EXCERPT—CHAPTER 7
SIMULTANEOUS PERTURBATION
STOCHASTIC APPROXIMATION (SPSA)

• Organization of chapter in ISSO
  – Problem setting
  – SPSA algorithm
  – Theoretical foundation
  – Asymptotic normality and efficiency
  – Practical guidelines—MATLAB code
  – Numerical examples
  – Extensions and further results
  – Adaptive simultaneous perturbation method

Additional information available at www.jhuapl.edu/SPSA
(reference list, background articles, MATLAB code, and video)
Consider standard minimization setting, i.e., find root $\theta^*$ to

$$
g(\theta) = \frac{\partial L(\theta)}{\partial \theta} = 0
$$

where $L(\theta)$ is scalar-valued loss function to be minimized and $\theta$ is $p$-dimensional vector

- Assume only (possibly noisy) measurements of $L(\theta)$ available
  - No direct measurements of $g(\theta)$ used, as are required in stochastic gradient methods

- Noisy measurements of $L(\theta)$ in areas such as Monte Carlo simulation, real-time control/estimation, etc.

- Interested in $p > 1$ setting (including $p \gg 1$)
SPSA Algorithm

• Let $\hat{g}_k(\theta)$ denote SP estimate of $g(\theta)$ at $k$th iteration
• Let $\hat{\theta}_k$ denote estimate for $\theta^*$ at $k$th iteration
• SPSA algorithm has form

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{g}_k(\hat{\theta}_k)$$

where \{a_k\} is nonnegative gain sequence

• Generic iterative form above is standard in SA; stochastic analogue to steepest descent

• Under conditions, $\hat{\theta}_k \rightarrow \theta^*$ in some stochastic sense as $k \rightarrow \infty$
Computation of $\hat{g}_k(\bullet)$ (Heart of SPSA)

- Let $\Delta_k$ be vector of $p$ independent random variables at $k$th iteration
  $$\Delta_k = \left[ \Delta_{k1}, \Delta_{k2}, \ldots, \Delta_{kp} \right]^T$$

- $\Delta_k$ typically generated by Monte Carlo
- Let $\{c_k\}$ be sequence of positive scalars
- For iteration $k \rightarrow k+1$, take measurements at design levels: $\hat{\theta}_k \pm c_k \Delta_k$

$$y(\hat{\theta}_k + c_k \Delta_k) = L(\hat{\theta}_k + c_k \Delta_k) + \varepsilon_k^{(+)}$$
$$y(\hat{\theta}_k - c_k \Delta_k) = L(\hat{\theta}_k - c_k \Delta_k) + \varepsilon_k^{(-)}$$

where $\varepsilon_k^{(\pm)}$ are measurement noise terms

- Common special case is when $\varepsilon_k^{(\pm)} = 0 \ \forall \ k$
  (e.g., system identification with perfect measurements of the likelihood function)
Computation of $\hat{g}_k(\cdot)$ (cont’d)

- The standard SP form for $\hat{g}_k(\cdot)$:

$$
\hat{g}_k(\hat{\theta}_k) = \begin{bmatrix}
\frac{y(\hat{\theta}_k + c_k \Delta_k) - y(\hat{\theta}_k - c_k \Delta_k)}{2c_k \Delta_k_1} \\
\vdots \\
\frac{y(\hat{\theta}_k + c_k \Delta_k) - y(\hat{\theta}_k - c_k \Delta_k)}{2c_k \Delta_k_p}
\end{bmatrix}
$$

- Note that $\hat{g}_k(\cdot)$ only requires two measurements of $L(\cdot)$ independent of $p$

- Above SP form contrasts with standard finite-difference approximations taking $2p$ (or $p+1$) measurements

- Intuitive reason why $\hat{g}_k(\cdot)$ is appropriate is that $E[\hat{g}_k(\hat{\theta}_k) | \hat{\theta}_k] \approx g(\hat{\theta}_k)$; formalized in Section B
Essential Conditions for SPSA

• To use SPSA, there are regularity conditions on $L(\theta)$, choice of $\Delta_k$, the gain sequences $\{a_k\}$, $\{c_k\}$, and the measurement noise
  – Sections 7.3 and 7.4 of ISSO present essential conditions
• Roughly speaking the conditions are:
  A. $L(\theta)$ smoothness: $L(\theta)$ is thrice differentiable function (can be relaxed—see Section 7.3 of ISSO)
  B. Choice of $\Delta_k$ distribution: For all $k$, $\Delta_k$ has independent components, symmetrically distributed around 0, and $E(\Delta_{ki}^2) < \infty$, $E(\Delta_{ki}^{-2}) < \infty$
    – Bounded inverse moments condition is critical (excludes $\Delta_{ki}$ being normally or uniformly distributed)
    – Symmetric Bernoulli $\Delta_{ki} = \pm 1$ (prob = $\frac{1}{2}$ for each outcome) is allowed; asymptotically optimal (see Section F or Section 7.7 of ISSO)
C. **Gain sequences:** standard SA conditions:

\[ a_k, c_k > 0, a_k, c_k \to 0 \text{ as } k \to \infty \]

\[ \sum_{k=0}^{\infty} a_k = \infty, \sum_{k=0}^{\infty} \left( \frac{a_k}{c_k} \right)^2 < \infty \]

