Of course, the crucial question of that approach is how to choose the stepsizes $\gamma_j$. Let $x_*$ be an optimal solution of (1.1). Note that since the set $X$ is compact and $f(x)$ is continuous, (1.1) has an optimal solution. Note also that the iterate $x_j = x_j(\xi_{j-1})$ is a function of the history $\xi_{[j-1]} = (\xi_1, \ldots, \xi_{j-1})$ of the generated random process and hence is random.

Denote

$$A_j = \frac{1}{2} \| x_j - x_* \|^2_2 \quad \text{and} \quad a_j = \mathbb{E}[A_j] = \frac{1}{2} \mathbb{E} \left[ \| x_j - x_* \|^2_2 \right].$$

By using (1.5) and since $x_* \in X$ and hence $\Pi_X(x_*) = x_*$, we can write

$$A_{j+1} = \frac{1}{2} \| \Pi_X(x_j - \gamma_j G(x_j, \xi_j)) - x_* \|^2_2$$
$$= \frac{1}{2} \| \Pi_X(x_j - \gamma_j G(x_j, \xi_j)) - \Pi_X(x_*) \|^2_2$$
$$\leq \frac{1}{2} \| x_j - \gamma_j G(x_j, \xi_j) - x_* \|^2_2$$
$$= A_j + \frac{1}{2} \gamma_j^2 \| G(x_j, \xi_j) \|^2_2 - \gamma_j (x_j - x_*)^T G(x_j, \xi_j).$$

Since $x_j = x_j(\xi_{j-1})$ is independent of $\xi_j$, we have

$$\mathbb{E} \left[ (x_j - x_*)^T G(x_j, \xi_j) \right] = \mathbb{E} \left\{ \mathbb{E} \left[ (x_j - x_*)^T G(x_j, \xi_j) \mid \xi_{j-1} \right] \right\}$$
$$\leq \mathbb{E} \left\{ (x_j - x_*)^T \mathbb{E} \left[ G(x_j, \xi_j) \mid \xi_{j-1} \right] \right\}$$
$$\leq \mathbb{E} \left\{ (x_j - x_*)^T g(x_j) \right\}.$$ 

(2.4)

Assume now that there is a positive number $M$ such that

$$\mathbb{E} \left[ \| G(x, \xi) \|^2_2 \right] \leq M^2 \quad \forall x \in X.$$ 

Then, by taking expectation of both sides of (2.3) and using (2.4), we obtain

$$a_{j+1} \leq a_j - \gamma_j \mathbb{E} \left[ (x_j - x_*)^T g(x_j) \right] + \frac{1}{2} \gamma_j^2 M^2.$$ 

(2.5)

Suppose further that the expectation function $f(x)$ is differentiable and strongly convex on $X$, i.e., there is constant $c > 0$ such that

$$f(x') \geq f(x) + (x' - x)^T \nabla f(x) + \frac{1}{2} c \| x' - x \|^2_2, \quad \forall x', x \in X,$$

or equivalently that

$$ (x' - x)^T (\nabla f(x') - \nabla f(x)) \geq c \| x' - x \|^2_2 \quad \forall x', x \in X. $$

(2.7)

Note that strong convexity of $f(x)$ implies that the minimizer $x_*$ is unique. By optimality of $x_*$, we have that

$$(x - x_*)^T \nabla f(x_*) \geq 0 \quad \forall x \in X,$$

which together with (2.7) implies that $$(x - x_*)^T \nabla f(x) \geq c \| x - x_* \|^2_2.$$ In turn, it follows that $$(x - x_*)^T g \geq c \| x - x_* \|^2_2$$ for all $x \in X$ and $g \in \partial f(x)$, and hence

$$\mathbb{E} \left[ (x_j - x_*)^T g(x_j) \right] \geq c \mathbb{E} \left[ \| x_j - x_* \|^2_2 \right] = 2ca_j.$$ 

Therefore, it follows from (2.6) that

$$a_{j+1} \leq (1 - 2c\gamma_j)a_j + \frac{1}{2} \gamma_j^2 M^2.$$ 

(2.8)
Let us take stepsizes \( \gamma_j = \theta/j \) for some constant \( \theta > 1/(2c) \). Then, by (2.8), we have

\[
a_{j+1} \leq (1 - 2c\theta/j)a_j + \frac{1}{2} \theta^2 M^2/j^2.
\]

It follows by induction that

\[
(2.9) \quad \mathbb{E} \left[ \|x_j - x_*\|_2^2 \right] = 2a_j \leq Q(\theta)/j,
\]

where

\[
(2.10) \quad Q(\theta) = \max \left\{ \theta^2 M^2(2c\theta - 1)^{-1}, \|x_1 - x_*\|_2^2 \right\}.
\]

