An optimal gradient method for smooth strongly convex minimization: numerical examples

Adrien Taylor · **Yoel Drori**

Date of current version: April 29, 2021

This short note contains numerical examples of the design procedure proposed in [Taylor and Drori, 2021] (based on optimizing a method's coefficients), as well as numerical comparisons with an alternate approach [Drori and Taylor, 2020] (based on mimicking conjugate-gradient type methods), and with lower bounds [Drori and Taylor, 2021].

The codes for reproducing those results can be found at

`https://github.com/AdrienTaylor/Optimal-Gradient-Method`

The examples were obtained by numerically solving the first-order method design problem in [Taylor and Drori, 2021] (for different design criterion), formulated as a linear semidefinite program using standard solvers [Löfberg, 2004, Mosek, 2010].

1 Numerical examples

1.1 Optimized method for $\|w_N - w_\star\|^2/\|w_0 - w_\star\|^2$; recovering ITEM numerically

The following list provides numerical examples obtained for $N = 1, ..., 5$ with $L = 1$ and $\mu = .1$ presented in the notations from [Taylor and Drori, 2021, Definition 2], together with the corresponding worst-case guarantees.

- For a single iteration, the step size optimization procedure produces a method with guarantee $\frac{\|w_1-w_\star\|^2}{\|w_0-w_\star\|^2} \leq 0.6694$, and a corresponding step size
$$[h^\star_{i,j}] = \begin{bmatrix}1.8182\end{bmatrix},$$
which corresponds to the step size $2/(L+\mu)$.
- For $N = 2$, the optimized method has a guarantee $\frac{\|w_2-w_\star\|^2}{\|w_0-w_\star\|^2} \leq 0.3769$, and the corresponding step sizes are
$$[h^\star_{i,j}] = \begin{bmatrix}1.5466 & \\ 0.2038 & 2.4961\end{bmatrix}.$$
- For $N = 3$, we obtain $\frac{\|w_3-w_\star\|^2}{\|w_0-w_\star\|^2} \leq 0.1932$ with
$$[h^\star_{i,j}] = \begin{bmatrix}1.5466 & & \\ 0.1142 & 1.8380 & \\ 0.0642 & 0.4712 & 2.8404\end{bmatrix}.$$

A. Taylor acknowledges support from the European Research Council (grant SEQUOIA 724063).This work was funded in part by the french government under management of Agence Nationale de la recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

Adrien Taylor
INRIA, Département d'informatique de l'ENS, École normale supérieure, CNRS, PSL Research University, Paris, France
Email: adrien.taylor@inria.fr
Yoel Drori
Google Research Israel. Email: dyoel@google.com

– For $N = 4$, we obtain $\frac{\|w_4-w_\star\|^2}{\|w_0-w_\star\|^2} \leq 0.0944$ with

$$[h^\star_{i,j}] = \begin{bmatrix} 1.5466 & & & \\ 0.1142 & 1.8380 & & \\ 0.0331 & 0.2432 & 1.9501 & \\ 0.0217 & 0.1593 & 0.6224 & 3.0093 \end{bmatrix}.$$

– Finally, for $N = 5$, we reach $\frac{\|w_5-w_\star\|^2}{\|w_0-w_\star\|^2} \leq 0.0451$ with the step sizes

$$[h^\star_{i,j}] = \begin{bmatrix} 1.5466 & & & & \\ 0.1142 & 1.8380 & & & \\ 0.0331 & 0.2432 & 1.9501 & & \\ 0.0108 & 0.0792 & 0.3093 & 1.9984 & \\ 0.0075 & 0.0554 & 0.2164 & 0.6985 & 3.0902 \end{bmatrix}.$$

It is straightforward to verify that numerical worst-case guarantees match

$$\|z_N - w_\star\|^2 \leq \|z_0 - w_\star\|^2/(1 + qA_N),$$

from [Taylor and Drori, 2021, Theorem 3]. Furthermore, one can observe an apparent strange step size pattern for going from one iteration to the next one: each time, the last line is replaced by another one, and an additional line is added. This behavior can be explained as follows: one can observe that ITEM is obtained by setting $y_k \leftarrow w_k$ $(k = 0, \ldots, N-1)$ and $z_N \leftarrow w_N$ in the [Taylor and Drori, 2021, Definition 2], as w_k's are the points where the gradients are evaluated for $k = 0, \ldots, N-1$, whereas w_N is obtained for optimizing its worst-case guarantee (but no gradient is evaluated at w_N).

