Supplementary Material to "A spherical analysis of Adam with Batch Normalization"

A Radial invariance of filters with BN

In this section, we show the radial invariance of a set of filters equipped with BN. Please note that the following notations are specific and restricted to this section.

For the sake of simplicity, we only consider the case of a convolutional layer that preserves the spatial extension of the input. We also focus on a single filter. Since all filters act independently on input data, the following calculation holds for any filter.

Let $\mathbf{x} \in \mathbb{R}^{C \times K}$ be the parameters of a single filter, where C is the number of input channels and K is the kernel size. During training, this layer is followed by BN and applied to a batch $\mathbf{s} \in \mathbb{R}^{B \times C \times D}$ of B inputs of spatial size D. The output of the convolution operator ϕ applied to a filter $\mathbf{x} \in \mathbb{R}^{C \times K}$ and to a given batch element $\mathbf{s}_b \in \mathbb{R}^{C \times D}$, with $b \in [\![1, B]\!]$, is thus:

$$\mathbf{t}_b \overset{\text{def}}{=} \phi(\mathbf{x}, \mathbf{s}_b) \in \mathbb{R}^D. \tag{21}$$

The application $(\mathbf{x}, \mathbf{s}_b) \mapsto \phi(\mathbf{x}, \mathbf{s}_b)$ is bilinear. BN then centers and normalizes the output $\mathbf{t}$ using the mean and variance over the batch and the spatial dimension:

$$\mu = \frac{1}{BD} \sum_{b,j} t_{b,j}, \tag{22}$$

$$\sigma^2 = \frac{1}{BD} \sum_{b,j} \left(t_{b,j} - \mu\right)^2, \tag{23}$$

$$\hat{\mathbf{t}}_b \overset{\text{def}}{=} \left(\sigma^2 + \epsilon\right)^{-1/2} \left(\mathbf{t}_b - \mu \mathbf{1}_D\right), \tag{24}$$

where $\mathbf{1}_D$ denotes the all-ones vector of dimension D and ϵ is a small constant.

Now if the coefficients of the filter are rescaled by $\rho > 0$, then, by bilinearity, the new output of the layer for this filter verifies:

$$\tilde{\mathbf{t}}_b = \phi(\rho \mathbf{x}, \mathbf{s}_b) = \rho \phi(\mathbf{x}, \mathbf{s}_b). \tag{25}$$

Since the variance of inputs is generally large in practice, for small ϵ, the mean and variance are:

$$\tilde{\mu} = \rho \mu, \tag{26}$$

$$\tilde{\sigma}^2 \approx \rho^2 \sigma^2. \tag{27}$$

It can then be considered that the subsequent BN layer is invariant to this rescaling, *i.e.*, $\hat{\tilde{\mathbf{t}}}_b \approx \hat{\mathbf{t}}_b$.

B Extension to other normalization layers

The radial invariance for BN described above in Appendix A applies as well to InstanceNorm (IN) (Ulyanov et al., 2016) as the normalization is also done with respect to channels but without the batch dimension. Regarding LayerNorm (Ba et al., 2016) (LN), the normalization is performed over all channels and the entire weight layer can thus be rescaled too, without impacting the output. As for GroupNorm (Wu & He, 2018) (GN), it associates several channels for normalization; the radial invariance in this case concerns the corresponding group of filters.

Thanks to this general property of radial invariance, the results in this paper not only concern BN but also IN. In fact, they apply as well to LN and GN when considering the suitable group of parameters. The optimization in this case concerns the proper slice of the parameter tensor of the layer, i.e., the whole tensor for LN, and the selected group of filters for GN.

C Results in Sections 2 and 3

In this section, we provide proofs and/or empirical results supporting the claims in Sections 2 and 3 of the paper.

In the following, the double parentheses around an equation number, e.g., ((10)), indicate that we recall an equation that was previously stated in the main paper, rather than introduce a new one, e.g., noted (28). Also, framed formulas actually refer to results stated in the main paper, thus with double-bracket equation numbering.

C.1 Proof of theorems and validity of assumptions

C.1.1 Proof of Theorem 2 (Image step on $\mathcal{S}_{d-1}$) in Section 2.3

We recall the main theorem in Section 2.3.

Theorem 2 (Image step on $\mathcal{S}_{d-1}$) *The update of a group of radially-invariant parameters $\mathbf{x}_k$ at step k corresponds to an update of its projection $\mathbf{u}_k$ on $\mathcal{S}_{d-1}$ through an exponential map at $\mathbf{u}_k$ with velocity $\eta_k^e \mathbf{c}_k^\perp$ at order 3:*

$$\mathbf{u}_{k+1} = \mathrm{Exp}_{\mathbf{u}_k}\left(-\left[1 + O\left(\left(\eta_k^e \|\mathbf{c}_k^\perp\|\right)^2\right)\right]\eta_k^e \mathbf{c}_k^\perp\right), \tag{(10)}$$

where $\mathrm{Exp}_{\mathbf{u}_k}$ is the exponential map on $\mathcal{S}_{d-1}$, and with

$$\mathbf{c}_k \overset{\text{def}}{=} r_k \mathbf{a}_k \oslash \frac{\mathbf{b}_k}{d^{-1/2}\|\mathbf{b}_k\|}, \quad \eta_k^e \overset{\text{def}}{=} \frac{\eta_k}{r_k^2 d^{-1/2}\|\mathbf{b}_k\|}\left(1 - \frac{\eta_k\langle \mathbf{c}_k, \mathbf{u}_k\rangle}{r_k^2 d^{-1/2}\|\mathbf{b}_k\|}\right)^{-1}. \tag{(11)}$$

More precisely:

$$\mathbf{u}_{k+1} = \frac{\mathbf{u}_k - \eta_k^e \mathbf{c}_k^\perp}{\sqrt{1 + (\eta_k^e\|\mathbf{c}_k^\perp\|)^2}}. \tag{(12)}$$

Proof. To simplify the calculation in the demonstration, we introduce the following notation:

$$A_k \overset{\text{def}}{=} \frac{\eta_k}{r_k^2 d^{-1/2}\|\mathbf{b}_k\|}. \tag{28}$$

We first demonstrate the expression for the radius dynamics in Eq. (14) and the precise step for $\mathbf{u}$ in Eq. (12). Then we use geometric arguments and a Taylor expansion to derive the update on the sphere stated in Eq.(10).

Radius dynamics. We first show Eq. (14), which we recall here using the A_k notation:

$$\boxed{\frac{r_{k+1}}{r_k} = (1 - A_k\langle \mathbf{c}_k, \mathbf{u}_k\rangle)\sqrt{1 + (\eta_k^e\|\mathbf{c}_k^\perp\|)^2}.} \tag{(14)}$$

First, we rewrite the step of a generic scheme in Eqs. (3-4) along the radial and tangential directions and separate the division vector $\mathbf{b}_k$ into its deformation $\frac{\mathbf{b}_k}{d^{-1/2}\|\mathbf{b}_k\|}$ and its scalar scheduling effect $d^{-1/2}\|\mathbf{b}_k\|$, as stated in the discussion:

$$\begin{aligned}
r_{k+1}\mathbf{u}_{k+1} &= r_k\mathbf{u}_k - \frac{\eta_k}{d^{-1/2}\|\mathbf{b}_k\|}\mathbf{a}_k \oslash \frac{\mathbf{b}_k}{d^{-1/2}\|\mathbf{b}_k\|} \\
&= r_k\left[\mathbf{u}_k - \frac{\eta_k}{r_k^2 d^{-1/2}\|\mathbf{b}_k\|} r_k\mathbf{a}_k \oslash \frac{\mathbf{b}_k}{d^{-1/2}\|\mathbf{b}_k\|}\right] \\
&= r_k\left[\mathbf{u}_k - A_k r_k \mathbf{a}_k \oslash \frac{\mathbf{b}_k}{d^{-1/2}\|\mathbf{b}_k\|}\right].
\end{aligned} \tag{29}$$

We can note the appearance of a new term $r_k\mathbf{a}_k$. The vector $\mathbf{a}_k$ is a gradient momentum and therefore homogeneous to a gradient. Using Lemma 1, $r_k\mathbf{a}_k$ is homogeneous to a gradient on the hypersphere and can be interpreted as the momentum on the hypersphere.

