A Geometric Interpretation of Stochastic Gradient Descent in Deep Learning and Bolzmann Machines

Rita Fioresi
University of Bologna
rita.fioresi@unibo.it
Pratik Chaudhari
University of Pennsylvania
pratikac@seas.upenn.edu

Abstract

Stochastic gradient descent (SGD) is an essential ingredient in training neural networks, though its geometric meaning is not completely understood. We describe a deterministic model in which the trajectories of our dynamical systems are geodesics of a family of metrics arising naturally and encoding the information on the highly non-isotropic gradient noise in SGD. We model our system through an analogy with General Relativity, where we replace the electromagnetic field with the gradient of the loss.

Introduction

Stochastic gradient descent performs un update of the weights $w \in \Omega \subset \mathbb{R}^d$ of a neural network, replacing the ordinary gradient of the loss function $f = \sum_{i=1}^{N} f_i$ with $\nabla_\mathcal{B} f$:

$$dw = -\nabla_\mathcal{B} f dt, \qquad \nabla_\mathcal{B} f = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \nabla f_i \tag{1}$$

where

- dw is the continuous version of the weight update at step j: $w_{j+1} = w_j - \eta \nabla_\mathcal{B} f(w_j)$.
- f_i is the loss relative to the i-th element in our dataset Σ of size $|\Sigma| = N$.
- $\mathcal{B}$ is the minibatch.

The *diffusion matrix* is the variance of $\nabla_\mathcal{B} f$, viewed as a random variable, $\phi : \Sigma \longrightarrow \mathbb{R}^d$, $\phi(z_i) = \nabla f_i$:

$$D(w) = \mathbb{E}[(\phi - \mathbb{E}[\phi])(\phi - \mathbb{E}[\phi])^t] \tag{2}$$

With a direct calculation one shows that:

$$D = \frac{1}{N} \sum_k (\nabla f_k)(\nabla f_k)^t - (\nabla f)(\nabla f)^t = \frac{1}{N^2} (\langle \partial_r \widehat{f}, \partial_s \widehat{f} \rangle) \tag{3}$$

where:

$$\widehat{f} = (f_1 - f_2, f_1 - f_3, \dots, f_{N-1} - f_N) \in \mathbb{R}^{N(N-1)/2}$$

The diffusion matrix measures effectively the *anisotropy* of our data:

$$D = 0 \quad \text{if and only if} \quad \partial_r(f_i) = \partial_r(f_j), \qquad \text{for all} \quad r = 1, \dots, d, \quad i, j = 1, \dots, N$$

Furthermore, D *is singular*: $\mathrm{rk}(D) \leq N - 1$.

Values for N and d for various architectures on CIFAR and SVHN datasets

Architecture	$d = \lvert\text{Weights}\rvert$	$N = \lvert\text{Data}\rvert$, CIFAR	$N = \lvert\text{Data}\rvert$, SVHN
ResNet	1.7M	60K	600K
Wide ResNet	11M	60K	600K
DenseNet (k=12)	1M	60K	600K
DenseNet (k=24)	27.2M	60K	600K

Diffusion Metric and General Relativity

The evolution of a dynamical system in general relativity occurs according to the geodesics with respect to the metric imposed on the Minkowski space by the presence of gravitational masses:

$$\frac{d^2 w^\mu}{dt^2} + \Gamma^\mu_{\rho\sigma} \frac{dw^\rho}{dt} \frac{dw^\sigma}{dt} = \frac{q}{m} F^\mu_\nu \frac{dw^\nu}{dt} \tag{4}$$

where $\Gamma^\mu_{\rho\sigma}$ are the Christoffel symbols for the Levi-Civita connection with metric $g = (g_{ij})$:

$$\Gamma^t_{uv} = \frac{1}{2} g^{tz} (\partial_u g_{vz} - \partial_z g_{uv} + \partial_v g_{uz}) \tag{5}$$

and $\frac{q}{m} F^\mu_\nu$ is a term regarding an external force, e.g. an electromagnetic field.

If we take time derivative of the differential equation ruling the ordinary (i.e. non stochastic) gradient descent:

$$\frac{d^2 w^\mu}{dt^2} = -\frac{d}{dt} \partial_\mu f$$

and we compare with (4), it is clear that $-\frac{d}{dt} \partial_\mu f$ effectively replaces the force term $(q/m) F^\mu_\nu \frac{dw^\nu}{dt}$.

