Stochastic gradient methods

Yuxin Chen
Princeton University, Spring 2018
Outline

- Stochastic gradient descent (stochastic approximation)
- Convergence analysis
- Reducing variance via iterate averaging
Stochastic programming

\[
\begin{align*}
\text{minimize}_x & \quad F(x) = \mathbb{E}[f(x; \xi)] \\
& \text{expected risk, population risk, ...}
\end{align*}
\]

- \(\xi\): randomness in problem
- suppose \(f(\cdot, \xi)\) is convex for every \(\xi\) (and hence \(F(\cdot)\) is convex)
Example: empirical risk minimization

Let \( \{a_i, y_i\}_{i=1}^n \) be \( n \) random samples, and consider

\[
\text{minimize}_{x} \quad F(x) := \frac{1}{n} \sum_{i=1}^{n} f(x; \{a_i, y_i\})
\]

\( \text{empirical risk} \)

e.g. quadratic loss \( f(x; \{a_i, y_i\}) = (a_i^\top x - y_i)^2 \)

If one draws index \( j \sim \text{Unif}(1, \cdots, n) \) uniformly at random, then

\[
F(x) = \mathbb{E}_j[f(x; \{a_j, y_j\})]
\]
A natural solution

Under “mild” technical conditions

\[ x^{t+1} = x^t - \eta_t \nabla F(x^t) \]
\[ = x^t - \eta_t \nabla \mathbb{E}[f(x^t; \xi)] \]
\[ = x^t - \eta_t \mathbb{E}[\nabla x f(x^t; \xi)] \]

issues:

- distribution of \( \xi \) may be unknown
- even if it is known, evaluating high-dimensional expectation is often expensive
Stochastic gradient descent
(stochastic approximation)
Stochastic gradient descent (SGD)

— Robbins, Monro ’51

stochastic approximation / stochastic gradient descent (SGD)

\[ x^{t+1} = x^t - \eta_t \, g(x^t; \xi^t) \quad (12.1) \]

where \( g(x^t; \xi^t) \) is \textit{unbiased} estimate of \( \nabla F(x^t) \), i.e.

\[ \mathbb{E}[g(x^t; \xi^t)] = \nabla F(x^t) \]
Stochastic gradient descent (SGD)

— Robbins, Monro ’51

stochastic approximation / stochastic gradient descent (SGD)

\[ x^{t+1} = x^t - \eta_t g(x^t; \xi^t) \]  

(12.1)

- a stochastic algorithm for finding critical point \( x \) obeying \( \nabla F(x) = 0 \)
- more generally, a stochastic algorithm for finding roots of \( G(x) := \mathbb{E}[g(x; \xi)] \)
Example: SGD for empirical risk minimization

\begin{align*}
\text{minimize}_x \quad F(x) := \frac{1}{n} \sum_{i=1}^{n} f(x; \{a_i, y_i\}) \\
\text{empirical risk}
\end{align*}

\textbf{for } t = 0, 1, \ldots \\
\text{choose } i_t \text{ uniformly at random}

\begin{align*}
x^{t+1} &= x^t - \eta_t \nabla_x f_{i_t}(x^t; \{a_i, y_i\})
\end{align*}
Example: SGD for empirical risk minimization

**benefits:** SGD exploits information more efficiently than batch methods

- practical data usually involve lots of redundancy; using all data simultaneously in each iteration might be inefficient
- SGD is particularly efficient at very beginning, as it achieves fast initial improvement with very cheap per-iteration cost
Example: SGD for empirical risk minimization

— Bottou, Curtis, Nocedal ’18

Fig. 3.1: Empirical risk $R_n$ as a function of the number of accessed data points (ADP) for a batch L-BFGS method and the stochastic gradient (SG) method ($\eta_t = 4$) on a binary classification problem with a logistic loss objective and the RCV1 dataset. SG was run with a fixed stepsize of $\eta = 4$. Theoretical Motivation

One can also cite theoretical arguments for a preference of SG over a batch approach. Let us give a preview of these arguments now, which are studied in more depth and further detail in §4.

