Exact convergence rates of the last iterate in subgradient methods

François Glineur and Moslem Zamani

Information and Communication Technologies, Electronics and Applied Mathematics Institute, and Center for Operations Research and Econometrics
UCLouvain

Results from preprint https://arxiv.org/abs/2307.11134

EUROPT 2024 June 27, 2024, Lund

Subgradient methods

Subgradient methods

Objective: minimize a function $f : \mathbb{R}^d \to \mathbb{R}$ that is

- convex

  $$\partial f(x) = \{g \text{ such that } f(y) \geq f(x) + g^T(y - x) \text{ for all } y\} \neq \emptyset$$

- B-Lipschitz continous

  $$g \in \partial f(x) \Rightarrow \|g\| \leq B$$

- with minimizer x^*

Method: subgradient method with fixed step sizes $\{h_k\}$

$$x_{k+1} = x_k - h_k g_k \text{ for some } g_k \in \partial f(x_k)$$

Performance criteria

Target: convergence rate after N iterations, either

- $f(x_N) - f(x_*)$
- $\min_{0 \le k \le N} f(x_k) - f(x_*)$ (method is not monotone)
- $f(\frac{1}{N+1}\sum_{k=0}^{N} x_k) - f(x_*)$ (average)

Initial iterate assumption:

$$\|x_0 - x_*\| \le R$$

Lower bound: no method can achieve better rate than

$$\frac{BR}{\sqrt{N+1}}$$

Lower bound proof (variation of [Drori, Teboulle 2016])

Consider following function with $B = 1$ and $x_* = 0$

$$f(x) = \max_{1 \leq k \leq N+1} x_k$$

Choose starting point $x_0 = (1, 1, \ldots, 1)$ with $R = \sqrt{N+1}$

- Subgradient $g \in \partial f(x)$ always picked as a basis vector e_i
- Induction argument:

  x_k must contain at least $N + 1 - k$ components equal to 1
- After N steps we have at least one component equal to 1
- Conclusion: we have $f(x_k) \geq 1$ for all $0 \leq k \leq N$ hence

$$f(x_N) \; \geq \; 1 = \frac{BR}{\sqrt{N+1}}$$

  (also for other criteria / for steps with several past subgradients)
- Note $f(x_k)$ and $\|g_k\|$ are constant throughout iterations

Standard convergence analysis

Only two ingredients:

subgradient inequality and *square distance telescoping*

$$
\begin{aligned}
\left\|x^{k+1} - x^\star\right\|^2 &= \left\|x^k - h_k g^k - x^\star\right\|^2 \\
&= \left\|x^k - x^\star\right\|^2 + h_k^2 \left\|g^k\right\|^2 - 2h_k \langle g^k, x^k - x^\star\rangle \\
&\le \left\|x^k - x^\star\right\|^2 + h_k^2 \left\|g^k\right\|^2 - 2h_k \left(f(x^k) - f(x^\star)\right).
\end{aligned}
$$

using subgradient inequality between x^* and x^k

$$f(x^\star) - f(x^k) \ge \langle g^k, x^\star - x^k\rangle$$

Standard convergence analysis (cont.)

Hence

$$h_k\big(f(x^k) - f(x^*)\big) \leq \tfrac{1}{2}\|x^{k+1} - x^*\|^2 - \tfrac{1}{2}\|x^k - x^*\|^2 + h_k^2 B^2$$

and *telescoping* (summing from $k = 0$ to $k = N$) gives

$$\sum_{k=0}^{N} h_k\big(f(x^k) - f(x^\star)\big) \leq \tfrac{1}{2}\left\|x^1 - x^\star\right\|^2 + \tfrac{1}{2} B^2 \sum_{k=1}^{N} h_k^2$$

hence

$$\min_{0 \leq k \leq N} f(x^k) - f(x^\star) \leq \frac{\frac{1}{2}\|x^0 - x^\star\|^2 + \frac{1}{2} B^2 \sum_{k=0}^{N} h_k^2}{\sum_{k=1}^{N} h_k}$$

Standard convergence analysis (end.)

$$\min_{0\le k\le N} f(x^k) - f(x^\star) \le \frac{\frac{1}{2}\|x^0 - x^\star\|^2 + \frac{1}{2}B^2\sum_{k=0}^N h_k^2}{\sum_{k=1}^N h_k}$$

- Right-hand side is convex in stepsizes h_k
- Optimal values are $h_k = \frac{R}{B}\frac{1}{\sqrt{N+1}}$
- Leads to

$$\min_{0\le k\le N} f(x^k) - f(x^\star) \le \frac{BR}{\sqrt{N+1}}$$

which is *optimal*

(and same rate holds for average iterate, using $f(\frac{1}{N+1}\sum_{k=0}^N x_k) - f(x_*) \le \frac{1}{N+1}\sum_{k=0}^N f(x_k) - f(x_*)$)

End of story?

