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).

June 24, 2016

Talk Overview

1. Problem Statement
2. Canonical Rates in Simulation Optimization
3. Stochastic Approximation; Sample Average Approximation (and its refinements)
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."

$$\begin{aligned} \text{minimize} \quad & f(x) = \mathbb{E}[F(x)] \\ \text{subject to} \quad & g(x) = \mathbb{E}[G(x)] \leq 0, x \in \mathcal{D}; \end{aligned}$$

- $f : \mathcal{D} \to \mathbb{R}$ (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

Simulation Optimization

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

$$\begin{aligned} \text{minimize} \quad & f(x) = \mathbb{E}[F(x)] \\ \text{subject to} \quad & g(x) = \mathbb{E}[G(x)] \leq 0, x \in \mathcal{D}; \end{aligned}$$

- $f : \mathcal{D} \to \mathbb{R}$ (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.

"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 $\mathbb{E}[X^2] < \infty$,

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

(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 $s = \Theta(m^{-1/6})$,

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

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

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

"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 $\mathbb{E}[X^2] < \infty$,

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

(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 $s = \Theta(m^{-1/6})$,

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

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

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

"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 $\mathbb{E}[X^2] < \infty$,

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

(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 $s = \Theta(m^{-1/6})$,

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

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

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

"Stochastic Complexity," Canonical Rates

Examples: ... contd.

(iii) Canonical rate for SO (with direct stochastic gradient observations):

$$\underbrace{\|X_k - x^*\|}_{\text{error}} = \mathcal{O}_p(1/\sqrt{W_k}),$$

where W_k is the *total* simulation effort by iteration k. (The rate will deteriorate if direct gradient observations are not available (Mokkadem and Pelletier, 2011).)

Stochastic and Sample-Average Approximation

1. SA (Robbins and Monro, 1951; Kiefer and Wolfowitz, 1952; Mokkadem and Pelletier, 2011; Djeddour et al., 2008; Polyak and Juditsky, 1992) is amongst the most used algorithms.

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

   (Google Scholar Hits: 200,000!)

2. "Canonical convergence" and simplicity are SA's strengths. (And, $\sqrt{k}(\overline{X} - x^*) \xrightarrow{d} \mathcal{N}(0, V)$.)
3. Automatic implementation has been challenging even after six decades of research (Pasupathy and Ghosh, 2013; Broadie et al., 2011).
4. Sample-Average Approximation (Shapiro et al., 2009; Kim et al., 2014) is not an algorithm.

Stochastic and Sample-Average Approximation

1. SA (Robbins and Monro, 1951; Kiefer and Wolfowitz, 1952; Mokkadem and Pelletier, 2011; Djeddour et al., 2008; Polyak and Juditsky, 1992) is amongst the most used algorithms.

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

   (Google Scholar Hits: 200,000!)

2. "Canonical convergence" and simplicity are SA's strengths. (And, $\sqrt{k}(\overline{X} - x^*) \xrightarrow{d} \mathcal{N}(0, V)$.)
3. Automatic implementation has been challenging even after six decades of research (Pasupathy and Ghosh, 2013; Broadie et al., 2011).
4. Sample-Average Approximation (Shapiro et al., 2009; Kim et al., 2014) is not an algorithm.

Stochastic and Sample-Average Approximation

1. SA (Robbins and Monro, 1951; Kiefer and Wolfowitz, 1952; Mokkadem and Pelletier, 2011; Djeddour et al., 2008; Polyak and Juditsky, 1992) is amongst the most used algorithms.

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

   (Google Scholar Hits: 200,000!)

2. "Canonical convergence" and simplicity are SA's strengths. (And, $\sqrt{k}(\overline{X} - x^*) \xrightarrow{d} \mathcal{N}(0, V)$.)
3. Automatic implementation has been challenging even after six decades of research (Pasupathy and Ghosh, 2013; Broadie et al., 2011).
4. Sample-Average Approximation (Shapiro et al., 2009; Kim et al., 2014) is not an algorithm.

Stochastic and Sample-Average Approximation

1. SA (Robbins and Monro, 1951; Kiefer and Wolfowitz, 1952; Mokkadem and Pelletier, 2011; Djeddour et al., 2008; Polyak and Juditsky, 1992) is amongst the most used algorithms.

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

   (Google Scholar Hits: 200,000!)

2. "Canonical convergence" and simplicity are SA's strengths. (And, $\sqrt{k}(\overline{X} - x^*) \xrightarrow{d} \mathcal{N}(0, V)$.)
3. Automatic implementation has been challenging even after six decades of research (Pasupathy and Ghosh, 2013; Broadie et al., 2011).
4. Sample-Average Approximation (Shapiro et al., 2009; Kim et al., 2014) is not an algorithm.

Sampling Controlled Stochastic Recursion (SCSR)

An Alternative to SA?

Two Crucial Ideas:

I.1 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?

I.2 SA "kills noise" using a parameter sequence; why not "kill noise" using sampling?

Loosely Stated Algorithmic Paradigm:

- Use a recursion (such as line search) as the underlying search mechanism;
- Sample judiciously and adaptively, more in the beginning and less later.

Sampling Controlled Stochastic Recursion (SCSR)

An Alternative to SA?

Two Crucial Ideas:

I.1 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?

I.2 SA "kills noise" using a parameter sequence; why not "kill noise" using sampling?

Loosely Stated Algorithmic Paradigm:

- Use a recursion (such as line search) as the underlying search mechanism;
- Sample judiciously and adaptively, more in the beginning and less later.

Sampling Controlled Stochastic Recursion (SCSR)

An Alternative to SA?

Two Crucial Ideas:

- I.1 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?
- I.2 SA "kills noise" using a parameter sequence; why not "kill noise" using sampling?

Loosely Stated Algorithmic Paradigm:

- Use a recursion (such as line search) as the underlying search mechanism;
- Sample judiciously and adaptively, more in the beginning and less later.

Sampling Controlled Stochastic Recursion (SCSR)

An Alternative to SA?

Two Crucial Ideas:

I.1 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?

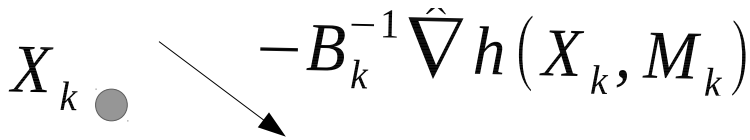
I.2 SA "kills noise" using a parameter sequence; why not "kill noise" using sampling?

Loosely Stated Algorithmic Paradigm:

- Use a recursion (such as line search) as the underlying search mechanism;
- Sample judiciously and adaptively, more in the beginning and less later.

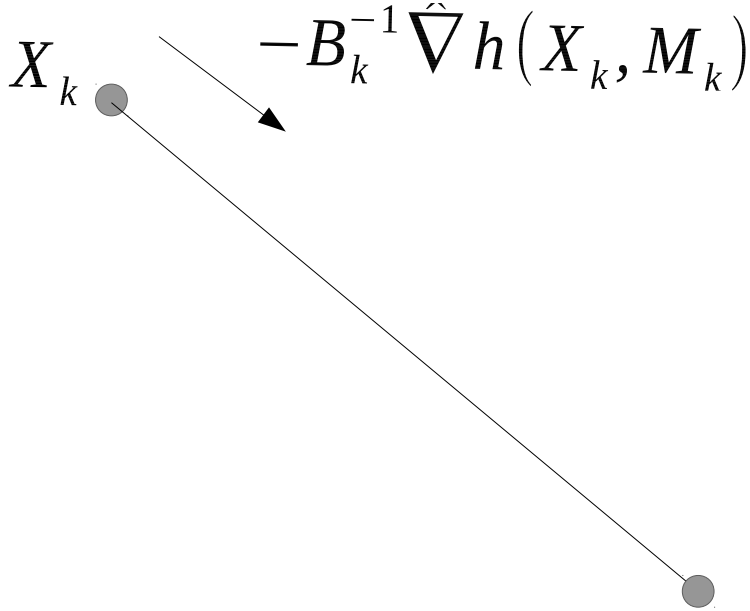
Adaptive SCSR: Line Search

Repeatedly perform an inexact line search with estimated gradient.