Suppose further that \( x_* \) is an interior point of \( X \) and \( \nabla f(x) \) is Lipschitz continuous, i.e., there is constant \( L > 0 \) such that

\[
(2.11) \quad \|\nabla f(x') - \nabla f(x)\|_2 \leq L\|x' - x\|_2 \quad \forall x', x \in X.
\]

Then

\[
(2.12) \quad f(x) \leq f(x_*) + \frac{1}{2} L\|x - x_*\|_2^2, \quad \forall x \in X,
\]

and hence

\[
(2.13) \quad \mathbb{E} [f(x_j) - f(x_*)] \leq \frac{1}{2} L \mathbb{E} \left[ \|x_j - x_*\|_2^2 \right] \leq \frac{1}{2} L Q(\theta)/j,
\]

where \( Q(\theta) \) is defined in (2.10).

Under the specified assumptions, it follows from (2.9) and (2.13), respectively, that after \( t \) iterations, the expected error of the current solution in terms of the distance to \( x_* \) is of order \( O(t^{-1/2}) \), and the expected error in terms of the objective value is of order \( O(t^{-1}) \), provided that \( \theta > 1/(2c) \). The simple example of \( X = \{ x : \|x\|_2 \leq 1 \} \), \( f(x) = \frac{1}{2} c x^T x \), and \( G(x, \xi) = \nabla f(x) + \xi \), with \( \xi \) having standard normal distribution \( \mathcal{N}(0, I_n) \), demonstrates that the outlined upper bounds on the expected errors are tight within factors independent of \( t \).

We have arrived at the \( O(t^{-1}) \) rate of convergence in terms of the expected value of the objective mentioned in the Introduction. Note, however, that the result is highly sensitive to a priori information on \( c \). What would happen if the parameter \( c \) of strong convexity is overestimated? As a simple example, consider \( f(x) = x^T/10 \), \( X = [-1, 1] \subset \mathbb{R} \), and assume that there is no noise, i.e., \( G(x, \xi) = \nabla f(x) \). Suppose, further that we take \( \theta = 1 \) (i.e., \( \gamma_j = 1/j \)), which will be the optimal choice for \( c = 1 \), while actually here \( c = 0.2 \). Then the iteration process becomes

\[
x_{j+1} = x_j - f'(x_j)/j = \left( 1 - \frac{1}{5j} \right) x_j,
\]

and hence starting with \( x_1 = 1 \),

\[
x_j = \prod_{s=1}^{j-1} \left( 1 - \frac{1}{5s} \right) = \exp \left\{ - \sum_{s=1}^{j-1} \ln \left( 1 + \frac{1}{5s-1} \right) \right\} \geq \exp \left\{ - \sum_{s=1}^{j-1} \frac{1}{5s-1} \right\} > \exp \left\{ - 0.25 + \int_1^{j-1} \frac{1}{5t-1} dt \right\} > \exp \left\{ - 0.25 + 0.2 \ln 1.25 - \frac{1}{5} \ln j \right\} \geq 0.8 j^{-1/5}.
\]
That is, the convergence is extremely slow. For example, for \( j = 10^9 \), the error of the iterated solution is greater than 0.015. On the other hand, for the optimal stepsize factor of \( \theta = 1/c = 5 \), the optimal solution \( x_\star = 0 \) is found in one iteration.

It could be added that the stepsizes \( \gamma_j = \theta/j \) may become completely unacceptable when \( f \) loses strong convexity. For example, when \( f(x) = x^4 \), \( X = [-1, 1] \), and there is no noise, these stepsizes result in a disastrously slow convergence: \(|x_j| \geq O([\ln(j+1)]^{-1/2})\). The precise statement here is that with \( \gamma_j = \theta/j \) and \( 0 < x_1 \leq \frac{1}{6\sqrt{\theta}} \), we have that \( x_j \geq \frac{x_1}{\sqrt{1+32\theta x_1^2[1+\ln(j+1)]}} \) for \( j = 1, 2, \ldots \).

We see that in order to make the SA “robust”—applicable to general convex objectives rather than to strongly convex ones—one should replace the classical stepsizes \( \gamma_j = O(j^{-1}) \), which can be too small to ensure a reasonable rate of convergence even in the “no noise” case, with “much larger” stepsizes. At the same time, a detailed analysis shows that “large” stepsizes poorly suppress noise. As early as in [15] it was realized that in order to resolve the arising difficulty, it makes sense to separate collecting information on the objective from generating approximate solutions. Specifically, we can use large stepsizes, say, \( \gamma_j = O(j^{-1/2}) \) in (2.1), thus avoiding too slow motion at the cost of making the trajectory “more noisy.” In order to suppress, to some extent, this noisiness, we take, as approximate solutions, appropriate averages of the search points \( x_j \) rather than these points themselves.