From Eq. (29), we introduce $\mathbf{c}_k$ (the deformed momentum on hypersphere) as in Eq. (11) and decompose it into the radial and tangential components. We have:

$$\begin{aligned}\frac{r_{k+1}}{r_k}\mathbf{u}_{k+1} &= \mathbf{u}_k - A_k\mathbf{c}_k \\ &= \left(1 - A_k\langle\mathbf{c}_k, \mathbf{u}_k\rangle\right)\mathbf{u}_k - A_k\mathbf{c}_k^{\perp}. \end{aligned} \tag{30}$$

By taking the squared norm of the equation, we obtain:

$$\frac{r_{k+1}^2}{r_k^2} = \left(1 - A_k\langle\mathbf{c}_k, \mathbf{u}_k\rangle\right)^2 + \left(A_k\|\mathbf{c}_k^{\perp}\|\right)^2. \tag{31}$$

Making the assumption that $1 - A_k\langle\mathbf{c}_k, \mathbf{u}_k\rangle > 0$, which is true in practice and discussed in the next subsection, we have:

$$\frac{r_{k+1}}{r_k} = \left(1 - A_k\langle\mathbf{c}_k, \mathbf{u}_k\rangle\right)\sqrt{1 + \left(\frac{A_k}{1 - A_k\langle\mathbf{c}_k, \mathbf{u}_k\rangle}\|\mathbf{c}_k^{\perp}\|\right)^2}. \tag{32}$$

After introducing $\eta_k^e = \frac{A_k}{(1-A_k\langle\mathbf{c}_k,\mathbf{u}_k\rangle)}$ as in Eq. (11), we obtain the result of (14).

Update of normalized parameters. We then show Eq. (12):

$$\boxed{\mathbf{u}_{k+1} = \frac{\mathbf{u}_k - \eta_k^e\mathbf{c}_k^{\perp}}{\sqrt{1 + (\eta_k^e\|\mathbf{c}_k^{\perp}\|)^2}}.} \tag{(12)}$$

Combining the radius dynamics previously calculated with Eq. (30), we have:

$$\begin{aligned}\mathbf{u}_{k+1} =& \frac{(1 - A_k\langle\mathbf{c}_k, \mathbf{u}_k\rangle)\mathbf{u}_k - A_k\mathbf{c}_k^{\perp}}{(1 - A_k\langle\mathbf{c}_k, \mathbf{u}_k\rangle)\sqrt{1 + (\eta_k^e\|\mathbf{c}_k^{\perp}\|)^2}} &(33)\\ =& \frac{\mathbf{u}_k - \frac{A_k}{1-A_k\langle\mathbf{c}_k,\mathbf{u}_k\rangle}\mathbf{c}_k^{\perp}}{\sqrt{1 + (\eta_k^e\|\mathbf{c}_k^{\perp}\|)^2}}. &(34)\end{aligned}$$

Hence the result (12) using the definition of η_k^e.

This result provides a unique decomposition of the generic step as a step in $\mathrm{span}(\mathbf{u}_k, \mathbf{c}_k^{\perp})$ for the normalized filter (Eq. (12)) and as a radius update (Eq. (14)).

We split the rest of the proof of the theorem in three parts.

Distance covered on the sphere. The distance covered on the hypersphere $\mathcal{S}_{d-1}$ by an optimization step is:

$$\mathrm{dist}_{\mathcal{S}_{d-1}}(\mathbf{u}_{k+1}, \mathbf{u}_k) = \arccos(\langle\mathbf{u}_{k+1}, \mathbf{u}_k\rangle). \tag{35}$$

From Eq. (12) and with Lemma 1, we also have:

$$\langle\mathbf{u}_{k+1}, \mathbf{u}_k\rangle = \frac{1}{\sqrt{1 + (\eta_k^e\|\mathbf{c}_k^{\perp}\|)^2}}. \tag{36}$$

Therefore, $\mathrm{dist}_{\mathcal{S}_{d-1}}(\mathbf{u}_{k+1}, \mathbf{u}_k) = \varphi(\eta_k^e\|\mathbf{c}_k^{\perp}\|)$ where $\varphi : z \mapsto \arccos\left(\frac{1}{\sqrt{1+z^2}}\right)$, which is equal to $\arctan$ on $\mathbb{R}_+$. Then a Taylor expansion at order 3 of $\arctan$ yields for $\eta_k^e\|\mathbf{c}_k^{\perp}\|$:

$$\mathrm{dist}_{\mathcal{S}_{d-1}}(\mathbf{u}_{k+1}, \mathbf{u}_k) = \eta_k^e\|\mathbf{c}_k^{\perp}\| + O\left(\left(\eta_k^e\|\mathbf{c}_k^{\perp}\|\right)^3\right). \tag{37}$$

The Taylor expansion validity is discussed in the next subsection.

Exponential map on the sphere. Given a Riemannian manifold $\mathcal{M}$, for a point $\mathbf{u} \in \mathcal{M}$ there exists an open set $\mathcal{O}$ of the tangent space $\mathcal{T}_{\mathbf{u}}\mathcal{M}$ containing $\mathbf{0}$, such that for any tangent vector $\mathbf{w} \in \mathcal{O}$ there is a unique geodesic (a path minimizing the local distance on $\mathcal{M}$ when conserving the tangent velocity) $\gamma : [-1, 1] \to \mathcal{M}$ that is differentiable and such that $\gamma(0) = \mathbf{u}$ and $\gamma'(0) = \mathbf{w}$. Then, the exponential map of $\mathbf{w}$ from $\mathbf{u}$ is defined as $\mathrm{Exp}_{\mathbf{u}}(\mathbf{w}) = \gamma(1)$.

In the case of the manifold $\mathcal{S}_{d-1}$, the geodesics are complete (they are well defined for any point $\mathbf{u} \in \mathcal{S}_{d-1}$ and any velocity $\mathbf{w} \in \mathcal{T}_{\mathbf{u}}\mathcal{S}_{d-1}$) and are the great circles: for any $\mathbf{u} \in \mathcal{S}_{d-1}$ and any $\mathbf{w} \in \mathcal{T}_{\mathbf{u}}\mathcal{S}_{d-1}$, the map $\psi : t \in \mathbb{R} \mapsto \mathrm{Exp}_{\mathbf{u}}(t\mathbf{w})$) verifies $\psi(\mathbb{R}) = \mathcal{S}_{d-1} \cap \mathrm{span}(\{\mathbf{u}, \mathbf{w}\})$ which is a great circle passing through $\mathbf{u}$ with tangent $\mathbf{w}$. Furthermore, since the circumference of the great circle is 2π, we have that for any $\mathbf{p} \in \mathcal{S}_{d-1} \backslash \{-\mathbf{u}\}$ there is a unique $\mathbf{w}$ verifying $\|\mathbf{w}\| < \pi$ such that $\mathbf{p} = \mathrm{Exp}_{\mathbf{u}}(\mathbf{w})$ and we have:

$$\mathrm{dist}_{\mathcal{S}_{d-1}}(\mathbf{u}, \mathbf{p}) = \|\mathbf{w}\| \text{ and } \langle \mathbf{p}, \mathbf{w} \rangle \geq 0. \tag{38}$$

Optimization step as an exponential map. We will use the previously stated differential geometry properties to prove:

$$\boxed{\mathbf{u}_{k+1} = \mathrm{Exp}_{\mathbf{u}_k}\left(-\left[1 + O\left(\left(\eta_k^e \|\mathbf{c}_k^\perp\|\right)^2\right)\right] \eta_k^e \mathbf{c}_k^\perp\right).} \tag{(10)}$$

For an optimization step we have:

- by construction, $\mathbf{c}_k^\perp \in \mathcal{T}_{\mathbf{u}_k}\mathcal{S}_{d-1}$;
- from Eq. (12), $\mathbf{u}_{k+1} \in \mathcal{S}_{d-1} \cap \mathrm{span}(\{\mathbf{u}_k, \mathbf{c}_k^\perp\})$;
- from Eq. (12), $\langle \mathbf{u}_{k+1}, \mathbf{c}_k^\perp \rangle \leq 0$.