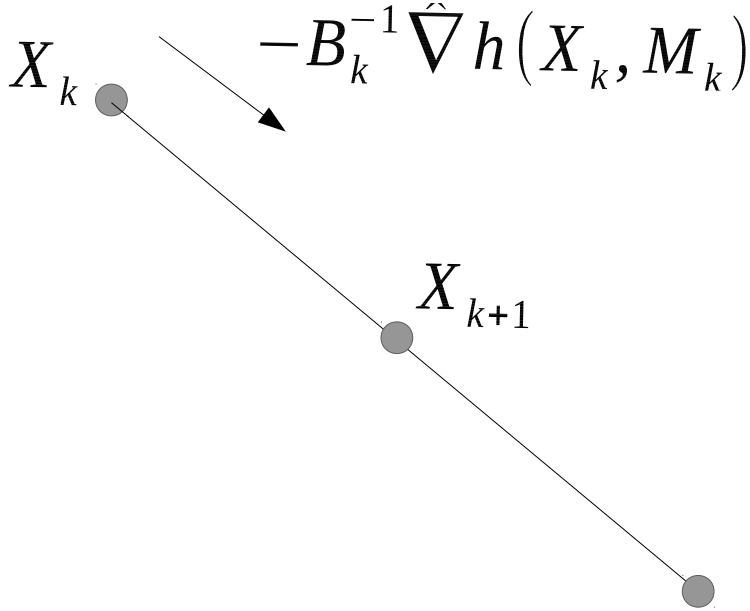
Adaptive SCSR: Line Search

Repeatedly perform an inexact line search with estimated gradient.



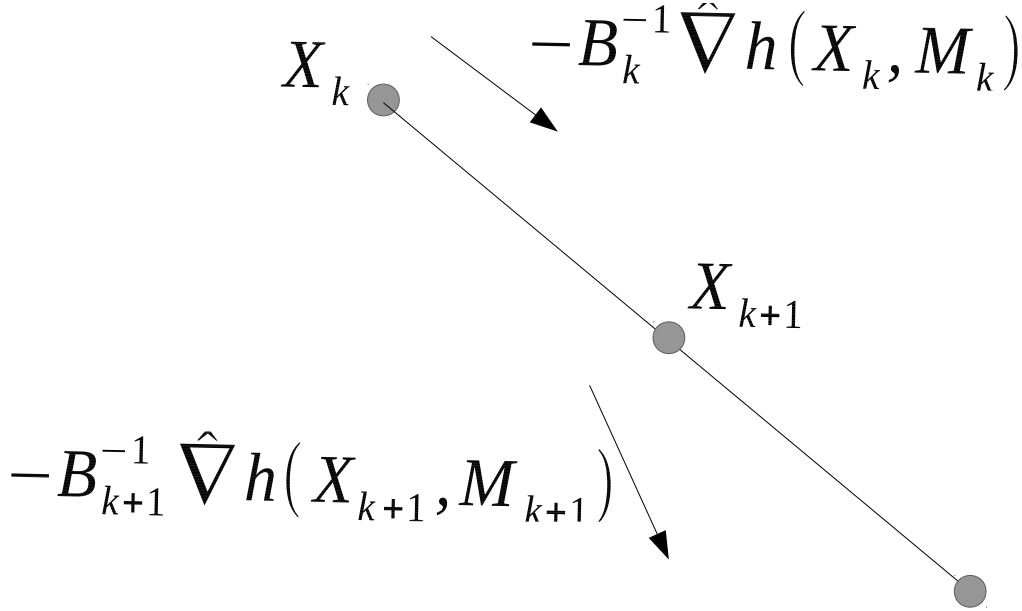
Adaptive SCSR: Line Search

Repeatedly perform an inexact line search with estimated gradient.



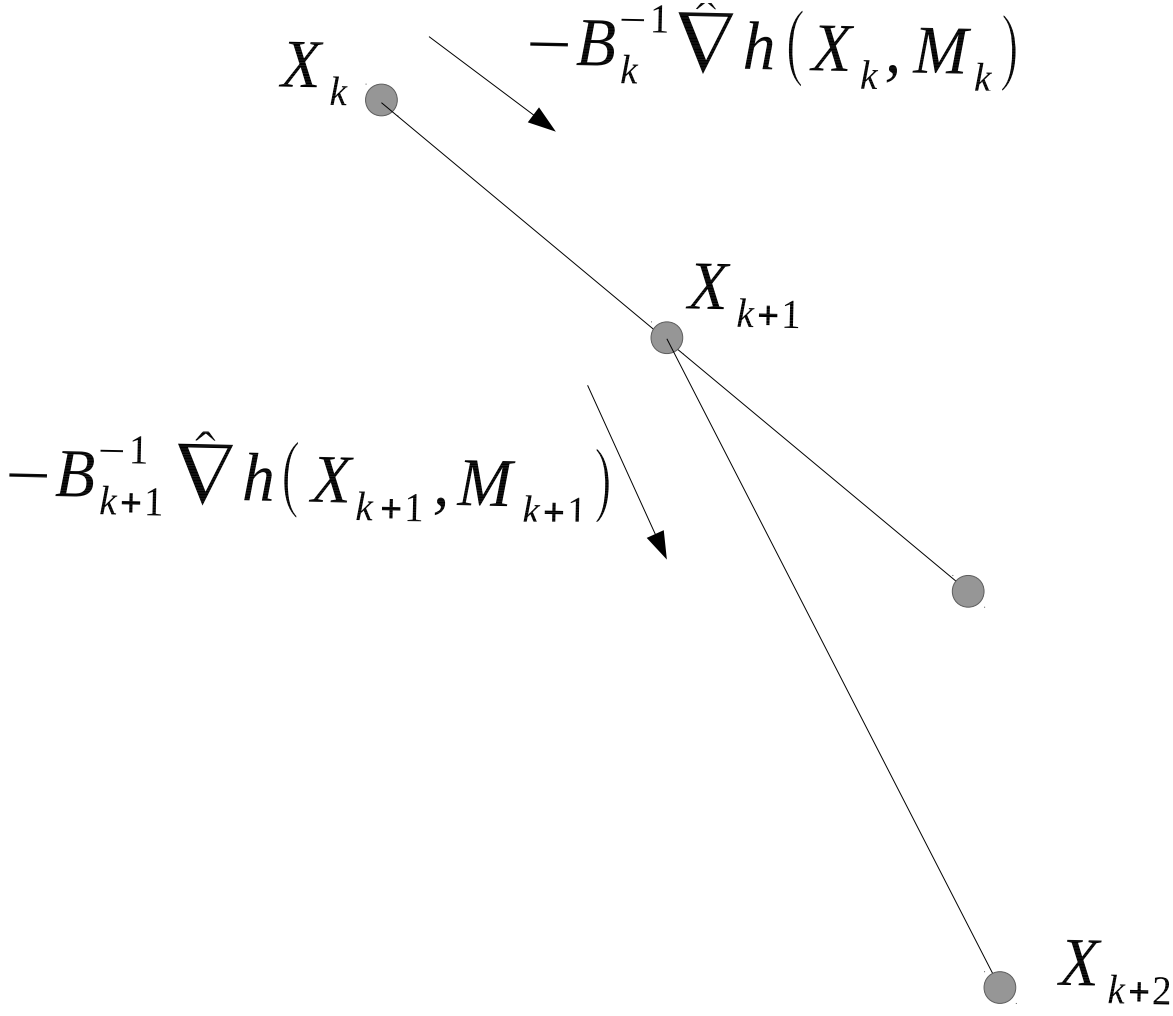
Adaptive SCSR: Line Search

Repeatedly perform an inexact line search with estimated gradient.



