Lecture 7

Stochastic approximation schemes

February 25, 2015
Introduction

A key challenge in the solution of stochastic optimization problems is the need to contend with a multi-dimensional integral posed by the expectation.

Two avenues:

- **Decomposition schemes**: assume $\Omega$ has a finite cardinality or $\Omega \triangleq \{\omega^1, \ldots, \omega^K\}$. Leverage problem structure in constructing primal/dual decomposition problems.

- **Monte-Carlo sampling schemes**: No finite cardinality assumption, generate samples from $\Omega$. This will be the focus of the next set of lectures.
Monte-Carlo sampling approaches

- Consider the stochastic program given by the following:

\[
\min_{x \in X} f(x) \triangleq \mathbb{E}[f(x, \xi)],
\]

where \( f : X \times \mathbb{R}^d \to \mathbb{R}, \) \( X \subseteq \mathbb{R}^n \) is a closed and convex set and \( \xi : \Omega \to \mathbb{R}^d \) is a \( d \)-dimensional random variable.

- Why do decomposition methods represent only a partial resolution. Suppose each component of \( \xi \) is independent and can take on 3 values. This implies that \(|\Omega| = 3^d\) scenarios. Even in modest problem settings, it is not unheard of \( d \) taking on values in the 100s, implying that \( K = 3^{100} \), a massive challenge when parallelism is leveraged.

- Several avenues are available, all of which leverage Monte-Carlo sampling:
(i) **Interior sampling techniques:** In such approaches, one attempts to modify the standard L-shaped method to allow for general sample-spaces [HS96].

(ii) **Exterior sampling techniques or sample-average approximation (SAA) techniques:** In such an approach, we generate a set of $N$ samples, $\xi^1, \ldots, \xi^N$ and consider the sample average problem:

$$\min_{x \in X} \hat{f}_N(x) \triangleq \frac{1}{N} \sum_{k=1}^{N} f(x, \xi^k). \quad (SP_N)$$

Then $\hat{x}_N$ and $u_N$ are statistical estimators of $x^*$ and $f^*$, assuming that $x^*$ denotes the unique solution of $(SP)$ and $f^*$, its value. One
question of interest:

$$\hat{x}_N \to x^*, \quad \text{with probability one, as } N \to \infty$$

$$v_N \to f^*, \quad \text{with probability one, as } N \to \infty$$

(Consistency)

(iii) **Stochastic approximation schemes:** In many instances, a set of a priori samples cannot be generated. Instead, one alternative lies in employing a sample to generate a new iterate. We focus on this scheme in this lecture and revisit the other schemes subsequently.
In the seminal paper by Robbins and Monro [RM51], the authors considered the question of finding the root of a stochastic system; we consider the variant of this problem in the context of solving stochastic optimization problems.

Suppose \( w_k \triangleq \nabla f(x_k; \xi) - \mathbb{E}[\nabla f(x_k; \xi_k)] \).

To motivate the scheme, suppose \( X \triangleq \mathbb{R}^n \), and given \( x_0 \), and consider a Newton method in which \( x_{k+1} \) is updated as follows:

\[
x_{k+1} := x_k - \left( \nabla^2 f(x_k) \right)^{-1} \nabla f(x_k, \xi_k), \quad k = 0, 1, \ldots
\]

By noting that \( \nabla f(x_k, \xi_k) \) can be expressed as \( \mathbb{E}[\nabla f(x_k, \xi)] + w_k \), we have that

\[
x_{k+1} := x_k - \left( \nabla^2 f(x_k) \right)^{-1} \mathbb{E}[\nabla f(x_k, \xi)] - \left( \nabla^2 f(x_k) \right)^{-1} w_k, \quad k = 0, 1, \ldots
\]
• It follows that if $x_k \to x^*$, then $\mathbb{E}[\nabla f(x_k, \xi)] \to 0$ and $\nabla^2 f(x_k) \to \nabla^2 f(x^*) \succ 0$. Consequently, it has to hold that $w_k \to 0$.

• **Challenge:** But this does not hold for a range of problems (such as settings with i.i.d. random variables with finite variance).

• To avert this issue, Robbins and Monro considered an alternate scheme in which $(\nabla^2 f(x_k))^{-1}$ by $\gamma_k$, a positive sequence; more specifically, the scheme reduces to the following:

$$x_{k+1} := x_k - \gamma_k \nabla f(x_k, \xi_k), \quad k = 0, 1, \ldots$$

This iteration can be equivalently written as

$$x_{k+1} := x_k - \gamma_k (\nabla f(x_k) + w_k), \quad k = 0, 1, \ldots$$
• Moreover, the sequence \( \{\gamma_k\} \) satisfies the following requirements:

\[
\sum_{k=0}^{\infty} \gamma_k = \infty, \quad \text{(Non-summability)}
\]

\[
\sum_{k=0}^{\infty} \gamma_k^2 < \infty, \quad \text{(Square-summability)}
\]

• Each of these properties has an important role to play as we will show.

  (i) The role of square summability ensures that \( \sum_k \gamma_k \omega_k \) is convergent in an almost sure sense for a range of problems.

  (ii) On the other hand summability allows for claiming that \( x_k \) converges to \( x^* \) almost surely.
Convergence of SA for strongly convex functions

- We begin by considering a setting where $f(x)$ is a strongly convex function over $X \subseteq \mathbb{R}^n$. Note that strong convexity over a closed and convex set suffices for the existence of a unique solution $x^\star$.

- Consider the following update rule:

$$x_{k+1} := \Pi_X (x_k - \gamma_k F(x_k, w_k)),$$

(1)

where $k \geq 0$, $\Pi_X(y)$ denotes the projection of $y$ on the closed and convex set, $\gamma_k$ denotes the steplength, and $F(x_k, \omega_k)$ denotes the sampled map. Notably, $w_k$ can be defined as $w_k \triangleq F(x_k, \omega_k) - F(x_k)$ and the aforementioned scheme can also be stated as

$$x_{k+1} := \Pi_X (x_k - \gamma_k (F(x_k) + w_k)),$$

(2)
• It is relatively easy to see that the conditional mean, given by $E[\omega_k \mid \mathcal{F}_k]$, is equal to zero for all $k \geq 0$.

