Modern Optimization Methods for Big Data Problems

MATH11146
The University of Edinburgh
Spring 2017

Peter Richtárik

Modern Optimization Methods for Big Data Problems

Lecture 4

Randomized Methods for Solving Linear Systems: Convergence Analysis of the Basic Method; Parallel and Accelerated Methods

January 25, 2017

Covariance Matrix and Total Variance of a Random Vector

Definition 24 (Covariance matrix)

If $x \in \mathbb{R}^n$ is a random vector, then the matrix

$$\mathsf{Var}\,(x) \stackrel{\text{def}}{=} \mathrm{E}\left[(x - \mathrm{E}\,[x])(x - \mathrm{E}\,[x])^\top\right]$$

is called the **covariance matrix** of x.

Definition 25 (Total Variance)

If $x \in \mathbb{R}^n$ is a random vector, then the value

$$\mathsf{TVar}\,(x) \stackrel{\text{def}}{=} \mathrm{E}\left[(x - \mathrm{E}\,[x])^\top (x - \mathrm{E}\,[x])\right] = \mathrm{E}\left[\|x - \mathrm{E}\,[x]\,\|^2\right]$$

is called the **total variance** of x.

Exercise 6

Let $x \in \mathbb{R}^n$ be a random vector. Show that:

(i) *The total variance is the trace of the covariance matrix:* $\mathsf{TVar}(x) = \mathsf{Tr}\,(\mathsf{Var}\,(x))$

(ii) $\mathsf{TVar}\,(\mathbf{U}^\top \mathbf{B}^{1/2} x) = \mathrm{E}\left[\|x - \mathrm{E}\,[x]\,\|_{\mathbf{B}}^2\right]$.

Strong vs Weak Convergence

Definition 26 (Strong and Weak Convergence)

We say that a sequence of random vectors $\{x_k\}$ converges to x_*

- **weakly** if $\|\mathrm{E}\,[x_k - x_*]\,\|_{\mathbf{B}}^2 \to 0$ as $k \to \infty$
- **strongly** if $\mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right] \to 0$ as $k \to \infty$ (aka *L*2 **convergence**)

The following lemma explains why **strong convergence** is a stronger convergence concept than **weak convergence.**

Lemma 27

For any random vector $x_k \in \mathbb{R}^n$ and any $x_* \in \mathbb{R}^n$ we have the identity

$$\mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right] = \|\mathrm{E}\,[x_k - x_*]\|_{\mathbf{B}}^2 + \underbrace{\mathrm{E}\left[\|x_k - \mathrm{E}\,[x_k]\|_{\mathbf{B}}^2\right]}_{\mathsf{TVar}\,(\mathbf{U}^\top \mathbf{B}^{1/2} x_k)}.$$

As a consequence, **strong convergence implies**

- weak convergence,
- convergence of $\mathsf{TVar}\,(\mathbf{U}^\top \mathbf{B}^{1/2} x_k)$ to zero.

Proof of Lemma 27

Let $\mu = \mathrm{E}\left[x_k\right]$. Then

$$\begin{aligned}
\mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right] &= \mathrm{E}\left[\|x_k - \mu + \mu - x_*\|_{\mathbf{B}}^2\right] \\
&= \mathrm{E}\left[\|x_k - \mu\|_{\mathbf{B}}^2 + \|\mu - x_*\|_{\mathbf{B}}^2 + 2\langle x_k - \mu, \mu - x_*\rangle_{\mathbf{B}}\right] \\
&= \mathrm{E}\left[\|x_k - \mu\|_{\mathbf{B}}^2\right] + \|\mu - x_*\|_{\mathbf{B}}^2 + 2\langle \underbrace{\mathrm{E}\left[x_k - \mu\right]}_{0}, \mu - x_*\rangle_{\mathbf{B}} \\
&= \mathrm{E}\left[\|x_k - \mu\|_{\mathbf{B}}^2\right] + \|\mu - x_*\|_{\mathbf{B}}^2.
\end{aligned}$$

In the first step we have expanded the square and in the second step we have used linearity of expectation.

Weak Convergence

Weak Convergence

Theorem 28 (Weak Convergence 1)

Choose any $x_0 \in \mathbb{R}^n$ and let $\{x_k\}$ be the random iterates produced by Algorithm 1. Let $x_* \in \mathcal{L}$ be chosen arbitrarily. Then

$$\mathrm{E}\left[x_{k+1} - x_*\right] = \left(\mathbf{I} - \omega \mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right]\right)\mathrm{E}\left[x_k - x_*\right]. \tag{33}$$

Moreover, by transforming the error via the linear mapping $h \to \mathbf{U}^\top \mathbf{B}^{1/2} h$, this can be written in the form

$$\mathrm{E}\left[\mathbf{U}^\top \mathbf{B}^{1/2}(x_k - x_*)\right] = (\mathbf{I} - \omega \Lambda)^k \mathbf{U}^\top \mathbf{B}^{1/2}(x_0 - x_*), \tag{34}$$

which is separable in the coordinates of the transformed error:

$$\mathrm{E}\left[u_i^\top \mathbf{B}^{1/2}(x_k - x_*)\right] = (1 - \omega\lambda_i)^k u_i^\top \mathbf{B}^{1/2}(x_0 - x_*), \qquad i = 1, 2, \ldots, n. \tag{35}$$

