Overview of Second-Order Stochastic Methods for Constrained Nonlinear Optimization

Kenneth Wu

December 16, 2024

Abstract

In this report, we give a high-level overview of some academic papers concerning two related, but distinct, approaches to making effective use of stochastic gradient in second-order methods for continuous optimization problems. The first method is an extension of sequential quadratic programming, while the latter is an extension of interior point methods. When substituting out gradients for stochastic gradients, some of the quantities that the standard version of these two methods rely on become unreliable. Both methods circumvent this by iteratively refining an estimate of a Lipschitz constant, which is used to perform alternative routines to the standard ones that are compromised under stochasticity.

Introduction

When formulating optimization problems with some application in mind, there may be desirable features that we'd like our optimal solution to have (or perhaps features we'd ideally like to avoid). Often we are content to merely encourage the desired behavior by employing a penalty term, whether it be because our requirements are lenient, or because the corresponding hard constraint function would be unpleasant to work with.

Nonetheless, there are cases where we simply cannot budge; perhaps there is a legal requirement in the scenario that must be met, or it may be a matter of safety. To accommodate these situations, it's important to have optimization methods that can handle hard constraints. Among such methods, second-order methods that use not just the gradient, but the Hessian (or at least an approximation) are quite popular, particularly for problems of moderate size. Second-order methods often converge in fewer steps and can better handle ill-conditioned problems. However, these methods have seen less widespread adoption among large-scale problems like those coming from machine learning where their ability to get solutions of high precision is less relevant, and where the complexity cost per iteration is intractable due to the size of the problem.

While some readers may express disappointment at the the relative lack of mathematical formulae or formal theorems or explicit pseudocode compared to the original papers, we stress that our aim is more so to build intrigue than to be exhaustive. We want to avoid getting boggled down in the minute details, at least at this stage; our hope is that by the end of this report, the reader will be in a better position to read the original papers, or at least be more motivated to learn the background material pertaining to these methods if they only have a passing familiarity as of yet.

Background

We denote the set of real numbers by $\mathbb{R}$ and the set of nonnegative numbers by $\mathbb{R}_{\geq 0}$. We use $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$ to denote the set of n-dimensional real vectors and the set of m-by-n-dimensional real matrices respectively. In this report, we will concern ourselves with optimization problems of the form:

$$\begin{array}{lll} \text{minimize} & f(x) & \\ \text{subject to} & g_i(x) \leq 0, & \forall i \in \{1, 2, \ldots, k_{\text{ineq}}\} \\ & h_i(x) = 0, & \forall i \in \{1, 2, \ldots, k_{\text{eq}}\} \end{array}$$

While the assumptions on f, g, and h can vary, we will usually require them to at least be continuously differentiable at the very least. We define the 0–∞ indicator function of a set $Q \subseteq \mathbb{R}^n$ to be

$$\delta_Q(x) = \begin{cases} 0 & \text{if } x \in Q, \\ \infty & \text{otherwise.} \end{cases}$$

In our case, we are primarily interested in the 0–∞ indicatior function of the feasible set $D \subseteq \mathbb{R}^n$ described by the inequality and equality constraints. Let ∂Q denote the boundary of a set $Q \subseteq \mathbb{R}^n$. We say that a continuous function $\psi \colon \mathbb{R}^n \to \mathbb{R}$ is a barrier function for D if and only if $\lim_{k\to\infty} \psi(x_k) = \infty$ for every sequence $\{x_k\} \subseteq D$ such that $\lim_{k\to\infty} x_k \in \partial D$ and $\psi(x) = \infty$ for $x \notin D$. The motivation behind barrier functions is that they provide a continuous approximation to δ_D. By adding a scaled barrier function to the original objective function, we can approximate our constrained optimization problem by an unconstrained one:

$$\min_{x \in D} f(x) = \min\{f(x) + \delta_D(x) \mid x \in \mathbb{R}^n\} = \min\left\{f(x) + \lim_{t \searrow 0} \big(t\psi(x)\big) \mid x \in \mathbb{R}^n\right\}$$

(The diagonal arrow here is to indicate that we only care about the behavior of positive t.) Smaller values of the barrier parameter t provide a more accurate approximation, but can be problematic numerically due to the steep slope along the boundary. For this reason, methods that utilize barrier functions will iteratively decrease the barrier parameter and try to drive it toward zero.