• The following Lemma is employed for establishing almost-sure convergence and may be found in [Pol87] (cf. Lemma 10, page 49).

**Lemma 1** Let $\{v_k\}$ be a sequence of nonnegative random variables, where $E[v_0] < \infty$, and let $\{u_k\}$ and $\{\mu_k\}$ be deterministic scalar sequences such that:

$$E[v_{k+1} \mid v_0, \ldots, v_k] \leq (1 - u_k)v_k + \mu_k \quad \text{a.s. for all } k \geq 0,$$

$$0 \leq u_k \leq 1, \quad \mu_k \geq 0, \quad \text{for all } k \geq 0,$$

$$\sum_{k=0}^{\infty} u_k = \infty, \quad \sum_{k=0}^{\infty} \mu_k < \infty, \quad \lim_{k \to \infty} \frac{\mu_k}{u_k} = 0.$$
Then, \( v_k \rightarrow 0 \) almost surely as \( k \rightarrow \infty \).

**Assumption 1** Suppose \( f(x) \) is a strongly convex, continuously differentiable with Lipschitz continuous gradients with convexity constant \( \eta > 0 \) and Lipschitz constant \( L \) over \( X \).

- Throughout, we assume that \( \mathbb{E}[\|x_0\|^2] < \infty \). Furthermore we let \( \mathcal{F}_k \) denote the history of the method up to time \( k \), i.e., \( \mathcal{F}_k = \{x_0, \xi_0, \xi_1, \ldots, \xi_{k-1}\} \) for \( k \geq 1 \) and \( \mathcal{F}_0 = \{x_0\} \).

**Proposition 1 (a.s. convergence under strong convexity)** Consider algorithm (1) and suppose Assumption 1 holds. Then the sequence \( \{x_k\} \) converges almost surely to \( x^* \), a unique solution of (SP).

**Proof:**

- Consider algorithm (1) and let \( F(x) \triangleq \nabla_x \mathbb{E}[f(x, \xi)] \). By the non-expansivity property of the Euclidean projection operator*, for all \( k \geq 0 \),

\[ \text{Recall that } \Pi_X(\bullet) \text{ is said to be non-expansive over } X \text{ if for all } x, y \in X, \|\Pi_X(x) - \Pi_X(y)\| \leq \|x - y\|. \]

\[ \]
\[ \|x_{k+1} - x^*\|^2 \] can be bounded as follows:

\[
\|x_{k+1} - x^*\|^2 = \|\Pi_x(x_k - \gamma_k(F(x_k) + w_k)) - \Pi_x(x^* - \gamma_k F(x^*))\|^2 \\
\leq \|x_k - x^* - \gamma_k(F(x_k) + w_k - F(x^*))\|^2.
\]

- By taking conditional expectations and invoking the fact that the conditional expectation of \(w_k\) was zero or \(\mathbb{E}[w_k \mid \mathcal{F}_k] = 0\), we have

\[
\mathbb{E}\left[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\right] \leq \|x_k - x^*\|^2 \\
+ \gamma_k^2 \|F(x_k) - F(x^*)\|^2 + \gamma_k^2 \mathbb{E}\left[\|w_k\|^2 \mid \mathcal{F}_k\right] \\
- 2\gamma_k(x_k - x^*)^T(F(x_k) - F(x^*)) \\
\leq (1 - 2\eta \gamma_k + \gamma_k^2 L^2)\|x_k - x^*\|^2 + \gamma_k^2 \nu^2,
\]

where the second inequality is a resulting of leveraging the strong
monotonicity and Lipschitz continuity of $F(x)$ over $X$ as well as the boundedness of $\mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k]$.

- It follows that for sufficiently large $k$, $\gamma_k < 2\eta/L^2$ and by considering a shifted sequence, it can be seen that Lemma 1 can be invoked, where $v_{k+1} = \|x_{k+1} - x^*\|^2$, $u_k = 2\eta\gamma_k - \gamma_k^2 L^2$ and $\mu_k = \gamma_k^2 L^2$. It is relatively straightforward to show that $u_k \leq 1$ for $k \geq K$.

- Additionally, $\sum_{k=0}^{\infty} \mu_k = L^2 \sum_{k=0}^{\infty} \gamma_k^2 < \infty$, and $\sum_{k=0}^{\infty} \gamma_k = \sum_{k=0}^{\infty} 2\eta\gamma_k - \sum_{k=0}^{\infty} \gamma_k^2 L^2 = \infty$, since $\gamma_k$ is not summable but is square summable.

- It follows that $v_k \to 0$ almost surely as $k \to \infty$ and the result holds. □
Weakening strong convexity assumption

**Assumption 2** Suppose $f(x)$ is a convex, continuously differentiable, with Lipschitz continuous gradients with Lipschitz constant $L$ over $X$.

We use the following result by Robbins and Siegmund that pertains to to *almost super-martingales*:

**Lemma 2** (Robbins-Siegmund) Let $v_k, u_k, \alpha_k,$ and $\beta_k$ be nonnegative random variables, and let the following relations hold almost surely:

$$
E[v_{k+1} | \tilde{F}_k] \leq (1+\alpha_k)v_k-u_k+\beta_k \quad \text{for all } k, \sum_{k=0}^{\infty} \alpha_k < \infty, \sum_{k=0}^{\infty} \beta_k < \infty,
$$

where $\tilde{F}_k$ denotes the collection $v_0, \ldots, v_k, u_0, \ldots, u_k, \alpha_0, \ldots, \alpha_k, \beta_0, \ldots, \beta_k$. 