1.2 Optimized methods for $(f(w_N) - f_\star)/\|w_0 - w_\star\|^2$

In this second example, we consider the criterion $(f(w_N) - f_\star)/\|w_0 - w_\star\|$. The following list provides solutions obtained by solving the corresponding design problem for $N = 1, ..., 5$ with $L = 1$ and $\mu = .1$. The solutions are presented using the notations from [Taylor and Drori, 2021] together with the corresponding worst-case guarantees.

– For a single iteration, by solving the corresponding optimization problem, we obtain a method with guarantee $\frac{f(w_1)-f_\star}{\|w_0-w_\star\|} \leq 0.1061$ and step size

$$[h^\star_{i,j}] = \begin{bmatrix} 1.4606 \end{bmatrix}.$$

This bound and the corresponding step size match the optimal step size $h_{1,0} = \frac{q+1-\sqrt{q^2-q+1}}{q}$, see [Taylor, 2017, Theorem 4.14].

– For $N = 2$ iterations, we obtain $\frac{f(w_2)-f_\star}{\|w_0-w_\star\|} \leq 0.0418$ with

$$[h^\star_{i,j}] = \begin{bmatrix} 1.5567 & \\ 0.1016 & 1.7016 \end{bmatrix}.$$

– For $N = 3$, we obtain $\frac{f(w_3)-f_\star}{\|w_0-w_\star\|} \leq 0.0189$ with

$$[h^\star_{i,j}] = \begin{bmatrix} 1.5512 & & \\ 0.1220 & 1.8708 & \\ 0.0316 & 0.2257 & 1.8019 \end{bmatrix}.$$

– For $N = 4$, we obtain $\frac{f(w_4)-f_\star}{\|w_0-w_\star\|} \leq 0.0089$, with

$$[h^\star_{i,j}] = \begin{bmatrix} 1.5487 & & & \\ 0.1178 & 1.8535 & & \\ 0.0371 & 0.2685 & 2.0018 & \\ 0.0110 & 0.0794 & 0.2963 & 1.8497 \end{bmatrix}.$$

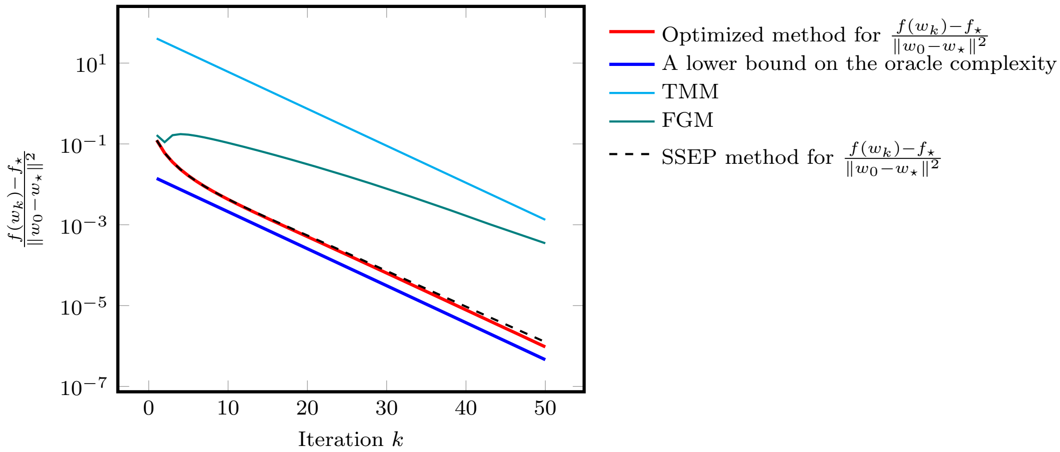

Fig. 1 Numerical comparison (for $L=1$, $\mu=0.01$) between (i) the worst-case guarantee of the optimized method for $\frac{f(w_k)-f_\star}{\|w_0-w_\star\|^2}$ (in red); (ii) a lower bound on the oracle complexity for this setup (in blue; presented in [Drori and Taylor, 2021, Corollary 3]), which corresponds to $\frac{f(w_k)-f_\star}{\|w_0-w_\star\|^2} \geq \mu\frac{2-\sqrt{q}}{1+\sqrt{q}}\left(1-\sqrt{q}\right)^{2k}$; (iii) the triple momentum method [Van Scoy et al., 2018] (cyan); (iv) Nesterov's fast gradient method (defined in [Nesterov, 2004, Section 2.2, "Constant Step Scheme, II"]; FGM, green), and (v) the method generated by the subspace-search elimination procedure (SSEP) from [Drori and Taylor, 2020] (dashed, black). All worst-case guarantees are tight in the sense that they were computed numerically using appropriate performance estimation problems.

