Adaptive Sampling Controlled Stochastic Recursions

Raghu Pasupathy {pasupath@purdue.edu},
Purdue Statistics, West Lafayette, IN

Co-authors:
Soumyadip Ghosh (IBM Watson Research);
Fatemeh Hashemi (Virginia Tech);
Peter Glynn (Stanford University).

January 7, 2016

The talk that did not make it ... !

THE TALK THAT DID NOT MAKE IT ... !

1. An Overview of Stochastic Approximation and Sample-Average Approximation Methods

THE TALK THAT DID NOT MAKE IT ... !

1. An Overview of Stochastic Approximation and Sample-Average Approximation Methods
2. Some References:
   2.1 A Guide to SAA [Kim et al., 2014]
   2.2 Lectures on Stochastic Prgramming: Modeling and Theory [Shapiro et al., 2009]
   2.3 Simulation Optimization: A Concise Overview and Implementation Guide [Pasupathy and Ghosh, 2013]
   2.4 Introduction to Stochastic Search and Optimization [Spall, 2003]

The talk that made it …

Adaptive Sampling Controlled Stochastic Recursions

1. Problem Statement
2. Canonical Rates in Simulation Optimization
3. Stochastic Approximation
4. Adaptive Sampling Controlled Stochastic Recursion (ASCSR)
5. The Optimality of ASCSR
6. Sample Numerical Experience
7. Final Remarks

Problem Context

Simulation Optimization

"Solve an optimization problem when only 'noisy' observations of the objective functions/constraints are available."

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g(x) \leq 0, x \in \mathcal{D}; \end{array}$$

Problem Context

Simulation Optimization

"Solve an optimization problem when only 'noisy' observations of the objective functions/constraints are available."

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g(x) \leq 0, x \in \mathcal{D}; \end{array}$$

- $f : \mathcal{D} \to \mathbb{R}$ (and its derivative) can only be estimated, e.g., $F_m(x) = m^{-1} \sum_{i=1}^{m} F_j(x)$, where $F_j(x)$ are iid random variables with mean $f(x)$;
- $g : \mathcal{D} \to \mathbb{R}^c$ can only be estimated using $G_m = m^{-1} \sum_{i=1}^{m} G_j(x)$, where $G_j(x)$ are iid random vectors with mean $g(x)$;
- unbiased observations of the derivative of f may or may not be available.

Problem Context

Stochastic Root Finding

"Find a zero of a function when only 'noisy' observations of the function are available."

$$\begin{aligned} \text{find} \quad & x \\ \text{such that} \quad & f(x) = 0, x \in \mathcal{D}; \end{aligned}$$

where

- $f : \mathcal{D} \to \mathbb{R}^c$ can only be estimated using $F_m = m^{-1} \sum_{i=1}^{m} F_j(x)$, where $F_j(x)$ are iid random vectors with mean $f(x)$.

"STOCHASTIC COMPLEXITY," CANONICAL RATES

Examples:

(i) $\xi = \mathbb{E}[X], \hat{\xi}(m) = m^{-1}\sum_{i=1}^{m} X_i$ where $X_i, i = 1, 2, \ldots$ are iid copies of X. Then, when $\mathbb{E}[X^2] < \infty$,

$$\mathrm{rmse}(\hat{\xi}(\mathrm{m}), \xi) = \mathcal{O}(\mathrm{m}^{-1/2}).$$

"Stochastic Complexity," Canonical Rates

Examples:

(i) $\xi = \mathbb{E}[X], \hat{\xi}(m) = m^{-1}\sum_{i=1}^{m} X_i$ where $X_i, i = 1, 2, \dots$ are iid copies of X. Then, when $\mathbb{E}[X^2] < \infty$,

$$\mathrm{rmse}(\hat{\xi}(\mathrm{m}), \xi) = \mathcal{O}(\mathrm{m}^{-1/2}).$$

(ii) $\xi = g'(x)$ and $\hat{\xi}(m) = \dfrac{\overline{Y}_m(x+s) - \overline{Y}_m(x-s)}{2s}$, where $g(\cdot) : \mathbb{R} \to \mathbb{R}$ and $Y_i(x), i = 1, 2, \dots$ are iid copies of $Y(x)$ satisfying $\mathbb{E}[Y(x)] = g(x)$. Then, when $s = \Theta(m^{-1/6})$,

$$\mathrm{rmse}(\hat{\xi}(\mathrm{m}), \xi) = \mathcal{O}(\mathrm{m}^{-1/3}).$$

"Stochastic Complexity," Canonical Rates

Examples:

(i) $\xi = \mathbb{E}[X], \hat{\xi}(m) = m^{-1}\sum_{i=1}^{m} X_i$ where $X_i, i = 1, 2, \ldots$ are iid copies of X. Then, when $\mathbb{E}[X^2] < \infty$,

$$\text{rmse}(\hat{\xi}(\text{m}), \xi) = \mathcal{O}(\text{m}^{-1/2}).$$

(ii) $\xi = g'(x)$ and $\hat{\xi}(m) = \dfrac{\overline{Y}_m(x+s) - \overline{Y}_m(x-s)}{2s}$, where $g(\cdot) : \mathbb{R} \to \mathbb{R}$ and $Y_i(x), i = 1, 2, \ldots$ are iid copies of $Y(x)$ satisfying $\mathbb{E}[Y(x)] = g(x)$. Then, when $s = \Theta(m^{-1/6})$,

$$\text{rmse}(\hat{\xi}(\text{m}), \xi) = \mathcal{O}(\text{m}^{-1/3}).$$

For forward differences, $s = \Theta(m^{-1/4})$,

$$\text{rmse}(\hat{\xi}(\text{m}), \xi) = \mathcal{O}(\text{m}^{-1/4}).$$

“Stochastic Complexity,” Canonical Rates

Examples: ... contd.