Then, there exists α that verifies $\|\alpha \mathbf{c}_k^\perp\| < \pi$ such that:

$$\mathbf{u}_{k+1} = \mathrm{Exp}_{\mathbf{u}_k}\left(\alpha \mathbf{c}_k^\perp\right). \tag{39}$$

From Eq. (38), because $\langle \mathbf{u}_{k+1}, \mathbf{c}_k^\perp \rangle \leq 0$, we have $\alpha < 0$. We also have that $\|\alpha \mathbf{c}_k^\perp\| = \mathrm{dist}_{\mathcal{S}_{d-1}}(\mathbf{u}_{k+1}, \mathbf{u}_k)$. Then, using the distance previously calculated in Eq. (37), we have:

$$|\alpha| \|\mathbf{c}_k^\perp\| = \eta_k^e \|\mathbf{c}_k^\perp\| + O\left(\left(\eta_k^e \|\mathbf{c}_k^\perp\|\right)^3\right), \tag{40}$$

$$|\alpha| = \eta_k^e \left[1 + O\left(\left(\eta_k^e \|\mathbf{c}_k^\perp\|\right)^2\right)\right]. \tag{41}$$

Combining the sign and absolute value of α, we get the final exponential map expression:

$$\mathbf{u}_{k+1} = \mathrm{Exp}_{\mathbf{u}_k}\left(-\left[1 + O\left(\left(\eta_k^e \|\mathbf{c}_k^\perp\|\right)^2\right)\right] \eta_k^e \mathbf{c}_k^\perp\right), \tag{(10)}$$

$$\approx \mathrm{Exp}_{\mathbf{u}_k}\left(-\eta_k^e \mathbf{c}_k^\perp\right). \tag{42}$$

Note that we implicitly assume here that $|\alpha| \|\mathbf{c}_k^\perp\| \approx \eta_k^e \|\mathbf{c}_k^\perp\| < \pi$, which is discussed in the next subsection.

□

C.1.2 Validity of the assumptions in Theorem 2

Sign of $1 - A_k \langle \mathbf{c}_k, \mathbf{u}_k \rangle$. We tracked the maximum of the quantity $A_k \langle \mathbf{c}_k, \mathbf{u}_k \rangle$ for all the filters of a ResNet20 CIFAR trained on CIFAR10 and optimized with SGD-M or Adam (see Appendix **??** for implementation details). As can be seen on Fig. 4, this quantity is always small compared to 1, making $1 - A_k \langle \mathbf{c}_k, \mathbf{u}_k \rangle$ always positive in practice. The order of magnitude of this quantity is roughly the same for different architectures and datasets.

Taylor expansion. We tracked the maximum of the quantity $\eta_k^e \|\mathbf{c}_k^\perp\|$ for all the filters of a ResNet20 CIFAR trained on CIFAR10 and optimized with SGD-M or Adam. The observed values justify the Taylor expansion and validate the assumption $|\alpha| \|\mathbf{c}_k^\perp\| \approx \eta_k^e \|\mathbf{c}_k^\perp\| < \pi$. (cf. Fig 5). The order of magnitude of this quantity is roughly the same for other different architectures and datasets.

C.1.3 ν_k, order 2 moment on the hypersphere for Adam

Scheduling effect of Adam division vector. With Eq. (64) and using Lemma 1, we can give the expression of the second-order moment on the sphere, defined as $\nu_k = r_k d^{-1/2} \|\mathbf{b}_k\|$:

$$\boxed{\nu_k = d^{-1/2} \frac{1 - \beta_1^{k+1}}{1 - \beta_1} \Big(\frac{1 - \beta_2}{1 - \beta_2^{k+1}}\Big)^{1/2} \Big(\sum_{i=0}^{k} \beta_2^{k-i} \frac{r_k^2}{r_i^2} \|\nabla \mathcal{L}(\mathbf{u}_i) + \lambda r_i^2 \mathbf{u}_i\|^2\Big)^{1/2}.} \tag{43}$$

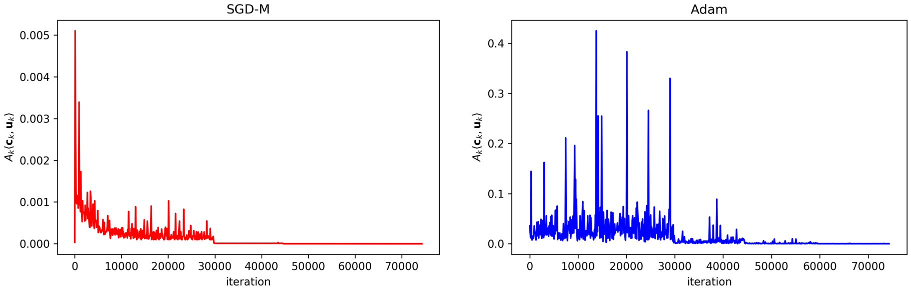

Figure 4: **Tracking of $A_k\langle \mathbf{c}_k, \mathbf{u}_k\rangle$ for SGD-M and Adam.** The above graphs show the maximum of the absolute value of $A_k\langle \mathbf{c}_k, \mathbf{u}_k\rangle$ for all filters in all layers of a ResNet20 CIFAR trained on CIFAR10 and optimized with SGD-M (left) or Adam (right). The quantity is always small compared to 1. Therefore we may assume that $1 - A_k\langle \mathbf{c}_k, \mathbf{u}_k\rangle \geq 0$.

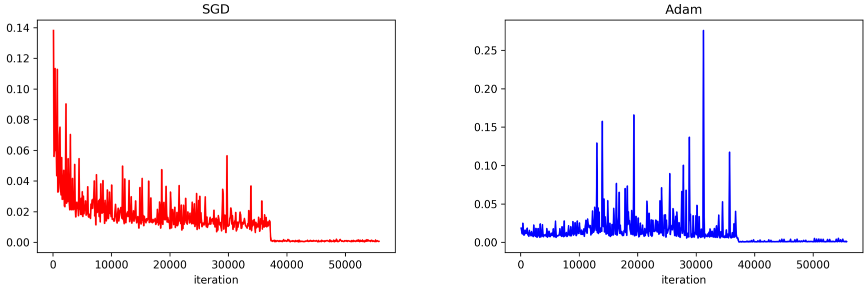

Figure 5: **Tracking of $\eta_k^e \|\mathbf{c}_k^\perp\|$ for SGD-M and Adam.** The above graphs show the maximum of the absolute value of $\eta_k^e \|\mathbf{c}_k^\perp\|$ for all filters in all layers of a ResNet20 CIFAR trained on CIFAR10 and optimized with SGD-M (left) or Adam (right).

