Théorie du Deep Learning le 28 juin 2021

Spherical perspective on learning with Batch Normalization

Simon Roburin

Yann de Mont-Marin

Andrei Bursuc

Renaud Marlet

Patrick Pérez

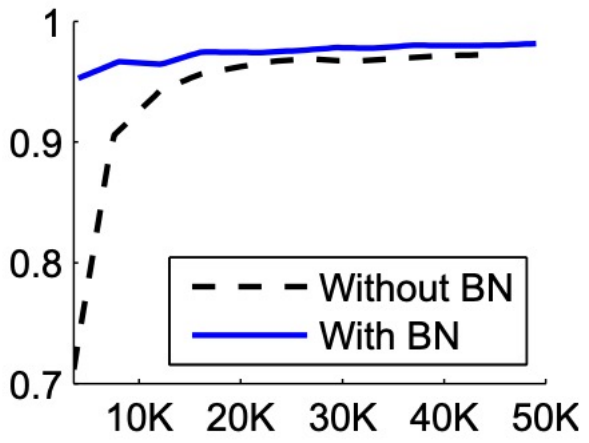
Mathieu Aubry

Overview

- Background on Radial Invariance
- Spherical Framework
- Application 1: theoretical link between SGD & variant of Adam
- Application 2: construction of new optimizers based on geometrical analysis

What is Batch Normalization (BN) ?

- Normalization layer
- Improves performances & Optimization stability
- Widely used in modern CNNs & MLPs architectures



Test accuracy of a simple MLP trained on MNIST with & without BN

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift – S. Ioffe, C. Szegedy (2015)≈

Batch Normalization in a CNN

Let's consider $\mathbf{x} \in \mathbb{R}^{C \times K}$ a single convolutional filter where C is the number of channels and K the kernel size.

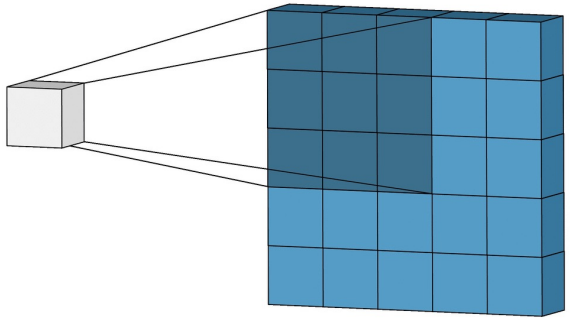


Illustration of a convolution on a channel of the input

On each batch $\mathbf{s}_b \in \mathbb{R}^{C \times D}$, the convolution ϕ reads $\mathbf{t}_b = \phi(\mathbf{s}_b, \mathbf{x}) \in \mathbb{R}^D$.

Batch Normalization in a CNN

Batch Normalization centers and normalizes the output over the batch and spatial dimension:

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

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

$$\hat{\mathbf{t}}_b \stackrel{\text{def}}{=} (\sigma^2 + \epsilon)^{-1/2} (\mathbf{t}_b - \mu \mathbf{1}_D)$$

Radial Invariance

If we rescale the convolutional filter by $\rho > 0$ the new output of convolution reads:

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

Radial Invariance

If we rescale the convolutional filter by $\rho > 0$ the new output of convolution reads:

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

Then $\tilde{\mu} = \rho\mu$ and $\tilde{\sigma} \approx \rho\sigma$.

We have the same output after Batch Norm !

The loss function $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ is radially invariant !

The loss function $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ is radially invariant !

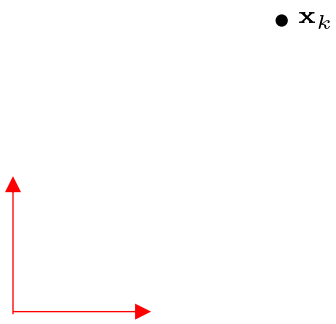
Lemma 1

If $\mathcal{L}$ is radially invariant and almost everywhere differentiable, for all $\rho > 0$ and for all $\mathbf{x} \in \mathbb{R}^d$ where the loss is differentiable, we have:

