A note on the natural gradient and its connections with the Riemannian gradient, the mirror descent, and the ordinary gradient

Frank Nielsen

August 2020

Given a real-valued function $L_\theta(\theta)$ (parameterized by a a D-dimensional vector θ) to minimize on parameter space $\theta \in \Theta \subset \mathbb{R}^D$, the *gradient descent* (GD) method (also called the *steepest descent* method) is a first-order local optimization procedure which starts by initializing the parameter to an arbitrary value $\theta_0 \in \Theta$, and then iteratively updates at stage t the current position θ_t to θ_{t+1} as follows:

$$\text{GD}: \quad \theta_{t+1} = \theta_t - \alpha_t \nabla_\theta L_\theta(\theta_t). \tag{1}$$

The scalar $\alpha_t > 0$ is called the *step size* or *learning rate* in machine learning. The ordinary gradient (OG) $\nabla_\theta F_\theta(\theta)$ (vector of partial derivatives) represents the *steepest vector* at θ of the function graph $\mathcal{L}_\theta = \{(\theta, L_\theta(\theta)) \ : \ \theta \in \Theta\}$. The GD method was pioneered by Cauchy [7] (1847) and its convergence proof to a *stationary point* was first reported in Curry [9] (1944).

If we *reparameterize* the function L_θ using a one-to-one and onto differentiable mapping $\eta = \eta(\theta)$ (with reciprocal inverse mapping $\theta = \theta(\eta)$), the GD update rule transforms as:

$$\eta_{t+1} = \eta_t - \alpha_t \nabla_\eta L_\eta(\eta_t), \tag{2}$$

where

$$L_\eta(\eta) := L_\theta(\theta(\eta)). \tag{3}$$

Thus in general, the two gradient descent position sequences $\{\theta_t\}_t$ and $\{\eta_t\}_t$ (initialized at $\theta_0 = \theta(\eta_0)$ and $\eta_0 = \eta(\theta_0)$) are different (because $\eta(\theta) \neq \theta$) and the two GDs may potentially reach different stationary points! In other words, the GD local optimization depends on the choice of the parameterization of the function L (i.e., L_θ or L_η). For example, minimizing with the GD a temperature function $L_\theta(\theta)$ with respect to Celsius degrees θ may yield a different result than minimizing the same temperature function $L_\eta(\eta) = L_\theta(\theta(\eta))$ expressed with respect to Farenheit degrees η. That is, the GD optimization is *extrinsic* since it depends on the choice of the parameterization of the function, and does not take into account the nature of the parameter space Θ.

The natural gradient precisely addresses this problem and solves it by choosing *intrinsically* the steepest direction with respect to a Riemannian metric tensor field on the parameter manifold.

1 Natural gradient: Connection with the Riemannian gradient

Let (M, g) be a D-dimensional Riemannian space [10] equipped with a metric tensor g, and $L \in C^\infty(M)$ a smooth function to minimize on the manifold M. The *Riemannian gradient* [4] uses the

Riemannian *exponential map* $\exp_p : T_p \to M$ to update the sequence of points p_t's on the manifold as follows:

$$\text{RG}: \quad p_{t+1} = \exp_{p_t}(-\alpha_t \nabla_M L(p_t)), \tag{4}$$

where the Riemannian gradient ∇_M is defined according to a *directional derivative* ∇_v by:

$$\nabla_M L(p) := \nabla_v \left(L\left(\exp_p(v)\right)\right)\big|_{v=0}, \tag{5}$$

with

$$\nabla_v L(p) := \lim_{h\to 0} \frac{L(p+hv) - L(p)}{h}. \tag{6}$$

However, the Riemannian exponential mapping $\exp_p(\cdot)$ is often computationally intractable since it requires to solve a system of second-order differential equations [10, 1]. Thus instead of using $\exp_p$, we shall rather use a computable *Euclidean retraction* $R : T_p \to \mathbb{R}^D$ of the exponential map expressed in a local θ-coordinate system:

$$\text{RetG}: \quad \theta_{t+1} = R_{\theta_t}\left(-\alpha_t \nabla_\theta L_\theta(\theta_t)\right). \tag{7}$$

Using the retraction [1] $R_p(v) = p + v$ which corresponds to a first-order Taylor approximation of the exponential map, we recover the *natural gradient descent* [2]:

$$\text{NG}: \theta_{t+1} = \theta_t - \alpha_t g_\theta^{-1}(\theta_t) \nabla_\theta L_\theta(\theta_t). \tag{8}$$

The *natural gradient* [2] (NG)

$$^{\text{NG}}\nabla L_\theta(\theta) := g_\theta^{-1}(\theta) \nabla_\theta L_\theta(\theta) \tag{9}$$

is the *Riemannian steepest descent*, and the natural gradient descent yields the following update rule

$$\text{NG}: \theta_{t+1} = \theta_t - \alpha_t \, ^{\text{NG}}\nabla L_\theta(\theta_t). \tag{10}$$

