



Gradient descent in phase retrieval

Massimo Cicognani¹ · Alberto Bernardini² · Alessandro Ilic Mezza² · Roberto Leone Cicognani³

Received: 31 July 2026 / Accepted: 1 September 2026
© The Author(s) 2026

Abstract

Alternating projections, like Griffin-Lim and Fast Griffin-Lim algorithms, are among the most widely used iterative methods for phase retrieval. In this paper, we build upon the appendix of the original paper by Griffin and Lim and formulate them as gradient descent algorithms. By this approach, we can give convergence guarantees, evaluate the rate of convergence, and introduce a new phase recovery method based on the Nesterov acceleration of the gradient descent.

Keywords Phase retrieval · Gradient descent · Griffin-Lim algorithm · Signal processing · Time-frequency analysis · Nesterov's optimization

Mathematics Subject Classification 94A12 · 42B10 · 90C26 · 65K05

1 Introduction

The problem of phase retrieval consists in recovering an unknown signal from the magnitude of its Fourier transform [1]. This is an important task in many engineering fields, including optical imaging [2], crystallography [3], and astronomy [4]. In audio

Alberto Bernardini, Alessandro Ilic Mezza, and Roberto Leone Cicognani have contributed equally to this work.

✉ Massimo Cicognani
massimo.cicognani@unibo.it

Alberto Bernardini
alberto.bernardini@polimi.it

Alessandro Ilic Mezza
alessandroilic.mezza@polimi.it

Roberto Leone Cicognani
robertoleonecicognani@gmail.com

¹ Dipartimento di Matematica, Università di Bologna, Bologna, Italy

² Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy

³ Korg Italy, Osimo, Italy

signal processing, estimating a time-domain signal from the (modified) magnitude of its time-frequency representation is key to many applications [5–13]. Several audio processing techniques such as time-scale modification and spectral subtraction, in fact, entail magnitude modifications which may result in *inconsistent* Short-Time Fourier Transform (STFT) spectrograms [14, 15]. Retrieving signals from synthetic magnitude STFTs, i.e., with phase information not being available in the first place, also plays a central role in many neural vocoders and text-to-speech models [16–19].

Introduced in 1984, the Griffin-Lim Algorithm (GLA) [20] is still one of the leading methods for signal estimation and phase retrieval [15, 21–24]. This celebrated alternating projections method, differently from the closely related Gerchberg-Saxton algorithm [25], which retrieves the phase of complex-valued signals in terms of the non-redundant Fourier transform, exploits the redundancy of STFT to estimate a real-valued sequence such that its magnitude is closest to a target magnitude spectrogram in a least squares sense. To accomplish this, GLA starts from an arbitrary initial phase and iteratively projects the resulting complex-valued matrix back and forth between time and time-frequency domains, while ensuring that the STFT magnitude matches the target at each step.

In 2013, the Fast Griffin-Lim Algorithm (FGLA) [26] was empirically shown to improve upon GLA, accelerating the original algorithm by means of an inertial term with constant momentum. In the past decade, FGLA has established itself as one of the most widely adopted iterative phase retrieval methods for audio signal processing applications. Nevertheless, theoretical convergence remains an open problem.

In 2023, Nenov et al. proposed the Accelerated Griffin-Lim Algorithm (AGLA) [24], which combines doubly-projected and non-projected STFTs. Furthermore, the authors provide convergence guarantees for FGLA and AGLA for restricted ranges of momenta, which, however, do not include the values that empirically proved to work best.

In this manuscript, we build upon the appendix of Griffin and Lim's 1984 paper [20]. From there, it follows directly that GLA reduces to a gradient descent algorithm under mild conditions on the analysis window. This result has been known since GLA's inception, at the same time, to the best of our knowledge, this interpretation has received limited attention in the literature on Griffin-Lim methods ever since. We refer to [22] for a second order algorithm based on the gradient of a generalization of the nonconvex objective function in GLA.

Here, we exploit the fact that the objective function of GLA is locally L -smooth, that is with Lipschitz continuous gradient, in convex sets. Among them, there are unbounded neighborhoods of any point with target magnitude. We then apply to audio phase recovery the theory of the standard and accelerated gradient descent which is well-established also for nonconvex objective functions [27–31]. This approach allows us to prove the convergence of a gradient descent phase recovery algorithm which contains GLA as a special case for an appropriate choice of the step-size. Namely, we prove that there exists a subsequence of the iterates convergent to a critical point of the objective function and that the whole value sequence converges monotonically decreasing [20]. Furthermore, we can evaluate as $\mathcal{O}(1/i)$ its convergence rate in neighborhoods of convexity [32], i being the iteration index. Experimental results consistently fit with this asymptotic behavior in most cases.

In this framework, we can look at FGLA as an accelerated gradient descent with constant momentum [32]. Under a local L -smoothness assumption, in [33] we have provided convergence guarantees for a broader range of momenta than the one discussed in [24]. Here, we present a slightly more general result, which applies also to some particular cases of AGLA.

In addition, our proof shows the convergence of the whole FGLA value sequence, which is nontrivial due to the lack of monotonicity in inertial methods [32, 34]. The critical limit value is most likely a local minimum because accelerated gradient descent algorithms escape saddle points faster [31, 35–37].

The achievement of a provably superior rate of convergence in neighborhoods of convexity is to be found in the Nesterov mathematical theory. In his impactful 1983 paper [34], Nesterov proved that the convergence rate of the convex gradient descent can be improved from $\mathcal{O}(1/i)$ to $\mathcal{O}(1/i^2)$ using a sequence of variable momenta depending on the iteration index i [28, 32, 38].

Nesterov acceleration gained renewed attention in the deep learning community when, in 2013, Sutskever and colleagues [39] popularized its use for stochastic gradient descent under the name Nesterov Accelerated Gradient (NAG).

Examples of proper Nesterov accelerated gradient methods can be found in earlier phase retrieval literature [40, 41]. Nesterov acceleration, though, was applied to the convex program that results from the semidefinite relaxation of the problem of estimating the rank-one outer product of the target signal with itself. While such a “lifting” operation leads to a tractable convex optimization program, it comes at the cost of dramatically increasing the dimensionality of the problem, thereby severely limiting its applicability.

Following the Nesterov theory, we propose an accelerated method, called Nesterov Accelerated Griffin-Lim Algorithm (NAGLA), that has $\mathcal{O}(1/i^2)$ rate of convergence in neighborhoods of convexity. Our experimental results are in line with this asymptotic behavior in most cases. Furthermore, NAGLA achieves a moderate but consistent improvement over the other fast algorithms in the short-time regime. The remainder of this manuscript is organized as follows. In Sect. 2, we focus on GLA and the standard gradient descent in phase retrieval. Our main results about the convergence of FGLA as an accelerated gradient descent algorithm are presented in Sect. 3. In Sect. 4, we propose NAGLA basing on Nesterov’s theory. Section 5 is devoted to numerical evaluation. Finally, Sect. 6 concludes this work.

2 Griffin-Lim algorithm

In this section, we consider the Griffin-Lim Algorithm [20] for audio signals phase recovery. We first recall the mathematical results about the uniqueness of the solution in phase retrieval, then introduce GLA in its original alternating projections form in the time-frequency domain. Analyzing the appendix of [20], we exploit the equivalent gradient descent formulation of GLA in the time domain to give an alternative proof of convergence, which extends to a generalized version of the algorithm. Our approach also provides estimates of the rate of convergence.

2.1 Uniqueness of the solution

The problem of phase retrieval does not have a unique solution. Let

$$|X| = (|X[0]|, |X[1]|, \dots, |X[N-1]|)$$

be the magnitude of the nonredundant discrete time Fourier transform

$$X[k] = \sum_n x[n] e^{-i2\pi kn/N},$$

of the time signal

$$x = (x[0], x[1], \dots, x[N-1]).$$

For any given x , the following y share the same magnitude, i.e., $|Y| = |X|$:

- Constant phase shift $y = cx$ with a unitary scalar factor $|c| = 1$;
- Time shift $y = (x[s], \dots, x[N-1], x[0], \dots, x[s-1])$;
- Mirror $y = (x^*[N-1], \dots, x^*[1], x^*[0])$.

From a physical viewpoint, these are acceptable ambiguities but in general infinitely many solutions can be obtained beyond these trivial associates of x [42].

Uniqueness up to a constant phase holds thanks to the redundancy of the STFT

$$X[m, k] = \sum_n x[n] w[Sm - n] e^{-i2\pi kn/K}, \quad (1)$$

where $w[n]$ is an analysis window of length K with stride S . From the results in [43], we have equal STFT magnitudes

$$|X| = |Y| \iff y = cx, \quad |c| = 1,$$

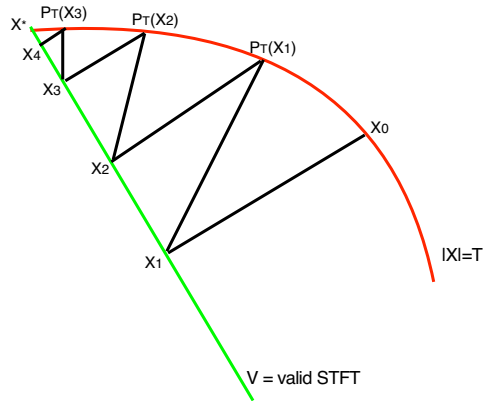
for a sufficiently large overlap ratio $(K - S)/K$.

2.2 Griffin-Lim double projection

Phase retrieval algorithms were introduced well before the mathematical proof in [43] of the uniqueness of the solution up to a constant phase. The Gerchberg-Saxton algorithm [25] for image reconstruction inspired the Griffin-Lim algorithm [20] for the recovery of an audio signal x_* from the magnitude $T = |X_*|$ of its STFT. For a noisy or modified T , GLA seeks for the closest estimate x_* in a least squares sense. As the Gerchberg-Saxton, GLA is an alternating projections algorithm. The first projection

$$P_T(X) = T \frac{X}{|X|} \quad (2)$$

Fig. 1 GLA alternating projections



is on the set $\mathcal{T} = \{X : |X| = T\}$. If $X[m, k] = 0$ the corresponding entry in $X/|X|$ is set to 1 (null phase). $P_{\mathcal{T}}$ is continuous in the complementary of the singular set

$$\mathcal{S} = \{x; \exists [m, k] : X[m, k] = 0, T[m, k] \neq 0\}. \quad (3)$$

The second projection

$$P_{\mathcal{V}}(X) = \text{STFT}(\text{iSTFT}(X)) \quad (4)$$

is the orthogonal projection on the linear subspace $\mathcal{V}$ spanned by the STFT. Here, iSTFT denotes the so-called inverse short-time Fourier transform, i.e., the pseudoinverse of the full column rank linear transformation in (1).