- It is well known that a batch approach can minimize $R_n$ at a fast rate; e.g., if $R_n$ is strongly convex (see Assumption 4.5) and one applies a batch gradient method, then there exists a constant $\theta_2(0, 1)$ such that, for all $k \geq N$, the training error satisfies

$$R_n(w_k) \leq \min R_n - O(\theta_2),$$

where $R_n$ denotes the minimal value of $R_n$. The rate of convergence exhibited here is referred to as R-linear convergence in the optimization literature [117] and geometric convergence in the machine learning research community; we shall simply refer to it as linear convergence.

From (3.9), one can conclude that, in the worst case, the total number of iterations in which the training error can be above a given $\epsilon > 0$ is proportional to $\log(1/\epsilon)$. This means that, with $\eta_t = 4$, the number of iterations required to achieve an error below $\epsilon$ is

$$\text{iterations} \propto \log(1/\epsilon).$$

Stochastic gradient methods 12-10
Reinforcement learning studies Markov decision process (MDP) with unknown model

**core problem:** estimate so-called “value function” under stationary policy $\pi$

$$V^\pi(s) = \mathbb{E}\left[r_0 + \gamma V^\pi(s_1) \mid s_0 = s\right]$$

(12.2)

for all $s \in S$, without knowing transition probabilities of MDP
Example: temporal difference (TD) learning

We won’t explain what equation (12.2) means, but remark that . . .

- $V^\pi(\cdot)$: value function under policy $\pi$
- $s_t$: state at time $t$
- $S$: state space
- $0 < \gamma < 1$: discount factor
- $r_t$: reward at time $t$
Definition of value function is equivalent to

\[
\mathbb{E} \left[ V^\pi(s) - r_0 - \gamma V^\pi(s_1) \bigg| s_0 = s \right] = 0
\]

**TD(0) algorithm:** for \( t = 0, 1, \ldots \)

draw new state \( s_{t+1} \), collect reward \( r_t \), then update

\[
\hat{V}^\pi(s_t) \leftarrow \hat{V}^\pi(s_t) - \eta_t \, g(\hat{V}^\pi)
\]

or

\[
\hat{V}^\pi(s_t) \leftarrow \hat{V}^\pi(s_t) - \eta_t \left\{ \hat{V}^\pi(s_t) - r_t - \hat{V}^\pi(s_{t+1}) \right\}
\]
Example: Q-learning

What if we also want to find optimal policy?

**core problem:** solve so-called “Bellman equation”

\[
V(s) = \max_{a \in A} \left\{ R(s, a) + \gamma \mathbb{E}[V(s_1) | s_0 = s, a_0 = a] \right\}
\]  \hspace{1cm} (12.3)

for all \( s \in S \), without knowing transition probabilities of MDP
Example: Q-learning

Again we won’t explain what Bellman equation means, but remark that ...

- $V(\cdot)$: value function
- $s_t$: state at time $t$
- $S$: state space
- $a_t$: action at time $t$
- $A$: action space
- $0 < \gamma < 1$: discount factor
- $R(\cdot, \cdot)$: reward function
Example: Q-learning

\[
V(s) = \max_{a \in A} \left\{ R(s, a) + \gamma \mathbb{E}[V(s_1) | s_0 = s, a_0 = a] \right\}
\]

- since transition probabilities are unknown, it is natural to resort to stochastic approximation methods
- issue: Bellman equation has \( \mathbb{E} \) inside max operator
- very cute idea: introduce so-called “Q function”
Example: Q-learning

\[ V(s) = \max_{a \in A} \left\{ R(s, a) + \gamma \mathbb{E}[V(s_1) \mid s_0 = s, a_0 = a] \right\} \]

Define Q function as

\[ Q(s, a) := R(s, a) + \gamma \mathbb{E}[V(s_1) \mid s_0 = s, a_0 = a] \]

\[ = R(s, a) + \gamma \mathbb{E}\left[ \max_{\tilde{a} \in A} Q(s_1, \tilde{a}) \mid s_0 = s, a_0 = a \right] \]

\[ = V(s_1) \quad (12.4) \]