Finally,

$$\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 = \sum_{i=1}^n (1 - \omega\lambda_i)^{2k}\left(u_i^\top \mathbf{B}^{1/2}(x_0 - x_*)\right)^2. \tag{36}$$

Weak Convergence

Theorem 29 (Convergence 2)

Let $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$. Then for all $i = 1, 2, \ldots, n$,

$$\mathrm{E}\left[u_i^\top \mathbf{B}^{1/2}(x_k - x_*)\right] = \begin{cases} 0 & \text{if} \quad \lambda_i = 0, \\ (1 - \omega\lambda_i)^k u_i^\top \mathbf{B}^{1/2}(x_0 - x_*) & \text{if} \quad \lambda_i > 0. \end{cases} \tag{37}$$

Moreover,

$$\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 \le \rho^k(\omega)\|x_0 - x_*\|_{\mathbf{B}}^2, \tag{38}$$

where the rate is given by

$$\rho(\omega) \overset{def}{=} \max_{i:\lambda_i > 0}(1 - \omega\lambda_i)^2. \tag{39}$$

Necessary and Sufficient Conditions for Convergence

Corollary 30 (Necessary and sufficient conditions)

Let Assumption 3 (exactness) hold. Choose any $x_0 \in \mathbb{R}^n$ *and let* $x_* = \Pi^{\mathbf{B}}_{\mathcal{L}}(x_0)$.

If $\{x_k\}$ *are the random iterates produced by Algorithm 1, then the following statements are equivalent:*

(i) $|1-\omega\lambda_i| < 1$ *for all* i *for which* $\lambda_i > 0$

(ii) $0 < \omega < 2/\lambda_{\max}$

(iii) $\mathrm{E}\left[u_i^\top \mathbf{B}^{1/2}(x_k - x_*)\right] \to 0$ *for all* i

(iv) $\|\mathrm{E}\left[x_k - x_*\right]\|^2_{\mathbf{B}} \to 0$

Proof of Theorems 28 and 29 - I

We first start with a lemma.

Lemma 31

Let Assumption 3 (exactness) hold. Consider arbitrary $x \in \mathbb{R}^n$ *and let* $x_* = \Pi^{\mathbf{B}}_{\mathcal{L}}(x)$. *If* $\lambda_i = 0$, *then* $u_i^\top \mathbf{B}^{1/2}(x - x_*) = 0$.

Proof.

From (17) we see that $x - x_* = \mathbf{B}^{-1}\mathbf{A}^\top w$ for some $w \in \mathbb{R}^m$. Therefore, $u_i^\top \mathbf{B}^{1/2}(x - x_*) = u_i^\top \mathbf{B}^{-1/2}\mathbf{A}^\top w$. By Theorem 18, we have $\mathrm{Range}\,(u_i \ : \ \lambda_i = 0) = \mathrm{Null}\left(\mathbf{A}\mathbf{B}^{-1/2}\right)$, from which it follows that $u_i^\top \mathbf{B}^{-1/2}\mathbf{A} = 0$. □

Proof of Theorem 28: Algorithm 1 can be written in the form

$$e_{k+1} = (\mathbf{I} - \omega \mathbf{B}^{-1}\mathbf{Z}_k)e_k, \tag{40}$$

where $e_k = x_k - x_*$. Multiplying both sides of this equation by $\mathbf{B}^{1/2}$ from the left, and taking expectation conditional on e_k, we obtain

$$\mathrm{E}\left[\mathbf{B}^{1/2}e_{k+1} \mid e_k\right] = (\mathbf{I} - \omega\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2})\mathbf{B}^{1/2}e_k.$$

Proof of Theorems 28 and 29 - II

Taking expectations on both sides and using the tower property, we get

$$\mathrm{E}\left[\mathbf{B}^{1/2}e_{k+1}\right] = \mathrm{E}\left[\mathrm{E}\left[\mathbf{B}^{1/2}e_{k+1} \mid e_k\right]\right] = (\mathbf{I}-\omega\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2})\mathrm{E}\left[\mathbf{B}^{1/2}e_k\right].$$

We now replace $\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}$ by its eigenvalue decomposition $\mathbf{U}\Lambda\mathbf{U}^\top$ (see (31)), multiply both sides of the last inequality by $\mathbf{U}^\top$ from the left, and use linearity of expectation to obtain

$$\mathrm{E}\left[\mathbf{U}^\top\mathbf{B}^{1/2}e_{k+1}\right] = (\mathbf{I}-\omega\Lambda)\mathrm{E}\left[\mathbf{U}^\top\mathbf{B}^{1/2}e_k\right].$$

Unrolling the recurrence, we get (34). When this is written coordinate-by-coordinate, (35) follows. Identity (36) follows immediately by equating standard Euclidean norms of both sides of (34).