The iSTFT is given by

$$x[n] = \frac{\sum_m w[Sm - n] \sum_k X[m, k] e^{i2\pi kn/K}}{K \sum_m w^2[Sm - n]}, \quad (5)$$

using the (normalized) synthesis window $\tilde{w}[n] = w[n]/\sum_m w^2[Sm - n]$. In order to simplify $\tilde{w}[n]$, the analysis window $w[n]$ can be chosen to satisfy the mild Constant Overlap-Add (COLA) constraint

$$\sum_m w^2[Sm - n] = C, \quad \forall n \in \mathbb{Z}, \quad C \in \mathbb{R}_{>0}, \quad (6)$$

as for the experiments in the original GLA paper [20].

The double projection GLA is defined by

$$X_{i+1} = P_{\mathcal{V}}(P_{\mathcal{T}}(X_i)) \quad (7)$$

with initial point X_0 such that $|X_0| = T$, Fig. 1.

If any phase information is not available, the standard choices for the initial step are the null or random phase [20, 23, 24, 26]. In the time domain, GLA takes the form

$$x_{i+1} = \text{iSTFT}((P_{\mathcal{T}}(X_i))). \quad (8)$$

2.3 The original proof of convergence for GLA

The convergence of GLA is proved in the appendix of [20] using the global convergence theorem [38] for a descending algorithm $x_{i+1} = A(x_i)$:

Theorem 1 (global convergence) For $x_{i+1} = A(x_i)$, assume that

- (i) x_i is a bounded sequence;
- (ii) $A(x)$ is a closed map;
- (iii) there exist a continuous real-valued function $d(x)$ and a set Γ such that $d(x_{i+1}) < d(x_i)$ if $x_i \notin \Gamma$, and $d(x_{i+1}) \leq d(x_i)$ if $x_i \in \Gamma$.

Then, there exists a subsequence of x_i converging to a point $x_* \in \Gamma$.

In the following, the set Γ is referred to as the *solution set*.

For GLA, the map $A(x)$ is given by

$$A(x) = \text{iSTFT}((P_{\mathcal{T}}(X))) \quad (9)$$

as stated in (8).

In [20], the function $d(x)$ is the spectral distance measure

$$d(x) = \| |X| - T \|_F^2, \quad (10)$$

where $\| \cdot \|_F$ denotes the Frobenius norm in the Hilbert space of complex matrices, and the solution set is proved to be the set of its critical points, i.e.,

$$\Gamma = \{x : \nabla d(x) = 0\}. \quad (11)$$

Therefore, applying Theorem 1, in the appendix of [20] it is proved that there exists a subsequence of x_i convergent to a critical point x_* of $d(x)$. Being $d(x)$ non-globally convex, x_* may turn out to be a local minimum or even a saddle point.

The hypothesis (i) of Theorem 1 is satisfied because the boundedness of the GLA sequence x_i in (8) follows from $\|P_{\mathcal{T}}(X)\|_F = \|T\|_F$ for all X .

Concerning hypothesis (ii), in [20] the authors claim that $A(x)$ in (9) is closed because it is a continuous map. In doing so, they implicitly assume that $x_i \notin \mathcal{S}$, $\mathcal{S}$ the singular set (3) of the projection $P_{\mathcal{T}}$.

In order to prove the validity of hypothesis (iii) and that the solution set Γ is given by (11), the authors deduce

$$d(x_{i+1}) < d(x_i) \iff x_{i+1} \neq x_i \quad (12)$$

from the definition of the two projections $P_{\mathcal{T}}$, $P_{\mathcal{V}}$. Then, they observe that

$$\nabla d(x_i) = L(x_i - x_{i+1}), \quad (13)$$

with

$$L = 2K \sum_m w^2 [Sm - n], \quad (14)$$

from (8) and the straightforward computation of the gradient

$$\frac{1}{L} \nabla d(x) = x - \text{iSTFT}((P_{\mathcal{T}}(X))), \quad x \notin \mathcal{S}. \quad (15)$$

Equations (12) and (13) imply that hypothesis (iii) in Theorem 1 is satisfied with Γ the set of critical points of $d(x)$ in (11).

2.4 Griffin-Lim gradient descent

The key point in the original proof of the convergence of GLA is the evaluation (13), formula (A11) in [20], of the gradient of the function $d(x)$ in (10) at the iterates x_i . Solving for x_{i+1} in equality (13), we can recast the GLA update (8) in the time domain as follows:

$$x_{i+1} = x_i - \frac{1}{L} \nabla d(x_i), \quad x_i \notin \mathcal{S}. \quad (16)$$

With a COLA window, the constant L does not depend on the time index n . Under this mild condition, the above update (16) shows that the GLA double projection (7)–(8) is equivalent to a gradient descent algorithm.

From now on, we take a window function that satisfies the COLA constraint, and we assume $x_i \notin \mathcal{S}$ [20]. In general, a gradient descent phase recovery algorithm does not require the analysis window to satisfy (6), see, e.g., [22]. However, the COLA assumption is needed to establish the equivalence of (16) with the double projection formulation given in (7) and (8).

Convergence guarantees of gradient descent methods hold under the assumption that the objective function is L -smooth. The function $d(x)$ in (10) does not have this global regularity but there exist convex sets $\mathcal{R}$, which we call regions of L -smoothness, where for any fixed $x, y \in \mathcal{R}$, and also for all interior points z of the line segment with endpoints x, y , we have

$$\frac{1}{L} \|\nabla d(x) - \nabla d(y)\| \leq \|x - y\|, \quad (17)$$

and

$$\frac{1}{L} \|\nabla d(z) - \nabla d(y)\| \leq \|z - y\|, \quad z = y + \lambda(x - y), \quad \lambda \in (0, 1). \quad (18)$$

Examples of said regions are presented in subsection 2.7.

2.5 Griffin-Lim convergence by L -smoothness

The theory of the L -smooth gradient descent applies to the algorithm

$$x_{i+1} = x_i - \frac{\gamma}{L} \nabla d(x_i) \quad (19)$$

with step-size γ/L , $0 < \gamma < 2$ [32]. In view of the representation (15) of the gradient, in the time-frequency domain this results in the linear combination

$$X_{i+1} = (1 - \gamma)X_i + \gamma P_{\mathcal{V}}(P_{\mathcal{T}}(X_i)). \quad (20)$$

With $\gamma = 1$, we have the GLA double projection, see (7).

Under assumptions of local L -smoothness, we can prove the convergence of the algorithm (19) in a similar way as under global Lipschitz regularity [29, 32].

We recall key inequalities at fixed points x, y that follow from (17) and (18) [32].

Lemma 1 *Let x, y satisfy (17) and (18). Then, we have*

$$d(x) \leq d(y) + \langle \nabla d(y), x - y \rangle + \frac{L}{2} \|x - y\|^2 \quad (21)$$

and

$$d(y) \leq d(x) - \langle \nabla d(y), x - y \rangle + \frac{L}{2} \|x - y\|^2. \quad (22)$$

Proof Both inequalities (21) and (22) follow from (17) and (18) via the interpolation integral formula

$$d(x) - d(y) = \int_0^1 \langle \nabla d(y + \lambda(x - y)), x - y \rangle d\lambda$$

given by the fundamental theorem of calculus. $\square$

In the proof of the following theorem, only (21) is needed. We have stated also (22) for later use in treating inertial methods.

Theorem 2 Assume that there exists $i_0 \in \mathbb{N}$ such that the iterates x_i in (19) belong to a region of L -smoothness for $i \geq i_0$. Then, for all $\gamma \in (0, 2)$:

- (i) The sequence $d(x_i)$ is decreasing for $i \geq i_0$, hence there exists $\lim_{i \rightarrow \infty} d(x_i) = d_*$.
- (ii) $\nabla d(x_i) \rightarrow 0$ with

$$\min_{0 \leq \ell \leq i-1} \|\nabla d(x_\ell)\|^2 \leq C/i, \quad i > i_0, \quad (23)$$

C a positive constant depending on L, γ, i_0 .

- (iii) x_i is a bounded sequence. Any limit point x_* of a convergent subsequence is a critical point for $d(x)$, hence a point such that

$$X_* = P_{\mathcal{V}}(P_{\mathcal{T}}(X_*)). \quad (24)$$

Proof With $(x, y) = (x_{i+1}, x_i)$, x_i, x_{i+1} being the iterates in (19), inequality (21) gives

$$d(x_{i+1}) \leq d(x_i) - \frac{\gamma(2 - \gamma)}{2L} \|\nabla d(x_i)\|^2, \quad i \geq i_0. \quad (25)$$

This proves the descent $d(x_{i+1}) \leq d(x_i)$ for $i \geq i_0$ with equality if and only if $\nabla d(x_i) = 0$ and $x_{i+1} = x_i$. The value sequence $d(x_i)$ converges to the lower bound $d_* = \inf\{d(x_i) : i \geq i_0\}$.