C.1.4 Proof of Theorem 4 (SGD equivalent scheme on the unit hypersphere) in Section 3.2

We prove the following theorem:

Theorem 4 (SGD equivalent scheme on the unit hypersphere.) *For any $\lambda > 0, \eta > 0, r_0 > 0$, we have the following equivalence at order 2 in the radius dynamics:*

$$\begin{cases} \text{(SGD)} \\ \mathbf{x}_0 = r_0 \mathbf{u}_0 \\ \lambda_k = \lambda \\ \eta_k = \eta \end{cases} \quad \textit{is scheme-equivalent at order 2 to} \quad \begin{cases} \text{(AdamG*)} \\ \mathbf{x}_0 = \mathbf{u}_0 \\ \beta = (1-\eta\lambda)^4 \\ \eta_k = (2\beta)^{-1/2} \\ v_0 = r_0^4 (2\eta^2 \beta^{1/2})^{-1} \end{cases}$$

Proof. As summarized in Table 1, the expressions of the effective learning rates and directions for SGD are $\mathbf{c}_k^\perp = r_k \nabla \mathcal{L}(\mathbf{x}_k) = \nabla \mathcal{L}(\mathbf{u}_k)$ and $\eta_k^e = \frac{\eta_k}{r_k^2(1-\eta_k\lambda_k)}$.

Equivalence with SGD and L_2 regularization. We look for conditions leading to an equivalence between SGD with L_2 regularization and SGD without L_2 regularization. Using Lemma 3, the equality of effective directions is trivial and the equality of effective learning rates for any step k yields the following equivalence:

$$\begin{cases} \text{(SGD)} \\ \tilde{\mathbf{x}}_0 = r_0 \mathbf{u}_0 \\ \tilde{\lambda}_k = \lambda \\ \tilde{\eta}_k = \eta \end{cases} \quad \text{is scheme-equivalent to} \quad \begin{cases} \text{(SGD)} \\ \mathbf{x}_0 = r_0 \mathbf{u}_0 \\ \lambda_k = 0 \\ \eta_k = \eta(1-\eta\lambda)^{-2k-1} \end{cases} \tag{44}$$

L_2 regularization is equivalent to an exponential scheduling of the learning rate, as found in Li & Arora (2020). Here, we provide a proof in a constructive manner. We are going to use Lemma 3 and find a sufficient condition to have:

$$\begin{cases} \text{(i) } \mathbf{u}_0 = \tilde{\mathbf{u}}_0 \\ \text{(ii) } \forall k \geq 0, \eta_k^e = \tilde{\eta}_k^e, \mathbf{c}_k^{\perp} = \tilde{\mathbf{c}}_k^{\perp}. \end{cases}$$

Equation (i) is trivially satisfied by simply taking the same starting point: $\tilde{\mathbf{x}}_0 = \mathbf{x}_0$.

Regarding (ii), because effective directions are the same and only depend on $\mathbf{u}_k$, we only need a sufficient condition on η_k^e. For effective learning rates, using Eq. ((14)) and expressions in Table 1, we have:

$$\eta_k^e = \tilde{\eta_k^e} \Leftrightarrow \frac{\eta_k}{r_k^2} = \frac{\tilde{\eta}_k}{\tilde{r}_k^2(1-\tilde{\eta}_k\lambda)}. \tag{45}$$

Since $\tilde{\eta_k} = \eta$, we obtain:

$$(45) \Leftrightarrow \eta_k = \left(\frac{r_k}{\tilde{r}_k}\right)^2 \frac{\eta}{(1-\eta\lambda)}.$$

Therefore:

$$\frac{\eta_{k+1}}{\eta_k} = \left(\frac{r_{k+1}\tilde{r}_k}{\tilde{r}_{k+1}r_k}\right)^2 = \left(\frac{r_{k+1}/r_k}{\tilde{r}_{k+1}/\tilde{r}_k}\right)^2.$$

By using the radius dynamics in Eq. (14) for the two schemes, SGD and SGD with L_2 regularization, and by the equality of effective learning rates and directions, we have:

$$\begin{aligned} \frac{\eta_{k+1}}{\eta_k} &= \left(\frac{\sqrt{1+(\eta_k^e\|\mathbf{c}_k^{\perp}\|)^2}}{(1-\eta\lambda)\sqrt{1+(\tilde{\eta_k^e}\|\tilde{\mathbf{c}_k^{\perp}}\|)^2}}\right)^2 \\ &= (1-\eta\lambda)^{-2}. \end{aligned}$$

By taking Eq. (45) for $k=0$, because $r_0 = \tilde{r}_0$ we have: $\eta_0 = \eta(1-\eta\lambda)^{-1}$. Combining the previous relation and the initialization case, we derive by induction that $\eta_k = \eta(1-\eta\lambda)^{-2k-1}$ is a sufficient condition. We can conclude, using Lemma 3, the equivalence stated in Eq. (44).

Resolution of the radius dynamics. Without L_2 regularization, the absence of radial component in $\mathbf{c}_k$ makes the radius dynamics simple:

$$r_{k+1}^2 = r_k^2 + \frac{(\eta_k\|\nabla\mathcal{L}(\mathbf{u}_k)\|)^2}{r_k^2}. \tag{46}$$

With a Taylor expansion at order 2, we can show that for $k \geq 1$ the solution $r_k^2 = \sqrt{2\sum_{i=0}^{k-1}(\eta_i\|\nabla\mathcal{L}(\mathbf{u}_i)|)^2 + r_0^4}$ satisfies the previous equation. Indeed using the expression at step $k+1$ gives:

$$\begin{aligned} r_{k+1}^2 &= \sqrt{2\sum_{i=0}^{k-1}(\eta_i\|\nabla\mathcal{L}(\mathbf{u}_i)|)^2 + r_0^4 + 2(\eta_k\|\nabla\mathcal{L}(\mathbf{u}_k)\|)^2} \\ &= r_k^2\sqrt{1 + 2\frac{(\eta_k\|\nabla\mathcal{L}(\mathbf{u}_k)\|)^2}{r_k^4}} \\ &= r_k^2\left(1 + (1/2)2\frac{(\eta_k\|\nabla\mathcal{L}(\mathbf{u}_k)\|)^2}{r_k^4} + o\left(\frac{(\eta_k\|\nabla\mathcal{L}(\mathbf{u}_k)\|)^2}{r_k^4}\right)\right) \\ &= r_k^2 + \frac{(\eta_k\|\nabla\mathcal{L}(\mathbf{u}_k)\|)^2}{r_k^2} + o\left(\frac{(\eta_k\|\nabla\mathcal{L}(\mathbf{u}_k)\|)^2}{r_k^2}\right). \end{aligned}$$

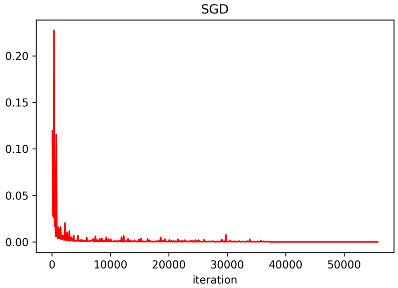

Figure 6: **Validity of Taylor expansion**. We tracked the maximum value of $(\eta_k \|\nabla\mathcal{L}(\mathbf{u}_k)\|)^2/r_k^2$ for all filters in all layers of a ResNet20 CIFAR trained on CIFAR10 with SGD. The order of magnitude of the gradient is roughly the same for other architectures or datasets. It empirically validates the approximation by the Taylor expansion.