If $d \in \mathbb{R}^n$ satisfies $\langle -\nabla f(x), d \rangle$ for some $x \in D$, we say that d is a descent direction at x. The idea behind this definition is that, as a consequence of Taylor's theorem, we are guaranteed a decrease in function value by starting at x and traveling a sufficiently small step size $\alpha \in \mathbb{R}_{>0}$ in the direction of d, i.e., $f(x + \alpha d) < f(x)$. The process of finding a descent direction and an appropriately small step size to match said descent direction is called a line search, and can be either performed exactly or inexactly. In an exact line search, we solve a univariate optimization problem to obtain $\tilde{\alpha} \in \arg\min_{\alpha>0} f(x + \alpha d)$.

In practice, the reduction in objective function value may not be enough to justify the cost of solving this univeriate minimization exactly, and so we settle for any step size that achieves a "reasonably" good reduction in objective function value. (For more details on the precise sense in which this is meant, look up "sufficient decrease" in [5].)

We say that a sequence of random variables $\{X_k\} \subseteq \mathbb{R}^n$ converges to a random variable almost surely if and only if $\mathbb{P}\big((\lim_{k\to\infty} X_k) = X\big) = 1$. This can be thought of as essentially being pointwise convergence, modulo a set of measure zero.

Literature Review

Sequential quadratic programming (commonly abbreviated as SQP) and interior-point methods and are two of the best-known methods for solving constrained continuous optimization problems with nonlinear constraint and objective functions. The former deals with inequality constraints by iteratively makes educated guesses of which constraints will be active at an optimal solution (i.e., $g_i(x^*) = 0$ or $h_i(x^*) = 0$), then linearizes the problem to get a quadratic subproblem. The latter deals with inequality constraints by using barrier functions and has unconstrained optimization subproblems (which amounts to solving a nonlinear system of equations). Both methods have seen widespread adoption among both commerical and open-source solvers, and have substantial convergence and complexity guarantees to back their effectiveness. However, both rely on using gradients and Hessians (approximate or otherwise), which can make them prohibitively expensive for solving large-scale problems.

One relatively underexplored avenue for addressing this concern is to replace the gradient by a stochastic gradient estimate, which is motivated in part by the success of stochastic gradient methods. Replicating their success in the constrained setting with second-order methods is unfortunately not so straightfoward, as existing parts of the interior-point and SQP framework may not behave properly when gradients are swapped out for stochastic gradients.

In a paper written in 2020 by Albert Berahas, Frank E. Curtis, Daniel P. Robinson, and Baoyu Zhou [1], the authors propose a stochastic SQP algorithm targeted at optimization problems of a particular structure, where it is assumed that the objective function and its gradient are only accessible via stochastic estimates. The objective function should be given by the expectation of a continuously differentiable function of both the decision variables and a random variable of unknown distribution. On the other hand, the nonlinear equality constraints should be deterministic. (Inequality constraints are not yet considered in the paper.)

Since a line search would not be tractable in the stochastic gradient setting, the authors instead utilize a step size scheme based on adaptively estimated Lipschitz constants. This scheme is first introduced without the use of stochastic gradients, where it is shown that an SQP algorithm based on their proposed step size scheme has convergence guarantees that put it on equal footing with that of an SQP algorithm using a standard backtracking line search. The authors then proceed to employ their step size scheme in tandem with stochastic gradients. They show that the resulting algorithm enjoys similar convergence guarantees to those of its non-stochastic counterpart, albeit in expectation. Said analysis relies on the merit function parameter not behaving "poorly," though the authors give evidence that such poor behavior either occurs with probability zero or only under extreme circumstances.

This work was later extended in [2], where for a simplified version of the algorithm presented in [1], the authors give almost-sure convergence guarantees for the primal iterates, Lagrange multipliers, and stationary measures (as opposed to in expectation).