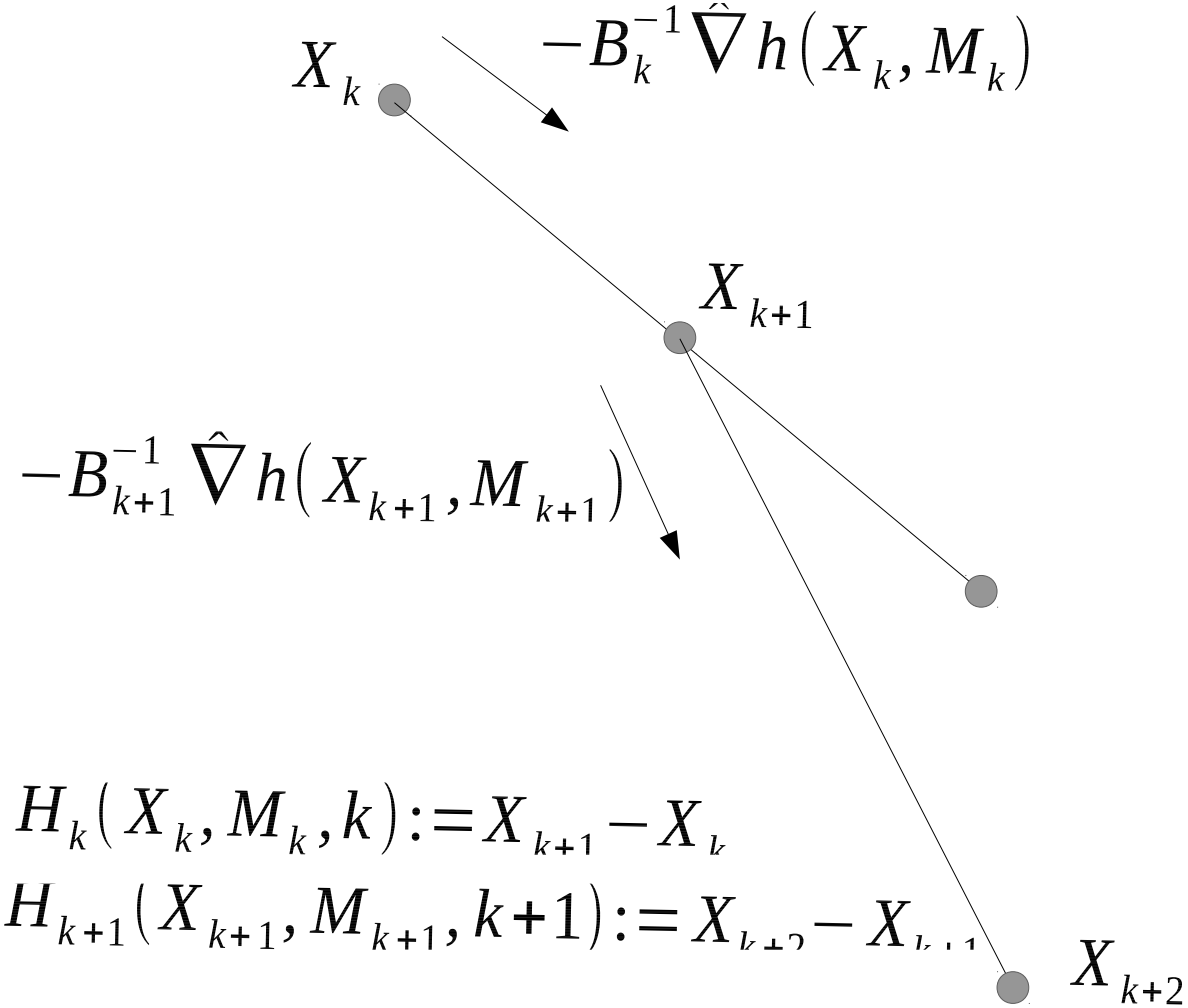
Adaptive SCSR: Line Search

Repeatedly perform an inexact line search with estimated gradient.



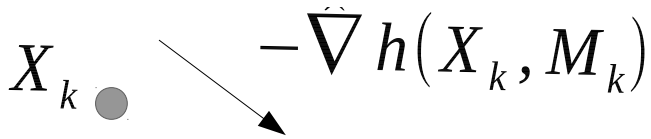
Adaptive SCSR: Line Search

Repeatedly perform an inexact line search with estimated gradient.



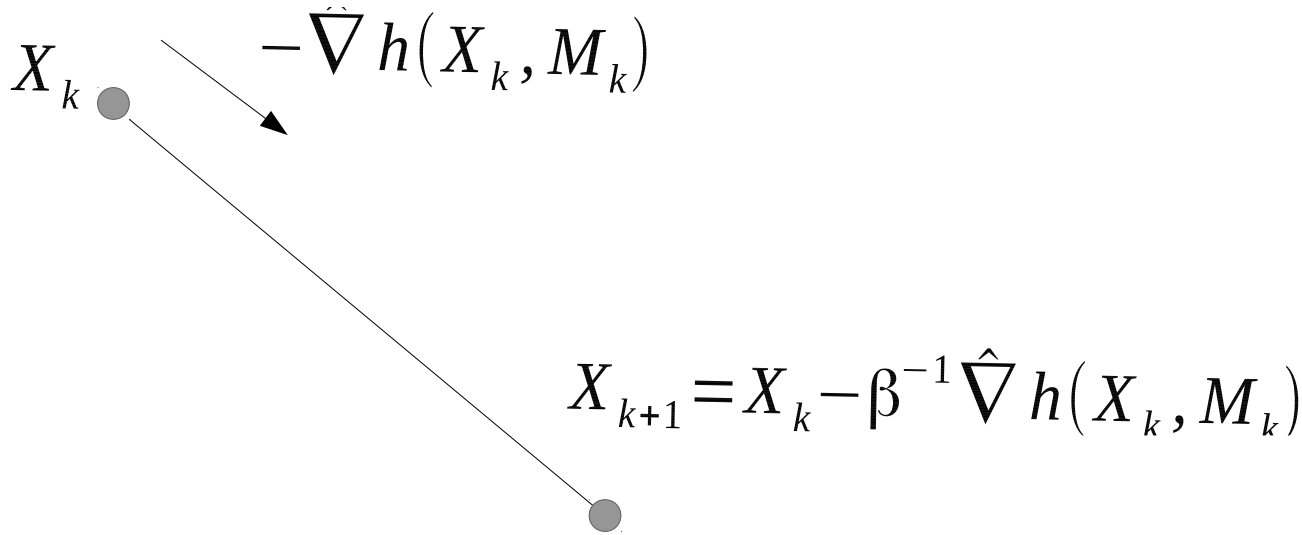
Adaptive SCSR: Gradient Search

Estimate gradient and take a fixed step.



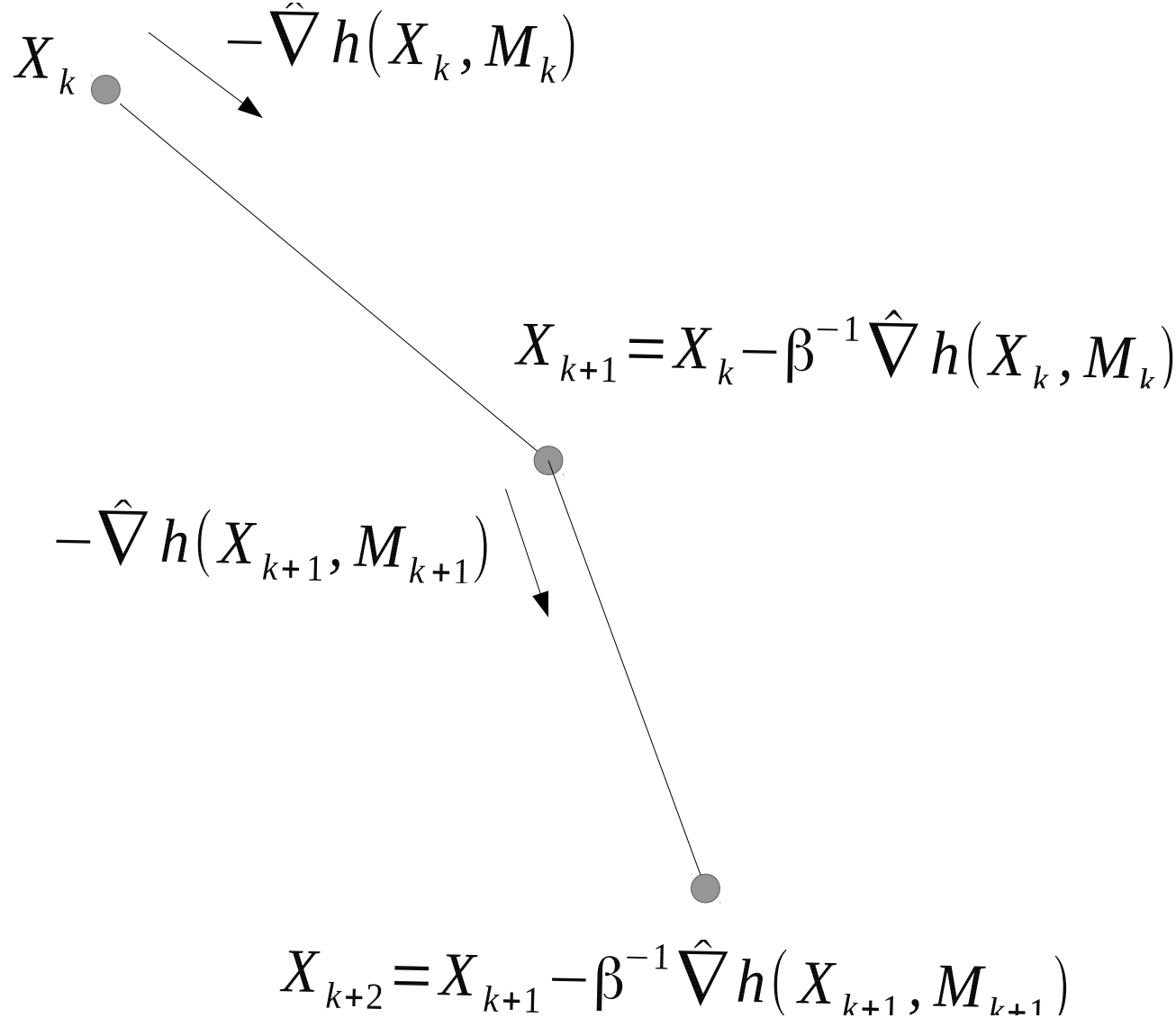
Adaptive SCSR: Gradient Search

Estimate gradient and take a fixed step.



