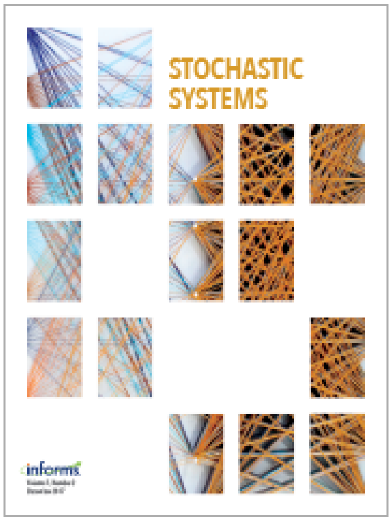

Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Open Problem—Iterative Schemes for Stochastic Optimization: Convergence Statements and Limit Theorems

Afrooz Jalilzadeh, Jinlong Lei, Uday V. Shanbhag

informs
http://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Vol. 9, No. 3, September 2019, pp. 299–302
ISSN 1946-5238 (online)

Special Section: Open Problems in Applied Probability

Open Problem—Iterative Schemes for Stochastic Optimization: Convergence Statements and Limit Theorems

Afrooz Jalilzadeh,[a] Jinlong Lei,[a] Uday V. Shanbhag[a]

[a] The Pennsylvania State University, State College, Pennsylvania 16801

Contact: azj5286@psu.edu (AJ); jxl800@psu.edu (JL); udaybag@psu.edu, https://orcid.org/0000-0001-5869-9561 (UVS)

Published Online in Articles in Advance: September 17, 2019

https://doi.org/10.1287/stsy.2019.0043

History: This paper was accepted for the *Stochastic Systems* Special Section on Open Problems in Applied Probability, presented at the 2018 INFORMS Annual Meeting in Phoenix, Arizona, November 4–7, 2018.

Keywords: stochastic optimization • stochastic approximation • central limit theorems

1. Central Limit Statements for Stochastic Gradient Methods and Their Variants

Central limit theorems represent among the most celebrated of limit theorems in probability theory (Lindeberg 1922, Feller 1945). It may be recalled that the sum of n independent and identically distributed zero mean square integrable random variables grows at the rate of $\sqrt{n}$. Consequently, by dividing this sum by $\sqrt{n}$, we expect its law to possibly approach a limit. This intuition is formalized by the central limit theorem (CLT; Chow and Teicher 2012, chapter 9), and in fact, this limit is the normal distribution. Our interest lies in developing limiting statements for estimators to the following stochastic optimization problem:

$$\min_{x\in\mathbb{R}^n} f(x) \triangleq \mathbb{E}[F(x,\xi)], \tag{SOpt}$$

where $f(x)$ is a strongly convex function on $\mathbb{R}^n$, $F:\mathbb{R}^n\times\mathbb{R}^d \to \mathbb{R}$ is a real-valued function, $\xi:\Omega\to\mathbb{R}^d$, and $(\Omega,\mathcal{F},\mathbb{P})$ denotes the associated probability space. We consider the limiting behavior of a rescaled estimation error $(x_k - x^*)$, where x^* is a solution to a (SOpt) and $\{x_k\}_{k\ge 0}$ is a sequence generated from a stochastic gradient scheme (SA), where $x_0\in\mathbb{R}^n$ and $\{\gamma_k\}$ is a positive step length sequence.

$$x_{k+1} := x_k - \gamma_k \nabla_x F(x_k,\xi_k), \qquad k\ge 0. \tag{SA}$$

We define $\mathcal{F}_k \triangleq \{x_0,\xi_0,\ldots,\xi_{k-1}\}$ and assume there exists an oracle that generates $\nabla_x F(x_k,\xi_k)$ at x_k such that, for any $k\ge 0$, $\mathbb{E}[M_{k+1}\mid\mathcal{F}_k]=0$ and $\mathbb{E}\big[\|M_{k+1}\|^2\mid\mathcal{F}_k\big]\le\nu^2$ *a.s.*, where $M_{k+1}\triangleq\nabla_x f(x_k)-\nabla_x F(x_k,\xi_k)$. Under suitable summability requirements on $\{\gamma_k\}$, differentiability of $f(x)$ (with twice differentiability at x^* and $\mathbf{H}\triangleq\nabla^2 f(x^*)$, $\|\nabla f(x)-\mathbf{H}(x-x^*)\|\le c\|x-x^*\|^{1+\beta}$ (with $c>0,\beta\in(0,1)$) and suitable spectral requirements on $\mathbf{H}$, we present a CLT.