- **Q learning**: use stochastic approximation methods to estimate Q function (rather than value function \( V(\cdot) \))
Example: Q-learning

Definition of Q-function is equivalent to

\[
\mathbb{E}\left[ Q(s, a) - R(s, a) - \gamma \max_{\tilde{a} \in A} Q(s_1, \tilde{a}) \mid s_0 = s, a_0 = a \right] = 0
\]

\[
g(Q)
\]

Q-learning algorithm: for \( t = 0, 1, \ldots \)

draw new state \( s_{t+1} \) using action \( a_t \), then update

\[
\hat{Q}(s_t, a_t) \leftarrow \hat{Q}(s_t, a_t) - \eta_t g(\hat{Q}) \quad \text{or}
\]

\[
\hat{Q}(s_t, a_t) \leftarrow \hat{Q}(s_t, a_t) - \eta_t \left\{ \hat{Q}(s_t, a_t) - R(s_t, a_t) - \gamma \max_{\tilde{a} \in A} \hat{Q}(s_{t+1}, \tilde{a}) \right\}
\]
Convergence analysis
Strongly convex and smooth problems

\[
\min_{x} \quad F(x) := \mathbb{E}[f(x; \xi)]
\]

- \(F\): \(\mu\)-strongly convex, \(L\)-smooth

- \(g(x^t; \xi^t)\): unbiased estimate of \(\nabla F(x^t)\) given \(\{\xi^0, \cdots, \xi^{t-1}\}\)

- for all \(x\),

\[
\mathbb{E}\left[\|g(x; \xi)\|_2^2\right] \leq \sigma_g^2 + c_g \|\nabla F(x)\|_2^2 \tag{12.5}
\]
Theorem 12.1 (Convergence of SGD for strongly convex problems; fixed stepsizes)

Under assumptions in Page 12-20, if $\eta_t \equiv \eta \leq \frac{1}{Lc_g}$, then SGD (12.1) achieves

$$\mathbb{E}[F(x^t) - F(x^*)] \leq \frac{\eta L\sigma_g^2}{2\mu} + (1 - \eta \mu)^t(F(x^0) - F(x^*))$$

- check Bottou, Curtis, Nocedal ’18 (Theorem 4.6) for proof
Implications: SGD with fixed stepsizes

\[ \mathbb{E}[F(x^t) - F(x^*)] \leq \frac{\eta L \sigma_g^2}{2\mu} + (1 - \eta \mu)^t \left(F(x^0) - F(x^*)\right) \]

- fast (linear) convergence at very beginning
- converges to some neighborhood of \( x^* \) — variation in gradient computation prevents further progress
- when gradient computation is noiseless (i.e. \( \sigma_g = 0 \)), it converges linearly to optimal
- smaller stepsizes \( \eta \) yield better converging points
One practical strategy

Run SGD with fixed stepsizes; whenever progress stalls, reduce stepsizes and continue SGD

― Bottou, Curtis, Nocedal ’18

whenever progress stalls, we half stepsizes and repeat
Convergence with diminishing stepsizes

**Theorem 12.2 (Convergence of SGD for strongly convex problems; diminishing stepsizes)**

Suppose $F$ is $\mu$-strongly convex, and (12.5) holds with $c_g = 0$. If $\eta_t = \frac{\theta}{t+1}$ for some $\theta > \frac{1}{2\mu}$, then SGD (12.1) achieves

$$
\mathbb{E}\left[\|x^t - x^*\|^2\right] \leq \frac{c_\theta}{t + 1}
$$

where $c_\theta = \max \left\{ \frac{2\theta^2\sigma^2_g}{2\mu\theta - 1}, \|x_0 - x^*\|^2 \right\}$

- convergence rate $O(1/t)$ with diminishing stepsize $\eta_t \asymp 1/t$
Proof of Theorem 12.2