Telescoping (25) we have

$$\frac{\gamma(2-\gamma)}{2L} \sum_{i=i_0}^{\infty} \|\nabla d(x_i)\|^2 \leq d(x_{i_0}) - d_*. \quad (26)$$

The convergence of the series implies $\nabla d(x_i) \rightarrow 0$. Then, from

$$(i - i_0) \cdot \min_{i_0 \leq \ell \leq i-1} \|\nabla d(x_\ell)\|^2 \leq \sum_{\ell=i_0}^i \|\nabla d(x_\ell)\|^2,$$

inequality (26), and

$$\min_{0 \leq \ell \leq i-1} \|\nabla d(x_\ell)\|^2 \leq \min_{i_0 \leq \ell \leq i-1} \|\nabla d(x_\ell)\|^2,$$

we get the rate estimate (23) with constant

$$C = \frac{2L(i_0 + 1)(d(x_{i_0}) - d_*)}{\gamma(2 - \gamma)}.$$

Finally, from the representation of the gradient (15) and $\nabla d(x_i) \rightarrow 0$, we have

$$x_i - \text{iSTFT}(P_{\mathcal{T}}(X_i)) \rightarrow 0$$

that shows that x_i is a bounded sequence as $P_{\mathcal{T}}(X_i)$. Any limit point x_* of a convergent subsequence x_{ℓ_i} is such that $d(x_*) = d_*$ and $\nabla d(x_*) = 0$ by uniqueness of the limits $\lim_{i \rightarrow \infty} d(x_{\ell_i})$ and $\lim_{i \rightarrow \infty} \nabla d(x_{\ell_i})$. X_* satisfies (24) since from (15) any critical point of $d(x)$ is a fixed point of the double projection. $\square$

Remark 1 The conclusions of Theorem 2 hold true under the more general assumption that inequality (21) is satisfied by $(x, y) = (x_{i+1}, x_i)$ for $i \geq i_0$.

2.6 GLA rate of convergence

For $\gamma = 1$, Theorem 2 provides an alternative proof of the convergence of GLA. On the one hand, it is a weaker proof because of the L -smooth assumptions for $i \geq i_0$. On the other hand, it works for the generalized algorithm defined by (19) and (20). Furthermore, inequality (23) provides a first information about the rate of convergence: it implies that GLA finds a point x such that

$$\|\nabla d(x)\| \leq \varepsilon, \quad (27)$$

an ε critical point, in at most $\mathcal{O}(\varepsilon^{-2})$ steps. Note that (27) is equivalent to

$$\|X - P_{\mathcal{V}}(P_{\mathcal{T}}(X))\| \leq \varepsilon/\sqrt{2L} \quad (28)$$

by (15) and Parseval identity.

The gradient descent theory originates in convex optimization. If for $i \geq i_c \geq i_0$ the iterates enter a region of local convexity for $d(x)$, then we can prove the same rate of convergence of the value sequence

$$d(x_i) - d(x_*) = \mathcal{O}(1/i)$$

to the local minimum $d(x_*)$, as in the case of a globally convex function to the optimal value [29, 32]. This is done in the following corollary to Theorem 2.

Corollary 1 *If the GLA (16) iterates x_i , $i \geq i_c \geq i_0$, and the limit point x_* of a convergent subsequence belong to a region of local convexity for $d(x)$, then x_* is a local minimizer, $d(x_i)$ converges to the local minimum value $d(x_*)$ with rate given by*

$$d(x_i) - d(x_*) \leq C_0/i, \quad (29)$$

$i > i_c$, C_0 a positive constant, and $\nabla d(x_i) \rightarrow 0$ with the rate

$$\min_{0 \leq \ell \leq i-1} \|\nabla d(x_\ell)\|^2 \leq C_1/i^2, \quad (30)$$

$C_1 > 0$. Moreover, the whole sequence x_i converges to x_* .

Proof Any critical point in a region of local convexity is a local minimizer.

For the GLA sequence (16), the inequality (21) gives

$$d(x_{i+1}) \leq d(x_i) - \frac{L}{2} \|x_{i+1} - x_i\|^2, \quad i \geq i_0.$$

Subtracting $d(x_*)$ in both sides we obtain

$$d(x_{i+1}) - d(x_*) \leq d(x_i) - d(x_*) - \frac{L}{2} \|x_{i+1} - x_i\|^2. \quad (31)$$

By convexity, for $i \geq i_c \geq i_0$, we have

$$d(x_i) - d(x_*) \leq \langle \nabla d(x_i), x_i - x_* \rangle = L \langle x_{i+1} - x_i, x_* - x_i \rangle. \quad (32)$$

Substituting (32) in (31) we get

$$d(x_{i+1}) - d(x_*) \leq \frac{L}{2} \left[2 \langle x_{i+1} - x_i, x_* - x_i \rangle - \|x_{i+1} - x_i\|^2 \right]. \quad (33)$$

Then, by $2\langle a, b \rangle - \|a\|^2 = \|b\|^2 - \|a - b\|^2$ with $a = x_{i+1} - x_i$ and $b = x_* - x_i$, for $i \geq i_c \geq i_0$, we obtain

$$0 \leq d(x_{i+1}) - d(x_*) \leq \frac{L}{2} \left(\|x_i - x_*\|^2 - \|x_{i+1} - x_*\|^2 \right). \quad (34)$$

In particular, $\|x_{i+1} - x_*\| \leq \|x_i - x_*\| \quad \forall i \geq i_c$, which yields the localization

$$\|x_{i_\varepsilon} - x_*\| < \varepsilon \implies \|x_i - x_*\| < \varepsilon \quad \forall i \geq i_\varepsilon \geq i_c.$$

The convergence to x_* of a subsequence ensures that for any $\varepsilon > 0$ there exists such an x_{i_ε} , hence the whole sequence x_i converges to x_* . Furthermore, from (34) we have also

$$d(x_\ell) - d(x_*) \leq \frac{L}{2} \left(\|x_i - x_*\|^2 - \|x_{i+1} - x_*\|^2 \right), \quad \ell \geq i + 1, \quad (35)$$

since $d(x_\ell) \leq d(x_{i+1})$. Summing up for $i = i_c, \dots, \ell - 1$ and telescoping the above inequalities (35), we obtain

$$(\ell - i_c)(d(x_\ell) - d(x_*)) \leq \frac{L}{2} \|x_{i_c} - x_*\|^2 \quad (36)$$

hence the estimate (29) with $C_0 = \frac{L(i_c+1)}{2} \|x_{i_c} - x_*\|^2$ using again i as index instead of ℓ .

Finally, telescoping (25) we have also

$$\frac{\gamma(2-\gamma)}{2L} \sum_{\ell=i}^{2i-1} \|\nabla d(x_\ell)\|^2 \leq d(x_i) - d(x_{2i-1}) \leq d(x_i) - d(x_*), \quad (37)$$

$i > i_0$, which gives (30) thanks to

$$i \cdot \min_{i \leq \ell \leq 2i-1} \|\nabla d(x_\ell)\|^2 \leq \sum_{\ell=i+1}^{2i} \|\nabla d(x_\ell)\|^2$$

and (29). □

2.7 On the existence of regions of L-smoothness

Theorem 2 assumes that, from some iteration i_0 onward, the iterates x_i in (19) belong to a region of L -smoothness, that is, a region in which (17) and (18) hold. Here, we discuss the applicability of our hypotheses by providing nontrivial examples of such regions.

First, we recast the inequality (17) in terms of phase variation and given magnitude:

Proposition 1 Let L be the constant in (14), and let $x, y \notin \mathcal{S}$. The condition

$$\sum_{m,k} T[m, k](1 - \psi[m, k])(|X[m, k]| + |Y[m, k]| - T[m, k]) \geq 0 \quad (38)$$

with

$$\psi[m, k] = \cos(\angle X[m, k] - \angle Y[m, k]) \quad (39)$$

is sufficient for inequality (17).

Proof The COLA condition (6) ensures the Parseval identity $\|X\|^2 = \frac{L}{2}\|x\|^2$ for any signal x whose STFT is X . Substituting (15) into (17), applying the STFT, and squaring both sides thus yields

$$\|(X - Y) - (P_{\mathcal{V}}(P_{\mathcal{T}}(X)) - P_{\mathcal{V}}(P_{\mathcal{T}}(Y)))\|^2 \leq \|X - Y\|^2. \quad (40)$$

Since (4) is an orthogonal projection, (40) is implied by

$$\|(X - Y) - (P_{\mathcal{T}}(X) - P_{\mathcal{T}}(Y))\|^2 \leq \|X - Y\|^2. \quad (41)$$

By taking $W = (X - Y)$, $Q = (P_{\mathcal{T}}(X) - P_{\mathcal{T}}(Y))$, and $\|W - Q\|^2 = \|W\|^2 - 2\Re\langle W, Q \rangle + \|Q\|^2$, with $\langle W, Q \rangle$ the Frobenius inner product, we can write (41) as

$$2\Re\langle W, Q \rangle - \|Q\|^2 \geq 0. \quad (42)$$

Substituting (2) and expanding in polar coordinates, after some straightforward algebraic manipulations, we obtain

$$\begin{aligned} \Re\langle W, Q \rangle &= \sum_{m,k} T[m, k](|X[m, k]| + |Y[m, k]|)(1 - \psi[m, k]), \\ \|Q\|^2 &= \sum_{m,k} 2T[m, k]^2(1 - \psi[m, k]). \end{aligned}$$

Factoring out $2T[m, k](1 - \psi[m, k])$ in (42) yields (38). It follows that (41) holds as long as (38) is satisfied. Hence, (38) is a sufficient condition for (17). $\square$

Condition (38) requires the weighted sum of $|X[m, k]| + |Y[m, k]| - T[m, k]$ to be nonnegative, $T[m, k](1 - \psi[m, k]) \geq 0$ being nonnegative weights. Therefore, (38) is readily satisfied for

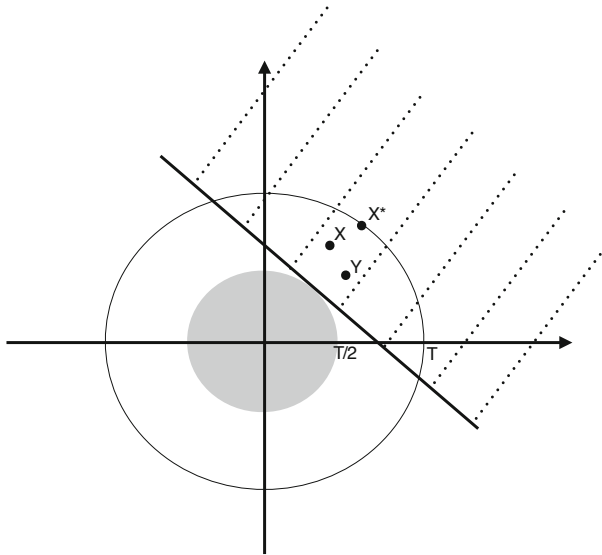


Fig. 2 Example of a region of L -smoothness covering signal estimates satisfying $|X_*| > T/2$

$$|X[m, k]|, |Y[m, k]| \geq T[m, k]/2 \quad (43)$$

because each term in the sum is then nonnegative. This remains true at any couple of points Z, Y , with Z in the line segment with endpoints X and Y , as long as $X[m, k]$ and $Y[m, k]$ belong to the same half-plane of complex numbers with modulus greater than or equal to $T[m, k]/2$. This constitute an example of unbounded regions $\mathcal{R}$ of L -smoothness that cover all optimal signal estimates x_* with $|X_*| = T$, as well as suboptimal estimates with $|X_*| > T/2$, see Fig. 2.