Stochastic Optimization
Then, almost surely we have

\[
\lim_{k \to \infty} v_k = v, \quad \sum_{k=0}^{\infty} u_k < \infty,
\]

where \( v \geq 0 \) is some random variable.

**Proposition 2 (a.s.convergence under convexity)** Let Assumption 2 hold. Assume that the optimal set \( X^* \) of (SP) is nonempty. Then, the sequence \( \{x_k\} \) generated by (1) converges almost surely to some random point in \( X^* \).

**Proof:**

- By definition of the method and the nonexpansive property of the
projection operation, we obtain for any $x^* \in X^*$ and $k \geq 0$,

\[
\|x_{k+1} - x^*\|^2 \leq \|x_k - x^* - \gamma_k (\nabla f(x_k) + w_k)\|^2 \\
= \|x_k - x^*\|^2 - 2\gamma_k (\nabla f(x_k) + w_k)^T (x_k - x^*) \\
+ \gamma_k^2 \|\nabla f(x_k) + w_k\|^2.
\]

- By leveraging convexity and the gradient inequality, we have that

\[
f(x^*) \geq f(x_k) + \nabla f(x_k)^T (x^* - x_k),
\]

implying that

\[
-\nabla f(x_k)^T (x_k - x^*) \leq -(f(x_k) - f(x^*)).
\]
• By the previous observation, we have the following:

\[
\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - 2\gamma_k (f(x_k) - f(x^*)) \\
- 2\gamma_k w_k^T (x_k - x^*) + \gamma^2_k \|\nabla f(x_k) + w_k\|^2.
\]

• Since \(\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2\) for any \(a, b \in \mathbb{R}^n\), by using \(f^* = f(x^*)\), and by adding and subtracting \(\nabla f(x^*)\) in the last term, we obtain

\[
\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - 2\gamma_k (f(x_k) - f^*) - 2\gamma_k w_k^T (x_k - x^*) \\
+ 2\gamma^2_k \|\nabla f(x_k) - \nabla f(x^*)\|^2 \\
+ 2\gamma^2_k \|\nabla f(x^*) + w_k\|^2.
\]

• Taking the conditional expectation given \(\mathcal{F}_k\), using \(\mathbb{E}[w_k \mid \mathcal{F}_k] = 0\) and
the Lipschitzian property of the gradient, we have

\[
\mathbb{E}\left[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k\right] \leq (1 + 2L^2\gamma_k^2)\|x_k - x^*\|^2 - 2\gamma_k(f(x_k) - f^*) \\
+ 2\gamma_k^2 \left(\|\nabla f(x^*)\|^2 + \mathbb{E}\left[\|w_k\|^2 \mid \mathcal{F}_k\right]\right).
\]

- The conditions of Lemma 2 are satisfied. Therefore, almost surely, the sequence \(\{\|x_{k+1} - x^*\|\}\) is convergent for any \(x^* \in X^*\) and \(\sum_{k=0}^{\infty} \gamma_k(f(x_k) - f^*) < \infty\).

- The former relation implies that \(\{x_k\}\) is bounded a.s., while the latter implies \(\lim \inf_{k \to \infty} f(x_k) = f^*\) a.s. in view of the condition \(\sum_{k=0}^{\infty} \gamma_k = \infty\).

- Since the set \(X\) is closed, all accumulation points of \(\{x_k\}\) lie in \(X\). Furthermore, since \(f(x_k) \to f^*\) along a subsequence a.s., by continuity
of $f$ it follows that $\{x_k\}$ has a subsequence converging to some random point in $X^*$ a.s. Moreover, since $\{\|x_{k+1} - x^*\|\}$ is convergent for any $x^* \in X^*$ a.s., the entire sequence $\{x_k\}$ converges to some random point in $X^*$ a.s.
Least-norm solutions via Tikhonov regularization

In the previous section, we showed that stochastic approximation allowed for almost sure convergence when the function is merely convex; unfortunately, in such a setting, the solution set of the problem is not a singleton.

One shortcoming: no properties are available for the type of solution obtained. Next, we consider a different scheme that leverages Tikhonov regularization scheme that derives a specific solution.

In such schemes, a sequence of iterates is generated, each element of which is a solution of the Tikhonov subproblem, defined as follows:

\[
\min_{x \in X} \left( f(x) + \frac{1}{2} \epsilon_k \| x \|^2 \right), \quad \text{(Tik}_k\text{)}
\]

with \( \epsilon_k \to 0 \).
Lemma 3 Suppose $f(x)$ is a convex and continuously differentiable function in $x$. Then, $f(x) + \frac{1}{2} \epsilon \|x\|^2$ is a strongly convex function with parameter $\epsilon$.

Proof: Omitted.

Since $(Tik_k)$ is a strongly convex subproblem, it must admit a unique solution. The next result formalizes the convergence of the generated sequence.

Proposition 3 (Tikhonov regularization scheme) Consider a convex optimization problem in which $\nabla f(x)$ is continuous and single-valued. Suppose $x_k$ is defined as the solution to $(Tik_k)$ and $\{\epsilon_k\}$ is a decreasing positive sequence converging to zero. If $x^*$ is the least-norm solution of $(SP)$, then the following holds:

$$\lim_{k \to \infty} x_k = x^*.$$
• Question: Could we use the Tikhonov regularization scheme to compute a solution to (SP). This requires increasingly accurate solutions to

\[
\min_{x \in X} \mathbb{E} \left[ f(x, \omega) + \frac{1}{2} \epsilon_k \|x\|_2^2 \right].
\]

When \( \mathbb{E}[\bullet] \) is over a general measure space, the Tikhonov subproblem requires using a Monte-Carlo scheme and it is difficult to get increasingly exact solutions.
Tikhonov and iterative Tikhonov regularization
\[ x_k = \prod_K (x_k - \gamma (F(x_k) + \epsilon_k x_k)) \]

\[ x_k \equiv x_k (\epsilon_k) \]

