MULTISCALE ANALYSIS OF ACCELERATED GRADIENT METHODS[*]

MOHAMMAD FARAZMAND[†]

Abstract. Accelerated gradient descent iterations are widely used in optimization. It is known that, in the continuous-time limit, these iterations converge to a second-order differential equation which we refer to as the accelerated gradient flow. Using geometric singular perturbation theory, we show that, under certain conditions, the accelerated gradient flow possesses an attracting invariant slow manifold to which the trajectories of the flow converge asymptotically. We obtain a general explicit expression in the form of functional series expansions that approximates the slow manifold to any arbitrary order of accuracy. To the leading order, the accelerated gradient flow reduced to this slow manifold coincides with the usual gradient descent. We illustrate the implications of our results on three examples.

Key words. convex optimization, accelerated gradient methods, invariant manifolds, singular perturbation theory

AMS subject classifications. 49-XX, 34-XX, 65-XX

DOI. 10.1137/18M1203997

1. Introduction. We consider the convex optimization problem $\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$ where $f : \mathcal{X} \to \mathbb{R}$ is a convex function of class C^r (i.e., f is r-time continuously differentiable) and $\mathcal{X} \subseteq \mathbb{R}^n$ is a compact convex set. We also assume that f attains its unique minimum at a point $\mathbf{x}^* \in \mathcal{X}$.

The most well-known method for obtaining the minimum $\mathbf{x}^*$ is the gradient (or steepest) descent iterations

$$(1.1) \qquad \mathbf{x}_{k+1} = \mathbf{x}_k - s \nabla f(\mathbf{x}_k),$$

where $s > 0$ is a parameter. Under certain assumptions, the sequence $\{\mathbf{x}_k\}$ is guaranteed to converge to the minimum $\mathbf{x}^*$ [1]. It is also well known that in the limit $s \to 0$, the discrete iterations (1.1) converge to the continuous-time gradient flow $\dot{\mathbf{x}} = -\nabla f(\mathbf{x})$, which is a first-order ordinary differential equation (ODE). In fact, the gradient descent iterations are an explicit Euler discretization of the gradient flow [16, 3].

The convergence rate of the gradient descent iterations is $\mathcal{O}(1/k)$, which is rather slow [13]. In order to improve this convergence rate, a number of accelerated gradient descent iterations have been developed [14]. Most notable perhaps is Nesterov's accelerated gradient descent,

$$(1.2) \qquad \mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k(\mathbf{x}_k - \mathbf{x}_{k-1}) - s_k \nabla f(\mathbf{x}_k + \lambda_k(\mathbf{x}_k - \mathbf{x}_{k-1})),$$

where $\{\lambda_k\}$ and $\{s_k\}$ are prescribed sequences of positive real numbers [12]. The convergence rate of this accelerated gradient descent is $\mathcal{O}(1/k^2)$. The terms involving λ_k are referred to as the acceleration terms. In practice, λ_k and s_k are sometimes assumed to be constants independent of k. For $\lambda_k = 0$, we recover the steepest descent iterations (1.1).

[†]Department of Mathematics, North Carolina State University, Raleigh, NC 27695 (farazmand @ncsu.edu).

Su, Boyd, and Candès [17] discovered that, in a small step size limit, the Nesterov iterations converge to the second-order differential equation

$$(1.3) \qquad \ddot{\mathbf{x}} + \frac{\rho}{t}\dot{\mathbf{x}} + \nabla f(\mathbf{x}) = 0,$$

where $\rho = 3$. We refer to this differential equation as the *Nesterov flow*. Recall that the gradient descent iterations converge, in the continuous-time limit, to the autonomous first-order differential equation $\dot{\mathbf{x}} = -\nabla f(\mathbf{x})$, while the Nesterov accelerated gradient descent converges to the nonautonomous second-order differential equation (1.3). (It is nonautonomous because of the explicit dependent of the coefficient ρ/t on time t.)