(iii) Owing to (i), SO and SRFP algorithms “declare victory” if the error $\|X_k - x^*\|$ in their solution estimator X_k decays as $\mathcal{O}_p(1/\sqrt{W_k})$, where W_k is the *total* simulation effort expended towards obtaining X_k.

“STOCHASTIC COMPLEXITY,” CANONICAL RATES

But I hasten to add...

- There is a now a well-understood relationship between smoothness and complexity in convex problems primarily due to the work of Alexander Shapiro, Arkadi Nemirovskii, and Yuri Nesterov — see Bubeck (2014) for a beautiful monograph.
- Is there an analogous theory to be developed based on the assumed structural property of the sample-paths?

Stochastic Approximation (SA)

Robbins and Monro (1951):

$$X_{k+1} = X_k - a_k H(X_k),$$

where $H(x)$ estimates $h(x) \triangleq \nabla f(x)$.

Stochastic Approximation (SA)

Robbins and Monro (1951):

$$X_{k+1} = X_k - a_k H(X_k),$$

where $H(x)$ estimates $h(x) \triangleq \nabla f(x)$.

Kiefer-Wolfowitz (1952) analogue for optimization:

$$X_{k+1} = X_k - a_k \left(\frac{F(X_k + s_k) - F(X_k)}{s_k} \right),$$

where $F(x)$ estimates $f(x)$.

Stochastic Approximation (SA)

Robbins and Monro (1951):

$$X_{k+1} = X_k - a_k H(X_k),$$

where $H(x)$ estimates $h(x) \triangleq \nabla f(x)$.

Kiefer-Wolfowitz (1952) analogue for optimization:

$$X_{k+1} = X_k - a_k \left(\frac{F(X_k + s_k) - F(X_k)}{s_k} \right),$$

where $F(x)$ estimates $f(x)$.

Modern Incarnations:

$$\begin{aligned} X_{k+1} &= \Pi_D[X_k - a_k B_k^{-1} H(X_k)], \quad \text{(RM)}; \\ X_{k+1} &= \Pi_D[X_k - a_k B_k^{-1} \hat{\nabla} F(X_k)], \quad \text{(KW)}; \end{aligned}$$

where D is the feasible space, and $\Pi_D[x]$ denotes projection.

Stochastic Approximation

SA is ubiquitous

1. SA is probably amongst most used algorithms. (Typing "Stochastic Approximation" in Google Scholar brings up about 1.77 million hits!)
2. SA is backed by more than six decades of research.
3. Enormous number of variations of SA have been created and studied.
4. SA is used in virtually every field where there is a need for stochastic optimization (Pasupathy (2014)).

SA: Asymptotics

1. Convergence ($\mathcal{L}_2$,wp1) guaranteed assuming
   - C.1 structural conditions on f, g;
   - C.2 $\sum_{k=1}^{\infty} a_k = \infty$;
   - C.3 $\sum_{k=1}^{\infty} a_k^2 < \infty$ for Robbins-Munro and $\sum_{k=1}^{\infty} a_k^2/s_k^2 < \infty, s_k \to 0$ for Kiefer-Wolfowitz.

   ($C.3$ can be weakened to $a_k \to 0$ [Broadie et al., 2011].)
2. The canonical rate of $O_p(1/\sqrt{k})$ is achievable for Robbins-Munro [Fabian, 1968, Polyak and Juditsky, 1992, Ruppert, 1985, Ruppert, 1991].
3. Deterioration in the Kiefer-Wolfowitz context [Mokkadem and Pelletier, 2011, Djeddour et al., 2008].
   (Loosely, when $\rho/v(s_k)$ is the deterministic bias of the recursion, the best achievable rate is $\Theta(1/\sqrt{ks_k^2})$ achieved when s_k is chosen so that $v(s_k)^{-1}\sqrt{kc_k^2}$ has a nonzero limit.)

Why an Alternative Paradigm?

1. SA's parameters are still difficult to choose.
   - Conditions $C.2$ and $C.3$ leave enormous classes feasible parameter sequences from which to choose. (See Broadie et al. [Broadie et al., 2011]; Vaidya and Bhatnagar [Vaidya and Bhatnagar, 2006] for further detail.)
   - Nemiroski, Juditsky, Lan and Shapiro [Nemirovski et al., 2009] demonstrate that there can be a severe degradation in the convergence rate of SA-type methods if the parameters inherent to the function are guessed incorrectly.
2. Shouldn't advances in nonlinear programming be exploited more fully?
3. SA does not lend itself to trivial parallelization.

Sampling Controlled Stochastic Recursion (SCSR)

An Alternative to SA?

Instead of SA, why not just employ your favorite deterministic recursion (e.g., quasi-Newton, trust region), and replace unknown quantities in the recursion by Monte Carlo estimators?

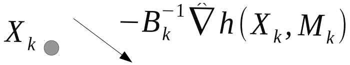
Sampling Controlled Stochastic Recursion (SCSR)

An Alternative to SA?

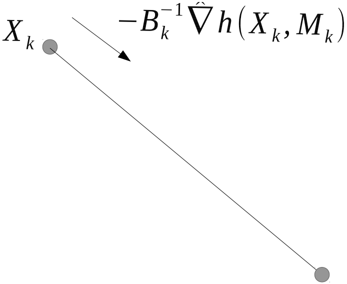
Instead of SA, why not just employ your favorite deterministic recursion (e.g., quasi-Newton, trust region), and replace unknown quantities in the recursion by Monte Carlo estimators?