Figure 1: Tikhonov regularization
\[
x_k \equiv x_k(\epsilon_k) \\
\begin{align*}
x_k &= \Pi_K(x_k - \gamma(F(x_k) + \epsilon_k x_k)) \\
x_{k+1}^j &= \Pi_K(x_k^j - \gamma(F(x_k^j) + \epsilon_k x_k^j))
\end{align*}
\]

Figure 2: Tikhonov regularization
\[ x_{k+1} = \Pi_K (x_k - \gamma_k (F(x_k) + \epsilon_k x_k)) \]

Update of regularization parameter and iterate at same point in time

**Figure 3: Iterative Tikhonov regularization**
Iterative Tikhonov regularization schemes

\[ x^{k+1} := \Pi_X \left( x^k - \gamma_k \left( \nabla f(x^k) + \epsilon_k x^k + w^k \right) \right). \] (ITR)

- Note that \( x^0 \in X \) is a random initial point with a finite expectation \( \mathbb{E}[\|x^0\|^2] \) and \( F(x^k) \triangleq \nabla f(x^k) = \nabla \mathbb{E}[f(x^k, \omega^k)]. \)

- The vector \( w^k \) is a stochastic error in evaluating \( F(x^k) \), while \( \gamma_k \) is the stepsize and \( \epsilon_k \) is the regularization parameter chosen by user \( i \) at the \( k \)th iteration. The iterate updates can be compactly written as

\[ x^{k+1} := \Pi_K \left( x^k - \gamma_k \left( F(x^k) + \epsilon_k x^k + w^k \right) \right), \] (3)

where \( \gamma_k \) and \( \epsilon_k \) denote the steplength and regularization sequence. The scheme specified by (ITR) (and its compact version (3)) is referred to as an iterative Tikhonov regularization (ITR) method.
Lemma 4 Let the set $X \subseteq \mathbb{R}^n$ be closed and convex, and let the map $F : K \to \mathbb{R}^n$ be continuous and monotone over $K$. Assume that $\text{SOL}(K, F)$ is nonempty. Let the sequence $\{\epsilon_k\}$ be monotonically decreasing to zero. Then, for the Tikhonov sequence $\{y^k\}$, the following hold:

(a) $\{y^k\}$ is bounded and every accumulation point of $\{y^k\}$ is a solution of (SP);

(b) The following inequality holds

$$||y^k - y^{k-1}|| \leq M_y \frac{\epsilon_{k-1} - \epsilon_k}{\epsilon_k} ||y^{k-1}|| \quad \text{for all } k \geq 1,$$

where $M_y$ is a norm bound on the Tikhonov sequence, i.e., $||y^k|| \leq M_y$ for all $k \geq 0$;

(c) Then $\{y^k\}$ converges to the smallest norm solution of (SP).
**Assumption 3** The set $X \subseteq \mathbb{R}^n$ is closed and convex; The mapping $\nabla f : K \to \mathbb{R}^n$ is Lipschitz continuous over $X$ with a constant $L$; The stochastic error is such that $\mathbb{E}[w_k \mid F_k] = 0$ for all $k \geq 0$ almost surely.

Expectedly, convergence of the method (ITR) does rely on some coordination across steplengths and the regularization parameters. Specifically, we impose the following conditions.

**Assumption 4** Let $\{\epsilon_k\}$ be a monotonically decreasing sequence. Let the following hold:

(a) $\lim_{k \to \infty} \frac{\gamma_k}{\epsilon_k} = 0$;

(b) $\lim_{k \to \infty} \gamma_k \epsilon_k = 0$ and $\lim_{k \to \infty} \epsilon_k = 0$;

(c) $\sum_{k=0}^{\infty} \gamma_k \epsilon_k = \infty$;
\( (d) \sum_{k=1}^{\infty} \frac{(\epsilon_k - 1 - \epsilon_k)^2}{\epsilon_k^{2 \min}} \left( 1 + \frac{1}{\gamma_k \epsilon_k} \right) < \infty; \)

\( (e) \lim_{k \to \infty} \frac{(\epsilon_k - 1 - \epsilon_k)^2}{\epsilon_k^{3 \gamma_k}} \left( 1 + \frac{1}{\gamma_k \epsilon_k} \right) = 0; \)

\( (f) \lim_{k \to \infty} \frac{\gamma_k}{\epsilon_k} \mathbb{E}[\|w^k\|^2 \mid \mathcal{F}_k] = 0 \) and \( \sum_{k=0}^{\infty} \gamma_k^2 \mathbb{E}[\|w^k\|^2 \mid \mathcal{F}_k] < \infty \) a.s.

Later in forthcoming Lemma 5, we demonstrate that Assumption 4 can be satisfied by a simple choice of steplength and regularization sequences of the form \((k + \eta_i)^{-a}\) and \((k + \zeta_i)^{-b}\) where \(\eta_i, \zeta_i, a\) and \(b\) are specified in Lemma 5.

In the following proposition, using Assumption 4, we prove that the random sequence \(\{x^k\}\) of the method (ITR) and the Tikhonov sequence \(\{y^k\}\) associated with the problems \(\text{VI}(K, F + D(\epsilon_k))\), \(k \geq 0\), have the same accumulation points a.s. Assumption 4 basically provides the conditions on
the sequences \( \{ \epsilon_{k,i} \} \) and \( \{ \gamma_{k,i} \} \) ensuring that the sequence \( \{ \| x^k - y^{k-1} \|_2 \} \) is a convergent supermartingale.\(^\dagger\)

**Proposition 4** Let Assumptions 3 and 4 hold. Also, assume that (SP) is solvable. Let the sequence \( \{ x^k \} \) be generated by stochastic iterative Tikhonov algorithm (ITR). Then,

\[
\lim_{k \to \infty} \| x^k - y^{k-1} \| = 0 \quad a.s.
\]