More recently, Wibisono, Wilson, and Jordan [18] showed that a large class of accelerated gradient methods can be formulated as temporal discretizations of a second-order differential equation. In their framework, the continuous-time accelerated gradient methods coincide with the Euler–Lagrange equations corresponding to a single Lagrangian,

$$(1.4) \qquad \mathcal{L}(\mathbf{x}, \dot{\mathbf{x}}, t) = e^{\alpha_t + \gamma_t}\left(D_h(\mathbf{x} + e^{-\alpha_t}\dot{\mathbf{x}}, \mathbf{x}) - e^{\beta_t}f(\mathbf{x})\right),$$

which we refer to as the *Bregman Lagrangian*. Here, α_t, β_t, and γ_t are potentially time-dependent scalar functions, and the Bregman divergence D_h is defined as

$$(1.5) \qquad D_h(\mathbf{y}, \mathbf{x}) = h(\mathbf{y}) - h(\mathbf{x}) - \langle \nabla h(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the usual Euclidean inner product, and the *distance-generating function* $h : \mathcal{X} \to \mathbb{R}$ is convex and continuously differentiable.

Wibisono, Wilson, and Jordan [18] assume the so-called ideal scaling, $\dot{\gamma}_t = e^{\alpha_t}$, and show that the Euler–Lagrange equation corresponding to the Bregman Lagrangian (1.4) reads

$$(1.6) \qquad \ddot{\mathbf{x}} + (e^{\alpha_t} - \dot{\alpha}_t)\dot{\mathbf{x}} + e^{2\alpha_t + \beta_t}\left[\nabla^2 h(\mathbf{x} + e^{-\alpha_t}\dot{\mathbf{x}})\right]^{-1}\nabla f(\mathbf{x}) = 0.$$

For the special choice $\alpha_t = \log((\rho - 1)/t)$, $\beta_t = -2\log((\rho - 1)/t)$, and $h(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$, the Euler–Lagrange equation (1.6) coincides with (1.3).

Wibisono, Wilson, and Jordan [18] also show that, for certain functions α_t and β_t, the trajectories of (1.6) converge to the minimum $\mathbf{x}^*$ asymptotically. Furthermore, they recover several well-known accelerated gradient methods by discretizing (1.6) in time in an appropriate fashion.

In this paper, we investigate the multiscale behavior of the accelerated gradient flows and their phase space structure using geometric singular perturbation theory. We divide our analysis into two parts: autonomous and nonautonomous cases. In the autonomous case (section 2), the choice of the functions α_t and β_t is restricted such that the Euler–Lagrange equation (1.6) does not have an explicit dependence on time. In this case, we show that the solutions of the Euler–Lagrange equation (1.6) converge exponentially quickly towards an n-dimensional, attracting, slow manifold embedded in the $2n$-dimensional phase space of the Euler–Lagrange equation. We derive explicit formulas for the slow manifold in terms of a functional series expansion. To the leading order, the Euler–Lagrange equation, reduced to the slow manifold, coincides with the usual gradient descent. We also investigate the reduced flow at higher orders and prove that the minimum $\mathbf{x}^*$ is a locally asymptotic stable fixed point of the reduced flow at any order.

The Nesterov accelerated gradient flow (1.3) is nonautonomous because of the explicit time dependence of the coefficient ρ/t. We treat this nonautonomous case

separately (section 3) using a nonautonomous extension of the singular perturbation theory. While the results are similar to the autonomous accelerated gradient flow, here the slow manifold is an $(n + 1)$-dimensional invariant manifold in the $(2n + 1)$-dimensional *extended* phase space of the flow.

The paper is organized as follows. Section 2 contains our main results, including the slow manifold reduction in Euclidean and non-Euclidean spaces in the autonomous case. Section 3 treats the nonautonomous case of (1.3). In section 4, we demonstrate the implications of our results on three examples. Section 5 contains our concluding remarks.

2. Multiscale analysis: Autonomous case.

2.1. Euclidean setting. In this section, we consider the Euclidean case where the distance-generating function in the Bregman Lagrangian (1.4) is given by $h(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$. In this case, the Lagrangian reads

$$(2.1) \qquad \mathcal{L}(\mathbf{x}, \dot{\mathbf{x}}, t) = e^{\alpha_t + \gamma_t} \left(\frac{1}{2}\|e^{-\alpha_t}\dot{\mathbf{x}}\|^2 - e^{\beta_t} f(\mathbf{x}) \right).$$

The corresponding Euler–Lagrange equation reads

$$(2.2) \qquad \ddot{\mathbf{x}} + (e^{\alpha_t} - \dot{\alpha}_t)\dot{\mathbf{x}} + e^{2\alpha_t + \beta_t}\nabla f(\mathbf{x}) = 0.$$

The following theorem, which constitutes our main result, states that under certain assumptions the trajectories of the Euler–Lagrange equation (2.2) converge asymptotically to an n-dimensional invariant slow manifold embedded in the $2n$-dimensional phase space $(\mathbf{x}, \dot{\mathbf{x}})$. This slow manifold is a graph over the $\mathbf{x}$ coordinates. We obtain an explicit expression for this graph in the form of a functional series expansion. Our results use Fenichel's geometric singular perturbation theory [4]. There are various statements of this theory with different extents of generality [19, 8, 9]. A statement of the geometric singular perturbation theory that most closely fits our purposes can be found in Chapter 3 (and Theorem 3.1.4) of [9].