- Use a recursion (such as line search) as the underlying search mechanism;
- Sample judiciously.

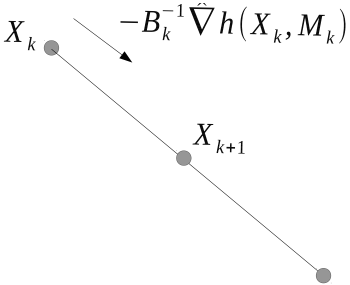
ADAPTIVE SCSR: LINE SEARCH



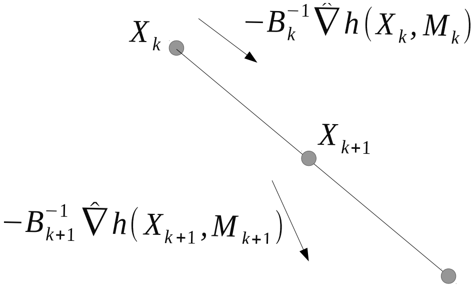
Adaptive SCSR: Line Search



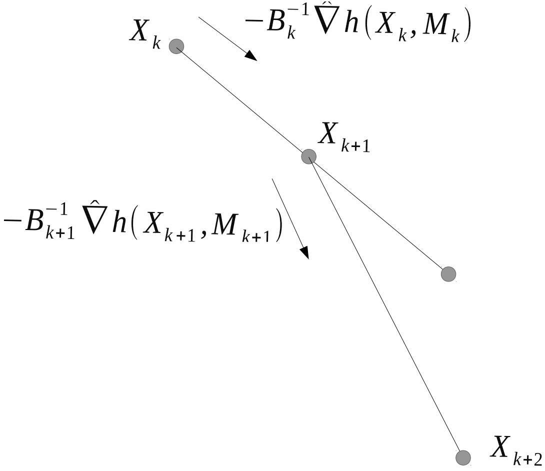
Adaptive SCSR: Line Search



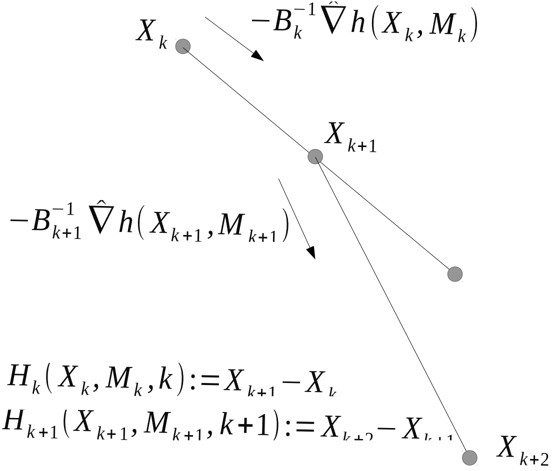
Adaptive SCSR: Line Search



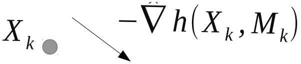
ADAPTIVE SCSR: LINE SEARCH



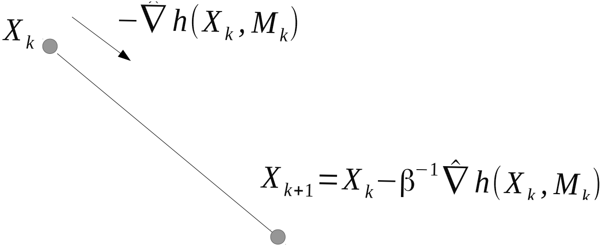
Adaptive SCSR: Line Search



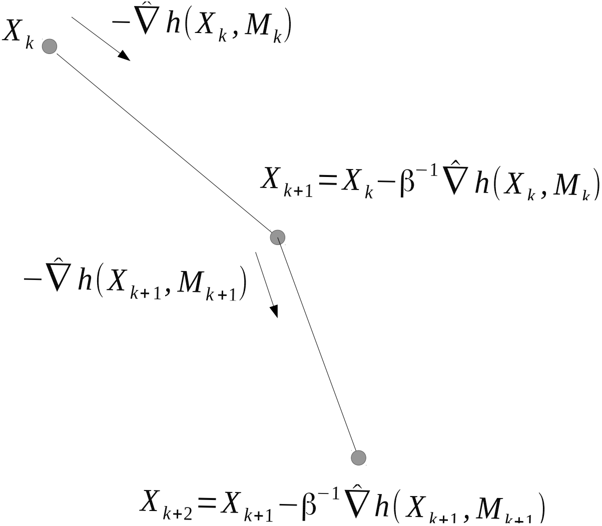
Adaptive SCSR: Gradient Search



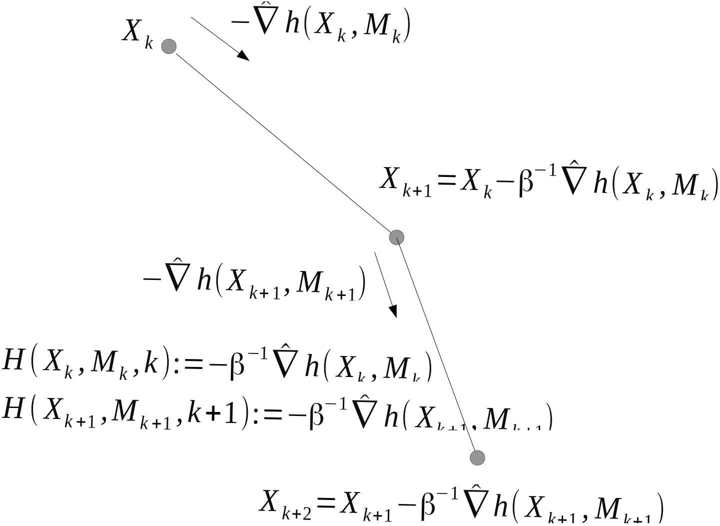
Adaptive SCSR: Gradient Search



Adaptive SCSR: Gradient Search



Adaptive SCSR: Gradient Search



Sampling-Controlled Stochastic Recursion (SCSR)