Theorem 1 (CLT for (SOpt); Chen 2002, theorem 3.3.2)**.** *Let algorithm (SA) be applied to the problem (SOpt). Under the aforementioned assumptions, the following holds, where $N(\mu,\mathbf{S})$ denotes the normal distribution with mean μ and covariance $\mathbf{S}$.*

$$\frac{x_k-x^*}{\sqrt{\gamma_k}} \xrightarrow[k\to\infty]{d} N(0,\mathbf{\Sigma}),\ \textit{where}\ \mathbf{\Sigma}=\int_0^\infty e^{\left(\mathbf{H}+\frac{\alpha}{2}\mathbf{I}_m\right)t}\mathbf{S}_0 e^{\left(\mathbf{H}+\frac{\alpha}{2}\mathbf{I}_m\right)t}dt. \tag{1}$$

Theorem 1 is merely an instance of CLT-type results, and stationary distributions of stochastic gradient methods with decreasing/constant step lengths have been investigated (Kushner and Yin 2003, Mandt et al. 2017). In the context of Theorem 1, open questions arise from weakening the assumptions to examining the impact of acceleration and variance reduction schemes that use increasing batch sizes of gradients.

1.1. Weakening Convexity and Smoothness

Can strong convexity and smoothness (away from x^*) requirements be weakened? One avenue may lie in using iteratively regularized and smoothed schemes (Koshal et al. 2013; Yousefian et al. 2013, 2017; Jalilzadeh et al. 2018a) with $\nabla_x f(x_k)$ replaced by $\nabla_x f_{\eta_k}(x_k) + \epsilon_k x_k$, where $f_\eta(x)$ is a smoothing of f (Jalilzadeh et al. 2018a) and ϵ_k denotes the regularization parameter at iteration k. Proceeding further, can smoothness be further weakened to allow for semismoothness (Facchinei and Pang 2007) at x^*?

1.2. Variance Reduction Schemes

Recently, there has been an interest in utilizing variance reduction schemes in which an increasing batch size of gradients, given by $\left(\sum_{j=1}^{N_k} \nabla f(x_k, \xi_{j,k})\right)/N_k$, is employed at each step of (SA). In such regimes, it has been shown that, under strong convexity, if the batch size N_k increases at a geometric rate, the mean-squared error decays at an appropriate geometric rate (Shanbhag and Blanchet 2015). Specifically, if $N_k = \lceil \rho^{-(k+1)} \rceil$, where $\rho \in (0,1)$, it seems possible that, for some $q \in (0,1)$, an appropriate CLT can be proven of the form $q^{-k}(x_k - x^*) \xrightarrow[k\to\infty]{d} N(0, \mathbf{\Sigma})$.

1.3. Accelerated and Quasi-Newton Schemes

Stochastic variants (Mokhtari and Ribeiro 2014, Jalilzadeh et al. 2018a) of quasi-Newton methods have been developed in the past few years with a view toward mitigating the concerns of ill-conditioning, where $\{x_k\}$ is updated according to the following rule:

$$x_{k+1} := x_k - \gamma_k \mathbf{P}_k \nabla_x F(x_k, \xi_k), \qquad k \geq 0, \qquad \text{(stochastic Quasi-Newton (SQN))}$$

where $\mathbf{P}_k$ is an approximation of the Hessian inverse from a modified Broyden-Fletcher-Goldfarb-Shanno (BFGS) update rule. Accelerated gradient methods for convex L−smooth problems achieve a rate of $\mathcal{O}(1/k^2)$ (Nesterov 1983, Beck and Teboulle 2009); an identical rate is achieved via constant step length variance reduced accelerated scheme for (SOpt) in Ghadimi and Lan (2016), Jofré and Thompson (2017), and Jalilzadeh et al. (2018b), while a rate of $\mathcal{O}(1/k)$ is achieved for nonsmooth (but smoothable) problems (Jalilzadeh et al. 2018b). Can CLTs be developed to characterize the limiting behavior of SQN and accelerated schemes?