Using SGD update rule, we have

\[
\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2_2 = \|\mathbf{x}^t - \eta_t \mathbf{g}(\mathbf{x}^t; \xi^t) - \mathbf{x}^*\|^2_2 \\
= \|\mathbf{x}^t - \mathbf{x}^*\|^2_2 - 2\eta_t (\mathbf{x}^t - \mathbf{x}^*)^\top \mathbf{g}(\mathbf{x}^t; \xi^t) + \eta_t^2 \|\mathbf{g}(\mathbf{x}^t; \xi^t)\|^2_2 \quad (12.6)
\]

Since \(\mathbf{x}^t\) is indep. of \(\xi_t\), apply law of total expectation to obtain

\[
\mathbb{E}[(\mathbf{x}^t - \mathbf{x}^*)^\top \mathbf{g}(\mathbf{x}^t; \xi^t)] = \mathbb{E}\left[\mathbb{E}[(\mathbf{x}^t - \mathbf{x}^*)^\top \mathbf{g}(\mathbf{x}^t; \xi^t) | \xi_1, \cdots, \xi_{t-1}]\right] \\
= \mathbb{E}[(\mathbf{x}^t - \mathbf{x}^*)^\top \mathbb{E}[\mathbf{g}(\mathbf{x}^t; \xi^t) | \xi_1, \cdots, \xi_{t-1}]] \\
= \mathbb{E}[(\mathbf{x}^t - \mathbf{x}^*)^\top \nabla F(\mathbf{x}^t)] \quad (12.7)
\]
Proof of Theorem 12.2 (cont.)

Furthermore, strong convexity gives

\[ \langle \nabla F(x^t), x^t - x^* \rangle = \langle \nabla F(x^t) - \nabla F(x^*), x^t - x^* \rangle \geq \mu \| x^t - x^* \|_2^2 = 0 \]

\[ \implies \mathbb{E} \left[ \langle \nabla F(x^t), x^t - x^* \rangle \right] \geq \mu \mathbb{E} \left[ \| x^t - x^* \|_2^2 \right] \quad (12.8) \]

Combine (12.6), (12.7), (12.8) and (12.5) (with \( c_g = 0 \)) to obtain

\[ \mathbb{E} \left[ \| x^{t+1} - x^* \|_2^2 \right] \leq (1 - 2\mu \eta_t) \mathbb{E} \left[ \| x^t - x^* \|_2^2 \right] + \eta_t^2 \sigma_g^2 \]

\( \text{does not vanish unless } \eta_t \to 0 \) \quad (12.9)

Take \( \eta_t = \frac{\theta}{t+1} \) and use induction to conclude proof (exercise!)
Informally, when minimizing strongly convex functions, no algorithm performing $t$ queries to noisy first-order oracles can achieve accuracy better than the order of $1/t$

$$\implies \text{SGD with stepsizes } \eta_t \asymp 1/t \text{ is optimal}$$
More precisely, consider a class of problems in which $f$ is $\mu$-strongly convex and $L$-smooth, and $\text{Var}(\|g(x^t; \xi^t)\|_2) \leq \sigma^2$. Then worst-case iteration complexity for (stochastic) first-order methods:

$\sqrt{\frac{L}{\mu}} \log \left( \frac{L\|x_0 - x^*\|_2^2}{\varepsilon} \right) + \frac{\sigma^2}{\mu \varepsilon}$

- for deterministic case: $\sigma = 0$, and hence lower bound is

$\sqrt{\frac{L}{\mu}} \log \left( \frac{L\|x_0 - x^*\|_2^2}{\varepsilon} \right)$

(achievable by Nesterov’s method)
More precisely, consider a class of problems in which $f$ is $\mu$-strongly convex and $L$-smooth, and $\text{Var}(\|g(x^t; \xi^t)\|_2) \leq \sigma^2$. Then worst-case iteration complexity for (stochastic) first-order methods:

$$\sqrt{\frac{L}{\mu}} \log \left( \frac{L\|x_0 - x^*\|^2}{\varepsilon} \right) + \frac{\sigma^2}{\mu \varepsilon}$$

