t+10})$ at time t. For EMA's BN statistics we follow Cai et al. (2021).

In our experiments, we compare *Baseline* against *EMA*. We refer as *Baseline* to the momentum SGD model on which we perform the EMA. For the EMA we consider two different early stopping epochs: at *best validation accuracy* and *lowest validation loss*, which are often not aligned and produce solutions with different properties. In both cases, we report the EMA with the largest decay ($\alpha = 0.998$) and recompute BN stats once after training. In App. B we report the full results including the two EMA variants with and without BN recompute.

4.2 Generalization

We start by investigating the performance of EMA models in terms of test accuracy. We find that averaging with EMA improves generalization, always outperforming the SGD baseline. This is not unexpected, as the generalization benefit of (uniform) averaging is well-known in deep learning (Izmailov et al., 2018). Nonetheless, to the best of our knowledge, we are the first to show this for EMA.

In Table 1 we report test accuracy and test loss for the momentum SGD baseline and its EMA, early stopped either at the epoch of best accuracy or lowest loss. We emphasize two takeaways. Firstly, EMA performs consistently better than the baseline. We include SWA results when available, which bring a generalization improvement approximately similar to EMA, but leave for future work a comprehensive comparison between the two methods. Secondly, the EMA model with the lowest loss does not correspond to the model with highest accuracy, as the early stopping point to minimize the loss is always earlier (see App. B). As we discuss in the next sections, the EMA with the lowest loss outperforms the best accuracy EMA in other metrics (*e.g.*,

Dataset	Architecture	Baseline		EMA (acc.)		EMA (loss)		SWA
		Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.
CIFAR-100	ResNet-18	77.63	1.02	**78.55**	0.84	78.05	0.82	-
	VGG-16	72.82	1.77	73.62	1.13	72.08	1.06	**74.27**
	WRN-28-10	81.07	0.78	**82.72**	0.67	81.90	0.64	82.15
CIFAR-10	ResNet-18	95.19	0.22	**95.59**	0.15	95.27	0.15	-
TinyImageNet	ResNet-18	66.03	1.60	**67.97**	1.35	67.05	1.36	-

Table 1: Test accuracy and loss on a baseline model and its EMA. The EMA model consistently outperforms the baseline in **accuracy** and loss. We explore EMA with two early stopping criteria: best accuracy and lowest loss. EMA models' BN statistics are recomputed once. SWA results from Izmailov et al. (2018).

calibration, repeatability, transferability), suggesting a trade-off between maximizing these metrics or model accuracy.

4.3 Label Noise

In this section, we study the case of training with label noise, *i.e.*, with a fraction of the training set wrongly annotated. We perform experiments on the benchmarks of CIFAR-10N and CIFAR-100N (Wei et al., 2022), two datasets with human annotator label noise of approximately 40%.

Interestingly, we find that the effect of implicit regularization is magnified in the presence of label noise. In Fig. 2 we observe that the EMA model performs best when averaging iterates at a relatively high learning rate. The EMA model peaks at 64.8% accuracy at epoch 100, with a learning rate around 0.4, and then decays until it plateaus at 55.0% at the end of training. Eventually, when the learning rate is decayed low enough, the model fits (*i.e.*, memorizes) all the noisy labels and reaches 100% train accuracy (App. E.1), but generalizes worse. Memorization in the EMA occurs as the learning rate is decayed, and not due to continued training (App. E.2). An explanation for the outstanding performance of EMA is that the regularizing effect of stochastic noise effectively prevents fitting the noisy labels, while it allows learning of general patterns in the data.