Using $\eta_k = \eta(1-\eta\lambda)^{-2k-1}$, introducing $\beta = (1-\eta\lambda)^4$, omitting the $o\left(\frac{(\eta_k\|\nabla\mathcal{L}(\mathbf{u}_k)\|)^2}{r_k^2}\right)$ and injecting the previous solution in the effective learning rate, we obtain the closed form:

$$\begin{aligned} \eta_k^e &= \frac{\eta(1-\eta\lambda)^{-2k-1}}{\sqrt{2\sum_{i=0}^{k-1}\eta^2(1-\eta\lambda)^{-4i-2}\|\nabla\mathcal{L}(\mathbf{u}_i)\|^2 + r_0^4}} \\ &= \frac{(2\beta)^{-\frac{1}{2}}}{\sqrt{\sum_{i=0}^{k-1}\beta^{(k-1)-i}\|\nabla\mathcal{L}(\mathbf{u}_i)\|^2 + \beta^k\frac{r_0^4}{2\eta^2\beta^{\frac{1}{2}}}}}. \end{aligned} \tag{47}$$

AdamG*. The AdamG* scheme is constrained on the hypersphere thanks to the normalization; the radius is therefore constant and equal to 1. The absence of radial component in the update gives: $\mathbf{c}_k^\perp = \nabla\mathcal{L}(\mathbf{u}_k)$ and $\eta_k^e = \frac{\eta_k}{\sqrt{v_k}}$. Thus, the resolution of the induction on v_k leads to the the closed form:

$$\eta_k^e = \frac{\eta_k}{\sqrt{\sum_{i=0}^{k-1}\beta^{(k-1)-i}\|\nabla\mathcal{L}(\mathbf{u}_i)\|^2 + \beta^k v_0}}. \tag{48}$$

Hence the final theorem, when identifying the closed-form expressions of effective learning rates and using Lemma 3. □

C.1.5 VALIDITY OF THE ASSUMPTIONS IN THEOREM 4

Validity of the Taylor expansion. We tracked, for a CNN trained with SGD, the quantity $(\eta_k\|\nabla\mathcal{L}(\mathbf{u}_k)\|)^2/r_k^2$, which is the variable of the Taylor expansion. As can be seen in Figure 6, the typical order of magnitude is 10^{-2}, justifying the Taylor expansion.

A quick formal analysis also suggests the validity of this hypothesis. Thanks to the expression of $\eta_k = (1-\eta\lambda)^{-2i-k}\eta$ shown in the previous section, if we replace $\|\nabla\mathcal{L}(\mathbf{u}_k)\|$ by a constant for asymptotic analysis, the comparison becomes:

$$(1-\eta\lambda)^{-4k-2} \ll (1-\eta\lambda)^{-2}\frac{1-(1-\eta\lambda)^{-4k}}{1-(1-\eta\lambda)^{-4}} \tag{49}$$

$$1 \ll \frac{1-(1-\eta\lambda)^{4k}}{(1-\eta\lambda)^{-4}-1}. \tag{50}$$

It is asymptotically true.

D Results in Section 4 (Geometric phenomena in Adam optimization)

D.1 Results in Section 4.2 (Identification of geometrical phenomena in Adam)

Decomposition of the effective direction. We decompose the effective direction as a gradient term and an L_2 regularization term:

$$\boxed{\mathbf{c}_k^{\text{grad}} = \nabla\mathcal{L}(\mathbf{u}_k) + \sum_{i=0}^{k-1} \beta^{k-i} \frac{r_k}{r_i} \nabla\mathcal{L}(\mathbf{u}_i),} \tag{(17)}$$

$$\boxed{\mathbf{c}_k^{L_2} = \mathbf{u}_k + \sum_{i=0}^{k-1} \beta^{k-i} \frac{r_i}{r_k} \mathbf{u}_i.} \tag{(17)}$$

Note that these expressions highlight the main terms at step k and the dependency on r_i.

Developing the recurrence in Eq (4), we obtain:

$$\mathbf{a}_k = \sum_{i=0}^{k} \beta^{k-i} \left(\nabla\mathcal{L}(\mathbf{x}_i) + \lambda \mathbf{x}_i\right). \tag{51}$$

Using Lemma 1 and decomposing on $\nabla\mathcal{L}(\mathbf{u}_i)$ and $\mathbf{u}_i$, we have:

$$\begin{aligned} \mathbf{a}_k &= \sum_{i=0}^{k} \beta^{k-i} \left(\frac{1}{r_i}\nabla\mathcal{L}(\mathbf{u}_i) + \lambda r_i \mathbf{u}_i\right) && (52) \\ &= \frac{1}{r_k}\left(\sum_{i=0}^{k} \beta^{k-i} \left(\frac{r_k}{r_i}\nabla\mathcal{L}(\mathbf{u}_i) + \lambda r_k r_i \mathbf{u}_i\right)\right). && (53) \end{aligned}$$

Thus:

$$r_k \mathbf{a}_k = \sum_{i=0}^{k} \beta^{k-i} \frac{r_k}{r_i} \nabla\mathcal{L}(\mathbf{u}_i) + \lambda r_k^2 \sum_{i=0}^{k} \beta^{k-i} \frac{r_i}{r_k} \mathbf{u}_i, \tag{54}$$

which leads to the expression of $\mathbf{c}_k^{\text{grad}}$ and $\mathbf{c}_k^{L_2}$ when we define $\mathbf{c}_k \overset{\text{def}}{=} r_k \mathbf{a}_k \oslash \frac{\mathbf{b}_k}{d^{-1/2}\|\mathbf{b}_k\|}$ (Eq. (11)).

D.2 Results in Section 4.2 (Empirical study)

Clarification on Adam without deformation of gradients (a). Following Theorem 2, the division vector $\mathbf{b}_k$ has two contributions in the decomposition:

- a deformation in $\mathbf{c}_k$ applied to $\mathbf{a}_k$: $\mathbf{c}_k = r_k \mathbf{a}_k \oslash \frac{\mathbf{b}_k}{d^{-1/2}\|\mathbf{b}_k\|}$;
- a scheduling effect in the effective learning rate $d^{-1/2}\|\mathbf{b}_k\|$ (Eq. (11)).

The goal is to find a new division vector $\mathsf{S}(\mathbf{b}_k)$ that does not create a deformation while preserving the scheduling effect of $\mathbf{b}_k$ in the effective learning rate. This means:

$$\frac{\mathsf{S}(\mathbf{b}_k)}{d^{-1/2}\|\mathsf{S}(\mathbf{b}_k)\|} = [1 \cdots 1]^\top, \tag{55}$$

$$d^{-1/2}\|\mathsf{S}(\mathbf{b}_k)\| = d^{-1/2}\|\mathbf{b}_k\|. \tag{56}$$

This leads to $\mathsf{S}(\mathbf{b}_k) = d^{-1/2}\|\mathbf{b}_k\|[1 \cdots 1]^\top$.

In the case of $\beta_1 = 0$, $\mathbf{a}_k = \nabla\mathcal{L}(\mathbf{x}_k)$, for any $\mathbf{b}_k$. When we apply the standardization, we obtain:

$$\mathbf{c}_k = r_k \nabla\mathcal{L}(\mathbf{x}_k) \oslash \frac{\mathsf{S}(\mathbf{b}_k)}{d^{-1/2}\|\mathsf{S}(\mathbf{b}_k)\|} = \nabla\mathcal{L}(\mathbf{u}_k) \oslash [1 \cdots 1]^\top = \nabla\mathcal{L}(\mathbf{u}_k). \tag{57}$$

The direction lies in the tangent space because, by Lemma 1, the gradient belongs to it.

In the generic scheme, using the standardization gives:

$$\begin{aligned}\mathbf{x}_{k+1} &= \mathbf{x}_k - \eta_k \mathbf{a}_k \oslash \mathsf{S}(\mathbf{b}_k) \qquad (58)\\ &= \mathbf{x}_k - \eta_k \mathbf{a}_k \oslash (d^{-1/2}\|\mathbf{b}_k\|[1\cdots 1]^\top) \qquad (59)\\ &= \mathbf{x}_k - \eta_k \mathbf{a}_k/(d^{-1/2}\|\mathbf{b}_k\|). \qquad (60)\end{aligned}$$

This means that the standardization consists in replacing the Hadamard division by $\mathbf{b}_k$ with a scalar division by $d^{-1/2}\|\mathbf{b}_k\|$.

