Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

RELIABLE ADAPTIVE STOCHASTIC OPTIMIZATION WITH HIGH PROBABILITY GUARANTEES

Miaolan Xie

Operations Research and Information Engineering
Cornell University
136 Hoy Rd
Ithaca, NY 14853, USA

ABSTRACT

To handle real-world data that is noisy, biased and even corrupted, we consider a simple adaptive framework for stochastic optimization where the step size is adaptively adjusted according to the algorithm's progress instead of manual tuning or using a pre-specified sequence. Function value, gradient and possibly Hessian estimates are provided by probabilistic oracles and can be biased and arbitrarily corrupted, capturing multiple settings including expected loss minimization in machine learning, zeroth-order and low-precision optimization. This framework is very general and encompasses stochastic variants of line search, quasi-Newton, cubic regularized Newton and SQP methods for unconstrained and constrained problems. Under reasonable conditions on the oracles, we show high probability bounds on the sample and iteration complexity of the algorithms.

1 USER-FRIENDLY AND ADAPTIVE ALGORITHMS

It is highly desirable for an optimization algorithm to be adaptive and robust to its input parameters. For example, in training machine learning models using optimization algorithms, one crucial step is to tune the sequence of step sizes or learning rates. The step sizes used directly affect the quality of the resulting model, and most algorithms are very sensitive to the input step size or learning rate, so it is important to choose them correctly. However, step size tuning is a rather tedious and manual process for the user.

There have been various attempts to resolve this issue. Algorithms like Adam and AdaGrad, which are widely used in the machine learning community, try to achieve this by estimating the second moment of the gradients. However, these algorithms still require the tuning of an initial step size parameter and can be sensitive to its magnitude. There have also been algorithms leveraging second-order information like Hessians or various approximations of the Hessians to achieve this goal, at the expense of more costly iterations. On the other hand, in deterministic optimization, there exist efficient adaptive algorithms, such as line search, trust region method, and adaptive cubic regularization. However, all these methods typically require exact function and gradient evaluations.

In this work, we adapt some of these algorithms to the setting where both function and gradient estimations are noisy and stochastic. The algorithms are designed to adapt to each specific problem and the initial input step size parameter. Specifically, the explicit or implicit step size of each iteration is adaptively adjusted by comparing the estimated progress of the algorithm to the expected progress that it is supposed to achieve. The estimated progress is obtained using stochastic function estimations. If the estimated progress is less than the expected progress, the step size will be decreased, otherwise, the step size will be increased. These quantities are actually easy to compute and do not require knowledge of any problem-specific constants. Stochastic function estimations are usually cheap to obtain; for instance, in deep learning, they can simply be obtained from a forward pass of the neural network, which is cheaper

than a stochastic gradient estimation (which requires a backward pass). The noisy function estimations serve two purposes: 1. safeguarding against "bad" stochastic gradients, 2. allowing the step sizes to be adaptive to the local landscape of the problem, and the progress of the algorithm. The algorithms are also adaptive in the accuracy requirements of the gradient estimates. We discuss this in the next section.

2 REALISTIC INPUT ASSUMPTIONS

In many real-life problems, the inputs are susceptible to the influence of outliers, noise, low-precision arithmetic, system failure, and even adversarial attacks. Hence, it is desirable to have a unifying framework that can model all these anomalies and consider algorithms that are robust to them. Thus, in this work, the inputs given to the algorithms are only assumed to be "sufficiently accurate" with a certain probability.

Specifically, we assume the inputs for the algorithms are provided by probabilistic oracles, which capture multiple standard settings including expected loss minimization in machine learning, zeroth-order (derivative-free) optimization, and low-precision optimization. In particular, the oracle output can be arbitrarily bad with some probability and otherwise is sufficiently accurate. This accuracy is adjusted adaptively by the algorithms up to some best achievable accuracy. These weak assumptions allow for bias and the possibility of unbounded error. For example, in (Jin, Scheinberg, and Xie 2021), for a continuous function $\phi(x)$, the stochastic gradient estimation $g(x,\xi)$ at iteration x (here ξ indicates the underlying randomness in the gradient estimation) just needs to satisfy:

$$\mathbb{P}_\xi\left(\|g(x,\xi)-\nabla\phi(x)\| \le \max\{M(x),\varepsilon\}\right) \ge 1-\delta,$$

where $M(x)$ is a quantity dependent on the norm of the (estimated) gradient at iterate x, and ε is the best accuracy achievable by the oracle.

3 GOOD THEORETICAL GUARANTEES

Despite the amount of bias and noise from the input, we show that good theoretical guarantees can nonetheless still be obtained by the algorithms, not only in expectation but with overwhelmingly high probability, under some reasonable assumptions on the stochastic function estimations (Jin, Scheinberg, and Xie 2021; Jin, Scheinberg, and Xie 2023; Scheinberg and Xie 2022). In particular, we show that the stopping time T_{stoc} (the number of iterations to reach a specified accuracy) of the stochastic algorithm is of the same order as the stopping time T_{det} of the classical algorithm with exact function and gradient evaluations with overwhelmingly high probability: For any $t \ge C \cdot T_{\mathsf{det}}$,

$$\mathbb{P}\left(T_{\mathsf{stoc}} \le t\right) \ge 1-2\exp\left(-ct\right).$$

It is intriguing to see the guarantees of the stochastic algorithms with realistic assumptions still match those of the classical algorithms with exact function information up to a constant factor C. The size of the constant factor C depends on the amount of noise and bias in the input. When the inputs are exact, the classical guarantees are recovered by the algorithms. Hence, the theoretical guarantees are adaptive to the amount of "noise" in the input. Another point of interest is that in the machine learning setting, all the results can be applied not only to the empirical risk minimization problem in the form of a finite sum but also to the expected risk minimization problem. It is desirable to have results for the expected risk minimization problem since it speaks directly to the generalization quality of the solution.

REFERENCES

Jin, B., K. Scheinberg, and M. Xie. 2021. "High Probability Complexity Bounds for Adaptive Step Search Based on Stochastic Oracles". *arXiv preprint arXiv:2106.06454*.

Jin, B., K. Scheinberg, and M. Xie. 2023. "Sample Complexity Analysis for Adaptive Optimization Algorithms with Stochastic Oracles". *arXiv preprint arXiv:2303.06838*.

Scheinberg, K., and M. Xie. 2022. "Stochastic Adaptive Regularization Method with Cubics: A High Probability Complexity Bound". In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*.