

Empirical and computer-aided robustness analysis of long-step and accelerated methods in smooth convex optimization

Pierre Vernimmen and François Glineur

Mathematical Engineering Department, UCLouvain, Louvain-la-Neuve, Belgium
{pierre.vernimmen, francois.glineur}@uclouvain.be

Abstract. This work assesses both empirically and theoretically, using the performance estimation methodology, how robust different first-order optimization methods are when subject to relative inexactness in their gradient computations. Relative inexactness occurs, for example, when compressing the gradient using fewer bits of information, which happens when dealing with large-scale problems on GPUs. Three major families of methods are analyzed: constant step gradient descent, long-step methods, and accelerated methods. The latter two are first shown to be theoretically not robust to inexactness. Then, a semi-heuristic shortening factor is introduced to improve their theoretical guarantees. All methods are subsequently tested on a concrete inexact problem, with two different types of relative inexactness, and it is observed that both accelerated methods are much more robust than expected, and that the shortening factor significantly helps the long-step methods. In the end, all shortened methods appear to be promising, even in this inexact setting.

Keywords: Performance Estimation, First-Order Optimization Methods, Long Steps, Accelerated Gradient, Inexact Gradient, Robustness

1 Introduction and Related Work

1.1 Inexact gradient methods

We first recall that a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is said to be *convex* if, for all $x, y \in \mathbb{R}^d$ and any $\alpha \in [0, 1]$, the following inequality holds:

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y).$$

This property ensures that the line segment between any two points on the graph of f lies above the graph itself, which implies that local minima are also global minima.

Moreover, we say that f is *L-smooth* (or equivalently, that its gradient is Lipschitz continuous) if there exists a constant $L > 0$ such that for all $x, y \in \mathbb{R}^d$,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

This condition guarantees that the gradient of f does not change too rapidly, which is essential for controlling the behavior and convergence rate of gradient-based methods.

In this work, we consider several first-order methods applied to L -smooth convex functions. Our goal is to study how convergence is affected when the gradient used at each iteration is computed inexactly. In particular, we will study the convergence rate of the methods after N iterations in (squared) gradient norm, which is the smallest quantity τ_N such that the following holds (note the use of a min to deal with non-monotone methods):

$$\frac{1}{L} \min_{k \in \{0, \dots, N\}} \|\nabla f(x_k)\|^2 \leq \tau_N(f(x_0) - f(x_*)). \quad (1)$$

Inexact gradients occur in a large variety of situations, such as the use of floating point computations with limited accuracy (see Section 1.2), dependence on data that is only known approximately, or more generally when facing the time-accuracy tradeoff that is almost always present when the objective function (and its gradient) is obtained through another iterative procedure (e.g. a simulation or another optimization process). Stochasticity can also be viewed as a particular case of inexactness.

Some first-order methods relying on an inexact gradient have been studied before theoretically, using several distinct notions of inexactness. The approximate gradient introduced in [4] is assumed to differ from the true gradient by some error whose norm is bounded. In [6], another notion of inexact gradient is developed, based on the maximal error incurred by the corresponding quadratic upper bound. In these two cases, the approximation error is not directly related to the scale of the gradient, i.e. error is measured in an *absolute* manner.

In this work, we focus on a *relative* notion of inexactness, where the norm of the difference between the true gradient and its approximate value is bounded by a certain fraction ¹ of the gradient norm. The inexact gradient at x_k is written d_k , and satisfies the following inequality:

$$\|d_k - \nabla f(x_k)\| \leq \delta \|\nabla f(x_k)\|, \text{ with relative inexactness level } \delta \in [0, 1) \quad (2)$$

This notion captures the idea that the precision of the gradient approximation should scale with the gradient norm, which is especially useful in optimization contexts where the gradient becomes small near an optimum. It was considered in [5] to analyze the effect of inexactness when dealing with smooth, strongly convex functions.