- for noisy case with large $\sigma$, lower bound is dominated by

$$\frac{\sigma^2}{\mu} \cdot \frac{1}{\varepsilon}$$
Comparisons with batch GD

Empirical risk minimization with $n$ samples:

<table>
<thead>
<tr>
<th></th>
<th>iteration complexity</th>
<th>per-iteration cost</th>
<th>total comput. cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>batch GD</td>
<td>$\log \frac{1}{\varepsilon}$</td>
<td>$n$</td>
<td>$n \log \frac{1}{\varepsilon}$</td>
</tr>
<tr>
<td>SGD</td>
<td>$\frac{1}{\varepsilon}$</td>
<td>1</td>
<td>$\frac{1}{\varepsilon}$</td>
</tr>
</tbody>
</table>

SGD is more appealing for large $n$ and moderate accuracy $\varepsilon$ (in which case $\frac{1}{\varepsilon} < n \log \frac{1}{\varepsilon}$)

— which often arises in *big data* regime!
Convex problems

What if we lose strong convexity?

\[
\text{minimize}_x \quad F(x) := \mathbb{E}[f(x; \xi)]
\]

- \( F \): convex
- \( \mathbb{E}[\|g(x; \xi)\|^2_2] \leq \sigma^2_g \) for all \( x \)
- \( g(x^t; \xi^t) \) is unbiased estimate of \( \nabla F(x^t) \) given \( \{\xi^0, \cdots, \xi^{t-1}\} \)
Convex problems

Suppose we return weighted average

\[ \tilde{x}^t := \sum_{k=0}^{t} \eta_k \sum_{j=0}^{t} \eta_j x^k \]

**Theorem 12.3**

*Under assumptions in Page 12-30, one has*

\[
\mathbb{E}[F(\tilde{x}^t) - F(x^*)] \leq \frac{1}{2} \mathbb{E}[\|x^0 - x^*\|^2] + \frac{1}{2} \sigma_g^2 \sum_{k=0}^{t} \eta_k^2 \sum_{k=0}^{t} \eta_k
\]

- if \( \eta_t \asymp 1/\sqrt{t} \), then

\[
\mathbb{E}[F(\tilde{x}^t) - F(x^*)] \lesssim \frac{\log t}{\sqrt{t}}
\]
Proof of Theorem 12.3

Remark: very similar to convergence analysis of subgradient methods

By convexity of $F$, we have $F(x) \geq F(x^t) + (x - x^t)^\top \nabla F(x^t)$

$$\implies \mathbb{E}[(x^t - x^*)^\top \nabla F(x^t)] \geq \mathbb{E}[F(x^t) - F(x^*)]$$

This together with (12.6) and (12.7) implies

$$2\eta_k \mathbb{E}[F(x^k) - F(x^*)] \leq \mathbb{E}[\|x^k - x^*\|^2_2] - \mathbb{E}[\|x^{k+1} - x^*\|^2_2] + \eta_k^2 M^2$$

Sum over $k = 0, \cdots, t$ to obtain

$$\sum_{k=0}^{t} 2\eta_k \mathbb{E}[F(x^k) - F(x^*)] \leq \mathbb{E}[\|x^0 - x^*\|^2_2] - \mathbb{E}[\|x^{t+1} - x^*\|^2_2] + M^2 \sum_{k=0}^{t} \eta_k^2$$

$$\leq \mathbb{E}[\|x^0 - x^*\|^2_2] + M^2 \sum_{k=0}^{t} \eta_k^2$$
Proof of Theorem 12.3 (cont.)