Proof of Theorem 29: If $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$, then from Lemma 31 we see that $\lambda_i = 0$ implies $u_i^\top\mathbf{B}^{1/2}(x_0 - x_*) = 0$. Using this in (35) gives (37).

Proof of Theorems 28 and 29 - III

Finally, inequality (38) follows from

$$\begin{aligned}
\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 &\overset{(36)}{=} \sum_{i=1}^{n}(1-\omega\lambda_i)^{2k}\left(u_i^\top\mathbf{B}^{1/2}(x_0-x_*)\right)^2 \\
&= \sum_{i:\lambda_i>0}(1-\omega\lambda_i)^{2k}\left(u_i^\top\mathbf{B}^{1/2}(x_0-x_*)\right)^2 \\
&\overset{(39)}{\leq} \rho^k(\omega)\sum_{i:\lambda_i>0}\left(u_i^\top\mathbf{B}^{1/2}(x_0-x_*)\right)^2 \\
&= \rho^k(\omega)\sum_{i:\lambda_i>0}\left(u_i^\top\mathbf{B}^{1/2}(x_0-x_*)\right)^2 + \rho^k(\omega)\sum_{i:\lambda_i=0}\left(u_i^\top\mathbf{B}^{1/2}(x_0-x_*)\right)^2 \\
&= \rho^k(\omega)\sum_{i}\left(u_i^\top\mathbf{B}^{1/2}(x_0-x_*)\right)^2 \\
&= \rho^k(\omega)\sum_{i}(x_0-x_*)^\top\mathbf{B}^{1/2}u_iu_i^\top\mathbf{B}^{1/2}(x_0-x_*) \\
&= \rho^k(\omega)\sum_{i}(x_0-x_*)^\top\mathbf{B}^{1/2}\left(\sum_i u_iu_i^\top\right)\mathbf{B}^{1/2}(x_0-x_*) = \rho^k(\omega)\|x_0-x_*\|_{\mathbf{B}}^2.
\end{aligned}$$

The last identity follows from the fact that $\sum_i u_iu_i^\top = \mathbf{U}\mathbf{U}^\top = \mathbf{I}$.

Optimal Stepsize Choice for Weak Convergence

Convergence Rate as a Function of ω

We now consider the problem of choosing the stepsize (relaxation) parameter ω.

In view of (38) and (39), the optimal relaxation parameter is the one solving the following optimization problem:

$$\min_{\omega \in \mathbb{R}} \left\{ \rho(\omega) = \max_{i:\lambda_i > 0} (1 - \omega\lambda_i)^2 \right\}. \tag{41}$$

We solve the above problem in the next result (Theorem 32).

Optimal Stepsize

Theorem 32 (Stepsize Choice)

Let $\omega^* \overset{def}{=} 2/(\lambda_{\min}^+ + \lambda_{\max})$. *Then the objective of* (41) *is given by*

$$\rho(\omega) = \begin{cases} (1-\omega\lambda_{\max})^2 & \text{if} \quad \omega \le 0 \\ (1-\omega\lambda_{\min}^+)^2 & \text{if} \quad 0 \le \omega \le \omega^* \\ (1-\omega\lambda_{\max})^2 & \text{if} \quad \omega \ge \omega^* \end{cases}. \tag{42}$$

Moreover, ρ *is decreasing on* $(-\infty, \omega^*]$ *and increasing on* $[\omega^*, +\infty)$*, and hence the optimal solution of* (41) *is* ω^*. *Further, we have:*

(i) *If we choose* $\omega = 1$ *(no over-relaxation), then*

$$\rho(1) = (1-\lambda_{\min}^+)^2. \tag{43}$$

(ii) *If we choose* $\omega = 1/\lambda_{\max}$ *(over-relaxation), then*

$$\rho(1/\lambda_{\max}) = \left(1 - \frac{\lambda_{\min}^+}{\lambda_{\max}}\right)^2 \overset{(32)}{=} \left(1 - \frac{1}{\zeta}\right)^2. \tag{44}$$

(iii) *If we choose* $\omega = \omega^*$ *(optimal over-relaxation), the optimal rate is*

$$\rho(\omega^*) = \left(1 - \frac{2\lambda_{\min}^+}{\lambda_{\min}^+ + \lambda_{\max}}\right)^2 \overset{(32)}{=} \left(1 - \frac{2}{\zeta+1}\right)^2. \tag{45}$$

Proof of Theorem 32

Recall that $\lambda_{\max} \le 1$. Letting

$$\rho_i(\omega) = (1-\omega\lambda_i)^2,$$

it can be shown that

$$\rho(\omega) = \max\{\rho_j(\omega), \rho_n(\omega)\},$$

where j is such that $\lambda_j = \lambda_{\min}^+$. Note that $\rho_j(\omega) = \rho_n(\omega)$ for $\omega \in \{0, \omega^*\}$. From this we deduce that $\rho_j \ge \rho_n$ on $(-\infty, 0]$, $\rho_j \le \rho_n$ on $[0, \omega^*]$, and $\rho_j \ge \rho_n$ on $[\omega^*, +\infty)$, obtaining (42). We see that ρ is decreasing on $(-\infty, \omega^*]$, and increasing on $[\omega^*, +\infty)$.

