Natural gradient descent with momentum

Anthony NOUY, Laboratoire de Mathématiques Jean Leray - Nantes
Agustin SOMACAL, Laboratoire de Mathématiques Jean Leray - Nantes

Natural gradient descent [1] can be seen as a preconditioned update where parameter changes are properly decorrelated from a functional perspective [8] (whitening transform). In other words, in a spirit similar to a second order (or Newton's) method, the Natural Gradient update uses the Gram matrix instead of the Hessian, defined as the metric of the approximation manifold (for example the manifold of neural networks) at the current iteration [3].
Although the assemblage and inversion of the Gram matrix is prohibitively expensive in the context of big machine learning models, it becomes not only feasible but necessary when we look at scientific machine learning problems [7]. For example when searching for the solution of a parametric partial differential equation (pPDE) by solving the forward (given parameters) or inverse (given measurements) problems using Physics Informed Neural Networks (PINNs) traditional ubiquitous solvers like Adam or L-BFGS do not yield reliable solutions or they take too long to converge [4]. However, taking the natural gradient perspective allows us to reinterpret gradient descent as a projection of the functional gradient into the tangent space of the approximation manifold [2] obtaining great improvements.
That being said, both gradient and natural gradient descent will still get stuck at any local minima. Furthermore, when the loss function is other than the euclidean distance (for example when minimising the residual of a pPDE as in PINNs) even the natural gradient might yield non-optimal directions at each step. The talk will focus on how we can tackle these situations by introducing a Natural version of classical inertial dynamic methods like Nestorov [5] or heavy-ball [6].

[1] S.-i. Amari. *Natural Gradient Works Efficiently in Learning.* **10(2)**, 251–276. doi : 10.1162/089976698300017746.

[2] R. Gruhlke, A. Nouy, P. Trunschke. *Optimal sampling for stochastic and natural gradient descent.*

[3] J. Martens. *New insights and perspectives on the natural gradient method.*

[4] J. Müller, M. Zeinhofer. *Achieving High Accuracy with PINNs via Energy Natural Gradients.* doi :10.48550/arXiv.2302.13163.

[5] Y. Nesterov. *A method for solving the convex programming problem with convergence rate O(1/k^2).* **269**, 543–547.

[6] B. Polyak. *Some methods of speeding up the convergence of iteration methods.* **4(5)**, 1–17. doi : 10.1016/0041-5553(64)90137-5.

[7] N. Schwencke, C. Furtlehner. *ANaGRAM : A Natural Gradient Relative to Adapted Model for efficient PINNs learning.* doi :10.48550/arXiv.2412.10782.

[8] J. Sohl-Dickstein. *The Natural Gradient by Analogy to Signal Whitening, and Recipes and Tricks for its Use.* doi :10.48550/arXiv.1205.1828.

Contact : agustin.somacal@ec-nantes.fr