Setting $v_t = \frac{\eta_t}{\sum_{k=0}^{t} \eta_k}$ yields

$$\sum_{k=0}^{t} v_k \mathbb{E}[F(x^k) - F(x^*)] \leq \frac{1}{2} \mathbb{E}[\|x^0 - x^*\|_2^2] + \frac{1}{2} M^2 \sum_{k=0}^{t} \eta_k^2 \sum_{k=0}^{t} \eta_k$$

By convexity of $F$, we arrive at

$$\mathbb{E}[F(\tilde{x}^t) - F(x^*)] \leq \frac{1}{2} \mathbb{E}[\|x^0 - x^*\|_2^2] + \frac{1}{2} M^2 \sum_{k=0}^{t} \eta_k^2 \sum_{k=0}^{t} \eta_k$$
Reducing variance via iterate averaging
Two conflicting regimes

- noiseless case (i.e. $g(x; \xi) = \nabla F(x)$): stepsizes $\eta_t \approx 1/t$ are way too conservative

- general noisy case: longer stepsizes ($\eta_t \gg 1/t$) might fail to suppress noise (and hence slow down convergence)

Can we modify SGD so as to allow for larger stepsizes without compromising convergence rate?
Motivation for iterate averaging

SGD with long stepsizes poorly suppresses noise, which tends to oscillate around global minimizers due to noisy nature of gradient computation

One may, however, average iterates to mitigate oscillation and reduce variance
Acceleration by averaging

— Ruppert ’88, Polyak ’90, Polyak, Juditsky ’92

\[
\text{return } \overline{x}^t := \frac{1}{t} \sum_{i=0}^{t-1} x^i
\]  

(12.10)

with larger stepsizes \( \eta_t \approx t^{-\alpha}, \alpha < 1 \)

**Key idea:** slow algorithms with suboptimal convergence rates need to be averaged
Example: toy quadratic problem

\[
\begin{align*}
\text{minimize}_{x \in \mathbb{R}^d} & \quad \frac{1}{2} \|x\|_2^2 \\
\end{align*}
\]

- Constant stepsizes: \( \eta_t \equiv \eta < 1 \)
- \( g(x^t; \xi^t) = x^t + \xi^t \) with
  - \( \mathbb{E}[\xi^t \mid \xi^0, \ldots, \xi^{t-1}] = 0 \)
  - \( \lim_{t \to \infty} \mathbb{E}[\xi^t \xi^t^\top \mid \xi^0, \ldots, \xi^{t-1}] = I \)
Example: toy quadratic problem

\[
\minimize_{x \in \mathbb{R}^d} \quad \frac{1}{2} \| x \|^2
\]

SGD iterates:

\[
x^1 = x^0 - \eta (x^0 + \xi^0) = (1 - \eta) x^0 - \eta \xi^0
\]

\[
x^2 = x^1 - \eta (x^1 + \xi^1) = (1 - \eta)^2 x^0 - \eta (1 - \eta) \xi^0 - \eta \xi^1
\]

\[
\vdots
\]

\[
x^t = (1 - \eta)^t x^0 - \eta (1 - \eta)^{t-1} \xi^0 - \eta (1 - \eta)^{t-2} \xi^1 - \ldots
\]
Example: toy quadratic problem

\[ \text{minimize}_{x \in \mathbb{R}^d} \quad \frac{1}{2} \| x \|^2 \]

\[ \overline{x}^t \approx \frac{1}{t} \sum_{k=0}^{t-1} (1 - \eta)^k x^0 - \eta \left\{ 1 + (1 - \eta) + \cdots \right\} \frac{1}{t} \sum_{k=0}^{t-1} \xi^k \]

imprecise; but close enough for large \( t \)

\[ \approx -\frac{1}{t} \sum_{k=0}^{t-1} \xi^0 \]

(since \( 1 + (1 - \eta) + \cdots = \eta^{-1} \))

\[ \rightarrow \sqrt{t} \mathcal{N}(0, I) \]

(central limit theorem for martingale)
Example: more general quadratic problems

\[ \text{minimize}_{x \in \mathbb{R}^d} \quad \frac{1}{2} x^\top A x - b^\top x \]

- \( A \succeq \mu I \succ 0 \) (strongly convex)
- constant stepsizes: \( \eta_t \equiv \eta < 1/\mu \)
- \( g(x^t; \xi^t) = A x^t - b + \xi^t \) with
  - \( \mathbb{E}[\xi^t | \xi^0, \ldots, \xi^{t-1}] = 0 \)
  - \( S := \lim_{t \to \infty} \mathbb{E}[\xi^t \xi^{t\top} | \xi^0, \ldots, \xi^{t-1}] \) is finite
Example: more general quadratic problems

\[
\text{minimize}_{x \in \mathbb{R}^d} \quad \frac{1}{2} x^\top Ax - b^\top x
\]