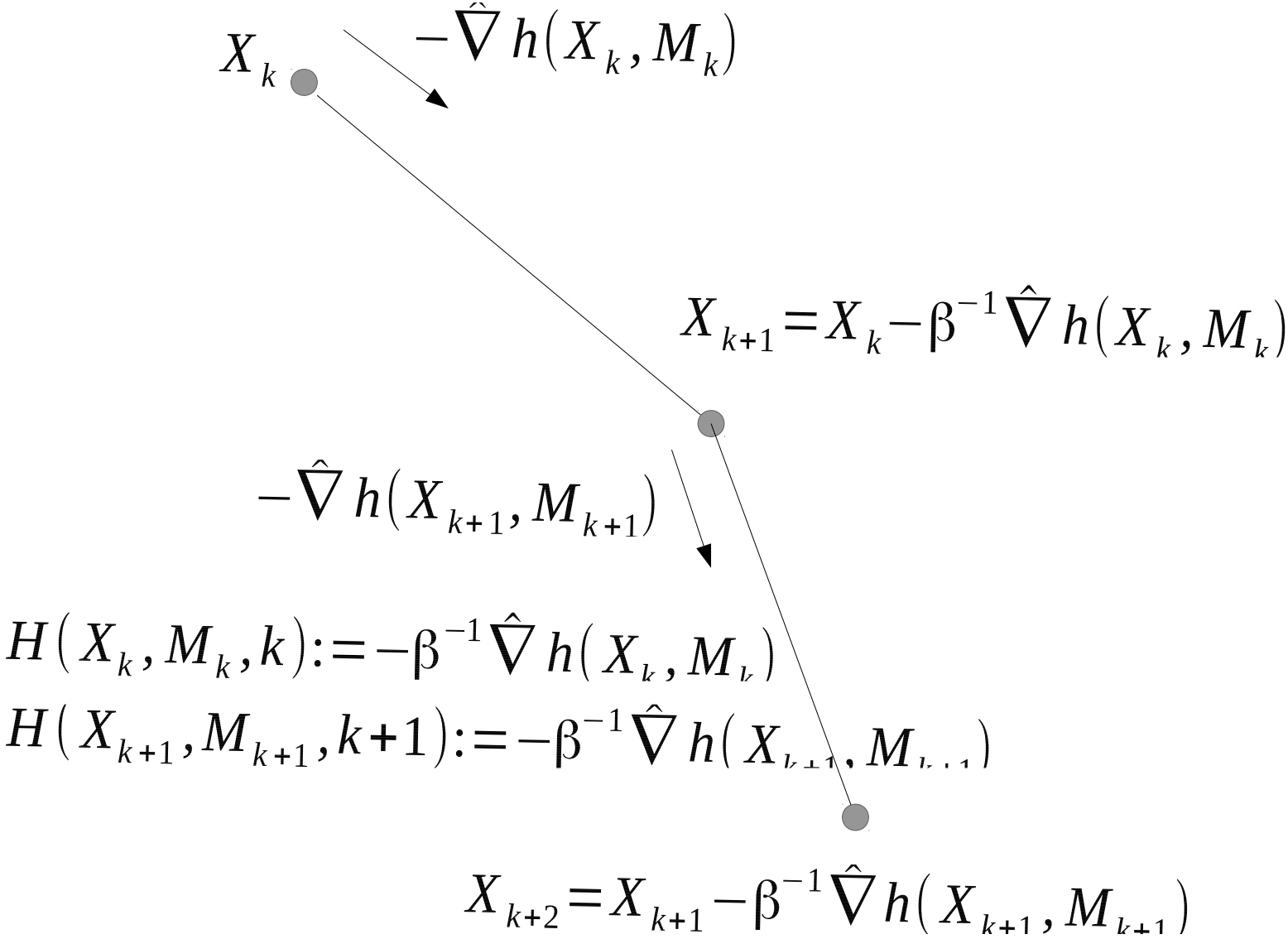
Adaptive SCSR: Gradient Search

Estimate gradient and take a fixed step.



Adaptive SCSR: Gradient Search

Estimate gradient and take a fixed step.



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. 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?

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.

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.

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?

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

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 constant $c > 0$.

Adaptive SCSR

Sample Size Determination

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

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

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 constant $c > 0$.

Adaptive SCSR

Sample Size Determination

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

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

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 constant $c > 0$.

Adaptive SCSR

Consider the adaptive sampling gradient method.

$$X_{k+1} = X_k - \beta^{-1} H(X_k, M_k, k), \quad k = 1, 2, \ldots, \quad \text{(SCSR)}$$

where H approximates the true gradient h. Assume the following.

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 $x \in \mathcal{D}$, $\ell_0 \|x - x^*\|_2^2 \leq h^T(x) h(x) \leq \ell_1 \|x - x^*\|_2^2$.

A.3 $H(x, m) \triangleq h(x) + \sum_{j=1}^{m} \xi_j(x)$, where $\xi_j(x)$ are iid copies of $\xi(x)$ and $\mathbb{E}[\xi(x)] = 0$.

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

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

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

Adaptive SCSR

Theoretical Results: Some Intuition on Iteration Evolution

Theorem (Polyak (1987))

Let $\{Y_k\}$ *be nonnegative random variables, where* $\mathbb{E}[Y_0] < \infty$, *and let* $\{\alpha_k\}$, $\{\beta_k\}$ *be deterministic scalar sequences such that*

$$\mathbb{E}[Y_{k+1}|Y_0, \cdots, Y_k] \leq (1 - \alpha_k)Y_k + \beta_k \overset{a.s.}{\to} \textit{for} k \geq 0,$$

where $0 \leq \alpha_k \leq 1$, $\beta_k \geq 0$, $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\sum_{k=0}^{\infty} \beta_k < \infty$, $\lim_{k\to\infty} \frac{\beta_k}{\alpha_k} = 0$.

Then $Y_k \overset{a.s.}{\to} 0$ *and* $\lim_{k\to\infty} \mathbb{E}[Y_k] = 0$.

Theorem (Consistency)

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

Adaptive SCSR

Theoretical Results: Some Intuition on Iteration Evolution

Theorem (Polyak (1987))

Let $\{Y_k\}$ *be nonnegative random variables, where* $\mathbb{E}[Y_0] < \infty$, *and let* $\{\alpha_k\}$, $\{\beta_k\}$ *be deterministic scalar sequences such that*

$$\mathbb{E}[Y_{k+1}|Y_0, \cdots, Y_k] \leq (1 - \alpha_k)Y_k + \beta_k \overset{a.s.}{\to} \textit{for} k \geq 0,$$

where $0 \leq \alpha_k \leq 1$, $\beta_k \geq 0$, $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\sum_{k=0}^{\infty} \beta_k < \infty$, $\lim_{k\to\infty} \frac{\beta_k}{\alpha_k} = 0$.

Then $Y_k \overset{a.s.}{\to} 0$ *and* $\lim_{k\to\infty} \mathbb{E}[Y_k] = 0$.

Theorem (Consistency)

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

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

A Fundamental Theorem for Adaptive Sampling

Theorem (Adaptive Sampling Theorem)

Let $X_1, X_2, \ldots$ be iid copies of X with $\mu := \mathbb{E}[X]$, $\sigma^2 := \mathbb{V}ar(X)$. For $\epsilon \in (0, 1/2)$, let:

$$M := \inf_{m \geq \nu} \left\{ m^{\epsilon} \frac{\hat{\sigma}(m)}{\sqrt{m}} < c|\overline{X}(m)| \right\}; \; n^* := \max\{\nu, (\frac{\sigma}{\mu})^{\eta}\}, \eta = \frac{2}{1-2\epsilon}.$$

The following assertions hold.