Our results show that in the smooth convex setting, even with substantial inexactness, classical first-order methods can still converge at a good rate, which is important for real-world optimization problems where exact gradients are often unavailable. We also illustrate how truncating the mantissa of the gradient components naturally leads to this type of relative inexactness (Section 1.2). More specifically, we analyze two types of gradient-type methods, namely fixed

¹ Values of the fraction $\delta \geq 1$ do not make sense, as they allow $d_k = 0$, leading to methods making no progress.

step-size gradient descent (Algorithm 1) and momentum-based accelerated gradient descent (Algorithm 2, from [10]), both using the worst-case theoretical PEP methodology (Sections 1.4 and 2) and on a smooth convex logistic classification problem (Section 3).

Algorithm 1 Inexact Gradient Descent with Relative Inexactness δ

- 1: Given an L -smooth convex function f , an initial iterate x_0 , a schedule of step sizes h_k , a number of iterations N , and an inexactness parameter $\delta \in [0, 1)$
- 2: $k \leftarrow 0$
- 3: **while** $k < N$ **do**
- 4: Compute an approximate gradient d_k at x_k with relative inexactness:

$$\|d_k - \nabla f(x_k)\| \leq \delta \|\nabla f(x_k)\|$$

- 5: Update parameter:
$$x_{k+1} = x_k - \frac{h_k}{L} d_k$$
 - 6: $k \leftarrow k + 1$
 - 7: **end while**
 - 8: **return** Iterate x_k with the smallest gradient norm
-

Algorithm 2 Inexact Fast Gradient Descent with Relative Inexactness δ

- 1: Given an L -smooth convex function f , an initial iterate x_0 , a step size h , a number of iterations N , and an inexactness parameter $\delta \in [0, 1)$
- 2: $k \leftarrow 0$
- 3: $y_0 = x_0$
- 4: **while** $k < N$ **do**
- 5: Compute an approximate gradient d_k at y_k with relative inexactness:

$$\|d_k - \nabla f(y_k)\| \leq \delta \|\nabla f(y_k)\|$$

- 6: Update parameters:
$$x_{k+1} = y_k - \frac{h}{L} d_k$$

$$y_{k+1} = x_{k+1} + \frac{k-1}{k+2} (x_{k+1} - x_k)$$
 - 7: $k \leftarrow k + 1$
 - 8: **end while**
 - 9: **return** Iterate x_k with the smallest gradient norm
-

1.2 An application of inexact gradient: compressed gradient descent

One way relative inexactness naturally arises is through compression techniques, which deliberately reduce gradient precision to save memory, such as in Compressed Gradient Descent, where each of the components of the gradient is compressed, as can be seen hereunder. Computers store numbers with limited precision. In 32-bit floating point, a number a is represented with a sign, a mantissa b , and an exponent e :

$$a = \underbrace{\pm}_{1 \text{ bit}} \times \underbrace{b}_{23 \text{ bits}} \times \underbrace{2^{e-127}}_{8 \text{ bits}}, \quad b \in [1, 2)$$

If we approximate a by \tilde{a} , using, for example, 0 bit of information instead of 23 for the mantissa (this amounts to fixing $b = 1$), the maximal possible relative error in this context of compressed gradient descent is as follows:

$$\delta_{\text{compressed}} = \frac{|a - \tilde{a}|}{|a|} = \frac{|\pm b2^{e-127} - \pm 2^{e-127}|}{|\pm b2^{e-127}|} = \frac{b-1}{b} \in \left[0; \frac{1}{2}\right] < \frac{1}{2}$$

This means we have a relative error of $\frac{1}{2}$ at most for each a . Furthermore, it can be shown that if each gradient component is δ -relatively inexact, then the whole gradient also is. In the example, when each gradient component’s mantissa is fixed to 1 (0 bit) reduces storage from $32d$ to $9d$ bits per gradient vector, where d is the number of components. If Algorithms 1 and 2 maintain strong performance compared to a version with more accurate gradients, the potential memory savings, reducing storage by more than a factor of three, can be substantial, particularly for high-dimensional optimization problems. More generally, it can be shown that keeping n_{bit} bits in each mantissa leads to an inexact gradient with $\delta_{\text{compressed}} = \left(\frac{1}{2}\right)^{n_{\text{bit}}+1}$. Such limited-precision floating-point computations are now frequently offered in modern hardware, such as in graphics processing unit (GPUs), e.g. with the FP64, FP32 and FP16 modes.