An Alternative to SA?

$$
\begin{aligned}
X_{k+1} &= X_k + H_k(X_k, M_k, k), \quad k = 1, 2, \ldots. \quad \text{(SCSR)} \\
x_{k+1} &= x_k + h_k(x_k, k), \quad k = 1, 2, \ldots. \quad \text{(DA)}
\end{aligned}
$$

SAMPLING-CONTROLLED STOCHASTIC RECURSION (SCSR)

AN ALTERNATIVE TO SA?

$$
\begin{aligned}
X_{k+1} &= X_k + H_k(X_k, M_k, k), \quad k = 1, 2, \ldots. \quad \text{(SCSR)} \\
x_{k+1} &= x_k + h_k(x_k, k), \quad k = 1, 2, \ldots. \quad \text{(DA)}
\end{aligned}
$$

1. How should the sample size M_k be chosen (adatively) to ensure convergence wp1 of the iterates $\{X_k\}$?
2. Can the canonical rate be achieved in such "practical" algorithms?

Sampling-Controlled Stochastic Recursion (SCSR)

An Alternative to SA?

$$\begin{aligned} X_{k+1} &= X_k + H_k(X_k, M_k, k), \quad k = 1, 2, \ldots. \quad \text{(SCSR)} \\ x_{k+1} &= x_k + h_k(x_k, k), \quad k = 1, 2, \ldots. \quad \text{(DA)} \end{aligned}$$

1. Some theory on non-adaptive "optimal sampling rates" has been developed recently [Pasupathy et al., 2014]. More)

Sampling-Controlled Stochastic Recursion (SCSR)

An Alternative to SA?

$$\begin{aligned} X_{k+1} &= X_k + H_k(X_k, M_k, k), \quad k = 1, 2, \ldots. \quad \text{(SCSR)} \\ x_{k+1} &= x_k + h_k(x_k, k), \quad k = 1, 2, \ldots. \quad \text{(DA)} \end{aligned}$$

1. Some theory on non-adaptive "optimal sampling rates" has been developed recently [Pasupathy et al., 2014]. ▸ More)
2. Virtually all recursions in [Ortega and Rheinboldt, 1970] and in [Duflo and Wilson, 1997] are subsumed.
3. Trust-region [Conn et al., 2000] and DFO-type recursions [Conn et al., 2009] are subsumed with effort!
4. Two prominent "realizations" of SCSR-type algorithms are [Byrd et al., 2012] and [Chang et al., 2013].

Adative SCSR

The Guiding Principle for Optimal Sampling

Write:

$$X_{k+1} = X_k + H_k(X_k, M_k, k), \quad k = 1, 2, \ldots . \quad \text{(SCSR)}$$

ADATIVE SCSR

THE GUIDING PRINCIPLE FOR OPTIMAL SAMPLING

Write:

$$X_{k+1} = X_k + H_k(X_k, M_k, k), \quad k = 1, 2, \ldots . \quad \text{(SCSR)}$$

as

$$X_{k+1} - x^* = \underbrace{X_k + h_k(X_k, k) - x^*}_{\text{structural error}} + \underbrace{H_k(X_k, M_k, k) - h_k(X_k, k)}_{\text{sampling error}}.$$

Adative SCSR

The Guiding Principle for Optimal Sampling

Write:

$$X_{k+1} = X_k + H_k(X_k, M_k, k), \quad k = 1, 2, \ldots. \quad \text{(SCSR)}$$

as

$$X_{k+1} - x^* = \underbrace{X_k + h_k(X_k, k) - x^*}_{\text{structural error}} + \underbrace{H_k(X_k, M_k, k) - h_k(X_k, k)}_{\text{sampling error}}.$$

(i) Sample so that $\|H_k(X_k, M_k) - h_k(X_k)\| \approx \|X_k + h_k(X_k, k) - x^*\|$ in some sense, for optimal evolution;
(ii) Fast structural recursion with (i) ensures efficiency, a fact that is not immediately evident.

Adaptive SCSR

Sample Size Determination

How much to sample? Sample until structural error estimate $\approx$ sampling error estimate?

Adaptive SCSR

Sample Size Determination

How much to sample? Sample until structural error estimate ≈ sampling error estimate?

$$M_k|\mathcal{F}_k = \inf_{m \geq \nu(k)} \left\{ m^{\epsilon} \hat{\text{se}}(H_k(X_k, m)) < c \|H_k(X_k, m)\| | \mathcal{F}_k \right\},$$

Adaptive SCSR

Sample Size Determination

How much to sample? Sample until structural error estimate $\approx$ sampling error estimate?

$$M_k|\mathcal{F}_k = \inf_{m \geq \nu(k)} \{m^{\epsilon}\hat{\text{se}}(H_k(X_k, m)) < c\|H_k(X_k, m)\||\mathcal{F}_k\},$$

which is usually,

$$M_k|\mathcal{F}_k = \inf_{m \geq \nu(k)} \left\{m^{\epsilon}\frac{\hat{\sigma}(X_k, m)}{\sqrt{m}} < c\|H_k(X_k, m)\||\mathcal{F}_k\right\}.$$

1. $\{\nu(k)\} \to \infty$ is the "escorting sequence," and ϵ is the "coercion" constant.
2. The constants $c, \beta > 0$.