T.1 *M is well-defined, that is, $\mathbb{P}(M < \infty) = 1$.*

T.2 *If $\mathbb{E}[|X-\mu|^{4(1+\epsilon)}] < \infty$, for identifiable $v_1, v_2 \in (0, \infty)$, $\mathbb{E}[M] \leq v_1 n^* + v_2$.*

T.3 *For identifiable v,*

$$\mathbb{E}[|\overline{X}(M) - \mu|^2] \leq \frac{v}{n^*}.$$

Adaptive SCSR

A Fundamental Theorem for Adaptive Sampling

Theorem (Adaptive Sampling Theorem)

Let $X_1, X_2, \ldots$ be iid copies of X with $\mu := \mathbb{E}[X]$, $\sigma^2 := \mathbb{V}ar(X)$. For $\epsilon \in (0, 1/2)$, let:

$$M := \inf_{m \geq \nu} \left\{ m^{\epsilon} \frac{\hat{\sigma}(m)}{\sqrt{m}} < c|\overline{X}(m)| \right\}; \; n^* := \max\{\nu, (\frac{\sigma}{\mu})^{\eta}\}, \eta = \frac{2}{1-2\epsilon}.$$

The following assertions hold.

T.1 *M is well-defined, that is, $\mathbb{P}(M < \infty) = 1$.*

T.2 *If $\mathbb{E}[|X-\mu|^{4(1+\epsilon)}] < \infty$, for identifiable $v_1, v_2 \in (0, \infty)$, $\mathbb{E}[M] \leq v_1 n^* + v_2$.*

T.3 *For identifiable v,*

$$\mathbb{E}[|\overline{X}(M) - \mu|^2] \leq \frac{v}{n^*}.$$

Adaptive SCSR

A Fundamental Theorem for Adaptive Sampling

Theorem (Adaptive Sampling Theorem)

Let $X_1, X_2, \ldots$ be iid copies of X with $\mu := \mathbb{E}[X]$, $\sigma^2 := \mathbb{V}ar(X)$. For $\epsilon \in (0, 1/2)$, let:

$$M := \inf_{m \geq \nu} \left\{ m^{\epsilon} \frac{\hat{\sigma}(m)}{\sqrt{m}} < c|\overline{X}(m)| \right\}; \; n^* := \max\{\nu, (\frac{\sigma}{\mu})^{\eta}\}, \eta = \frac{2}{1-2\epsilon}.$$

The following assertions hold.

T.1 *M is well-defined, that is, $\mathbb{P}(M < \infty) = 1$.*

T.2 *If $\mathbb{E}[|X-\mu|^{4(1+\epsilon)}] < \infty$, for identifiable $v_1, v_2 \in (0, \infty)$, $\mathbb{E}[M] \leq v_1 n^* + v_2$.*

T.3 *For identifiable v,*

$$\mathbb{E}[|\overline{X}(M) - \mu|^2] \leq \frac{v}{n^*}.$$

Adaptive SCSR

A Fundamental Theorem for Adaptive Sampling

Theorem (Adaptive Sampling Theorem)

Let $X_1, X_2, \ldots$ be iid copies of X with $\mu := \mathbb{E}[X]$, $\sigma^2 := \mathbb{V}ar(X)$. For $\epsilon \in (0, 1/2)$, let:

$$M := \inf_{m \geq \nu} \left\{ m^{\epsilon} \frac{\hat{\sigma}(m)}{\sqrt{m}} < c|\overline{X}(m)| \right\}; \; n^* := \max\{\nu, (\frac{\sigma}{\mu})^{\eta}\}, \eta = \frac{2}{1-2\epsilon}.$$

The following assertions hold.

T.1 *M is well-defined, that is, $\mathbb{P}(M < \infty) = 1$.*

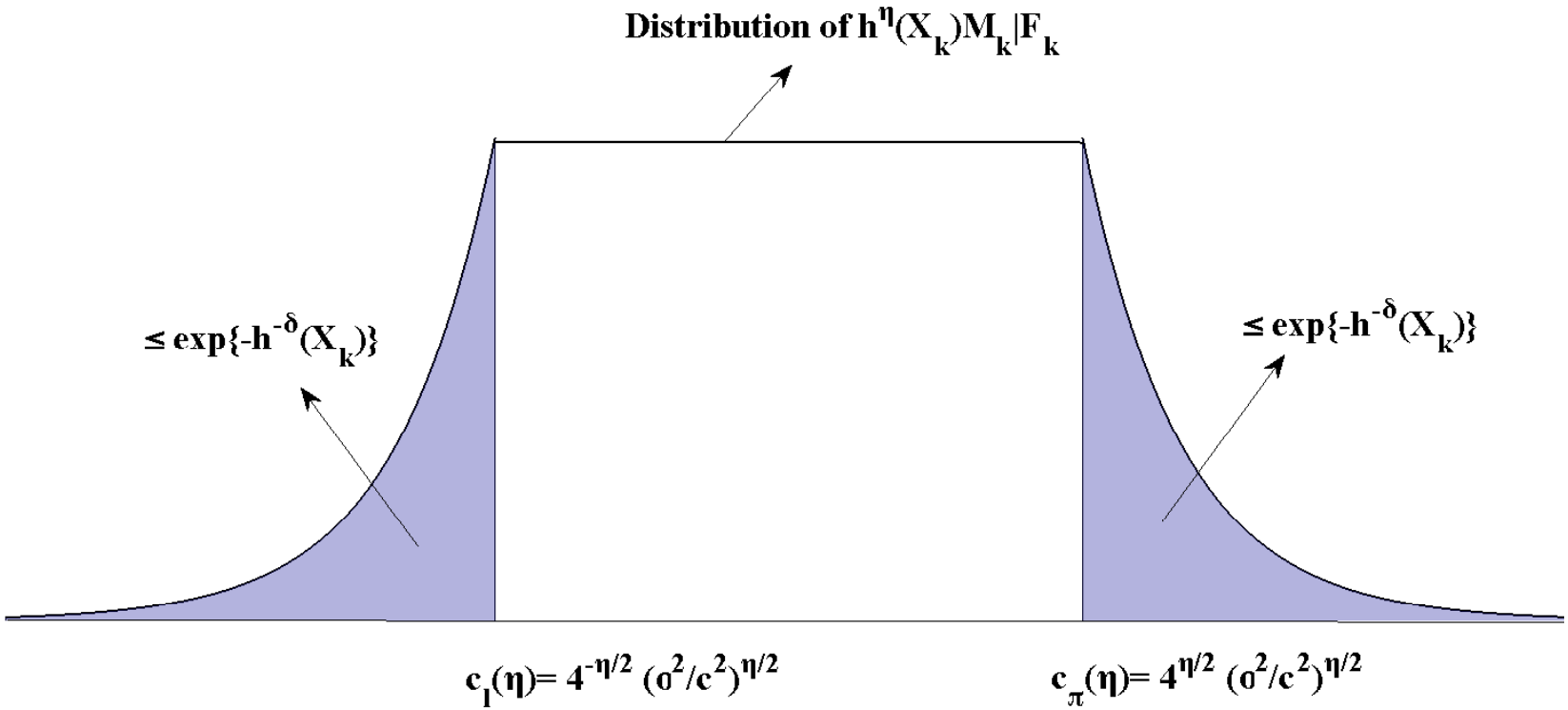
T.2 *If $\mathbb{E}[|X-\mu|^{4(1+\epsilon)}] < \infty$, for identifiable $v_1, v_2 \in (0, \infty)$, $\mathbb{E}[M] \leq v_1 n^* + v_2$.*

T.3 *For identifiable v,*

$$\mathbb{E}[|\overline{X}(M) - \mu|^2] \leq \frac{v}{n^*}.$$

Adaptive SCSR

A Fundamental Theorem for Adaptive Sampling



Adaptive SCSR

Theoretical Results: Some Intuition on Iteration Evolution

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

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

$$\begin{aligned}
&\mathbb{E}_\Omega[Z_{k+1}^2 | \mathcal{F}_k] \\
&\leq \left(1 - \frac{2\ell_0}{\beta} + \frac{\ell_1^2}{\beta^2}\right) Z_k^2 + \frac{1}{\beta^2} \mathbb{E}_\Omega[\|H(X_k, M_k) - h(X_k)\|^2 | \mathcal{F}_k] \\
&\leq \left(1 - \frac{2\ell_0}{\beta} + \frac{\ell_1^2}{\beta^2}\right) Z_k^2 + \frac{1}{\beta^2} \frac{\tilde{c}}{n_k^*} \\
&\leq \left(1 - \frac{2\ell_0}{\beta} + \frac{\ell_1^2}{\beta^2}(1 + \tilde{c}\alpha^{-2})\right) Z_k^2.
\end{aligned}$$

Linear Convergence of the error's L_2! How much work have we done though?

