A Comparison of First- and Second-Order Optimization Methods in Machine Learning and Statistical Inference

Authors and Affiliations

Achraf Cherkaoui, Evrim Oral
LSUHSC BSDS Program

Abstract

Optimization lies at the core of modern statistical inference and machine learning, governing parameter estimation in likelihood-based models, M-estimators, and high-dimensional predictive systems. Within the framework of Empirical Risk Minimization (ERM), estimation is formulated as the minimization of an empirical objective function, encompassing maximum likelihood estimation, penalized likelihood, and general M-estimation. The evolution of optimization methods for these problems has been driven by two principal paradigms: first-order methods, typified by Gradient Descent (GD), and second-order methods, most notably the Newton–Raphson (NR) algorithm.

The essential distinction between these approaches stems from the order of local information exploited via Taylor series expansions of the objective function. Gradient Descent relies solely on first-order derivative information, corresponding to a local linear approximation. In contrast, Newton–Raphson incorporates the Hessian matrix of second derivatives, yielding a local quadratic approximation. In likelihood-based settings, the Hessian is closely related to the observed Fisher information, providing curvature information that reflects statistical efficiency. This difference in local modeling leads to fundamentally distinct convergence behavior: under standard regularity conditions and in a neighborhood of a nondegenerate minimizer, GD exhibits linear convergence, whereas NR achieves quadratic convergence. Moreover, Newton's method is affine invariant under nonsingular linear reparameterizations, while Gradient Descent is sensitive to parameter scaling and conditioning—properties that directly affect statistical stability and inference.

Despite its asymptotic efficiency and favorable local convergence properties, classical Newton-type optimization requires $O(p^2)$memory and $O(p^3)$computational complexity in p-dimensional problems, limiting its feasibility in high-dimensional statistical models. First-order methods, requiring only gradient evaluations, scale linearly in parameter dimension and sample size per iteration, rendering them more suitable for large-scale estimation problems. This study presents a rigorous derivation of these structural differences, elucidating the trade-offs between curvature exploitation, scalability, and convergence efficiency in contemporary machine learning.