Adaptive SCSR

Heuristic Interpretation I: Byrd, Chin, Nocedal and Wu (2012)

1. At X_k, $d = H_k(X_k, M_k)$ is a descent direction at X_k if $\|H_k(X_k, m) - h_k(X_k)\|_2 \leq c\|H(X_k, m)\|_2$ for some $c \in [0, 1)$.
2. Notice:

$$\mathbb{E}[\|H_k(X_k, m) - h_k(X_k)\|_2^2 | \mathcal{F}_k] = \mathbb{V}(H_k(X_k, m) | \mathcal{F}_k).$$

The above two points inspires the heuristic:

$$M_k | \mathcal{F}_k = \inf_m \{\sqrt{\hat{\mathbb{V}}(H_k(X_k) | \mathcal{F}_k)} \leq c\|H_k(X_k, m)\|_2 | \mathcal{F}_k\}. \qquad (1)$$

(Sample until estimated error in gradient is less than c times gradient estimate, i.e., until you are confident you have a descent direction.)

Adaptive SCSR

Heuristic Interpretation II: Pasupathy and Schmeiser (2010)

1. The coefficient of variation of $H_k(X_k, m)|\mathcal{F}_k$ can be estimated as

$$\hat{\text{cv}}\left(H_k(X_k, m)|\mathcal{F}_k\right) = \frac{\sqrt{\hat{\mathbb{V}}\left(H_k(X_k, m)|\mathcal{F}_k\right)}}{H_k(X_k, m)}.$$

2. A "reasonable" heuristic is to then continue sampling until the absolute value of the estimated coefficient of variation drops below the fixed threshold c.

Adaptive SCSR

Theoretical Results: Standing Assumptions and Notation

A.1 There exists a unique root x^* such that $h(x^*) = 0$.

A.2 There exists ℓ_0, ℓ_1 such that for all $x \in \mathcal{D}$,
$\ell_0 \|x - x^*\|_2^2 \leq h^T(x)h(x) \geq \ell_1 \|x - x^*\|_2^2$.

A.3 $H_k(X_k, m) \triangleq h(X_k) + \sum_{j=1}^{m} \xi_{kj}$, where ξ_k is a martingale-difference process defined on the probability space $(\Omega, \mathcal{F}, \mathcal{F}_k, P)$, and ξ_{kj} are iid copies of ξ_k.

Adaptive SCSR

Theoretical Results: Some Intuition on Iteration Evolution

Letting $Z_k = X_k - x^*$, we see that

$$Z_{k+1} = Z_k + \frac{1}{\beta}h(X_k) + \frac{1}{\beta}(H(X_k, M_k) - h(X_k)), \text{ and}$$

Adaptive SCSR

Theoretical Results: Some Intuition on Iteration Evolution

Letting $Z_k = X_k - x^*$, we see that

$$Z_{k+1} = Z_k + \frac{1}{\beta} h(X_k) + \frac{1}{\beta}(H(X_k, M_k) - h(X_k)), \text{ and}$$

$$\mathbb{E}_\Omega[Z_{k+1}^2 | \mathcal{F}_k] \leq \underbrace{\left(1 - \frac{2\ell_0}{\beta} + \frac{\ell_1^2}{\beta^2}\right) Z_k^2}_{\text{structural error}} + \underbrace{\frac{1}{\beta^2} \mathbb{E}_\Omega[\|H(X_k, M_k) - h(X_k)\|^2 | \mathcal{F}_k]}_{\text{sampling error}}.$$

Adaptive SCSR

Theoretical Results: Some Intuition on Iteration Evolution

Letting $Z_k = X_k - x^*$, we see that

$$Z_{k+1} = Z_k + \frac{1}{\beta}h(X_k) + \frac{1}{\beta}(H(X_k, M_k) - h(X_k)), \text{ and}$$

$$\mathbb{E}_\Omega[Z_{k+1}^2|\mathcal{F}_k] \le \underbrace{\left(1 - \frac{2\ell_0}{\beta} + \frac{\ell_1^2}{\beta^2}\right) Z_k^2}_{\text{structural error}} + \underbrace{\frac{1}{\beta^2}\mathbb{E}_\Omega[\|H(X_k, M_k) - h(X_k)\|^2|\mathcal{F}_k]}_{\text{sampling error}}.$$

Recall Guiding Principles:
(i) $\mathbb{E}_\Omega[\|H(X_k, M_k) - h(X_k)\|^2|\mathcal{F}_k] \approx h^2(X_k)$ for opt. evolution;
(ii) fast structural recursion with (i) for efficiency.

Adaptive SCSR

Theoretical Results: Consistency

Theorem

Let the sequence $\{\nu_k\}$ *satisfy* $\sum_k \nu_k^{-1} < \infty$. *Then the A-SCSR iterates* $\{X_k\}$ *satisfy* $\{X_k\} \xrightarrow{a.s.} x^*$ *as* $k \to \infty$.

Adaptive SCSR

Theoretical Results: Consistency

Theorem

Let the sequence $\{\nu_k\}$ satisfy $\sum_k \nu_k^{-1} < \infty$. Then the A-SCSR iterates $\{X_k\}$ satisfy $\{X_k\} \stackrel{a.s.}{\to} x^*$ as $k \to \infty$.