More generally, (38) being less restrictive than (43), x, y may also satisfy (17) and (18) in sets containing signal estimates x_* for which $|X_*[m, k]| \leq T[m, k]/2$ at some time-frequency indices. In fact, condition (38) holds true at endpoints X, Y , and for every pair Z, Y with Z on the line segment joining them, whenever the positively weighted contributions from bins at which $X[m, k]$ and $Y[m, k]$ lie in the same half-plane are sufficiently large to compensate for the negative contributions from the remaining bins.

3 Fast Griffin-Lim algorithm

In this section, we consider GLA-based inertial methods that speed up the double projection by means of a momentum term. Specifically, we focus on the Fast Griffin-Lim Algorithm [26]. In [33], by looking at FGLA as the acceleration with constant momentum of GLA, we have proved its convergence for a broad range of momenta under the local L -smoothness assumption. Here, we prove a slightly more general result.

3.1 Fast Griffin-Lim double projection

In [26], Perraudin et al. introduced the double projection method

$$\begin{cases} X_{i+1} = P_{\mathcal{V}}(P_{\mathcal{T}}(Y_i)), \\ Y_{i+1} = X_{i+1} + \alpha(X_{i+1} - X_i), \end{cases} \quad (44)$$

with $X_0 = Y_0$, $|Y_0| = T$, $\alpha \geq 0$ a constant momentum coefficient. With $\alpha = 0$, FGLA coincides with the original GLA.

It was noted that FGLA achieves optimal performance when α is close to, but less than, one (e.g., $\alpha = 0.99$). By contrast, setting $\alpha \geq 1$ can lead to instability or prevent convergence altogether. FGLA consistently outperforms GLA in experiments and is one of the most used iterative algorithms in audio phase recovery.

In [24], Nenov et al. introduced a method called Accelerated Griffin-Lim Algorithm. For initial points X_0 , Y_0 , Z_0 and constant parameters α , β , $\gamma > 0$, AGLA iterates are defined by

$$\begin{cases} X_{i+1} = (1 - \gamma)Z_i + \gamma P_{\mathcal{V}}(P_{\mathcal{T}}(Y_i)), \\ Y_{i+1} = X_{i+1} + \alpha(X_{i+1} - X_i), \\ Z_{i+1} = X_{i+1} + \beta(X_{i+1} - X_i). \end{cases} \quad (45)$$

With $\gamma = 1$ AGLA coincides with FGLA.

From the numerical evaluation in [24], the AGLA configurations which perform best have $\alpha \in [0.95, 1.05]$, $\beta \in [0.99, 1.40]$, and $\gamma \in [1.25, 1.30]$. The convergence of the algorithm is proved for [24]

$$\begin{aligned} 0 &< \gamma < 2, \\ 0 &\leq 2\beta|1 - \gamma| < 2 - \gamma, \\ 0 &\leq \alpha < \begin{cases} \left(1 - \frac{1}{\gamma}\right)\beta + \frac{1}{\gamma} - \frac{1}{2}, & 0 < \gamma \leq 1, \\ \frac{1}{2\beta(\gamma-1)+\gamma} - \frac{1}{2}, & 1 < \gamma < 2, \end{cases} \end{aligned} \quad (46)$$

hence all the best-performing parameter settings do not satisfy the convergence guarantee (46). In particular, taking $\gamma = 1$ in (46), the convergence of FGLA follows for $\alpha \in [0, 1/2]$.

As such, the convergence of the most effective FGLA and AGLA remains to be proved.

3.2 Fast Griffin-Lim gradient descent

From the representation of the gradient (15), the FGLA update (44) in the time domain is formulated as the acceleration with constant momentum [32] of the GLA gradient descent (16):

$$\begin{cases} x_{i+1} = y_i - \frac{1}{L}\nabla d(y_i), \\ y_{i+1} = x_{i+1} + \alpha(x_{i+1} - x_i). \end{cases} \quad (47)$$

In the above algorithm (47) the step-size can be taken from the interval $(0, 1/L]$ [32]. Replacing $1/L$ with γ/L , $\gamma \leq 1$, we have

$$\begin{cases} x_{i+1} = y_i - \frac{\gamma}{L} \nabla d(y_i), \\ y_{i+1} = x_{i+1} + \alpha(x_{i+1} - x_i), \end{cases} \quad (48)$$

that corresponds to

$$\begin{cases} X_{i+1} = (1 - \gamma)Y_i + \gamma P_V(P_T(Y_i)), \\ Y_{i+1} = X_{i+1} + \alpha(X_{i+1} - X_i), \end{cases} \quad (49)$$

in the time-frequency domain, hence to AGLA (45) with parameters $\beta = \alpha$, $0 < \gamma \leq 1$, and $Z_i = Y_i$.

In order to prove results of convergence for FGLA together with its generalization with parameter $\gamma \neq 1$ in (49), we look at (48) as the discretization of the dynamical system

$$\ddot{x} + (1 - \alpha)\dot{x} + \nabla d(x) = 0 \quad (50)$$

with $x = x(t)$ a continuous-time signal [44–46].

A Lyapunov function for (50) is given by

$$\hat{d}(x(t)) = d(x(t)) + \frac{1}{2} \|\dot{x}(t)\|^2, \quad (51)$$

where $d(x)$ and $\|\dot{x}\|^2/2$ are the potential and kinetic energy, respectively [31]. Differentiating with respect to time, we get

$$\frac{d}{dt} \hat{d}(x(t)) = -(1 - \alpha) \|\dot{x}(t)\|^2 \leq 0 \quad (52)$$

hence the monotonic descent of $\hat{d}(x(t))$. For the discrete system (48) of the generalized FGLA we introduce the Lyapunov sequence

$$\hat{d}(x_i) = d(x_i) + \frac{L}{2\gamma} \|x_i - x_{i-1}\|^2. \quad (53)$$

Theorem 3 In the momentum-based gradient descent algorithm (48), let $\alpha \in [0, 1)$ and $\gamma \in (0, 1]$ be such that

$$\alpha < \frac{1}{\sqrt{1 + \gamma}}, \quad (54)$$

in particular

$$\alpha < \frac{1}{\sqrt{2}} \quad (55)$$

in the FGLA case $\gamma = 1$. Assume that there exists $i_0 \in \mathbb{N}$ such that the iterates x_i, y_ℓ belong to a region of L -smoothness for $i, \ell \geq i_0$. Then, we have:

- (i) The sequence $\hat{d}(x_i)$ in (53) is decreasing for $i \geq i_0$, hence converging to the lower bound $d_* = \inf\{\hat{d}(x_i) : i \geq i_0\}$.
- (ii) The positive sequence $\|x_i - x_{i-1}\|$ converges to zero. In particular, $d(x_i)$ converges to the same limit d_* of $\hat{d}(x_i)$.
- (iii) $\nabla d(y_i) \rightarrow 0$ with

$$\min_{i_0 \leq \ell \leq i-1} \|\nabla d(y_\ell)\|^2 \leq \frac{2L(\hat{d}(x_{i_0}) - d_*)}{C_{\alpha, \gamma}(i - i_0)}, \quad (56)$$

$$C_{\alpha, \gamma} = \begin{cases} \gamma(1 - \gamma), & \gamma < 1, \\ (1 - 2\alpha^2)/2(1 + \alpha^2), & \gamma = 1. \end{cases}$$

- (iv) x_i, y_i are bounded sequences. Any limit point x_* of convergent subsequences is a critical point for $d(x)$ hence such that $X_* = P_V(P_T(X_*))$.

Proof From inequality (21) with $x = x_{i+1}$, $y = y_i$, $i \geq i_0$, and the update in (48), we have

$$d(x_{i+1}) \leq d(y_i) - \frac{\gamma(2 - \gamma)}{2L} \|\nabla d(y_i)\|^2 \quad (57)$$

whereas (21) with $x = x_i$, $y = y_i$ implies

$$d(y_i) \leq d(x_i) - \langle \nabla d(y_i), x_i - y_i \rangle + \frac{L}{2} \|x_i - y_i\|^2. \quad (58)$$

From inequality (58) and the identity

$$-\langle a, b \rangle = -\frac{\|a + b\|^2}{2} + \frac{\|a\|^2}{2} + \frac{\|b\|^2}{2},$$

we obtain

$$\begin{aligned} d(y_i) &\leq d(x_i) - \frac{L}{\gamma} \langle \frac{\gamma}{L} \nabla d(y_i), x_i - y_i \rangle + \frac{L}{2} \|x_i - y_i\|^2 = \\ &= d(x_i) - \frac{L}{2\gamma} \|\frac{\gamma}{L} \nabla d(y_i) - y_i + x_i\|^2 + \frac{\gamma}{2L} \|\nabla d(y_i)\|^2 + \frac{L(1 + \gamma)}{2\gamma} \|x_i - y_i\|^2 = \\ &= d(x_i) - \frac{L}{2\gamma} \|x_i - x_{i+1}\|^2 + \frac{\gamma}{2L} \|\nabla d(y_i)\|^2 + \frac{L(1 + \gamma)}{2\gamma} \|x_i - y_i\|^2. \end{aligned} \quad (59)$$

Combining inequalities (57) and (59), it follows that

$$d(x_{i+1}) + \frac{L}{2\gamma} \|x_i - x_{i+1}\|^2 \leq d(x_i) - \frac{\gamma(1 - \gamma)}{2L} \|\nabla d(y_i)\|^2 + \frac{L(1 + \gamma)}{2\gamma} \|x_i - y_i\|^2. \quad (60)$$

Now, substituting $y_i - x_i = \alpha(x_i - x_{i-1})$ and adding $L\|x_i - x_{i-1}\|^2/2\gamma$ to both sides in (60), we have

$$\hat{d}(x_i) - \hat{d}(x_{i+1}) \geq \frac{\gamma(1-\gamma)}{2L} \|\nabla d(y_i)\|^2 + (1-\alpha^2 - \gamma\alpha^2) \frac{L\|x_i - x_{i-1}\|^2}{2\gamma}. \quad (61)$$

This proves the decrease $\hat{d}(x_{i+1}) \leq \hat{d}(x_i)$ for $i \geq i_0$ thanks to $\gamma \in (0, 1]$ and (54).