The remaining results follow directly by plugging specific values of ω into (42).

Strong Convergence

Decrease of Distance is Proportional to $f_{\mathbf{S}}$

Lemma 33 (Decrease of Distance)

Choose $x_0 \in \mathbb{R}^n$ and let $\{x_k\}_{k=0}^{\infty}$ be the random iterates produced by Algorithm 1, with an arbitrary relaxation parameter $\omega \in \mathbb{R}$. Let $x_* \in \mathcal{L}$.

Then we have the identities $\|x_{k+1} - x_k\|_{\mathbf{B}}^2 = 2\omega^2 f_{\mathbf{S}_k}(x_k)$, and

$$\|x_{k+1} - x_*\|_{\mathbf{B}}^2 = \|x_k - x_*\|_{\mathbf{B}}^2 - 2\omega(2-\omega) f_{\mathbf{S}_k}(x_k). \tag{46}$$

Moreover, $\mathrm{E}\left[\|x_{k+1} - x_k\|_{\mathbf{B}}^2\right] = 2\omega^2 \mathrm{E}\left[f(x_k)\right]$, and

$$\mathrm{E}\left[\|x_{k+1} - x_*\|_{\mathbf{B}}^2\right] = \mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right] - 2\omega(2-\omega)\mathrm{E}\left[f(x_k)\right]. \tag{47}$$

Remarks: Equation (46) says that for any $x_* \in \mathcal{L}$, in the k-th iteration of Algorithm 1 the distance of the current iterate from x_* decreases by the amount $2\omega(2-\omega) f_{\mathbf{S}_k}(x_k)$.

Lower Bound on a Quadratic

Lemma 34

Let Assumption 3 be satisfied. Then the inequality

$$x^\top \mathbf{B}^{-1/2}\mathrm{E}[\mathbf{Z}]\mathbf{B}^{-1/2}x \geq \lambda_{\min}^+(\mathbf{B}^{-1/2}\mathrm{E}[\mathbf{Z}]\mathbf{B}^{-1/2})x^\top x \qquad (48)$$

holds for all $x \in \mathrm{Range}\left(\mathbf{B}^{-1/2}\mathbf{A}^\top\right)$.

Proof.

It is known that for any matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$, the inequality

$$x^\top \mathbf{M}^\top \mathbf{M} x \geq \lambda_{\min}^+(\mathbf{M}^\top\mathbf{M})x^\top x$$

holds for all $x \in \mathrm{Range}\left(\mathbf{M}^\top\right)$. Applying this with $\mathbf{M} = (\mathrm{E}[\mathbf{Z}])^{1/2}\mathbf{B}^{-1/2}$, we see that (48) holds for all $x \in \mathrm{Range}\left(\mathbf{B}^{-1/2}(\mathrm{E}[\mathbf{Z}])^{1/2}\right)$. However,

$$\begin{aligned}
\mathrm{Range}\left(\mathbf{B}^{-1/2}(\mathrm{E}[\mathbf{Z}])^{1/2}\right) &= \mathrm{Range}\left(\mathbf{B}^{-1/2}(\mathrm{E}[\mathbf{Z}])^{1/2}(\mathbf{B}^{-1/2}(\mathrm{E}[\mathbf{Z}])^{1/2})^\top\right) \\
&= \mathrm{Range}\left(\mathbf{B}^{-1/2}\mathrm{E}[\mathbf{Z}]\mathbf{B}^{-1/2}\right) = \mathrm{Range}\left(\mathbf{B}^{-1/2}\mathbf{A}^\top\right),
\end{aligned}$$

where the last identity follows by combining Assumption 3 and Theorem 18. □

Proof of Lemma 33 - I

Recall that Algorithm 1 performs the update

$$x_{k+1} = x_k - \omega \mathbf{B}^{-1}\mathbf{Z}_k(x_k - x_*).$$

From this we get

$$\begin{aligned}
\|x_{k+1} - x_k\|_{\mathbf{B}}^2 &= \omega^2 \|\mathbf{B}^{-1}\mathbf{Z}_k(x_k - x_*)\|_{\mathbf{B}}^2 \\
&\overset{(19)}{=} \omega^2 (x_k - x_*)^\top \mathbf{Z}_k (x_k - x_*) \\
&\overset{(20)}{=} 2\omega^2 f_{\mathbf{S}_k}(x_k). \qquad (49)
\end{aligned}$$

In a similar vein,

$$\begin{aligned}
\|x_{k+1} - x_*\|_{\mathbf{B}}^2 &= \|(\mathbf{I} - \omega\mathbf{B}^{-1}\mathbf{Z}_k)(x_k - x_*)\|_{\mathbf{B}}^2 \\
&= (x_k - x_*)^\top(\mathbf{I} - \omega\mathbf{Z}_k\mathbf{B}^{-1})\mathbf{B}(\mathbf{I} - \omega\mathbf{B}^{-1}\mathbf{Z}_k)(x_k - x_*) \\
&\overset{(19)}{=} (x_k - x_*)^\top(\mathbf{B} - \omega(2-\omega)\mathbf{Z}_k)(x_k - x_*) \\
&\overset{(20)}{=} \|x_k - x_*\|_{\mathbf{B}}^2 - 2\omega(2-\omega)f_{\mathbf{S}_k}(x_k), \qquad (50)
\end{aligned}$$

Proof of Lemma 33 - II

establishing (46).