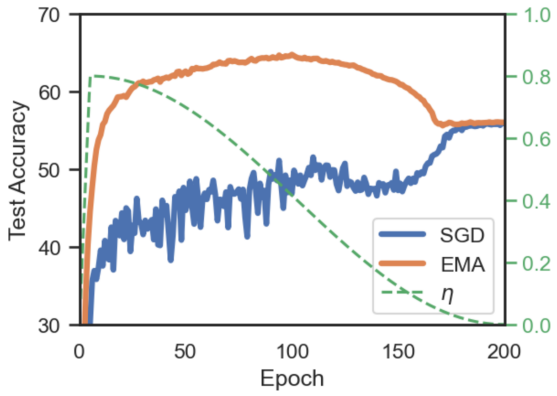


Figure 2: CIFAR-100N on ResNet-34. EMA vs SGD baseline, and learning rate (η). EMA dominates SGD and outperforms the best model. Training on data with 40% of label noise, evaluating on clean test set, mean of 3 runs.

Method	CIFAR-10N	CIFAR-100N
DivideMix	92.56	71.13
PES(semi)	92.68	70.36
ELR+	91.09	66.70
EMA (acc.)	**86.71**	**65.15**
CAL	85.36	61.73
CORES	83.60	61.15
Co-Teaching	83.83	60.30
JoCor	83.37	59.97
...	...	...
CE (standard)	77.69	55.50

Table 2: Selection of best-performing methods on CIFAR-10N (Worse) and CIFAR-100N, with 40% label noise in train data, using a ResNet-34. Ours is **highlighted**, all other results are from Wei et al. (2022). Leaderboard available at `http://www.yliuu.com/web-cifarN/Leaderboard.html`

We compare our results to the leaderboard for robust training to label noise (see Tab. 2). The performance of the EMA under label noise not only is a good example of the effect of implicit regularization, but also it is an extremely simple method which is actually competitive with state-of-the-art. The other methods in Tab. 2 are complex and specialized frameworks, often computationally demanding. Most of them, including the top-performing DivideMix (Li et al., 2020), train two networks simultaneously while refining labels based on their predictions, and use advanced augmentation strategies such as MixUp. In contrast, we do not adopt any specialized technique or data augmentation, we only keep an EMA model on top of vanilla momentum SGD training. Despite its simplicity, EMA outperforms multiple specialized methods and gets reasonably close to the state-of-the-art. We believe this can be particularly relevant when the presence or the level of label noise is unknown. While specialized (costly) methods need to be justified by heavy label noise, (lightweight) EMA can simply be adopted by default.

4.4 Prediction disagreement

Training of modern neural networks includes multiple sources of randomness, such as batch ordering, initialization and data augmentations. As a result, two independent runs (with exact same algorithm, architecture, training data and hyperparamter configuration) can converge to very different solutions. Even if their accuracy is usually similar, the resulting models will differ in many predictions of individual samples (Jiang et al., 2021b; Bhojanapalli et al., 2021). This prediction disagreement, also known as *churn*, poses a challenge for reproducibility and repeatability in deep learning. Moreover, in real-world systems where the production model is often replaced, it is desirable that the new model, expected to be ever so slightly more accurate, behaves similarly to the previous ones – that is, has a low predictive churn (Jiang et al., 2021a).

We denote the churn between two functions f_{θ_1} and f_{θ_2} as the fraction of test samples with different prediction, the lower the better. That is, $\frac{1}{N}\sum_{n=1}^{N} 1[f_{\theta_1}(x_n) \neq f_{\theta_2}(x_n)]$ for the N samples in the test set, where $f(x_n)$ is the top-class predicted. We also propose to use the Jensen-Shannon (JS) divergence as a metric for prediction disagreement, which considers the difference in the entire class probability vector. The JS divergence is a symmetrized version of the Kullback–Leibler (KL) divergence defined as $\mathrm{JS}(\mathbf{p}\|\mathbf{q}) = {}^1\!/_2\,\mathrm{KL}(\mathbf{p}\|\mathbf{m}) + {}^1\!/_2\,\mathrm{KL}(\mathbf{q}\|\mathbf{m})$, where $\mathbf{m} = {}^1\!/_2(\mathbf{p}+\mathbf{q})$.