Next, telescoping in (61), we obtain both

$$(1-\alpha^2 - \gamma\alpha^2) \sum_{i=i_0}^{N-1} \frac{L\|x_i - x_{i-1}\|^2}{2\gamma} \leq \hat{d}(x_{i_0}) - \hat{d}(x_N) \leq \hat{d}(x_{i_0}) - d_* \quad (62)$$

and

$$\frac{\gamma(1-\gamma)}{2L} \sum_{i=i_0}^{N-1} \|\nabla d(y_i)\|^2 \leq \hat{d}(x_{i_0}) - \hat{d}(x_N) \leq \hat{d}(x_{i_0}) - d_*. \quad (63)$$

From (62), it follows that the kinetic energy $L\|x_i - x_{i-1}\|^2/2\gamma \rightarrow 0$ as the general term of a convergent series. In particular, the potential energy $d(x_i)$ converges to the equilibrium energy level d_* .

As far as $\nabla d(y_i)$ is concerned, inequality (63) implies $\nabla d(y_i) \rightarrow 0$ with rate (56) for $\gamma < 1$ and $C_{\alpha,\gamma} = \gamma(1-\gamma)$. To include the FGLA case $\gamma = 1$, we combine

$$\begin{aligned} \|\nabla d(y_i)\|^2 &= \frac{L^2}{\gamma^2} \|x_{i+1} - x_i - \alpha(x_i - x_{i-1})\|^2 \leq \\ &\leq \frac{4L}{\gamma} \left(\frac{L\|x_{i+1} - x_i\|^2}{2\gamma} + \alpha^2 \frac{L\|x_i - x_{i-1}\|^2}{2\gamma} \right), \end{aligned} \quad (64)$$

with (62), obtaining

$$\sum_{i=i_0}^{N-1} \|\nabla d(y_i)\|^2 \leq \frac{4L(1+\alpha^2)(\hat{d}(x_{i_0}) - d_*)}{\gamma(1-\alpha^2 - \gamma\alpha^2)} \quad (65)$$

since inequality (62) is valid for N and $N+1$. This proves that $\nabla d(y_i) \rightarrow 0$ and the estimate (56) with $C_{\alpha,\gamma} = (1-2\alpha^2)/(2(1+\alpha^2))$ also for $\gamma = 1$.

From the representation of the gradient (15) and $\nabla d(y_i) \rightarrow 0$, we have

$$y_i - \text{iSTFT}(P_{\mathcal{T}}(Y_i)) \rightarrow 0$$

which shows that y_i is a bounded sequence as $P_{\mathcal{T}}(Y_i)$. Then, from

$$y_i - x_i = \alpha(x_i - x_{i-1}) \rightarrow 0$$

also the sequence x_i is bounded as y_i and the two sequences have convergent subsequences with same indices. Any same limit point x_* of convergent subsequences y_{ℓ_i}

and x_{ℓ_i} is a critical point for $d(x)$ with value $d(x_*) = d_* = \lim_{i \rightarrow \infty} d(x_i)$ and it is a fixed point of the double projection. $\square$

Remark 2 The conclusions of Theorem 3 hold true under the more general assumption that inequalities (21) and (22) are satisfied by $(x, y) = (x_{i+1}, y_i)$ and $(x, y) = (x_i, y_i)$, respectively, for $i \geq i_0$.

In [33], we proved Theorem 3 for the case $\gamma = 1$. We showed that, under the local L -smoothness assumption, FGLA convergence is guaranteed for momenta $\alpha \in [0, 1/\sqrt{2})$, thereby extending the guarantees in [24] that hold for $\alpha \in [0, 1/2)$ by taking $\gamma = 1$ in (46). For $\gamma < 1$, inequality (54) yields a sufficient condition for the convergence of particular cases of AGLA (45) which is more general than $\alpha < 1 - \gamma/2$, which follows from (46) with $\gamma \leq 1$ and $\beta = \alpha$. Nevertheless, it is worth emphasizing that the convergence guarantees presented in [24] do not require any Lipschitz assumption.

In general, the monotone descent

$$d(x_{i+1}) \leq d(x_i), \quad i \geq 0,$$

that characterizes GLA sequences [20] is not valid for FGLA. Yet, the decrease of the Lyapunov sequence $\hat{d}(x_i)$ provides the convergence $d(x_i) \rightarrow d(x_*)$ of the entire FGLA value sequence.

The convergence of FGLA can be extended to all $\alpha \in [0, 1)$ in the case that the limit point of a convergent subsequence lies in a neighborhood of convexity:

Corollary 2 *If x_i, y_ℓ , for $i, \ell \geq i_c \geq i_0$, and the limit point x_* of a convergent subsequence belong to a region of local convexity for $d(x)$, then the conclusions of Theorem 3 hold true for all $\alpha \in [0, 1)$ and $\gamma \in (0, 1]$ without assuming condition (54). Moreover, x_* is a local minimizer, and the whole sequence x_i converges to x_* .*

Proof If x_i, y_i belong to a region of local convexity for $d(x)$, then we have

$$d(y_i) \leq d(x_i) - \langle \nabla d(y_i), x_i - y_i \rangle. \quad (66)$$

This gives the stronger descent

$$\hat{d}(x_i) - \hat{d}(x_{i+1}) \geq \frac{\gamma(1-\gamma)}{2L} \|\nabla d(y_i)\|^2 + (1-\alpha^2) \frac{L\|x_i - x_{i-1}\|^2}{2\gamma} \quad (67)$$

for the Lyapunov sequence and all the conclusions hold true for any $\alpha \in [0, 1)$ and $\gamma \in (0, 1]$.

Moreover, convexity localizes the iterates

$$\|x_{i_\varepsilon} - x_*\| < \varepsilon \implies \|x_i - x_*\| < \varepsilon \quad \forall i \geq i_\varepsilon \geq i_c,$$

as we have seen in the proof of Corollary 1 for GLA. We refer to Lemma 6.4 in [31] for a unified proof that works both for the classic and accelerated gradient descent

methods. The convergence to x_* of a subsequence ensures that for any $\varepsilon > 0$ there exists such an x_{i_ε} , hence the whole sequence x_i converges to x_* . $\square$

Not all critical points are minimum points and not all minimum points lie in a neighborhood of convexity. However, it is known that inertial methods escape saddle points faster than the classic gradient descent [31, 35–37].

4 Nesterov accelerated Griffin-Lim algorithm

In this section, we propose an inertial method that accelerates GLA according to the Nesterov theory [34] of the gradient descent. We call it Nesterov Accelerated Griffin-Lim Algorithm (NAGLA). In the time-frequency domain, the implementation is similar to that of FGLA (44) but with a sequence of variable momenta instead of a constant one. Nesterov's principles endow NAGLA with a rate of convergence of $\mathcal{O}(1/i^2)$ in neighborhoods of convexity, improving the $\mathcal{O}(1/i)$ rate of GLA.

4.1 Nesterov accelerated gradient descent

We first briefly recall Nesterov's theory. Inertial gradient descent algorithms with constant momentum accelerate the standard gradient descent in case of a L -smooth and μ -strongly convex objective function $f(x)$, that is a function such that $f(x) - \mu\|x\|^2/2$ is convex, $0 < \mu < L$. In this case, the standard gradient descent converges linearly with contraction factor $1 - \mu/L$ [32]. The inertial method with the precise choice $\alpha = (\sqrt{L} - \sqrt{\mu})/(\sqrt{L} + \sqrt{\mu})$ for the momentum coefficient improves the contraction factor to $1 - \sqrt{\mu/L}$ [32]. Acceleration fails in the convex case $\mu = 0$ [32]. Inertial methods with constant momentum exhibit the same $\mathcal{O}(1/i)$ convergence rate as the standard algorithm for any $\alpha \in (0, 1)$, even in a weaker form [47].

In his seminal paper [34], Nesterov introduced an accelerated gradient descent with variable momenta

$$\begin{cases} x_{i+1} = x_i - \frac{\gamma_i}{L} \nabla f(y_i), \\ y_{i+1} = x_{i+1} + \alpha_i (x_{i+1} - x_i). \end{cases} \quad (68)$$

The rate of convergence of the standard gradient descent is improved from $\mathcal{O}(1/i)$ to

$$f(x_i) - f(x_*) = \mathcal{O}(1/i^2).$$

The momenta sequence α_i is increasing and convergent to 1. It is given by

$$\alpha_i = \theta_{i+1}(\theta_i^{-1} - 1) \quad (69)$$

with $\gamma_i, \theta_i \in (0, 1]$ so that

$$\frac{\gamma_{i+1}(1 - \theta_{i+1})}{\theta_{i+1}^2} \leq \frac{\gamma_i}{\theta_i^2}, \quad (70)$$

$$\theta_i = \mathcal{O}(1/i). \quad (71)$$

The original choice in [34] is $\theta_0 = 1$ and equality in (70). Also

$$\theta_i = \frac{2}{i+2} \implies \alpha_i = \frac{i}{i+3}$$

is frequently used [28, 44, 48]. In convex settings, the standard choice for the step-size is given by $\gamma_i = 1$, $\forall i$ [32, 34].

Nesterov's accelerated method together with its variants and generalizations is widely used also in stochastic gradient descent and nonconvex or nonsmooth optimization [27, 28, 30, 39]. In nonconvex landscapes, it escapes saddle points faster than any other first-order algorithm [31] maintaining its optimal rate of convergence in neighborhoods of convexity.

