Proceedings of Machine Learning Research vol 75:1–44, 2018 31st Annual Conference on Learning Theory

Accelerated Gradient Descent Escapes Saddle Points Faster than Gradient Descent

Chi Jin CHIJIN@CS.BERKELEY.EDU
University of California, Berkeley

Praneeth Netrapalli PRANEETH@MICROSOFT.COM
Microsoft Research India

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
University of California, Berkeley

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

Nesterov's accelerated gradient descent (AGD), an instance of the general family of "momentum methods," provably achieves faster convergence rate than gradient descent (GD) in the convex setting. While these methods are widely used in modern *nonconvex* applications, including training of deep neural networks, whether they are provably superior to GD in the nonconvex setting remains open. This paper studies a simple variant of Nesterov's AGD, and shows that it escapes saddle points and finds a second-order stationary point in $\tilde{O}(1/\epsilon^{7/4})$ iterations, matching the best known convergence rate, which is faster than the $\tilde{O}(1/\epsilon^2)$ iterations required by GD. To the best of our knowledge, this is the first direct acceleration (single-loop) algorithm that is provably faster than GD in general nonconvex setting—all previous nonconvex accelerated algorithms rely on more complex mechanisms such as nested loops and proximal terms. Our analysis is based on two key ideas: (1) the use of a simple Hamiltonian function, inspired by a continuous-time perspective, which AGD monotonically decreases on each step even for nonconvex functions, and (2) a novel framework called *improve or localize*, which is useful for tracking the long-term behavior of gradient-based optimization algorithms. We believe that these techniques may deepen our understanding of both acceleration algorithms and nonconvex optimization.

1. Introduction

Nonconvex optimization problems are ubiquitous in modern machine learning. While it is NP-hard to find global minima of a nonconvex function in the worst case, in the setting of machine learning it has proved useful to consider a less stringent notion of success, namely that of convergence to a *first-order stationary point* (where $\nabla f(\mathbf{x}) = 0$). Gradient descent (GD), a simple and fundamental optimization algorithm that has proved its value in large-scale machine learning, is known to find an ϵ-first-order stationary point (where $\|\nabla f(\mathbf{x})\| \le \epsilon$) in $O(1/\epsilon^2)$ iterations (Nesterov, 1998), and this rate is sharp (Cartis et al., 2010). Such results, however, do not seem to address the practical success of gradient descent; first-order stationarity includes local minima, saddle points or even local maxima, and a mere guarantee of convergence to such points seems unsatisfying. Indeed, architectures such as deep neural networks induce optimization surfaces that can be teeming with such highly suboptimal saddle points (Dauphin et al., 2014). It is important to study to what ex-

tent gradient descent avoids such points, particularly in the high-dimensional setting in which the directions of escape from saddle points may be few.

This paper focuses on convergence to a *second-order stationary point* (where $\nabla f(\mathbf{x}) = 0$ and $\nabla^2 f(\mathbf{x}) \succeq 0$). Second-order stationarity rules out many common types of saddle points (*strict* saddle points where $\lambda_{\min}(\nabla^2 f(\mathbf{x})) < 0$), allowing only local minima and higher-order saddle points. A significant body of recent work, some theoretical and some empirical, shows that for a large class of well-studied machine learning problems, neither higher-order saddle points nor spurious local minima exist. That is, *all second-order stationary points are (approximate) global minima* for these problems. Choromanska et al. (2014) and Kawaguchi (2016) present such a result for learning neural networks, Bandeira et al. (2016); Mei et al. (2017) for synchronization and MaxCut, Boumal et al. (2016) for smooth semidefinite programs, Bhojanapalli et al. (2016) for matrix sensing, Ge et al. (2016) for matrix completion, and Ge et al. (2017) for robust PCA. These results strongly motivate the quest for *efficient algorithms* to find second-order stationary points.

Hessian-based algorithms can explicitly compute curvatures and thereby avoid saddle points (see, e.g., Nesterov and Polyak, 2006; Curtis et al., 2014), but these algorithms are computationally infeasible in the high-dimensional regime. GD, by contrast, is known to get stuck at strict saddle points (Nesterov, 1998, Section 1.2.3). Recent work has reconciled this conundrum in favor of GD; Jin et al. (2017), building on earlier work of Ge et al. (2015), show that a perturbed version of GD converges to an ϵ-relaxed version of a second-order stationary point (see Definition 6) in $\tilde{O}(1/\epsilon^2)$ iterations. That is, perturbed GD in fact finds second-order stationary points as fast as standard GD finds first-order stationary points, up to logarithmic factors in dimension.

On the other hand, GD is known to be suboptimal in the convex case. In a celebrated paper, Nesterov (1983) showed that an accelerated version of gradient descent (AGD), which takes "momentum steps" in addition to gradient steps, finds an ϵ-suboptimal point (see Section 2.2) in $O(1/\sqrt{\epsilon})$ steps, while gradient descent takes $O(1/\epsilon)$ steps. The basic idea of momentum acceleration has been used to design faster algorithms for a range of other convex optimization problems (Beck and Teboulle, 2009; Nesterov, 2012; Lee and Sidford, 2013). We will refer to this general family as "momentum-based methods." Such results have focused on the convex setting.

In the nonconvex setting, while there has been recent work on designing accelerated algorithms that converge faster than GD (see, e.g., Agarwal et al., 2017; Carmon et al., 2016, 2017), these proposals rely on more complex mechanisms, including nested loops and the incorporation of proximal terms to achieve the faster rate. In contrast, empirically, Nesterov's classical, single-loop AGD and the related family of "momentum-based algorithms"—have been used successfully in modern large-scale nonconvex applications, where they have been observed to perform better than GD (Sutskever et al., 2013). Providing a theoretical understanding of the scope of this phenomenon, and its possible limitations, is an interesting and important open problem. We are thus led to ask the following question:

Does Nesterov's AGD (or a simple variant) yield faster convergence than GD in general nonconvex setting?

This paper answers this question in the affirmative. We present a simple momentum-based algorithm (PAGD for "perturbed AGD") that finds an ϵ-second order stationary point in $\tilde{O}(1/\epsilon^{7/4})$ iterations, faster than the $\tilde{O}(1/\epsilon^2)$ iterations required by GD. The pseudocode of our algorithm is presented in Algorithm 2.[1] PAGD adds two algorithmic features to AGD (Algorithm 1):

1. See Section 3 for values of various parameters.

Algorithm 1 Nesterov's Accelerated Gradient Descent $(\mathbf{x}_0, \eta, \theta)$

$\mathbf{v}_0 \leftarrow 0$
for $t = 0, 1, \ldots,$ **do**
  $\mathbf{y}_t \leftarrow \mathbf{x}_t + (1-\theta)\mathbf{v}_t$
  $\mathbf{x}_{t+1} \leftarrow \mathbf{y}_t - \eta \nabla f(\mathbf{y}_t)$
  $\mathbf{v}_{t+1} \leftarrow \mathbf{x}_{t+1} - \mathbf{x}_t$

Algorithm 2 Perturbed Accelerated Gradient Descent $(\mathbf{x}_0, \eta, \theta, \gamma, s, r, \mathscr{T})$

1: $\mathbf{v}_0 \leftarrow 0$
2: **for** $t = 0, 1, \ldots,$ **do**
3: **if** $\|\nabla f(\mathbf{x}_t)\| \le \epsilon$ and *no perturbation in last* $\mathscr{T}$ *steps* **then** } Perturbation
4: $\mathbf{x}_t \leftarrow \mathbf{x}_t + \xi_t \qquad \xi_t \sim \text{Unif}\left(\mathbb{B}_0(r)\right)$ } Perturbation
5: $\mathbf{y}_t \leftarrow \mathbf{x}_t + (1-\theta)\mathbf{v}_t$ } AGD
6: $\mathbf{x}_{t+1} \leftarrow \mathbf{y}_t - \eta \nabla f(\mathbf{y}_t)$ } AGD
7: $\mathbf{v}_{t+1} \leftarrow \mathbf{x}_{t+1} - \mathbf{x}_t$ } AGD
8: **if** $f(\mathbf{x}_t) \le f(\mathbf{y}_t) + \langle \nabla f(\mathbf{y}_t), \mathbf{x}_t - \mathbf{y}_t \rangle - \frac{\gamma}{2}\|\mathbf{x}_t - \mathbf{y}_t\|^2$ **then** } Negative curvature exploitation
9: $(\mathbf{x}_{t+1}, \mathbf{v}_{t+1}) \leftarrow$ Negative-Curvature-Exploitation$(\mathbf{x}_t, \mathbf{v}_t, s)$ } Negative curvature exploitation

- Perturbation (Lines 3-4): when the gradient is small, we add a small perturbation sampled uniformly from a d-dimensional ball with radius r. The homogeneous nature of this perturbation mitigates our lack of knowledge of the curvature tensor at or near saddle points.
- Negative Curvature Exploitation (NCE, Lines 8-9; pseudocode in Algorithm 3): when the function is observed to have "a lot" of negative curvature along $\mathbf{y}_t$ to $\mathbf{x}_t$ direction, we simply move $\mathbf{x}$ along this direction based on current momentum $\mathbf{v}_t$, and then reset $\mathbf{v}$.

We note that both components are straightforward to implement and increase computation by a constant factor. The perturbation idea follows from Ge et al. (2015) and Jin et al. (2017), while NCE is inspired by Carmon et al. (2017). We note that the analysis of the NCE subroutine is rather straightforward (see Section 3 and Section 4.1 for more details); the main challenge of this paper is to understand the behavior of perturbation and Nesterov's AGD steps in the nonconvex setting. To the best of our knowledge, PAGD is the first instance of a single-loop algorithm (meaning that it does not require an inner loop of optimization of a surrogate function) that is provably faster than GD in a general nonconvex setting.

1.1. Related Work

In this section, we review related work from the perspective of both nonconvex optimization and momentum/acceleration. For clarity of presentation, when discussing rates, we focus on the dependence on the accuracy ϵ and the dimension d while assuming all other problem parameters are constant. Table 1 presents a comparison of the current work with previous work.

Convergence to first-order stationary points: Traditional analyses in this case assume only Lipschitz gradients (see Definition 1). Nesterov (1998) shows that GD finds an ϵ-first-order stationary point in $O(1/\epsilon^2)$ steps. Ghadimi and Lan (2016) guarantee that AGD also converges in $\widetilde{O}\left(1/\epsilon^2\right)$ steps. Under the additional assumption of Lipschitz Hessians (see Definition 5), Carmon

et al. (2017) propose NCE steps, and develop a new algorithm that converges in $\widetilde{O}\left(1/\epsilon^{7/4}\right)$ iterations. We emphasize that while the analysis of the NCE step is rather straightforward, the main challenge is to deal with the case when NCE is not triggered. Carmon et al. (2017) study a nested-loop algorithm, repeatedly adding proximal terms to the objective in the inner loop and solving the modified functions using AGD. When AGD is applied to these modified functions, they can be analyzed similarly to the standard analysis in the convex case. In contrast, the current work studies AGD in its original, single-loop form, in a general nonconvex setting, and develops a new analysis based on a Hamiltonian perspective.

Convergence to second-order stationary points: All results in this setting assume Lipschitz continuity of both gradient and Hessian. Classical approaches, such as cubic regularization (Nesterov and Polyak, 2006) and trust region algorithms (Curtis et al., 2014), require access to Hessians, and are known to find ϵ-second-order stationary points in $O(1/\epsilon^{1.5})$ steps. However, the requirement of these algorithms to form the Hessian makes them infeasible for high-dimensional problems. A second set of algorithms utilize only Hessian-vector products instead of the full Hessian. Such products can be approximated by differentiating the gradients at two very close points, and in many applications they can also be directly computed efficiently. Rates of $\widetilde{O}(1/\epsilon^{7/4})$ have been established for such algorithms with nested loops (Carmon et al., 2016; Agarwal et al., 2017; Royer and Wright, 2017). Finally, in the realm of purely gradient-based algorithms, Ge et al. (2015) present the first polynomial guarantees for a perturbed version of GD, and Jin et al. (2017) sharpen it to $\widetilde{O}(1/\epsilon^2)$. For the special case of quadratic functions, O'Neill and Wright (2017) analyze the behavior of AGD around critical points and show that it escapes saddle points faster than GD, but it does not address the setting of general nonconvex functions. Parallel to this work, Allen-Zhu and Li (2017); Xu et al. (2017) also propose gradient-based algorithms achieving $\widetilde{O}(1/\epsilon^{7/4})$ rate, but their algorithms are also based on the framework of nested loops and repeatedly adding proximal terms, similar to Carmon et al. (2016).

Acceleration: There is also a rich literature that aims to understand momentum methods; e.g., Allen-Zhu and Orecchia (2014) view AGD as a linear coupling of GD and mirror descent, Su et al. (2016) and Wibisono et al. (2016) view AGD as the discretization of a second-order differential equation, and Bubeck et al. (2015) view AGD from a geometric perspective. Most of this work is tailored to the convex setting, and it is unclear and nontrivial to generalize the results to a nonconvex setting. There are also several papers that study AGD with relaxed versions of convexity—see Necoara et al. (2015); Li and Lin (2017) and references therein for overviews of these results.

1.2. Main Techniques

Our results rely on the following three key ideas. To the best of our knowledge, the first two are novel, while the third one was delineated in Jin et al. (2017).

Hamiltonian: A major challenge in analyzing momentum-based algorithms is that the objective function does not decrease monotonically as is the case for GD. To overcome this in the convex setting, several Lyapunov functions have been proposed (Wilson et al., 2016). However these Lyapunov functions involve the global minimum $\mathbf{x}^\star$, which cannot be computed by the algorithm, and is thus of limited value in the nonconvex setting. A key technical contribution of this paper is the design of a function which is both computable and tracks the progress of AGD. The function takes the form of a Hamiltonian:

$$E_t := f(\mathbf{x}_t) + \frac{1}{2\eta}\left\|\mathbf{v}_t\right\|^2; \tag{1}$$

Guarantees	Simplicity	Algorithm	Iterations
First-order Stationary Point	Nested-loop	Carmon et al. (2017)	$\widetilde{O}\left(1/\epsilon^{7/4}\right)$
	Single-loop	GD (Nesterov, 1998)	$O(1/\epsilon^2)$
		AGD (Ghadimi and Lan, 2016)	$O(1/\epsilon^2)$
Second-order Stationary Point	Nested-loop	Carmon et al. (2016)	$\widetilde{O}(1/\epsilon^{7/4})$
		Agarwal et al. (2017)	$\widetilde{O}(1/\epsilon^{7/4})$
	Single-loop	Noisy GD (Ge et al., 2015)	$O(\text{poly}(d/\epsilon))$
		Perturbed GD (Jin et al., 2017)	$\widetilde{O}(1/\epsilon^2)$
		Perturbed AGD [This Work]	$\widetilde{O}(1/\epsilon^{7/4})$

Table 1: Complexity of finding stationary points. $\widetilde{O}(\cdot)$ ignores polylog factors in d and ϵ.

i.e., a sum of potential energy and kinetic energy terms. It is monotonically decreasing in the continuous-time setting *regardless of the convexity of* $f(\cdot)$. This is *not* the case in general in the discrete-time setting, a fact which requires us to incorporate the NCE step—see Section 4.1 for more details. We note that monotonic decrease of the Hamiltonian, by itself, does not give any convergence rate, which brings us to our second key technical contribution.

Improve or localize: This paper formalizes a simple but powerful framework for analyzing long-term behavior of nonconvex optimization algorithms. This framework requires us to show that for a given algorithm, *either the algorithm makes significant progress or the iterates do not move much*. We call this the *improve-or-localize* phenomenon. For instance, when progress is measured by function value, it is easy to show that for GD, with proper choice of learning rate, we have:

$$\frac{1}{2\eta}\sum_{\tau=0}^{t-1}\|\mathbf{x}_{\tau+1}-\mathbf{x}_\tau\|^2 \le f(\mathbf{x}_0)-f(\mathbf{x}_t).$$

For AGD, a similar lemma can be shown by replacing the objective function with the Hamiltonian (see Lemma 9). Once this phenomenon is established, we can conclude that if an algorithm does not make much progress, it is localized to a small ball, and we can then approximate the objective function by either a linear or a quadratic function (depending on smoothness assumptions) in this small local region. Moreover, an upper bound on $\sum_{\tau=0}^{t-1}\|\mathbf{x}_{\tau+1}-\mathbf{x}_\tau\|^2$ lets us conclude that iterates do not oscillate much in this local region (oscillation is a unique phenomenon of momentum algorithms as can be seen even in the convex case). This gives us better control of approximation error.

Coupling sequences for escaping saddle points: When an algorithm arrives in the neighborhood of a strict saddle point, where $\lambda_{\min}(\nabla^2 f(\mathbf{x})) < 0$, all we know is that there exists a direction of escape (the direction of the minimum eigenvector of $\nabla^2 f(\mathbf{x})$); denote it by $\mathbf{e}_{\text{esc}}$. To avoid such points, the algorithm randomly perturbs the current iterate uniformly in a small ball, and runs AGD starting from this point $\tilde{\mathbf{x}}_0$. As in Jin et al. (2017), we can divide this ball into a "stuck region," $\mathcal{X}_{\text{stuck}}$, starting from which AGD does not escape the saddle quickly, and its complement

from which AGD escapes quickly. In order to show quick escape from a saddle point, we must show that the volume of $\mathcal{X}_{\text{stuck}}$ is very small compared to that of the ball. Though $\mathcal{X}_{\text{stuck}}$ may be without an analytical form, one can control the rate of escape by studying two AGD sequences that start from two realizations of perturbation, $\tilde{\mathbf{x}}_0$ and $\tilde{\mathbf{x}}_0'$, which are separated along $\mathbf{e}_{\text{esc}}$ by a small distance r_0. In this case, at least one of the sequences escapes the saddle point quickly, which proves that the width of $\mathcal{X}_{\text{stuck}}$ along $\mathbf{e}_{\text{esc}}$ can not be greater than r_0, and hence $\mathcal{X}_{\text{stuck}}$ has small volume.

2. Preliminaries

In this section, we will review some well-known results on GD and AGD in the strongly convex setting, and existing results on convergence of GD to second-order stationary points.

2.1. Notation

Bold upper-case letters ($\mathbf{A}, \mathbf{B}$) denote matrices and bold lower-case letters ($\mathbf{x}, \mathbf{y}$) denote vectors. For vectors $\|\cdot\|$ denotes the ℓ_2-norm. For matrices, $\|\cdot\|$ denotes the spectral norm and $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue. For $f : \mathbb{R}^d \to \mathbb{R}$, $\nabla f(\cdot)$ and $\nabla^2 f(\cdot)$ denote its gradient and Hessian respectively, and $f^\star$ denotes its global minimum. We use $O(\cdot), \Theta(\cdot), \Omega(\cdot)$ to hide absolute constants, and $\tilde{O}(\cdot), \tilde{\Theta}(\cdot), \tilde{\Omega}(\cdot)$ to hide absolute constants and polylog factors for all problem parameters.

2.2. Convex Setting

To minimize a function $f(\cdot)$, GD performs the following sequence of steps:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t).$$

The suboptimality of GD and the improvement achieved by AGD can be clearly illustrated for the case of smooth and strongly convex functions.

Definition 1 *A differentiable function $f(\cdot)$ is ℓ-**smooth (or ℓ-gradient Lipschitz)** if:*

$$\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\| \le \ell \|\mathbf{x}_1 - \mathbf{x}_2\| \quad \forall\, \mathbf{x}_1, \mathbf{x}_2.$$

The gradient Lipschitz property asserts that the gradient can not change too rapidly in a small local region.

Definition 2 *A twice-differentiable function $f(\cdot)$ is α-**strongly convex** if $\lambda_{\min}(\nabla^2 f(\mathbf{x})) \ge \alpha,\ \forall\, \mathbf{x}$.*

Let $f^* := \min_{\mathbf{y}} f(\mathbf{y})$. A point $\mathbf{x}$ is said to be ϵ-**suboptimal** if $f(\mathbf{x}) \le f^* + \epsilon$. The following theorem gives the convergence rate of GD and AGD for smooth and strongly convex functions.

Theorem 3 (Nesterov (2004)) *Assume that the function $f(\cdot)$ is ℓ-smooth and α-strongly convex. Then, for any $\epsilon > 0$, the iteration complexities to find an ϵ-suboptimal point are as follows:*

- *GD with $\eta = 1/\ell$:* $\quad O((\ell/\alpha) \cdot \log((f(\mathbf{x}_0) - f^*)/\epsilon))$
- *AGD (Algorithm 1) with $\eta = 1/\ell$ and $\theta = \sqrt{\alpha/\ell}$:* $\quad O(\sqrt{\ell/\alpha} \cdot \log((f(\mathbf{x}_0) - f^*)/\epsilon))$.

The number of iterations of GD depends linearly on the ratio ℓ/α, which is called the condition number of $f(\cdot)$ since $\alpha\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq \ell\mathbf{I}$. Clearly $\ell \ge \alpha$ and hence condition number is always at least one. Denoting the condition number by κ, we highlight two important aspects of AGD: (1) the momentum parameter satisfies $\theta = 1/\sqrt{\kappa}$ and (2) AGD improves upon GD by a factor of $\sqrt{\kappa}$.

Algorithm 3 Negative Curvature Exploitation$(\mathbf{x}_t, \mathbf{v}_t, s)$

if $\|\mathbf{v}_t\| \geq s$ **then**
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t$;
else
  $\delta = s \cdot \mathbf{v}_t / \|\mathbf{v}_t\|$
  $\mathbf{x}_{t+1} \leftarrow \operatorname{argmin}_{\mathbf{x} \in \{\mathbf{x}_t + \delta, \mathbf{x}_t - \delta\}} f(\mathbf{x})$
return $(\mathbf{x}_{t+1}, 0)$

2.3. Nonconvex Setting

For nonconvex functions finding global minima is NP-hard in the worst case. The best one can hope for in this setting is convergence to stationary points. There are various levels of stationarity.

Definition 4 $\mathbf{x}$ *is an* ϵ***-first-order stationary point*** *of function* $f(\cdot)$ *if* $\|\nabla f(\mathbf{x})\| \leq \epsilon$.

As mentioned in Section 1, for most nonconvex problems encountered in practice, a majority of first-order stationary points turn out to be saddle points. Second-order stationary points require not only zero gradient, but also positive semidefinite Hessian, ruling out most saddle points. Second-order stationary points are meaningful, however, only when the Hessian is continuous.