Notice that the natural gradient is a *contravariant vector*[1] while the ordinary gradient is a *covariant vector*. A covariant vector $[v_i]$ is transformed into a contravariant vector $[v^i]$ by $v^i = \sum_j g^{ij} v_i$, that is by using the dual Riemannian metric $g^*_\eta(\eta) = g_\theta(\theta)^{-1}$, see [13]. The natural gradient is *invariant* under an invertible smooth change of parameterization. However, the natural gradient *descent* does not guarantee that the positions θ_t's always stay on the manifold: Indeed, it may happen that for some t, $\theta_t \notin \Theta$ when $\Theta \neq \mathbb{R}^D$.

Property 1 ([4]) *The natural gradient descent approximates the intrinsic Riemannian gradient descent.*

Let us emphasize that the natural gradient descent is not intrinsic because of the step sizes α_t.

Next, we shall explain how the natural gradient descent is related to the *mirror descent* and the *ordinary gradient* when the Riemannian space Θ is dually flat.

[1]Recall that the *inner product* between two vectors u and v in a tangent plane T_p for $p \in M$ is expressed equivalently as $\langle u, v\rangle_p = g_p(u,v) = \sum_{i=1}^D u^i v_i = \sum_{i=1}^D u_i v^i = \sum_{i,j} g_{ij} u^i v^j = \sum_{i,j} g^{ij} u_i v_j$, where $[w^i]$ and $[w_i]$ denote the contravariant and covariant components of a vector w, respectively. The metric tensor $g^* = g^{ij}$ is called the *dual Riemannian metric*. In a local coordinate chart θ, we have $[g_{ij}][g^{ij}] = I$, where $g = [g(e_i, e_j)]$ with $\{e_1, \ldots, e_D\}$ the natural basis of the vector space T_p.

2 Natural gradient in dually flat spaces: Connections to mirror descent and ordinary gradient

A dually flat space (M, g, ∇, ∇^*) is a manifold M equipped with a pair (∇, ∇^*) of dual torsion-free flat connections which are coupled to the Riemannian metric tensor g [3, 13, 14] in the sense that $\frac{\nabla+\nabla^*}{2} = {}^{LC}\nabla$, where ${}^{LC}\nabla$ denotes the unique metric torsion-free Levi-Civita connection (see the fundamental theorem of Riemannian geometry [13]).

On a dually flat space, there exists a pair of dual global *Hessian structures* [17] with dual canonical Bregman divergences [5, 3]. The dual Riemannian metrics can be expressed as the Hessians of dual convex potential functions. Examples of Hessian manifolds are the *manifolds of exponential families* or the *manifolds of mixture families* [15]. On a dually flat space induced by a strictly convex and C^3 function F (Bregman generator), we have two dual global coordinate system: $\theta(\eta) = \nabla F^*(\eta)$ and $\eta(\theta) = \nabla F(\theta)$, where F^* denotes the Legendre-Fenchel convex conjugate function [11, 12]. The Hessian metric expressed in the primal θ-coordinate system is $g_\theta(\theta) = \nabla^2 F(\theta)$, and the dual Hessian metric expressed in the dual coordinate system is $g^*_\eta(\eta) = \nabla^2 F^*(\eta)$. Crouzeix's identity [8, 13] shows that $g_\theta(\theta) g_\eta(\eta) = I$, where I denotes the $D \times D$ matrix identity.

2.1 Natural gradient: Connection with Bregman mirror descent methods

The ordinary gradient descent method can be extended using a *proximity function* $\Phi(\cdot,\cdot)$ as follows:

$$\text{PGD}: \quad \theta_{t+1} = \arg\min_{\theta\in\Theta} \left\{ \langle \theta, \nabla L_\theta(\theta_t)\rangle + \frac{1}{\alpha_t}\Phi(\theta,\theta_t) \right\}. \tag{11}$$

When $\Phi(\theta,\theta_t) = \frac{1}{2}\|\theta-\theta_t\|^2$, the PGD update rule becomes the GD update rule.

Consider a Bregman divergence [5] B_F for the proximity function Φ: $\Phi(p,q) = B_F(p:q)$. Then the PGD yields the following *mirror descent* (MD):

$$\text{MD}: \quad \theta_{t+1} = \arg\min_{\theta\in\Theta} \left\{ \langle \theta, \nabla L(\theta_t)\rangle + \frac{1}{\alpha_t} B_F(\theta:\theta_t) \right\}. \tag{12}$$

This mirror descent can be interpreted as a natural gradient descent as follows:

Property 2 ([16]) *Mirror descent on the Hessian manifold (M, g) is equivalent to natural gradient descent on the dual Hessian manifold (M, g^*).*

Indeed, the mirror descent rule yields the following natural gradient update rule:

$$\text{NG}^*: \eta_{t+1} = \eta_t - \alpha_t (g^*_\eta)^{-1}(\eta_t)\nabla_\eta L_\theta(\theta(\eta_t)), \tag{13}$$

$$= \eta_t - \alpha_t (g^*_\eta)^{-1}(\eta_t)\nabla_\eta L_\eta(\eta_t), \tag{14}$$

where $g^*_\eta(\eta) = \nabla^2 F^*(\eta) = (\nabla^2_\theta F(\theta))^{-1}$ and $\theta(\eta) = \nabla F^*(\theta)$.