4.2 Nesterov Griffin-Lim double projection

We propose to accelerate the gradient descent (19) following Nesterov's principles. NAGLA is defined by (68) with $f(x)$ the objective function $d(x)$ of GLA, i.e., (10). In the time-frequency domain, this corresponds to

$$\begin{cases} X_{i+1} = (1 - \gamma_i)Y_i + \gamma_i P_{\mathcal{V}}(P_{\mathcal{T}}(Y_i)), \\ Y_{i+1} = X_{i+1} + \alpha_i(X_{i+1} - X_i). \end{cases} \quad (72)$$

By choosing $\gamma_i = 1$, we obtain the FGLA-like double projection

$$\begin{cases} X_{i+1} = P_{\mathcal{V}}(P_{\mathcal{T}}(Y_i)), \\ Y_{i+1} = X_{i+1} + \alpha_i(X_{i+1} - X_i). \end{cases} \quad (73)$$

with variable momenta as in (69)–(70) instead of the constant momentum in (44).

We can give a concrete realization of the algorithm by taking

$$\alpha_i = \frac{i + H - 2}{i + H + 1} \quad (74)$$

for any fixed hyperparameter $H \geq 2$, which comes from substituting

$$\theta_i = \frac{2}{i + H} \quad (75)$$

into (69).

The convergence of NAGLA (72) under the local L -smoothness assumption can be proved with the arguments of the proof of Theorem 3 taking a decreasing step-size sequence $\gamma_i \in (0, 1]$ such that

$$\alpha_i^2 < \frac{\gamma_{i+1}}{\gamma_i(1 + \gamma_i)}, \quad \sum_{i=1}^{+\infty} \gamma_i = +\infty. \quad (76)$$

Note that inequality $\alpha_i^2 < \gamma_{i+1}/\gamma_i(1 + \gamma_i)$ implies $\gamma_i \rightarrow 0$.

According to Nesterov's theory, NAGLA (73) improves the $\mathcal{O}(1/i)$ convergence rate of GLA to $\mathcal{O}(1/i^2)$ in neighborhoods of convexity for $d(x)$. In these neighborhoods, NAGLA converges with $\gamma_i = 1, \forall i$.

Theorem 4 Let the NAGLA iterates x_i, y_ℓ , with $i, \ell \geq i_0$, in (73) enter and remain in a region of L -smoothness for $d(x)$ and for $i, \ell \geq i_c \geq i_0$ belong to a region of local convexity with local minimum interior points.

Then, the limit point x_* of a convergent subsequence of the bounded sequence x_i is a local minimizer, and $d(x_i)$ converges to the local minimum value $d(x_*)$ with rate given by

$$d(x_i) - d(x_*) \leq C_0/i^2, \quad (77)$$

$i > i_c$, C_0 a positive constant. Moreover, the whole sequence x_i converges to x_* .

Proof As the double projection (73), the NAGLA iterates x_i form a bounded sequence. From Nesterov's theory, the limit point x_* of a convergent subsequence is a critical point [29, 49], hence a local minimizer, and $d(x_i) \rightarrow d(x_*)$ with rate given by (77) [34]. The convergence of the entire sequence x_i to x_* follows from the localization of the iterates as in the proofs of Corollary 1 and Corollary 2. $\square$

5 Numerical evaluation

In this section, we investigate by numerical experiments the Lipschitz inequality (17) at the iterates, the expected rates of convergence of GLA and NAGLA, and compare the short-term performance of the considered inertial methods.

In subsections 5.1, 5.2, and 5.3, we use signals from the TIMIT speech corpus,¹ In subsection 5.4, we use music signals from the StarNet dataset.² For the TIMIT experiments, we assume no prior knowledge of the signal phase and initialize the phase to zero. In contrast, the StarNet experiments consider a simple music source separation task, where we evaluate the methods using an informed phase initialization.

TIMIT data. For the experiments presented in subsections 5.1 through 5.3, we use the 72 phonetically diverse (SI) utterances from the TIMIT core test set, spoken by a total of 24 speakers and made available as monophonic WAV files at a sampling rate of 16 kHz. The utterance durations range from 1.33 s to 6.21 s, with an average of 3.32 s. The STFT and iSTFT are computed using a 32 ms Hann window with an 8 ms hop size. This choice of window and analysis parameters satisfies the COLA constraint [20] and provides sufficient overlap to guarantee uniqueness of the solution up to a constant phase [43].

StarNet data. For the experiments presented in subsection 5.4, we use a portion of the StarNet dataset [50] consisting of 52 of its 104 music pieces. The pieces, provided as 16bit-44.1kHz stereo WAV files, include two instrument tracks, and each piece

¹ J. S. Garofolo et al. 1993, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," distributed by Linguistic Data Consortium. [Online]. Available: <https://catalog.ldc.upenn.edu/LDC93S1>.

² M. Alinoori and V. Tzerpos, August 20, 2022, "StarNet," Zenodo. [Online]. Available: <https://dx.doi.org/10.5281/zenodo.6917099>.

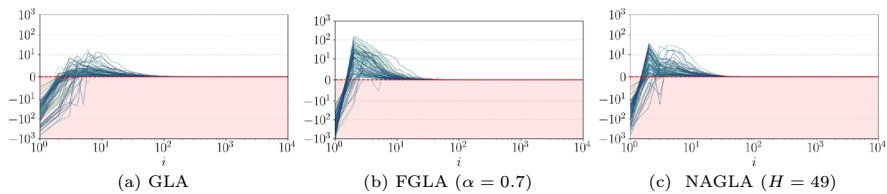


Fig. 3 $\sum_{m,k} T[m, k](1 - \psi[m, k])(|X_{i+1}[m, k]| + |Y_i[m, k]| - T[m, k])$ as a function of i for GLA, FGLA and NAGLA sequences of the 72 SI utterances from TIMIT

is rendered using two instrument sets: *strings-piano* and *clarinet-vibraphone*. Each music piece is several minutes long. First, we convert the stereo tracks to mono by averaging the channels. Then, we extract the first non-overlapping 10-second excerpt where the power of both instruments is above a given threshold. As in the speech experiments, we choose the STFT and iSTFT parameters to achieve a 75% overlap between consecutive analysis windows, this time using a 92 ms Hann window and a 23 ms hop size.

Parameter choices. In subsection 5.1, FGLA is tested with $\alpha = 0.7$, which satisfies the convergence condition $\alpha < \sqrt{1/2}$ established in Theorem 3.

For the comparative study of the performance of inertial methods presented in subsections 5.3 and 5.4, we do not impose parameter choices that guarantee convergence. Instead, FGLA is run with $\alpha = 0.99$, following the recommendation of [26], while AGLA uses the best parameter combination from [24, Table II, p. 197].

As far as NAGLA is concerned, $H \geq 2$ is a tunable parameter that sets the initial momentum. For the case of null initial phase, we choose $H = 49$ to ensure a data-independent setup. This produces a relatively high initial momentum ($\alpha_0 = 47/50 = 0.94$), allowing NAGLA to shorten the transient regime [39]. At the same time, α_0 remains noticeably below one relative to the growth rate determined by (74). Specifically, $H = 49$ yields an initial momentum α_0 that is 0.05 lower than the standard choice $\alpha = 0.99$ of the FGLA momentum and reaches it after 250 iterations.

In the experiments with informed initial phase, instead, we use $H = 29$. This choice reflects the fact that, when some prior phase information is available, the transient regime tends to be shorter, thereby motivating a lower initial momentum ($\alpha_0 = 27/30 = 0.90$) than in the null-phase initialization experiments ($\alpha_0 = 0.94$).

5.1 Lipschitz inequalities

The left-hand side of (38) is evaluated at $(X, Y) = (X_{i+1}, Y_i)$, with $Y_i = X_i$ for GLA, for each of the 72 TIMIT utterances described above. By Proposition 1, nonnegativity of this quantity is sufficient for inequality (17) to hold at points $(x, y) = (x_{i+1}, y_i)$.

Figure 3 reports the resulting values for (a) GLA, (b) FGLA ($\alpha = 0.7$), and (c) NAGLA ($H = 49$). In all three cases, the summation in (38) becomes nonnegative after only a few iterations. Thereafter, it converges toward zero as phase differences diminish across time-frequency bins, with this behavior appearing at relatively small iteration indices. We emphasize, however, that this pairwise test does not verify condition (18) at the interior points of the corresponding line segments.

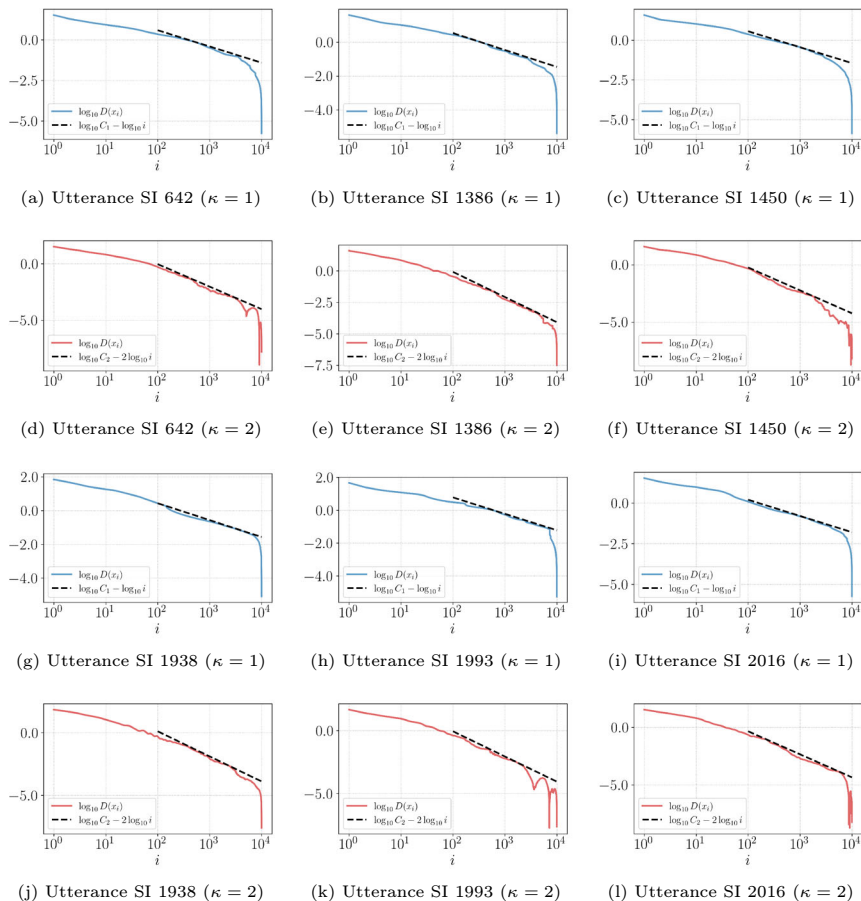


Fig. 4 Logarithm of the absolute difference, $\log_{10} D(x_i)$, for GLA (blue) and NAGLA (red) sequences. The dashed black lines with slope $-\kappa$ represent reference bounds corresponding to the theoretical asymptotic convergence rates $\mathcal{O}(1/i^\kappa)$, $\kappa = 1$ for GLA, $\kappa = 2$ for NAGLA