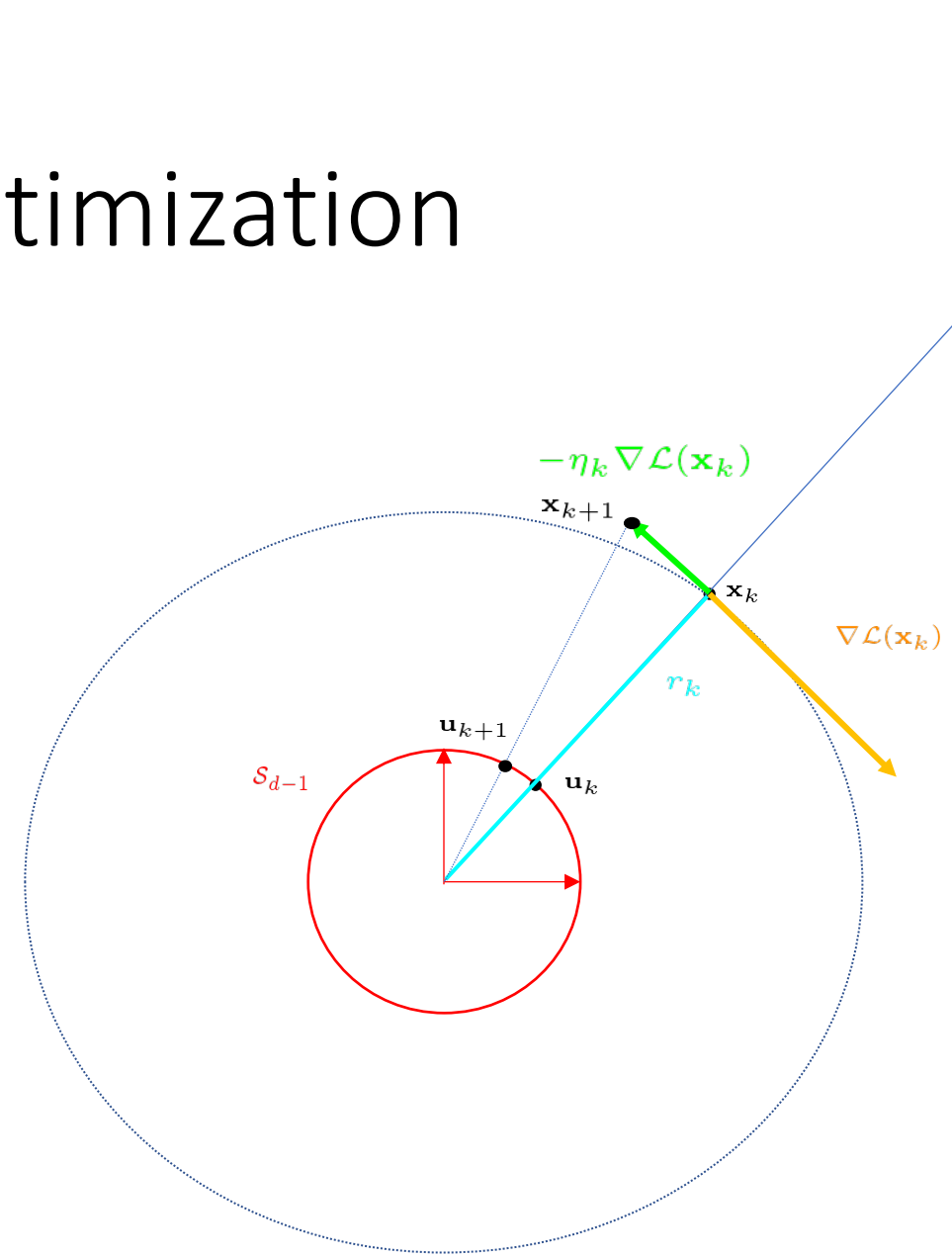
$$\langle \nabla \mathcal{L}(\mathbf{x}), \mathbf{x} \rangle = 0$$

$$\nabla \mathcal{L}(\mathbf{x}) = \rho \nabla \mathcal{L}(\rho \mathbf{x})$$

Impact on Optimization



Impact on Optimization



Spherical Framework

For any radially invariant model

Generic optimization scheme

For a group of radially invariant parameter $\mathbf{x} \in \mathbb{R}^d$:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \mathbf{a}_k \oslash \mathbf{b}_k,$$
$$\mathbf{a}_k = \beta \mathbf{a}_{k-1} + \nabla \mathcal{L}(\mathbf{x}_k) + \lambda \mathbf{x}_k,$$

SGD Momentum

Adam

$$\mathbf{a}_k = \frac{\mathbf{m}_k}{1-\beta_1}, \qquad \mathbf{b}_k = \frac{1-\beta_1^{k+1}}{1-\beta_1}\sqrt{\frac{\mathbf{v}_k}{1-\beta_2^{k+1}} + \epsilon}.$$

Theorem 1

The update of corresponds to an update of its projection through an exponential map at with velocity at order 3:

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

With:

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

Effective learning quantities

Scheme	η^e	$\mathbf{c}^\perp$
SGD	$\frac{\eta}{r^2}$	$\nabla\mathcal{L}(\mathbf{u})$
SGD + L_2	$\frac{\eta}{r^2(1-\eta\lambda)}$	$\nabla\mathcal{L}(\mathbf{u})$
SGD-M	$\frac{\eta}{r^2}(1-\frac{\eta\langle\mathbf{c},\mathbf{u}\rangle}{r^2})^{-1}$	$\mathbf{c}^\perp$
Adam	$\frac{\eta}{r\nu}(1-\frac{\eta\langle\mathbf{c},\mathbf{u}\rangle}{r\nu})^{-1}$	$\mathbf{c}^\perp$

Table 1

Notion of equivalent schemes

Lemma 2

We consider two optimization schemes S and $\tilde{S}$ charecterized by their trajectories $(\mathbf{u}_k)_{k\geq 0}$ and $(\tilde{\mathbf{u}}_k)_{k\geq 0}$, we have:

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

SGD is a variant of Adam on the unit hypersphere

Theorem 2

For any $\lambda > 0, \eta > 0, r_0 > 0$, we have the following equivalence when using the radius dynamic at order 2 in:

$$\begin{cases} \text{(SGD)} \\ \mathbf{x}_0 = r_0 \mathbf{u}_0 \\ \lambda_k = \lambda \\ \eta_k = \eta \end{cases}$$

is scheme equivalent in step with

$$\begin{cases} \text{(AdamG*)} \\ \mathbf{x}_0 = \mathbf{u}_0 \\ \beta = (1 - \eta\lambda)^4 \\ \eta_k = (2\beta)^{-1/2} \\ v_0 = r_0^4 (2\eta^2 \beta^{1/2})^{-1}. \end{cases}$$

Where

$$\text{(AdamG*)}: \begin{cases} \hat{\mathbf{x}}_{k+1} = \mathbf{x}_k - \eta_k \frac{\nabla \mathcal{L}(\mathbf{x}_k)}{\sqrt{v_k}}, \\ \mathbf{x}_{k+1} = \frac{\hat{\mathbf{x}}_{k+1}}{\|\hat{\mathbf{x}}_{k+1}\|}, \\ v_{k+1} = \beta v_k + \|\nabla \mathcal{L}(\mathbf{x}_k)\|^2. \end{cases}$$

**A Riemanian approach to Optimization, M. Cho, J. Lee (2017)*

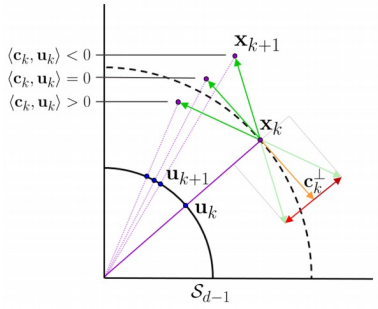
Empirical analysis of geometrical phenomena with Adam

Geometrical phenomena in Adam

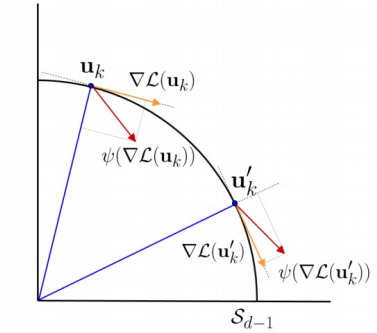
1. In optimization on manifold, direction only depends on trajectory on manifold

2. In Adam direction not only depends on traj on manifold but also on deformed gradients and radial terms

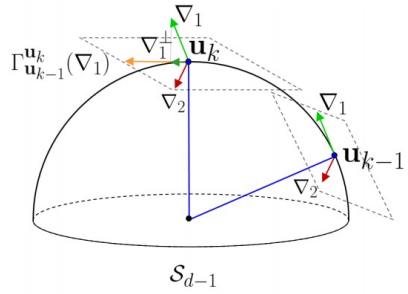
3. We and introduce incremental variants of Adam to nulify these geometric phenomena



(a) Radial scheduling



(b) Anisotropy of deformation



(c) Gradient-history contribution

Empirical results

Method	CIFAR10 ResNet20	CIFAR10 ResNet18	CIFAR10 VGG16	CIFAR100 ResNet18	CIFAR100 VGG16	SVHN ResNet18	SVHN VGG16
Adam	90.98 ± 0.06	93.77 ± 0.20	92.83 ± 0.17	71.30 ± 0.36	68.43 ± 0.16	95.32 ± 0.23	95.57 ± 0.20
AdamW†	90.19 ± 0.24	93.61 ± 0.12	92.53 ± 0.25	67.39 ± 0.27	71.37 ± 0.22	95.13 ± 0.15	94.97 ± 0.08
AdamG†	91.64 ± 0.17	94.67 ± 0.12	93.41 ± 0.17	73.76 ± 0.34	70.17 ± 0.20	95.73 ± 0.05	95.70 ± 0.25
Adam w/o (a)	91.15 ± 0.11	93.95 ± 0.23	92.92 ± 0.11	74.44 ± 0.22	68.73 ± 0.27	95.75 ± 0.09	95.66 ± 0.09
Adam w/o (ab)	$\mathbf{91.92} \pm 0.18$	$\mathbf{95.11} \pm 0.10$	$\mathbf{93.89} \pm 0.09$	$\mathbf{76.15} \pm 0.25$	$\mathbf{71.53} \pm 0.19$	$\mathbf{96.05} \pm 0.12$	$\mathbf{96.22} \pm 0.09$
Adam w/o (abc)	91.81 ± 0.20	94.92 ± 0.05	93.75 ± 0.06	75.28 ± 0.35	71.45 ± 0.13	95.84 ± 0.07	95.82 ± 0.05

- Adam (a) : Adam direction intrinsic to the unit hypersphere
- Adam (b): Adam with radial terms decoupled to the direction
- Adam (c): Adam with corrected radius ratio

Questions ?

Thanks for your attention