**Proof:**

- Recall that the Tikhonov trajectory is defined by the following fixed point relationship: 
  \[ y^k = \Pi_X [y^k - \gamma_k (F(y^k) + \epsilon_k y^k)] \]

\(^\dagger\)Recall that a discrete-time Martingale is a sequence of random variables \( X_1, X_2, \ldots \) such that \( \mathbb{E}[X_{k+1} \mid X_1, \ldots, X_k] = X_k \) for all \( k \). A submartingale and a super-martingale satisfy \( \mathbb{E}[X_{k+1} \mid X_1, \ldots, X_k] \geq X_k \) and \( \mathbb{E}[X_{k+1} \mid X_1, \ldots, X_k] \leq X_k \), respectively for all \( k \).
• By leveraging the nonexpansivity property of the projection operator, we have

\[
\|x^{k+1} - y^k\|^2 = \|\Pi_X [x^k - \gamma_k (F(x^k) + \epsilon_k x^k + w^k)] \\
- \Pi_X [y^k - \gamma_k (F(y^k) + \epsilon_k y^k)]\|^2 \\
\leq \|x^k - \gamma_k (F(x^k) + \epsilon_k x^k + w^k) - y^k + \gamma_k (F(y^k) + \epsilon_k y^k)\|^2.
\]

• By expanding the left hand expression of the preceding relation, it can be seen that

\[
\|x^{k+1} - y^k\|^2 \leq \|x^k - y^k\|^2 - 2\gamma_k (x^k - y^k)^T (F(x^k) - F(y^k)) \\
- 2\gamma_k \epsilon_k \|x^k - y^k\|^2 - 2\gamma_k (x^k - y^k)^T w^k + w^k \quad (4) \\
+ \gamma_k^2 \|F(x^k) - F(y^k) + \epsilon_k (x^k - y^k)\|^2. \quad (5)
\]
• The last term in the inequality can be expanded as

\[
\| F(x^k) - F(y^k) + w^k + \epsilon_k (x^k - y^k) \|^2 \\
= \| F(x^k) - F(y^k) \|^2 + \| w^k \|^2 + \epsilon_k^2 \| x^k - y^k \|^2 \\
+ 2(F(x^k) - F(y^k))^T w^k \\
+ 2\epsilon_k ((x^k - y^k)^T w^k + (F(x^k) - F(y^k))^T (x^k - y^k)).
\]

• Now, we take the expectation of (4) and (6) conditional on the filtration \( \mathcal{F}_k \), and leverage \( \mathbb{E}[w^k_i \mid \mathcal{F}_k] = 0 \) (cf. Assumption 3). By combining
these two relations we obtain

\[ \mathbb{E}[\|x^{k+1} - y^k\|^2 \mid \mathcal{F}_k] \leq (1 - 2\gamma_k \epsilon_k + \gamma_k^2 \epsilon_k^2)\|x^k - y^k\|^2 \]

\[ + \gamma_k^2 \left(\|F(x^k) - F(y^k)\|^2 + \mathbb{E}[\|w^k\|^2 \mid \mathcal{F}_k]\right) \]

\[ + 2\gamma_k^2 \epsilon_k(x^k - y^k)^T(F(x^k) - F(y^k)) \]

\[ - 2\gamma_k(x^k - y^k)^T(F(x^k) - F(y^k)). \]

- By leveraging the Lipschitz continuity and monotonicity of $F$ over $X$, we have the following inequality:

\[ \mathbb{E}[\|x^{k+1} - y^k\|^2 \mid \mathcal{F}_k] \leq (1 - 2\gamma_k \epsilon_k + \gamma_k^2 \epsilon_k^2 + 2\gamma_k^2 \epsilon_k L \]

\[ + \gamma_k^2 L^2)\|x^k - y^k\|^2 + \gamma_k^2 \mathbb{E}[\|w^k\|^2 \mid \mathcal{F}_k]. \]
Letting $q_k \triangleq 1 - 2\gamma_k\epsilon_k + \gamma_k^2(\epsilon_k + L)^2$, it can be concluded that

$$
\mathbb{E}[\|x^{k+1} - y^k\|^2 | \mathcal{F}_k] \leq q_k\|x^k - y^k\|^2 + \gamma_k^2 \mathbb{E}[\|w^k\|^2 | \mathcal{F}_k]. \tag{9}
$$

Next we derive a relationship between $\|x^k - y^k\|$ and $\|x^k - y^{k-1}\|$. By the triangle inequality, we have $\|x^k - y^k\| \leq \|x^k - y^{k-1}\| + \|y^{k-1} - y^k\|$, while from Lemma 4 we have

$$
\|y^k - y^{k-1}\| \leq M_y \frac{\epsilon_{k-1} - \epsilon_k}{\epsilon_k} \quad \text{for all } k \geq 1.
$$
Consequently, we have that

\[ \|x^k - y^k\|^2 \leq \|x^k - y^{k-1}\|^2 + \|y^k - y^{k-1}\|^2 + 2\|x^k - y^{k-1}\|\|y^k - y^{k-1}\| \]
\[ \leq \|x^k - y^{k-1}\|^2 \]
\[ + \left( M_y \frac{\epsilon_{k-1} - \epsilon_k}{\epsilon_k} \right)^2 + 2M_y \frac{\epsilon_{k-1} - \epsilon_k}{\epsilon_k}\|x^k - y^{k-1}\|. \quad (10) \]

The Cauchy-Schwartz inequality allows for estimating the last term as follows:

\[ M_y \frac{\epsilon_{k-1} - \epsilon_k}{\epsilon_k}\|x^k - y^{k-1}\| \leq \gamma_k \epsilon_k \|x^k - y^{k-1}\|^2 + \frac{(\epsilon_{k-1} - \epsilon_k)^2}{\gamma_k \epsilon_k^2} M_y^2. \]
• Using this in the preceding relation we obtain

\[
\|x^k - y^k\|^2 \leq (1 + \gamma_k \epsilon_k) \|x^k - y^{k-1}\|^2 \\
+ \left( M_y \frac{\epsilon_{k-1} - \epsilon_k}{\epsilon_k} \right)^2 \left( 1 + \frac{1}{\gamma_k \epsilon_k} \right).
\]  
(11)