In the case of Adam, we recall that:

$$\mathbf{b}_k = \frac{1-\beta_1^{k+1}}{1-\beta_1}\sqrt{\frac{\mathbf{v}_k}{1-\beta_2^{k+1}}+\epsilon}\,. \tag{(9)}$$

Omitting ϵ for simplicity we have:

$$d^{-1/2}\|\mathbf{b}_k\| = \frac{1-\beta_1^{k+1}}{1-\beta_1}\left(\frac{1}{1-\beta_2^{k+1}}\right)^{\frac{1}{2}} d^{-1/2}\|\sqrt{\mathbf{v}_k}\|. \tag{61}$$

Let us calculate $\|\sqrt{\mathbf{v}_k}\|$. Developing the recursion of $\mathbf{v}_k$, as defined in Eq. (8), leads to:

$$\mathbf{v}_k = (1-\beta_2)\sum_{i=0}^{k}\beta_2^{k-i}\left(\nabla\mathcal{L}(\mathbf{x}_i)+\lambda\mathbf{x}_i\right)^2, \tag{62}$$

$$\sqrt{\mathbf{v}_k} = \sqrt{1-\beta_2}\sqrt{\sum_{i=0}^{k}\beta_2^{k-i}\left(\nabla\mathcal{L}(\mathbf{x}_i)+\lambda\mathbf{x}_i\right)^2}, \tag{63}$$

where the square and the square-root are element-wise operations. Hence, if we take the square norm:

$$\begin{aligned}\|\sqrt{\mathbf{v}_k}\|^2 &= (1-\beta_2)\sum_{j=1}^{d}\left(\sqrt{\sum_{i=0}^{k}\beta_2^{k-i}\left(\nabla\mathcal{L}(\mathbf{x}_i)+\lambda\mathbf{x}_i\right)^2}\right)_j^2\\ &= (1-\beta_2)\sum_{j=1}^{d}\sum_{i=0}^{k}\beta_2^{k-i}\left(\nabla\mathcal{L}(\mathbf{x}_i)+\lambda\mathbf{x}_i\right)_j^2\\ &= (1-\beta_2)\sum_{i=0}^{k}\beta_2^{k-i}\sum_{j=1}^{d}\left(\nabla\mathcal{L}(\mathbf{x}_i)+\lambda\mathbf{x}_i\right)_j^2\\ &= (1-\beta_2)\sum_{i=0}^{k}\beta_2^{k-i}\|\nabla\mathcal{L}(\mathbf{x}_i)+\lambda\mathbf{x}_i\|^2, \qquad (64)\end{aligned}$$

where the j subscript denotes the j-th element of the vector. It is exactly the order-2 moment of the gradient norm.

Therefore, we define the scalar v_k:

$$v_k = \beta_2 v_{k-1} + (1-\beta_2)d^{-1}\|\nabla\mathcal{L}(\mathbf{x}_k)+\lambda\mathbf{x}_k\|^2, \tag{65}$$

which is the order-2 moment of the gradient norm with a factor d^{-1}. It verifies $\sqrt{v_k} = d^{-1/2}\|\sqrt{\mathbf{v}_k}\|$, needed for the scalar division stated in Eq. (61). By applying the bias correction, it gives the formula given in the paper of Adam w/o (a):

$$\begin{aligned}\mathbf{x}_{k+1} &= \mathbf{x}_k - \eta_k\frac{\mathbf{m}_k}{1-\beta_1^{k+1}}\Big/\sqrt{\frac{v_k}{1-\beta_2^{k+1}}+\epsilon}, \qquad (66)\\ \mathbf{m}_k &= \beta_1\mathbf{m}_{k-1}+(1-\beta_1)(\nabla\mathcal{L}(\mathbf{x}_k)+\lambda\mathbf{x}_k), \qquad (67)\\ v_k &= \beta_2 v_{k-1} + (1-\beta_2)d^{-1}\|\nabla\mathcal{L}(\mathbf{x}_k)+\lambda\mathbf{x}_k\|^2. \qquad (68)\end{aligned}$$

Note that the previous demonstration makes the factor d^{-1} appear in v_k to have exactly the scheduling effect of Adam without the deformation.

Clarification on Adam without deformed gradients and no additional radial terms (ab).

We introduce the rescaling and transport transformation of the momentum to neutralize the identified effects on the effective direction (cf. Section 4.2). The resulting, new $\mathbf{c}_k$ is orthogonal to $\mathbf{u}_k$ and does not contribute in the effective learning rate tuning with its radial part.

To avoid gradient history leaving the tangent space and thus neutralize **(b)**, we perform a parallel transport of the momentum $\mathbf{a}_{k-1}$ from the corresponding point on the sphere $\mathbf{u}_{k-1}$ to the new point $\mathbf{u}_k$ denoted as $\Gamma^{\mathbf{u}_k}_{\mathbf{u}_{k-1}}(\mathbf{a}_{k-1})$ at each iteration $k \geq 1$. Figure 12(c) illustrates the transport of a gradient. The parallel transport between two points associates each vector of the tangent space of the first point to a vector of the second tangent space by preserving the scalar product with the derivatives along the geodesic. Consequently, the gradients accumulated in the resulting momentum now lie in the tangent space of $\mathbf{u}_k$ at each step. This neutralizes the additional radial terms phenomena from $\mathbf{c}_k^{\text{grad}}$. Since $\mathbf{u}_{k-1}$, $\mathbf{u}_k$ and $\mathbf{a}_k$ are coplanar, the transport of the momentum on the hypersphere can be expressed as a rotation:

$$\mathsf{T}(\mathbf{a}_{k-1}) \overset{\text{def}}{=} \Gamma^{\mathbf{u}_k}_{\mathbf{u}_{k-1}}(\mathbf{a}_{k-1}) = \langle \mathbf{u}_{k-1}, \mathbf{u}_k \rangle \mathbf{a}_{k-1} - \langle \mathbf{a}_{k-1}, \mathbf{u}_k \rangle \mathbf{u}_{k-1}, \tag{69}$$

$$\mathbf{a}_k = \beta \mathsf{T}(\mathbf{a}_{k-1}) + \nabla \mathcal{L}(\mathbf{x}_k) + \lambda \mathbf{x}_k. \tag{70}$$

Although the transport operation is strictly defined on the tangent space only, the scalar product formulation enables its extension to the whole space. The transformation is linear and $\mathsf{T}(\mathbf{u}_{k-1}) = 0$. We thus have:

$$\mathsf{T}(\mathbf{a}_{k-1} - \lambda \mathbf{u}_{k-1}) = \mathsf{T}(\mathbf{a}_{k-1}). \tag{71}$$

In the previous formulation, we see that the L_2 component is not transported and does not contribute in the new momentum. Finally, the momentum only contains the contribution of the current L_2 regularization. This means that the RT transformation decouples the L_2 regularization and thus neutralizes the additional radial terms from $\mathbf{c}_k^{L_2}$.

Clarification on Adam without deformed gradients, no additional radial terms and no radius ratio (abc).

To avoid the ratio $\frac{r_k}{r_i}$ in the effective learning direction and thus to cancel **(c)**, we rescale the momentum in the update by the factor $\frac{r_{k-1}}{r_k}$ at each iteration $k \geq 1$. From Lemma. 1, we obtain:

$$\mathsf{R}(\mathbf{a}_{k-1}) \overset{\text{def}}{=} \frac{r_{k-1}}{r_k} \mathbf{a}_{k-1} \tag{72}$$

$$\mathbf{a}_k = \beta \mathsf{R}(\mathbf{a}_{k-1}) + \nabla \mathcal{L}(\mathbf{x}_k) + \lambda \mathbf{x}_k \tag{73}$$

$$= \frac{1}{r_k} \Big(\sum_{i=0}^{k} \beta^{k-i} (\nabla \mathcal{L}(\mathbf{u}_k) + \lambda r_k r_i \mathbf{u}_i) \Big). \tag{74}$$

Note that now, the factor $\frac{r_k}{r_i}$ is not contained anymore in the gradient contribution of $\mathbf{c}_k = r_k \mathbf{a}_k$, which neutralizes the radius ratio phenomenon.