Taking expectation in (49) and using the tower property, we get

$$\begin{aligned} \mathrm{E}\left[\|x_{k+1} - x_k\|_{\mathbf{B}}^2\right] &= \mathrm{E}\left[\mathrm{E}\left[\|x_{k+1} - x_k\|_{\mathbf{B}}^2 \mid x_k\right]\right] \\ &\overset{(49)}{=} 2\omega^2 \mathrm{E}\left[\mathrm{E}\left[f_{\mathbf{S}_k}(x_k) \mid x_k\right]\right] \\ &= 2\omega^2 \mathrm{E}\left[f(x_k)\right], \end{aligned}$$

where in the last step we have used the definition of f.

Taking expectation in (46), we get

$$\begin{aligned} \mathrm{E}\left[\|x_{k+1} - x_*\|_{\mathbf{B}}^2\right] &= \mathrm{E}\left[\mathrm{E}\left[\|x_{k+1} - x_*\|_{\mathbf{B}}^2 \mid x_k\right]\right] \\ &\overset{(50)}{=} \mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2 - 2\omega(2-\omega) f(x_k)\right] \\ &= \mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right] - 2\omega(2-\omega)\mathrm{E}\left[f(x_k)\right]. \end{aligned}$$

Quadratic Bounds

Lemma 35 (Quadratic bounds)

For all $x \in \mathbb{R}^n$ and $x_* \in \mathcal{L}$ we have

$$\lambda_{\min}^+ \cdot f(x) \le \frac{1}{2}\|\nabla f(x)\|_{\mathbf{B}}^2 \le \lambda_{\max} \cdot f(x). \tag{51}$$

and

$$f(x) \le \frac{\lambda_{\max}}{2}\|x - x_*\|_{\mathbf{B}}^2. \tag{52}$$

Moreover, if Assumption 3 holds, then for all $x \in \mathbb{R}^n$ and $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x)$ we have

$$\frac{\lambda_{\min}^+}{2}\|x - x_*\|_{\mathbf{B}}^2 \le f(x). \tag{53}$$

Proof of Lemma 35 - I

In view of (15) and (31), we obtain a spectral characterization of f:

$$f(x) = \frac{1}{2}\sum_{i=1}^{n} \lambda_i \left(u_i^\top \mathbf{B}^{1/2}(x - x_*)\right)^2, \tag{54}$$

where x_* is any point in $\mathcal{L}$. On the other hand, in view of (26) and (31), we have

$$\begin{aligned}
\|\nabla f(x)\|_{\mathbf{B}}^2 &= \|\mathbf{B}^{-1}\mathrm{E}[\mathbf{Z}](x - x_*)\|_{\mathbf{B}}^2 \qquad (55)\\
&= (x - x_*)^\top \mathrm{E}[\mathbf{Z}]\,\mathbf{B}^{-1}\mathrm{E}[\mathbf{Z}](x - x_*)\\
&= (x - x_*)^\top \mathbf{B}^{1/2}(\mathbf{B}^{-1/2}\mathrm{E}[\mathbf{Z}]\,\mathbf{B}^{-1/2})(\mathbf{B}^{-1/2}\mathrm{E}[\mathbf{Z}]\,\mathbf{B}^{-1/2})\mathbf{B}^{1/2}(x - x_*)\\
&= (x - x_*)^\top \mathbf{B}^{1/2}\mathbf{U}(\mathbf{U}^\top\mathbf{B}^{-1/2}\mathrm{E}[\mathbf{Z}]\,\mathbf{B}^{-1/2}\mathbf{U})^2\mathbf{U}^\top\mathbf{B}^{1/2}(x - x_*)\\
&\overset{(31)}{=} (x - x_*)^\top \mathbf{B}^{1/2}\mathbf{U}\Lambda^2\mathbf{U}^\top\mathbf{B}^{1/2}(x - x_*)\\
&= \sum_{i=1}^{n} \lambda_i^2 \left(u_i^\top \mathbf{B}^{1/2}(x - x_*)\right)^2. \qquad (56)
\end{aligned}$$

Inequality (51) follows by comparing (54) and (55), using the bounds

$$\lambda_{\min}^{+}\lambda_i \le \lambda_i^2 \le \lambda_{\max}\lambda_i,$$

which hold for i for which $\lambda_i > 0$.