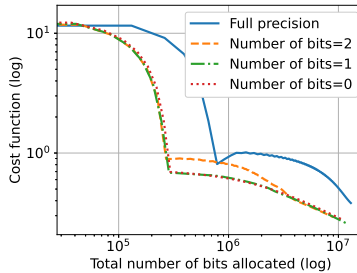


Fig. 1: Convergence of the cost function in Gradient Descent (full precision) and Compressed Gradient Descent (with 2-, 1- and 0-bit mantissa), as a function of the total number of bits allocated throughout the execution of the algorithm

A preliminary numerical experiment was conducted on a logistic regression problem of classifying hand pictures showing numbers 0 and 1 [9]. In Figure 1, we plot the cost function as a function of the total number of bits allocated in memory throughout the execution of both Gradient Descent (full precision) and Compressed Gradient Descent². It can be seen that using Compressed Gradient Descent improves the method’s performance compared to classical Gradient Descent, thus showing the usefulness of analyzing such relatively inexact methods. In experiments (Section 3.1), we show that such a compression leading to significant memory savings still maintains a competitive performance in practical problems.

² Total number of bits = [number of iterations performed] × [9 + bits used for the mantissa] × [# elements in the gradient]

1.3 Description of the considered first-order methods

In this work, we analyze several first-order methods with distinct step size schedules, all of which were originally designed for the exact case ($\delta = 0$), and are first evaluated in this context to establish a baseline. Following this, they are modified and adjusted to handle the specific requirements of relative inexactness. Evaluating them in the inexact setting allows us to measure their robustness and effectiveness under relative inexactness. Each step size is normalized by the Lipschitz constant L of the gradient [11].

- **Constant Step Size:** We use the simple choice where $h_n = 1.5 \quad \forall n$. This is optimal for one step of Algorithm 1 without inexactness ($\delta = 0$) [13]. The (squared) gradient norm of the iterates can be shown to converge in $\mathcal{O}(\frac{1}{N})$ after N iterations in the exact case (using the τ convergence measure described above).
- **Dynamic Step Size:** From Teboulle and Vaisbourd [14], this fixed, constant stepsize schedule is defined as:

$$h_0 = \sqrt{2}, \quad h_n = \frac{-H_{n-1} + \sqrt{H_{n-1}^2 + 8(H_{n-1} + 1)}}{2},$$

where $H_{n-1} := \sum_{i=0}^{n-1} h_i$, and also converges in $\mathcal{O}(\frac{1}{N})$. Its worst-case performance is close to the one achieved by an optimal choice of a constant stepsize, but without requiring knowing the number of performed iterations in advance [14]. As $N \rightarrow \infty$, step sizes tend towards 2, which is known to be the optimal constant step size in the exact case when performing $N \rightarrow \infty$ iterations [13].

- **Silver Step Size Schedule:** this long-step schedule was recently proposed by Altschuler and Parrilo [2]. For $n = 2^k - 1$, it can be defined recursively:

$$\begin{aligned} h_1 &:= \lceil \sqrt{2} \rceil \\ h_{1:2n+1} &:= [h_{1:n}, 1 + \rho^{k-1}, h_{1:n}], \end{aligned}$$

where $\rho := 1 + \sqrt{2}$ is the *silver ratio*. This schedule relies on performing very long steps from time to time, interspersed with a lot of smaller step sizes. It accelerates convergence to $\mathcal{O}(\frac{1}{N^{1.27}})$ in the exact case, but only guarantees this worst-case performance after given numbers of iterations (those satisfying $n = 2^k - 1$).

