entropy

MDPI

Article

A Geometric Interpretation of Stochastic Gradient Descent Using Diffusion Metrics

Rita Fioresi [1,*], Pratik Chaudhari [2] and Stefano Soatto [3]

[1] Dipartimento di Matematica, piazza Porta San Donato 5, University of Bologna, 40126 Bologna, Italy
[2] Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA; pratikac@seas.upenn.edu
[3] Computer Science Department, University of California, Los Angeles, CA 90095, USA; soatto@cs.ucla.edu
* Correspondence: rita.fioresi@unibo.it

Received: 30 November 2019; Accepted: 10 January 2020; Published: 15 January 2020

Abstract: This paper is a step towards developing a geometric understanding of a popular algorithm for training deep neural networks named stochastic gradient descent (SGD). We built upon a recent result which observed that the noise in SGD while training typical networks is highly non-isotropic. That motivated a deterministic model in which the trajectories of our dynamical systems are described via geodesics of a family of metrics arising from a certain diffusion matrix; namely, the covariance of the stochastic gradients in SGD. Our model is analogous to models in general relativity: the role of the electromagnetic field in the latter is played by the gradient of the loss function of a deep network in the former.

Keywords: stochastic gradient descent; deep learning; general relativity

1. Introduction

Deep neural networks are high-dimensional machine learning models that have demonstrated impressive performance on a number of challenging tasks in computer vision, natural language processing and reinforcement learning [1,2]. These models are typically trained to minimize the misprediction error compared to large amounts of human-annotated data. A large number of diverse models with varying properties are prevalent in these application domains. Despite this diversity, stochastic gradient descent (SGD) is the gold standard for training deep neural networks. It has been shown to obtain good generalization performance; i.e., to train a model that performs well on new data, across a wide range of applications. In spite of this popularity and efficacy, a precise understanding of SGD for deep learning remains elusive.

This paper develops a geometric understanding of stochastic gradient descent. We build upon the work of [3], wherein the authors model the dynamics of SGD as a stochastic differential equation with state-dependent Gaussian noise. We interpret the covariance of this noise, called the diffusion matrix henceforth, as a metric on the parameter space. Our result provides a deterministic Equation (10) that can be compared to SGD near equilibrium points Equation (1). We write the diffusion matrix $D(x)$ in the form Equation (3) to show how it fundamentally captures the anisotropy of the dynamical system underlying SGD. This clarifies how $D(x)$ is one of the key factors that differentiates steady-state solutions of SGD from those of ordinary gradient descent GD (see comparison in [3–5]). Using the diffusion matrix, we then define a family of metrics on the parameter space that we call *diffusion metrics*. We then take the Einstein equation describing the geodesic on a Riemannian manifold, for the motion of a particle subject to a gravitational and electromagnetic field. We replace the electromagnetic force by the ordinary gradient while gravity is taken into account using the diffusion metric itself. After some mild hypotheses on the architecture of the neural network, we obtain the result that geodesics

with respect to this equation correspond precisely to the evolution of a dynamical system, which is not subject to Euclidean gradient descent but to relativistic gradient descent (RGD) with respect to the family of diffusion metrics.

In the end, we obtain Equation (10), which is along the same vein as natural gradient descent in [6], but whose significance is much deeper in the context of SGD, since it stems from the anisotropy of the gradients with respect to the various parameters. This anisotropy encodes the difference between the dynamics of GD and those of SGD. We also compare our result with the ones in [3] and show them to be perfectly compatible. In the end we provide Appendix A for the reader convenience, with a quick review of some key facts of Riemannian geometry.

2. Continuous-Time SGD and the Diffusion Matrix

Stochastic gradient descent performs an update of the weights x of a neural network, replacing the ordinary gradient of the loss function $f = \frac{1}{N}\sum_{i=1}^{N} f_i$ with $\nabla_{\mathcal{B}} f$:

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

where dx represents the continuous version of the weight update at step j: $x_{j+1} = x_j - \eta \nabla_{\mathcal{B}} f(x_j)$, with the learning rate η incorporated into the expression of $\nabla_{\mathcal{B}} f$, and $\mathcal{B}$ is the mini-batch. In the expression of the loss function $f = \frac{1}{N}\sum_{i=1}^{N} f_i$, f_i is the loss relative to the i-th element in our dataset Σ of size $|\Sigma| = N$. We assume that weights belong to a compact subset $\Omega \subset \mathbb{R}^d$ and that the f_is satisfy suitable regularity conditions (see [3] Section 2 for more details).