2. Stochastic Iterative Schemes for Stochastic Nonconvex Optimization

Over the last 15 years, there has been a surge of advances in stochastic first-order schemes, including extragradient methods (Juditsky et al. 2011, Yousefian et al. 2014, Jalilzadeh and Shanbhag 2016), accelerated schemes (Ghadimi and Lan 2016, Jalilzadeh et al. 2018b), variance-reduced schemes (Shanbhag and Blanchet 2015, Jofré and Thompson 2019), deterministic step length schemes for nonconvex programs (Ghadimi and Lan 2016, Ghadimi et al. 2016), and SQN schemes (Lucchi et al. 2015, Zhou et al. 2017, Jalilzadeh et al. 2018a). We review some recent advances in stochastic trust-region and line-search methods.

2.1. Trust-Region Methods

Trust-region (TR) methods represent an important avenue for computing stationary points of nonconvex nonlinear programs. When considering the minimization of $f(x)$ over $\mathbb{R}^n$, given a current iterate x_k, TR methods compute a direction p_k by solving (2).

$$\min_{\|p_k\| \leq \Delta_k} m_k(p_k) \triangleq f(x_k) + \nabla_x f(x_k)^T p_k + \tfrac{1}{2} p_k^T B_k p_k, \tag{2}$$

where Δ_k is the trust-region radius and B_k is generated via a quasi-Newton update rule in first-order regimes or $B_k \triangleq \nabla^2 f(x_k)$ when second derivatives are available. Suppose the ratio between actual reduction $f(x_k) - f(x_{k+1})$ and predicted reduction $m_k - m_{k+1}$ is denoted by $c_k \triangleq \frac{f(x_k) - f(x_{k+1})}{m(x_k) - m(x_{k+1})}$. If $c_k \geq \bar{c}$, for some $\bar{c} \in (0,1)$, then $x_{k+1} = x_k + p_k$, and the trust region is maintained constant or raised if $c_k \approx 1$; otherwise, $x_{k+1} = x_k$, and the trust-region radius Δ_k is reduced by a constant factor. In Chang et al. (2013) and Shashaani et al. (2018), asymptotic convergence is guaranteed by sampling the function values repeatedly to compute estimates of the function (and possibly gradient) values and consequently reduce the error of estimates without any rate result. A fully stochastic variant of the TR method with random models (STORM) proposed in Chen et al. (2018) is developed in Blanchet et al. (2019) covering settings in which the objective and gradient estimates might be biased. Under suitable assumptions, including a requirement that random models m_k are probabilistically "fully linear" and

Hessians H_k are bounded, both a.s. convergence to stationary points and an iteration complexity of $\mathcal{O}(1/\epsilon^2)$ are proven.

2.2. Line-Search Methods

Line-search methods compute a direction p_k based on minimizing $f(x_k) + \nabla_x f(x_k)^T p_k + \frac{1}{2} p_k^T B_k p_k$, where $B_k \succeq 0$ could be generated via a quasi-Newton update or could be the true Hessian $\nabla^2 f(x_k)$. Then α_k is computed by conducting a search satisfying the Armijo condition $f(x_k + \alpha_k p_k) \le f(x_k) + \sigma \alpha_k \nabla_x f(x_k)^T p_k$, where $\sigma \in (0, 1)$. Consequently, $x_{k+1} = x_k + \alpha_k p_k$. An important aspect of the gradient method is to appropriately choose a step size. When applying the line search, backtracking introduces a non-Martingale difference sequence, and a direct application of the law of large numbers is not possible. To overcome this challenge, empirical process theory is utilized to analyze the resulting processes (Iusem et al. 2019), allowing for deriving optimal iteration complexity $\mathcal{O}(1/\epsilon)$ and near-optimal oracle complexity $\mathcal{O}(1/\epsilon^2)$ (also see Paquette and Scheinberg (2018)).