Definition 5 *A twice-differentiable function* $f(\cdot)$ *is* ρ***-Hessian Lipschitz*** *if:*

$$\left\|\nabla^2 f(\mathbf{x}_1) - \nabla^2 f(\mathbf{x}_2)\right\| \leq \rho \|\mathbf{x}_1 - \mathbf{x}_2\| \quad \forall\, \mathbf{x}_1, \mathbf{x}_2.$$

Definition 6 (Nesterov and Polyak (2006)) *For a* ρ*-Hessian Lipschitz function* $f(\cdot)$, $\mathbf{x}$ *is an* ϵ***-second-order stationary point*** *if:*

$$\|\nabla f(\mathbf{x})\| \leq \epsilon \quad \textit{and} \quad \lambda_{\min}(\nabla^2 f(\mathbf{x})) \geq -\sqrt{\rho\epsilon}.$$

The following theorem gives convergence rate of perturbed GD to second-order stationary points.

Theorem 7 ((Jin et al., 2017)) *Assume that the function* $f(\cdot)$ *is* ℓ*-smooth and* ρ*-Hessian Lipschitz. Then, for any* $\epsilon > 0$*, perturbed GD outputs an* ϵ*-second-order stationary point w.h.p in iterations:*

$$\widetilde{O}\left(\frac{\ell(f(\mathbf{x}_0) - f^*)}{\epsilon^2}\right).$$

Note that this rate is essentially the same as that of GD for convergence to first-order stationary points. In particular, it only has polylogarithmic dependence on the dimension.

3. Main Result

In this section, we present our main result providing a convergence rate of PAGD(Algorithm 2). As mentioned in Section 1, PAGD is essentially AGD with two key differences: perturbation and NCE. Perturbation is added (to escape saddle points) when the gradient is small, and no more frequently than once in $\mathscr{T}$ steps. The perturbation ξ_t is sampled uniformly from a d-dimensional ball with radius r. The specific choices of gap and uniform distribution are for technical convenience (they are sufficient for our theoretical result but not necessary).

NCE (Algorithm 3) is explicitly designed to guarantee decrease of the Hamiltonian (1), whenever AGD steps are not guaranteed to do so. In particular, AGD steps might not decrease the Hamiltonian when

$$f(\mathbf{x}_t) \le f(\mathbf{y}_t) + \langle \nabla f(\mathbf{y}_t), \mathbf{x}_t - \mathbf{y}_t \rangle - \frac{\gamma}{2} \left\| \mathbf{x}_t - \mathbf{y}_t \right\|^2, \tag{2}$$

i.e., the function has a large negative curvature between the current iterates $\mathbf{x}_t$ and $\mathbf{y}_t$. In this case, NCE is triggered. NCE does the following: if the momentum $\mathbf{v}_t$ is small, then $\mathbf{y}_t$ and $\mathbf{x}_t$ are close, so the large negative curvature also carries over to the Hessian at $\mathbf{x}_t$ due to the Lipschitz Hessian property. Since one of the directions $\pm(\mathbf{y}_t - \mathbf{x}_t)$ is negatively aligned with $\nabla f(\mathbf{x}_t)$, moving from $\mathbf{x}_t$ along this direction decreases function value and Hamiltonian. If the momentum $\mathbf{v}_t$ is large, negative curvature can no longer be exploited, but resetting the momentum to zero kills the second term in (1), significantly decreasing the Hamiltonian.

Setting of hyperparameters: Let ϵ be the target accuracy for a second-order stationary point, let ℓ and ρ be gradient/Hessian-Lipschitz parameters, and let c, χ be absolute constant and log factor to be specified later. Let $\kappa := \ell/\sqrt{\rho\epsilon}$, and set

$$\eta = \frac{1}{4\ell}, \quad \theta = \frac{1}{4\sqrt{\kappa}}, \quad \gamma = \frac{\theta^2}{\eta}, \quad s = \frac{\gamma}{4\rho}, \quad \mathscr{T} = \sqrt{\kappa} \cdot \chi c, \quad r = \eta\epsilon \cdot \chi^{-5} c^{-8}. \tag{3}$$

The following theorem is the main result of this paper.

Theorem 8 *Assume that the function $f(\cdot)$ is ℓ-smooth and ρ-Hessian Lipschitz. There exists an absolute constant $c_{\max}$ such that for any $\delta > 0$, $\epsilon \le \frac{\ell^2}{\rho}$, $\Delta_f \ge f(\mathbf{x}_0) - f^\star$, if $\chi = \max\{1, \log \frac{d\ell\Delta_f}{\rho\epsilon\delta}\}$, $c \ge c_{\max}$, if we run PAGD (Algorithm 2) with choice of parameters according to (3), then with probability at least $1 - \delta$, one of the iterates $\mathbf{x}_t$ will be an ϵ-second order stationary point in the following number of iterations:*

$$O\left(\frac{\ell^{1/2}\rho^{1/4}(f(\mathbf{x}_0) - f^*)}{\epsilon^{7/4}} \log^6\left(\frac{d\ell\Delta_f}{\rho\epsilon\delta} \right) \right).$$

Theorem 8 says that when PAGD is run for the designated number of steps (a number which is polylogarithmic in dimension[2]), at least one of the iterates is an ϵ-second-order stationary point. We focus on the case of small ϵ (i.e., $\epsilon \le \ell^2/\rho$) so that the Hessian requirement for the ϵ-second-order stationary point ($\lambda_{\min}(\nabla^2 f(\mathbf{x})) \ge -\sqrt{\rho\epsilon}$) is nontrivial. Note that $\left\|\nabla^2 f(\mathbf{x})\right\| \le \ell$ implies $\kappa = \ell/\sqrt{\rho\epsilon}$, which can be viewed as a condition number, akin to that in convex setting. Comparing Theorem 8 with Theorem 7, PAGD, with a momentum parameter $\theta = \Theta(1/\sqrt{\kappa})$, achieves $\tilde{\Theta}(\sqrt{\kappa})$ better iteration complexity compared to PGD.

Output ϵ-second order stationary point: Although Theorem 8 only guarantees that one of the iterates is an ϵ-second order stationary point, it is straightforward to identify one of them by adding a proper termination condition: once the gradient is small and satisfies the pre-condition to add a perturbation, we can keep track of the point $\mathbf{x}_{t_0}$ prior to adding perturbation, and compare the Hamiltonian at t_0 with the one $\mathscr{T}$ steps after. If the Hamiltonian decreases by $\mathscr{E} = \tilde{\Theta}(\sqrt{\epsilon^3/\rho})$, then the algorithm has made progress, otherwise $\mathbf{x}_{t_0}$ is an ϵ-second-order stationary point according to Lemma 13. Doing so will add a hyperparameter (threshold $\mathscr{E}$) but does not increase complexity.

2. This logarithmic dimension dependency can be removed if the primary target is to only find ϵ-first order stationary point, in which case the perturbation component (Lines 3-4) in Algorithm 2 need not be executed (since an ϵ-first order stationary point has been found), resulting in a completely deterministic algorithm.

4. Overview of Analysis

In this section, we will present an overview of the proof of Theorem 8. Section 4.1 presents the Hamiltonian for AGD and its key property of monotonic decrease. Section 4.2 presents *improve-or-localize* lemma, as well as the main intuition behind acceleration. Section 4.3 demonstrates how to apply these tools to prove Theorem 8. Complete details can be found in the appendix.

4.1. Hamiltonian

While GD guarantees decrease of function value in every step (even for nonconvex problems), the biggest stumbling block to analyzing AGD is that it is less clear how to keep track of "progress." Known Lyapunov functions for AGD (Wilson et al., 2016) are restricted to the convex setting and furthermore are not computable by the algorithm (as they depend on $\mathbf{x}^\star$).

To deepen the understanding of AGD in a nonconvex setting, we inspect it from a dynamical systems perspective, where we fix the ratio $\tilde{\theta} = \theta/\sqrt{\eta}$ to be a constant, while letting $\eta \to 0$. This leads to an ODE which is the continuous limit of AGD (Su et al., 2016):

$$\ddot{\mathbf{x}} + \tilde{\theta}\dot{\mathbf{x}} + \nabla f(\mathbf{x}) = 0, \tag{4}$$

where $\ddot{\mathbf{x}}$ and $\dot{\mathbf{x}}$ are derivatives with respect to time t. This equation is a second-order dynamical equation with *dissipative forces* $-\tilde{\theta}\dot{\mathbf{x}}$. Integrating both sides, we obtain:

$$f(\mathbf{x}(t_2)) + \frac{1}{2}\dot{\mathbf{x}}(t_2)^2 = f(\mathbf{x}(t_1)) + \frac{1}{2}\dot{\mathbf{x}}(t_1)^2 - \tilde{\theta}\int_{t_1}^{t_2} \dot{\mathbf{x}}(t)^2 \mathrm{d}t. \tag{5}$$

Using physical language, $f(\mathbf{x})$ is a *potential energy* while $\dot{\mathbf{x}}^2/2$ is a *kinetic energy*, and the sum is a *Hamiltonian*. The integral shows that the Hamiltonian decreases monotonically with time t, and the decrease is given by the *dissipation* term, $\tilde{\theta}\int_{t_1}^{t_2} \dot{\mathbf{x}}(t)^2 \mathrm{d}t$. Note that (5) holds regardless of the convexity of $f(\cdot)$. This monotonic decrease of the Hamiltonian can in fact be extended to the discretized version of AGD when the function is convex, or mildly nonconvex:

Lemma 9 (Hamiltonian decreases monotonically) *Assume that the function $f(\cdot)$ is ℓ-smooth, the learning rate $\eta \le \frac{1}{2\ell}$, and $\theta \in [2\eta\gamma, \frac{1}{2}]$ in AGD (Algorithm 1). Then, for every iteration t where (2) does not hold, we have:*

$$f(\mathbf{x}_{t+1}) + \frac{1}{2\eta}\left\|\mathbf{v}_{t+1}\right\|^2 \le f(\mathbf{x}_t) + \frac{1}{2\eta}\left\|\mathbf{v}_t\right\|^2 - \frac{\theta}{2\eta}\left\|\mathbf{v}_t\right\|^2 - \frac{\eta}{4}\left\|\nabla f(\mathbf{y}_t)\right\|^2. \tag{6}$$

Denote the discrete Hamiltonian as $E_t := f(\mathbf{x}_t) + \frac{1}{2\eta}\left\|\mathbf{v}_t\right\|^2$, and note that in AGD, $\mathbf{v}_t = \mathbf{x}_t - \mathbf{x}_{t-1}$. Lemma 9 tolerates nonconvexity with curvature at most $\gamma = \Theta(\theta/\eta)$. Unfortunately, when the function has large negative curvature between current iterates $\mathbf{x}_t$ and $\mathbf{y}_t$ (so that (2) holds), the analogy between the continuous and discretized versions breaks and (6) no longer holds. In fact, standard AGD can even increase the Hamiltonian in this regime (see Appendix A.1 for more details). However, we also note condition (2) is indeed an easy case, since it is in general challenging to efficiently find a negative curvature direction, but straightforward to exploit it when we observe one. This motivates us to modify the algorithm by adding the NCE step, which addresses this issue. We have the following simple result about NCE:

Lemma 10 *Assume that $f(\cdot)$ is ℓ-smooth and ρ-Hessian Lipschitz. For every iteration t of Algorithm 2 where (2) holds (thus running NCE), we have:*

$$E_{t+1} \le E_t - \min\{\frac{s^2}{2\eta}, \frac{1}{2}(\gamma - 2\rho s)s^2\}.$$

Lemmas 9 and 10 jointly assert that the Hamiltonian decreases monotonically in all situations, and are the main tools in the proof of Theorem 8. They not only give us a way of tracking progress, but also quantitatively measure the amount of progress.

4.2. Improve or Localize

One significant challenge in the analysis of gradient-based algorithms for nonconvex optimation is that many phenomena—such as accumulation of momentum and the escape from saddle points via perturbation—are multiple-step behaviors; they do not happen in a single step. We address this issue by developing a general technique for analyzing long-term behavior of such algorithms.

In our case, to track the long-term behavior of AGD, one key observation from Lemma 9 is that the amount of progress relates to movement of the iterates, which leads to the following *improve-or-localize* corollary:

Corollary 11 (Improve or localize) *Under the same setting as in Lemma 9, if (2) does not hold for all steps in $[t, t+T]$, we have:*

$$\sum_{\tau=t+1}^{t+T} \|\mathbf{x}_\tau - \mathbf{x}_{\tau-1}\|^2 \le \frac{2\eta}{\theta}(E_t - E_{t+T}).$$

Corollary 11 says that the algorithm either makes progress in terms of the Hamiltonian, or the iterates do not move much. In the second case, Corollary 11 allows us to approximate the dynamics of $\{\mathbf{x}_\tau\}_{\tau=t}^{t+T}$ with a *quadratic approximation* of $f(\cdot)$.

The acceleration phenomenon is rooted in and can be seen clearly for a quadratic, where the function can be decomposed into eigen-directions. Consider an eigen-direction with eigenvalue λ, and linear term g (i.e., in this direction $f(x) = \frac{\lambda}{2}x^2 + gx$). The GD update becomes $x_{\tau+1} = (1-\eta\lambda)x_\tau - \eta g$, with $\mu_{\text{GD}}(\lambda) := 1 - \eta\lambda$ determining the rate of GD. The update of AGD is $(x_{\tau+1}, x_\tau) = (x_\tau, x_{\tau-1})\mathbf{A}^\top - (\eta g, 0)$ with matrix $\mathbf{A}$ defined as follows:

$$\mathbf{A} := \begin{pmatrix} (2-\theta)(1-\eta\lambda) & -(1-\theta)(1-\eta\lambda) \\ 1 & 0 \end{pmatrix}.$$

The rate of AGD is determined by the largest eigenvalue of $\mathbf{A}$, denoted by $\mu_{\text{AGD}}(\lambda)$. Recall the choice of parameter (3), and divide the eigen-directions into the following three categories.

- **Strongly convex directions** $\lambda \in [\sqrt{\rho\epsilon}, \ell]$**:** the slowest case is $\lambda = \sqrt{\rho\epsilon}$, where $\mu_{\text{GD}}(\lambda) = 1 - \Theta(1/\kappa)$ while $\mu_{\text{AGD}}(\lambda) = 1 - \Theta(1/\sqrt{\kappa})$, which results in AGD converging faster than GD.

- **Flat directions** $\lambda \in [-\sqrt{\rho\epsilon}, \sqrt{\rho\epsilon}]$**:** the representative case is $\lambda = 0$ where AGD update becomes $x_{\tau+1} - x_\tau = (1-\theta)(x_\tau - x_{\tau-1}) - \eta g$. For $\tau \le 1/\theta$, we have $|x_{t+\tau} - x_t| = \Theta(\tau)$ for GD while $|x_{t+\tau} - x_t| = \Theta(\tau^2)$ for AGD, which results in AGD moving along negative gradient directions faster than GD.

- **Strongly nonconvex directions** $\lambda \in [-\ell, -\sqrt{\rho\epsilon}]$**:** similar to the strongly convex case, the slowest rate is for $\lambda = -\sqrt{\rho\epsilon}$ where $\mu_{\text{GD}}(\lambda) = 1 + \Theta(1/\kappa)$ while $\mu_{\text{AGD}}(\lambda) = 1 + \Theta(1/\sqrt{\kappa})$, which results in AGD escaping saddle point faster than GD.

Finally, the approximation error (from a quadratic) is also under control in this framework. With appropriate choice of T and threshold for $E_t - E_{t+T}$ in Corollary 11, by the Cauchy-Swartz inequality we can restrict iterates $\{\mathbf{x}_\tau\}_{\tau=t}^{t+T}$ to all lie within a local ball around $\mathbf{x}_t$ with radius $\sqrt{\epsilon/\rho}$, where both the gradient and Hessian of $f(\cdot)$ and its quadratic approximation $\tilde{f}_t(\mathbf{x}) = f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle + \frac{1}{2}(\mathbf{x} - \mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{x} - \mathbf{x}_t)$ are close:

Fact *Assume $f(\cdot)$ is ρ-Hessian Lipschitz, then for all $\mathbf{x}$ so that $\|\mathbf{x} - \mathbf{x}_t\| \le \sqrt{\epsilon/\rho}$, we have $\|\nabla f(\mathbf{x}) - \nabla \tilde{f}_t(\mathbf{x})\| \le \epsilon$ and $\|\nabla^2 f(\mathbf{x}) - \nabla^2 \tilde{f}_t(\mathbf{x})\| = \|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{x}_t)\| \le \sqrt{\rho\epsilon}$.*

4.3. Main Framework

For simplicity of presentation, recall $\mathscr{T} := \sqrt{\kappa} \cdot \chi c = \tilde{\Theta}(\sqrt{\kappa})$ and denote $\mathscr{E} := \sqrt{\epsilon^3/\rho} \cdot \chi^{-5} c^{-7} = \tilde{\Theta}(\sqrt{\epsilon^3/\rho})$, where c is sufficiently large constant as in Theorem 8. Our overall proof strategy will be to show the following "average descent claim": *Algorithm 2 decreases the Hamiltonian by $\mathscr{E}$ in every set of $\mathscr{T}$ iterations as long as it does not reach an ϵ-second-order stationary point*. Since the Hamiltonian cannot decrease more than $E_0 - E^\star = f(\mathbf{x}_0) - f^\star$, this immediately shows that it has to reach an ϵ-second-order stationary point in $O((f(\mathbf{x}_0) - f^\star)\mathscr{T}/\mathscr{E})$ steps, proving Theorem 8.

It can be verified by the choice of parameters (3) and Lemma 9 that whenever (2) holds so that NCE is triggered, the Hamiltonian decreases by at least $\mathscr{E}$ in one step. So, if NCE step is performed even once in each round of $\mathscr{T}$ steps, we achieve enough average decrease. The troublesome case is when in some time interval of $\mathscr{T}$ steps starting with $\mathbf{x}_t$, only AGD steps are performed without NCE. If $\mathbf{x}_t$ is not an ϵ-second order stationary point, either the gradient is large or the Hessian has a large negative direction. We prove the average decrease claim by considering these two cases.

Lemma 12 (Large gradient) *Consider the setting of Theorem 8. If $\|\nabla f(\mathbf{x}_\tau)\| \ge \epsilon$ for all $\tau \in [t, t + \mathscr{T}]$, then by running Algorithm 2 we have $E_{t+\mathscr{T}} - E_t \le -\mathscr{E}$.*

Lemma 13 (Negative curvature) *Consider the setting of Theorem 8. If $\|\nabla f(\mathbf{x}_t)\| \le \epsilon$, $\lambda_{\min}(\nabla^2 f(\mathbf{x}_t)) < -\sqrt{\rho\epsilon}$, and perturbation has not been added in iterations $\tau \in [t - \mathscr{T}, t)$, then by running Algorithm 2, we have $E_{t+\mathscr{T}} - E_t \le -\mathscr{E}$ with high probability.*

We note that an important aspect of these two lemmas is that the Hamiltonian decreases by $\Omega(\mathscr{E})$ in $\mathscr{T} = \tilde{\Theta}(\sqrt{\kappa})$ steps, which is faster compared to PGD which decreases the function value by $\Omega(\mathscr{E})$ in $\mathscr{T}^2 = \tilde{\Theta}(\kappa)$ steps (Jin et al., 2017). That is, the acceleration phenomenon in PAGD happens in both cases. We also stress that under both of these settings, PAGD cannot achieve $\Omega(\mathscr{E}/\mathscr{T})$ decrease in each step—it has to accumulate momentum over time to achieve $\Omega(\mathscr{E}/\mathscr{T})$ amortized decrease.

4.3.1. LARGE GRADIENT SCENARIO

For AGD, gradient and momentum interact, and both play important roles in the dynamics. Fortunately, according to Lemma 9, the Hamiltonian decreases sufficiently whenever the momentum $\mathbf{v}_t$ is large, so it is sufficient to discuss the case where the momentum is small.

One difficulty in proving Lemma 12 lies in enforcing the precondition that gradients of all iterates are large even with quadratic approximation. Intuitively we hope that the large initial gradient $\|\nabla f(\mathbf{x}_t)\| \ge \epsilon$ suffices to give a sufficient decrease of the Hamiltonian. Unfortunately, this is not true. Let $\mathcal{S}$ be the subspace of eigenvectors of $\nabla^2 f(\mathbf{x}_t)$ with eigenvalues in $[\sqrt{\rho\epsilon}, \ell]$, consisting of

all the strongly convex directions, and let $\mathcal{S}^c$ be the orthogonal subspace. It turns out that the initial gradient component in $\mathcal{S}$ is not very helpful in decreasing the Hamiltonian since AGD rapidly decreases the gradient in these directions. We instead prove Lemma 12 in two steps.

Lemma 14 *(informal) If $\mathbf{v}_t$ is small, $\|\nabla f(\mathbf{x}_t)\|$ not too large and $E_{t+\mathscr{T}/2} - E_t \geq -\mathscr{E}$, then for all $\tau \in [t + \mathscr{T}/4, t + \mathscr{T}/2]$ we have $\|\mathcal{P}_{\mathcal{S}} \nabla f(\mathbf{x}_\tau)\| \leq \epsilon/2$.*

Lemma 15 *(informal) If $\mathbf{v}_t$ is small and $\|\mathcal{P}_{\mathcal{S}^c} \nabla f(\mathbf{x}_t)\| \geq \epsilon/2$, then we have $E_{t+\mathscr{T}/4} - E_t \leq -\mathscr{E}$.*

See the formal versions, Lemma 21 and Lemma 22, for more details. We see that if the Hamiltonian does not decrease much (and so is localized in a small ball), the gradient in the strongly convex subspace $\|\mathcal{P}_{\mathcal{S}} \nabla f(\mathbf{x}_\tau)\|$ vanishes in $\mathscr{T}/4$ steps by Lemma 14. Since the hypothesis of Lemma 12 guarantees a large gradient for all of the $\mathscr{T}$ steps, this means that $\|\mathcal{P}_{\mathcal{S}^c} \nabla f(\mathbf{x}_t)\|$ is large after $\mathscr{T}/4$ steps, thereby decreasing the Hamiltonian in the next $\mathscr{T}/4$ steps (by Lemma 15).