- **Fast Gradient Method (FGM):** This is the classical fast gradient method of Nesterov (Algorithm 2) with $h = 1$, which achieves the optimal $\mathcal{O}(\frac{1}{N^2})$ convergence rate in objective function values in the exact case ($\delta = 0$). It however requires keeping an additional previous iterate in memory and will serve as a baseline to determine how relative inexactness impacts accelerated methods.

1.4 Introduction to the PEP methodology

All algorithms described in Section 1.3 are first analyzed using the so-called Performance Estimation methodology [7], which allows the automated computation of exact worst-case performance guarantees for a very large class of first-order optimization algorithms. The approach relies on the observation that, by definition, the worst-case behavior of a first-order black-box optimization algorithm is by itself an optimization problem that consists of finding the worst (maximal) convergence rate (recall its definition in (1)) over all possible inputs to the algorithm. In conclusion, x_N being the output of an algorithm \mathcal{A} after making N calls to the oracle \mathcal{O}_f , we look after the worst possible function f (with iterates x_0, \dots, x_N) within a given function class \mathcal{F} , equivalent to solving the following Performance Estimation Problem (PEP) [13]:

$$\begin{aligned} \tau_N = & \max_{f \in \mathcal{F}, \{x_0, \dots, x_N, x_*\} \subset \mathbb{R}^d} \min_{k \in \{0, \dots, N\}} \|\nabla f(x_k)\|^2 \\ & \text{s.t. } x_1, \dots, x_N \text{ are generated from } x_0 \text{ by algorithm } \mathcal{A} \\ & x_* \in \arg \min_{x \in \mathbb{R}^d} f(x), f(x_0) - f(x_*) \leq 1. \end{aligned} \tag{PEP}$$

Solving the optimization problem (PEP) gives the tight worst-case performance of a first-order algorithm \mathcal{A} on the class of functions \mathcal{F} , such as the class of smooth convex functions. This worst-case performance is said to be tight, as it is exactly equal to the worst-case rate. The last constraint bounds the difference in objective function value between the initial point x_0 and the optimal solution x_* , as it is well known that in most situations, the performance of a first-order method cannot be sensibly assessed without such a constraint. The distance between x_0 and x_* can also be considered, leading to potentially different rates. Still, in this work, we focus on the objective function difference, leading to the convergence rate τ_N introduced earlier.

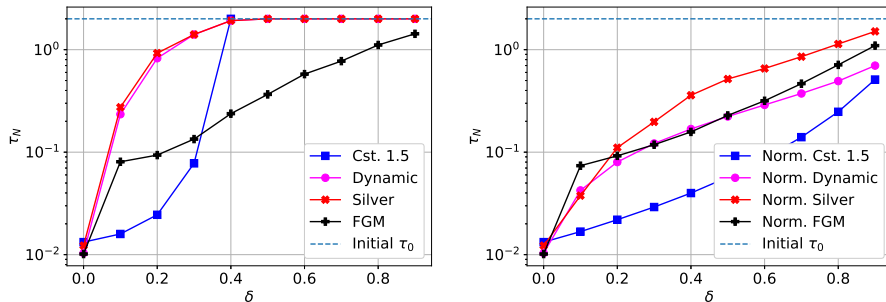
To solve problem (PEP), which contains an infinite-dimensional variable (the function f), it must first be reformulated using interpolation conditions characterizing the class of functions \mathcal{F} , as described in [13]. Inexactness of the gradient can also be dealt with in this setting by writing condition (2) as an additional constraint [5,12]. This leads to a tractable semidefinite program to compute the tight convergence rate τ_N and obtain insights about worst-case behaviors.

2 Worst-case Results

The above-described PEP methodology allows us to get the numerical worst-case performance of a given method with fixed parameters. In practice we used the PEPit toolbox [8] with the MOSEK solver [3]. Given that the schedule of step sizes presented in Section 1.3 is already fixed, fixing different values of inexactness δ allows us to get the theoretical convergence rate τ_N .

A method is said to have a *better* convergence rate than another one if its value τ_N is *smaller*. Indeed, as $f(x_0) - f(x_*)$ is fixed, the convergence rate indicates the minimum size of the best squared gradient norm among all iterates.