What about last-iterate convergence?

$$\min_{0 \leq k \leq N} f(x^k) - f(x^\star) \leq \frac{BR}{\sqrt{N+1}}$$

- Says nothing about convergence of last iterate x_N
- O. Shamir, Open problem: Is averaging needed for strongly convex stochastic gradient descent? *JMLR* (2012)
- Practitioners often use the last iterate
- Storing best iterate might not be feasible (storage requirements, objective computation)
- Algorithm may correspond to a real-word dynamical system

Goal of this talk: study last-iterate convergence with and without performance estimation

This talk

Take-home messages:

- Performance estimation applied to subgradient methods
- Exact convergence rates can be obtained for the last iterate: suboptimal by a factor $O(\sqrt{\log(N)})$
- New last-iterate optimal method can be designed with linearly decreasing step sizes
- Extensions to constrained case, to normalized steps
- Inspiration for results provided by performance estimation but ultimately all proofs converted to classical style using a new key lemma

Contents

Last-iterate convergence

Tool: performance estimation

For a given PEP (Performance Estimation Problem) we can

- compute the exact value of the performance criteria's worst-case = optimal value of PEP problem
- identify an explicit function (and starting point) achieving this worst-case value = primal solution of PEP problem + interpolation
- obtain an independently-checkable proof that this worst-case value is a valid (upper) bound on the performance criteria = dual multiplier of PEP problem
- all three steps can be done either numerically or analytically

For a large class of first-order methods, including fixed-step subgradient methods, these can be computed *exactly* using a semidefinite programming (SDP) problem.

Interpolation conditions for nonsmooth convex functions

To perform PEP for subgradient methods on a class of functions

we need the corresponding *interpolation conditions* explicitly

given a list of values $(x_i, f_i, g_i)_{i \in I}$,
does there exist a convex f with B-bounded subgradients such that
$f(x_i) = f_i$ and $g_i \in \partial f(x_i)$ for all $i \in I = \{*, 0, 1, \dots N\}$

Necessary and sufficient conditions:

$$f(x_i) = f_i \text{ and } g_i \in \partial f(x_i) \text{ for every } i \in I$$

$$\Leftrightarrow$$

$$f_j \geq f_i + g_i^T(x_j - x_i) \text{ for every } i, j \in I$$

$$\|g_i\| \leq B \text{ for every } i \in I$$

Leads to a convex, tractable formulation as a SDP

Results: average iterate

Worst-case for fixed-step subgradient method

$$x_{i+1} = x_i - h(\tfrac{R}{B})g_i$$

applied to convex function with B-bounded subgradients

- For *average* value of iterates $\hat{f}_N = \frac{f(x_0)+f(x_1)+\ldots+f(x_N)}{N+1}$, tight worst-case is

  $$\hat{f}_N - f(x_*) \le \begin{cases} BR\left(\frac{1}{2}h + \frac{1}{2(N+1)}\frac{1}{h}\right) & \text{when } h \ge \frac{1}{N+1} \\ BR\left(1 - \frac{N}{2}h\right) & \text{when } h \le \frac{1}{N+1} \end{cases}$$

  (recovers result shown earlier for large h)

- Optimal constant step-size is then $h^* = \frac{1}{\sqrt{N+1}}$ (belongs to "large step" case) leading to tight worst-case

  $$\hat{f}_N - f(x_*) \le \frac{BR}{\sqrt{N+1}}$$

Results: last iterate

- Define sequence $\{s_N\}_{N\geq 0} = \{1, 2, \frac{5}{2}, \frac{29}{10}, \ldots\}$ with
$$s_0 = 1, s_{i+1} = s_i + \frac{1}{s_i} \text{ for all } i \geq 0$$
- No closed form, s_N^2 grows like $2(N+1) + \frac{1}{2}\log(N)$, also appears in [Nesterov 2009] for primal-dual subgradient
- For value of *last* iterate $f(x_N)$, tight worst-case is
$$f(x_N) - f(x_*) \leq \begin{cases} BR\left[(\frac{1}{2}s_N^2 - N)h + \frac{1}{2s_N^2}\frac{1}{h}\right] & \text{when } h \geq \frac{1}{s_N^2} \\ BR(1 - Nh) & \text{when } h \leq \frac{1}{s_N^2} \end{cases}$$
- No previous result with correct asymptotic rate for last iterate
- [Harvey,Liaw,Plan,Randhawa 2019] prove a $\frac{\log N}{32\sqrt{N}}$ lower bound when $B = 1$ with stepsize $h_i = \frac{1}{\sqrt{i}}$, and prove a high probability $\mathcal{O}(\frac{\log N}{\sqrt{N}})$ upper bound in stochastic case