A few attempts have been made to reduce churn with algorithmic variations (Bhojanapalli et al., 2021; Jiang et al., 2021a; Madhyastha & Jain, 2019). In our experiments, we train 3 models on independent runs with different seeds and measure the pair-wise churn and JS div. between their predictions. In Table 3 we compare the prediction disagreement of the momentum SGD baseline with the EMA model. We use the EMA with early stopping at the epoch of lowest validation loss (which outperforms EMA of best accuracy, see App. B) and recompute BN stats. Using EMA brings a great improvement in repeatability, even outperforming the state of the art (Bhojanapalli et al., 2021), a specialized method that uses co-distillation and has a $\times 2$ training

		Test Acc.		Churn		JS divergence	
		Baseline	Method	Baseline	Method	Baseline	Method
Method: EMA (lowest loss)							
CIFAR-100	ResNet-18	77.63	**78.05**	18.84	**12.71**	0.32	**0.09**
	WRN-28-10	81.07	**81.90**	15.69	**8.91**	0.10	**0.055**
	VGG-16	**72.82**	72.08	23.7	**20.07**	0.676	**0.134**
CIFAR-10	ResNet-18	95.25	**95.46**	3.78	**2.71**	0.017	**0.01**
Tiny-ImageNet	ResNet-18	66.03	**67.05**	29.36	**15.32**	0.85	**0.14**
Method: Co-distillation KL (Bhojanapalli et al., 2021)							
CIFAR-100	ResNet-56	73.26	**76.53**	26.77	**17.09**	-	-
CIFAR-10	ResNet-56	93.97	**94.63**	5.72	**4.21**	-	-

Table 3: Prediction disagreement (churn, JS div.) results. Using an EMA model substantially improves the consistency of predictions between independent runs, achieving a lower prediction disagreement than methods designed specifically for this goal.

Linear eval on:	CIFAR-10		CIFAR-100	
Pretraining on:	Tiny-ImageNet	CIFAR-100	Tiny-ImageNet	CIFAR-10
Baseline	74.61 ± 0.26	82.10 ± 0.21	52.77 ± 0.15	33.72 ± 1.67
EMA (best accuracy)	78.70 ± 0.45	84.03 ± 1.07	57.30 ± 0.39	$\mathbf{37.09} \pm 1.06$
EMA (lowest loss)	$\mathbf{79.13} \pm 0.76$	$\mathbf{85.02} \pm 0.03$	$\mathbf{57.78} \pm 0.09$	36.09 ± 0.86
Supervised	95.25 ± 0.11		77.63 ± 0.14	

Table 4: Linear evaluation on CIFAR-10/100 with a frozen ResNet-18 backbone pretrained on another dataset. Mean and std deviation for 3 seeds. The significant improvements in accuracy using EMA pretrained models indicates that the representations learned are more general and transferable.

cost. The EMA model consistently reduces the classification churn across different datasets and architectures. Note that we gain a large factor in repeatability, as most samples are already correctly predicted by both models, and so don't move. We also find a consistent improvement in the continuous metric of JS divergence.

4.5 Transfer Learning

In order to assess the quality and generalizability of the learned representations, we test their ability to transfer to other datasets. We investigate whether the implicit regularization with stochastic noise promotes the learning of more general representations that generalize across datasets, instead of relying on patterns specific to the training distribution.

We evaluate transfer learning via linear evaluation, similarly to Chen et al. (2020). We use a frozen pretrained model as a feature extractor, all layers but for the last one, and add a linear classification head for another dataset on top. Then, we train the classification head for 50 epochs with SGD with Nesterov momentum of 0.9, without weight decay and with a tuned learning rate of 0.01 without warmup. We do not use EMA on the classification head.

We find that the EMA models learn more general representations which better transfer to other datasets, compared to the SGD baseline. Table 4 shows the results for linear evaluation on the frozen feature extractors, demonstrating that EMA models' representations are more linearly separable when transferred to other tasks. Interestingly, EMA with early stopping at the epoch of lowest validation loss often outperforms the epoch of best accuracy, likely because stopping at an earlier epoch increases the strength of implicit regularization.