4.3.2. NEGATIVE CURVATURE SCENARIO

In this section, we will show that the volume of the set around a strict saddle point from which AGD does not escape quickly is very small (Lemma 13). We do this using the coupling mechanism introduced in Jin et al. (2017), which gives a fine-grained understanding of the geometry around saddle points. More concretely, letting the perturbation radius $r = \tilde{\Theta}(\epsilon/\ell)$ as specified in (3), we show the following lemma.

Lemma 16 *(informal) Suppose $\|\nabla f(\tilde{\mathbf{x}})\| \leq \epsilon$ and $\lambda_{\min}(\nabla^2 f(\tilde{\mathbf{x}})) \leq -\sqrt{\rho\epsilon}$. Let $\mathbf{x}_0, \mathbf{x}_0'$ be at distance at most r from $\tilde{\mathbf{x}}$, and $\mathbf{x}_0 - \mathbf{x}_0' = r_0 \mathbf{e}_1$ where $\mathbf{e}_1$ is the minimum eigen-direction of $\nabla^2 f(\tilde{\mathbf{x}})$ and $r_0 \geq \delta r/\sqrt{d}$. Then for AGD starting at $(\mathbf{x}_0, \mathbf{v})$ and $(\mathbf{x}_0', \mathbf{v})$, we have:*

$$\min\{E_{\mathscr{T}} - \widetilde{E}, E'_{\mathscr{T}} - \widetilde{E}\} \leq -\mathscr{E},$$

where $\widetilde{E}, E_{\mathscr{T}}$ and $E'_{\mathscr{T}}$ are the Hamiltonians at $(\tilde{\mathbf{x}}, \mathbf{v}), (\mathbf{x}_{\mathscr{T}}, \mathbf{v}_{\mathscr{T}})$ and $(\mathbf{x}'_{\mathscr{T}}, \mathbf{v}'_{\mathscr{T}})$ respectively.

See the formal version in Lemma 23. We note that δ in this lemma is a small number that characterizes the failure probability of the algorithm (as defined in Theorem 8), and $\mathscr{T}$ has logarithmic dependence on δ according to (3). Lemma 16 says that around any strict saddle, for any two points that are separated along the smallest eigen-direction by at least $\delta r/\sqrt{d}$, PAGD, starting from at least one of those points, decreases the Hamiltonian, and hence escapes the strict saddle. This implies that the width of the region starting from where AGD is stuck has width at most $\delta r/\sqrt{d}$, and thus has small volume.

5. Conclusions

In this paper, we show that a variant of AGD can escape saddle points faster than GD, demonstrating that momentum techniques can indeed accelerate convergence even for nonconvex optimization. Our algorithm finds an ϵ-second order stationary point in $\widetilde{O}\left(1/\epsilon^{7/4}\right)$ iterations, faster than the $\widetilde{O}\left(1/\epsilon^2\right)$ iterations taken by GD. This is the first single-loop algorithm that achieves this rate. Our analysis relies on novel techniques that lead to a better understanding of momentum techniques as well as nonconvex optimization.

References

Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1195–1199. ACM, 2017.

Zeyuan Allen-Zhu and Yuanzhi Li. Neon2: Finding local minima via first-order oracles. *arXiv preprint arXiv:1711.06673*, 2017.

Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

Afonso S Bandeira, Nicolas Boumal, and Vladislav Voroninski. On the low-rank approach for semidefinite programs arising in synchronization and community detection. In *Conference on Learning Theory*, pages 361–382, 2016.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira. The non-convex Burer-Monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pages 2757–2765, 2016.

Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to Nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for non-convex optimization. *arXiv preprint arXiv:1611.00756*, 2016.

Yair Carmon, Oliver Hinder, John C Duchi, and Aaron Sidford. Convex until Proven Guilty: Dimension-free acceleration of gradient descent on non-convex functions. *arXiv preprint arXiv:1705.02766*, 2017.

Coralia Cartis, Nicholas Gould, and Ph L Toint. On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems. *Siam journal on optimization*, 20(6):2833–2852, 2010.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surface of multilayer networks. *arXiv:1412.0233*, 2014.

Frank E Curtis, Daniel P Robinson, and Mohammadreza Samadi. A trust region algorithm with a worst-case iteration complexity of $o(\epsilon^{-3/2})$ for nonconvex optimization. *Mathematical Programming*, pages 1–32, 2014.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, pages 2933–2941, 2014.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Computational Learning Theory (COLT)*, 2015.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017.

Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning (ICML)*, 2017.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances In Neural Information Processing Systems*, pages 586–594, 2016.

Yin Tat Lee and Aaron Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. In *Foundations of Computer Science (FOCS)*, pages 147–156. IEEE, 2013.

Huan Li and Zhouchen Lin. Provable accelerated gradient method for nonconvex low rank optimization. *arXiv preprint arXiv:1702.04959*, 2017.

Song Mei, Theodor Misiakiewicz, Andrea Montanari, and Roberto I Oliveira. Solving SDPs for synchronization and maxcut problems via the Grothendieck inequality. In *Conference on Learning Theory (COLT)*, pages 1476–1515, 2017.

Ion Necoara, Yurii Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *arXiv preprint arXiv:1504.06298*, 2015.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate o (1/k2). *Soviet Mathematics Doklady*, 27:372–376, 1983.

Yurii Nesterov. *Introductory Lectures on Convex Programming Volume I: Basic course*. Springer, 1998.

Yurii Nesterov. *Introductory Lectures on Convex Optimization*, volume 87. Springer Science & Business Media, 2004.

Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Michael O'Neill and Stephen J Wright. Behavior of accelerated gradient methods near critical points of nonconvex problems. *arXiv preprint arXiv:1706.07993*, 2017.

Clément W Royer and Stephen J Wright. Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. *arXiv preprint arXiv:1706.03131*, 2017.

Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *Journal of Machine Learning Research*, 17 (153):1–43, 2016.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

A. Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 133:E7351–E7358, 2016.

Ashia C Wilson, Benjamin Recht, and Michael I Jordan. A Lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635*, 2016.

Yi Xu, Rong Jin, and Tianbao Yang. Neon+: Accelerated gradient methods for extracting negative curvature for non-convex optimization. *arXiv preprint arXiv:1712.01033*, 2017.

Appendix A. Proof of Hamiltonian Lemmas

In this section, we prove Lemma 9, Lemma 10 and Corollary 11, which are presented in Section 4.1 and Section 4.2. In section A.1 we also give an example where standard AGD with negative curvature exploitation can increase the Hamiltonian.

Recall that we define the Hamiltonian as $E_t := f(\mathbf{x}_t) + \frac{1}{2\eta}\|\mathbf{v}_t\|^2$, where, for AGD, we define $\mathbf{v}_t = \mathbf{x}_t - \mathbf{x}_{t-1}$. The first lemma shows that this Hamiltonian decreases in every step of AGD for mildly nonconvex functions.

Lemma 9 (Hamiltonian decreases monotonically) *Assume that the function $f(\cdot)$ is ℓ-smooth and set the learning rate to be $\eta \le \frac{1}{2\ell}$, $\theta \in [2\eta\gamma, \frac{1}{2}]$ in AGD (Algorithm 1). Then, for every iteration t where (2) does not hold, we have:*

$$E_{t+1} \le E_t - \frac{\theta}{2\eta}\|\mathbf{v}_t\|^2 - \frac{\eta}{4}\|\nabla f(\mathbf{y}_t)\|^2 .$$

Proof Recall that the update equation of accelerated gradient descent has following form:

$$\begin{aligned}\mathbf{x}_{t+1} &\leftarrow \mathbf{y}_t - \eta\nabla f(\mathbf{y}_t)\\ \mathbf{y}_{t+1} &\leftarrow \mathbf{x}_{t+1} + (1-\theta)(\mathbf{x}_{t+1} - \mathbf{x}_t).\end{aligned}$$

By smoothness, with $\eta \le \frac{1}{2\ell}$:

$$f(\mathbf{x}_{t+1}) \le f(\mathbf{y}_t) - \eta\|\nabla f(\mathbf{y}_t)\|^2 + \frac{\ell\eta^2}{2}\|\nabla f(\mathbf{y}_t)\|^2 \le f(\mathbf{y}_t) - \frac{3\eta}{4}\|\nabla f(\mathbf{y}_t)\|^2 , \tag{7}$$

assuming that the precondition (2) does not hold:

$$f(\mathbf{x}_t) \ge f(\mathbf{y}_t) + \langle\nabla f(\mathbf{y}_t), \mathbf{x}_t - \mathbf{y}_t\rangle - \frac{\gamma}{2}\|\mathbf{y}_t - \mathbf{x}_t\|^2 , \tag{8}$$

and given the following update equation:

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 &= \|\mathbf{y}_t - \mathbf{x}_t - \eta\nabla f(\mathbf{y}_t)\|^2\\ &= \left[(1-\theta)^2\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 - 2\eta\langle\nabla f(\mathbf{y}_t), \mathbf{y}_t - \mathbf{x}_t\rangle + \eta^2\|\nabla f(\mathbf{y}_t)\|^2\right],\end{aligned} \tag{9}$$

we have:

$$\begin{aligned}f(\mathbf{x}_{t+1}) + \frac{1}{2\eta}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \le& f(\mathbf{x}_t) + \langle\nabla f(\mathbf{y}_t), \mathbf{y}_t - \mathbf{x}_t\rangle - \frac{3\eta}{4}\|\nabla f(\mathbf{y}_t)\|^2\\ &+ \frac{1+\eta\gamma}{2\eta}(1-\theta)^2\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 - \langle\nabla f(\mathbf{y}_t), \mathbf{y}_t - \mathbf{x}_t\rangle + \frac{\eta}{2}\|\nabla f(\mathbf{y}_t)\|^2\\ \le& f(\mathbf{x}_t) + \frac{1}{2\eta}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 - \frac{2\theta - \theta^2 - \eta\gamma(1-\theta)^2}{2\eta}\|\mathbf{v}_t\|^2 - \frac{\eta}{4}\|\nabla f(\mathbf{y}_t)\|^2\\ \le& f(\mathbf{x}_t) + \frac{1}{2\eta}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 - \frac{\theta}{2\eta}\|\mathbf{v}_t\|^2 - \frac{\eta}{4}\|\nabla f(\mathbf{y}_t)\|^2 .\end{aligned}$$

The last inequality uses the fact that $\theta \in [2\eta\gamma, \frac{1}{2}]$ so that $\theta^2 \le \frac{\theta}{2}$ and $\eta\gamma \le \frac{\theta}{2}$. We substitute in the definition of $\mathbf{v}_t$ and E_t to finish the proof. ■

We see from this proof that (8) relies on approximate convexity of $f(\cdot)$, which explains why in all existing proofs, the convexity between $\mathbf{x}_t$ and $\mathbf{y}_t$ is so important. A perhaps surprising fact to note is that the above proof can in fact go through even with mild nonconvexity (captured in line 8 of Algorithm 2). Thus, high nonconvexity is the problematic situation. To overcome this, we need to slightly modify AGD so that the Hamiltonian is decreasing. This is formalized in the following lemma.

Lemma 10 *Assume that $f(\cdot)$ is ℓ-smooth and ρ-Hessian Lipschitz. For every iteration t of Algorithm 2 where (2) holds (thus running NCE), we have:*

$$E_{t+1} \le E_t - \min\{\frac{s^2}{2\eta}, \frac{1}{2}(\gamma - 2\rho s)s^2\}.$$

Proof When we perform an NCE step, we know that (2) holds. In the first case ($\|\mathbf{v}_t\| \ge s$), we set $\mathbf{x}_{t+1} = \mathbf{x}_t$ and set the momentum $\mathbf{v}_{t+1}$ to zero, which gives:

$$E_{t+1} = f(\mathbf{x}_{t+1}) = f(\mathbf{x}_t) = E_t - \frac{1}{2\eta}\|\mathbf{v}_t\|^2 \le E_t - \frac{s^2}{2\eta}.$$

In the second case ($\|\mathbf{v}_t\| \le s$), expanding in a Taylor series with Lagrange remainder, we have:

$$f(\mathbf{x}_t) = f(\mathbf{y}_t) + \langle \nabla f(\mathbf{y}_t), \mathbf{x}_t - \mathbf{y}_t\rangle + \frac{1}{2}(\mathbf{x}_t - \mathbf{y}_t)^\top \nabla^2 f(\zeta_t)(\mathbf{x}_t - \mathbf{y}_t),$$

where $\zeta_t = \phi\mathbf{x}_t + (1-\phi)\mathbf{y}_t$ and $\phi \in [0,1]$. Due to the certificate (2) we have

$$\frac{1}{2}(\mathbf{x}_t - \mathbf{y}_t)^\top \nabla^2 f(\zeta_t)(\mathbf{x}_t - \mathbf{y}_t) \le -\frac{\gamma}{2}\|\mathbf{x}_t - \mathbf{y}_t\|^2.$$

On the other hand, clearly $\min\{\langle \nabla f(\mathbf{x}_t), \delta\rangle, \langle \nabla f(\mathbf{x}_t), -\delta\rangle\} \le 0$. WLOG, suppose $\langle \nabla f(\mathbf{x}_t), \delta\rangle \le 0$, then, by definition of $\mathbf{x}_{t+1}$, we have:

$$f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t + \delta) = f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \delta\rangle + \frac{1}{2}\delta^\top \nabla^2 f(\zeta_t')\delta \le f(\mathbf{x}_t) + \frac{1}{2}\delta^\top \nabla^2 f(\zeta_t')\delta,$$

where $\zeta_t' = \mathbf{x}_t + \phi'\delta$ and $\phi' \in [0,1]$. Since $\|\zeta_t - \zeta_t'\| \le 2s$, δ also lines up with $\mathbf{y}_t - \mathbf{x}_t$:

$$\delta^\top \nabla^2 f(\zeta_t')\delta \le \delta^\top \nabla^2 f(\zeta_t)\delta + \left\|\nabla^2 f(\zeta_t') - \nabla^2 f(\zeta_t)\right\|\|\delta\|^2 \le -\gamma\|\delta\|^2 + 2\rho s\|\delta\|^2.$$

Therefore, this gives

$$E_{t+1} = f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \frac{1}{2}(\gamma - \rho s)s^2 \le E_t - \frac{1}{2}(\gamma - 2\rho s)s^2,$$

which finishes the proof. ■

The Hamiltonian decrease has an important consequence: if the Hamiltonian does not decrease much, then all the iterates are localized in a small ball around the starting point. Moreover, the iterates do not oscillate much in this ball. We called this the improve-or-localize phenomenon.

Corollary 11 (Improve or localize) *Under the same setting as in Lemma 9, if (2) does not hold for all steps in $[t, t+T]$, we have:*

$$\sum_{\tau=t+1}^{t+T}\|\mathbf{x}_\tau - \mathbf{x}_{\tau-1}\|^2 \le \frac{2\eta}{\theta}(E_t - E_{t+T}).$$

Proof The proof follows immediately from telescoping the argument of Lemma 9. ■

A.1. AGD can increase the Hamiltonian under nonconvexity

In the previous section, we proved Lemma 9 which requires $\theta \geq 2\eta\gamma$, that is, $\gamma \leq \theta/(2\eta)$. In this section, we show Lemma 9 is almost tight in the sense that when $\gamma \geq 4\theta/\eta$ in (2), we have:

$$f(\mathbf{x}_t) \leq f(\mathbf{y}_t) + \langle \nabla f(\mathbf{y}_t), \mathbf{x}_t - \mathbf{y}_t \rangle - \frac{\gamma}{2} \|\mathbf{x}_t - \mathbf{y}_t\|^2 .$$

Monotonic decrease of the Hamiltonian may no longer hold, indeed, AGD can increase the Hamiltonian for those steps.

Consider a simple one-dimensional example, $f(x) = -\frac{1}{2}\gamma x^2$, where (2) always holds. Define the initial condition $x_0 = -1, v_0 = 1/(1-\theta)$. By update equation in Algorithm 1, the next iterate will be $x_1 = y_0 = 0$, and $v_1 = x_1 - x_0 = 1$. By the definition of Hamiltonian, we have

$$\begin{aligned} E_0 =& f(x_0) + \frac{1}{2\eta}|v_0|^2 = -\frac{\gamma}{2} + \frac{1}{2\eta(1-\theta)^2} \\ E_1 =& f(x_1) + \frac{1}{2\eta}|v_1|^2 = \frac{1}{2\eta}, \end{aligned}$$

since $\theta \leq 1/4$. It is not hard to verify that whenever $\gamma \geq 4\theta/\eta$, we will have $E_1 \geq E_0$; that is, the Hamiltonian increases in this step.

This fact implies that when we pick a large learning rate η and small momentum parameter θ (both are essential for acceleration), standard AGD does not decrease the Hamiltonian in a very nonconvex region. We need another mechanism such as NCE to fix the monotonically decreasing property.

Appendix B. Proof of Main Result

In this section, we set up the machinery needed to prove our main result, Theorem 8. We first present the generic setup, then, as in Section 4.3, we split the proof into two cases, one where gradient is large and the other where the Hessian has negative curvature. In the end, we put everything together and prove Theorem 8.

To simplify the proof, we introduce some notation for this section, and state a convention regarding absolute constants. Recall the choice of parameters in Eq.(3):

$$\eta = \frac{1}{4\ell}, \quad \theta = \frac{1}{4\sqrt{\kappa}}, \quad \gamma = \frac{\theta^2}{\eta} = \frac{\sqrt{\rho\epsilon}}{4}, \quad s = \frac{\gamma}{4\rho} = \frac{1}{16}\sqrt{\frac{\epsilon}{\rho}}, \quad r = \eta\epsilon \cdot \chi^{-5} c^{-8},$$

where $\kappa = \frac{\ell}{\sqrt{\rho\epsilon}}, \chi = \max\{1, \log \frac{d\ell\Delta_f}{\rho\epsilon\delta}\}$, and c is a sufficiently large constant as stated in the precondition of Theorem 8. Throughout this section, we also always denote

$$\mathscr{T} := \sqrt{\kappa} \cdot \chi c, \quad \mathscr{E} := \sqrt{\frac{\epsilon^3}{\rho}} \cdot \chi^{-5} c^{-7}, \quad \mathscr{S} := \sqrt{\frac{2\eta\mathscr{T}\mathscr{E}}{\theta}} = \sqrt{\frac{2\epsilon}{\rho}} \cdot \chi^{-2} c^{-3}, \quad \mathscr{M} := \frac{\epsilon\sqrt{\kappa}}{\ell} c^{-1},$$

which represent the special units for time, the Hamiltonian, the parameter space and the momentum. All the lemmas in this section hold when the constant c is picked to be sufficiently large. To avoid ambiguity, throughout this section $O(\cdot), \Omega(\cdot), \Theta(\cdot)$ notation **only hides an absolute constant which is independent of the choice of sufficiently large constant** c, which is defined in the precondition of Theorem 8. That is, we will always make c dependence explicit in $O(\cdot), \Omega(\cdot), \Theta(\cdot)$ notation. Therefore, for a quantity like $O(c^{-1})$, we can always pick c large enough so that it cancels out the absolute constant in the $O(\cdot)$ notation, and make $O(c^{-1})$ smaller than any fixed required constant.

B.1. Common setup

Our general strategy in the proof is to show that if none of the iterates $\mathbf{x}_t$ is a SOSP, then in all $\mathscr{T}$ steps, the Hamiltonian always decreases by at least $\mathscr{E}$. This gives an average decrease of $\mathscr{E}/\mathscr{T}$. In this section, we establish some facts which will be used throughout the entire proof, including the decrease of the Hamiltonian in NCE step, the update of AGD in matrix form, and upper bounds on approximation error for a local quadratic approximation.

The first lemma shows if negative curvature exploitation is used, then in a single step, the Hamiltonian will decrease by $\mathscr{E}$.

Lemma 17 *Under the same setting as Theorem 8, for every iteration t of Algorithm 2 where (2) holds (thus running NCE), we have:*

$$E_{t+1} - E_t \le -2\mathscr{E}.$$

Proof It is also easy to check that the precondition of Lemma 10 holds, and by the particular choice of parameters in Theorem 8, we have:

$$\min\{\frac{s^2}{2\eta}, \frac{1}{2}(\gamma - 2\rho s)s^2\} \ge \Omega(\mathscr{E}c^7) \ge 2\mathscr{E},$$

where the last inequality is by picking c in Theorem 8 large enough, which finishes the proof. ■

Therefore, whenever NCE is called, the decrease of the Hamiltonian is already sufficient. We thus only need to focus on AGD steps. The next lemma derives a general expression for $\mathbf{x}_t$ after an AGD update, which is very useful in multiple-step analysis. The general form is expressed with respect to a reference point $\mathbf{0}$, which can be any arbitrary point (in many cases we choose it to be $\mathbf{x}_0$).