We define the *diffusion matrix* as the product of the size of the mini-batch $|\mathcal{B}|$ and the variance of $\nabla_{\mathcal{B}} f$, viewed as a random variable, $\phi : \Sigma \longrightarrow \mathbb{R}^d$, $\phi(z_i) = \nabla f_i$:

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

Notice that $D(x) \in \mathbb{R}^{d\times d}$ and does not depend on the size of the mini-batch; it only depends on the weights x, loss function f and the dataset Σ. With a direct calculation one can show 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:

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

and $\langle , \rangle$ is the euclidean scalar product. In fact:

$$\begin{aligned}
D_{rs} &= \frac{1}{N}\sum_{k=1}^{N} \partial_r f_k\, \partial_s f_k - \frac{1}{N^2}\sum_{i,j=1}^{N} \partial_r f_i\, \partial_s f_j \\
&= \frac{1}{N^2}[N(\partial_r f_1\, \partial_s f_1 + \cdots + \partial_r f_N\, \partial_s f_N) + \\
&\quad -(\partial_r f_1\, \partial_s f_1 + \partial_r f_1\, \partial_s f_2 + \cdots + \partial_r f_N\, \partial_s f_N)] = \\
&= \frac{1}{N^2}[(N-1)\partial_r f_1\, \partial_s f_1 - \partial_r f_1\, \partial_s f_2 - \cdots - \partial_r f_1\, \partial_s f_N + \\
&\quad -\partial_r f_2\, \partial_s f_1 + (N-1)\partial_r f_2\, \partial_s f_2 - \cdots - \partial_r f_2\, \partial_s f_N + \ldots \\
&\quad -\partial_r f_N\, \partial_s f_1 - \partial_r f_N\, \partial_s f_2 + \cdots + (N-1)\partial_r f_N\, \partial_s f_N]
\end{aligned}$$

which gives:

$$\begin{aligned} D_{rs} &= \tfrac{1}{N^2}[(\partial_r f_1 - \partial_r f_2)(\partial_s f_1 - \partial_s f_2) + (\partial_r f_1 - \partial_r f_3)(\partial_s f_1 - \partial_s f_3) + \dots \\ &+ (\partial_r f_1 - \partial_r f_N)(\partial_s f_1 - \partial_s f_N) + (\partial_r f_2 - \partial_r f_3)(\partial_s f_2 - \partial_s f_3) + \dots \\ &+ (\partial_r f_{N-1} - \partial_r f_N)(\partial_s f_{N-1} - \partial_s f_N) = \tfrac{1}{N^2}(\langle \partial_r \widehat{f}, \partial_s \widehat{f} \rangle). \end{aligned}$$

The diffusion matrix measures effectively the *anisotropy* of our data: $D = 0$ if and only if $\partial_r(f_i) = \partial_r(f_j)$ for all $r = 1, \dots, d$ and $i, j = 1, \dots, N$. In other words, the diffusion matrix measures how the loss of each datum depends, at first order, on the weights in a different way with respect to the loss of each of the other data-points. So, it tells us how much we should expect the SGD dynamics to differ from the GD one. Notice that the expression (Equation (3)) gives us immediately a bound on the rank of D; namely, $\mathrm{rk}(D) \leq N - 1$.

The Table 1 suggests that in many algorithms currently available, the diffusion matrix has low rank; hence, it is singular; this just by comparing the size d of D and its rank which bound by N. This fact turns out to be very important in the construction of the diffusion metrics, as we will see below.