Proof Sketch.

$$\begin{aligned}
&\mathbb{E}_\Omega[\|H(X_k, M_k) - h(X_k)\|^2|\mathcal{F}_k] \\
&\quad\leq \frac{1}{\nu_k}\mathbb{E}_\Omega[M_k\|H(X_k, M_k) - h(X_k)\|^2|\mathcal{F}_k] \\
&\quad\leq \frac{1}{\nu_k}\mathbb{E}_\Omega[\sup_m \|\sqrt{m}\,(H(X_k, m) - h(X_k))\,\|^2|\mathcal{F}_k] \\
&\quad= O(\frac{1}{\nu_k}).
\end{aligned}$$

Adaptive SCSR

Theoretical Results: Quality of Estimator

Theorem

Let $\sigma^2 = \mathbb{V}(Y_1(x^*)) < \infty$. *Recalling that* $M_k|\mathcal{F}_k = \inf_{m \geq \nu(k)} \left\{ m^{\epsilon} \frac{\hat{\sigma}(X_k, m)}{\sqrt{m}} < c\|H(X_k, m)\| | \mathcal{F}_k \right\}$, *we have as* $k \to \infty$,

$$\frac{\mathbb{E}[\|H(X_{k+1}, M_{k+1})\|^2 | \mathcal{F}_k]}{\mathbb{E}[M_{k+1}^{-1+2\epsilon} | \mathcal{F}_k]} \overset{a.s.}{\to} \frac{\sigma^2}{c^2}.$$

1. Proof relies on the fact that the conditional second moment of the excess is uniformly bounded away from infinity.
2. The theorem essentially connects the sampling error with the sequential sample size.

Adaptive SCSR

Theoretical Results: Behavior of Sample Size

Theorem

Denote $\eta = 2/(1-2\epsilon)$. The following hold as $k \to \infty$ and for some $\delta > 0$.

(i) *If $x \leq 4^{-\eta/2}(\frac{\sigma^2}{c^2})^{\eta/2}$, then*

$$\mathbb{P}\{h^{\eta}(X_k)M_k \leq x | \mathcal{F}_k\} \leq \exp\{-h^{-\delta}(X_k)\}.$$

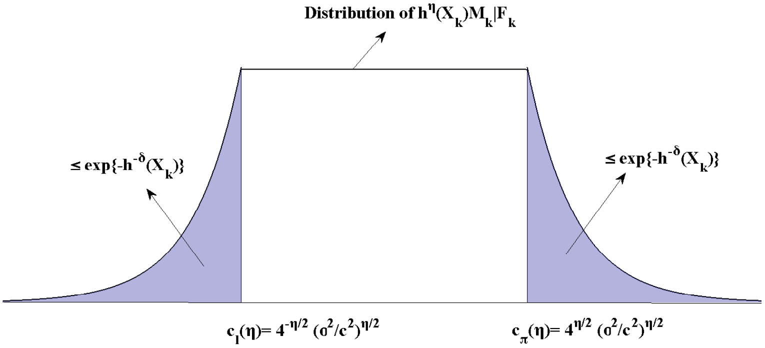
(ii) *If $x \geq 4^{\eta/2}(\frac{\sigma^2}{c^2})^{\eta/2}$, then*

$$\mathbb{P}\{h^{\eta}(X_k)M_k \geq x | \mathcal{F}_k\} \leq \exp\{-h^{-\delta}(X_k)\}.$$

(In English, M_k concentrates around $h^{-\eta}(X_k)$.)

ADAPTIVE SCSR

THEORETICAL RESULTS: BEHAVIOR OF SAMPLE SIZE



Adaptive SCSR

Theoretical Results: Behavior of Sample Size

Theorem

Denote $\eta = 2/(1-2\epsilon)$. Then following hold almost surely.

(i) $\liminf_k h^{\eta}(X_k)\mathbb{E}[M_k|\mathcal{F}_k] \geq 4^{-\eta/2}(\frac{\sigma^2}{c^2})^{\eta/2}$.

(ii) $\limsup_k h^{\eta}(X_k)\mathbb{E}[M_k|\mathcal{F}_k] \leq 4^{\eta/2}(\frac{\sigma^2}{c^2})^{\eta/2}$.

Adaptive SCSR

Theoretical Results: Behavior of Sample Size

Theorem

Denote $\eta = 2/(1-2\epsilon)$. Then following hold almost surely.

(i) $\liminf_k h^\eta(X_k)\mathbb{E}[M_k|\mathcal{F}_k] \geq 4^{-\eta/2}(\frac{\sigma^2}{c^2})^{\eta/2}$.

(ii) $\limsup_k h^\eta(X_k)\mathbb{E}[M_k|\mathcal{F}_k] \leq 4^{\eta/2}(\frac{\sigma^2}{c^2})^{\eta/2}$.

Theorem

Denote $\eta = 2/(1-2\epsilon)$. Then following hold almost surely.

(i) $\liminf_k h^{-2}(X_k)\mathbb{E}[M_k^{-1+2\epsilon}|\mathcal{F}_k] \geq 1/4$.

(ii) $\limsup_k h^{-2}(X_k)\mathbb{E}[M_k^{-1+2\epsilon}|\mathcal{F}_k] \leq 4$.

(Loosely, $\mathbb{E}[M_k^{-1+2\epsilon}|\mathcal{F}_k] \approx h^2(X_k)$.)