2.3. Expectation-Valued Constrained Problems

Suppose the problem is complicated by stochastic constraints of the form $\mathbb{E}[C_i(x, \xi)] = 0, i = 1, \ldots, m, x \ge 0$, where $C_i : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ is a twice-continuously differentiable function for $i = 1, \ldots, m$. Although nonlinear programming has been processed via interior-point schemes, sequential quadratic programming schemes, and Lagrangian methods, stochastic extensions remain unaddressed. For instance, in line-search interior-point schemes, the subproblem requires solving the linearized Newton system, requiring both the Hessian of the (stochastic) objective and each (stochastic) nonlinear constraint function. Subsequently, choosing a step length requires conducting a line search with respect to a suitable merit function defined as $\mathcal{M}(z_k + \alpha_k p_k, \rho_k) \le \mathcal{M}(z_k, \rho_k) + \sigma \nabla_x \mathcal{M}(z_k, \rho_k)^T p_k$, where $\mathcal{M}(z, \rho) = \mathbb{E}[F(x, \xi)] - \sum_{i=1}^m \mathbb{E}[C_i(x, \xi)] + \frac{1}{2} \rho \sum_{i=1}^m \|\mathbb{E}[C_i(x, \xi)]\|^2$. Developing asymptotics and rate statements (even local) remain open questions for stochastic extensions of such schemes.

2.4. Stochastic Quasi-Newton Enhancements

In deterministic regimes, second derivatives may either be unavailable or challenging to compute. As a result, the matrix B_k in the TR or the LS subproblem is no longer the true Hessian but may be constructed as through a quasi-Newton update, an example being the celebrated BFGS update rule (Nocedal and Wright 2006). Can one integrate recently developed stochastic BFGS updates (Jalilzadeh et al. 2018a) to develop BFGS variants? Can variance-reduced counterparts allow for achieving deterministic rates of convergence?

2.5. Further Challenges

Beyond Sections 2.3 and 2.4, there are a host of challenges that remain, including ensuring convergence to second-order Karush–Kuhn–Tucker points (such as by combining negative curvature directions with Newton directions in a stochastic curvilinear search) (Moguerza and Prieto 2003, Curtis and Robinson 2019) to allowing for nonsmoothness and semismoothness in the problem definition (Facchinei and Pang 2007).

References

Beck A, Teboulle M (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* 2(1):183–202.

Blanchet J, Cartis C, Menickelly M, Scheinberg K (2019) Convergence rate analysis of a stochastic trust region method via supermartingales. *INFORMS J. Optim.* 1(2):91–183.

Chang KH, Hong LJ, Wan H (2013) Stochastic trust-region response-surface method (STRONG)—A new response-surface framework for simulation optimization. *INFORMS J. Comput.* 25(2):230–243.

Chen H-F (2002) *Stochastic Approximation and Its Applications*, vol. 64 (Kluwer Academic Publishers, Dordrecht, Netherlands).

Chen R, Menickelly M, Scheinberg K (2018) Stochastic optimization using a trust-region method and random models. *Math. Programming* 169(2): 447–487.

Chow YS, Teicher H (2012) *Probability Theory: Independence, Interchangeability, Martingales* (Springer-Verlag, New York).

Curtis FE, Robinson DP (2019) Exploiting negative curvature in deterministic and stochastic optimization. *Math. Programming* 176(1–2):69–94.

Facchinei F, Pang JS (2007) *Finite-Dimensional Variational Inequalities and Complementarity Problems* (Springer-Verlag, New York).

Feller W (1945) The fundamental limit theorems in probability. *Bull. Amer. Math. Soc.* 51(11):800–832.

Ghadimi S, Lan G (2016) Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Programming* 156(1–2): 59–99.

Ghadimi S, Lan G, Zhang H (2016) Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Math. Programming* 155(1–2):267–305.

Iusem AN, Jofré A, Oliveira RI, Thompson P (2019) Variance-based extragradient methods with line search for stochastic variational inequalities. *SIAM J. Optim.* 29(1):175–206.

Jalilzadeh A, Shanbhag UV (2016) eg-VSSA: An extragradient variable sample-size stochastic approximation scheme: Error analysis and complexity trade-offs. *2016 Winter Simulation Conf. (WSC)* (Institute of Electrical and Electronics Engineers, Piscataway, NJ), 690–701.