• Combining (9) and (11), we obtain the following estimate:

\[
\mathbb{E}[\|x^{k+1} - y^{k}\|^2 \mid F_k] \leq q_k (1 + \gamma_k \epsilon_k) \|x^k - y^{k-1}\|^2 + \gamma_k^2 \mathbb{E}[\|w^k\|^2 \mid F_k] \\
+ q_k \left( M_y \frac{\epsilon_{k-1} - \epsilon_k}{\epsilon_k} \right)^2 \left( 1 + \frac{1}{\gamma_k \epsilon_k} \right).
\]  
(12)

\[
+ q_k \left( M_y \frac{\epsilon_{k-1} - \epsilon_k}{\epsilon_k} \right)^2 \left( 1 + \frac{1}{\gamma_k \epsilon_k} \right).
\]  
(13)

• Next, we estimate the coefficient of \(\|x^k - y^{k-1}\|^2\) in (12). Recalling the definition \(q_k = 1 - 2\gamma_k \epsilon_k + \gamma_k^2 (\epsilon_k + L)^2\), we show that \(q_k \in (0, 1)\) for
sufficiently large $k$, allowing us to express $q_k$ as follows:

$$q_k = 1 - \gamma_k \epsilon_k \left(2 - \frac{\gamma_k^2 (\epsilon_k + L)^2}{\gamma_k \epsilon_k}\right).$$

By Assumption 4(a) we have

$$\lim_{k \to \infty} \left(\frac{\gamma_k}{\epsilon_k} (\epsilon_k + L)^2\right) = 0,$$

implying that there exists a large enough integer $\tilde{k} \geq 0$ such that

$$\frac{\gamma_k}{\epsilon_k} (\epsilon_k + L)^2 \leq c \quad (14)$$

holds for all $k \geq \tilde{k}$ and some $c \in (0, 1)$. 
• Thus, $1 \leq 2 - \frac{\gamma_k}{\epsilon_k}(\epsilon_k + L)^2 \leq 2$ for all $k \geq \tilde{k}$, implying that for $q_k$ we have $1 - 2\gamma_k \epsilon_k \leq q_k \leq 1 - \gamma_k \epsilon_k$ for all $k \geq \tilde{k}$.

• Furthermore, since $\gamma_k \epsilon_k \to 0$ by Assumption 4(b), we can choose $\tilde{k}$ large enough so that $q_k \in (0, 1)$ for $k \geq \tilde{k}$. Hence, for $k \geq \tilde{k}$ we obtain

$$0 \leq q_k (1 + \gamma_k \epsilon_k) \leq q_k + \gamma_k \epsilon_k$$

and using the definition of $q_k$ we further have for $k \geq \tilde{k}$,

$$0 \leq q_k (1 + \gamma_k \epsilon_k) \leq 1 - \gamma_k \epsilon_k \left(1 - \frac{\gamma_k (\epsilon_k + L)^2}{\epsilon_k}\right) \leq 1 - \gamma_k \epsilon_k (1 - c),$$

(15)

where the last inequality follows from (14).

• Using relations (15) and (12), we obtain

$$\mathbb{E}[\|x^{k+1} - y^k\|^2 \mid \mathcal{F}_k] \leq (1 - u_k)\|x^k - y^{k-1}\|^2 + v_k \quad \text{for all } k \geq \tilde{k},$$
where \( u_k \triangleq (1 - c)\gamma_k\epsilon_k \) and

\[
v_k = q_k \left( M_y \frac{\epsilon_{k-1} - \epsilon_k}{\epsilon_k} \right)^2 \left( 1 + \frac{1}{\gamma_k\epsilon_k} \right) r + \gamma_k^2 \mathbb{E}[\|w^k\|^2 | \mathcal{F}_k].
\]

- We now verify that the conditions of Lemma 2 are satisfied for \( k \geq \tilde{k} \). Since \( c < 1 \), from (15) we have \( 0 \leq u_k \leq 1 \) for all \( k \geq \tilde{k} \), while from Assumption 4(c) we have \( \sum_{k=\tilde{k}}^{\infty} u_k = \infty \).

- Under stepsize conditions Assumption 4(d)–(f), it can be verified that \( \lim_{k \to \infty} \frac{v_k}{u_k} = 0 \) and \( \sum_{k=0}^{\infty} v_k < \infty \).

- Thus, the conditions of Lemma 2 are satisfied for \( k \geq \tilde{k} \). Noting that Lemma 2 applies to a process delayed by a deterministic time-offset, we can conclude that \( \|x^k - y^{k-1}\| \to 0 \text{ a.s.} \).
Lemma 5 Consider $\gamma_k = (k + \eta)^{-a}$ and $\epsilon_k = (k + \zeta)^{-b}$ for $k \geq 0$, where each $\eta$ and $\zeta$ are selected from a uniform distribution on the intervals $[\underline{\eta}, \bar{\eta}]$ and $[\underline{\zeta}, \bar{\zeta}]$, respectively, for some $0 < \underline{\eta} < \bar{\eta}$ and $0 < \underline{\zeta} < \bar{\zeta}$. Let $a, b \in (0, 1)$, $a + b < 1$, and $a > b$. Then $\{\gamma_k\}$ and $\{\epsilon_k\}$ satisfy Assumption 4(a)–(e).
Constant steplength error bounds

Let us begin with an assumption:

**Assumption 5** Suppose the function $f(x)$ is strongly convex with constant $c$ and its sampled gradients satisfy $\mathbb{E}[\|\nabla f(x, \xi)\|^2] \leq M^2$ for all $x \in X$.