3. Diffusion Metrics and General Relativity

The evolution of a dynamical system in general relativity takes place along the geodesics according to the metric imposed on the Minkowski space by the presence of gravitational masses. The equation for such geodesics, once Einstein's equation is solved, is:

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

where $\Gamma^\mu_{\rho\sigma}$ are the Christoffel symbols for the Levi–Civita connection:

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

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

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

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

and compare with Equation (4), observe that $-\frac{d}{dt}\,\partial_\mu f$ effectively replaces the force term $(q/m)F^\mu_\nu\frac{dx^\nu}{dt}$. Hence, the geodesic equation Equation (4) models the ordinary GD equation if we take a constant metric and replace the force term with the gradient of the loss; furthermore, this corresponds to the condition $D = 0$ in SGD dynamics Equation (1).

This suggests that one may define a metric dependent on the diffusion matrix; this metric should become constant when $D = 0$. As a side remark, notice that since D is singular in many important practical applications (see Table 1), it is not reasonable to use it to define the metric itself. On the other hand, since D measures the anisotropy of the weight space, it is reasonable to employ it to perturb the euclidean metric. So the stochastic nature of the dynamical system ruled by the SGD is replaced by a perturbation of the dynamics for the ordinary gradient descent. As an analogy, the presence of (small) masses in space, in the weak field approximation (see [7]) of general relativity, generates gravity; this motivates our (small) deformation of the euclidean metric using the diffusion matrix.

Table 1. Values for N and d for various architectures on CIFAR and SVHN datasets (see [8]).

Architecture	$d = \|$**Weights**$\|$	$N = \|$**Data**$\|$, **CIFAR**	$N = \|$**Data**$\|$, **SVHN**
ResNet	1.7 M	60 K	600 K
Wide ResNet	11 M	60 K	600 K
DenseNet (k = 12)	1 M	60 K	600 K
DenseNet (k = 24)	27.2 M	60 K	600 K

At each point $x \in \Omega$, we define a metric called *diffusion metric* as

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

with $\mathcal{E}(x) < 1/M_x$, where $M_x = \max\{\lambda_k : \lambda_k \text{ is an eigenvalue of } D(x)\}$. This ensures that g is non-singular at each $x \in \Omega$. We have, thus, defined a family of metrics that depend on the real parameter $\mathcal{E}$. We expect this model to approximate the solution to the Fokker–Planck equation when the parameter β^{-1}, whose interpretation is related with the temperature, is very small (see [3] for the notation and more details).

Notice that our heuristic hypothesis on $\mathcal{E}$ allows us to make the so called *weak field approximation* (see [7]):

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

Hence, we have the following expression for the Christoffel's symbols (in this approximation we discard $\mathcal{O}(\mathcal{E}^2)$):

$$\Gamma^w_{uv} = \frac{1}{2}\sum_z (\delta_{wz} - \mathcal{E}d_{wz})\,\mathcal{E}\,(\partial_u d_{vz} - \partial_z d_{uv} + \partial_v d_{uz}) = \frac{\mathcal{E}}{2}\,(\partial_u d_{vw} - \partial_w d_{uv} + \partial_v d_{uw}) \tag{7}$$

where d_{ij} are the coefficients of D, and δ_{wz} is the Kronecker delta.

Let us now compute the Christoffel's symbols and then substitute them into the geodesic equation given by Equation (4).

$$\begin{aligned}\Gamma^k_{ij} &= \tfrac{\mathcal{E}}{2N^2}[\langle \partial_i\partial_j \widehat{f}, \partial_k \widehat{f}\rangle + \langle \partial_j \widehat{f}, \partial_i\partial_k \widehat{f}\rangle - \langle \partial_i\partial_k \widehat{f}, \partial_j \widehat{f}\rangle - \langle \partial_i \widehat{f}, \partial_k\partial_j \widehat{f}\rangle \\ &\quad + \langle \partial_j\partial_k \widehat{f}, \partial_i \widehat{f}\rangle + \langle \partial_k \widehat{f}, \partial_i\partial_j \widehat{f}\rangle] = \tfrac{\mathcal{E}}{N^2}\langle \partial_i\partial_j \widehat{f}, \partial_k \widehat{f}\rangle.\end{aligned}$$

Let us substitute in Equation (4) (writing the sum now):

$$\frac{d^2x^k}{dt^2} + \frac{\mathcal{E}}{N^2}\sum_{i,j}\langle \partial_i\partial_j \widehat{f}, \partial_k \widehat{f}\rangle \frac{dx^i}{dt}\frac{dx^j}{dt} = \frac{d}{dt}\partial_k f. \tag{8}$$