As expected, since we only train a linear layer, the accuracy of linear evaluation is far from the supervised performance when training an entire model from scratch. Nonetheless, we believe that our results are insightful for understanding the differences between averaged solutions and SGD solutions decaying learning rate to zero, showcasing how implicit regularization promotes the learning of more general representations. In future work, it would be of interest to extend this study to EMA of self-supervised methods and investigate if it can readily improve their performance.

4.6 Calibration

Calibration of a model is the property that the predicted probabilities reflect the true likelihood of the ground-truth. While calibrated models are important for high-stakes decision making, for example in medical domains, modern deep neural networks are generally not well-calibrated. Multiple methods have been proposed to improve calibration. Guo et al. (2017) use a post-hoc temperature scaling tuned on a hold-out set. Deep ensembles are also known to improve uncertainty estimation and calibration (Lakshminarayanan et al., 2017; Jiang et al., 2021b), but are an expensive solution. Another class of methods directly optimize for low calibration error during training with auxiliary objectives (Karandikar et al., 2021).

In Table 5 we report the test accuracy and calibration error for a baseline model, trained on SGD with Nesterov momentum, and its EMA. We use the Expected Calibration Error (ECE) metric, widely used in the literature. We fix the number of bins to $M = 100$ and compute ECE with equal-mass binning (Nixon

		Baseline	EMA
ResNet-18 CIFAR-100	Accuracy	75.83	76.31
	ECE	11.75	9.46
	ECE w/ TS	4.67	**3.13**
VGG-16 CIFAR-100	Accuracy	70.57	70.46
	ECE	20.57	8.12
	ECE w/ TS	12.17	**3.64**
WideResNet-28-10 CIFAR-100	Accuracy	79.29	80.12
	ECE	6.38	6.20
	ECE w/ TS	5.60	**3.37**
ResNet-18 CIFAR-10	Accuracy	94.54	95.01
	ECE	3.58	1.99
	ECE w/ TS	1.99	**1.04**
ResNet-18 Tiny-ImageNet	Accuracy	63.74	65.81
	ECE	12.62	8.88
	ECE w/ TS	**3.35**	3.57

Table 5: Expected Calibration Error (ECE) and ECE after Temperature Scaling (TS) results, lower is better. EMA consistently outperforms the SGD baseline.

et al., 2019). We also report ECE after temperature scaling (TS) as proposed by (Guo et al., 2017). We train on an 80% split of the full training dataset, tune the temperature in the remaining 20% hold-out set, and evaluate on test. For the EMA we use early stopping at the epoch of lowest loss and recompute BN stats after training.

We find that using an EMA considerably reduces the calibration error across all models and datasets tried, compared to the SGD baseline. The improvement that EMA brings seems to be orthogonal to the popular post-hoc operation of temperature scaling, which corrects for an average over/under-confidence. Combining temperature scaling and EMA reduces ECE by a factor of $\times 4$ in some cases. We hypothesize that a temporal ensemble of model weights represents a high diversity of solutions, which leads to an improved uncertainty estimation.

5 Conclusion

In this work, we have performed a thorough study weight averaging in deep learning through EMA models, that was lacking in the literature despite its extensive use. We proposed a training framework to limit the overhead induced by keeping EMA models, and provided a insights to improve the understanding of weight averaging dynamics. We showed that EMA allows to maintain high learning rates, which leads to implicit regularization that favors learning more general representations. Then, we highlighted the surprising early performance of EMA models, and we showed that too large averaging windows cannot be used for EMA teachers, since they require recomputing Batch Norm statistics. On the benefits of using EMA on top of SGD iterates, we showcase the strong performance of EMA models with a comprehensive empirical evaluation. We illustrate that not only EMA models generalize better, which was known in the weight averaging literature with methods such as SWA, but they also are more robust to label noise (beating many specialized methods), improve consistency of predictions across runs, produce more general and transferable representations, and are better calibrated.