Suppose $A_j := \frac{1}{2}\|x_j - x^*\|^2$ and $a_j := \mathbb{E}[A_j]$.

Next, we observe that

$$A_{j+1} = \frac{1}{2}\|x_j - x^*\|^2$$

$$= \frac{1}{2}\|\Pi_X (x_j - \gamma_j \nabla f(x_j, \xi_j)) - \Pi_X (x^*)\|^2$$

$$\leq \frac{1}{2}\| (x_j - \gamma_j \nabla f(x_j, \xi_j)) - x^*\|^2$$

$$= A_j + \frac{1}{2}\gamma_j^2\|\nabla f(x_j, \xi_j)\|^2 - \gamma_j (x_j - x^*)^T \nabla f(x_j, \xi_j).$$
It can be seen that \( x_j = x_j(\xi_1, \ldots, \xi_{j-1}) = x_j(\xi_{[j-1]}) \), implying that \( x_j \) is independent of \( \xi_j \). As a consequence,

\[
\mathbb{E}\left[(x_j - x^*)^T \nabla f(x_j, \xi_j)\right] = \mathbb{E}\left[\mathbb{E}\left[(x_j - x^*)^T \nabla f(x_j, \xi_j) \mid \xi_{[j-1]}\right]\right] \\
= \mathbb{E}\left[(x_j - x^*)^T \mathbb{E}\left[\nabla f(x_j, \xi_j) \mid \xi_{[j-1]}\right]\right] \\
= \mathbb{E}\left[(x_j - x^*)^T \nabla f(x_j)\right].
\]

It follows from taking expectations on both sides of

\[
A_{j+1} \leq A_j + \frac{1}{2} \gamma_j^2 \| \nabla f(x_j, \xi_j) \|^2 - \gamma_j (x_j - x^*)^T \nabla f(x_j, \xi_j),
\]

leading to

\[
a_{j+1} \leq a_j + \frac{1}{2} \gamma^2 M^2 - \gamma_j \mathbb{E}\left[(x_j - x^*)^T \nabla f(x_j, \xi_j)\right].
\]
By assumption, we have that $f(x)$ is strongly convex over $X$ with a constant $c$. Recall by the optimality of $x^*$,

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

As a consequence,

$$
\mathbb{E}[(x_j - x^*)^T \nabla f(x_j, \xi_j)] = \mathbb{E}[(x_j - x^*)^T (\nabla f(x_j, \xi_j) - \nabla f(x^*\xi_j))] \\
+ (x_j - x^*) \nabla f(x^*) \\
\geq \mathbb{E}[(x_j - x^*)^T (\nabla f(x_j, \xi_j) - \nabla f(x^*\xi_j))] \\
\geq c \|x_j - x^*\|^2 = 2ca_j.
$$

It follows that

$$
a_{j+1} \leq (1 - 2c\gamma_j)a_j + \frac{1}{2}\gamma_j^2 M^2.
$$
Generally in stochastic approximation problems, the steplength sequence employed is $\gamma_j = \theta/j$ where $\theta$ is a positive scalar, allowing us to claim that

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

Suppose we now additionally impose a requirement that $\theta > 1/2c$, we obtain that

$$2a_j \leq \frac{\max\{\theta^2M^2(2c\theta - 1)^{-1}, 2a_1\}}{j}.$$ 

As a result,

$$\mathbb{E}[\|x_j - x^*\|^2] \leq \frac{Q(\theta)}{j},$$

where

$$Q(\theta) \triangleq \max\{\theta^2M^2(2c\theta - 1)^{-1}, \|x_1 - x^*\|^2\}.$$ 

Furthermore, $Q(\theta)$ is minimized at $\theta = 1/c$. 

Next, we assume that $x^* \in \text{int}(X)$. Furthermore, $\nabla f(x)$ is assumed to be Lipschitz continuous with a constant $L$. As a result, we have that

$$f(x^*) \geq f(x) + \nabla f(x)^T(x^* - x) = f(x) + (\nabla f(x) - \nabla f(x^*))^T(x - x^*)$$

$$f(x) \leq f(x^*) - (\nabla f(x) - \nabla f(x^*))^T(x - x^*) \leq f(x^*) + L\|x - x^*\|^2,$$

for all $x \in X$.

It follows that

$$\mathbb{E}[f(x_j) - f(x^*)] \leq L\mathbb{E}[\|x_j - x^*\|^2] \leq \frac{Q(\theta)L}{j}.$$

Findings: ($\theta > 1/(2c)$)

- After $j$ iterations, expected error of solution is of the order of $O(j^{-\frac{1}{2}})$. 

Stochastic Optimization
After \( j \) iterations, expected error of function is of the order of \( O(j^{-1}) \).

Note that if \( c \) is overestimated or \( \theta < 1/(2c) \), this can prove devastating.

**Example 1** Consider a noise-free problem in which \( f(x) = \frac{1}{2} \kappa x^2 \) with \( \kappa > 0 \). Clearly \( x^* = 0 \). Furthermore, \( G(x, \xi) = \nabla f(x) = \kappa x \). Suppose \( \theta = 1 \) and \( \gamma_j = 1/j \). Then we have

\[
x_{j+1} = x_j - \kappa x_j / j = x_j (1 - \kappa / j).
\]

- Suppose \( \kappa = 1 \). Then the optimal solution is found in one iteration.

- Suppose \( \kappa < 1 \). Then we have

\[
x_{j+1} = x_1 \prod_{s=1}^{j} \left(1 - \frac{\kappa}{s}\right) = x_1 \exp \left( - \sum_{s=1}^{j} \ln \left(1 + \frac{\kappa}{s - \kappa}\right) \right).
\]
A bit of algebra reveals that

\[ x_{j+1} = O(1)j^{-\kappa} \text{ and } f(x_{j+1}) > O(1)j^{-2\kappa}. \]

In effect the convergence becomes very poor as \( \kappa \to 0 \). Specifically, to reduce the error in \( x_j \) by a factor of 10, we need to increase the number of iterations by \( 10^{1/\kappa} \).