We can note that R and T are commutative and that we can combine them in a simple concise scalar expression:

$$\mathsf{RT}(\mathbf{a}_{k-1}) \overset{\text{def}}{=} \frac{\langle \mathbf{x}_k, \mathbf{x}_{k-1} \rangle \mathbf{a}_{k-1} - \langle \mathbf{x}_k, \mathbf{a}_{k-1} \rangle \mathbf{x}_{k-1}}{\langle \mathbf{x}_k, \mathbf{x}_k \rangle}, \tag{75}$$

$$\mathbf{a}_k = \beta \mathsf{RT}(\mathbf{a}_{k-1}) + \nabla \mathcal{L}(\mathbf{x}_k) + \lambda \mathbf{x}_k. \tag{76}$$

This new momentum leads to $\mathbf{c}_k = \mathbf{c}_k^{\mathsf{RT}} + r_k^2 \lambda \mathbf{u}_k$ with $\langle \mathbf{c}_k, \mathbf{u}_k \rangle = \lambda r_k^2$ and $\mathbf{c}_k^{\perp} = \mathbf{c}_k^{\mathsf{RT}}$. The latter relies only on the trajectory on the hypersphere and always lies in the tangent space:

$$\mathbf{c}_k^{\mathsf{RT}} = \beta \Gamma^{\mathbf{u}_k}_{\mathbf{u}_{k-1}}(\mathbf{c}_{k-1}^{\mathsf{RT}}) + \nabla \mathcal{L}(\mathbf{u}_k). \tag{77}$$

The final Adam w/o (abc) scheme reads:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \frac{\mathbf{m}_k}{1-\beta_1^{k+1}} / \sqrt{\frac{v_k}{1-\beta_2^{k+1}} + \epsilon}, \tag{78}$$

$$\mathbf{m}_k = \beta_1 \mathsf{RT}(\mathbf{m}_{k-1}) + (1-\beta_1)(\nabla\mathcal{L}(\mathbf{x}_k) + \lambda\mathbf{x}_k), \tag{79}$$

$$v_k = \beta_2 \frac{r_{k-1}^2}{r_k^2} v_{k-1} + (1-\beta_2) d^{-1} \|\nabla\mathcal{L}(\mathbf{x}_k) + \lambda\mathbf{x}_k\|^2. \tag{80}$$

We also rescale the introduced scalar v_k at each step with the factor $\frac{r_{k-1}^2}{r_k^2}$. This removes the radius from the gradient contribution of the scheduling $\nu^{\mathsf{R}} = r_k v_k$, in contrast with ν_k from Eq. (43). The new scheduling effect reads:

$$\boxed{\nu_k^{\mathsf{R}} = d^{-1/2} \frac{1-\beta_1^{k+1}}{1-\beta_1} \sqrt{\frac{1-\beta_2}{1-\beta_2^{k+1}}} \Big(\sum_{i=0}^{k} \beta_2^{k-i} \|\nabla\mathcal{L}(\mathbf{u}_i) + \lambda r_i r_k \mathbf{u}_i\|^2 \Big)^{1/2}}.$$

D.3 Training and implementation details

To assess empirically the significance of the above phenomena in the context of CNNs with BN, we evaluate the different variants of AdamW, AdamG, Adam w/o (a), w/o (ab), w/o (abc) over a variety of datasets and architectures.

Note that the set of parameters $\boldsymbol{\theta}$ of a CNN with BN layers can be split in two disjoint subsets: $\boldsymbol{\theta} = \mathcal{F} \cup \mathcal{R}$, where $\mathcal{F}$ is the set of groups of radially-invariant parameters and $\mathcal{R}$ the remaining parameters. As demonstrated in Appendix A, the subset $\mathcal{F}$ includes parameters of all filters followed by BN. Since we are only interested in comparing optimization on $\mathcal{F}$, Adam variants w/o (a), w/o (ab), w/o (abc), AdamW AdamG are applied only to the optimization of the parameters in $\mathcal{F}$ whereas the ones in $\mathcal{R}$ are optimized with the original Adam scheme. The algorithm of Adam w/o (abc) is illustrated in Algorithm 1.

For each optimization scheme, each dataset and each architecture, the same grid search range and budget was performed while mini-batch size was fixed. We used a mini-batch size of 128 for SVHN, CIFAR10 and CIFAR100. The learning rates η varied in $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, the weight decay in $10^{-3} \cdot \{0, \frac{1}{128}, \frac{1}{64}, \frac{1}{32}, \frac{1}{32}, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}\}$ (similar to Loshchilov & Hutter (2019)), the momentum was fixed to 0.9 (β_1 for variants of Adam) and the order-two moment β_2 in $\{0.99, 0.999, 0.9999\}$ (as in Kingma & Ba (2015)).

We used the same step-wise learning rate scheduler for each method. For SVHN, CIFAR10 and CIFAR100, models were trained for 405 epochs, and the learning rate multiplied by 0.1 at epochs 135, 225 and 315.

The optimization schemes introduced in this paper do not change the complexity in time of the algorithm. During the update of parameters in a layer, we only do a temporary copy of the parameter tensor just before the update to perform the RT transformation. This temporary copy is flushed after the RT transformation. Nothing permanent is stored in the optimizer.

Note that, for each architecture and each dataset, the same learning rate was systematically found for each method while the momentum factor was fixed at 0.9 (cf. Table 3).

Other hyperparameters, e.g., L_2 regularization and order-2 moment, are illustrated in Table 4.

D.4 Additionnal empirical results

In this section we provide the mean loss training curves associated to Table 2 for every Adam variant.

Table 3: Best learning rate and momentum factor. We systematically found the same learning rate for each dataset and architecture while the momentum factor was fixed to 0.9.

Method	η_0	β, β_1
Adam w/o (a)	0.001	0.9
Adam w/o (ab)	0.001	0.9
Adam w/o (abc)	0.001	0.9
Adam	0.001	0.9
AdamW	0.001	0.9
AdamG	0.01	0.9

Algorithm 1 Adam w/o (abc) and its algorithm illustrated for a filter $\mathbf{x} \in \mathbb{R}^d$ followed by BN. Steps that are different from Adam are shown in highlight. For non-convolutional layers we use standard Adam.

```
Require: $\beta_1, \beta_2 \in [0, 1)$; $\lambda, \eta \in \mathbb{R}$; $\mathcal{L}(\mathbf{x})$
initialize step $k \leftarrow -1$; $\mathbf{m}_k \leftarrow 0$; $v_k \leftarrow 0$; $\mathbf{x} \in \mathbb{R}^d$
while stopping criterion not met do
    $k \leftarrow k + 1$
    $\mathbf{g} \leftarrow \nabla\mathcal{L}(\mathbf{x}_k) + \lambda\mathbf{x}_k$
    $\mathbf{m}_k \leftarrow \beta_1\mathbf{m}_{k-1} + (1 - \beta_1)\mathbf{g}$
    $v_k \leftarrow \beta_2 v_{k-1} + (1 - \beta_2)d^{-1}\mathbf{g}^\top\mathbf{g}$
    $\hat{\mathbf{m}} \leftarrow \mathbf{m}_k/(1 - \beta_1^{k+1})$
    $\hat{v} \leftarrow v_k/(1 - \beta_2^{k+1})$
    $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \eta\hat{\mathbf{m}}/(\sqrt{\hat{v}} + \epsilon)$
    $\mathbf{m}_k \leftarrow \mathbf{m}_k(\mathbf{x}_{k+1}^\top\mathbf{x}_k\mathbf{m}_k - \mathbf{m}_k^\top\mathbf{x}_{k+1}\mathbf{x}_k)/(\mathbf{x}_{k+1}^\top\mathbf{x}_{k+1})$
    $v_k \leftarrow v_k(\mathbf{x}_k^\top\mathbf{x}_k/\mathbf{x}_{k+1}^\top\mathbf{x}_{k+1})$
return resulting parameters $\mathbf{x}_k$
```

Table 4: Best L_2 regularization (λ) and order-2 moment factors (β_2).