(better to violate some of these gain conditions in certain practical problems; e.g., nonstationary tracking and control where \( a_k = a > 0, c_k = c > 0 \ \forall \ k, i \))

D. **Measurement Noise:** Martingale difference

\[ E[\varepsilon_k^{(+) - \varepsilon_k^{(-)}} | \hat{\theta}_k, \Delta_k] = 0 \]

\( \forall \ k \) sufficiently large. (Noises not required to be independent of each other or of current/previous \( \hat{\theta}_k \) and \( \Delta_k \) values.) **Alternative** condition (no martingale mean 0 assumption needed) is that \( \varepsilon_k^{(\pm)} \) be bounded \( \forall \ k \)
B. THEORETICAL FOUNDATION

Three Questions

**Question 1:** Is \( \hat{g}_k(\cdot) \) a valid estimator for \( g(\cdot) \)?

**Answer:** Yes, under modest conditions.

**Question 2:** Will the algorithm converge to \( \theta^* \)?

**Answer:** Yes, under reasonable conditions.

**Question 3:** Do savings in data/iteration lead to a corresponding savings in converging to optimum?

**Answer:** Yes, under reasonable conditions.
Theoretical Basis (Sects. 7.3 – 7.4 of ISSO)

- Under appropriate regularity conditions (e.g., $E(\Delta_{ki}^{-2}) < \infty, L(\theta)$ thrice continuously differentiable, $\epsilon_k^{(\pm)}$ is martingale difference noise, etc.), we have:

- **Near Unbiasedness**
  
  $$E[\hat{g}_k(\hat{\theta}_k) | \hat{\theta}_k] = g(\hat{\theta}_k) + O(c_k^2) \text{ a.s.}$$

  where $c_k \to 0$

- **Convergence:**
  
  $$\hat{\theta}_k \to \theta^* \text{ a.s. as } k \to \infty$$

- **Asymptotic Normality:**
  
  $$k^{\beta/2} (\hat{\theta}_k - \theta^*) \xrightarrow{\text{dist.}} N(\mu, \Sigma), \quad 0 < \beta \leq \frac{2}{3}$$

  where $\mu, \Sigma,$ and $\beta$ depend on SA gains, $\Delta_k$ distribution, and shape of $L(\theta)$
Efficiency Analysis

• Can use asymptotic normality to analyze relative efficiency of SPSA and FDSA (Spall, 1992; Sect. 7.4 of ISSO)

• Analogous to SPSA asymptotic normality result, FDSA is also asymptotically normal (Chap. 6 of ISSO)

The critical cost in comparing relative efficiency of SPSA and FDSA is number of loss function measurements $y(\cdots)$, not number of iterations per se

• Loss function measurements represent main cost (by far)—other costs are trivial

• Full efficiency story is fairly complex—see Section 7.4 of ISSO and references therein
Efficiency Analysis (cont’d)

• Will compare SPSA and FDSA by looking at relative mean square error (MSE) of θ estimate

• Consider relative MSE for same no. of measurements, \( n \) (not same no. of iterations). Under regularity conditions above:

\[
\frac{E\left(\left\| \hat{\theta}_{\text{SPSA}(n)} - \theta^* \right\|^2\right)}{E\left(\left\| \hat{\theta}_{\text{FDSA}(n)} - \theta^* \right\|^2\right)} \to \frac{1}{p^\beta}, \; 0 < \beta \leq \frac{2}{3}
\]

as \( n \to \infty \)

• Equivalently, to achieve **same asymptotic MSE**

\[
\frac{\text{no. meas. } y(\theta) \text{ in SPSA}}{\text{no. meas. } y(\theta) \text{ in FDSA}} = \frac{1}{p}
\]

• Results ☺ and ☺ ☺ are main theoretical results justifying SPSA

7-11
Paraphrase of ☺ ☺ above:

- SPSA and FDSA converge in same number of iterations despite \( p \)-fold savings in cost/iteration for SPSA

— or —

- One properly generated simultaneous random change of all variables in a problem contains as much information for optimization as a full set of one-at-a-time changes of each variable