THEOREM 1. *Let $f \in C^{r+1}(\mathcal{X})$ with $r \geq 0$. Define $\mu = e^{\alpha_t} - \dot{\alpha}_t$ and $\eta = e^{2\alpha_t + \beta_t}$. If μ and η are constant, then there exists $\mu_0 > 0$ such that for all $\mu > \mu_0$ the following are true.*

(i) *The trajectories of the Euler–Lagrange equation* (2.2) *converge exponentially quickly to an n-dimensional invariant manifold embedded in the $2n$-dimensional phase space $(\mathbf{x}, \dot{\mathbf{x}})$. Furthermore, this invariant manifold is a graph over the $\mathbf{x}$ coordinates (See Figure* 2.1 *for an illustration).*

(ii) *The flow of* (2.2) *on the invariant manifold $\mathcal{M}_\epsilon$ is given by*

$$(2.3) \qquad \dot{\mathbf{x}} = \sum_{k=0}^{r} \epsilon^{2k+1}\mathbf{g}_{2k+1}(\mathbf{x}) + \mathcal{O}(\epsilon^{2r+3}),$$

where $\epsilon = \mu^{-1}$ and the maps $\mathbf{g}_k : \mathcal{X} \to \mathbb{R}^n$ are defined recursively by

$$\mathbf{g}_1(\mathbf{x}) = -\eta\nabla f(\mathbf{x}),$$
$$\mathbf{g}_{2k}(\mathbf{x}) = \mathbf{0},$$
$$(2.4) \qquad \mathbf{g}_{2k+1}(\mathbf{x}) = -\sum_{\ell=1}^{2k-1} \nabla\mathbf{g}_\ell(\mathbf{x})\mathbf{g}_{2k-\ell}(\mathbf{x}), \quad k \geq 1.$$
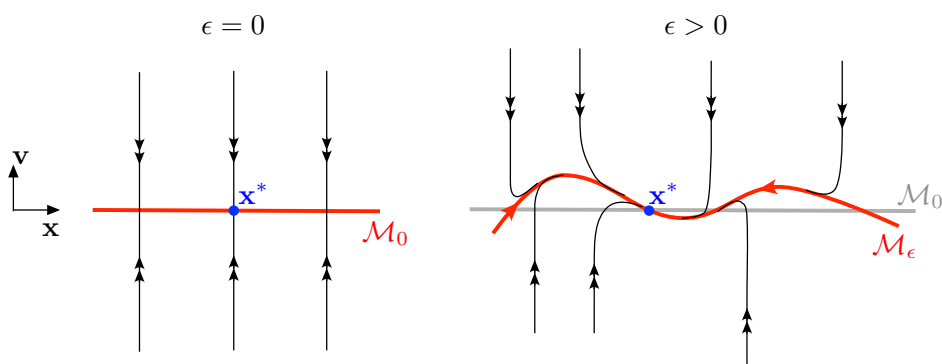
FIG. 2.1. *A sketch of the geometric singular perturbation theory. In the singular limit, $\epsilon = 0$, the set $\mathcal{M}_0 = \{(\mathbf{x}, \mathbf{v}) \in \mathbb{R}^{2n} : \mathbf{v} = 0\}$ is filled with fixed points. The trajectories $(\mathbf{x}(\tau), \mathbf{v}(\tau)) = (\mathbf{x}_0, \mathbf{v}_0 e^{-\tau})$ converge exponentially quickly to $(\mathbf{x}_0, \mathbf{0}) \in \mathcal{M}_0$ from any initial condition $(\mathbf{x}_0, \mathbf{v}_0)$. In other words, the set $\mathcal{M}_0$ is invariant and the global attractor of the system. For $0 < \epsilon \ll 1$, the manifold $\mathcal{M}_0$ deforms into a nearby manifold $\mathcal{M}_\epsilon$ which is also invariant and globally attracting. However, $\mathcal{M}_\epsilon$ is not necessarily a collection of fixed points.*

Before presenting the proof of this theorem, we make a few remarks.