– Finally, for $N=5$, we obtain $\frac{f(w_5)-f_\star}{\|w_0-w_\star\|} \leq 0.0042$ with

$$[h^\star_{i,j}] = \begin{bmatrix} 1.5476 & & & & \\ 0.1159 & 1.8454 & & & \\ 0.0350 & 0.2551 & 1.9748 & & \\ 0.0125 & 0.0913 & 0.3489 & 2.0625 & \\ 0.0039 & 0.0287 & 0.1095 & 0.3334 & 1.8732 \end{bmatrix}.$$

Note that when $\mu=0$, we recover the step size policy of the OGM by Kim and Fessler [2016]. When setting $\mu>0$, we observe that the resulting optimized method is apparently less practical as the step sizes critically depend on the horizon N. In particular, one can observe that $h^\star_{1,0}$ varies with the horizon N.

Figure 1 illustrates the behavior of the worst-case guarantee for larger values of N and compares it to the currently best known corresponding lower bound, as well as to worst-case guarantees for TMM, Nesterov's Fast Gradient Method (FGM) for strongly convex functions, as well as to the methods generated with the SSEP procedure from [Drori and Taylor, 2020]. All the worst-case guarantees are computed numerically using the corresponding performance estimation problems (see e.g., the toolbox [Taylor et al., 2017]).

1.3 Optimized methods for $(f(w_N)-f_\star)/(f(w_0)-f_\star)$

In this second example, we consider the criterion $(f(w_N)-f_\star)/(f(w_0)-f_\star)$. The following step sizes were obtained by setting $L=1$ and $\mu=.1$ and solving the resulting optimization problem from different values of N.

– For a single iteration, $N=1$, we obtain a guarantee $\frac{f(w_1)-f_\star}{f(w_0)-f_\star} \leq 0.6694$ with the corresponding step size

$$[h^\star_{i,j}] = \begin{bmatrix} 1.8182 \end{bmatrix},$$

which matches the known optimal step size $2/(L+\mu)$ for this setup [De Klerk et al., 2017, Theorem 4.2].