Proof of Lemma 35 - II

We now move to the bounds involving norms. First, note that for any $x_* \in \mathcal{L}$ we have

$$\begin{aligned}
f(x) &\overset{(15)}{=} \frac{1}{2}(x - x_*)^\top \mathrm{E}[\mathbf{Z}](x - x_*) \qquad (57)\\
&= \frac{1}{2}(\mathbf{B}^{1/2}(x - x_*))^\top(\mathbf{B}^{-1/2}\mathrm{E}[\mathbf{Z}]\,\mathbf{B}^{-1/2})\mathbf{B}^{1/2}(x - x_*).
\end{aligned}$$

The upper bound follows by applying the inequality

$$\mathbf{B}^{-1/2}\mathrm{E}[\mathbf{Z}]\,\mathbf{B}^{-1/2} \preceq \lambda_{\max}\mathbf{I}.$$

If $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x)$, then in view of (17), we have

$$\mathbf{B}^{1/2}(x - x_*) \in \mathrm{Range}\left(\mathbf{B}^{-1/2}\mathbf{A}^\top\right).$$

Applying Lemma 34 to (57), we get the lower bound.

Strong Convergence

Theorem 36 (Strong convergence)

Let Assumption 3 (exactness) hold and set $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$. *Let* $\{x_k\}$ *be the random iterates produced by Algorithm 1, where the relaxation parameter satisfies* $0 < \omega < 2$, *and let* $r_k \overset{def}{=} \mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right]$. *Then for all* $k \geq 0$ *we have*

$$(1 - \omega(2-\omega)\lambda_{\max})^k r_0 \leq r_k \leq (1 - \omega(2-\omega)\lambda_{\min}^+)^k r_0. \tag{58}$$

The best rate is achieved when $\omega = 1$.

Proof.

Let $\phi_k = \mathrm{E}[f(x_k)]$. We have

$$r_{k+1} \overset{(47)}{=} r_k - 2\omega(2-\omega)\phi_k \overset{(53)}{\leq} r_k - \omega(2-\omega)\lambda_{\min}^+ r_k,$$

and

$$r_{k+1} \overset{(47)}{=} r_k - 2\omega(2-\omega)\phi_k \overset{(52)}{\geq} r_k - \omega(2-\omega)\lambda_{\max} r_k.$$

Inequalities (58) follow from this by unrolling the recurrences. □

Convergence of $f(x_k)$

Convergence of $f(x_k)$

Theorem 37 (Convergence of f)

Choose $x_0 \in \mathbb{R}^n$, and let $\{x_k\}_{k=0}^{\infty}$ be the random iterates produced by Algorithm 1, where the relaxation parameter satisfies $0 < \omega < 2$.

(i) *Let $x_* \in \mathcal{L}$. The average iterate $\hat{x}_k \overset{def}{=} \frac{1}{k}\sum_{t=0}^{k-1} x_t$ for all $k \geq 1$ satisfies*

$$\mathrm{E}\left[f(\hat{x}_k)\right] \leq \frac{\|x_0 - x_*\|_{\mathbf{B}}^2}{2\omega(2-\omega)k}. \tag{59}$$

(ii) *Now let Assumption 3 hold. For $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$ and $k \geq 0$ we have*

$$\mathrm{E}\left[f(x_k)\right] \leq \left(1 - \omega(2-\omega)\lambda_{\min}^+\right)^k \frac{\lambda_{\max}\|x_0 - x_*\|_{\mathbf{B}}^2}{2}. \tag{60}$$

The best rate is achieved when $\omega = 1$.

Proof of Theorem 37

(i) Let $\phi_k = \mathrm{E}\left[f(x_k)\right]$ and $r_k = \mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right]$. By summing up the identities from (47), we get

$$2\omega(2-\omega)\sum_{t=0}^{k-1}\phi_t = r_0 - r_k.$$

Therefore, using Jensen's inequality, we get

$$\mathrm{E}\left[f(\hat{x}_k)\right] \leq \mathrm{E}\left[\frac{1}{k}\sum_{t=0}^{k-1} f(x_t)\right] = \frac{1}{k}\sum_{t=0}^{k-1}\phi_t = \frac{r_0 - r_k}{2\omega(2-\omega)k} \leq \frac{r_0}{2\omega(2-\omega)k}.$$

(ii) Combining inequality (52) with Theorem 36, we get

$$\mathrm{E}\left[f(x_k)\right] \leq \frac{\lambda_{\max}}{2}\mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right] \overset{(58)}{\leq} \left(1 - \omega(2-\omega)\lambda_{\min}^+\right)^k \frac{\lambda_{\max}\|x_0 - x_*\|_{\mathbf{B}}^2}{2}.$$

Parallel Method ("Minibatch Method")

Parallel Method ("Minibatch Method")