Let us concentrate on the expression:

$$\begin{aligned}\tfrac{\mathcal{E}}{N^2}\textstyle\sum_{i,j}\langle \partial_i\partial_j \widehat{f}, \partial_k \widehat{f}\rangle \frac{dx^i}{dt}\frac{dx^j}{dt} &= \tfrac{\mathcal{E}}{N^2}\textstyle\sum_{i,j,\alpha} \partial_i\partial_j \widehat{f}_\alpha\, \partial_k \widehat{f}_\alpha \frac{dx^i}{dt}\frac{dx^j}{dt} \\ &= \tfrac{\mathcal{E}}{N^2}\textstyle\sum_\alpha \frac{d^2 \widehat{f}_\alpha}{dt^2}\partial_k \widehat{f}_\alpha.\end{aligned}$$

Now, we take the integral in dt (we compute the parts twice):

$$\begin{aligned}\tfrac{\mathcal{E}}{N^2}\textstyle\int\sum_\alpha \frac{d^2 \widehat{f}_\alpha}{dt^2}\partial_k \widehat{f}_\alpha\, dt &= \tfrac{\mathcal{E}}{N^2}\textstyle\sum_\alpha \left[\frac{d \widehat{f}_\alpha}{dt}\partial_k \widehat{f}_\alpha - \int \frac{d \widehat{f}_\alpha}{dt}\frac{\partial_k \widehat{f}_\alpha}{dt}\, dt\right] \\ &= \tfrac{\mathcal{E}}{N^2}\textstyle\sum_\alpha \left[\frac{d \widehat{f}_\alpha}{dt}\partial_k \widehat{f}_\alpha - \widehat{f}_\alpha \frac{d}{dt}\partial_k \widehat{f}_\alpha + \int \widehat{f}_\alpha \frac{d^2}{dt^2}\partial_k \widehat{f}_\alpha\, dt\right].\end{aligned}$$

Notice that, in many practical applications we have:

$$\frac{d^2}{dt^2}\partial_k \widehat{f}_\alpha = 0 \tag{9}$$

because $\partial_i \partial_j \partial_k \widehat{f} = 0$.

Hence

$$\begin{aligned} \frac{\mathcal{E}}{N^2}\int \sum_\alpha \frac{d^2\widehat{f}_\alpha}{dt^2}\partial_k \widehat{f}_\alpha \, dt &= \frac{\mathcal{E}}{N^2}\sum_{\alpha,\ell}\left[\partial_\ell \widehat{f}_\alpha \frac{dx^\ell}{dt}\partial_k \widehat{f}_\alpha - \widehat{f}_\alpha \frac{d}{dt}\partial_k \widehat{f}_\alpha\right] \\ &= \mathcal{E}\sum_\ell d_{k,\ell}\frac{dx^\ell}{dt} - \frac{\mathcal{E}}{N^2}\sum_\alpha \widehat{f}_\alpha \frac{d}{dt}\partial_k \widehat{f}_\alpha. \end{aligned}$$

We now substitute the obtained expression into the Equation (8), taking the integral in dt:

$$\frac{dx^k}{dt} + \mathcal{E}\sum_\ell d_{k,\ell}\frac{dx^\ell}{dt} - \frac{\mathcal{E}}{N^2}\sum_\alpha \widehat{f}_\alpha \frac{d}{dt}\partial_k \widehat{f}_\alpha = -\partial_k f.$$

We may assume $\frac{\mathcal{E}}{N^2}$ very small, as motivated by Equation (1); hence, discard this term. Writing the equation into vector form, we have:

$$\frac{dx}{dt} + \mathcal{E}D\frac{dx}{dt} = -\nabla f.$$

By the weak field approximation $(I + \mathcal{E}D)^{-1} \cong (I - \mathcal{E}D)$, we can write:

$$\frac{dx}{dt} = -(I - \mathcal{E}D)\nabla f = -\nabla_D f, \tag{10}$$

where $\nabla_D f$ is the gradient computed according to the diffusion metric Equation (6).