The method is called mirror descent [6] because it performs that gradient step in the *dual space* (mirror space) $H = \{\eta = \nabla F(\theta) \ : \ \theta \in \Theta\}$, and thus solves the inconsistency contravariant/covariant type problem of subtracting a covariant vector from a contravariant vector of the GD (see Eq. 1).

2.2 Natural gradient: Connection with the ordinary gradient descent

Let us prove now the following property of the natural gradient in a dually flat space (or Bregman manifold [14]):

Property 3 ([18]) *In a dually flat space induced by potential convex function F, the natural gradient amounts to the ordinary gradient on the dually parameterized function:* ${}^{\mathrm{NG}}\nabla L_\theta(\theta) = \nabla_\eta L_\eta(\eta)$ *where* $\eta = \nabla_\theta F(\theta)$ *and* $L_\eta(\eta) = L_\theta(\theta(\eta))$.

Proof: Let (M, g, ∇, ∇^*) be a dually flat space. We have $g_\theta(\theta) = \nabla^2 F(\theta) = \nabla_\theta \nabla_\theta F(\theta) = \nabla_\theta \eta$ since $\eta = \nabla_\theta F(\theta)$. The function to minimize can be written either as $L_\theta(\theta) = L_\theta(\theta(\eta))$ or as $L_\eta(\eta) = L_\eta(\eta(\theta))$. Recall the chain rule in the calculus of differentiation:

$$\nabla_\theta L_\theta(\theta) = \nabla_\theta (L_\eta(\eta(\theta))) = (\nabla_\theta \eta)(\nabla_\eta L_\eta(\eta)). \tag{15}$$

We have:

$$\begin{aligned} {}^{\mathrm{NG}}\nabla L_\theta(\theta) &:= g_\theta^{-1}(\theta) \nabla_\theta L_\theta(\theta), && (16)\\ &= (\nabla_\theta \eta)^{-1} (\nabla_\theta \eta) \nabla_\eta L_\eta(\eta), && (17)\\ &= \nabla_\eta L_\eta(\eta). && (18) \end{aligned}$$

□

Thus the natural gradient descent on a loss function $L_\theta(\theta)$ amounts to an ordinary gradient descent on the *dually parameterized* loss function $L_\eta(\eta) := L_\theta(\theta(\eta))$. In short, ${}^{\mathrm{NG}}\nabla_\theta L_\theta = \nabla_\eta L_\eta$.

References

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[2] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

[3] Shun-ichi Amari. *Information Geometry and Its Applications*. Applied Mathematical Sciences. Springer Japan, 2016.

[4] Silvere Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

[5] Lev M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[6] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[7] AL Cauchy. Methode générale pour la résolution des systèmes d'équations simultanées. *Comptes Rendus de l'Académie des Sciences*, 25:536–538, 1847.

[8] Jean-Pierre Crouzeix. A relationship between the second derivatives of a convex function and of its conjugate. *Mathematical Programming*, 13(1):364–365, 1977.

[9] Haskell B Curry. The method of steepest descent for non-linear minimization problems. *Quarterly of Applied Mathematics*, 2(3):258–261, 1944.

[10] Manfredo P Do Carmo. *Differential geometry of curves and surfaces: revised and updated second edition*. Courier Dover Publications, 2016.

[11] Frank Nielsen. Legendre transformation and information geometry, 2010.

[12] Frank Nielsen. Cramér-rao lower bound and information geometry. In *Connected at Infinity II*, pages 18–37. Springer, 2013.

[13] Frank Nielsen. An elementary introduction to information geometry. Technical report, 2018.

[14] Frank Nielsen. On geodesic triangles with right angles in a dually flat space. *arXiv preprint arXiv:1910.03935*, 2019.

[15] Frank Nielsen and Gaëtan Hadjeres. Monte Carlo information-geometric structures. In *Geometric Structures of Information*, pages 69–103. Springer, 2019.

[16] Garvesh Raskutti and Sayan Mukherjee. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015.

[17] Hirohiko Shima. *The geometry of Hessian structures*. World Scientific, 2007.

[18] Guodong Zhang, Shengyang Sun, David Duvenaud, and Roger Grosse. Noisy natural gradient as variational inference. In *International Conference on Machine Learning*, pages 5852–5861, 2018.