Lemma 18 *Let $\mathbf{0}$ be an origin (which can be fixed at an arbitrary point). Let $\mathcal{H} = \nabla^2 f(\mathbf{0})$. Then an AGD (Algorithm 1) update can be written as:*

$$\begin{pmatrix} \mathbf{x}_{t+1} \\ \mathbf{x}_t \end{pmatrix} = \mathbf{A}^t \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_0 \end{pmatrix} - \eta \sum_{\tau=1}^{t} \mathbf{A}^{t-\tau} \begin{pmatrix} \nabla f(\mathbf{0}) + \delta_\tau \\ 0 \end{pmatrix}, \tag{10}$$

where $\delta_\tau = \nabla f(\mathbf{y}_\tau) - \nabla f(\mathbf{0}) - \mathcal{H}\mathbf{y}_\tau$, and

$$\mathbf{A} = \begin{pmatrix} (2-\theta)(\mathbf{I} - \eta\mathcal{H}) & -(1-\theta)(\mathbf{I} - \eta\mathcal{H}) \\ \mathbf{I} & 0 \end{pmatrix}.$$

Proof Substituting for $(\mathbf{y}_t, \mathbf{v}_t)$ in Algorithm 1, we have a recursive equation for $\mathbf{x}_t$:

$$\mathbf{x}_{t+1} = (2-\theta)\mathbf{x}_t - (1-\theta)\mathbf{x}_{t-1} - \eta\nabla f((2-\theta)\mathbf{x}_t - (1-\theta)\mathbf{x}_{t-1}). \tag{11}$$

By definition of δ_τ, we also have:

$$\nabla f(\mathbf{y}_\tau) = \nabla f(\mathbf{0}) + \mathcal{H}\mathbf{y}_\tau + \delta_\tau.$$

Therefore, in matrix form, we have:

$$\begin{aligned}\begin{pmatrix}\mathbf{x}_{t+1}\\ \mathbf{x}_t\end{pmatrix} =& \begin{pmatrix}(2-\theta)(\mathbf{I}-\eta\mathcal{H}) & -(1-\theta)(\mathbf{I}-\eta\mathcal{H})\\ \mathbf{I} & 0\end{pmatrix}\begin{pmatrix}\mathbf{x}_t\\ \mathbf{x}_{t-1}\end{pmatrix} - \eta\begin{pmatrix}\nabla f(0)+\delta_t\\ 0\end{pmatrix}\\ =& \mathbf{A}^t\begin{pmatrix}\mathbf{x}_1\\ \mathbf{x}_0\end{pmatrix} - \eta\sum_{\tau=1}^{t}\mathbf{A}^{t-\tau}\begin{pmatrix}\nabla f(0)+\delta_\tau\\ 0\end{pmatrix},\end{aligned}$$

which finishes the proof. ■

Clearly $\mathbf{A}$ in Lemma 18 is a $2d\times 2d$ matrix, and if we expand $\mathbf{A}$ according to the eigenvector directions of $\begin{pmatrix}\mathcal{H} & 0\\ 0 & \mathcal{H}\end{pmatrix}$, $\mathbf{A}$ can be reorganized as a block-diagonal matrix consisting of d 2×2 matrices. Let the jth eigenvalue of $\mathcal{H}$ be denoted λ_j, and denote $\mathbf{A}_j$ as the jth 2×2 matrix with corresponding eigendirections:

$$\mathbf{A}_j = \begin{pmatrix}(2-\theta)(1-\eta\lambda_j) & -(1-\theta)(1-\eta\lambda_j)\\ 1 & 0\end{pmatrix}. \tag{12}$$

We note that the choice of reference point $\mathbf{0}$ is mainly to simplify mathmatical expressions involving $\mathbf{x}_t-\mathbf{0}$.

Lemma 18 can be viewed as update from a quadratic expansion around origin $\mathbf{0}$, and δ_τ is the approximation error which marks the difference between true function and its quadratic approximation. The next lemma shows that when sequence $\mathbf{x}_0,\cdots,\mathbf{x}_t$ are all close to $\mathbf{0}$, then the approximation error is under control:

Proposition 19 *Using the notation of Lemma 18, if for any $\tau\le t$, we have $\|\mathbf{x}_\tau\|\le R$, then for any $\tau\le t$, we also have*

1. $\|\delta_\tau\|\le O(\rho R^2)$;
2. $\|\delta_\tau-\delta_{\tau-1}\|\le O(\rho R)(\|\mathbf{x}_t-\mathbf{x}_{\tau-1}\|+\|\mathbf{x}_{\tau-1}-\mathbf{x}_{\tau-2}\|)$;
3. $\sum_{\tau=1}^t\|\delta_\tau-\delta_{\tau-1}\|^2\le O(\rho^2R^2)\sum_{\tau=1}^t\|\mathbf{x}_\tau-\mathbf{x}_{\tau-1}\|^2$.

Proof Let $\Delta_\tau=\int_0^1(\nabla^2 f(\phi\mathbf{y}_\tau)-\mathcal{H})\mathrm{d}\phi$. The first inequality is true because $\delta_\tau=\Delta_\tau\mathbf{y}_\tau$, thus:

$$\begin{aligned}\|\delta_\tau\| =& \|\Delta_\tau\mathbf{y}_\tau\|\le\|\Delta_\tau\|\|\mathbf{y}_\tau\| = \left\|\int_0^1(\nabla^2 f(\phi\mathbf{y}_\tau)-\mathcal{H})\mathrm{d}\phi\right\|\|\mathbf{y}_\tau\|\\ \le& \int_0^1\left\|(\nabla^2 f(\phi\mathbf{y}_\tau)-\mathcal{H})\right\|\mathrm{d}\phi\cdot\|\mathbf{y}_\tau\|\le\rho\|\mathbf{y}_\tau\|^2\le\rho\|(2-\theta)\mathbf{x}_\tau-(1-\theta)\mathbf{x}_{\tau-1}\|^2\le O(\rho R^2).\end{aligned}$$

For the second inequality, we have:

$$\delta_\tau-\delta_{\tau-1}=\nabla f(\mathbf{y}_\tau)-\nabla f(\mathbf{y}_{\tau-1})-\mathcal{H}(\mathbf{y}_\tau-\mathbf{y}_{\tau-1})=\Delta'_\tau(\mathbf{y}_\tau-\mathbf{y}_{\tau-1}),$$

where $\Delta'_\tau=\int_0^1(\nabla^2 f(\mathbf{y}_{\tau-1}+\phi(\mathbf{y}_\tau-\mathbf{y}_{\tau-1}))-\mathcal{H})\mathrm{d}\phi$. As in the proof of the first inequality, we have:

$$\begin{aligned}\|\delta_\tau-\delta_{\tau-1}\|\le\left\|\Delta'_\tau\right\|\|\mathbf{y}_\tau-\mathbf{y}_{\tau-1}\| =& \left\|\int_0^1(\nabla^2 f(\mathbf{y}_{\tau-1}+\phi(\mathbf{y}_\tau-\mathbf{y}_{\tau-1}))-\mathcal{H})\mathrm{d}\phi\right\|\|\mathbf{y}_\tau-\mathbf{y}_{\tau-1}\|\\ \le&\rho\max\{\|\mathbf{y}_\tau\|,\|\mathbf{y}_{\tau-1}\|\}\|\mathbf{y}_\tau-\mathbf{y}_{\tau-1}\|\le O(\rho R)(\|\mathbf{x}_\tau-\mathbf{x}_{\tau-1}\|+\|\mathbf{x}_{\tau-1}-\mathbf{x}_{\tau-2}\|).\end{aligned}$$

Finally, since $(\|\mathbf{x}_\tau - \mathbf{x}_{\tau-1}\| + \|\mathbf{x}_{\tau-1} - \mathbf{x}_{\tau-2}\|)^2 \le 2(\|\mathbf{x}_\tau - \mathbf{x}_{\tau-1}\|^2 + \|\mathbf{x}_{\tau-1} - \mathbf{x}_{\tau-2}\|^2)$, the third inequality is immediately implied by the second inequality. ■

B.2. Proof for large-gradient scenario

We prove Lemma 12 in this subsection. Throughout this subsection, we let $\mathcal{S}$ be the subspace with eigenvalues in $(\theta^2/[\eta(2-\theta)^2], \ell]$, and let $\mathcal{S}^c$ be the complementary subspace. Also let $\mathcal{P}_\mathcal{S}$ and $\mathcal{P}_{\mathcal{S}^c}$ be the corresponding projections. We note $\theta^2/[\eta(2-\theta)^2] = \Theta(\sqrt{\rho\epsilon})$, and this particular choice lies at the boundary between the real eigenvalues and complex eigenvalues of the matrix $\mathbf{A}_j$, as shown in Lemma 26.

The first lemma shows that if momentum or gradient is very large, then the Hamiltonian already has sufficient decrease on average.

Lemma 20 *Under the setting of Theorem 8, if $\|\mathbf{v}_t\| \ge \mathscr{M}$ or $\|\nabla f(\mathbf{x}_t)\| \ge 2\ell\mathscr{M}$, and at time step t only AGD is used without NCE or perturbation, then:*

$$E_{t+1} - E_t \le -4\mathscr{E}/\mathscr{T}.$$

Proof When $\|\mathbf{v}_t\| \ge \frac{\epsilon\sqrt{\kappa}}{10\ell}$, by Lemma 9, we have:

$$E_{t+1} - E_t \le -\frac{\theta}{2\eta}\|\mathbf{v}_t\|^2 \le -\Omega\left(\frac{\ell}{\sqrt{\kappa}}\frac{\epsilon^2\kappa}{\ell^2}c^{-2}\right) = -\Omega\left(\frac{\epsilon^2\sqrt{\kappa}}{2\ell}c^{-2}\right) \le -\Omega(\frac{\mathscr{E}}{\mathscr{T}}c^6) \le -\frac{4\mathscr{E}}{\mathscr{T}}.$$

The last step is by picking c to be a large enough constant. When $\|\mathbf{v}_t\| \le \mathscr{M}$ but $\|\nabla f(\mathbf{x}_t)\| \ge 2\ell\mathscr{M}$, by the gradient Lipschitz assumption, we have:

$$\|\nabla f(\mathbf{y}_t)\| \ge \|\nabla f(\mathbf{x}_t)\| - (1-\theta)\ell\|\mathbf{v}_t\| \ge \ell\mathscr{M}.$$

Similarly, by Lemma 9, we have:

$$E_{t+1} - E_t \le -\frac{\eta}{4}\|\nabla f(\mathbf{y}_t)\|^2 \le -\Omega(\frac{\epsilon^2\kappa}{\ell}c^{-2}) \le -\Omega(\frac{\mathscr{E}}{\mathscr{T}}c^6) \le -\frac{4\mathscr{E}}{\mathscr{T}}.$$

Again the last step is by picking c to be a large enough constant, which finishes the proof. ■

Next, we show that if the initial momentum is small, but the initial gradient on the nonconvex subspace $\mathcal{S}^c$ is large enough, then within $O(\mathscr{T})$ steps, the Hamiltonian will decrease by at least $\mathscr{E}$.

Lemma 21 (Formal Version of Lemma 15) *Under the setting of Theorem 8, if $\|\mathcal{P}_{\mathcal{S}^c}\nabla f(\mathbf{x}_0)\| \ge \frac{\epsilon}{2}$, $\|\mathbf{v}_0\| \le \mathscr{M}$, $\mathbf{v}_0^\top[\mathcal{P}_\mathcal{S}^\top\nabla^2 f(\mathbf{x}_0)\mathcal{P}_\mathcal{S}]\mathbf{v}_0 \le 2\sqrt{\rho\epsilon}\mathscr{M}^2$, and for $t \in [0, \mathscr{T}/4]$ only AGD steps are used without NCE or perturbation, then:*

$$E_{\mathscr{T}/4} - E_0 \le -\mathscr{E}.$$

Proof The high-level plan is a proof by contradiction. We first assume that the energy doesn't decrease very much; that is, $E_{\mathscr{T}/4} - E_0 \ge -\mathscr{E}$ for a small enough constant μ. By Corollary 11 and the Cauchy-Swartz inequality, this immediately implies that for all $t \le \mathscr{T}$, we have $\|\mathbf{x}_t - \mathbf{x}_0\| \le \sqrt{2\eta\mathscr{T}\mathscr{E}/(4\theta)} = \mathscr{S}/2$. In the rest of the proof we will show that this leads to a contradiction.

Given initial $\mathbf{x}_0$ and $\mathbf{v}_0$, we define $\mathbf{x}_{-1} = \mathbf{x}_0 - \mathbf{v}_0$. Without loss of generality, set $\mathbf{x}_0$ as the origin $\mathbf{0}$. Using the notation and results of Lemma 18, we have the following update equation:

$$\begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-1} \end{pmatrix} = \mathbf{A}^t \begin{pmatrix} 0 \\ -\mathbf{v}_0 \end{pmatrix} - \eta \sum_{\tau=0}^{t-1} \mathbf{A}^{t-1-\tau} \begin{pmatrix} \nabla f(0) + \delta_\tau \\ 0 \end{pmatrix}.$$

Consider the j-th eigen-direction of $\mathcal{H} = \nabla^2 f(\mathbf{0})$, recall the definition of the 2×2 block matrix $\mathbf{A}_j$ as in (12), and denote

$$(a_t^{(j)}, \ -b_t^{(j)}) = \begin{pmatrix} 1 & 0 \end{pmatrix} \mathbf{A}_j^t.$$

Then we have for the j-th eigen-direction:

$$\begin{aligned} x_t^{(j)} =& b_t^{(j)} v_0^{(j)} - \eta \sum_{\tau=0}^{t-1} a_{t-1-\tau}^{(j)} (\nabla f(0)^{(j)} + \delta_\tau^{(j)}) \\ =& -\eta \left[\sum_{\tau=0}^{t-1} a_\tau^{(j)} \right] \left(\nabla f(0)^{(j)} + \sum_{\tau=0}^{t-1} p_\tau^{(j)} \delta_\tau^{(j)} + q_t^{(j)} v_0^{(j)} \right), \end{aligned}$$

where

$$p_\tau^{(j)} = \frac{a_{t-1-\tau}^{(j)}}{\sum_{\tau=0}^{t-1} a_\tau^{(j)}} \quad \text{and} \quad q_t^{(j)} = -\frac{b_t^{(j)}}{\eta \sum_{\tau=0}^{t-1} a_\tau^{(j)}}.$$

Clearly $\sum_{\tau=0}^{t-1} p_\tau^{(j)} = 1$. For $j \in \mathcal{S}^c$, by Lemma 30, we know $\sum_{\tau=0}^{t-1} a_\tau^{(j)} \geq \Omega(\frac{1}{\theta^2})$. We can thus further write the above equation as:

$$x_t^{(j)} = -\eta \left[\sum_{\tau=0}^{t-1} a_\tau^{(j)} \right] \left(\nabla f(0)^{(j)} + \tilde{\delta}^{(j)} + \tilde{v}^{(j)} \right),$$

where $\tilde{\delta}^{(j)} = \sum_{\tau=0}^{t-1} p_\tau^{(j)} \delta_\tau^{(j)}$ and $\tilde{v}^{(j)} = q_t^{(j)} v_0^{(j)}$, coming from the Hessian Lipschitz assumption and the initial momentum respectively. For the remaining part, we would like to bound $\|\mathcal{P}_{\mathcal{S}^c} \tilde{\delta}\|$ and $\|\mathcal{P}_{\mathcal{S}^c} \tilde{\mathbf{v}}\|$, and show that both of them are small compared to $\|\mathcal{P}_{\mathcal{S}^c} \nabla f(\mathbf{x}_0)\|$.

First, for the $\|\mathcal{P}_{\mathcal{S}^c} \tilde{\delta}\|$ term, we know by definition of the subspace $\mathcal{S}^c$, and given that both eigenvalues of $\mathbf{A}_j$ are real and positive according to Lemma 26, such that $p_\tau^{(j)}$ is positive by Lemma 24, we have for any $j \in \mathcal{S}^c$:

$$\begin{aligned} |\tilde{\delta}^{(j)}| =& |\sum_{\tau=0}^{t-1} p_\tau^{(j)} \delta_\tau^{(j)}| \leq \sum_{\tau=0}^{t-1} p_\tau^{(j)} (|\delta_0^{(j)}| + |\delta_\tau^{(j)} - \delta_0^{(j)}|) \\ \leq& \left[\sum_{\tau=0}^{t-1} p_\tau^{(j)} \right] \left(|\delta_0^{(j)}| + \sum_{\tau=1}^{t-1} |\delta_\tau^{(j)} - \delta_{\tau-1}^{(j)}| \right) \leq |\delta_0^{(j)}| + \sum_{\tau=1}^{t-1} |\delta_\tau^{(j)} - \delta_{\tau-1}^{(j)}|. \end{aligned}$$

By the Cauchy-Swartz inequality, this gives:

$$\left\|\mathcal{P}_{\mathcal{S}^c}\tilde{\delta}\right\|^2 = \sum_{j\in\mathcal{S}^c}|\tilde{\delta}^{(j)}|^2 \le \sum_{j\in\mathcal{S}^c}(|\delta_0^{(j)}| + \sum_{\tau=1}^{t-1}|\delta_\tau^{(j)} - \delta_{\tau-1}^{(j)}|)^2 \le 2\left[\sum_{j\in\mathcal{S}^c}|\delta_0^{(j)}|^2 + \sum_{j\in\mathcal{S}^c}(\sum_{\tau=1}^{t-1}|\delta_\tau^{(j)} - \delta_{\tau-1}^{(j)}|)^2\right]$$
$$\le 2\left[\sum_{j\in\mathcal{S}^c}|\delta_0^{(j)}|^2 + t\sum_{j\in\mathcal{S}^c}\sum_{\tau=1}^{t-1}|\delta_\tau^{(j)} - \delta_{\tau-1}^{(j)}|^2\right] \le 2\left\|\delta_0\right\|^2 + 2t\sum_{\tau=1}^{t-1}\left\|\delta_\tau - \delta_{\tau-1}\right\|^2.$$

Recall that for $t \le \mathscr{T}$, we have $\|\mathbf{x}_t\| \le \mathscr{S}/2$. By Proposition 19, we know: $\|\delta_0\| \le O(\rho\mathscr{S}^2)$, and by Corollary 11 and Proposition 19:

$$t\sum_{\tau=1}^{t-1}\left\|\delta_\tau - \delta_{\tau-1}\right\|^2 \le O(\rho^2\mathscr{S}^2)t\sum_{\tau=1}^{t-1}\left\|\mathbf{x}_\tau - \mathbf{x}_{\tau-1}\right\|^2 \le O(\rho^2\mathscr{S}^4).$$

This gives $\|\mathcal{P}_{\mathcal{S}^c}\tilde{\delta}\| \le O(\rho\mathscr{S}^2) \le O(\epsilon\cdot c^{-6}) \le \epsilon/10$.

Next we consider the $\|\mathcal{P}_{\mathcal{S}^c}\tilde{\mathbf{v}}\|$ term. By Lemma 30, we have

$$-\eta q_t^{(j)} = \frac{b_t}{\sum_{\tau=0}^{t-1}a_\tau} \le O(1)\max\{\theta, \sqrt{\eta|\lambda_j|}\}.$$

This gives:

$$\|\mathcal{P}_{\mathcal{S}^c}\tilde{\mathbf{v}}\|^2 = \sum_{j\in\mathcal{S}^c}[q_t^{(j)}v_0^{(j)}]^2 \le O(1)\sum_{j\in\mathcal{S}^c}\frac{\max\{\eta|\lambda_j|,\theta^2\}}{\eta^2}[v_0^{(j)}]^2. \tag{13}$$

Recall that we have assumed by way of contradiction that $E_{\mathscr{T}/4} - E_0 \le -\mathscr{E}$. By the precondition that NCE is not used at $t=0$, due to the certificate (2), we have:

$$\frac{1}{2}\mathbf{v}_0^\top\nabla^2 f(\zeta_0)\mathbf{v}_0 \ge -\frac{\gamma}{2}\left\|\mathbf{v}_0\right\|^2 = -\frac{\sqrt{\rho\epsilon}}{8}\left\|\mathbf{v}_0\right\|^2,$$

where $\zeta_0 = \phi\mathbf{x}_0 + (1-\phi)\mathbf{y}_0$ and $\phi\in[0,1]$. Noting that we fix $\mathbf{x}_0$ as the origin $\mathbf{0}$, by the Hessian Lipschitz property, it is easy to show that $\left\|\nabla^2 f(\zeta_0) - \mathcal{H}\right\| \le \rho\left\|\mathbf{y}_0\right\| \le \rho\left\|\mathbf{v}_0\right\| \le \rho\mathscr{M} \le \sqrt{\rho\epsilon}$. This gives:

$$\mathbf{v}_0\mathcal{H}\mathbf{v}_0 \ge -2\sqrt{\rho\epsilon}\left\|\mathbf{v}_0\right\|^2.$$

Again letting λ_j denote the eigenvalues of $\mathcal{H}$, rearranging the above sum give:

$$\sum_{j:\lambda_j\le 0}|\lambda_j|[v_0^{(j)}]^2 \le O(\sqrt{\rho\epsilon})\left\|\mathbf{v}_0\right\|^2 + \sum_{j:\lambda_j>0}\lambda_j[v_0^{(j)}]^2$$
$$\le O(\sqrt{\rho\epsilon})\left\|\mathbf{v}_0\right\|^2 + \sum_{j:\lambda_j>\theta^2/\eta(2-\theta)^2}\lambda_j[v_0^{(j)}]^2 \le O(\sqrt{\rho\epsilon})\left\|\mathbf{v}_0\right\|^2 + \mathbf{v}_0^\top[\mathcal{P}_{\mathcal{S}}^\top\mathcal{H}\mathcal{P}_{\mathcal{S}}]\mathbf{v}_0.$$

The second inequality uses the fact that $\theta^2/\eta(2-\theta)^2 \le O(\sqrt{\rho\epsilon})$. Substituting into (13) gives:

$$\|\mathcal{P}_{\mathcal{S}^c}\tilde{\mathbf{v}}\|^2 \le O(\frac{1}{\eta})\left[\sqrt{\rho\epsilon}\left\|\mathbf{v}_0\right\|^2 + \mathbf{v}_0^\top[\mathcal{P}_{\mathcal{S}}^\top\mathcal{H}\mathcal{P}_{\mathcal{S}}]\mathbf{v}_0\right] \le O(\ell\sqrt{\rho\epsilon}\mathscr{M}^2) = O(\epsilon^2 c^{-2}) \le \epsilon^2/100.$$

Finally, putting all pieces together, we have:

$$\begin{aligned}\|\mathbf{x}_t\| \geq \|\mathcal{P}_{\mathcal{S}^c}\mathbf{x}_t\| \geq \eta \left[\min_{j\in\mathcal{S}^c}\sum_{\tau=0}^{t-1} a_\tau^{(j)}\right] \left\|\mathcal{P}_{\mathcal{S}^c}(\nabla f(0)+\tilde{\delta}+\tilde{\mathbf{v}})\right\| \\ \geq \Omega(\frac{\eta}{\theta^2})\left[\|\mathcal{P}_{\mathcal{S}^c}\nabla f(0)\| - \left\|\mathcal{P}_{\mathcal{S}^c}\tilde{\delta}\right\| - \|\mathcal{P}_{\mathcal{S}^c}\tilde{\mathbf{v}})\|\right] \geq \Omega(\frac{\eta\epsilon}{\theta^2}) \geq \Omega(\mathscr{S}c^3) \geq \mathscr{S}\end{aligned}$$

which contradicts the fact $\|\mathbf{x}_t\|$ that remains inside the ball around $\mathbf{0}$ with radius $\mathscr{S}/2$. ■

The next lemma shows that if the initial momentum and gradient are reasonably small, and the Hamitonian does not have sufficient decrease over the next $\mathscr{T}$ iterations, then both the gradient and momentum of the strongly convex component $\mathcal{S}$ will vanish in $\mathscr{T}/4$ iterations.