Adaptive SCSR

Theoretical Results: Efficiency

Theorem

Let $W_k = \sum_j M_j$ denote the total simulation effort after k iterations. Then,

(i) $E[\|X_k - x^*\|^2 W_k^{1-2\epsilon}] = O(1)$ *as* $k \to \infty$*;*

(ii) *If* $M_k = o_p(W_k)$*, then* $W_k^{1-2\epsilon}\|X_k - x^*\|^2 \xrightarrow{p} \infty$*.*

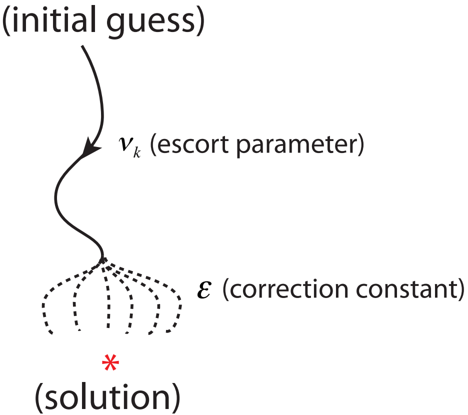
1. The result says that the mean squared error $\mathbb{E}[\|X_k - x^*\|^2] \approx (\mathbb{E}[W_k])^{-1}$, coinciding with the estimation rate.
2. Sampling should be atleast "geometric," irrespective of error!

Adaptive SCSR

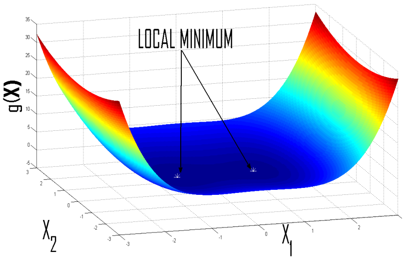
The Escort Sequence and the Coercion Constant

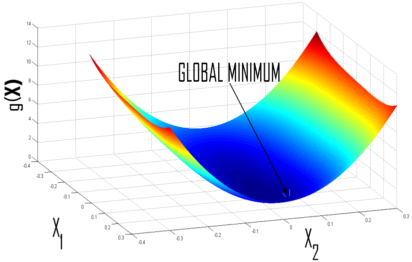
Theorem

Let $W_k = \sum_j M_j$ denote the total simulation effort after k iterations. Then, $\mathbb{P}\{M_k = \nu_k \quad i.o.\} = 0$.



Numerical Illustration





AluffiPentini Function

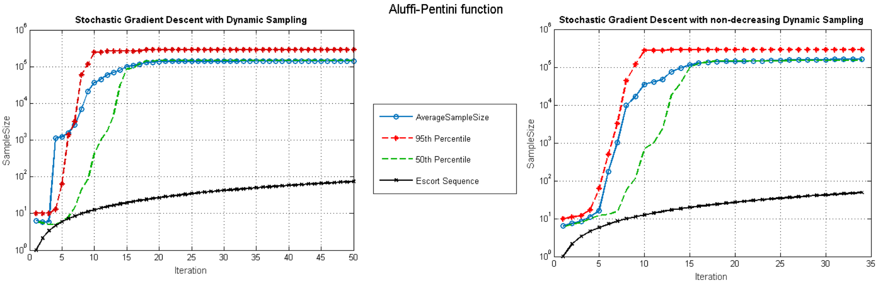
Rosenbrock Function

$g(\boldsymbol{x}) = \mathbb{E}_\xi[0.25(x_1\xi)^4 - 0.5(x_1\xi)^2 + 0.1(x_1\xi) + 0.5x_2^2]$, $\xi \sim N(1, 0.1)$

$g(\boldsymbol{x}) = \mathbb{E}_\xi[100(x_2 - (x_1\xi)^2)^2 + (x_1\xi - 1)^2]$, $\xi \sim N(1, 0.1)$

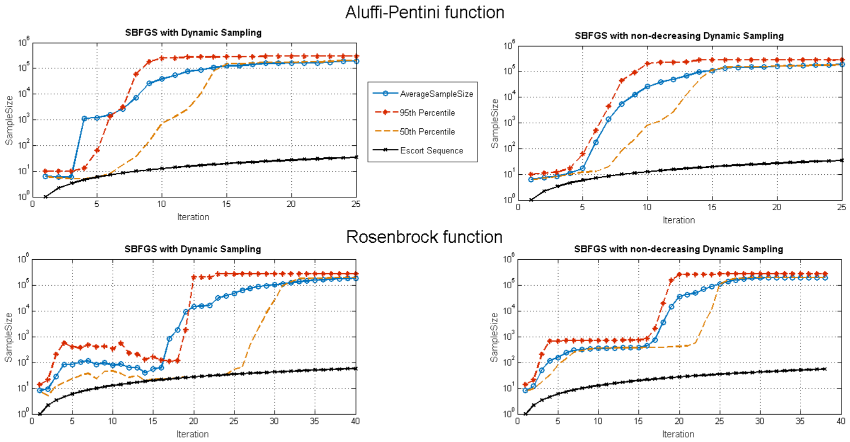
Numerical Illustration

Sample Size Behavior

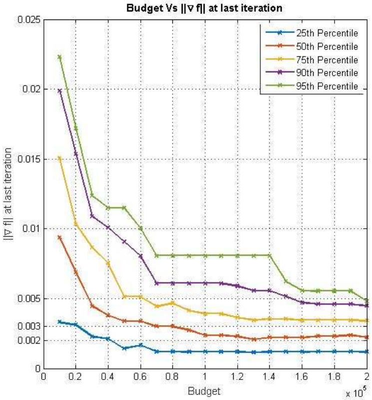


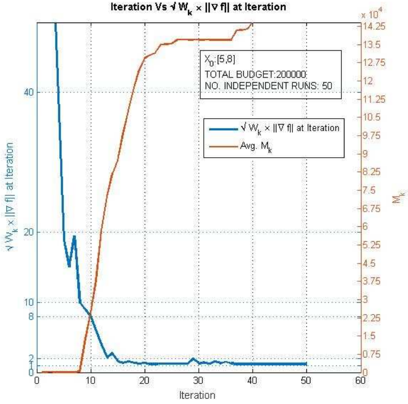
Numerical Illustration

Sample Size Behavior



Numerical Illustration





SUMMARY AND FINAL REMARKS

1. Main Insight for Canonical Rates:
   "Sample until the standard error estimate (of the object being estimated within the recursion) is in lock step with the estimate itself."