Adaptive SCSR

Theoretical Results: Some Intuition on Iteration Evolution

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

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

$$\begin{aligned}
&\mathbb{E}_\Omega[Z_{k+1}^2 | \mathcal{F}_k] \\
&\leq \left(1 - \frac{2\ell_0}{\beta} + \frac{\ell_1^2}{\beta^2}\right) Z_k^2 + \frac{1}{\beta^2} \mathbb{E}_\Omega[\|H(X_k, M_k) - h(X_k)\|^2 | \mathcal{F}_k] \\
&\leq \left(1 - \frac{2\ell_0}{\beta} + \frac{\ell_1^2}{\beta^2}\right) Z_k^2 + \frac{1}{\beta^2} \frac{\tilde{c}}{n_k^*} \\
&\leq \left(1 - \frac{2\ell_0}{\beta} + \frac{\ell_1^2}{\beta^2}(1 + \tilde{c}\alpha^{-2})\right) Z_k^2.
\end{aligned}$$

Linear Convergence of the error's L_2! How much work have we done though?

Adaptive SCSR

Theoretical Results: Some Intuition on Iteration Evolution

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

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

$$\begin{aligned}
&\mathbb{E}_\Omega[Z_{k+1}^2 | \mathcal{F}_k] \\
&\leq \left(1 - \frac{2\ell_0}{\beta} + \frac{\ell_1^2}{\beta^2}\right) Z_k^2 + \frac{1}{\beta^2} \mathbb{E}_\Omega[\|H(X_k, M_k) - h(X_k)\|^2 | \mathcal{F}_k] \\
&\leq \left(1 - \frac{2\ell_0}{\beta} + \frac{\ell_1^2}{\beta^2}\right) Z_k^2 + \frac{1}{\beta^2} \frac{\tilde{c}}{n_k^*} \\
&\leq \left(1 - \frac{2\ell_0}{\beta} + \frac{\ell_1^2}{\beta^2}(1 + \tilde{c}\alpha^{-2})\right) Z_k^2.
\end{aligned}$$

Linear Convergence of the error's L_2! How much work have we done though?

Adaptive SCSR

Theoretical Results: Some Intuition on Iteration Evolution

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

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

$$\begin{aligned}
&\mathbb{E}_\Omega[Z_{k+1}^2 | \mathcal{F}_k] \\
&\leq \left(1 - \frac{2\ell_0}{\beta} + \frac{\ell_1^2}{\beta^2}\right) Z_k^2 + \frac{1}{\beta^2}\mathbb{E}_\Omega[\|H(X_k, M_k) - h(X_k)\|^2 | \mathcal{F}_k] \\
&\leq \left(1 - \frac{2\ell_0}{\beta} + \frac{\ell_1^2}{\beta^2}\right) Z_k^2 + \frac{1}{\beta^2}\frac{\tilde{c}}{n_k^*} \\
&\leq \left(1 - \frac{2\ell_0}{\beta} + \frac{\ell_1^2}{\beta^2}(1 + \tilde{c}\alpha^{-2})\right) Z_k^2.
\end{aligned}$$

Linear Convergence of the error's L_2! How much work have we done though?

Adaptive SCSR

Theoretical Results: Efficiency

Theorem (Canonical Rate)

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* $W_k^{1-2\epsilon}\|X_k - x^*\|^2 \xrightarrow{p} \infty$.

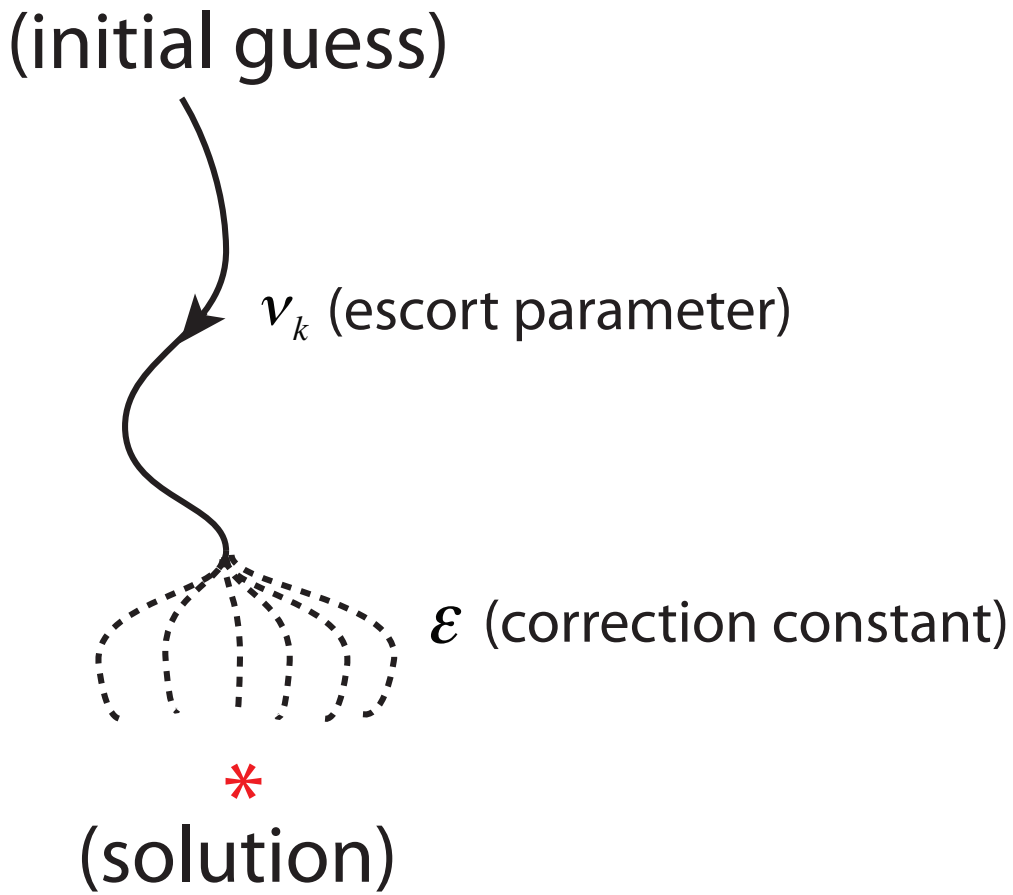
1. The result says that the mean squared error $\mathbb{E}[\|X_k - x^*\|^2] \approx (\mathbb{E}[W_k])^{-1}$, coinciding with the estimation rate.
2. Sampling should be atleast "geometric," irrespective of error!

Adaptive SCSR

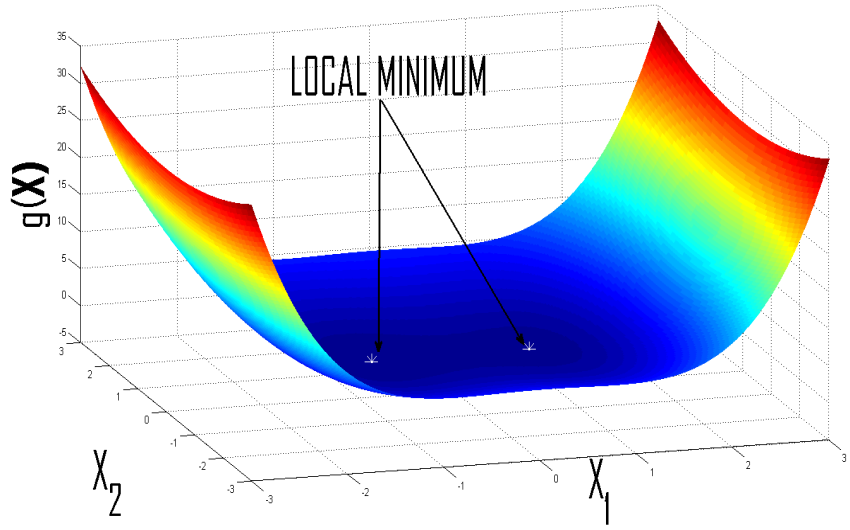
The Escort Sequence and the Coercion Constant