Hence, the geodesic equation (4) models the ordinary GD equation, if we replace the force term with the time derivative of the *gradient of the loss*; furthermore this corresponds to the condition $D = 0$ in SGD dynamics equation (1).

This suggests the definition of a metric, modelling the anisotropy of the system, hence depending on the diffusion matrix:

$$g(w) = \mathrm{id} + \mathcal{E}(w) D(w) \tag{6}$$

with $\mathcal{E}(w) < 1/M_x$, where $M_w = \max\{\lambda\}$ with λ eigenvalues of $D(w)$. This ensures that $g(w)$ is non singular.

As in General Relativity *weak field approximation* (see [1]):

$$g^{-1} = \mathrm{id} - \mathcal{E} D(w)$$

We then have to solve the equation:

$$\frac{d^2 x^k}{dt^2} + \frac{\mathcal{E}}{N^2} \sum_{i,j} \Gamma^k_{ij} \frac{dx^i}{dt} \frac{dx^j}{dt} = \frac{d}{dt} \partial_k f \qquad \text{with} \qquad \Gamma^k_{ij} = \frac{\mathcal{E}}{N^2} \langle \partial_i \partial_j \widehat{f}, \partial_k \widehat{f} \rangle.$$

This leads to the NGD (natural gradient descent) with respect to the diffusion metric:

$$\frac{dw}{dt} = -(I - \mathcal{E} D) \nabla f$$

Main Result

The anisotropy of the SGD:

$$\frac{dw}{dt} = -\nabla_\mathcal{B} f dt,$$

modelled by GRD (General Relativity Descent):

$$\frac{d^2 w^\mu}{dt^2} + \Gamma^\mu_{\rho\sigma} \frac{dw^\rho}{dt} \frac{dw^\sigma}{dt} = -\frac{d}{dt} \partial_\mu f$$

gives the natural gradient descent with respect to the diffusion metric:

$$\frac{dw}{dt} = -(I - \mathcal{E} D) \nabla f = -\nabla_D f \tag{7}$$

provided the approximation:

$$\frac{d^2}{dt^2} \partial_k \widehat{f}_\alpha = 0 \tag{8}$$

holds.

Application: For a two-layer network, commonly used for Deep Learning, (8) holds.

Conclusions

The General Relativity model helps to provide with a deterministic approach to the evolution of the dynamical system described by SGD, leading to the NGD with respect to a new metric: *the diffusion metric*. The results are compatible with [3].

Forthcoming Research

In Restricted Boltzmann machines (RBM) the training occurs via three distinct phases:

1. Positive phase
2. Negative phase
3. Weight update

The update of the weight occurs via the *contrastive divergence*:

$$\frac{dw}{dt} = -\frac{1}{T} \nabla G(w)$$

where G is the loss function and represents the KL divergence of the two probabilities p and p' in the positive and negative phases respectively:

$$G(w) = \sum_{\alpha \in \text{pos/neg conf}} p(v_\alpha) \log \frac{p(v_\alpha)}{p'(v_\alpha)}$$

This is an analog of SGD: not all possible configurations are reached in positive/negative phases:

$$\widehat{G}(w) = \sum_{\alpha \in \text{all conf}} p(v_\alpha) \log \frac{p(v_\alpha)}{p'(v_\alpha)}$$

Hence, $\nabla G(w)$ represents only part of $\nabla \widehat{G}$ taking into account all configurations, similarly to $\nabla_\mathcal{B} f$ in (1).

We plan to explore RGD in this context and establish a connection with the NGD as in (7) in this context.

References

[1] Adler, R., Bazin, M., Schiffer, M., *Introduction to General Relativity*. New York: McGraw-Hill, (1965).

[2] S. Amari, *Natural Gradient Works Efficiently in Learning*, Neural Computation 10, 251-276 (1998).

[3] P. Chaudhari, S. Soatto *Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks*, arXiv:1710.11029.

[4] Chaudhari, P. and Soatto, S. *On the energy landscape of deep networks*. arXiv:1511.06485, (2015).

[5] Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. *Entropy-SGD: biasing gradient descent into wide valleys*. arXiv:1611.01838, (2016).

[6] R. Fioresi, P. Chaudhari, S. Soatto, *A geometric interpretation of stochastic gradient descent using diffusion metrics*, Entropy, 2020.

[7] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). *Gradient-based learning applied to document recognition*. Proceedings of the IEEE, 86(11):22782324.

[8] P. Petersen, *Riemannian Geometry*, GTM, Springer, (1998).