5.2 Numerical evaluation of the rate of convergence

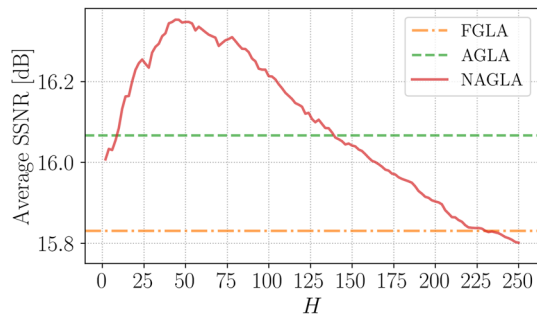
In this subsection, we empirically examine the asymptotic convergence rates of GLA and NAGLA. We use the same dataset and the same parameter choices as in previous subsection 5.1.

Figure 4 depicts the difference

$$D(x_i) = |d(x_i) - d(x_{N_*})|. \quad (78)$$

With $N_* = 10^4$, the value $d(x_{N_*})$ serves as an approximation of the local minimum $d(x_*)$. Although the minimum index at which the iterates start to exhibit the asymptotic rate may vary from signal to signal, we set $i_c = 100$ to illustrate the overall trend. For better visualization, the analysis is carried out in logarithmic scale with a machine

Fig. 5 Average SSNR at 100 iterations as a function of H . The maximum value is taken at $H = 44$.



epsilon so that $D(x_{N_*}) \neq 0$, and absolute value to prevent effects from the possible nonmonotonic behavior of the inertial method NAGLA. The rate is verified when $\log_{10} D(x_i)$ lies below a line of slope $-\kappa$, where $\kappa = 1$ for GLA, and $\kappa = 2$ for NAGLA.

It turns out that, for most signals, the observed behavior of GLA and NAGLA is consistent with the asymptotic rates prescribed by Corollary 1 and Theorem 4, respectively.

5.3 Short-term performance of inertial methods

Having performed the convergence analysis for $i \geq 100$, we now examine the early-iteration performance of NAGLA, and compare it with FGLA, implemented with a constant momentum of $\alpha = 0.99$ as specified in [26], and AGLA, where we set $\alpha = 0.99$, $\beta = 1$, and $\gamma = 1.3$, i.e., the best parameter combination from [24, Table II, p. 197]. For this purpose, we report the Spectrogram Signal-to-Noise Ratio (SSNR) [23, 24, 26]

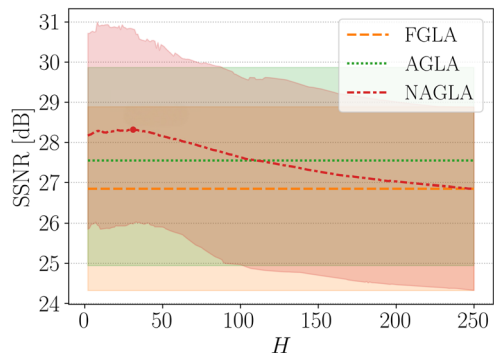
$$\text{SSNR}(x) = 10 \log_{10} \frac{\|T\|_F}{\| |\text{STFT}(x)| - T \|_F}, \quad (79)$$

computed between the ground-truth signals from TIMIT and the estimates at iteration $i = 100$. Notably, minimizing the function $d(x)$ in (10) is equivalent to maximizing $\text{SSNR}(x)$ in (79).

NAGLA turns out to have the same computational complexity of FGLA since, at iteration i , the constant momentum α in (44) is simply replaced by the prescribed value from the sequence α_i . This operation is $\mathcal{O}(1)$ in both time and space. Comparatively, the update of Z_i in AGLA (45) entails an additional $\mathcal{O}(MK)$ complexity factor, MK the total number of time-frequency bins.

NAGLA achieves a moderate but consistent improvement over the other fast algorithms. At $H = 49$ (Fig. 5), the average SSNR values are 15.83 dB, 16.06 dB, and 16.35 dB for FGLA, AGLA, and NAGLA, respectively. Moreover, all fast methods outperform GLA by approximately 4 dB.

Fig. 6 SSNR after 100 iterations as a function of H on the considered StarNet partition. The shaded areas denote the interquartile ranges



We analyze the sensitivity of NAGLA to variations in its hyperparameter and demonstrate that a broad range of values yields consistently good performance. A grid search was conducted over integer values of H in the interval $[2, 250]$. For each value, NAGLA was run for 100 iterations on all SI utterances in the TIMIT core test set. The grid search range was chosen so that its upper limit, $H = 250$, yields a momentum sequence $\{\alpha_i\}$ that remains nearly constant ($\pm 10^{-3}$) around its central value $\alpha_{50} \approx 0.99$, which corresponds to the constant momentum used in FGLA.

As shown in Fig. 5, NAGLA achieves an average SSNR higher than that of AGLA and FGLA across a wide range of H values, outperforming AGLA for $10 \leq H \leq 140$, and surpassing FGLA for $2 \leq H \leq 232$. Although the overall improvement is modest, the breadth of these intervals indicates a substantial margin for the practical choice of H and that NAGLA is robust with respect to such a choice.

5.4 Informed initial phase

In this subsection, we investigate the use of an alternative to zero-phase initialization. For all 52 StarNet excerpts, we compute the magnitude T of the solo clip and initialize the algorithm with the phase of the corresponding two-instrument mixture. For example, given a solo *strings* clip, the methods are initialized with the phase of the corresponding *strings-piano* mixture.

We compute the SSNR at 100 iterations as a function of H in the range $[2, 250]$ for all clips. Figure 6 depicts the average value and the corresponding interquartile range for FGLA, AGLA, and NAGLA.

At $H = 29$, the average SSNR is 24 dB for GLA, 26.85 for FGLA, 27.55 for AGLA, and 28.15 for NAGLA. The gain of NAGLA with respect to FGLA is of 4.8%, higher than the 3.3% showed in the experiments from TIMIT with null phase initialization reported in subsection 5.3. The red dot in Fig. 6 marks the highest average SSNR achieved by NAGLA, which occurs at $H = 31$ ($\alpha_0 \approx 0.906$).

Finally, Fig. 7 presents illustrative examples of individual signals that we consider particularly representative of recurring trends observed across the dataset.

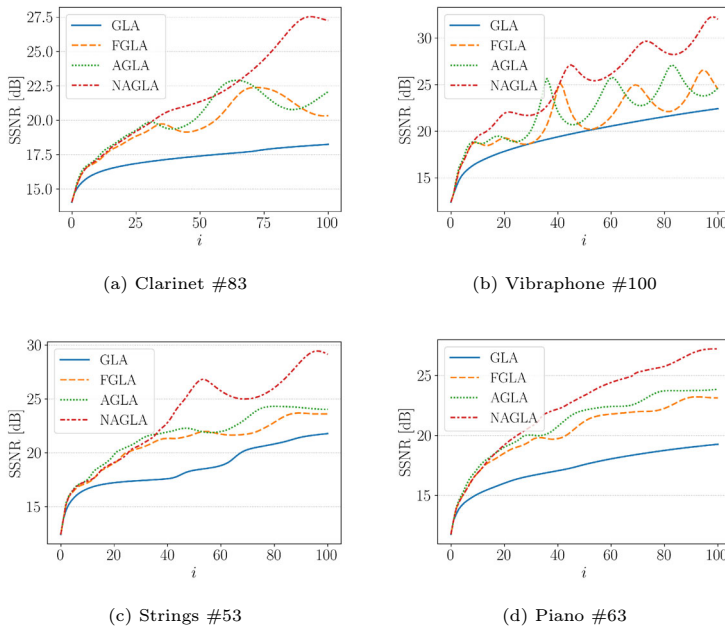


Fig. 7 SSNR as a function of the iteration index i for four audio held-out music pieces from StarNet ($H = 29$)

6 Conclusions

We have formulated the Griffin-Lim Algorithm (GLA) as a gradient descent algorithm and identified regions of L -smoothness for the objective function that contain all optimal and a full class of suboptimal signal estimates.

Under local L -smoothness assumptions, we provided an alternative proof of convergence for GLA that also applies to generalized gradient descent formulations with different step sizes. We also established convergence guarantees for the Fast Griffin-Lim Algorithm (FGLA) over a broad range of momenta. Furthermore, inspired by Nesterov's acceleration theory, we introduced the Nesterov Accelerated Griffin-Lim Algorithm (NAGLA), an accelerated gradient descent method with variable momenta, and proposed a principled choice for its momentum sequence.

Under the further assumption that the iterates enter a region of local convexity, we also showed that GLA achieves a rate of convergence of $\mathcal{O}(1/i)$, whereas NAGLA achieves the accelerated rate of $\mathcal{O}(1/i^2)$.