Setup			Adam	AdamW	AdamG	Adam w/o (a)	Adam w/o (ab)	Adam w/o (abc)
CIFAR10	ResNet20	λ	0.000250	0.000500	0.000125	0.000031	000125	0.000500
		β_2	0.99	0.99	0.99	0.99	0.99	0.99
	ResNet18	λ	0.000250	0.000008	0.000063	0.000250	000125	0.000016
		β_2	0.999	0.99	0.99	0.99	0.99	0.99
	VGG16	λ	0.000250	0.000031	0.000250	0.000063	0.0	0.000031
		β_2	0.999	0.999	0.999	0.999	0.99	0.999
CIFAR100	ResNet18	λ	0.000125	0.000125	0.000125	0.000125	000125	0.0
		β_2	0.999	0.99	0.99	0.99	0.99	0.999
	VGG16	λ	0.000063	0.000016	0.000063	0.000063	000125	0.000008
		β_2	0.99	0.99	0.99	0.99	0.99	0.99
SVHN	ResNet18	λ	0.0	0.000008	0.000500	0.000031	0.0005	0.000008
		β_2	0.999	0.999	0.99	0.99	0.999	0.999
	VGG16	λ	0.0	0.000031	0.000500	0.000008	0.00025,	0.000250
		β_2	0.99	0.99	0.99	0.99	0.99	0.999

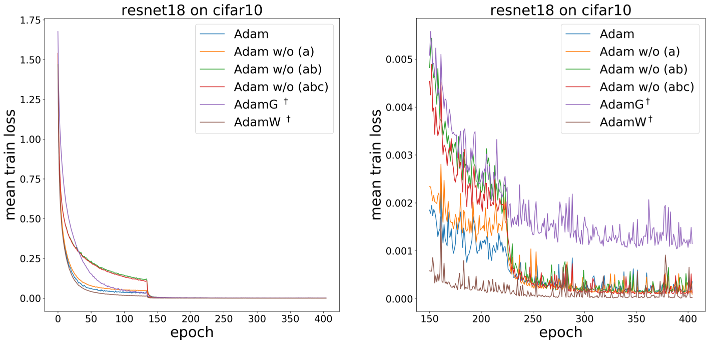

Figure 7: Training speed comparison with ResNet18 on CIFAR10. *Left:* Mean training loss over all training epochs (averaged across 5 seeds) for different Adam variants. *Right:* Zoom-in on the last epochs. Please refer to Table 2 for the corresponding accuracies.

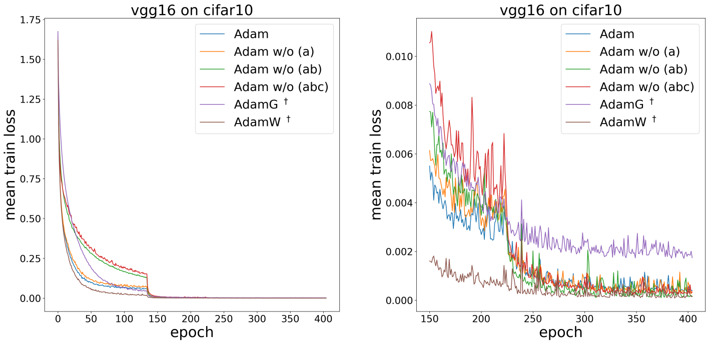

Figure 8: Training speed comparison with VGG16 on CIFAR10. *Left:* Mean training loss over all training epochs (averaged across 5 seeds) for different Adam variants. *Right:* Zoom-in on the last epochs. Please refer to Table 2 for the corresponding accuracies.

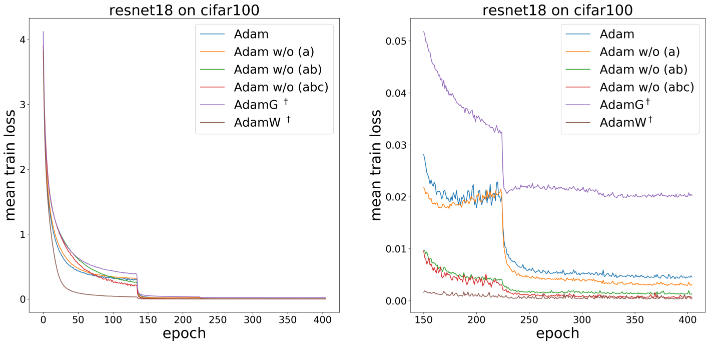

Figure 9: Training speed comparison with ResNet18 on CIFAR100. *Left:* Mean training loss over all training epochs (averaged across 5 seeds) for different Adam variants. *Right:* Zoom-in on the last epochs. Please refer to Table 2 for the corresponding accuracies.

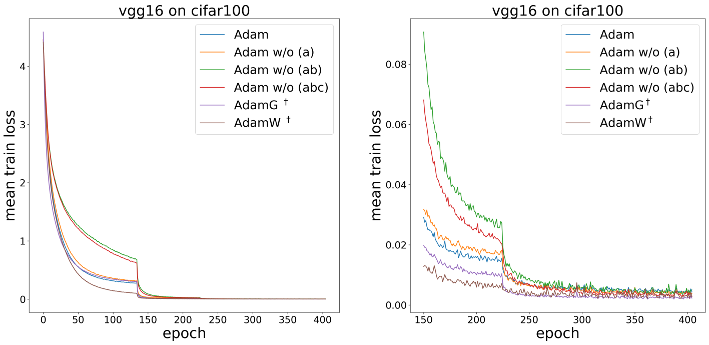

Figure 10: Training speed comparison with VGG16 on CIFAR100. *Left:* Mean training loss over all training epochs (averaged across 5 seeds) for different Adam variants. *Right:* Zoom-in on the last epochs. Please refer to Table 2 for the corresponding accuracies.

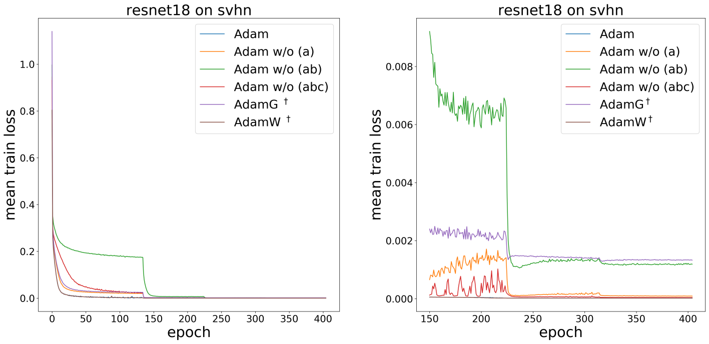

Figure 11: Training speed comparison with ResNet18 on SVHN. *Left:* Mean training loss over all training epochs (averaged across 5 seeds) for different Adam variants. *Right:* Zoom-in on the last epochs. Please refer to Table 2 for the corresponding accuracies.

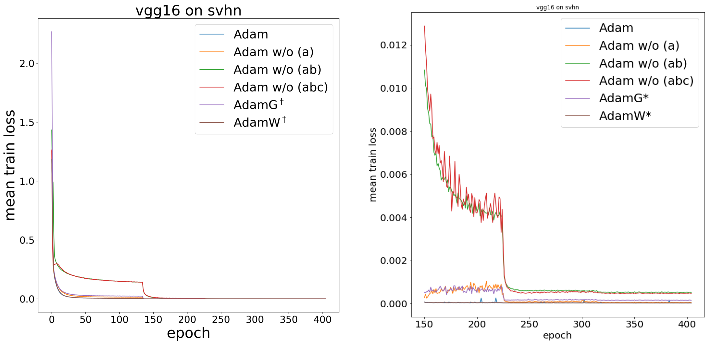

Figure 12: Training speed comparison with VGG16 on SVHN. *Left:* Mean training loss over all training epochs (averaged across 5 seeds) for different Adam variants. *Right:* Zoom-in on the last epochs. Please refer to Table 2 for the corresponding accuracies.