Lemma 22 (Formal Version of Lemma 14) *Under the setting of Theorem 8, suppose $\|\mathbf{v}_0\| \leq \mathscr{M}$ and $\|\nabla f(\mathbf{x}_0)\| \leq 2\ell\mathscr{M}$, $E_{\mathscr{T}/2} - E_0 \geq -\mathscr{E}$, and for $t \in [0, \mathscr{T}/2]$ only AGD steps are used, without NCE or perturbation. Then $\forall\, t \in [\mathscr{T}/4, \mathscr{T}/2]$:*

$$\|\mathcal{P}_{\mathcal{S}}\nabla f(\mathbf{x}_t)\| \leq \frac{\epsilon}{2} \;\; \textit{and} \;\; \mathbf{v}_t^\top[\mathcal{P}_{\mathcal{S}}^\top\nabla^2 f(\mathbf{x}_0)\mathcal{P}_{\mathcal{S}}]\mathbf{v}_t \leq \sqrt{\rho\epsilon}\mathscr{M}^2.$$

Proof Since $E_{\mathscr{T}} - E_0 \geq -\mathscr{E}$, by Corollary 11 and the Cauchy-Swartz inequality, we see that for all $t \leq \mathscr{T}$ we have $\|\mathbf{x}_t - \mathbf{x}_0\| \leq \sqrt{2\eta\mathscr{T}\mathscr{E}/\theta} = \mathscr{S}$.

Given initial $\mathbf{x}_0$ and $\mathbf{v}_0$, we define $\mathbf{x}_{-1} = \mathbf{x}_0 - \mathbf{v}_0$. Without loss of generality, setting $\mathbf{x}_0$ as the origin $\mathbf{0}$, by the notation and results of Lemma 18, we have the update equation:

$$\begin{pmatrix}\mathbf{x}_t \\ \mathbf{x}_{t-1}\end{pmatrix} = \mathbf{A}^t\begin{pmatrix}0 \\ -\mathbf{v}_0\end{pmatrix} - \eta\sum_{\tau=0}^{t-1}\mathbf{A}^{t-1-\tau}\begin{pmatrix}\nabla f(0)+\delta_\tau \\ 0\end{pmatrix}. \tag{14}$$

First we prove the upper bound on the gradient: $\forall\, t \in [\mathscr{T}/4, \mathscr{T}]$, we have $\|\mathcal{P}_{\mathcal{S}}\nabla f(\mathbf{x}_t)\| \leq \frac{\epsilon}{2}$. Let $\Delta_t = \int_0^1(\nabla^2 f(\phi\mathbf{x}_t) - \mathcal{H})\mathrm{d}\phi$. According to (14), we have:

$$\begin{aligned}\nabla f(\mathbf{x}_t) =& \nabla f(0) + (\mathcal{H}+\Delta_t)\mathbf{x}_t \\ =& \underbrace{\left(\mathbf{I} - \eta\mathcal{H}\begin{pmatrix}\mathbf{I} & 0\end{pmatrix}\sum_{\tau=0}^{t-1}\mathbf{A}^{t-1-\tau}\begin{pmatrix}\mathbf{I}\\0\end{pmatrix}\right)\nabla f(0)}_{\boldsymbol{g}_1} + \underbrace{\mathcal{H}\begin{pmatrix}\mathbf{I} & 0\end{pmatrix}\mathbf{A}^t\begin{pmatrix}0\\-\mathbf{v}_0\end{pmatrix}}_{\boldsymbol{g}_2} \\ & \underbrace{-\eta\mathcal{H}\begin{pmatrix}\mathbf{I} & 0\end{pmatrix}\sum_{\tau=0}^{t-1}\mathbf{A}^{t-1-\tau}\begin{pmatrix}\delta_t\\0\end{pmatrix}}_{\boldsymbol{g}_3} + \underbrace{\Delta_t\mathbf{x}_t}_{\boldsymbol{g}_4}.\end{aligned}$$

We will upper bound four terms $\boldsymbol{g}_1, \boldsymbol{g}_2, \boldsymbol{g}_3, \boldsymbol{g}_4$ separately. Clearly, for the last term $\boldsymbol{g}_4$, we have:

$$\|\boldsymbol{g}_4\| \leq \rho\|\mathbf{x}_t\|^2 \leq O(\rho\mathscr{S}^2) = O(\epsilon c^{-6}) \leq \epsilon/8.$$

Next, we show that the first two terms $\boldsymbol{g}_1, \boldsymbol{g}_2$ become very small for $t \in [\mathscr{T}/4, \mathscr{T}]$. Consider coordinate $j \in \mathcal{S}$ and the 2×2 block matrix $\mathbf{A}_j$. By Lemma 25 we have:

$$1 - \eta\lambda_j\begin{pmatrix}1 & 0\end{pmatrix}\sum_{\tau=0}^{t-1}\mathbf{A}_j^{t-1-\tau}\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}1 & 0\end{pmatrix}\mathbf{A}_j^t\begin{pmatrix}1\\1\end{pmatrix}.$$

Denote:

$$(a_t^{(j)}, \ -b_t^{(j)}) = \begin{pmatrix} 1 & 0 \end{pmatrix} \mathbf{A}_j^t.$$

By Lemma 32, we know:

$$\max_{j \in \mathcal{S}} \left\{ |a_t^{(j)}|, |b_t^{(j)}| \right\} \le (t+1)(1-\theta)^{\frac{t}{2}}.$$

This immediately gives when $t \ge \mathscr{T}/4 = \Omega(\frac{c}{\theta}\log\frac{1}{\theta})$ for c sufficiently large:

$$\|\mathcal{P}_{\mathcal{S}}\boldsymbol{g}_1\|^2 = \sum_{j\in\mathcal{S}} |(a_t^{(j)} - b_t^{(j)})\nabla f(0)^{(j)}|^2 \le (t+1)^2(1-\theta)^t \|\nabla f(0)\|^2 \le \epsilon^2/64$$
$$\|\mathcal{P}_{\mathcal{S}}\boldsymbol{g}_2\|^2 = \sum_{j\in\mathcal{S}} |\lambda_j b_t^{(j)} \mathbf{v}_0^{(j)}|^2 \le \ell^2(t+1)^2(1-\theta)^t \|\mathbf{v}_0\|^2 \le \epsilon^2/64.$$

Finally, for $\boldsymbol{g}_3$, by Lemma 34, for all $j \in \mathcal{S}$, we have

$$|\boldsymbol{g}_3^{(j)}| = \left| \eta\lambda_j \sum_{\tau=0}^{t-1} a_\tau^{(j)} \delta_{t-1-\tau} \right| \le |\delta_{t-1}^{(j)}| + \sum_{\tau=1}^{t-1} |\delta_\tau^{(j)} - \delta_{\tau-1}^{(j)}|.$$

By Proposition 19, this gives:

$$\|\mathcal{P}_{\mathcal{S}}\boldsymbol{g}_3\|^2 \le 2\|\delta_{t-1}\|^2 + 2t\sum_{\tau=1}^{t-1}\|\delta_\tau - \delta_{\tau-1}\|^2 \le O(\rho^2\mathscr{S}^4) \le O(\epsilon^2 \cdot c^{-12}) \le \epsilon^2/64.$$

In sum, this gives for any fixed $t \in [\mathscr{T}/4, \mathscr{T}]$:

$$\|\mathcal{P}_{\mathcal{S}}\nabla f(\mathbf{x}_t)\| \le \|\mathcal{P}_{\mathcal{S}}\boldsymbol{g}_1\| + \|\mathcal{P}_{\mathcal{S}}\boldsymbol{g}_2\| + \|\mathcal{P}_{\mathcal{S}}\boldsymbol{g}_3\| + \|\boldsymbol{g}_4\| \le \frac{\epsilon}{2}.$$

We now provide a similar argument to prove the upper bound for the momentum. That is, $\forall\, t \in [\mathscr{T}/4, \mathscr{T}]$, we show $\mathbf{v}_t^\top[\mathcal{P}_{\mathcal{S}}^\top\nabla^2 f(\mathbf{x}_0)\mathcal{P}_{\mathcal{S}}]\mathbf{v}_t \le \sqrt{\rho\epsilon}\mathscr{M}^2$. According to (14), we have:

$$\mathbf{v}_t = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-1} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & -1 \end{pmatrix} \mathbf{A}^t \begin{pmatrix} 0 \\ -\mathbf{v}_0 \end{pmatrix}}_{\boldsymbol{m}_1} \underbrace{- \eta \begin{pmatrix} 1 & -1 \end{pmatrix} \sum_{\tau=0}^{t-1} \mathbf{A}^{t-1-\tau} \begin{pmatrix} \nabla f(0) \\ 0 \end{pmatrix}}_{\boldsymbol{m}_2}$$
$$\underbrace{- \eta \begin{pmatrix} 1 & -1 \end{pmatrix} \sum_{\tau=0}^{t-1} \mathbf{A}^{t-1-\tau} \begin{pmatrix} \delta_\tau \\ 0 \end{pmatrix}}_{\boldsymbol{m}_3}.$$

Consider the j-th eigendirection, so that $j \in \mathcal{S}$, and recall the 2×2 block matrix $\mathbf{A}_j$. Denoting

$$(a_t^{(j)}, \ -b_t^{(j)}) = \begin{pmatrix} 1 & 0 \end{pmatrix} \mathbf{A}_j^t,$$

by Lemma 24 and 32, we have for $t \ge \mathscr{T}/4 = \Omega(\frac{c}{\theta}\log\frac{1}{\theta})$ with c sufficiently large:

$$\left\| [\mathcal{P}_{\mathcal{S}}^\top\nabla^2 f(\mathbf{x}_0)\mathcal{P}_{\mathcal{S}}]^{\frac{1}{2}} \boldsymbol{m}_1 \right\|^2 = \sum_{j\in\mathcal{S}} |\lambda_j^{\frac{1}{2}}(b_t^{(j)} - b_{t-1}^{(j)})\mathbf{v}_0^{(j)}|^2 \le \ell(t+1)^2(1-\theta)^t\|\mathbf{v}_0\|^2 \le O(\frac{\epsilon^2}{\ell}c^{-3}) \le \frac{1}{3}\sqrt{\rho\epsilon}\mathscr{M}^2.$$

On the other hand, by Lemma 25, we have:

$$\left|\eta\lambda_j \begin{pmatrix}1 & -1\end{pmatrix} \sum_{\tau=0}^{t-1} \mathbf{A}_j^{t-1-\tau} \begin{pmatrix}1\\0\end{pmatrix}\right| = \left|\eta\lambda_j \begin{pmatrix}1 & 0\end{pmatrix} \sum_{\tau=0}^{t-1} (\mathbf{A}_j^{t-1-\tau} - \mathbf{A}_j^{t-2-\tau}) \begin{pmatrix}1\\0\end{pmatrix}\right| = \left|\begin{pmatrix}1 & 0\end{pmatrix} (\mathbf{A}_j^{t} - \mathbf{A}_j^{t-1}) \begin{pmatrix}1\\1\end{pmatrix}\right|.$$

This gives, for $t \geq \mathscr{T}/4 = \Omega(\frac{c}{\theta}\log\frac{1}{\theta})$, and for c sufficiently large:

$$\begin{aligned}\left\|[\mathcal{P}_{\mathcal{S}}^\top \nabla^2 f(\mathbf{x}_0)\mathcal{P}_{\mathcal{S}}]^{\frac{1}{2}}\boldsymbol{m}_2\right\|^2 =& \sum_{j\in\mathcal{S}} |\lambda_j^{-\frac{1}{2}}(a_t^{(j)} - a_{t-1}^{(j)} - b_t^{(j)} + b_{t-1}^{(j)})\nabla f(0)^{(j)}|^2 \\ \leq& O(\frac{1}{\sqrt{\rho\epsilon}})(t+1)^2(1-\theta)^t \left\|\nabla f(0)\right\|^2 \leq O(\frac{\epsilon^2}{\ell}c^{-3}) \leq \frac{1}{3}\sqrt{\rho\epsilon}\mathscr{M}^2.\end{aligned}$$

Finally, for any $j \in \mathcal{S}$, by Lemma 34, we have:

$$|(\mathcal{H}^{\frac{1}{2}}\boldsymbol{m}_3)^{(j)}| = |\eta\lambda_j^{\frac{1}{2}} \sum_{\tau=0}^{t-1}(a_\tau - a_{\tau-1})\delta_{t-1-\tau}| \leq \sqrt{\eta}\left[\sum |\delta_{t-1}^{(j)}| + \sum_{\tau=1}^{t-1} |\delta_\tau^{(j)} - \delta_{\tau-1}^{(j)}|\right].$$

Again by Proposition 19:

$$\left\|[\mathcal{P}_{\mathcal{S}}^\top \nabla^2 f(\mathbf{x}_0)\mathcal{P}_{\mathcal{S}}]^{\frac{1}{2}}\boldsymbol{m}_3\right\|^2 = \eta\left[2\left\|\delta_{t-1}\right\|^2 + 2t\sum_{\tau=1}^{t-1}\left\|\delta_\tau - \delta_{\tau-1}\right\|^2\right] \leq O(\eta\rho^2\mathscr{S}^4) \leq O(\frac{\epsilon^2}{\ell}c^{-6}) \leq \frac{1}{3}\sqrt{\rho\epsilon}\mathscr{M}^2.$$

Putting everything together, we have:

$$\begin{aligned}\mathbf{v}_t^\top[\mathcal{P}_{\mathcal{S}}^\top \nabla^2 f(\mathbf{x}_0)\mathcal{P}_{\mathcal{S}}]\mathbf{v}_t \leq& \left\|[\mathcal{P}_{\mathcal{S}}^\top \nabla^2 f(\mathbf{x}_0)\mathcal{P}_{\mathcal{S}}]^{\frac{1}{2}}\boldsymbol{m}_1\right\|^2 + \left\|[\mathcal{P}_{\mathcal{S}}^\top \nabla^2 f(\mathbf{x}_0)\mathcal{P}_{\mathcal{S}}]^{\frac{1}{2}}\boldsymbol{m}_2\right\|^2 \\ &+ \left\|[\mathcal{P}_{\mathcal{S}}^\top \nabla^2 f(\mathbf{x}_0)\mathcal{P}_{\mathcal{S}}]^{\frac{1}{2}}\boldsymbol{m}_3\right\|^2 \leq \sqrt{\rho\epsilon}\mathscr{M}^2.\end{aligned}$$

This finishes the proof. ■

Finally, we are ready to prove the main lemma of this subsection (Lemma 12), which claims that if gradients in $\mathscr{T}$ iterations are always large, then the Hamiltonian will decrease sufficiently within a small number of steps.

Lemma 12 (Large gradient) *Consider the setting of Theorem 8. If $\|\nabla f(\mathbf{x}_\tau)\| \geq \epsilon$ for all $\tau \in [0, \mathscr{T}]$, then by running Algorithm 2 we have $E_{\mathscr{T}} - E_0 \leq -\mathscr{E}$.*

Proof Since $\|\nabla f(\mathbf{x}_\tau)\| \geq \epsilon$ for all $\tau \in [0, \mathscr{T}]$, according to Algorithm 2, the precondition to add perturbation never holds, so Algorithm will not add any perturbation in these $\mathscr{T}$ iterations.

Next, suppose there is at least one iteration where NCE is used. Then by Lemma 17, we know that that step alone gives $\mathscr{E}$ decrease in the Hamiltonian. According to Lemma 9 and Lemma 17 we know that without perturbation, the Hamiltonian decreases monotonically in the remaining steps. This means whenever at least one NCE step is performed, Lemma 12 immediately holds.

For the remainder of the proof, we can restrict the discussion to the case where NCE is never performed in steps $\tau \in [0, \mathscr{T}]$. Letting

$$\tau_1 = \arg\min_{t\in[0,\mathscr{T}]} \left\{t \,|\, \|\mathbf{v}_t\| \leq \mathscr{M} \text{ and } \|\nabla f(\mathbf{x}_t)\| \leq 2\ell\mathscr{M}\right\},$$

we know in case $\tau_1 \geq \frac{\mathscr{T}}{4}$, that Lemma 20 ensures $E_{\mathscr{T}} - E_0 \leq E_{\frac{\mathscr{T}}{4}} - E_0 \leq -\mathscr{E}$. Thus, we only need to discuss the case $\tau_1 \leq \frac{\mathscr{T}}{4}$. Again, if $E_{\tau_1+\mathscr{T}/2} - E_{\tau_1} \leq -\mathscr{E}$, Lemma 12 immediately holds. For the remaining case, $E_{\tau_1+\mathscr{T}/2} - E_{\tau_1} \leq -\mathscr{E}$, we apply Lemma 22 starting at τ_1, and obtain

$$\|\mathcal{P}_{\mathcal{S}}\nabla f(\mathbf{x}_t)\| \leq \frac{\epsilon}{2} \text{ and } \mathbf{v}_t^\top[\mathcal{P}_{\mathcal{S}}^\top\nabla^2 f(\mathbf{x}_{\tau_1})\mathcal{P}_{\mathcal{S}}]\mathbf{v}_t \leq \sqrt{\rho\epsilon}\mathscr{M}^2. \quad \forall t \in [\tau_1 + \frac{\mathscr{T}}{4}, \tau_1 + \frac{\mathscr{T}}{2}].$$

Letting:

$$\tau_2 = \arg\min_{t\in[\tau_1+\frac{\mathscr{T}}{4},\mathscr{T}]} \{t \,|\, \|\mathbf{v}_t\| \leq \mathscr{M}\},$$

by Lemma 20 we again know we only need to discuss the case where $\tau_2 \leq \tau_1 + \frac{\mathscr{T}}{2}$; otherwise, we already guarantee sufficient decrease in the Hamiltonian. Then, we clearly have $\|\mathcal{P}_{\mathcal{S}}\nabla f(\mathbf{x}_{\tau_2})\| \leq \frac{\epsilon}{2}$, also by the precondition of Lemma 12, we know $\|\nabla f(\mathbf{x}_{\tau_2})\| \geq \epsilon$, thus $\|\mathcal{P}_{\mathcal{S}^c}\nabla f(\mathbf{x}_{\tau_2})\| \geq \frac{\epsilon}{2}$. On the other hand, since if the Hamiltonian does not decrease enough, $E_{\tau_2} - E_0 \geq -\mathscr{E}$, by Lemma 11, we have $\|\mathbf{x}_{\tau_1} - \mathbf{x}_{\tau_2}\| \leq 2\mathscr{S}$, by the Hessian Lipschitz property, which gives:

$$\mathbf{v}_{\tau_2}^\top[\mathcal{P}_{\mathcal{S}}^\top\nabla^2 f(\mathbf{x}_{\tau_2})\mathcal{P}_{\mathcal{S}}]\mathbf{v}_{\tau_2} \leq \mathbf{v}_{\tau_2}^\top[\mathcal{P}_{\mathcal{S}}^\top\nabla^2 f(\mathbf{x}_{\tau_1})\mathcal{P}_{\mathcal{S}}]\mathbf{v}_{\tau_2} + \left\|\nabla^2 f(\mathbf{x}_{\tau_1}) - \nabla^2 f(\mathbf{x}_{\tau_2})\right\| \|\mathbf{v}_{\tau_2}\|^2 \leq 2\sqrt{\rho\epsilon}\mathscr{M}^2.$$

Now $\mathbf{x}_{\tau_2}$ satisfies all the preconditions of Lemma 21, and by applying Lemma 21 we finish the proof. ■

B.3. Proof for negative-curvature scenario

We prove Lemma 13 in this section. We consider two trajectories, starting at $\mathbf{x}_0$ and $\mathbf{x}_0'$, with $\mathbf{v}_0 = \mathbf{v}_0'$, where $\mathbf{w}_0 = \mathbf{x}_0 - \mathbf{x}_0' = r_0\mathbf{e}_1$, where $\mathbf{e}_1$ is the minimum eigenvector direction of $\mathcal{H}$, and where r_0 is not too small. We show that at least one of the trajectories will escape saddle points efficiently.

Lemma 23 (Formal Version of Lemma 16)
Under the same setting as Theorem 8, suppose $\|\nabla f(\tilde{\mathbf{x}})\| \leq \epsilon$ and $\lambda_{\min}(\nabla^2 f(\tilde{\mathbf{x}})) \leq -\sqrt{\rho\epsilon}$. Let $\mathbf{x}_0$ and $\mathbf{x}_0'$ be at distance at most r from $\tilde{\mathbf{x}}$. Let $\mathbf{x}_0 - \mathbf{x}_0' = r_0 \cdot \mathbf{e}_1$ and let $\mathbf{v}_0 = \mathbf{v}_0' = \tilde{\mathbf{v}}$ where $\mathbf{e}_1$ is the minimum eigen-direction of $\nabla^2 f(\tilde{\mathbf{x}})$. Let $r_0 \geq \frac{\delta\mathscr{E}}{2\Delta_f} \cdot \frac{r}{\sqrt{d}}$. Then, running AGD starting at $(\mathbf{x}_0, \mathbf{v}_0)$ and $(\mathbf{x}_0', \mathbf{v}_0')$ respectively, we have:

$$\min\{E_{\mathscr{T}} - \widetilde{E}, E'_{\mathscr{T}} - \widetilde{E}\} \leq -\mathscr{E},$$

where $\widetilde{E}, E_{\mathscr{T}}$ and $E'_{\mathscr{T}}$ are the Hamiltonians at $(\tilde{\mathbf{x}}, \tilde{\mathbf{v}}), (\mathbf{x}_{\mathscr{T}}, \mathbf{v}_{\mathscr{T}})$ and $(\mathbf{x}'_{\mathscr{T}}, \mathbf{v}'_{\mathscr{T}})$ respectively.