Theorem

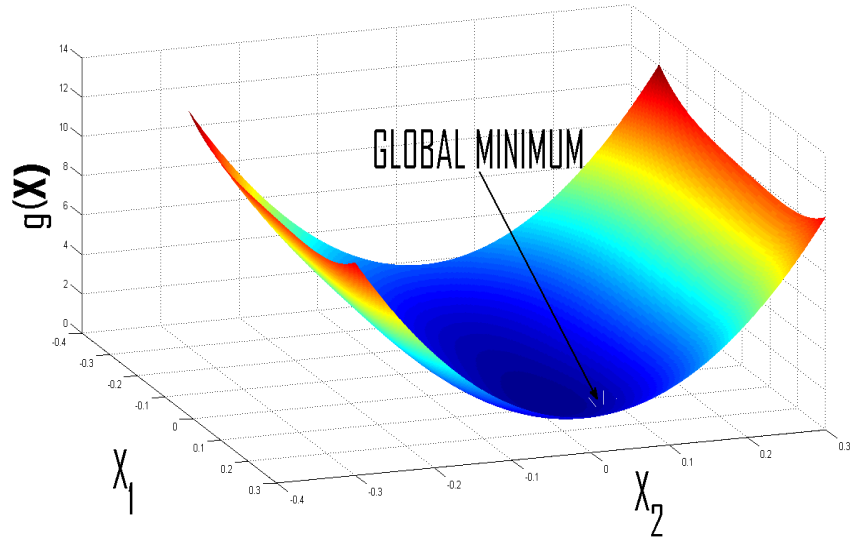
Let $W_k = \sum_j M_j$ denote the total simulation effort after k iterations. Then, $\mathbb{P}\{M_k = \nu_k \quad i.o.\} = 0$.