– For $N=2$, we obtain $\frac{f(w_2)-f_\star}{f(w_0)-f_\star} \leq 0.3554$ with

$$[h^\star_{i,j}] = \begin{bmatrix} 2.0095 & \\ 0.4229 & 2.0095 \end{bmatrix}.$$

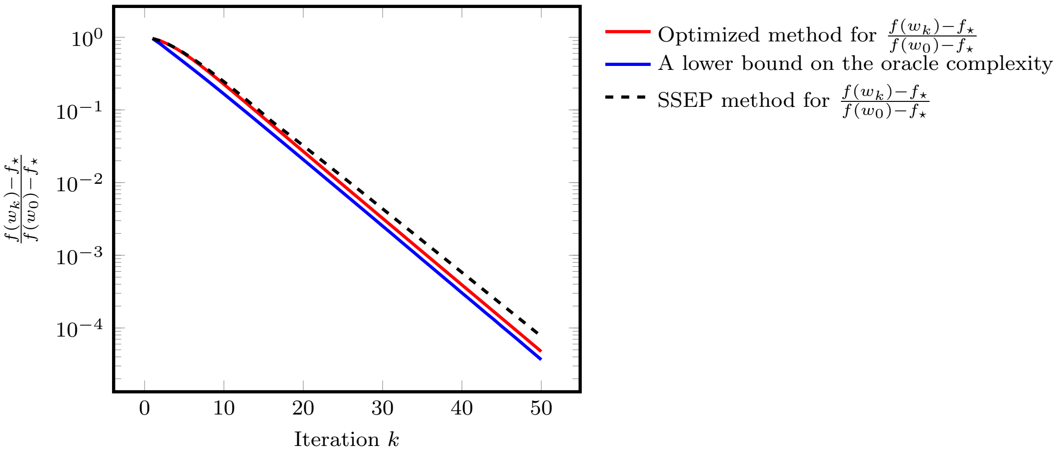

Fig. 2 Numerical comparison (for $L = 1$, $\mu = 0.01$) between (i) the worst-case guarantee of the optimized method for $\frac{f(w_k)-f_\star}{f(w_0)-f_\star}$ (in red); (ii) a lower bound on the oracle complexity for this setup (in blue; computed numerically using the procedure from [Drori and Taylor, 2021]); and (iii) a method generated by the subspace-search elimination procedure (SSEP) from [Drori and Taylor, 2020] (dashed, black). All worst-case guarantees are tight in the sense that they were computed numerically using appropriate performance estimation problems.

– For $N = 3$, we obtain $\frac{f(w_3)-f_\star}{f(w_0)-f_\star} \leq 0.1698$ with

$$[h^\star_{i,j}] = \begin{bmatrix} 1.9470 & & \\ 0.4599 & 2.2406 & \\ 0.1705 & 0.4599 & 1.9470 \end{bmatrix}.$$

– For $N = 4$, we obtain $\frac{f(w_4)-f_\star}{f(w_0)-f_\star} \leq 0.0789$ with

$$[h^\star_{i,j}] = \begin{bmatrix} 1.9187 & & & \\ 0.4098 & 2.1746 & & \\ 0.1796 & 0.5147 & 2.1746 & \\ 0.0627 & 0.1796 & 0.4098 & 1.9187 \end{bmatrix}.$$

– Finally, for $N = 5$, we reach $\frac{f(w_5)-f_\star}{f(w_0)-f_\star} \leq 0.0365$ with

$$[h^\star_{i,j}] = \begin{bmatrix} 1.9060 & & & & \\ 0.3879 & 2.1439 & & & \\ 0.1585 & 0.4673 & 2.1227 & & \\ 0.0660 & 0.1945 & 0.4673 & 2.1439 & \\ 0.0224 & 0.0660 & 0.1585 & 0.3879 & 1.9060 \end{bmatrix}.$$

Note that the resulting method is again apparently less practical than ITEM, as step sizes also critically depend on the horizon N; for example, observe again that the value of $h_{1,0}$ depends on N. Interestingly, one can observe that the corresponding step sizes are symmetric, and that the worst-case guarantees seem to behave slightly better than in the distance problem $\|w_N - w_\star\|^2/\|w_0 - w_\star\|^2$, although their asymptotic rate has to be the same, due to the properties of strongly-convex functions. Figure 2 illustrates the worst-case guarantees of the corresponding method for larger numbers of iterations, and compares it to the lower bound.

References

Etienne De Klerk, François Glineur, and Adrien B. Taylor. On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions. *Optimization Letters*, 11(7): 1185–1199, 2017.

Yoel Drori and Adrien B. Taylor. Efficient first-order methods for convex minimization: a constructive approach. *Mathematical Programming*, 184(1):183–220, 2020.

Yoel Drori and Adrien B. Taylor. On the oracle complexity of smooth strongly convex minimization. *preprint arXiv:2101.09740*, 2021.

Donghwan Kim and Jeffrey A. Fessler. Optimized first-order methods for smooth convex minimization. *Math. Program. Ser. A*, 159(1):81–107, 2016.

J. Löfberg. YALMIP : A toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD Conference*, 2004.

APS Mosek. The MOSEK optimization software. *Online at http://www.mosek.com*, 54, 2010.

Yurii Nesterov. *Introductory lectures on convex optimization: a basic course.* Applied optimization. Kluwer Academic Publishers, 2004. ISBN 9781402075537.

Adrien Taylor. *Convex Interpolation and Performance Estimation of First-order Methods for Convex Optimization*. PhD thesis, Université catholique de Louvain, 2017.

Adrien B. Taylor and Yoel Drori. An optimal gradient method for smooth strongly convex minimization. *preprint arXiv:2101.09741*, 2021.

Adrien B. Taylor, Julien M Hendrickx, and François Glineur. Performance estimation toolbox (PESTO): automated worst-case analysis of first-order optimization methods. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 1278–1283. IEEE, 2017.

Bryan Van Scoy, Randy A. Freeman, and Kevin M. Lynch. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters*, 2(1):49–54, 2018.