Proof Assume none of the two sequences decrease the Hamiltonian fast enough; that is,

$$\min\{E_{\mathscr{T}} - E_0, E'_{\mathscr{T}} - E'_0\} \geq -2\mathscr{E},$$

where E_0 and E'_0 are the Hamiltonians at $(\mathbf{x}_0, \mathbf{v}_0)$ and $(\mathbf{x}_0', \mathbf{v}_0')$. Then, by Corollary 11 and the Cauchy-Swartz inequality, we have for any $t \leq \mathscr{T}$:

$$\max\{\|\mathbf{x}_t - \tilde{\mathbf{x}}\|, \|\mathbf{x}_t' - \tilde{\mathbf{x}}\|\} \leq r + \max\{\|\mathbf{x}_t - \mathbf{x}_0\|, \|\mathbf{x}_t' - \mathbf{x}_0'\|\} \leq r + \sqrt{4\eta\mathscr{T}\mathscr{E}/\theta} \leq 2\mathscr{S}.$$

Fix the origin $\mathbf{0}$ at $\tilde{\mathbf{x}}$ and let $\mathcal{H}$ be the Hessian at $\tilde{\mathbf{x}}$. Recall that the update equation of AGD (Algorithm 1) can be re-written as:

$$\mathbf{x}_{t+1} = (2-\theta)\mathbf{x}_t - (1-\theta)\mathbf{x}_{t-1} - \eta\nabla f((2-\theta)\mathbf{x}_t - (1-\theta)\mathbf{x}_{t-1})$$

Taking the difference of two AGD sequences starting from $\mathbf{x}_0, \mathbf{x}_0'$, and let $\mathbf{w}_t = \mathbf{x}_t - \mathbf{x}_t'$, we have:

$$\begin{aligned}\mathbf{w}_{t+1} =& (2-\theta)\mathbf{w}_t - (1-\theta)\mathbf{w}_{t-1} - \eta\nabla f(\mathbf{y}_t) + \eta\nabla f(\mathbf{y}_t') \\ =& (2-\theta)(I - \eta\mathcal{H} - \eta\Delta_t)\mathbf{w}_t - (1-\theta)(I-\eta\mathcal{H}-\eta\Delta_t)\mathbf{w}_{t-1},\end{aligned}$$

where $\Delta_t = \int_0^1 (\nabla^2 f(\phi\mathbf{y}_t + (1-\phi)\mathbf{y}_t') - \mathcal{H})\mathrm{d}\phi$. In the last step, we used

$$\nabla f(\mathbf{y}_t) - \nabla f(\mathbf{y}_t') = (\mathcal{H}+\Delta_t)(\mathbf{y}_t - \mathbf{y}_t') = (\mathcal{H}+\Delta_t)[(2-\theta)\mathbf{w}_t - (1-\theta)\mathbf{w}_{t-1}].$$

We thus obtain the update of the $\mathbf{w}_t$ sequence in matrix form:

$$\begin{aligned}\begin{pmatrix}\mathbf{w}_{t+1}\\ \mathbf{w}_t\end{pmatrix} =& \begin{pmatrix}(2-\theta)(\mathbf{I}-\eta\mathcal{H}) & -(1-\theta)(\mathbf{I}-\eta\mathcal{H})\\ \mathbf{I} & 0\end{pmatrix}\begin{pmatrix}\mathbf{w}_t\\ \mathbf{w}_{t-1}\end{pmatrix} \\ & - \eta\begin{pmatrix}(2-\theta)\Delta_t\mathbf{w}_t - (1-\theta)\Delta_t\mathbf{w}_{t-1}\\ 0\end{pmatrix} \\ =& \mathbf{A}\begin{pmatrix}\mathbf{w}_t\\ \mathbf{w}_{t-1}\end{pmatrix} - \eta\begin{pmatrix}\delta_t\\ 0\end{pmatrix} = \mathbf{A}^{t+1}\begin{pmatrix}\mathbf{w}_0\\ \mathbf{w}_{-1}\end{pmatrix} - \eta\sum_{\tau=0}^{t}\mathbf{A}^{t-\tau}\begin{pmatrix}\delta_\tau\\ 0\end{pmatrix},\end{aligned} \tag{15}$$

where $\delta_t = (2-\theta)\Delta_t\mathbf{w}_t - (1-\theta)\Delta_t\mathbf{w}_{t-1}$. Since $\mathbf{v}_0 = \mathbf{v}_0'$, we have $\mathbf{w}_{-1} = \mathbf{w}_0$, and $\|\Delta_t\| \le \rho\max\{\|\mathbf{x}_t - \tilde{\mathbf{x}}\|, \|\mathbf{x}_t' - \tilde{\mathbf{x}}\|\} \le 2\rho\mathscr{S}$, as well as $\|\delta_\tau\| \le 6\rho\mathscr{S}(\|\mathbf{w}_\tau\| + \|\mathbf{w}_{\tau-1}\|)$. According to (15):

$$\mathbf{w}_t = \begin{pmatrix}\mathbf{I} & 0\end{pmatrix}\mathbf{A}^t\begin{pmatrix}\mathbf{w}_0\\ \mathbf{w}_0\end{pmatrix} - \eta\begin{pmatrix}\mathbf{I} & 0\end{pmatrix}\sum_{\tau=0}^{t-1}\mathbf{A}^{t-1-\tau}\begin{pmatrix}\delta_\tau\\ 0\end{pmatrix}.$$

Intuitively, we want to say that the first term dominates. Technically, we will set up an induction based on the following fact:

$$\left\|\eta\begin{pmatrix}\mathbf{I}, 0\end{pmatrix}\sum_{\tau=0}^{t-1}\mathbf{A}^{t-1-\tau}\begin{pmatrix}\delta_\tau\\ 0\end{pmatrix}\right\| \le \frac{1}{2}\left\|\begin{pmatrix}\mathbf{I}, 0\end{pmatrix}\mathbf{A}^t\begin{pmatrix}\mathbf{w}_0\\ \mathbf{w}_0\end{pmatrix}\right\|.$$

It is easy to check the base case holds for $t = 0$. Then, assume that for all time steps less than or equal to t, the induction assumption hold. We have:

$$\begin{aligned}\|\mathbf{w}_t\| \le& \left\|\begin{pmatrix}\mathbf{I} & 0\end{pmatrix}\mathbf{A}^t\begin{pmatrix}\mathbf{w}_0\\ \mathbf{w}_0\end{pmatrix}\right\| + \left\|\eta\begin{pmatrix}\mathbf{I} & 0\end{pmatrix}\sum_{\tau=0}^{t-1}\mathbf{A}^{t-1-\tau}\begin{pmatrix}\delta_\tau\\ 0\end{pmatrix}\right\| \\ \le& 2\left\|\begin{pmatrix}\mathbf{I} & 0\end{pmatrix}\mathbf{A}^t\begin{pmatrix}\mathbf{w}_0\\ \mathbf{w}_0\end{pmatrix}\right\|,\end{aligned}$$

which gives:

$$\begin{aligned}\|\delta_t\| \leq & O(\rho\mathscr{S})(\|\mathbf{w}_t\| + \|\mathbf{w}_{t-1}\|) \leq O(\rho\mathscr{S}) \left[\left\|\begin{pmatrix}\mathbf{I} & 0\end{pmatrix}\mathbf{A}^t\begin{pmatrix}\mathbf{w}_0\\ \mathbf{w}_0\end{pmatrix}\right\| + \left\|\begin{pmatrix}\mathbf{I} & 0\end{pmatrix}\mathbf{A}^{t-1}\begin{pmatrix}\mathbf{w}_0\\ \mathbf{w}_0\end{pmatrix}\right\|\right] \\ \leq & O(\rho\mathscr{S})\left\|\begin{pmatrix}\mathbf{I} & 0\end{pmatrix}\mathbf{A}^t\begin{pmatrix}\mathbf{w}_0\\ \mathbf{w}_0\end{pmatrix}\right\|,\end{aligned}$$

where in the last inequality, we used Lemma 38 for monotonicity in t.

To prove that the induction assumption holds for $t+1$ we compute:

$$\begin{aligned}\left\|\eta\begin{pmatrix}\mathbf{I}, 0\end{pmatrix}\sum_{\tau=0}^{t}\mathbf{A}^{t-\tau}\begin{pmatrix}\delta_\tau\\ 0\end{pmatrix}\right\| \leq & \eta\sum_{\tau=0}^{t}\left\|\begin{pmatrix}\mathbf{I}, 0\end{pmatrix}\mathbf{A}^{t-\tau}\begin{pmatrix}\mathbf{I}\\ 0\end{pmatrix}\right\|\|\delta_\tau\| \\ \leq & O(\eta\rho\mathscr{S})\sum_{\tau=0}^{t}\left\|\begin{pmatrix}\mathbf{I}, 0\end{pmatrix}\mathbf{A}^{t-\tau}\begin{pmatrix}\mathbf{I}\\ 0\end{pmatrix}\right\|\left\|\begin{pmatrix}\mathbf{I} & 0\end{pmatrix}\mathbf{A}^{\tau}\begin{pmatrix}\mathbf{w}_0\\ \mathbf{w}_0\end{pmatrix}\right\|. \qquad (16)\end{aligned}$$

By the precondition we have $\lambda_{\min}(\mathcal{H}) \leq -\sqrt{\rho\epsilon}$. Without loss of generality, assume that the minimum eigenvector direction of $\mathcal{H}$ is along he first coordinate $\mathbf{e}_1$, and denote the corresponding 2×2 matrix as $\mathbf{A}_1$ (as in the convention of (12). Let:

$$(a_t^{(1)}, \ -b_t^{(1)}) = \begin{pmatrix}1 & 0\end{pmatrix}\mathbf{A}_1^t.$$

We then see that (1) $\mathbf{w}_0$ is along the $\mathbf{e}_1$ direction, and (2) according to Lemma 37, the matrix $\begin{pmatrix}\mathbf{I}, 0\end{pmatrix}\mathbf{A}^{t-\tau}\begin{pmatrix}\mathbf{I}\\ 0\end{pmatrix}$ is a diagonal matrix, where the spectral norm is achieved along the first coordinate which corresponds to the eigenvalue $\lambda_{\min}(\mathcal{H})$. Therefore, using Equation (16), we have:

$$\begin{aligned}\left\|\eta\begin{pmatrix}\mathbf{I}, 0\end{pmatrix}\sum_{\tau=0}^{t}\mathbf{A}^{t-\tau}\begin{pmatrix}\delta_\tau\\ 0\end{pmatrix}\right\| \leq & O(\eta\rho\mathscr{S})\sum_{\tau=0}^{t}a_{t-\tau}^{(1)}(a_\tau^{(1)} - b_\tau^{(1)})\|\mathbf{w}_0\| \\ \leq & O(\eta\rho\mathscr{S})\sum_{\tau=0}^{t}[\frac{2}{\theta} + (t+1)]|a_{t+1}^{(1)} - b_{t+1}^{(1)}|\,\|\mathbf{w}_0\| \\ \leq & O(\eta\rho\mathscr{S}\mathscr{T}^2)\left\|\begin{pmatrix}\mathbf{I}, 0\end{pmatrix}\mathbf{A}^{t+1}\begin{pmatrix}\mathbf{w}_0\\ \mathbf{w}_0\end{pmatrix}\right\|,\end{aligned}$$

where, in the second to last step, we used Lemma 36, and in the last step we used $1/\theta \leq \mathscr{T}$. Finally, $O(\eta\rho\mathscr{S}\mathscr{T}^2) \leq O(c^{-1}) \leq 1/2$ by choosing a sufficiently large constant c. Therefore, we have proved the induction, which gives us:

$$\|\mathbf{w}_t\| = \left\|\begin{pmatrix}\mathbf{I} & 0\end{pmatrix}\mathbf{A}^t\begin{pmatrix}\mathbf{w}_0\\ \mathbf{w}_0\end{pmatrix}\right\| - \left\|\eta\begin{pmatrix}\mathbf{I} & 0\end{pmatrix}\sum_{\tau=0}^{t-1}\mathbf{A}^{t-1-\tau}\begin{pmatrix}\delta_\tau\\ 0\end{pmatrix}\right\| \geq \frac{1}{2}\left\|\begin{pmatrix}\mathbf{I} & 0\end{pmatrix}\mathbf{A}^t\begin{pmatrix}\mathbf{w}_0\\ \mathbf{w}_0\end{pmatrix}\right\|.$$

Noting that $\lambda_{\min}(\mathcal{H}) \leq -\sqrt{\rho\epsilon}$, by applying Lemma 38 we have

$$\frac{1}{2}\left\|\begin{pmatrix}\mathbf{I} & 0\end{pmatrix}\mathbf{A}^t\begin{pmatrix}\mathbf{w}_0\\ \mathbf{w}_0\end{pmatrix}\right\| \geq \frac{\theta}{4}(1+\Omega(\theta))^t r_0,$$

which grows exponentially. Therefore, for $r_0 \geq \frac{\delta\mathscr{E}}{2\Delta_f} \cdot \frac{r}{\sqrt{d}}$, and $\mathscr{T} = \Omega(\frac{1}{\theta} \cdot \chi c)$ where $\chi = \max\{1, \log \frac{d\ell\Delta_f}{\rho\epsilon\delta}\}$, where the constant c is sufficiently large, we have

$$\left\|\mathbf{x}_{\mathscr{T}} - \mathbf{x}'_{\mathscr{T}}\right\| = \|\mathbf{w}_{\mathscr{T}}\| \geq \frac{\theta}{4}(1+\Omega(\theta))^{\mathscr{T}} r_0 \geq 4\mathscr{S},$$

which contradicts the fact that:

$$\forall t \leq \mathscr{T}, \max\{\|\mathbf{x}_t - \tilde{\mathbf{x}}\|, \left\|\mathbf{x}'_t - \tilde{\mathbf{x}}\right\|\} \leq O(\mathscr{S}).$$

This means our assumption is wrong, and we can therefore conclude:

$$\min\{E_{\mathscr{T}} - E_0, E'_{\mathscr{T}} - E'_0\} \leq -2\mathscr{E}.$$

On the other hand, by the precondition on $\tilde{x}$ and the gradient Lipschitz property, we have:

$$\max\{E_0 - \tilde{E}, E'_0 - \tilde{E}\} \leq \epsilon r + \frac{\ell r^2}{2} \leq \mathscr{E},$$

where the last step is due to our choice of $r = \eta\epsilon \cdot \chi^{-5} c^{-8}$ in (3). Combining these two facts:

$$\min\{E_{\mathscr{T}} - \tilde{E}, E'_{\mathscr{T}} - \tilde{E}\} \leq \min\{E_{\mathscr{T}} - E_0, E'_{\mathscr{T}} - E'_0\} + \max\{E_0 - \tilde{E}, E'_0 - \tilde{E}\} \leq -\mathscr{E},$$

which finishes the proof. ■

We are now ready to prove the main lemma in this subsection, which states with that random perturbation, PAGD will escape saddle points efficiently with high probability.

Lemma 13 (Negative curvature) *Consider the setting of Theorem 8. If $\|\nabla f(\mathbf{x}_0)\| \leq \epsilon$, $\lambda_{\min}(\nabla^2 f(\mathbf{x}_0)) < -\sqrt{\rho\epsilon}$, and a perturbation has not been added in iterations $\tau \in [-\mathscr{T}, 0)$, then, by running Algorithm 2, we have $E_{\mathscr{T}} - E_0 \leq -\mathscr{E}$ with probability at least $1 - \frac{\delta\mathscr{E}}{2\Delta_f}$.*

Proof Since a perturbation has not been added in iterations $\tau \in [-\mathscr{T}, 0)$, according to PAGD (Algorithm 2), we add perturbation at $t = 0$, the Hamiltonian will increase by at most:

$$\Delta E \leq \epsilon r + \frac{\ell r^2}{2} \leq \mathscr{E},$$

where the last step is due to our choice of $r = \eta\epsilon \cdot \chi^{-5} c^{-8}$ in (3) with constant c sufficiently large. Again by Algorithm 2, a perturbation will never be added in the remaining iterations, and by Lemma 9 and Lemma 17 we know the Hamiltonian always decreases for the remaining steps. Therefore, if at least one NCE step is performed in iteration $\tau \in [0, \mathscr{T}]$, by Lemma 17 we will decrease $2\mathscr{E}$ in that NCE step, and at most increase by $\mathscr{E}$ due to the perturbation. This immediately gives $E_{\mathscr{T}} - E_0 \leq -\mathscr{E}$.

Therefore, we only need to focus on the case where NCE is never used in iterations $\tau \in [0, \mathscr{T}]$. Let $\mathbb{B}_{\mathbf{x}_0}(r)$ denote the ball with radius r around $\mathbf{x}_0$. According to algorithm 2, we know the iterate after adding perturbation to $\mathbf{x}_0$ is uniformly sampled from the ball $\mathbb{B}_{\mathbf{x}_0}(r)$. Let $\mathcal{X}_{\text{stuck}} \subset \mathbb{B}_{\mathbf{x}_0}(r)$ be the region where AGD is stuck (does not decrease the Hamiltonian $\mathscr{E}$ in $\mathscr{T}$ steps). Formally, for any point $\mathbf{x} \in \mathcal{X}_{\text{stuck}}$, let $\mathbf{x}_1, \cdots, \mathbf{x}_{\mathscr{T}}$ be the AGD sequence starting at $(\mathbf{x}, \mathbf{v}_0)$, then $E_{\mathscr{T}} - E_0 \geq -\mathscr{E}$. By

Lemma 23, $\mathcal{X}_{\text{stuck}}$ can have at most width $r_0 = \frac{\delta\mathscr{E}}{2\Delta_f} \cdot \frac{r}{\sqrt{d}}$ along the minimum eigenvalue direction. Therefore,

$$\frac{\text{Vol}(\mathcal{X}_{\text{stuck}})}{\text{Vol}(\mathbb{B}^{(d)}_{\mathbf{x}_0}(r))} \leq \frac{r_0 \times \text{Vol}(\mathbb{B}^{(d-1)}_0(r))}{\text{Vol}(\mathbb{B}^{(d)}_0(r))} = \frac{r_0}{r\sqrt{\pi}} \frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d}{2}+\frac{1}{2})} \leq \frac{r_0}{r\sqrt{\pi}} \cdot \sqrt{\frac{d}{2}+\frac{1}{2}} \leq \frac{\delta\mathscr{E}}{2\Delta_f}.$$

Thus, with probability at least $1 - \frac{\delta\mathscr{E}}{\Delta_f}$, the perturbation will end up outside of $\mathcal{X}_{\text{stuck}}$, which give $E_{\mathscr{T}} - E_0 \leq -\mathscr{E}$. This finishes the proof.

■

B.4. Proof of Theorem 8

Our main result is now easily obtained from Lemma 12 and Lemma 13.

Proof [Proof of Theorem 8] Suppose we never encounter any ϵ-second-order stationary point. Consider the set $\mathfrak{T} = \{\tau | \tau \in [0, \mathscr{T}] \text{ and } \|\nabla f(\mathbf{x}_\tau)\| \leq \epsilon\}$, and two cases: (1) $\mathfrak{T} = \varnothing$, in which case we know all gradients are large and by Lemma 12 we have $E_{\mathscr{T}} - E_0 \leq -\mathscr{E}$; (2) $\mathfrak{T} \neq \varnothing$. In this case, define $\tau' = \min \mathfrak{T}$; i.e., the earliest iteration where the gradient is small. Since by assumption, $\mathbf{x}'_\tau$ is not an ϵ-second-order stationary point, this gives $\nabla^2 f(\mathbf{x}_{\tau'}) \leq -\sqrt{\rho\epsilon}$, and by Lemma 13, we can conclude $E_{\tau'+\mathscr{T}} - E_0 \leq E_{\tau'+\mathscr{T}} - E_{\tau'} \leq -\mathscr{E}$. Clearly $\tau' + \mathscr{T} \leq 2\mathscr{T}$. That is, in either case, we will decrease the Hamiltonian by $\mathscr{E}$ in at most $2\mathscr{T}$ steps.

Then, for the the first case, we can repeat this argument starting at iteration $\mathscr{T}$, and for the second case, we can repeat the argument starting at iteration $\tau' + \mathscr{T}$. Therefore, we will continue to obtain a decrease of the Hamiltonian by an average of $\mathscr{E}/(2\mathscr{T})$ per step. Since the function f is lower bounded, we know the Hamiltonian can not decrease beyond $E_0 - E^\star = f(\mathbf{x}_0) - f^\star$, which means that in $\frac{2(f(\mathbf{x}_0)-f^\star)\mathscr{T}}{\mathscr{E}}$ steps, we must encounter an ϵ-second-order stationary point at least once.

Finally, in $\frac{2(f(\mathbf{x}_0)-f^\star)\mathscr{T}}{\mathscr{E}}$ steps, we will call Lemma 13 at most $\frac{2\Delta_f}{\mathscr{E}}$ times, and since Lemma 13 holds with probability $1 - \frac{\delta\mathscr{E}}{2\Delta_f}$, by a union bound, we know that the argument above is true with probability at least:

$$1 - \frac{\delta\mathscr{E}}{2\Delta_f} \cdot \frac{2\Delta_f}{\mathscr{E}} = 1 - \delta,$$

which finishes the proof. ■

Appendix C. Auxiliary Lemma

In this section, we present some auxiliary lemmas which are used in proving Lemma 21, Lemma 22 and Lemma 23. These deal with the large-gradient scenario (nonconvex component), the large-gradient scenario (strongly convex component), and the negative curvature scenario, respectively.

The first two lemmas establish some facts about powers of the structured matrices arising in AGD.