Algorithm 2 Parallel Method

Parameters: distribution $\mathcal{D}$ from which to sample matrices; positive definite matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$; stepsize/relaxation parameter $\omega \in \mathbb{R}$; parallelism parameter τ (aka "minibatch size")
Choose $x_0 \in \mathbb{R}^n$ ▷ Initialization
for $k = 0, 1, 2, \dots$ **do**
  for $i = 1, 2, \dots, \tau$ **do**
    Draw $\mathbf{S}_{ki} \sim \mathcal{D}$
    Set $z_{k+1,i} = x_k - \omega \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}_{ki}(\mathbf{S}_{ki}^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}_{ki})^\dagger \mathbf{S}_{ki}^\top(\mathbf{A}x_k - b)$
  Set $x_{k+1} = \frac{1}{\tau}\sum_{i=1}^{\tau} z_{k+1,i}$ ▷ Average the results

- Note that for $\tau = 1$, the parallel method (Algorithm 2) **reduces to the basic method (Algorithm 1).**
- We take one step of the basic method τ times, independently, started from x_k. The results are then averaged to obtain x_{k+1}.
- The τ computations **can (but do not have to!) be performed in parallel,** whence the name of the method.

Convergence of the Parallel Method

Theorem 38

Let Assumption 3 hold and set $x_* = \Pi^{\mathbf{B}}_{\mathcal{L}}(x_0)$. *Let* $\{x_k\}_{k=0}^{\infty}$ *be the random iterates produced by Algorithm 2, where the relaxation parameter satisfies* $0 < \omega < 2/\xi(\tau)$, *where* $\xi(\tau) \overset{def}{=} \frac{1}{\tau} + \left(1 - \frac{1}{\tau}\right)\lambda_{\max}$. *Then*

$$\mathrm{E}\left[\|x_{k+1} - x_*\|^2_{\mathbf{B}}\right] \le \rho(\omega, \tau) \cdot \mathrm{E}\left[\|x_k - x_*\|^2_{\mathbf{B}}\right],$$

and

$$\mathrm{E}\left[f(x_k)\right] \le \rho(\omega,\tau)^k \frac{\lambda_{\max}}{2}\|x_0 - x_*\|^2_{\mathbf{B}},$$

where

$$\rho(\omega,\tau) \overset{def}{=} 1 - \omega\left[2 - \omega\xi(\tau)\right]\lambda^+_{\min}.$$

Understanding the Behaviour of the Parallel Method - I

The **convergence factor**

$$\rho(\omega,\tau) = 1 - \omega\left[2 - \omega\underbrace{\left(\tfrac{1}{\tau} + \left(1 - \tfrac{1}{\tau}\right)\lambda_{\max}\right)}_{\xi(\tau)}\right]\lambda^+_{\min}$$

depends on the choice of the stepsize ω and on the minibatch size τ.

- The **stepsize rate function**

  $$\omega \mapsto \rho(\omega,\tau),$$

  is minimized for $\omega(\tau) \overset{\text{def}}{=} 1/\xi(\tau)$ and the associated **optimal rate** is

  $$\rho(\omega(\tau),\tau) = 1 - \frac{\lambda^+_{\min}}{\frac{1}{\tau} + \left(1 - \frac{1}{\tau}\right)\lambda_{\max}}. \tag{61}$$

- The **minibatch rate function**

  $$\tau \mapsto \rho(\omega(\tau),\tau)$$

  is **decreasing on** $[1,\infty)$, with

  $$\rho(\omega(1),1) = 1 - \lambda^+_{\min}, \qquad \lim_{\tau\to\infty}\rho(\omega(\tau),\tau) = 1 - \frac{\lambda^+_{\min}}{\lambda_{\max}}.$$

Understanding the Behaviour of the Parallel Method - II

Convergence Rate for $\tau = 1$ (with optimal stepsize $\omega = \omega(\tau)$):

$$k \geq \frac{1}{\lambda_{\min}^{+}} \log\left(\frac{\|x_0 - x_*\|_{\mathbf{B}}^2}{\epsilon}\right) \quad \Rightarrow \quad \mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right] \leq \epsilon$$

Convergence Rate for $\tau = +\infty$ (with optimal stepsize $\omega = \omega(\tau)$):

$$k \geq \frac{\lambda_{\max}}{\lambda_{\min}^{+}} \log\left(\frac{\|x_0 - x_*\|_{\mathbf{B}}^2}{\epsilon}\right) \quad \Rightarrow \quad \mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right] \leq \epsilon$$

Recall what we proved about the basic method:

- The **weak convergence rate of the basic method is "fast"**:

$$\tilde{\mathcal{O}}\left(\lambda_{\max}/\lambda_{\min}^{+}\right)$$

- The **strong convergence rate of the basic method is "slow"**:

$$\tilde{\mathcal{O}}\left(1/\lambda_{\min}^{+}\right)$$

So, how does minibatching improve the basic method?

- The **strong convergence rate of the parallel method** interpolates between slow and fast!