Going forward, the performance of the proposed and other Nesterov accelerated phase retrieval methods is worth being investigated further, considering different types of signals, initialization strategies, and audio quality metrics.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement. The first author was partially supported by the Italian Ministry of University and Research, under PRIN2022 (Scorrimento) Anomalies in partial differential equations and applications, 2022HCLAZ8 002, J53C24002560006

Data Availability All data supporting the findings of this study are available within the paper.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

Ethical approval All principles of ethical and professional conduct have been followed.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Klibanov, M.V., Sacks, P.E., Tikhonravov, A.V.: The phase retrieval problem. *Inverse Prob.* **11**(1), 1 (1995). <https://doi.org/10.1088/0266-5611/11/1/001>
2. Shechtman, Y., Eldar, Y.C., Cohen, O., Chapman, H.N., Miao, J., Segev, M.: Phase retrieval with application to optical imaging: a contemporary overview. *IEEE Signal Process. Mag.* **32**(3), 87–109 (2015). <https://doi.org/10.1109/MSP.2014.2352673>
3. Millane, R.P.: Phase retrieval in crystallography and optics. *J. Opt. Soc. Am. A* **7**(3), 394–411 (1990)
4. Fienup, C., Dainty, J.: Phase retrieval and image reconstruction for astronomy. *Image Recovery Theory Appl.* **231**, 275 (1987)
5. Magron, P., Badeau, R., David, B.: Phase recovery in NMF for audio source separation: an insightful benchmark. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 81–85. IEEE, South Brisbane, Queensland, Australia (2015)
6. Takaki, S., Kameoka, H., Yamagishi, J.: Direct modeling of frequency spectra and waveform generation based on phase recovery for DNN-based speech synthesis. In: INTERSPEECH 2017, Stockholm, Sweden, pp. 1128–1132 (2017)
7. Wang, Z.-Q., Roux, J.L., Wang, D., Hershey, J.R.: End-to-end speech separation with unfolded iterative phase reconstruction. In: INTERSPEECH 2018, Hyderabad, India (2018)
8. Drossos, K., Magron, P., Mimilakis, S.I., Virtanen, T.: Harmonic-percussive source separation with deep neural networks and phase recovery. In: 16th International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 421–425. IEEE, Tokyo, Japan (2018)
9. Magron, P., Badeau, R., David, B.: Model-based STFT phase recovery for audio source separation. *IEEE/ACM Trans. Audio Speech Language Process.* **26**(6), 1095–1105 (2018)
10. Cano, E., FitzGerald, D., Liutkus, A., Plumbley, M.D., Stöter, F.-R.: Musical source separation: An introduction. *IEEE Signal Process. Mag.* **36**(1), 31–40 (2018)
11. Mezza, A.I., Giampiccolo, R., Bernardini, A., Sarti, A.: Toward deep drum source separation. *Pattern Recogn. Lett.* **183**, 86–91 (2024). <https://doi.org/10.1016/j.patrec.2024.04.026>
12. Tan, X., Qin, T., Soong, F., Liu, T.-Y.: A survey on neural speech synthesis, (2021). Preprint at [arXiv:2106.15561](https://arxiv.org/abs/2106.15561)
13. Comanducci, L., Antonacci, F., Sarti, A.: Timbre transfer using image-to-image denoising diffusion implicit models. In: Proc. of the 24th International Society for Music Information Retrieval Conference (ISMIR), pp. 257–263., Milan, Italy (2024)
14. Le Roux, J., Ono, N., Sagayama, S.: Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction. In: Proceedings of the ITRW on Statistical and Perceptual Audio Processing, pp. 23–28., Brisbane, Queensland, Australia (2008)

15. Le Roux, J., Kameoka, H., Ono, N., Sagayama, S.: Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency. In: Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10), pp. 1–7., Graz, Austria (2010)
16. Wang, Y., Skerry-Ryan, R.J., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyriannakis, Y., Clark, R., Saurous, R.A.: Tacotron: Towards end-to-end speech synthesis. In: INTERSPEECH 2017, Stockholm, Sweden, pp. 4006–4010 (2017). <https://doi.org/10.21437/Interspeech.2017-1452>
17. Ping, W., Peng, K., Gibiansky, A., Arik, S.O., Kannan, A., Narang, S., Raiman, J., Miller, J.: Deep voice 3: Scaling text-to-speech with convolutional sequence learning. In: International Conference on Learning Representations (ICLR), Vancouver, BC, Canada (2018)
18. Liu, H., Baoueb, T., Fontaine, M., Le Roux, J., Richard, G.: GLA-GRAD: A Griffin-Lim extended waveform generation diffusion model. In: 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, pp. 11611–11615 (2024). <https://doi.org/10.1109/ICASSP48485.2024.10446058> . IEEE
19. Baoueb, T., Bie, X., Fontaine, M., Richard, G.: GLA-GRAD++: An improved Griffin-Lim guided diffusion model for speech synthesis. In: ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, pp. 17732–17736 (2026)
20. Griffin, D., Lim, J.: Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust. Speech Signal Process.* **32**(2), 236–243 (1984)
21. Zhu, X., Beaugregard, G.T., Wyse, L.L.: Real-time signal estimation from modified short-time Fourier transform magnitude spectra. *IEEE Trans. Audio Speech Lang. Process.* **15**(5), 1645–1653 (2007)
22. Decorsière, R., Søndergaard, P.L., MacDonald, E.N., Dau, T.: Inversion of auditory spectrograms, traditional spectrograms, and other envelope representations. *IEEE/ACM Trans. Audio Speech Language Process.* **23**(1), 46–56 (2014)
23. Peer, T., Welker, S., Gerkmann, T.: Beyond Griffin-Lim: Improved iterative phase retrieval for speech. In: 2022 International Workshop on Acoustic Signal Enhancement (IWAENC), Bamberg, Germany, pp. 1–5 (2022). IEEE
24. Nenov, R., Nguyen, D.-K., Balazs, P., Boş, R.I.: Accelerated Griffin-Lim algorithm: A fast and provably converging numerical method for phase retrieval. *IEEE Trans. Signal Process.*, 190–202 (2023)
25. Gerchberg, R.W., Saxton, W.O.: A practical algorithm for the determination of the phase from image and diffraction plane pictures. *Optik* **2**, 237–250 (1972)
26. Perraudin, N., Balazs, P., Søndergaard, P.L.: A fast Griffin-Lim algorithm. In: 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, pp. 1–4 (2013). IEEE
27. Ghadimi, S., Lan, G.: Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Program.* **156**(1), 59–99 (2016)
28. Guminov, S.V., Nesterov, Y.E., Dvurechensky, P.E., Gasnikov, A.V.: Accelerated primal-dual gradient descent with linesearch for convex, nonconvex, and nonsmooth optimization problems. *Dokl. Math.* **99**, 125–128 (2019) Springer
29. Nesterov, Y.E.: How to make the gradients small. *Optima* **88**, 10–11 (2012)
30. Li, H., Lin, Z.: Accelerated proximal gradient methods for nonconvex programming. In: Proc. Adv. Neural Inf. Process. Syst., Montreal, Canada, pp. 379–387 (2015). NIPS
31. Dixit, R., Gurbuzbalaban, M., Bajwa, W.U.: Accelerated gradient methods for nonconvex optimization: Escape trajectories from strict saddle points and convergence to local minima, (2023). Preprint at [arXiv:2307.07030](https://arxiv.org/abs/2307.07030)
32. Nesterov, Y.E.: Introductory Lectures on Convex Optimization. Applied Optimization. Springer, New York, USA (2004)
33. Mezza, A.I., Cicognani, M., Cicognani, R.L., Bernardini, A.: On the convergence of the fast griffin-lim algorithm. *IEEE Signal Processing Letters* **33**, 2006–2010 (2026) <https://doi.org/10.1109/LSP.2026.3693202>
34. Nesterov, Y.E.: A method for solving the convex programming problem convergence rate $O(1/k^2)$. *Soviet Math. Doklady* **27**, 372–376 (1983)
35. Jin, C., Netrapalli, P., Jordan, M.I.: Accelerated gradient descent escapes saddle points faster than gradient descent. In: Proceedings of the 31st Conference On Learning Theory, pp. 1042–1085., Stockholm, Sweden, PMLR (2018)
36. O'Neill, M., Wright, S.J.: Behavior of accelerated gradient methods near critical points of nonconvex functions. *Math. Program.* **176**, 403–427 (2019)

37. Wang, J.-K., Lin, C.-H., Abernethy, J.: Escaping saddle points faster with stochastic momentum. In: International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia (2020)
38. Luenberger, D.G., Ye, Y.: Linear and Nonlinear Programming, vol. 2, Springer, Cham, Switzerland (2021). (5th edn)
39. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: Proc. of the 30th International Conference on Machine Learning (ICML), vol. 28, pp. 1139–1147. PMLR, Atlanta, Georgia, USA (2013)
40. Candès, E.J., Eldar, Y.C., Strohmer, T., Voroninski, V.: Phase retrieval via matrix completion. *SIAM J. Imag. Sci.* **6**(1), 199–225 (2013). <https://doi.org/10.1137/110848074>
41. Sun, D.L., Smith, L., J.O.: Estimating a signal from a magnitude spectrogram via convex optimization. *Journal of the Audio Engineering Society* 8785, (2012)
42. Sanz, J.L.C.: Mathematical considerations for the problem of fourier transform phase retrieval from magnitude. *SIAM J. Appl. Math.* **45**, 651–664 (1985)
43. Balan, R., Casazza, P., Edidin, D.: On signal reconstruction without phase. *Appl. Comput. Harmon. Anal.* **20**, 345–356 (2006)
44. Su, W., Boyd, S., Candès, E.: A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *J. Mach. Learn. Res.* **17**(153), 1–43 (2016)
45. Shi, B., Du, S.S., Jordan, M.I., Su, W.J.: Understanding the acceleration phenomenon via high-resolution differential equations. *Math. Program.* **195**, 79–148 (2022)
46. Chen, S., Shi, B., Yuan, Y.-X.: Gradient norm minimization of Nesterov acceleration: $o(1/k^3)$, (2022). Preprint at [arXiv:2209.08862](https://arxiv.org/abs/2209.08862)
47. Ghadimi, E., Feyzmahdavian, H.R., Johansson, M.: Global convergence of the heavy-ball method for convex optimization. In: European Control Conference (ECC), pp. 310–315., Linz, Austria (2015)
48. Allen-Zhu, Z., Orecchia, L.: Linear coupling: An ultimate unification of gradient and mirror descent. In: 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), pp. 3–1322., Berkeley, CA, USA (2017)
49. Walkington, N.J.: Nesterov's method for convex optimization. *SIAM Rev.* **65**(2), 539–562 (2023)
50. Alinoori, M., Tzerpos, V.: Music-STAR: A style translation system for audio-based re-instrumentation. In: Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR), pp. 419–426., Bengaluru, India (2022)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.