Lemma 24 *Let the* 2×2 *matrix* $\mathbf{A}$ *have following form, for arbitrary* $a, b \in \mathbb{R}$*:*

$$\mathbf{A} = \begin{pmatrix} a & b \\ 1 & 0 \end{pmatrix}.$$

Letting μ_1, μ_2 denote the two eigenvalues of $\mathbf{A}$ (can be repeated or complex eigenvalues), then, for any $t \in \mathbb{N}$:

$$\begin{pmatrix} 1 & 0 \end{pmatrix} \mathbf{A}^t = \left(\sum_{i=0}^{t} \mu_1^i \mu_2^{t-i}, \quad -\mu_1\mu_2 \sum_{i=0}^{t-1} \mu_1^i \mu_2^{t-1-i} \right)$$
$$\begin{pmatrix} 0 & 1 \end{pmatrix} \mathbf{A}^t = \begin{pmatrix} 1 & 0 \end{pmatrix} \mathbf{A}^{t-1}.$$

Proof When the eigenvalues μ_1 and μ_2 are distinct, the matrix $\mathbf{A}$ can be rewritten as $\begin{pmatrix} \mu_1+\mu_2 & -\mu_1\mu_2 \\ 1 & 0 \end{pmatrix}$, and it is easy to check that the two eigenvectors have the form $\begin{pmatrix} \mu_1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} \mu_2 \\ 1 \end{pmatrix}$. Therefore, we can write the eigen-decomposition as:

$$\mathbf{A} = \frac{1}{\mu_1 - \mu_2} \begin{pmatrix} \mu_1 & \mu_2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{pmatrix} \begin{pmatrix} 1 & -\mu_2 \\ -1 & \mu_1 \end{pmatrix},$$

and the tth power has the general form:

$$\mathbf{A}^t = \frac{1}{\mu_1 - \mu_2} \begin{pmatrix} \mu_1 & \mu_2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \mu_1^t & 0 \\ 0 & \mu_2^t \end{pmatrix} \begin{pmatrix} 1 & -\mu_2 \\ -1 & \mu_1 \end{pmatrix}$$

When there are two repeated eigenvalue μ_1, the matrix $\begin{pmatrix} a & b \\ 1 & 0 \end{pmatrix}$ can be rewritten as $\begin{pmatrix} 2\mu_1 & -\mu_1^2 \\ 1 & 0 \end{pmatrix}$. It is easy to check that $\mathbf{A}$ has the following Jordan normal form:

$$\mathbf{A} = -\begin{pmatrix} \mu_1 & \mu_1+1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 & 1 \\ 0 & \mu_1 \end{pmatrix} \begin{pmatrix} 1 & -(\mu_1+1) \\ -1 & \mu_1 \end{pmatrix},$$

which yields:

$$\mathbf{A}^t = -\begin{pmatrix} \mu_1 & \mu_1+1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \mu_1^t & t\mu_1^{t-1} \\ 0 & \mu_1^t \end{pmatrix} \begin{pmatrix} 1 & -(\mu_1+1) \\ -1 & \mu_1 \end{pmatrix}.$$

The remainder of the proof follows from simple linear algebra calculations for both cases. ■

Lemma 25 *Under the same setting as Lemma 24, for any $t \in \mathbb{N}$:*

$$(\mu_1 - 1)(\mu_2 - 1) \begin{pmatrix} 1 & 0 \end{pmatrix} \sum_{\tau=0}^{t-1} \mathbf{A}^\tau \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 1 - \begin{pmatrix} 1 & 0 \end{pmatrix} \mathbf{A}^t \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Proof When μ_1 and μ_2 are distinct, we have:

$$\begin{pmatrix} 1 & 0 \end{pmatrix} \mathbf{A}^t = \left(\frac{\mu_1^{t+1} - \mu_2^{t+1}}{\mu_1 - \mu_2}, \quad -\frac{\mu_1\mu_2(\mu_1^t - \mu_2^t)}{\mu_1 - \mu_2} \right).$$

When μ_1, μ_2 are repeated, we have:

$$\begin{pmatrix} 1 & 0 \end{pmatrix} \mathbf{A}^t = \left((t+1)\mu_1^t, \quad -t\mu_1^{t+1} \right).$$

The remainder of the proof follows from Lemma 27 and linear algebra. ■

The next lemma tells us when the eigenvalues of the AGD matrix are real and when they are complex.

Lemma 26 *Let $\theta \in (0, \frac{1}{4}]$, $\mathbf{x} \in [-\frac{1}{4}, \frac{1}{4}]$ and define the 2×2 matrix $\mathbf{A}$ as follows:*

$$\mathbf{A} = \begin{pmatrix} (2-\theta)(1-x) & -(1-\theta)(1-x) \\ 1 & 0 \end{pmatrix}$$

Then the two eigenvalues μ_1 and μ_2 of $\mathbf{A}$ are solutions of the following equation:

$$\mu^2 - (2-\theta)(1-x)\mu + (1-\theta)(1-x) = 0.$$

Moreover, when $x \in [-\frac{1}{4}, \frac{\theta^2}{(2-\theta)^2}]$, μ_1 and μ_2 are real numbers, and when $x \in (\frac{\theta^2}{(2-\theta)^2}, \frac{1}{4}]$, μ_1 and μ_2 are conjugate complex numbers.

Proof An eigenvalue μ of the matrix $\mathbf{A}$ must satisfy the following equation:

$$\det(\mathbf{A} - \mu\mathbf{I}) = \mu^2 - (2-\theta)(1-x)\mu + (1-\theta)(1-x) = 0.$$

The discriminant is equal to

$$\begin{aligned} \Delta =& (2-\theta)^2(1-x)^2 - 4(1-\theta)(1-x) \\ =& (1-x)(\theta^2 - (2-\theta^2)x). \end{aligned}$$

Then μ_1 and μ_2 are real if and only if $\Delta \geq 0$, which finishes the proof. ■

Finally, we need a simple lemma for geometric sums.

Lemma 27 *For any $\lambda > 0$ and fixed t, we have:*

$$\sum_{\tau=0}^{t-1} (\tau+1)\lambda^\tau = \frac{1-\lambda^t}{(1-\lambda)^2} - \frac{t\lambda^t}{1-\lambda}.$$

Proof Consider the truncated geometric series:

$$\sum_{\tau=0}^{t-1} \lambda^\tau = \frac{1-\lambda^t}{1-\lambda}.$$

Taking derivatives, we have:

$$\sum_{\tau=0}^{t-1} (\tau+1)\lambda^\tau = \frac{\mathrm{d}}{\mathrm{d}\lambda} \sum_{\tau=0}^{t-1} \lambda^{\tau+1} = \frac{\mathrm{d}}{\mathrm{d}\lambda} \left[\lambda \cdot \frac{1-\lambda^t}{1-\lambda} \right] = \frac{1-\lambda^t}{(1-\lambda)^2} - \frac{t\lambda^t}{1-\lambda}.$$

■

C.1. Large-gradient scenario (nonconvex component)

All the lemmas in this section are concerned with the behavior of the AGD matrix for eigen-directions of the Hessian with eigenvalues being negative or small and positive, as used in proving Lemma 21. The following lemma bounds the smallest eigenvalue of the AGD matrix for those directions.

Lemma 28 *Under the same setting as Lemma 26, and for $x \in [-\frac{1}{4}, \frac{\theta^2}{(2-\theta)^2}]$, where $\mu_1 \geq \mu_2$, we have:*

$$\mu_2 \leq 1 - \frac{1}{2}\max\{\theta, \sqrt{|x|}\}.$$

Proof The eigenvalues satisfy:

$$\det(\mathbf{A} - \mu\mathbf{I}) = \mu^2 - (2-\theta)(1-x)\mu + (1-\theta)(1-x) = 0.$$

Let $\mu = 1 + u$. We have

$$\begin{aligned} & (1+u)^2 - (2-\theta)(1-x)(1+u) + (1-\theta)(1-x) = 0 \\ \Rightarrow \quad & u^2 + ((1-x)\theta + 2x)u + x = 0. \end{aligned}$$

Let $f(u) = u^2 + \theta u + 2xu - x\theta u + x$. To prove $\mu_2(\mathbf{A}) \leq 1 - \frac{\sqrt{|x|}}{2}$ when $x \in [-\frac{1}{4}, -\theta^2]$, we only need to verify $f(-\frac{\sqrt{|x|}}{2}) \leq 0$:

$$\begin{aligned} f(-\frac{\sqrt{|x|}}{2}) =& \frac{|x|}{4} - \frac{\theta\sqrt{|x|}}{2} + |x|\sqrt{|x|} - \frac{|x|\sqrt{|x|}\theta}{2} - |x| \\ \leq& |x|\sqrt{|x|}(1 - \frac{\theta}{2}) - \frac{3|x|}{4} \leq 0 \end{aligned}$$

The last inequality follows because $|x| \leq \frac{1}{4}$ by assumption.

For $x \in [-\theta^2, 0]$, we have:

$$f(-\frac{\theta}{2}) = \frac{\theta^2}{4} - \frac{\theta^2}{2} - x\theta + \frac{x\theta^2}{2} + x = -\frac{\theta^2}{4} + x(1-\theta) + \frac{x\theta^2}{2} \leq 0.$$

On the other hand, when $x \in [0, \theta^2/(2-\theta)^2]$, both eigenvalues are still real, and the midpoint of the two roots is:

$$\frac{u_1 + u_2}{2} = -\frac{(1-x)\theta + 2x}{2} = -\frac{\theta + (2-\theta)x}{2} \leq -\frac{\theta}{2}.$$

Combining the two cases, we have shown that when $x \in [-\theta^2, \theta^2/(2-\theta)^2]$ we have $\mu_2(\mathbf{A}) \leq 1 - \frac{\theta}{2}$.

In summary, we have proved that

$$\mu_2(\mathbf{A}) \leq \begin{cases} 1 - \frac{\sqrt{|x|}}{2}, & x \in [-\frac{1}{4}, -\theta^2] \\ 1 - \frac{\theta}{2}. & x \in [-\theta^2, \theta^2/(2-\theta)^2], \end{cases}$$

which finishes the proof. ■

In the same setting as above, the following lemma bounds the largest eigenvalue.

Lemma 29 *Under the same setting as Lemma 26, and with $x \in [-\frac{1}{4}, \frac{\theta^2}{(2-\theta)^2}]$, and letting $\mu_1 \geq \mu_2$, we have:*

$$\mu_1 \leq 1 + 2\min\{\frac{|x|}{\theta}, \sqrt{|x|}\}.$$

Proof By Lemma 26 and Vieta's formula we have:

$$(\mu_1 - 1)(\mu_2 - 1) = \mu_1\mu_2 - (\mu_1 + \mu_2) + 1 = x.$$

An application of Lemma 28 finishes the proof. ■

The following lemma establishes some properties of the powers of the AGD matrix.

Lemma 30 *Consider the same setting as Lemma 26, and let $x \in [-\frac{1}{4}, \frac{\theta^2}{(2-\theta)^2}]$. Denote:*

$$(a_t, \ -b_t) = \begin{pmatrix} 1 & 0 \end{pmatrix} \mathbf{A}^t.$$

Then, for any $t \geq \frac{2}{\theta} + 1$, we have:

$$\begin{aligned} \sum_{\tau=0}^{t-1} a_\tau \geq & \Omega(\frac{1}{\theta^2}) \\ \frac{1}{b_t}\left(\sum_{\tau=0}^{t-1} a_\tau\right) \geq & \Omega(1)\min\left\{\frac{1}{\theta}, \frac{1}{\sqrt{|x|}}\right\}. \end{aligned}$$

Proof We prove the two inequalities seperately.
First Inequality: By Lemma 24:

$$\begin{aligned} \sum_{\tau=0}^{t} \begin{pmatrix} 1 & 0 \end{pmatrix} \mathbf{A}^\tau \begin{pmatrix} 1 \\ 0 \end{pmatrix} = & \sum_{\tau=0}^{t}\sum_{i=0}^{\tau} \mu_1^{\tau-i}\mu_2^i = \sum_{\tau=0}^{t}(\mu_1\mu_2)^{\frac{\tau}{2}} \sum_{i=0}^{\tau}(\frac{\mu_1}{\mu_2})^{\frac{\tau}{2}-i} \\ \geq & \sum_{\tau=0}^{t}[(1-\theta)(1-x)]^{\frac{\tau}{2}} \cdot \frac{\tau}{2} \end{aligned}$$

The last inequality holds because in $\sum_{i=0}^{\tau}(\frac{\mu_1}{\mu_2})^{\frac{\tau}{2}-i}$ at least $\frac{\tau}{2}$ terms are greater than one. Finally, since $x \leq \theta^2/(2-\theta)^2 \leq \theta^2 \leq \theta$, we have $1 - x \geq 1 - \theta$, thus:

$$\begin{aligned} \sum_{\tau=0}^{t}[(1-\theta)(1-x)]^{\frac{\tau}{2}} \cdot \frac{\tau}{2} \geq & \sum_{\tau=0}^{t}(1-\theta)^\tau \cdot \frac{\tau}{2} \geq \sum_{\tau=0}^{1/\theta}(1-\theta)^\tau \cdot \frac{\tau}{2} \\ \geq & (1-\theta)^{\frac{1}{\theta}} \sum_{\tau=0}^{1/\theta} \frac{\tau}{2} \geq \Omega(\frac{1}{\theta^2}), \end{aligned}$$

which finishes the proof.

Second Inequality: Without loss of generality, assume $\mu_1 \geq \mu_2$. Again by Lemma 24:

$$\begin{aligned}\frac{\sum_{\tau=0}^{t-1} a_\tau}{b_t} =& \frac{\sum_{\tau=0}^{t-1}\sum_{i=0}^{\tau}\mu_1^i\mu_2^{\tau-i}}{\mu_1\mu_2\sum_{i=0}^{t-1}\mu_1^i\mu_2^{t-1-i}} = \frac{1}{\mu_1\mu_2}\sum_{\tau=0}^{t-1}\frac{\sum_{i=0}^{\tau}\mu_1^i\mu_2^{\tau-i}}{\sum_{i=0}^{t-1}\mu_1^i\mu_2^{t-1-i}} \\ \geq& \frac{1}{\mu_1\mu_2}\sum_{\tau=(t-1)/2}^{t-1}\frac{\sum_{i=0}^{\tau}\mu_1^i\mu_2^{\tau-i}}{\sum_{i=0}^{t-1}\mu_1^i\mu_2^{t-1-i}} \geq \frac{1}{\mu_1\mu_2}\sum_{\tau=(t-1)/2}^{t-1}\frac{1}{2\mu_1^{t-1-\tau}} \\ =& \frac{1}{2\mu_1\mu_2}\left[1+\frac{1}{\mu_1}+\cdots+\frac{1}{\mu_1^{(t-1)/2}}\right] \geq \frac{1}{2\mu_1\mu_2}\left[1+\frac{1}{\mu_1}+\cdots+\frac{1}{\mu_1^{1/\theta}}\right].\end{aligned}$$

The second-to-last inequality holds because it is easy to check

$$2\mu_1^{t-1-\tau}\sum_{i=0}^{\tau}\mu_1^i\mu_2^{\tau-i} \geq \sum_{i=0}^{t-1}\mu_1^i\mu_2^{t-1-i},$$

for any $\tau \geq (t-1)/2$. Finally, by Lemma 29, we have

$$\mu_1 \leq 1 + 2\min\{\frac{|x|}{\theta}, \sqrt{|x|}\}.$$

Since $\mu_1 = \Theta(1)$, $\mu_2 = \Theta(1)$, we have that when $|x| \leq \theta^2$,

$$\frac{\sum_{\tau=0}^{t-1} a_\tau}{b_t} \geq \Omega(1)\left[1+\frac{1}{\mu_1}+\cdots+\frac{1}{\mu_1^{1/\theta}}\right] \geq \Omega(1)\cdot\frac{1}{\theta}\cdot\frac{1}{(1+\theta)^{\frac{1}{\theta}}} \geq \Omega(\frac{1}{\theta}).$$

When $|x| > \theta^2$, we have:

$$\frac{\sum_{\tau=0}^{t-1} a_\tau}{b_t} \geq \Omega(1)\left[1+\frac{1}{\mu_1}+\cdots+\frac{1}{\mu_1^{1/\theta}}\right] = \Omega(1)\frac{1-\frac{1}{\mu_1^{1/\theta+1}}}{1-\frac{1}{\mu_1}} = \Omega(\frac{1}{\mu_1-1}) = \Omega(\frac{1}{\sqrt{|x|}}).$$

Combining the two cases finishes the proof. ■

C.2. Large-gradient scenario (strongly convex component)

All the lemmas in this section are concerned with the behavior of the AGD matrix for eigen-directions of the Hessian with eigenvalues being large and positive, as used in proving Lemma 22. The following lemma gives eigenvalues of the AGD matrix for those directions.

Lemma 31 *Under the same setting as Lemma 26, and with $x \in (\frac{\theta^2}{(2-\theta)^2}, \frac{1}{4}]$, we have $\mu_1 = re^{i\phi}$ and $\mu_2 = re^{-i\phi}$, where:*

$$r = \sqrt{(1-\theta)(1-x)}, \qquad \sin\phi = \sqrt{((2-\theta)^2x-\theta^2)(1-x)}/2r.$$

Proof By Lemma 26, we know that μ_1 and μ_2 are two solutions of

$$\mu^2 - (2-\theta)(1-x)\mu + (1-\theta)(1-x) = 0.$$

This gives $r^2 = \mu_1\mu_2 = (1-\theta)(1-x)$. On the other hand, discriminant is equal to

$$\begin{aligned}\Delta =& (2-\theta)^2(1-x)^2 - 4(1-\theta)(1-x)\\ =& (1-x)(\theta^2 - (2-\theta^2)x).\end{aligned}$$

Since $\mathrm{Im}(\mu_1) = r\sin\phi = \frac{\sqrt{-\Delta}}{2}$, the proof is finished. ■

Under the same setting as above, the following lemma delineates some properties of powers of the AGD matrix.

Lemma 32 *Under the same setting as in Lemma 26, and with $x \in (\frac{\theta^2}{(2-\theta)^2}, \frac{1}{4}]$, denote:*

$$(a_t, \ -b_t) = \begin{pmatrix}1 & 0\end{pmatrix}\mathbf{A}^t.$$

Then, for any $t \geq 0$, we have:

$$\max\{|a_t|, \ |b_t|\} \leq (t+1)(1-\theta)^{\frac{t}{2}}.$$

Proof By Lemma 24 and Lemma 31, using $|\cdot|$ to denote the magnitude of a complex number, we have:

$$\begin{aligned}|a_t| &= \left|\sum_{i=0}^{t}\mu_1^i\mu_2^{t-i}\right| \leq \sum_{i=0}^{t}|\mu_1^i\mu_2^{t-i}| = (t+1)r^t \leq (t+1)(1-\theta)^{\frac{t}{2}}\\ |b_t| &= \left|\mu_1\mu_2\sum_{i=0}^{t-1}\mu_1^i\mu_2^{t-1-i}\right| \leq \sum_{i=0}^{t-1}|\mu_1^{i+1}\mu_2^{t-i}| \leq tr^{t+1} \leq t(1-\theta)^{\frac{t+1}{2}}.\end{aligned}$$

Reorganizing these two equations finishes the proof. ■

The following is a technical lemma which is useful in bounding the change in the Hessian by the amount of oscillation in the iterates.

Lemma 33 *Under the same setting as Lemma 31, for any $T \geq 0$, any sequence $\{\epsilon_t\}$, and any $\varphi_0 \in [0, 2\pi]$:*

$$\sum_{t=0}^{T} r^t\sin(\phi t + \varphi_0)\epsilon_t \leq O(\frac{1}{\sin\phi})\left(|\epsilon_0| + \sum_{t=1}^{T}|\epsilon_t - \epsilon_{t-1}|\right).$$

Proof Let $\tau = \lfloor 2\pi/\phi \rfloor$ be the approximate period, and $J = \lfloor T/\tau \rfloor$ be the number of periods that exist within time T. Then, we can group the summation by each period:

$$\begin{aligned}\sum_{t=0}^{T} r^t \sin(\phi t)\epsilon_t =& \sum_{j=0}^{J}\left[\sum_{t=j\tau}^{\min\{(j+1)\tau-1,T\}} r^t \sin(\phi t+\varphi_0)\epsilon_t\right]\\ =& \sum_{j=0}^{J}\left[\sum_{t=j\tau}^{\min\{(j+1)\tau-1,T\}} r^t \sin(\phi t+\varphi_0)[\epsilon_{j\tau}+(\epsilon_t-\epsilon_{j\tau})]\right]\\ \le& \underbrace{\sum_{j=0}^{J}\left[\sum_{t=j\tau}^{\min\{(j+1)\tau-1,T\}} r^t \sin(\phi t+\varphi_0)\right]\epsilon_{j\tau}}_{\text{Term 1}} + \underbrace{\sum_{j=0}^{J}\left[\sum_{t=j\tau}^{\min\{(j+1)\tau-1,T\}} r^t|\epsilon_t-\epsilon_{j\tau}|\right]}_{\text{Term 2}}.\end{aligned}$$

We prove the lemma by bounding the first term and the second term on the right-hand-side of this equation separately.