In Figure 2a, convergence rates of the methods presented in Section 1.3 are presented after $N = 50$ iterations for different levels of inexactness. The horizontal line τ_0 represents the value if no iteration is performed (i.e., when staying at x_0).



(a) Stepsize schedules from Section 1.3 (b) Shortened schedules (divided by $1 + \delta$)
 Fig. 2: Exact convergence rate versus inexactness level δ for $N = 50$ iterations

Convergence rates of all methods strongly depend on the inexactness level δ , and no method is particularly robust to the inexactness of the gradient. For example, for an inexactness level $\delta \geq 0.4$, most methods do worse than performing no iterations. Indeed, too long step sizes hurt the inexact methods. In the context of constant step size gradient descent, it is shown in [15] that steps with $h > \frac{2}{1+\delta}$ diverge (see Algorithm 1 on the simple univariate function $f(x) = \frac{Lx^2}{2}$). We also observed in [16] that the structure of the convergence proof of gradient descent on smooth convex functions can be preserved to tackle inexactness as long as the constant step size satisfies $h \leq \frac{2}{1+\delta}$. Combining both observations naturally leads to the idea that dividing all step sizes by the factor $1 + \delta$ can be helpful to tackle inexactness; such schedules will be referred to as *shortened* schedules. Computing the corresponding worst-case rates with PEP leads to Figure 2b, with much more encouraging rates, which suggests using those methods in a practical context. In particular, the shortened constant step size $h = \frac{1.5}{1+\delta}$ appears to be very robust against inexactness, especially for small values of inexactness level δ . The other methods on Figure 2b are less robust to inexactness, but still benefit from the shortening of the step sizes.

3 Empirical Results

After a theoretical comparison of worst-case convergence rates in the previous section, we now test³ all methods empirically in a more practical setting, both in their original and shortened versions.

The numerical experiment we present is a *logistic classification problem*, which is formulated as a smooth and convex minimization task, aligned with the theoretical framework introduced in the previous section.

We apply this logistic regression model to a binary image classification task, using a dataset of handwritten images of digits 0 and 1 [9]. (We also ran similar

³ All codes from this section and the previous one can be found on https://github.com/Verpierre/Inexact_PEP_Experiment

tests on another dataset involving phishing websites [1], and observed similar conclusions.) Each image is represented as a vector of pixel values x_k , and the associated label $y_k \in \{0, 1\}$ indicates whether the image shows a zero ($y_k = 0$) or a one ($y_k = 1$). We denote by K the total number of training images.

The goal is to learn the best classifier—defined by a weight vector w and a bias term b —to distinguish between the two digits. This is done by minimizing the logistic loss over all the training examples, leading to the following optimization problem:

$$\min_{w,b} - \sum_{k=1}^K \left[y_k \log \left(\frac{1}{1 + e^{-(b+w^T x_k)}} \right) + (1 - y_k) \log \left(1 - \frac{1}{1 + e^{-(b+w^T x_k)}} \right) \right]$$

This loss function encourages the model to assign high probabilities to the correct labels. Thanks to the properties of the logistic function, this objective is smooth and convex.

For those experiments, the four algorithms presented in Section 1.3, as well as their shortened versions, are tested with $N = 100$ iterations on the logistic classification problem, averaging the results obtained on the same 6 different initial points x_0 . Several levels of inexactness are analyzed, and two main types of relative inexactness are considered: the one from the compressed gradient descent in Section 3.1, and an adversarial corruption of the gradient (trying to achieve the worst-case behavior) in Section 3.2.

In each plot, and for each value of δ , the best value of the squared norm among all iterates is shown (lower is better), as well as the best accuracy percentage on the training set (higher is better), which may be achieved at different iterates. We also display the corresponding accuracy on the test set (i.e., using the iterate with the best training accuracy).