Results: optimal stepsize and variants

- To perform N subgradient iterations, optimal stepsize is then

$$h^* = \frac{1}{\sqrt{s_N^2(s_N^2 - 2N)}}$$

and corresponding worst-case value satisfies

$$f(x_N) - f(x_*) \leq BR\sqrt{1 - \frac{2N}{s_N^2}} \lesssim BR \cdot \sqrt{\frac{1 + \frac{1}{4}\log(N)}{N+1}}$$

- Using suboptimal $h = \frac{1}{\sqrt{N+1}}$ leads to slightly worse

$$f(x_N) - f(x_*) \leq BR \cdot \left(\frac{\frac{5}{4} + \frac{1}{4}\log(N)}{\sqrt{N+1}}\right)$$

Results were obtained using the following PEP

$$
\begin{aligned}
\max\ & f^{N+1} - f^\star \\
\text{s.t. } & f^i - f^j - \left\langle \tfrac{B}{Rh}(x^j - x^{j+1}), x^i - x^j \right\rangle \geq 0 \quad i \in \{1, \dots, N+1, \star\}, j \in \{1, \dots, N\} \\
& f^i - f^{N+1} - \left\langle g^{N+1}, x^i - x^{N+1} \right\rangle \geq 0 \quad i \in \{1, \dots, N+1, \star\} \\
& f^i - f^\star \geq 0 \quad i \in \{1, \dots, N+1\} \\
& R^2 h^2 - \left\| x^k - x^{k+1} \right\|^2 \geq 0 \qquad k \in \{1, \dots, N\} \\
& B^2 - \left\| g^{N+1} \right\|^2 \geq 0 \\
& R^2 - \left\| x^1 - x^\star \right\|^2 \geq 0.
\end{aligned}
$$

PEP-based proof is ... straightforward?

Define $\sigma_i = \frac{1}{s_{i+1}}, \quad i \in \{0, 1, ..., N\}$ and observe that

$$
\begin{aligned}
& f^{N+1} - BR\left(\left(\tfrac{1}{2}s_{N+1}^2 - N\right)h + \tfrac{1}{2s_{N+1}^2 h}\right) + \sum_{i=1}^{N} \tfrac{B\sigma_{N-i}^2}{2Rh}\left(R^2h^2 - \left\|x^i - x^{i+1}\right\|^2\right) \\
& + \sum_{i=1}^{N}\sum_{j=i+1}^{N} \sigma_{N-j}\left(\sigma_{N-i} - \sigma_{N+1-i}\right)\left(f^i - f^j - \left\langle \tfrac{B}{Rh}(x^j - x^{j+1}), x^i - x^j \right\rangle\right) \\
& + \sum_{i=1}^{N}\left(\sigma_{N-i} - \sigma_{N+1-i}\right)\left(f^i - f^{N+1} - \left\langle g^{N+1}, x^i - x^{N+1}\right\rangle\right) + \tfrac{B\sigma_N^2}{2Rh}\left(R^2 - \left\|x^1\right\|^2\right) \\
& + \sigma_N \sum_{i=1}^{N} \sigma_{N-i}\left(-f^i - \left\langle \tfrac{B}{Rh}(x^i - x^{i+1}), -x^i \right\rangle\right) + \tfrac{Rh}{2B}\left(B^2 - \left\|g^{N+1}\right\|^2\right) \\
& + \sigma_N\left(-f^{N+1} + \left\langle g^{N+1}, x^i \right\rangle\right) \\
& = \tfrac{-Rh}{2B}\left\| g^{N+1} - \tfrac{B}{Rh}x^{N+1} + \tfrac{B}{Rh}\sum_{i=1}^{N}\left(\sigma_{N-i} - \sigma_{N+1-i}\right)x^i \right\|^2 \leq 0.
\end{aligned}
$$

Post-PEP reflections

- After staring at the PEP proof, we noticed similarities between inequality multipliers
- Grouping similar terms, we obtain Jensen-like inequalities
  (insight: applying Jensen $\leftrightarrow$ some sum of interpolation inequalities)
- Simplifying further we obtain a classic-style proof, that is no longer looking computer generated
- We encapsulate the main part of the proof in a *key Lemma*
- Key Lemma fully reverse-engineered from PEP but can be easily check by hand

Key Lemma for subgradient methods

Lemma ([Zamani,G 2023])