References

Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. There are many consistent explanations of unlabeled data: Why you should average. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Francis Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Srinadh Bhojanapalli, Kimberly Wilber, Andreas Veit, Ankit Singh Rawat, Seungyeon Kim, Aditya Menon, and Sanjiv Kumar. On the reproducibility of neural network predictions. *arXiv preprint arXiv:2102.03349*, 2021.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pp. 177–186. Springer, 2010.

Zhaowei Cai, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Zhuowen Tu, and Stefano Soatto. Exponential moving average normalization for self-supervised and semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 194–203, 2021.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1):3520–3570, 2017.

Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (s) gd over diagonal linear networks: Implicit regularisation, large stepsizes and edge of stability. *arXiv preprint arXiv:2302.08982*, 2023.

Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019.

Sébastien Gadat and Fabien Panloup. Optimal non-asymptotic analysis of the ruppert–polyak averaging stochastic algorithm. *Stochastic Processes and their Applications*, 156:312–348, 2023.

Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Vipul Gupta, Santiago Akle Serrano, and Dennis DeCoste. Stochastic weight averaging in parallel: Large-batch training that generalizes well. In *International Conference on Learning Representations*, 2020.

Haowei He, Gao Huang, and Yang Yuan. Asymmetric valleys: Beyond sharp and flat local minima. *Advances in neural information processing systems*, 32, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9924–9935, 2022.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pp. 876–885. Association For Uncertainty in Artificial Intelligence (AUAI), 2018.

Heinrich Jiang, Harikrishna Narasimhan, Dara Bahri, Andrew Cotter, and Afshin Rostamizadeh. Churn reduction via distillation. *arXiv preprint arXiv:2106.02654*, 2021a.

Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of sgd via disagreement. *arXiv preprint arXiv:2106.13799*, 2021b.

Jean Kaddour. Stop wasting my time! saving days of imagenet and bert training with latest weight averaging. *arXiv preprint arXiv:2209.14981*, 2022.

Archit Karandikar, Nicholas Cain, Dustin Tran, Balaji Lakshminarayanan, Jonathon Shlens, Michael C Mozer, and Becca Roelofs. Soft calibration objectives for neural networks. *Advances in Neural Information Processing Systems*, 34:29768–29779, 2021.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Chandrashekar Lakshminarayanan and Csaba Szepesvari. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pp. 1347–1355. PMLR, 2018.

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, pp. 5639–5650. PMLR, 2020.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

Tao Li, Zhehao Huang, Qinghua Tao, Yingwen Wu, and Xiaolin Huang. Trainable weight averaging for fast convergence and better generalization. *arXiv preprint arXiv:2205.13104*, 2022.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33: 20331–20342, 2020.

Pranava Madhyastha and Rishabh Jain. On model stability as a function of random seed. *arXiv preprint arXiv:1909.10447*, 2019.

Nicole Mücke, Gergely Neu, and Lorenzo Rosasco. Beating sgd saturation with tail-averaging and minibatching. *Advances in Neural Information Processing Systems*, 32, 2019.

Gergely Neu and Lorenzo Rosasco. Iterate averaging as regularization for stochastic gradient descent. In *Conference On Learning Theory*, pp. 3222–3242. PMLR, 2018.

Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. Self: Learning to filter noisy labels with self-ensembling. *arXiv preprint arXiv:1910.01842*, 2019.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR workshops*, volume 2, 2019.

Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pp. 3918–3926. PMLR, 2018.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.

Boris T Polyak. New stochastic approximation type procedures. *Automat. i Telemekh*, 7(98-107):2, 1990.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948, 2021.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TBWA6PLJZQm.

Guandao Yang, Tianyi Zhang, Polina Kirichenko, Junwen Bai, Andrew Gordon Wilson, and Chris De Sa. Swalp: Stochastic weight averaging in low precision training. In *International Conference on Machine Learning*, pp. 7015–7024. PMLR, 2019.

Yasin Yaz, Chuan-Sheng Foo, Stefan Winkler, Kim-Hui Yap, Georgios Piliouras, Vijay Chandrasekhar, et al. The unusual effectiveness of averaging in gan training. In *International Conference on Learning Representations*, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.