The smoothness constant L of the objective function, required for normalized step sizes, is estimated numerically by tracking the local curvature observed during an optimization process. The procedure begins by selecting an initial parameter x_0 and setting an initial estimate for the smoothness constant, $L_0 = 0$. We then generate a sequence of iterates x_1, \dots, x_N using an optimization method that does not require an explicit knowledge of the smoothness constant L . As we progress through the iterates, we incrementally refine our estimate of L by measuring the local curvature between successive iterates. Specifically, at each step k , the local curvature is calculated as the ratio of the norm of the difference between the gradients at two consecutive iterates to the norm of the difference between the iterates themselves: $L_{\text{local}} = \frac{\|\nabla f(x_k) - \nabla f(x_{k+1})\|}{\|x_k - x_{k+1}\|}$.

We then update our global estimate of the smoothness constant by taking the maximum of the current global estimate and the newly computed local curvature: $L_{k+1} = \max(L_k, L_{\text{local}})$.

This process continues until we have processed all the iterates, at which point the final estimate L_N represents an upper bound on the local curvature observed throughout the optimization run. In other words, L_N is the smallest value that satisfies the smoothness inequality for the observed sequence of iterates.

If a subsequent optimization method (such as an inexact gradient method) is now used with the estimated value L_N and the local curvature observed from these new iterates does not exceed this estimate, then the value of L_N is likely a good approximation of the true smoothness constant — at least for the set of iterates observed.

3.1 Empirical analysis of the compressed gradient descent

In this first experiment, we consider the compressed gradient descent (Section 1.2) using different numbers of bits n_{bit} to store the mantissa. In Section 1.2, we saw that the corresponding worst-case value of δ is equal to $(\frac{1}{2})^{n_{\text{bit}}+1}$, with n_{bit} being the number of bits used to store each mantissa. The results in this case, for different values of δ , are shown in Figure 3.

