Spherical perspective on learning with Batch Normalization
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 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.

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

Illustration of a convolution on a channel of the input
Batch Normalization in a CNN

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

\[ \mu = \frac{1}{BD} \sum_{b,j} t_{b,j}, \]
\[ \sigma^2 = \frac{1}{BD} \sum_{b,j} (t_{b,j} - \mu)^2, \]
\[ \hat{t}_b \overset{\text{def}}{=} (\sigma^2 + \epsilon)^{-1/2} (t_b - \mu 1_D) \]
Radial Invariance

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

\[
\hat{t}_b = \rho \phi(x, s_b) = \rho t_b
\]
Radial Invariance

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

$$\tilde{t}_b = \rho \phi(x, s_b) = \rho t_b$$

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

We have the same output after Batch Norm!
The loss function $\mathcal{L} : \mathbb{R}^d \rightarrow \mathbb{R}$ 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

Spherical perspective on learning with Batch Normalization
Impact on Optimization

Spherical perspective on learning with Batch Normalization
Spherical Framework

For any radially invariant model
Generic optimization scheme

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

$$
\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \mathbf{a}_k \otimes \mathbf{b}_k,
$$

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

**SGD momentum**

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

**Adam**
Theorem 1

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

\[ u_{k+1} = \text{Exp}_{u_k} \left( - \left[ 1 + O \left( \left( \eta_k^e \|c_k^\perp\| \right)^2 \right) \right] \eta_k^e c_k^\perp \right), \]
\[ \approx \text{Exp}_{u_k} \left( -\eta_k^e c_k^\perp \right) \]

With:

\[ c_k \overset{\text{def}}{=} r_k a_k \otimes \frac{b_k}{d^{-1/2}\|b_k\|}, \quad \eta_k^e \overset{\text{def}}{=} \frac{\eta_k}{r_k^2 d^{-1/2}\|b_k\|} \left( 1 - \frac{\eta_k \langle c_k, u_k \rangle}{r_k^2 d^{-1/2}\|b_k\|} \right)^{-1} \]
Effective learning quantities

<table>
<thead>
<tr>
<th>Scheme</th>
<th>$\eta^e$</th>
<th>$c^\perp$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGD</td>
<td>$\frac{\eta}{r^2}$</td>
<td>$\nabla \mathcal{L}(u)$</td>
</tr>
<tr>
<td>SGD + $L_2$</td>
<td>$\frac{\eta}{r^2(1-\eta \lambda)}$</td>
<td>$\nabla \mathcal{L}(u)$</td>
</tr>
<tr>
<td>SGD-M</td>
<td>$\frac{\eta}{r^2} (1 - \frac{\eta \langle c, u \rangle}{r^2})^{-1}$</td>
<td>$c^\perp$</td>
</tr>
<tr>
<td>Adam</td>
<td>$\frac{\eta}{r \nu} (1 - \frac{\eta \langle c, u \rangle}{r \nu})^{-1}$</td>
<td>$c^\perp$</td>
</tr>
</tbody>
</table>

*Table 1*
Notion of equivalent schemes

Lemma 2
We consider two optimization schemes $\mathcal{S}$ and $\tilde{\mathcal{S}}$ characterized by their trajectories $(u_k)_{k \geq 0}$ and $(\tilde{u}_k)_{k \geq 0}$, we have:

$$\begin{cases}
  u_0 = \tilde{u}_0 \\
  \forall k \geq 0, \eta_k^e = \tilde{\eta}_k^e, c_k^\perp = \tilde{c}_k^\perp
\end{cases} \Rightarrow \forall k \geq 0, u_k = \tilde{u}_k.$$
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:

\[
\begin{align*}
(SGD) & \quad \begin{cases} 
    x_0 = r_0u_0 \\
    \lambda_k = \lambda \\
    \eta_k = \eta 
\end{cases} \quad \text{is scheme equivalent in step with}
\end{align*}
\]

\[
\begin{align*}
(AdamG^*) & \quad \begin{cases} 
    x_0 = u_0 \\
    \beta = (1 - \eta \lambda)^4 \\
    \eta_k = (2\beta)^{-1/2} \\
    v_0 = r_0^4(2\eta^2\beta^{1/2})^{-1}. 
\end{cases}
\end{align*}
\]

Where \( (AdamG^*) : \)

\[
\begin{align*}
\hat{x}_{k+1} &= x_k - \eta_k \frac{\nabla L(x_k)}{\sqrt{v_k}} , \\
x_{k+1} &= \frac{\hat{x}_{k+1}}{\|\hat{x}_{k+1}\|} , \\
v_{k+1} &= \beta v_k + \|\nabla L(x_k)\|^2.
\end{align*}
\]

*A Riemannian 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
Empirical results

<table>
<thead>
<tr>
<th>Method</th>
<th>CIFAR10</th>
<th>CIFAR100</th>
<th>SVHN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ResNet20</td>
<td>VGG16</td>
<td>ResNet18</td>
</tr>
<tr>
<td>Adam</td>
<td>90.98 ± 0.06</td>
<td>93.77 ± 0.20</td>
<td>71.30 ± 0.36</td>
</tr>
<tr>
<td>AdamW&lt;sup&gt;+&lt;/sup&gt;</td>
<td>90.19 ± 0.24</td>
<td>93.61 ± 0.12</td>
<td>67.39 ± 0.27</td>
</tr>
<tr>
<td>AdamG&lt;sup&gt;+&lt;/sup&gt;</td>
<td>91.64 ± 0.17</td>
<td>94.67 ± 0.12</td>
<td>73.76 ± 0.34</td>
</tr>
<tr>
<td>Adam w/o (a)</td>
<td>91.15 ± 0.11</td>
<td>93.95 ± 0.23</td>
<td>74.44 ± 0.22</td>
</tr>
<tr>
<td>Adam w/o (ab)</td>
<td>91.92 ± 0.18</td>
<td>95.11 ± 0.10</td>
<td>76.15 ± 0.25</td>
</tr>
<tr>
<td>Adam w/o (abc)</td>
<td>91.81 ± 0.20</td>
<td>94.92 ± 0.05</td>
<td>75.28 ± 0.35</td>
</tr>
</tbody>
</table>

- 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