Stochastic Systems, 2019, vol. 9, no. 3, pp. 299–302, © 2019 The Author(s)

Jalilzadeh A, Nedić A, Shanbhag UV, Yousefian F (2018a) A variable sample-size stochastic quasi-Newton method for smooth and nonsmooth stochastic convex optimization. *2018 IEEE Conf. Decision Control (CDC)* (Institute of Electrical and Electronics Engineers, Piscataway, NJ), 4097–4102.

Jalilzadeh A, Shanbhag UV, Blanchet JH, Glynn PW (2018b) Optimal smoothed variable sample-size accelerated proximal methods for structured nonsmooth stochastic convex programs. Preprint arXiv:1803.00718, submitted March 2, https://arxiv.org/abs/1803.00718.

Jofré A, Thompson P (2019) On variance reduction for stochastic smooth convex optimization with multiplicative noise. *Math. Program.* 174(1–2):253–292.

Juditsky A, Nemirovski A, Tauvel C (2011) Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems* 1(1):17–58.

Koshal J, Nedić A, Shanbhag UV (2013) Regularized iterative stochastic approximation methods for stochastic variational inequality problems. *IEEE Trans. Automatic Control* 58(3):594–609.

Kushner H, Yin GG (2003) *Stochastic Approximation and Recursive Algorithms and Applications*, vol. 35 (Springer-Verlag, New York).

Lindeberg JW (1922) Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. Mathematische Zeitschrift 15(1): 211–225.

Lucchi A, McWilliams B, Hofmann T (2015) A variance reduced stochastic Newton method. Preprint arXiv:1503.08316, submitted March 28, https://arxiv.org/abs/1503.08316.

Mandt S, Hoffman MD, Blei DM (2017) Stochastic gradient descent as approximate Bayesian inference. *J. Machine Learn. Res.* 18(1):4873–4907.

Moguerza JM, Prieto FJ (2003) An augmented Lagrangian interior-point method using directions of negative curvature. *Math. Programming* 95(3): 573–616.

Mokhtari A, Ribeiro A (2014) RES: Regularized stochastic BFGS algorithm. *IEEE Trans. Signal Processing* 62(23):6089–6104.

Nesterov YE (1983) A method for solving the convex programming problem with convergence rate of $(1/k^2)$. *Soviet Math. Doklady* 27(2):372–376.

Nocedal J, Wright S (2006) *Numerical Optimization*. (Springer-Verlag, New York).

Paquette C, Scheinberg K (2018) A stochastic line search method with convergence rate analysis. Preprint, submitted July 20, https://arxiv.org/abs/1807.07994.

Shanbhag UV, Blanchet JH (2015) Budget-constrained stochastic approximation. *Proc. 2015 Winter Simulation Conf.* (Institute of Electrical and Electronics Engineers, Piscataway, NJ), 368–379.

Shashaani S, Hashemi FS, Pasupathy R (2018) ASTRO-DF: A class of adaptive sampling trust-region algorithms for derivative-free stochastic optimization. *SIAM J. Optim.* 28(4):3145–3176.

Yousefian F, Nedić A, Shanbhag UV (2013) A regularized smoothing stochastic approximation (RSSA) algorithm for stochastic variational inequality problems. *Winter Simulations Conf.: Simulation Making Decisions Complex World, WSC 2013* (Institute of Electrical and Electronics Engineers, Piscataway, NJ), 933–944.

Yousefian F, Nedić A, Shanbhag UV (2014) Optimal robust smoothing extragradient algorithms for stochastic variational inequality problems. *53rd IEEE Conf. Decision Control* (Institute of Electrical and Electronics Engineers, Piscataway, NJ), 5831–5836.

Yousefian F, Nedić A, Shanbhag UV (2017) On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems. *Math. Programming* 165(1):391–431.

Zhou C, Gao W, Goldfarb D (2017) Stochastic adaptive quasi-Newton methods for minimizing expected values. *Proc. 34th Internat. Conf. Machine Learn.* vol. 70, 4150–4159.