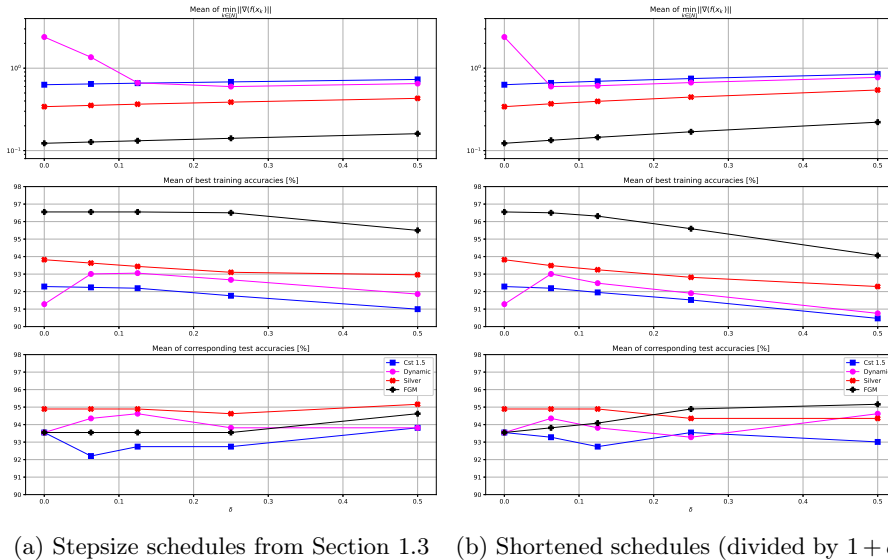


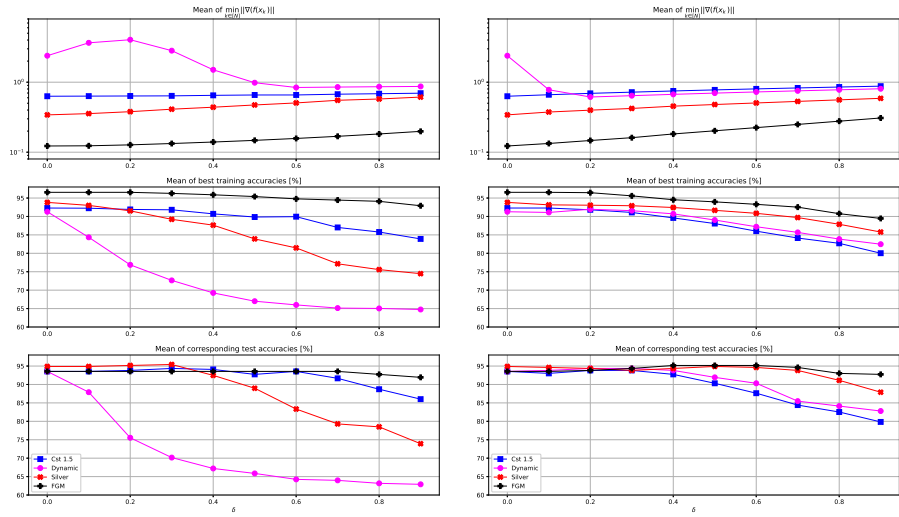
Fig. 3: Empirical results of the compressed gradient descent for different values of inexactness level $\delta_{\text{compressed}} = (\frac{1}{2})^{n_{\text{bit}}+1}$

This experiment shows that compressing the mantissa does not seem to cause any significant harm for any method. Shortened schedules also do not seem to make much difference. Furthermore, the ordering of performances in gradient norm and test accuracy between the different methods does not seem to change compared to the exact case. Indeed, FGM beats the Silver Stepsize Schedules method, which in turn beats the other two methods (matching their respective exact complexities $\mathcal{O}(\frac{1}{N^2})$, $\mathcal{O}(\frac{1}{N^{1.27}})$ and $\mathcal{O}(\frac{1}{N})$). It is however interesting to notice that test accuracies for FGM are sometimes worse than for silver step sizes. In the case of Dynamic Stepsizes, some compression seems to even help

the method as δ grows. This is probably because this compression reduces a bit the size of each component of the gradient, which can equivalently be seen, given the structure of Algorithm 1, as reducing the step size⁴. All in all, those results suggest that the error caused by compression is not harmful at all.

3.2 Empirical analysis of the gradient descent with corruption error

We now consider a more adversarial model of relative inexactness. Specifically, we add to the gradient the longest possible vector (depending on δ) pointing towards the *worst* direction. This *worst* direction is chosen as the direction taking us away from the true minimizer x_* . In some contexts, for example in cases of corruption error, it is useful to consider a worst-case analysis, as performed in this experiment. Indeed, we can imagine that, in some distributed computing situations, intruders try to make our algorithm fail, see for example [17]. The results given by the experiment for different values of δ can be seen in Figure 4 (left without the shortening factor $\frac{1}{1+\delta}$, right with it).



(a) Stepsize schedules from Section 1.3 (b) Shortened schedules (divided by $1 + \delta$)

Fig. 4: Empirical results of gradient descent with corruption (adversarial) error

In this context, we observe results that are more in line with the theoretical guarantees, as expected. Looking at non-shortened methods, Silver steps and Dynamic steps have more difficulties in handling the inexactness when it is designed to cause them problems, leading to very poor accuracies. Constant steps are also impacted, albeit less significantly. Surprisingly, FGM appears to

⁴ In this sense, it was observed that if the mantissa is truncated by rounding it upwards instead of downwards, the results deteriorated significantly.

work well with this kind of adversarial error, even with great inexactness, which is not explained by the worst-case theoretical analysis.

Adding the shortening factor greatly improves the performance of Silver and Dynamic steps, especially for the training and test accuracies (Silver steps being even able to rival FGM for $\delta \leq 0.3$). On the other hand, this shortening factor has a small negative effect on the other two methods. Overall, these results show the great performance (and robustness) in practical contexts of both accelerated methods, and we recommend the use of FGM, and of shortened Silver steps when a memoryless method is desired.

4 Conclusion

We conclude below about the worst-case and practical analysis of the different methods, both without and with the shortening factor $\frac{1}{1+\delta}$. The robustness of the methods is evaluated qualitatively in Table 1, for *small* inexactness ($\delta < 0.4$) and larger ones ($\delta \geq 0.4$).