In 2024, the same core research group that developed the aforementioned stochastic SQP method from [1] and [2] proposed an interior point method that incorporates stochastic gradients [4]. In order to streamline the analysis, the authors limit their scope to bound constraints (i.e., each constraint variable is prescribed a lower and upper threshold), though they state that their overall approach was designed with extensions to nonlinear equality and other types of nonlinear inequality constraints in mind.

As with the stochastic SQP method from before, it is assumed that the (potentially nonconvex) objective function and its gradient cannot reasonably be exactly evaluated (only stochastic gradients are avaliable), however here the objective function is not assumed to take the form of an expectation. In a similar fashion to that of [1] and [2], the authors present their approach in the deterministic setting first since their approach already differs from what might be considered standard, then follow up with the stochastic version. They highlight three particular features that distinguish the algorithm at hand from other similar-looking methods in the literature:

- Since stationary measures are no longer reliable due to noise, their algorithm uses a prescribed monotonically nonincreasing and vanishing barrier parameter sequence to determine when to decrease the barrier parameter.

- Instead of a fraction-to-the-boundary rule, they employ a novel "inner neighborhood" strategy for ensuring that iterates do not get too close to the boundary, where said neighborhoods are generated by a monotonically nonincreasing and vanishing sequence.

- They eschew a step acceptance criteria (such as e.g., a line search) in favor of a step size that uses Lipschitz constant estimates of the objective function's gradient (as was the case with [1] and [2]).

Convergence results are provided for both the deterministic and stochastic settings. Additionally, numerical results showed favorable performance, both on a set of well-known test problems, and when pitted against the projected stochastic gradient method.

In a recent development, the authors of [4] made an extension that replaces the usual nested-loop structure characteristic of a typical interior-point with a single loop [3]. This paper additionally allows for affine equality constraints and nonlinear inequality constraints, but does not yet support general nonlinear equality constraints. One drawback of the current method is that it requires the user to supply a feasible initial point that, in particular, is strictly feasible with respect to the nonaffine constraints. While there are algorithms for finding such a strictly feasible point, the authors pose the question of whether it is possible to design a single-loop infeasible interior-point without compromising on the convergence guarantees.

References

[1] Albert Berahas et al. *Sequential Quadratic Optimization for Nonlinear Equality Constrained Stochastic Optimization.* 2020. arXiv: 2007.10525 [math.OC]. URL: https://arxiv.org/abs/2007.10525.

[2] Frank E. Curtis, Xin Jiang, and Qi Wang. *Almost-sure convergence of iterates and multipliers in stochastic sequential quadratic optimization.* 2023. arXiv: 2308.03687 [math.OC]. URL: https://arxiv.org/abs/2308.03687.

[3] Frank E. Curtis, Xin Jiang, and Qi Wang. *Single-Loop Deterministic and Stochastic Interior-Point Algorithms for Nonlinearly Constrained Optimization.* 2024. arXiv: 2408.16186 [math.OC]. URL: https://arxiv.org/abs/2408.16186.

[4] Frank E. Curtis et al. *A Stochastic-Gradient-based Interior-Point Algorithm for Solving Smooth Bound-Constrained Optimization Problems.* 2024. arXiv: 2304.14907 [math.OC]. URL: https://arxiv.org/abs/2304.14907.

[5] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization.* 2nd ed. Springer Series in Operations Research and Financial Engineering. New York: Springer, 2006. ISBN: 9780387303031. DOI: 10.1007/978-0-387-40065-5. URL: http://dx.doi.org/10.1007/978-0-387-40065-5.

Biography

Kenneth Wu was born in Redmond, Washington but spent the majority of his upbringing in Olympia. He first became interested in mathematics while taking dual enrollment classes at South Puget Sound Community College, which is home to several instructors whom he credits with fostering his academic interests. He went on to obtain his Bachelor of Arts in mathematics from the University of Washington, where he served two quarters as an undergraduate teaching assistant for a course on designing and analyzing algorithms aimed at non-CS majors. In fall 2024, Kenneth began a Ph.D program in Industrial and Systems Engineering at Lehigh University. As of the time of writing, he is beginning research related to stochastic set-valued optimization.