We can summarize our result as follows: *the SGD equation* Equation (1) can be replaced, provided the approximation Equation (9) holds, by the deterministic equation Equation (10), where the dynamical system evolves with respect to the gradient computed according to the diffusion metric Equation (6).

We now want to compare our result Equation (10) with [3] Section 1, in order to understand how the steady state solutions of Equation (10) compare to the SGD steady state solutions described by Equation (1). In [3], the authors regard SGD as minimizing the function Φ instead of our loss f. Let us focus on (8) in [3], where the relation between f and Φ is discussed. In our case, we take $\nabla\Phi = \nabla_D f$ so that Equation (10) becomes

$$-\nabla f(x) + \widetilde{D}(x)\nabla\Phi(x) = 0, \tag{11}$$

where $\widetilde{D}(x) = (I + \mathcal{E}D(x))$ is the diffusion metric. If the term $D(x)$ in (8) in [3] is spelled out as our $\widetilde{D}(x)$, we can write such equation as:

$$j(x) = -\nabla f(x) + \widetilde{D}(x)\nabla\Phi(x) - \beta^{-1}\nabla\cdot\widetilde{D}(x). \tag{12}$$

Notice that according to our approximation Equation (9), $\nabla\cdot\widetilde{D}(x) = 0$. Hence, Equation (12) (that is (8) in [3]) is perfectly compatible with our treatment, and furthermore, the assumption 4 in [3] is fully justified by the fact that $j(x) = 0$.

4. Conclusions

The general relativity (GR) model helps to provide a deterministic approach the evolution of the dynamical system described by SGD, through the use of the diffusion metric accounting for the anisotropy of the system. Our results are compatible with [3]; moreover, they give GD for an isotropic system. We plan to explore deep neural networks mixing our GR model with energy models in a forthcoming paper (see also [9]). This will provide new geometric insight to the theory.

Author Contributions: Conceptualization, R.F. and P.C.; Investigation, R.F. and P.C.; Methodology, R.F.; Supervision, S.S.; Writing and original draft, R.F.; Writing and review editing, P.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by MSCA EU project GHAIA grant number 777822.

Acknowledgments: The authors wish to thank A. Achille and F. Faglioni for many illuminating discussions.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Riemannian Geometry

We collected a few well known facts of Riemannian geometry, and invite the reader to consult [10] for more details.

In Riemannian geometry, we define a metric g on a smooth manifold M, as a smooth assignment $p \mapsto g_p$, which gives, for each $p \in M$, a (non degenerate) scalar product on $T_p(M)$, the tangent space of M at p. Usually, this scalar product is assumed to be a positive definite; however, for general relativity, it is necessary to drop this assumption, so that we speak of a *pseudometric*, because our main example is the Minkowski metric. To ease the terminology, we say "metric" to include this more general setting as well.

Once a metric is given, we say that M is a *Riemannian manifold*. In local coordinates, $x^1, \dots, x^n$, we express the metric using 1-forms:

$$g = \sum_{i,j} g_{ij}\, \mathrm{d}x^i \otimes \mathrm{d}x^j.$$

where

$$g_{ij}|_p := g_p\left(\left.\frac{\partial}{\partial x^i}\right|_p, \left.\frac{\partial}{\partial x^j}\right|_p\right) \qquad \text{and} \qquad \left\{\frac{\partial}{\partial x^1}|_p, \dots, \frac{\partial}{\partial x^n}|_p\right\}$$

is a basis of the tangent space T_pM.

For example $\mathbb{R}^n$, identified with its tangent space at every point, has a canonical or standard metric given by:

$$g^{\mathrm{can}}_p \colon T_p\mathbb{R}^n \times T_p\mathbb{R}^n \longrightarrow \mathbf{R}, \qquad \left(\sum_i a_i \frac{\partial}{\partial x^i}, \sum_j b_j \frac{\partial}{\partial x^j}\right) \longmapsto \sum_i a_i b_i.$$

Here, $g^{\mathrm{can}}_{ij} = \delta_{ij}$.

Usually, we drop the $\sum$ symbol, following Einstein's convention.