Table 1: Qualitative evaluation of robustness based on worst-case and experiments for different δ . \times =bad, \sim =intermediate, \checkmark =good.

Method	$\delta < 0.4$	$\delta \geq 0.4$	$\delta < 0.4$	$\delta \geq 0.4$
	Worst-case		Experiment	
Cst. 1.5	\times	\times	\checkmark	\checkmark
Short. Cst. 1.5	\checkmark	\checkmark	\checkmark	\checkmark
Dynamic	\sim	\times	\sim	\times
Short. Dynamic	\sim	\sim	\checkmark	\checkmark
Silver	\sim	\times	\checkmark	\sim
Short. Silver	\sim	\sim	\checkmark	\checkmark
FGM	\sim	\sim	\checkmark	\checkmark
Short. FGM	\sim	\sim	\checkmark	\checkmark

This shows that, even in the deliberately worst practical scenarios, the methods behave better than what was expected in worst case, which may raise the question of the usefulness of worst-case analysis. However, the worst-case analysis also suggested the use of the shortening factor, which helped all non-accelerated methods in the second set of numerical experiments, still highlighting the relevance of such theoretical worst-case results.

Future work Encouraging results from this work open up some possibility of future work. The use of compression of the mantissa could be explored on more complex problems, such as small neural networks with sigmoid activation functions, generalizing the concept of logistic classification. Furthermore, a stochastic analysis of compressed gradient descent, instead of only considering the worst-case compression error, would also be interesting. Indeed, on average, errors caused by compression are much smaller than the worst-possible ones.

Acknowledgments The first author is supported by the UCLouvain *Fonds spéciaux de Recherche* (FSR). He would like to thank Nicolas Mil-Homens Cavaco for the fruitful discussions and Hugo Siberdt for the initial code in the exact case on the second dataset.

References

1. Neda Abdelhamid. Website Phishing. UCI Machine Learning Repository, 2014. DOI: <https://doi.org/10.24432/C5B301>.
2. Jason M Altschuler and Pablo A Parrilo. Acceleration by stepsize hedging: Silver stepsize schedule for smooth convex optimization. *Mathematical Programming*, 2024.
3. MOSEK ApS. *The MOSEK Python Fusion API manual. Version 11.0.*, 2025.
4. Alexandre d’Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, 2008.
5. Etienne De Klerk, François Glineur, and Adrien B Taylor. Worst-case convergence analysis of inexact gradient and newton methods through semidefinite programming performance estimation. *SIAM Journal on Optimization*, 30(3):2053–2082, 2020.
6. Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146:37–75, 2014.
7. Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1-2):451–482, 2014.
8. Baptiste Goujaud, Céline Moucer, François Glineur, Julien M Hendrickx, Adrien B Taylor, and Aymeric Dieuleveut. Pepit: computer-assisted worst-case analyses of first-order optimization methods in python. *Mathematical Programming Computation*, 16(3):337–367, 2024.
9. Arda Mavi. Sign language digits dataset, 2017.
10. Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$. *Doklady AN USSR*, 269:543–547, 1983.
11. Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.
12. Adrien B Taylor, Julien M Hendrickx, and François Glineur. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization*, 27(3):1283–1313, 2017.
13. Adrien B Taylor, Julien M Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161:307–345, 2017.
14. Marc Teboulle and Yakov Vaisbourd. An elementary approach to tight worst case complexity analysis of gradient based methods. *Mathematical Programming*, 201(1):63–96, 2023.
15. Pierre Vernimmen. Tight convergence analysis of exact and inexact gradient methods with constant and silver schedules. Master’s thesis, UCLouvain, 2024.
16. Pierre Vernimmen and François Glineur. Worst-case convergence analysis of relatively inexact gradient descent on smooth convex functions. *arXiv:2406.17506*, 2025.
17. Shuche Wang and Vincent YF Tan. Robust distributed gradient descent to corruption over noisy channels. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 2520–2525. IEEE, 2024.