Lightlike Pseudo-Riemannian Adaptive Gradient Method

Mastane Achab*
Deep Gambit Limited, Masdar City, Abu Dhabi, UAE
www.deepgambit.com

April 26, 2025

Abstract

We consider the problem of minimizing a function in a pseudo-Riemannian space. We show that under a lightlike constraint, the steepest descent produces an adaptive gradient method.

Notations The Euclidean norm of any vector $x \in \mathbb{R}^m$ ($m \geqslant 1$) is denoted $\|x\|$. Given a symmetric positive definite matrix $\mathsf{W} \in \mathbb{R}^{m \times m}$, we define the weighted norm $\| \cdot \|_{\mathsf{W}}$ as

$$\|x\|_{\mathsf{W}} = \sqrt{x^T \mathsf{W} x}\,. \tag{1}$$

1 Introduction

The purpose of this paper is to show that a certain pseudo-Riemannian minimization problem leads to an adaptive gradient method. Let us first briefly recall a few related notions of Riemannian optimization (Bonnabel [2013],Zhang and Sra [2016]) and of adaptivity.

Riemannian optimization. In a Riemannian space $(\mathbb{R}^m, \mathsf{A})$ with A a positive definite metric tensor, it was shown by Amari [1998] that the steepest descent direction of some function $f(x)$ is given by the negative natural gradient $-\mathsf{A}(x)^{-1}\nabla f(x)$, which is also known as Riemannian gradient descent in more general Riemannian manifolds (see e.g. Boumal [2023]). In the Euclidean scenario $\mathsf{A} \equiv \mathrm{I}_m$, it boils down to gradient descent.

*mastane@deepgambit.com

Adaptive gradient methods. The idea of adaptivity in gradient-based optimization of a vector parameter x is to use, at every iteration t, a separate learning rate $\lambda_{t,i}$ for each coordinate x_i. In popular approaches such as AdaGrad (Duchi et al. [2011]) and RMSprop (Tieleman [2012]), this is generally achieved by setting $\lambda_{t,i} = \lambda/\sqrt{G_{t,i}}$ where $G_{t,i}$ is a scalar summary of the history of squared i-th partial derivatives up to time t.

As in Bécigneul and Ganea [2018], this work proposes to combine both topics under a block-diagonal metric assumption.

2 Steepest lightlike descent

Pseudo-Riemannian setting. We consider the problem of minimizing a differentiable real-valued function f defined over the pseudo-Riemannian space $\mathbb{R}^{m+n}$ (with $m, n \geqslant 1$) equipped with the block-diagonal metric tensor given by

$$\mathsf{M}(z) = \begin{pmatrix} \mathsf{A}(z) & 0_{m\times n} \\ 0_{n\times m} & -\mathsf{B}(z) \end{pmatrix} \tag{2}$$

where for all $z = (x, y) \in \mathbb{R}^m \times \mathbb{R}^n$, both $\mathsf{A}(z)$ and $\mathsf{B}(z)$ are symmetric positive definite matrices. Clearly, M has the metric signature $(m, n, 0)$.

Example 1 (Spacetime). Flat Minkowski spacetime; Lorentzian geometry (see O'neill [1983],Bär [2004]).

Example 2 (Euclidean-Information). See Achab [2024] for an example of an optimization problem under the product metric composed of the Euclidean and Fisher-Rao metrics (Rao [2009]).

Radiant descent. Let us start by introducing a vector we call the "radiant", which will play in our pseudo-Riemannian framework the same role that the standard gradient plays in Euclidean steepest descent.

Assumption 1 (Regular point). A point $z = (x, y) \in \mathbb{R}^m \times \mathbb{R}^n$ is said regular if

$$\frac{\partial f}{\partial x}(z) = \left(\frac{\partial f}{\partial x_i}(z)\right)_{1\leqslant i\leqslant m} \neq 0_m \quad \text{and} \quad \frac{\partial f}{\partial y}(z) = \left(\frac{\partial f}{\partial y_j}(z)\right)_{1\leqslant j\leqslant n} \neq 0_n\,.$$

Definition 1 (Radiant vector). Under Assumption 1, we define the radiant of f at z as:

$$\mathbb{X} f(z) = \begin{pmatrix} \|\frac{\partial f}{\partial x}(z)\|^{-1}_{\mathsf{A}(z)^{-1}} \mathsf{A}(z)^{-1} \frac{\partial f}{\partial x}(z) \\ \|\frac{\partial f}{\partial y}(z)\|^{-1}_{\mathsf{B}(z)^{-1}} \mathsf{B}(z)^{-1} \frac{\partial f}{\partial y}(z) \end{pmatrix} \in \mathbb{R}^{m+n}.$$

By analogy with Minkowski spacetime, we say that a vector $v \in \mathbb{R}^{m+n}$ is lightlike[1] at z if

$$\boxed{v^\intercal \mathsf{M}(z) v = 0}. \tag{Lightlike}$$

[1] Recall that in Minkowski spacetime, a 4D vector $v = (x, y, z, ct)$ is said null or lightlike if $v^\intercal \mathrm{Diag}(1, 1, 1, -1)v = x^2 + y^2 + z^2 - c^2t^2 = 0$.