Theorem 12.4

Fix \(d\). Then as \(t \to \infty\), iterate average \(\bar{x}^t\) obeys

\[
\sqrt{t}(\bar{x}^t - x^*) \xrightarrow{D} \mathcal{N}(0, A^{-1} S A^{-1})
\]

convergence in distribution
Example: quadratic problems

\[ \sqrt{t}(\overline{x}^t - x^*) \xrightarrow{D} \mathcal{N}(0, A^{-1} S A^{-1}), \quad t \to \infty \]

- asymptotically, \( \|\overline{x}^t - x^*\|_2^2 \asymp \frac{1}{t} \), which matches convergence rate in Theorem 12.2

- much longer stepsizes \( (\eta_t \asymp 1) \)
  \[ \implies \text{faster convergence for less noisy cases (e.g. } \xi^t = 0) \]
Proof sketch of Theorem 12.5

(1) Let $\Delta^t = x^t - x^*$ and $\overline{\Delta}^t = \overline{x}^t - x^*$. SGD update rule gives

$$\Delta^{t+1} = \Delta^t - \eta(A\Delta^t + \xi^t) = (I - \eta A)\Delta^t - \eta\xi^t$$

$$\implies \Delta^{t+1} = (I - \eta A)^{t+1}\Delta^0 - \eta \sum_{k=0}^{t} (I - \eta A)^{t-k}\xi^k$$

(2) Simple calculation gives (check Polyak, Juditsky ’92)

$$\overline{\Delta}^t = \frac{1}{t\eta}G_0^t\Delta^0 + \frac{1}{t} \sum_{j=0}^{t-2} A^{-1}\xi^j + \frac{1}{t} \sum_{j=0}^{t-2} (G_j^t - A^{-1})\xi^j$$

where

$$G_j^t := \eta \sum_{i=0}^{t-1-j} (I - \eta A)^i$$
(3) From central limit theorem for martingales,

\[
\frac{1}{\sqrt{t}} \sum_{j=0}^{t-2} A^{-1} \xi_j \overset{D}{\to} \mathcal{N} \left( 0, A^{-1} S A^{-1} \right)
\]

(4) With proper stepsizes, one has (check Polyak, Juditsky ’92)

\[
\| G^t_0 \| < \infty, \quad \| G^t_j - A^{-1} \| < \infty \quad \text{and} \quad \lim_{t \to \infty} \frac{1}{t} \sum_{j=0}^{t-1} \| G^t_j - A^{-1} \| = 0
\]

(5) Combining these bounds establishes Theorem 12.5
More general strongly convex problems

\[
\minimize_{x \in \mathbb{R}^d} \ f(x)
\]

- \( f \): strongly convex
- stepsizes: \( \eta_t \asymp t^{-\alpha} \) with \( \alpha \in (0.5, 1) \)
- \( g(x^t; \xi^t) = \nabla f(x^t) + \xi^t \)
  - \( \mathbb{E}[\xi^t | \xi^0, \ldots, \xi^{t-1}] = 0 \)
  - \( S := \lim_{t \to \infty} \mathbb{E}[\xi^t \xi^{t\top} | \xi^0, \ldots, \xi^{t-1}] \) is finite
More general strongly convex problems

\[
\text{minimize}_{x \in \mathbb{R}^d} \quad f(x)
\]

**Theorem 12.5 (informal, Polyak, Juditsky ’92)**

Fix \( d \) and let \( t \to \infty \). For large class of strongly convex problems,

\[
\sqrt{t}(\bar{x}^t - x^*) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, (\nabla^2 f(x^*))^{-1} S(\nabla^2 f(x^*))^{-1}\right)
\]

- depending on local curvature at / around minimizer


Reference