Numerical Illust



AluffiPentini Function

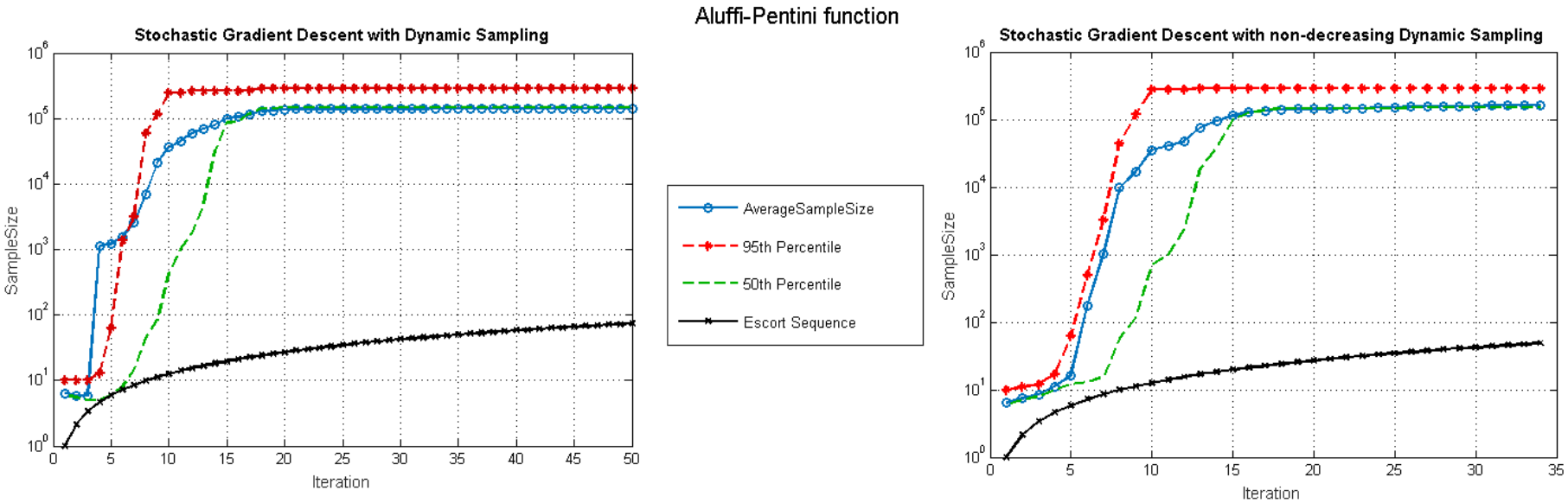


Rosenbrock Function

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

Numerical Illustration

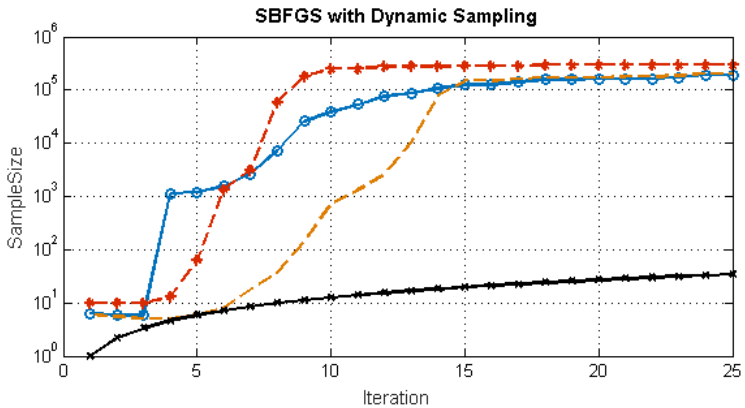
Sample Size Behavior

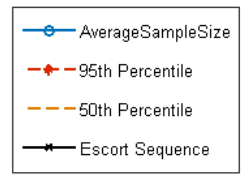


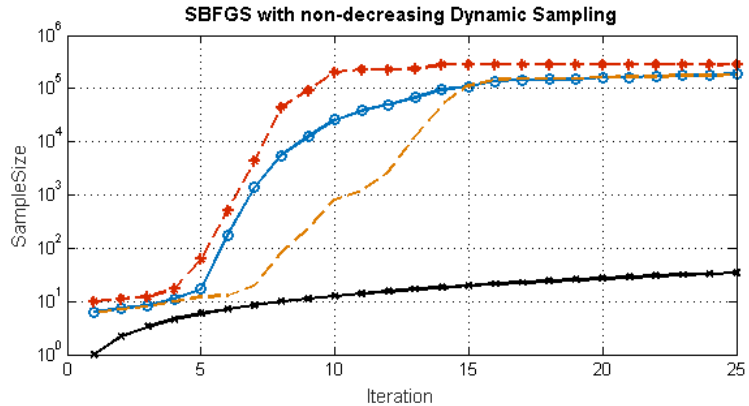
Numerical Illustration

Sample Size Behavior

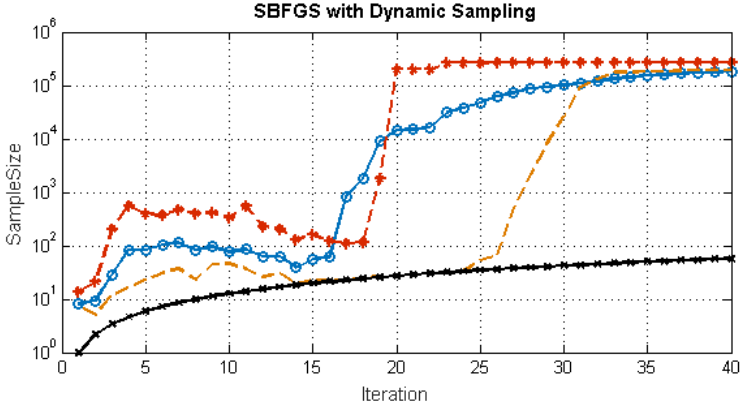
Aluffi-Pentini function

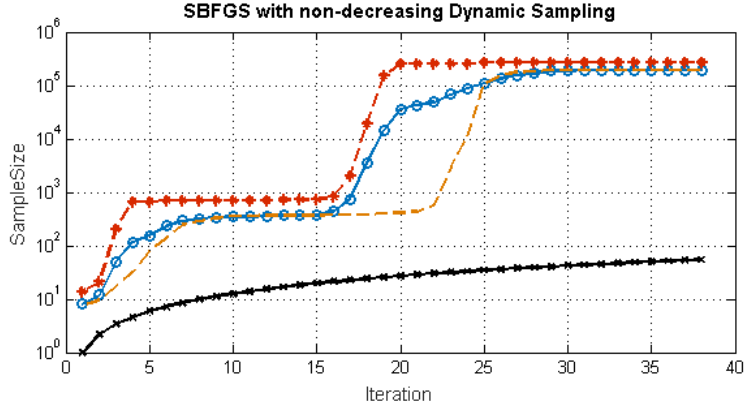




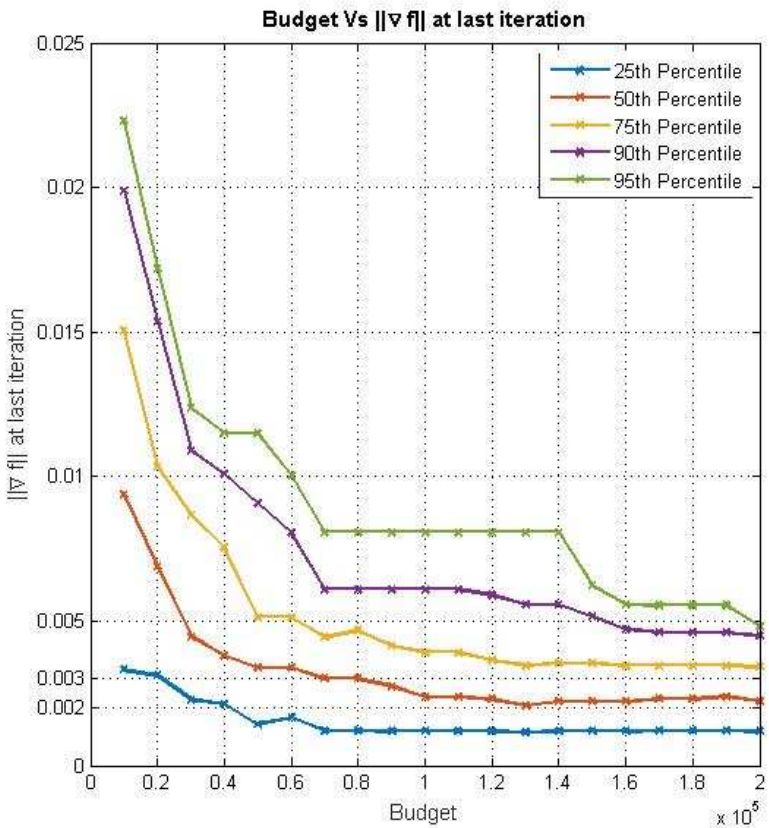


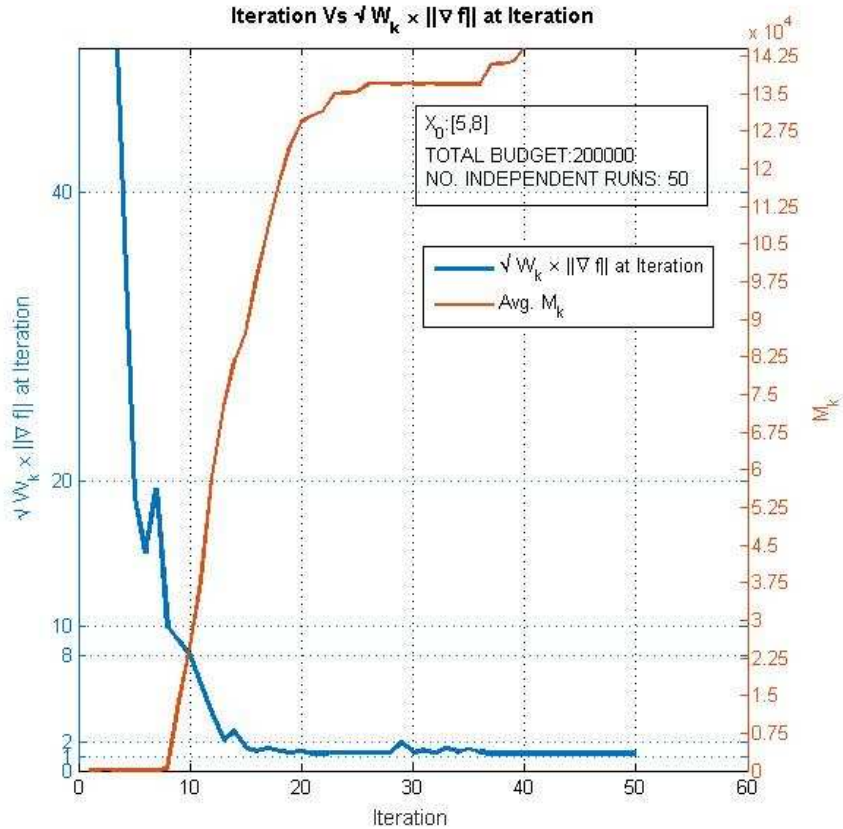
Rosenbrock function





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."
   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.

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.

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.

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.

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.

Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23:462–466.

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.

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

Polyak, B. T. (1987). *Introduction to optimization*. Optimization Software New York.

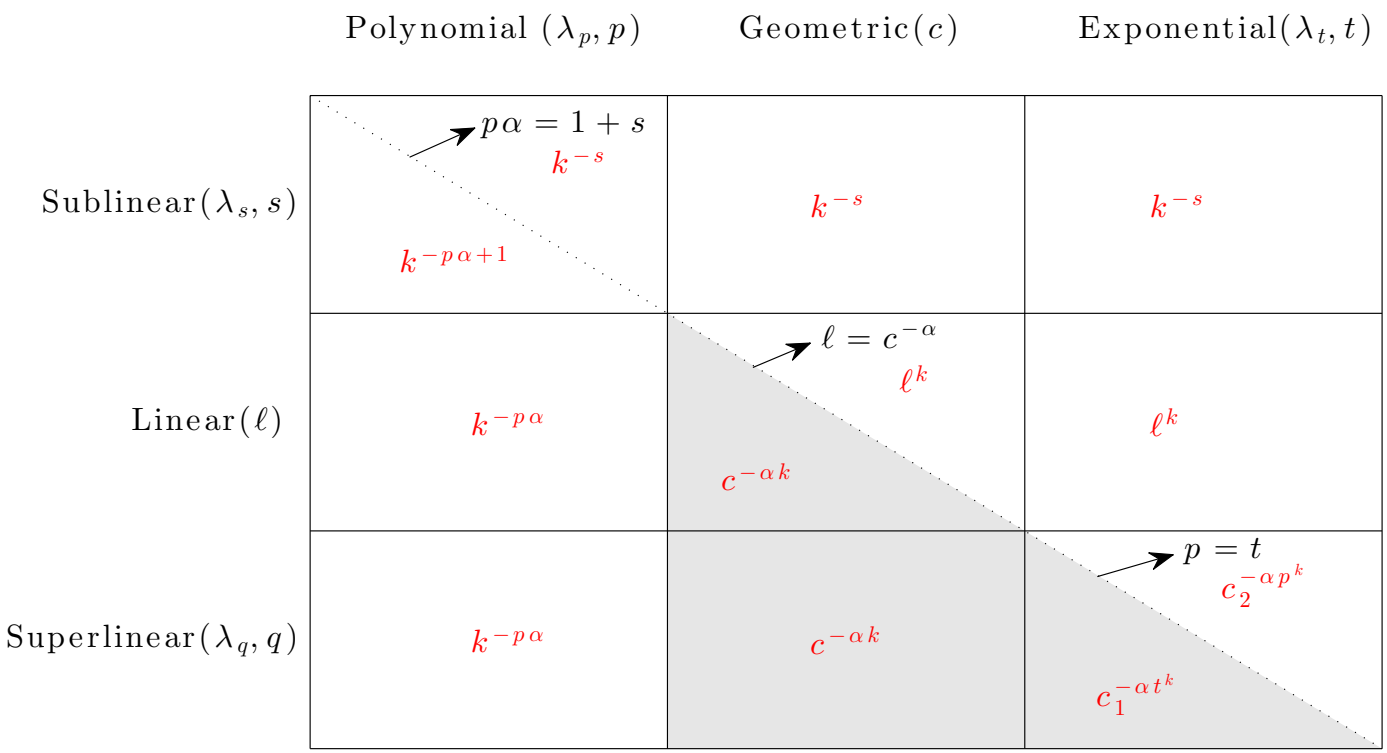
Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407.

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

SCSR: How Much to Sample?

Theoretical Guidance



Back