Summary and Final Remarks

1. Main Insight for Canonical Rates:
   "Sample until the standard error estimate (of the object being estimated within the recursion) is in lock step with the estimate itself."
   Some details, however, seem important.
   - The escorting sequence $\{\nu_k\}$ is needed to bring iterates to the vicinity of the root.
   - The coercion constant ϵ is needed, unfortunately, to make sure that the sampling error drops at the requisite rate.

Summary and Final Remarks

1. Main Insight for Canonical Rates:
   "Sample until the standard error estimate (of the object being estimated within the recursion) is in lock step with the estimate itself."
   Some details, however, seem important.
   - The escorting sequence $\{\nu_k\}$ is needed to bring iterates to the vicinity of the root.
   - The coercion constant ϵ is needed, unfortunately, to make sure that the sampling error drops at the requisite rate.
2. Generalization to faster recursions will involve a corresponding higher power of the object estimate.

Summary and Final Remarks

1. Main Insight for Canonical Rates:
   "Sample until the standard error estimate (of the object being estimated within the recursion) is in lock step with the estimate itself."
   Some details, however, seem important.
   - The escorting sequence $\{\nu_k\}$ is needed to bring iterates to the vicinity of the root.
   - The coercion constant ϵ is needed, unfortunately, to make sure that the sampling error drops at the requisite rate.
2. Generalization to faster recursions will involve a corresponding higher power of the object estimate.
3. Incorportaion of biased estimators, non-stationary recursions that include more than just the current point seems within reach.

Broadie, M., Cicek, D. M., and Zeevi, A. (2011).
General bounds and finite-time improvement for the kiefer-wolfowitz stochastic approximation algorithm.
Operations Research, 59(5):1211–1224.

Byrd, R. H., Chin, G. M., Nocedal, J., and Wu, Y. (2012).
Sample size selection for optimization methods for machine learning.
Mathematical Programming, Series B, 134:127–155.

Chang, K., Hong, J., and Wan, H. (2013).
Stochastic trust-region response-surface method (strong) – a new response-surface framework for simulation optimization.
INFORMS Journal on Computing.
To appear.

Conn, A. R., Gould, N. I. M., and Toint, P. L. (2000).
Trust-Region Methods.

SIAM, Philadelphia, PA.

- Conn, A. R., Scheinberg, K., and Vincente, L. N. (2009).
  Introduction to Derivative-Free Optimization.
  SIAM, Philadelphia, PA.

- Djeddour, K., Mokkadem, A., and Pelletier, M. (2008).
  On the recursive estimation of the location and of the size of the mode of a probability density.
  Serdica Mathematics Journal, 34:651–688.

- Duflo, M. and Wilson, S. S. (1997).
  Random Iterative Models.
  Springer, New York, NY.

- Fabian, V. (1968).
  On asymptotic normality in stochastic approximation.
  Annals of Mathematical Statistics, 39:1327–1332.

- Kim, S., Pasupathy, R., and Henderson, S. G. (2014).
  A guide to SAA.

Frederick Hilliers OR Series. Elsevier.

Mokkadem, A. and Pelletier, M. (2011).
A generalization of the averaging procedure: The use of two-time-scale algorithms.
SIAM Journal on Control and Optimization, 49:1523.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009).
Robust stochastic approximation approach to stochastic programming.
SIAM Journal on Optimization, 19(4):1574–1609.

Ortega, J. M. and Rheinboldt, W. C. (1970).
Iterative Solution of Nonlinear Equations in Several Variables.
Academic Press, New York, NY.

Pasupathy, R. and Ghosh, S. (2013).
Simulation optimization: A concise overview and implementation guide.

INFORMS TutORials. INFORMS.

Pasupathy, R., Glynn, P. W., Ghosh, S. G., and Hahemi, F. S. (2014).
How much to sample in simulation-based stochastic recursions?
Under Review.

Polyak, B. T. and Juditsky, A. B. (1992).
Acceleration of stochastic approximation by averaging.
SIAM Journal on Control and Optimization, 30(4):838–855.

Ruppert, D. (1985).
A Newton-Raphson version of the multivariate Robbins-Monro procedure.
Annals of Statistics, 13:236–245.

Ruppert, D. (1991).
Stochastic approximation.
Handbook in Sequential Analysis, pages 503–529. Dekker, New York, NY.

- Shapiro, A., Dentcheva, D., and Ruszczynski, A. (2009).
  Lectures on Stochastic Programming: Modeling and Theory.
  SIAM, Philadelphia, PA.

- Spall, J. C. (2003).
  Introduction to Stochastic Search and Optimization.
  John Wiley & Sons, Inc., Hoboken, NJ.

- Vaidya, R. and Bhatnagar, S. (2006).
  Robust optimization of random early detection.
  Telecommunication Systems, 33(4):291–316.

SCSR: How Much to Sample?

Theoretical Guidance

	Polynomial (λ_p, p)	Geometric(c)	Exponential(λ_t, t)
Sublinear(λ_s, s)	$p\alpha = 1 + s$: k^{-s} / $k^{-p\alpha+1}$	k^{-s}	k^{-s}
Linear(ℓ)	$k^{-p\alpha}$	$\ell = c^{-\alpha}$: ℓ^k / $c^{-\alpha k}$	ℓ^k
Superlinear(λ_q, q)	$k^{-p\alpha}$	$c^{-\alpha k}$	$p = t$: $c_2^{-\alpha p^k}$ / $c_1^{-\alpha t^k}$

Back