An *affine connection* ∇ on a smooth manifold M is an bilinear map $(X, Y) \mapsto \nabla_X Y$ associating with a pair of vector fields X, Y on M another vector field $\nabla_X Y$, satisfying:

1. $\nabla_{fX} Y = f \nabla_X Y$ for all functions f on M;
2. $\nabla_X (fY) = df(X)Y + f\nabla_X Y$.

Once this definition is given, it is possible to define ∇ on tensors of every order.

On a Riemannian manifold we have a unique affine connection, the *Levi–Civita connection* ∇, which is torsion free and preserves the metric; i.e., $\nabla g = 0$. In local coordinates the components of the connection are called the *Christoffel symbols*:

$$\nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} = \Gamma^k_{ij} \frac{\partial}{\partial x^k}.$$

By the above -mentioned uniqueness, the Γ^k_{ij}s are expressed in terms of the metric components g_{ij}:

$$\Gamma^l_{jk} = \tfrac{1}{2} g^{lr} \left(\partial_k g_{rj} + \partial_j g_{rk} - \partial_r g_{jk}\right),$$

where as usual, g^{ij} are the coefficients of the dual metric tensor; i.e., the entries of the inverse of the matrix (g_{kl}). The torsion freeness is equivalent to the symmetry

$$\Gamma^l_{jk} = \Gamma^l_{kj}.$$

In $\mathbb{R}^n$, the *gradient* of a scalar function f is the vector field characterized by the property: $\mathrm{grad}(f) \cdot v = D_v$ (we shall denote it with $\nabla(f)$ so that no confusion arises). In other words, its scalar product (in the euclidean metric) with a tangent vector v gives the directional derivative of f along v. When M is a Riemannian manifold with metric g, the gradient of a function f on M is defined in the same way, except that the scalar product is now given by g. So, in local coordinates, we have:

$$\nabla_g f = \frac{\partial f}{\partial x^i} g^{ij} \frac{\partial}{\partial x^j}.$$

Notice that when $g_{ij} = \delta_{ij}$, we retrieve the usual definition in $\mathbb{R}^n$.

We end our short summary of the key concepts, with the notion of *geodesic*.

A geodesic γ on a smooth manifold M with an affine connection ∇ is a curve defined by the following equation:

$$\nabla_{\dot\gamma} \dot\gamma = 0$$

Geometrically, this expresses the fact that the parallel transport, given by ∇, along the curve γ preserves any tangent vector to the curve. In local coordinates this becomes:

$$\frac{d^2\gamma^\lambda}{dt^2} + \Gamma^\lambda_{\mu\nu} \frac{d\gamma^\mu}{dt} \frac{d\gamma^\nu}{dt} = 0\,.$$

Notice that when the metric is constant, we have the familiar equation:

$$\frac{d^2\gamma^\lambda}{dt^2} = 0;$$

that is, the geodesics are straight lines.

References

1. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature*, **2015**, *521*, 436. [CrossRef] [PubMed]
2. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
3. Chaudhari, P.; Soatto, S. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. *arXiv* **2017**, arXiv:1710.11029.
4. Chaudhari, P.; Soatto, S. On the energy landscape of deep networks. *arXiv* **2015**, arXiv:1511.06485.
5. Chaudhari, P.; Choromanska, A.; Soatto, S.; LeCun, Y.; Baldassi, C.; Borgs, C.; Chayes, J.; Sagun, L.; Zecchina, R. Entropy-SGD: Biasing gradient descent into wide valleys. *arXiv* **2016**, arXiv:1611.01838.
6. Amari, S. Natural Gradient Works Efficiently in Learning. *Neural Comput.* **1998**, *10*, 251–276. [CrossRef]
7. Adler, R.; Bazin, M.; Schiffer, M. *Introduction to General Relativity*; McGraw-Hill: New York, NY, USA, 1965.
8. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. *arXiv* **2016**, arxiv:1608.06993.
9. Achille, A.; Soatto, S. On the emergence of invariance and disentangling in deep representations. *arXiv* **2017**, arXiv:1706.01350.
10. Petersen, P. *Riemannian Geometry*; (GTM); Springer: Cham, Switzerland, 1998.