For example, if \( \kappa = 0.1, x_1 = 1, j = 1e5 \), we have that \( x_j > 0.28 \). To reduce this to 0.028, we need to increase the number of iterations by \( 10^{10} \) or \( j = 10^{15} \).

Furthermore if the function loses strong convexity, the parameter \( c \) degenerates to zero and no choice of \( \theta > 1/2c \) may be available.
Robust Stochastic Approximation

An observation drawn from the previous subsection is that $\gamma_j = O(1/j)$ can be far too small to achieve a reasonable convergence rate.

An important advance was made by Polyak and Juditsky in generating a robust stepsize policy.

By convexity of $f(x)$, we have that for any $x \in X$,

$$f(x) \geq f(x_j) + (x - x_j)^T g(x_j),$$

where $g(x_j) \in \partial f(x_j)$. Taking expectations, we obtain

$$\mathbb{E}[(x_j - x^*)^T g(x_j)] \geq \mathbb{E}[f(x_j) - f(x^*)].$$
But, we have that

\[
\gamma_j \mathbb{E}[(x_j - x^*)^T g(x_j)] \leq a_j - a_{j+1} + \frac{1}{2} \gamma^2_j M^2,
\]

implying that

\[
\gamma_j \mathbb{E}[f(x_j) - f(x^*)] \leq a_j - a_{j+1} + \frac{1}{2} \gamma^2_j M^2.
\]

As a result, for \(1 \leq i \leq j\), we have the following:

\[
\sum_{t=i}^{j} \gamma_t \mathbb{E}[f(x_t) - f(x^*)] \leq \sum_{t=i}^{j} (a_t - a_{t+1}) + \frac{1}{2} \sum_{t=i}^{j} \gamma^2_t M^2 \leq a_i + \frac{1}{2} \sum_{t=i}^{j} \gamma^2_t M^2.
\]
Next, we define $v_t$ and $D_X$ as

$$v_t \triangleq \frac{\gamma_t}{\sum_{\tau=1}^{j} \gamma_{\tau}} \quad \text{and} \quad D_X \triangleq \max_{x \in X} \| x - x_1 \|_2.$$ 

It follows from invoking these definitions, that

$$\mathbb{E}\left[ \sum_{t=i}^{j} v_t f(x_t) - f(x^*) \right] \leq a_i + \frac{1}{2} \sum_{t=i}^{j} \frac{\gamma_t^2 M^2}{\sum_{t=i}^{j} \gamma_t}. \quad (16)$$

Next, we consider points given by $\tilde{x}_{i,j} := \sum_{t=i}^{j} v_t x_t$. By the convexity of $X$, we have that $\tilde{x}_{i,j} \in X$ and by the convexity of $f(\bullet)$:

$$f(\tilde{x}_{i,j}) \leq \sum_{t=i}^{j} v_t f(x_t).$$
From (16) and by noting that $a_1 \leq \frac{1}{2}D_X^2$ and $a_i \leq 2D_x^2$ for $i > 1$, we obtain the following:

$$\mathbb{E}[f(\tilde{x}_{1,j}) - f(x^*)] \leq \frac{D_X^2 + \frac{1}{2} \sum_{t=i}^{j} \gamma_t^2 M^2}{2 \sum_{t=i}^{j} \gamma_t}, \quad 1 \leq j$$  \hspace{1cm} (17)

$$\mathbb{E}[f(\tilde{x}_{i,j}) - f(x^*)] \leq \frac{4D_X^2 + \frac{1}{2} \sum_{t=i}^{j} \gamma_t^2 M^2}{2 \sum_{t=i}^{j} \gamma_t}, \quad 1 \leq i \leq j$$  \hspace{1cm} (18)

We are now in a position to develop constant stepsize schemes. Suppose $\gamma_t = \gamma$ for all $t = 1, \ldots, N$. Then it follows that

$$\mathbb{E}[f(\tilde{x}_{1,N}) - f(x^*)] \leq \frac{D_X^2 + M^2 N \gamma^2}{2N \gamma}.$$
By minimizing the right hand side in $\gamma > 0$, we obtain that

$$\gamma^* = \frac{D_x}{M \sqrt{N}}.$$  

This leads to the following efficiency estimate:

$$\mathbb{E}[f(\tilde{x}_{1,N}) - f(x^*)] \leq \frac{D_x M}{\sqrt{N}}.$$  \hspace{1cm} (19)

Next, we can also claim that for $1 \leq K \leq N$,  

$$\mathbb{E}[f(\tilde{x}_{K,N}) - f(x^*)] \leq \frac{C_{N,K} D_x M}{\sqrt{N}},$$  \hspace{1cm} (20)

$$C_{N,K} := \frac{2N}{N - K + 1} + \frac{1}{2}.$$
If we scale $\gamma$ by $\theta$, a positive scalar, implying that

$$
\gamma_t = \frac{\theta D_X}{M \sqrt{N}}, \quad t = 1, \ldots, N.
$$

The resulting efficiency estimate then becomes the following:

$$
\mathbb{E}[f(\tilde{x}_{K,N}) - f(x^*)] \leq \max\{\theta, \frac{1}{\theta}\} \frac{C_{N,K} D_X M}{\sqrt{N}}. \quad (21)
$$

Naturally, the expected error $O(N^{-\frac{1}{2}})$ is worse than the $O(N^{-1})$ obtained in the classical approach. But in that setting, the function was a smooth and strongly convex function. Here the error bound is guaranteed regardless of smoothness and the strong convexity assumptions. Further, changing $\theta$ does not have a devastating impact; it just rescales the error. Thus the qualifier “robust.”
References