Accelerated Method

Accelerated Method

In order to obtain further acceleration, we suggest to perform an update step in which x_{k+1} depends on both x_k and x_{k-1}. In particular, we take two *dependent* steps of Algorithm 1, one from x_k and one from x_{k-1}, and then take an affine combination of the results. That is, the process is started with $x_0, x_1 \in \mathbb{R}^n$, and for $k \geq 1$ involves an iteration of the form

$$\boxed{x_{k+1} = \gamma \phi_\omega(x_k, \mathbf{S}_k) + (1-\gamma)\phi_\omega(x_{k-1}, \mathbf{S}_{k-1})} \tag{62}$$

where the matrices $\{\mathbf{S}_k\}$ are independent samples from $\mathcal{D}$, and $\gamma \in \mathbb{R}$ is an **acceleration parameter.**

Remarks:

- By choosing $\gamma = 1$ (no acceleration), we recover the Basic Method.
- Theory suggests that γ should be always between 1 and 2. In particular, for well conditioned problems (small ζ), one should choose $\gamma \approx 1$, and for ill conditioned problems (large ζ), one should choose $\gamma \approx 2$.
- By a proper combination of overrelaxation (choice of ω) with acceleration (choice of γ), Algorithm 3 enjoys the **accelerated convergence rate of $\tilde{\mathcal{O}}(\sqrt{\zeta})$, where ζ is the condition number.**

Accelerated Method

Algorithm 3 Accelerated Method

Parameters: distribution $\mathcal{D}$ from which to sample matrices; positive definite matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$; stepsize/relaxation parameter $\omega > 0$; acceleration parameter $\gamma > 0$
Choose $x_0, x_1 \in \mathbb{R}^n$ such that $x_0 - x_1 \in \mathrm{Range}\left(\mathbf{B}^{-1}\mathbf{A}^\top\right)$ (for instance, choose $x_0 = x_1$)
Draw $\mathbf{S}_0 \sim \mathcal{D}$
Set $z_0 = \phi_\omega(x_0, \mathbf{S}_0)$
for $k = 1, 2, \ldots$ **do**
    Draw a fresh sample $\mathbf{S}_k \sim \mathcal{D}$
    Set $z_k = \phi_\omega(x_k, \mathbf{S}_k)$
    Set $x_{k+1} = \gamma z_k + (1-\gamma) z_{k-1}$ ▷ Main update step
Output x_k

Convergence

Theorem 39 (Complexity of Algorithm 3)

Let Assumption 3 (exactness) be satisfied and let $\{x_k\}_{k=0}^{\infty}$ *be the sequence of random iterates produced by Algorithm 3, started with* $x_0, x_1 \in \mathbb{R}^n$ *satisfying the relation* $x_0 - x_1 \in \mathrm{Range}\left(\mathbf{B}^{-1}\mathbf{A}^\top\right)$, *with* **relaxation parameter** $0 < \omega \leq 1/\lambda_{\max}$ *and* **acceleration parameter** $\gamma = 2/(1+\sqrt{\mu})$, *where* $\mu \in (0, \omega\lambda_{\min}^+)$. *Let* $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$. *Then there exists a constant* $C > 0$, *such that for all* $k \geq 2$ *we have*

$$\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 \leq (1-\sqrt{\mu})^{2k} C. \quad (63)$$

(i) *If we choose* $\omega = 1/\lambda_{\max}$ *(overrelaxation), then we can pick* $\mu = 0.99/\zeta$ *(recall that* $\zeta = \lambda_{\max}/\lambda_{\min}^+$ *is the condition number), which leads to the rate*

$$\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 \leq \left(1 - \sqrt{\frac{0.99\lambda_{\min}^+}{\lambda_{\max}}}\right)^{2k} C. \quad (64)$$

(ii) *If we choose* $\omega = 1$ *(no overrelaxation), then we can pick* $\mu = 0.99\lambda_{\min}^+$, *which leads to the rate*

$$\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 \leq \left(1 - \sqrt{0.99\lambda_{\min}^+}\right)^{2k} C. \quad (65)$$

Comments

Alternative Way of Writing Convergence Rate (64):

$$k \geq \frac{1}{2\sqrt{0.99}}\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}^+}}\log\left(\frac{C}{\epsilon}\right) \quad \Rightarrow \quad \|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 \leq \epsilon$$

Alternative Way of Writing Convergence Rate (65):

$$k \geq \frac{1}{2\sqrt{0.99}}\sqrt{\frac{\mathbf{1}}{\lambda_{\min}^+}}\log\left(\frac{C}{\epsilon}\right) \quad \Rightarrow \quad \|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 \leq \epsilon$$

- All three methods: basic (Algorithm 1), parallel (Algorithm 2) and accelerated (Algorithm 3) enjoy **linear convergence.** That is, their complexity has logarithmic dependence on $1/\epsilon$. This means that the error decays exponentially fast.
- However, **the leading constants in the complexity bounds are different.**
- Both the **basic and parallel methods** depend either on $1/\lambda_{\min}^+$ (slow) or $\lambda_{\max}/\lambda_{\min}^+$ (fast), depending on how we set the parameters ω, τ and γ, and whether we are interested in weak or strong convergence.
- However, the **accelerated method depends on the square root of these quantities.** This is why the method is called accelerated.