Suppose $h_{N+1} > 0$ *and introduce weights* v_k *that satisfy*

$$0 < v_0 \le v_1 \le \cdots \le v_N \le v_{N+1}$$

Then iterates of the subgradient methods satisfy

$$\sum_{k=0}^{N} \left(h_k v_k^2 - (v_k - v_{k-1}) \sum_{i=k}^{N} h_i v_i \right) \left(f(x^k) - f(\hat{x}) \right)$$
$$\le \frac{v_0^2}{2} \underbrace{\left\| x^0 - \hat{x} \right\|}_{R}^2 + \frac{1}{2} \sum_{k=1}^{N+1} h_k^2 v_k^2 \underbrace{\left\| g^k \right\|}_{B}^2$$

for any $\hat{x}$, including $\hat{x} = x_*$

Idea of the proof of the key Lemma

Generalizes the standard telescoping proof

From weights weights v_k define *auxiliary sequence z^k*

$$z^0 = \hat{x} \quad \text{and} \quad z^k = \left(1 - \frac{v_{k-1}}{v_k}\right)x^k + \left(\frac{v_{k-1}}{v_k}\right)z^{k-1}$$

for which we have

$$h_k v_k^2 \big(f(z^k) - f(x^k)\big) \leq \tfrac{1}{2} v_k^2 \|z^k - x^{k+1}\|^2 - \tfrac{1}{2} v_{k-1}^2 \|z^{k-1} - x^k\|^2 - h_k^2 v_k^2 B^2$$

which can be telescoped, and then apply Jensen on the result

Using the key Lemma

Lemma

Iterates of the subgradient methods satisfy

$$\sum_{k=0}^{N}\left(h_k v_k^2 - (v_k - v_{k-1})\sum_{i=k}^{N} h_i v_i\right)\left(f(x^k) - f(\hat{x})\right)$$
$$\leq \tfrac{v_0^2}{2}R^2 + \tfrac{1}{2}B^2\sum_{k=1}^{N+1} h_k^2 v_k^2$$

Proof of last-iterate convergence rate:

Choose weights v_k that cancel all coefficients of $f(x^k)$ except $f(x^N)$, which are

$$v_k = \frac{1}{s_{N+1-k}}$$

Exactness of rate

Follows from PEP, can be proved independently

- Explicit worst-case function can be obtained from PEP
- Defined recursively, coefficients are not straightforward
- Sugbradients for all iterates have maximum norm
- Subgradient inequality is satisfied between all pairs of iterates
- Matches exactly the announced convergence rate for the last iterate

Extensions

Last-iterate optimal subgradient method

Define the following new linearly decreasing stepsize schedule

$$x_{k+1} = x_k - \frac{R}{B}\frac{(N+1-k)}{(N+1)^{3/2}} g_k$$

Leads the optimal rate for the last iterate [Zamani,G 2023]

$$f(x_N) - f(x_*) \leq \frac{BR}{\sqrt{N+1}}$$

- Improves $\frac{15BD}{\sqrt{N+1}}$ [Jain,Nagaraj,Netrapalli 2021] for diameter D
- Same proof technique, using key lemma with other weights v_k
- *Schedule dependence on N* is forced for optimal method (already impossible to find fixed stepsizes h_1 and h_2 that are optimal for both $N = 1$ and $N = 2$)
- Existence of a last-iterate optimal method with stesizes independent from N and with momentum terms?

Subgradient method with normalized step sizes

Stepsizes so far feature a $\frac{R}{B}$ factor, require knowledge of R and B

- constant stepsizes $h_k = \frac{R}{B} h$ for some h
- optimal stepsizes $h_k = \frac{R}{B} \frac{(N+1-k)}{(N+1)^{3/2}}$

Need for B can be removed using normalized step sizes $\{t_k\}$

$$x_{k+1} = x_k - t_k \frac{g_k}{\|g_k\|} \text{ for some } g_k \in \partial f(x_k)$$

- All previous results are also valid with exactly the same rates if we assume $t_k = h_k B$
- constant stepsizes $t_k = Rh$ for some h
- optimal stepsizes $t_k = R \frac{(N+1-k)}{(N+1)^{3/2}}$
- Proof using key Lemma with adapted weights
- Removing dependence on R harder to achieve

Projected subgradient method

Solve convex constrained optimization

$$\min_{x \in X} f(x)$$

with the projected subgradient method with fixed step sizes $\{h_k\}$

$$x_{k+1} = \mathbb{P}\left[x_k - h_k g_k\right] \text{ for some } g_k \in \partial f(x_k)$$

($\mathbb{P}$ is orthogonal projection on convex set X)

- All results are also valid, with exactly the same rates (both constant and optimal stepsizes, also normalized)
- Straightforward adaptation of the key Lemma using non-expansiveness of the projection operator

Conclusions

Conclusions

Take-home messages:

- Performance estimation applied to subgradient methods
- Exact convergence rates can be obtained for the last iterate: suboptimal by a factor $O(\sqrt{\log(N)})$
- New last-iterate optimal method can be designed with linearly decreasing step sizes
- Extensions to constrained case, to normalized steps
- Inspiration for results provided by performance estimation but ultimately all proofs converted to classical style using a new key lemma