Proposition 2. The radiant vector $\text{⧗}f(z)$ is lightlike at z.

Proof. It holds:

$$\text{⧗}f(z)^\intercal \mathsf{M}(z)\text{⧗}f(z) = \frac{\|\frac{\partial f}{\partial x}(z)\|^2_{\mathsf{A}(z)^{-1}}}{\|\frac{\partial f}{\partial x}(z)\|^2_{\mathsf{A}(z)^{-1}}} - \frac{\|\frac{\partial f}{\partial y}(z)\|^2_{\mathsf{B}(z)^{-1}}}{\|\frac{\partial f}{\partial y}(z)\|^2_{\mathsf{B}(z)^{-1}}} = 0.$$

□

We are now ready to state our main result.

Theorem 3. Under Assumption 1, the steepest Lightlike descent direction of f at z is given by the negative radiant vector $-\text{⧗}f(z)$.

Proof. Let us follow the derivation of the natural gradient from Amari [1998]. Given $z = (x, y)$ and infinitesimal $\epsilon > 0$, we search for $v = (a, b)$ that minimizes $f(z + \epsilon v) \approx f(z) + \epsilon \nabla f(z)^\intercal v$ under the constraints:

$$\begin{cases} a^\intercal \mathsf{A}(z)a = 1 \\ b^\intercal \mathsf{B}(z)b = 1 \end{cases}, \tag{3}$$

ensuring that v satisfies the Lightlike condition

$$v^\intercal \mathsf{M}(z)v = 0.$$

By the Lagrangian method, we have

$$\begin{cases} \frac{\partial}{\partial a}\{\epsilon \nabla f(z)^\intercal v + \mu_1 a^\intercal \mathsf{A}(z)a\} = \mathsf{0}_m \\ \frac{\partial}{\partial b}\{\epsilon \nabla f(z)^\intercal v + \mu_2 b^\intercal \mathsf{B}(z)b\} = \mathsf{0}_n \end{cases} \Leftrightarrow \begin{cases} a = -\frac{\epsilon}{2\mu_1}\mathsf{A}(z)^{-1}\frac{\partial f}{\partial x}(z) \\ b = -\frac{\epsilon}{2\mu_2}\mathsf{B}(z)^{-1}\frac{\partial f}{\partial y}(z) \end{cases}. \tag{4}$$

Finally, we deduce by combining Eqs. 3-4 that

$$\frac{\epsilon^2\|\frac{\partial f}{\partial x}(z)\|^2_{\mathsf{A}(z)^{-1}}}{4\mu_1^2} = \frac{\epsilon^2\|\frac{\partial f}{\partial y}(z)\|^2_{\mathsf{B}(z)^{-1}}}{4\mu_2^2} = 1, \tag{5}$$

which concludes the proof since

$$v = -\text{⧗}f(z).$$

□

Theorem 3 motivates the following iterative pseudo-Riemannian minimization algorithm.

Definition 4 (Radiant Descent). For step-size $\lambda > 0$, time step $t \in \mathbb{N}$, we define the radiant descent iteration as

$$\boxed{z_{t+1} = z_t - \lambda \text{⧗}f(z_t)}. \tag{RD}$$

The radiant descent (RD) algorithm borrows ideas from Riemannian optimization, by computing the Riemannian gradient in each of the two groups of coordinates x and y, and from the concept of adaptivity by rescaling the effective learning rates for x and y in terms of the weighted norms of the partial derivatives of the objective function f.

Remark 1 (Adaptive Stochastic RD). A natural way to turn RD into a stochastic algorithm is to replace the squared weighted norms of the full batch gradients with some adaptive moving averages, akin to RMSprop.

References

Mastane Achab. A bregman firmly nonexpansive proximal operator for baryconvex optimization. arXiv preprint arXiv:2411.00928, 2024.

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

Christian Bär. Lorentzian geometry. Lecture Notes, Summer Term, 2004.

Gary Bécigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. arXiv preprint arXiv:1810.00760, 2018.

Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. IEEE Transactions on Automatic Control, 58(9):2217–2229, 2013.

Nicolas Boumal. An introduction to optimization on smooth manifolds. Cambridge University Press, 2023.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. Journal of machine learning research, 12(7), 2011.

Tingran Gao, Lek-Heng Lim, and Ke Ye. Semi-riemannian manifold optimization. arXiv preprint arXiv:1812.07643, 2018.

Barrett O'neill. Semi-Riemannian geometry with applications to relativity, volume 103. Academic press, 1983.

C. Radhakrishna Rao. Fisher-Rao metric. Scholarpedia, 4(2):7085, 2009. doi: 10.4249/scholarpedia.7085. revision #91265.

Sebastian Ruder. An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747, 2016.

Tijmen Tieleman. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning, 4(2):26, 2012.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In Conference on learning theory, pages 1617–1638. PMLR, 2016.