Term 2: Since $r \le 1$, it is not hard to see:

$$\begin{aligned}\text{Term 2} =& \sum_{j=0}^{J}\left[\sum_{t=j\tau}^{\min\{(j+1)\tau-1,T\}} r^t|\epsilon_t-\epsilon_{j\tau}|\right]\\ \le& \sum_{j=0}^{J}\left[\sum_{t=j\tau}^{\min\{(j+1)\tau-1,T\}} r^t\right]\left[\sum_{t=j\tau+1}^{\min\{(j+1)\tau-1,T\}} |\epsilon_t-\epsilon_{t-1}|\right]\\ \le& \tau\sum_{j=0}^{J}\left[\sum_{t=j\tau+1}^{\min\{(j+1)\tau-1,T\}} |\epsilon_t-\epsilon_{t-1}|\right] \le \tau\sum_{t=1}^{T}|\epsilon_t-\epsilon_{t-1}|.\end{aligned}$$

Term 1: We first study the inner-loop factor, $\sum_{t=j\tau}^{(j+1)\tau-1} r^t \sin(\phi t)$. Letting $\psi = 2\pi - \tau\phi$ be the offset for each approximate period, we have that for any $j < J$:

$$\begin{aligned}\left|\sum_{t=j\tau}^{(j+1)\tau-1} r^t \sin(\phi t+\varphi_0)\right| =& \left|\mathrm{Im}\left[\sum_{t=0}^{\tau-1} r^{j\tau+t} e^{i\cdot[\phi(j\tau+t)+\varphi_0]}\right]\right|\\ \le& r^{j\tau}\left\|\sum_{t=0}^{\tau-1} r^t e^{i\cdot\phi t}\right\| \le r^{j\tau}\left\|\frac{1-r^\tau e^{i\cdot(2\pi-\psi)}}{1-re^{i\cdot\phi}}\right\|\\ =& r^{j\tau}\sqrt{\frac{(1-r^\tau\cos\psi)^2+(r^\tau\sin\psi)^2}{(1-r\cos\phi)^2+(r\sin\phi)^2}}.\end{aligned}$$

Combined with the fact that for all $y \in [0,1]$ we have $e^{-3y} \le 1-y \le e^{-y}$, we obtain the following:

$$1-r^\tau = 1-[(1-\theta)(1-x)]^{\frac{\tau}{2}} = 1-e^{-\Theta((\theta+x)\tau)} = \Theta((\theta+x)\tau) = \Theta\left(\frac{(\theta+x)}{\phi}\right) \qquad (17)$$

Also, for any $a, b \in [0,1]$, we have $(1-ab)^2 \le (1-\min\{a,b\})^2 \le (1-a^2)^2 + (1-b^2)^2$, and by definition of τ, we immediately have $\psi \le \phi$. This yields:

$$\begin{aligned}\frac{(1-r^\tau\cos\psi)^2+(r^\tau\sin\psi)^2}{(1-r\cos\phi)^2+(r\sin\phi)^2} \le& \frac{2(1-r^{2\tau})^2+2(1-\cos^2\psi)^2+(r^\tau\sin\psi)^2}{(r\sin\phi)^2}\\ \le& O\left(\frac{1}{\sin^2\phi}\right)\left[\frac{(\theta+x)^2}{\phi^2}+\sin^4\phi+\sin^2\phi\right] \le O\left(\frac{(\theta+x)^2}{\sin^4\phi}\right)\end{aligned}$$

The second last inequality used the fact that $r = \Theta(1)$ (although note r^τ is not $\Theta(1)$). The last inequality is true since by Lemma 31, we know $(\theta+x)/\sin^2\phi \ge \Omega(1)$. This gives:

$$\left|\sum_{t=j\tau}^{(j+1)\tau-1} r^t \sin(\phi t+\varphi_0)\right| \le r^{j\tau}\cdot\frac{\theta+x}{\sin^2\phi},$$

and therefore, we can now bound the first term:

$$\begin{aligned}\text{Term 1} =& \sum_{j=0}^{J}\sum_{t=j\tau}^{\min\{(j+1)\tau-1,T\}} r^t\sin(\phi t+\varphi_0)\epsilon_{j\tau} = \sum_{j=0}^{J}\left[\sum_{t=j\tau}^{\min\{(j+1)\tau-1,T\}} r^t\sin(\phi t+\varphi_0)\right](\epsilon_0+\epsilon_{j\tau}-\epsilon_0)\\ \le& O(1)\sum_{j=0}^{J-1}\left[r^{j\tau}\frac{\theta+x}{\sin^2\phi}\right](|\epsilon_0|+|\epsilon_{j\tau}-\epsilon_0|)+\sum_{t=J\tau}^{T}(|\epsilon_0|+|\epsilon_{J\tau}-\epsilon_0|)\\ \le& O(1)\left[\frac{1}{1-r^\tau}\frac{\theta+x}{\sin^2\phi}+\tau\right]\cdot\left[|\epsilon_0|+\sum_{t=1}^{T}|\epsilon_t-\epsilon_{t-1}|\right] \le \left[O(\frac{1}{\sin\phi})+\tau\right]\cdot\left[|\epsilon_0|+\sum_{t=1}^{T}|\epsilon_t-\epsilon_{t-1}|\right].\end{aligned}$$

The second-to-last inequality used Eq.(17). In conclusion, since $\tau \le \frac{2\pi}{\phi} \le \frac{2\pi}{\sin\phi}$, we have:

$$\begin{aligned}\sum_{t=0}^{T} r^t\sin(\phi t+\varphi_0)\epsilon_t \le& \text{Term 1} + \text{Term 2} \le \left[O(\frac{1}{\sin\phi})+2\tau\right]\cdot\left[|\epsilon_0|+\sum_{t=1}^{T}|\epsilon_t-\epsilon_{t-1}|\right]\\ \le& O\left(\frac{1}{\sin\phi}\right)\left[|\epsilon_0|+\sum_{t=1}^{T}|\epsilon_t-\epsilon_{t-1}|\right].\end{aligned}$$

■

The following lemma combines the previous two lemmas to bound the approximation error in the quadratic.

Lemma 34 *Under the same setting as Lemma 26, and with $x \in (\frac{\theta^2}{(2-\theta)^2}, \frac{1}{4}]$, denote:*

$$(a_t,\ -b_t) = \begin{pmatrix}1 & 0\end{pmatrix}\mathbf{A}^t.$$

Then, for any sequence $\{\epsilon_\tau\}$, any $t \ge \Omega(\frac{1}{\theta})$, we have:

$$\begin{aligned}\sum_{\tau=0}^{t-1} a_\tau\epsilon_\tau \le& O(\frac{1}{x})\left(|\epsilon_0|+\sum_{\tau=1}^{t-1}|\epsilon_\tau-\epsilon_{\tau-1}|\right)\\ \sum_{\tau=0}^{t-1}(a_\tau-a_{\tau-1})\epsilon_\tau \le& O(\frac{1}{\sqrt{x}})\left(|\epsilon_0|+\sum_{\tau=1}^{t-1}|\epsilon_\tau-\epsilon_{\tau-1}|\right).\end{aligned}$$

Proof We prove the two inequalities separately.
First Inequality: Since $x \in (\frac{\theta^2}{(2-\theta)^2}, \frac{1}{4}]$, we further split the analysis into two cases:
Case $x \in (\frac{\theta^2}{(2-\theta)^2}, \frac{2\theta^2}{(2-\theta)^2}]$**:** By Lemma 24, we can expand dthe left-hand-side as:

$$\sum_{\tau=0}^{t-1} a_\tau \epsilon_\tau \le \sum_{\tau=0}^{t-1} |a_\tau|(|\epsilon_0| + |\epsilon_\tau - \epsilon_0|) \le \left[\sum_{\tau=0}^{t-1} |a_\tau|\right] \left(|\epsilon_0| + \sum_{\tau=1}^{t-1} |\epsilon_\tau - \epsilon_{\tau-1}|\right).$$

Noting that in this case $x = \Theta(\theta^2)$, by Lemma 32 and Lemma 27, we have for $t \ge O(1/\theta)$:

$$\sum_{\tau=0}^{t-1} |a_\tau| \le \sum_{\tau=0}^{t-1} (\tau+1)(1-\theta)^{\frac{\tau}{2}} \le O(\frac{1}{\theta^2}) = O(\frac{1}{x}).$$

Case $x \in (\frac{2\theta^2}{(2-\theta)^2}, \frac{1}{4}]$**:** Again, we expand the left-hand-side as:

$$\sum_{\tau=0}^{t-1} a_\tau \epsilon_\tau = \sum_{\tau=0}^{t-1} \frac{\mu_1^{\tau+1} - \mu_2^{\tau+1}}{\mu_1 - \mu_2} \epsilon_\tau = \sum_{\tau=0}^{t-1} \frac{r^{\tau+1} \sin[(\tau+1)\phi]}{r \sin[\phi]} \epsilon_\tau.$$

Noting in this case that $x = \Theta(\sin^2 \phi)$ by Lemma 31, then by Lemma 33 we have:

$$\sum_{\tau=0}^{t-1} a_\tau \epsilon_\tau \le O(\frac{1}{\sin^2 \phi}) \left(|\epsilon_0| + \sum_{\tau=1}^{t-1} |\epsilon_\tau - \epsilon_{\tau-1}|\right) \le O(\frac{1}{x}) \left(|\epsilon_0| + \sum_{\tau=1}^{t-1} |\epsilon_\tau - \epsilon_{\tau-1}|\right).$$

Second Inequality: Using Lemma 24, we know:

$$\begin{aligned} a_\tau - a_{\tau-1} =& \frac{(\mu_1^{\tau+1} - \mu_2^{\tau+1}) - (\mu_1^\tau - \mu_2^\tau)}{\mu_1 - \mu_2} \\ =& \frac{r^{\tau+1} \sin[(\tau+1)\phi] - r^\tau \sin[\tau\phi]}{r \sin[\phi]} \\ =& \frac{r^\tau \sin[\tau\phi](r\cos\phi - 1) + r^{\tau+1} \cos[\tau\phi] \sin\phi}{r \sin[\phi]} \\ =& \frac{r\cos\phi - 1}{r \sin\phi} \cdot r^\tau \sin[\tau\phi] + r^\tau \cos[\tau\phi], \end{aligned}$$

where we note $r = \Theta(1)$ and the coefficient of the first term is upper bounded by the following:

$$\left|\frac{r\cos\phi - 1}{r\sin\phi}\right| \le \frac{(1-\cos^2\phi) + (1-r^2)}{r\sin\phi} \le O\left(\frac{\theta + x}{\sin\phi}\right).$$

As in the proof of the first inequality, we split the analysis into two cases:
Case $x \in (\frac{\theta^2}{(2-\theta)^2}, \frac{2\theta^2}{(2-\theta)^2}]$**:** Again, we use

$$\sum_{\tau=0}^{t-1} (a_\tau - a_{\tau-1})\epsilon_\tau \le \sum_{\tau=0}^{t-1} |a_\tau - a_{\tau-1}|(|\epsilon_0| + |\epsilon_\tau - \epsilon_0|) \le \left[\sum_{\tau=0}^{t-1} |a_\tau - a_{\tau-1}|\right] \left(|\epsilon_0| + \sum_{\tau=1}^{t-1} |\epsilon_\tau - \epsilon_{\tau-1}|\right).$$

Noting $x = \Theta(\theta^2)$, again by Lemma 27 and $|\frac{\sin \tau\phi}{\sin \phi}| \leq \tau$, we have:

$$\left[\sum_{\tau=0}^{t-1} |a_\tau - a_{\tau-1}|\right] \leq O(\theta + x)\sum_{\tau=0}^{t-1} \tau(1-\theta)^{\frac{\tau}{2}} + \sum_{\tau=0}^{t-1}(1-\theta)^{\frac{\tau}{2}} \leq O(\frac{1}{\theta}) = O(\frac{1}{\sqrt{x}}).$$

Case $x \in (\frac{2\theta^2}{(2-\theta)^2}, \frac{1}{4}]$**:** From the above derivation, we have:

$$\sum_{\tau=0}^{t-1}(a_\tau - a_{\tau-1})\epsilon_\tau = \frac{r\cos\phi - 1}{r\sin\phi}\sum_{\tau=0}^{t-1} r^\tau \sin[\tau\phi]\epsilon_\tau + \sum_{\tau=0}^{t-1} r^\tau \cos[\tau\phi]\epsilon_\tau.$$

According to Lemma 31, in this case $x = \Theta(\sin^2\phi)$, $r = \Theta(1)$ and since $\Omega(\theta^2) \leq x \leq O(1)$, we have:

$$\left|\frac{r\cos\phi - 1}{r\sin\phi}\right| \leq O\left(\frac{\theta + x}{\sin\phi}\right) \leq O\left(\frac{\theta + x}{\sqrt{x}}\right) \leq O(1).$$

Combined with Lemma 33, this gives:

$$\sum_{\tau=0}^{t-1}(a_\tau - a_{\tau-1})\epsilon_\tau \leq O(\frac{1}{\sin\phi})\left(|\epsilon_0| + \sum_{\tau=1}^{t-1}|\epsilon_\tau - \epsilon_{\tau-1}|\right) \leq O(\frac{1}{\sqrt{x}})\left(|\epsilon_0| + \sum_{\tau=1}^{t-1}|\epsilon_\tau - \epsilon_{\tau-1}|\right).$$

Putting all the pieces together finishes the proof. ■

C.3. Negative-curvature scenario

In this section, we will prove the auxiliary lemmas required for proving Lemma 23.

The first lemma lower bounds the largest eigenvalue of the AGD matrix for eigen-directions whose eigenvalues are negative.

Lemma 35 *Under the same setting as Lemma 26, and with $x \in [-\frac{1}{4}, 0]$, and $\mu_1 \geq \mu_2$, we have:*

$$\mu_1 \geq 1 + \frac{1}{2}\min\{\frac{|x|}{\theta}, \sqrt{|x|}\}.$$

Proof The eigenvalues satisfy:

$$\det(\mathbf{A} - \mu\mathbf{I}) = \mu^2 - (2-\theta)(1-x)\mu + (1-\theta)(1-x) = 0.$$

Let $\mu = 1 + u$. We have

$$\begin{aligned} & (1+u)^2 - (2-\theta)(1-x)(1+u) + (1-\theta)(1-x) = 0 \\ \Rightarrow \quad & u^2 + ((1-x)\theta + 2x)u + x = 0. \end{aligned}$$

Let $f(u) = u^2 + \theta u + 2xu - x\theta u + x$. To prove $\mu_1(\mathbf{A}) \geq 1 + \frac{\sqrt{|x|}}{2}$ when $x \in [-\frac{1}{4}, -\theta^2]$, we only need to verify $f(\frac{\sqrt{|x|}}{2}) \leq 0$:

$$\begin{aligned} f(\frac{\sqrt{|x|}}{2}) =& \frac{|x|}{4} + \frac{\theta\sqrt{|x|}}{2} - |x|\sqrt{|x|} + \frac{|x|\sqrt{|x|}\theta}{2} - |x| \\ \leq& \frac{\theta\sqrt{|x|}}{2} - \frac{3|x|}{4} - |x|\sqrt{|x|}(1 - \frac{\theta}{2}) \leq 0 \end{aligned}$$

The last inequality holds because $\theta \le \sqrt{|x|}$ in this case.

For $x \in [-\theta^2, 0]$, we have:

$$f(\frac{|x|}{2\theta}) = \frac{|x|^2}{4\theta^2} + \frac{|x|}{2} - \frac{|x|^2}{\theta} + \frac{|x|^2}{2} - |x| = \frac{|x|^2}{4\theta^2} - \frac{|x|}{2} - |x|^2(\frac{1}{\theta} - \frac{1}{2}) \le 0,$$

where the last inequality is due to $\theta^2 \ge |x|$.

In summary, we have proved

$$\mu_1(\mathbf{A}) \ge \begin{cases} 1 + \frac{\sqrt{|x|}}{2}, & x \in [-\frac{1}{4}, -\theta^2] \\ 1 + \frac{|x|}{2\theta}. & x \in [-\theta^2, 0], \end{cases}$$

which finishes the proof. ■

The next lemma is a technical lemma on large powers.

Lemma 36 *Under the same setting as Lemma 26, and with $x \in [-\frac{1}{4}, 0]$, denote*

$$(a_t, \ -b_t) = \begin{pmatrix} 1 & 0 \end{pmatrix} \mathbf{A}^t.$$

Then, for any $0 \le \tau \le t$, we have

$$|a_{t-\tau}^{(1)}||a_\tau^{(1)} - b_\tau^{(1)}| \le [\frac{2}{\theta} + (t+1)]|a_{t+1}^{(1)} - b_{t+1}^{(1)}|.$$

Proof Let μ_1 and μ_2 be the two eigenvalues of the matrix $\mathbf{A}$, where $\mu_1 \ge \mu_2$. Since $x \in [-\frac{1}{4}, 0]$, according to Lemma 26 and Lemma 28, we have $0 \le \mu_2 \le 1 - \frac{\theta}{2} \le 1 \le \mu_1$, and thus expanding both sides using Lemma 24 yields:

$$\begin{aligned}
\text{LHS} =& \left[\sum_{i=0}^{t-\tau} \mu_1^{t-\tau-i}\mu_2^i\right] \left[(1-\mu_2)\left(\sum_{i=0}^{\tau-1} \mu_1^{\tau-i}\mu_2^i\right) + \mu_2^\tau\right] \\
=& \left[\sum_{i=0}^{t-\tau} \mu_1^{t-\tau-i}\mu_2^i\right](1-\mu_2)\left(\sum_{i=0}^{\tau-1} \mu_1^{\tau-i}\mu_2^i\right) + \left[\sum_{i=0}^{t-\tau} \mu_1^{t-\tau-i}\mu_2^i\right]\mu_2^\tau \\
\le& (t-\tau+1)\mu_1^{t-\tau}(1-\mu_2)\left(\sum_{i=0}^{\tau-1} \mu_1^{\tau-i}\mu_2^i\right) + \left[\sum_{i=0}^{t-\tau} \mu_1^{t-\tau-i}\mu_2^i\right] \\
\le& (t+1)(1-\mu_2)\left(\sum_{i=0}^{\tau-1} \mu_1^{t+1-i}\mu_2^i\right) + \frac{2}{\theta}(1-\mu_2)\left[\sum_{i=0}^{t-\tau} \mu_1^{t+1-i}\mu_2^i\right] \\
\le& [\frac{2}{\theta} + (t+1)]\left[(1-\mu_2)\sum_{i=0}^{t} \mu_1^{t+1-i}\mu_2^i + \mu_2^{t+1}\right] = \text{RHS},
\end{aligned}$$

which finishes the proof. ■

The following lemma gives properties of the $(1,1)$ element of large powers of the AGD matrix.

Lemma 37 *Let the 2×2 matrix $\mathbf{A}(x)$ be defined as follows and let $x \in [-\frac{1}{4}, 0]$ and $\theta \in (0, \frac{1}{4}]$.*

$$\mathbf{A}(x) = \begin{pmatrix} (2-\theta)(1-x) & -(1-\theta)(1-x) \\ 1 & 0 \end{pmatrix}.$$

For any fixed $t > 0$, letting $g(x) = \left| \begin{pmatrix} 1 & 0 \end{pmatrix} [\mathbf{A}(x)]^t \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right|$, then we have:

1. *$g(x)$ is a monotonically decreasing function for $x \in [-1, \theta^2/(2-\theta)^2]$.*
2. *For any $x \in [\theta^2/(2-\theta)^2, 1]$, we have $g(x) \leq g(\theta^2/(2-\theta)^2)$.*

Proof For $x \in [-1, \theta^2/(2-\theta)^2]$, we know that $\mathbf{A}(x)$ has two real eigenvalues $\mu_1(x)$ and $\mu_2(x)$, Without loss of generality, we can assume $\mu_1(x) \geq \mu_2(x)$. By Lemma 24, we know:

$$g(x) = \left| \begin{pmatrix} 1 & 0 \end{pmatrix} [\mathbf{A}(x)]^t \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right| = \sum_{i=0}^{t} [\mu_1(x)]^i [\mu_2(x)]^{t-i} = [\mu_1(x)\mu_2(x)]^{\frac{t}{2}} \sum_{i=0}^{t} \left[\frac{\mu_1(x)}{\mu_2(x)}\right]^{\frac{t}{2}-i}.$$

By Lemma 26 and Vieta's formulas, we know that $[\mu_1(x)\mu_2(x)]^{\frac{t}{2}} = [(1-\theta)(1-x)]^{\frac{t}{2}}$ is monotonically decreasing in x. On the other hand, we have that:

$$\frac{\mu_1(x)}{\mu_2(x)} + \frac{\mu_2(x)}{\mu_1(x)} + 2 = \frac{[\mu_1(x) + \mu_2(x)]^2}{\mu_1(x)\mu_2(x)} = \frac{(2-\theta)^2(1-x)}{1-\theta}$$

is monotonically decreasing in x, implying that $\sum_{i=0}^{t} \left[\frac{\mu_1(x)}{\mu_2(x)}\right]^{\frac{t}{2}-i}$ is monotonically decreasing in x. Since both terms are positive, this implies the product is also monotonically decreasing in x, which finishes the proof of the first part.

For $x \in [\theta^2/(2-\theta)^2, 1]$, the two eigenvalues $\mu_1(x)$ and $\mu_2(x)$ are conjugate, and we have:

$$[\mu_1(x)\mu_2(x)]^{\frac{t}{2}} = [(1-\theta)(1-x)]^{\frac{t}{2}} \leq [\mu_1(\theta^2/(2-\theta)^2)\mu_2(\theta^2/(2-\theta)^2)]^{\frac{t}{2}}$$

which yields:

$$\sum_{i=0}^{t} \left[\frac{\mu_1(x)}{\mu_2(x)}\right]^{\frac{t}{2}-i} \leq \left\| \sum_{i=0}^{t} \left[\frac{\mu_1(x)}{\mu_2(x)}\right]^{\frac{t}{2}-i} \right\| \leq \sum_{i=0}^{t} \left\| \frac{\mu_1(x)}{\mu_2(x)} \right\|^{\frac{t}{2}-i} = t+1 = \sum_{i=0}^{t} \left[\frac{\mu_1(\theta^2/(2-\theta)^2)}{\mu_2(\theta^2/(2-\theta)^2)}\right]^{\frac{t}{2}-i},$$

and this finishes the proof of the second part. ■

The following lemma gives properties of the sum of the first row of large powers of the AGD matrix.

Lemma 38 *Under the same setting as Lemma 26, and with $x \in [-\frac{1}{4}, 0]$, denote*

$$(a_t, \ -b_t) = \begin{pmatrix} 1 & 0 \end{pmatrix} \mathbf{A}^t.$$

Then we have

$$|a_{t+1} - b_{t+1}| \geq |a_t - b_t|$$

and

$$|a_t - b_t| \geq \frac{\theta}{2} \left(1 + \frac{1}{2} \min\{\frac{|x|}{\theta}, \sqrt{|x|}\}\right)^t.$$

Proof Since $x < 0$, we know that $\mathbf{A}$ has two distinct real eigenvalues. Let μ_1 and μ_2 be the two eigenvalues of $\mathbf{A}$. For the first inequality, by Lemma 24, we only need to prove:

$$\mu_1^{t+1} - \mu_2^{t+1} - \mu_1\mu_2(\mu_1^t - \mu_2^t) \geq \mu_1^t - \mu_2^t - \mu_1\mu_2(\mu_1^{t-1} - \mu_2^{t-1}).$$

Taking the difference of the LHS and RHS, we have:

$$\begin{aligned}
&\mu_1^{t+1} - \mu_2^{t+1} - \mu_1\mu_2(\mu_1^t - \mu_2^t) - (\mu_1^t - \mu_2^t) + \mu_1\mu_2(\mu_1^{t-1} - \mu_2^{t-1}) \\
=&\mu_1^t(\mu_1 - \mu_1\mu_2 - 1 + \mu_2) - \mu_2^t(\mu_2 - \mu_1\mu_2 - 1 + \mu_1) \\
=&(\mu_1^t - \mu_2^t)(\mu_1 - 1)(1 - \mu_2).
\end{aligned}$$

According to Lemma 26 and Lemma 28, $\mu_1 \geq 1 \geq \mu_2 \geq 0$, which finishes the proof of the first claim.

For the second inequality, again by Lemma 24, since both μ_1 and μ_2 are positive, we have:

$$a_t - b_t = \sum_{i=0}^{t} \mu_1^i \mu_2^{t-i} - \mu_1\mu_2 \sum_{i=0}^{t-1} \mu_1^i \mu_2^{t-1-i} \geq (1 - \mu_2) \sum_{i=0}^{t} \mu_1^i \mu_2^{t-i} \geq (1 - \mu_2)\mu_1^t.$$

By Lemma 28 we have $1 - \mu_2 \geq \frac{\theta}{2}$, By Lemma 35 we know $\mu_1 \geq 1 + \frac{1}{2}\min\{\frac{|x|}{\theta}, \sqrt{|x|}\}$. Combining these facts finishes the proof. ■