Remark 1. The conditions requiring $\mu = e^{\alpha_t} - \dot{\alpha}_t$ and $\eta = e^{2\alpha_t + \beta_t}$ to be constant are equivalent to

$$(2.5) \qquad \alpha_t = \log\left[\frac{\mu}{1 + (\mu e^{-\alpha_0} - 1)\, e^{\mu t}}\right], \quad \beta_t = \log \eta - 2\alpha_t,$$

where $\alpha_0, \eta \in \mathbb{R}$ and $\mu > 0$ are arbitrary constants satisfying $\alpha_0 \leq \log \mu$. For instance, the class of accelerated gradient flows considered in equation (7) of [20] and in equation (4) of [7] satisfy these conditions. We point out that these conditions are sufficient for the results of Theorem 1 to hold but are not necessary and can possibly be relaxed.

Remark 2. Note that the minimum $\mathbf{x}^*$ is a fixed point of (2.3) since $\nabla f(\mathbf{x}^*) = \mathbf{g}_k(\mathbf{x}^*) = \mathbf{0}$ for all $k \in \mathbb{N}$.

Remark 3. Polyak's heavy ball iterations are a particular discretization of the ODE (1.6) with constant $\mu = e^{\alpha_t} - \dot{\alpha}_t$ and $\eta = e^{2\alpha_t + \beta_t}$ [14]. The convergence of Polyak's heavy ball iterations is only guaranteed if f is strongly convex [11]. However, the fixed point $(\mathbf{x}, \dot{\mathbf{x}}) = (\mathbf{x}^*, \mathbf{0})$ is a globally asymptotically stable fixed point of the ODE (2.2) for all $\mu > 0$ and C^1 convex functions f. One can verify this by considering the Lyapunov function

$$(2.6) \qquad \mathcal{E}(\mathbf{x}, \dot{\mathbf{x}}) = \eta\left[f(\mathbf{x}) - f(\mathbf{x}^*)\right] + \frac{1}{2}|\dot{\mathbf{x}}|^2.$$

Note that $\mathcal{E}(\mathbf{x}, \dot{\mathbf{x}}) > 0$ for all $(\mathbf{x}, \dot{\mathbf{x}}) \in \mathcal{X} \times \mathbb{R}^n \setminus (\mathbf{x}^*, \mathbf{0})$ and $\frac{d}{dt}\mathcal{E} = -\mu|\dot{\mathbf{x}}|^2 \leq 0$. By LaSalle's invariant set theorem [10], and the fact that $\mathcal{E}(\mathbf{x}, \dot{\mathbf{x}}) \to \infty$ as $|(\mathbf{x}, \dot{\mathbf{x}})| \to \infty$, the fixed point $(\mathbf{x}^*, \mathbf{0})$ is globally asymptotically stable. The fact that Polyak's iterations fail to converge for certain convex functions is an artifact of the discretization.

Remark 4. To the leading order (neglecting $\mathcal{O}(\epsilon^3)$ terms), (2.3) reduces to the usual gradient flow

$$(2.7) \qquad \dot{\mathbf{x}} = -\epsilon \eta \nabla f(\mathbf{x}),$$

with a rescaling of time with the constant $\epsilon\eta$. This connection to the gradient flow was already speculated in an appendix of [18] through a heuristic argument.

Proof of Theorem 1. We first rewrite (2.2) as a system of first-order differential equations by introducing a new variable $\mathbf{v} = \dot{\mathbf{x}}$, so that

$$
\begin{aligned}
\dot{\mathbf{x}} &= \mathbf{v}, \\
\dot{\mathbf{v}} &= -\mu\mathbf{v} - \eta\nabla f(\mathbf{x}).
\end{aligned}
\tag{2.8}
$$

Next, we introduce the fast time $\tau = t/\epsilon$ where $\epsilon = \mu^{-1} \ll 1$. In terms of the fast time τ, (2.8) can be written as

$$
\begin{aligned}
\mathbf{x}' &= \epsilon\mathbf{v}, \\
\mathbf{v}' &= -\mathbf{v} - \epsilon\eta\nabla f(\mathbf{x}),
\end{aligned}
\tag{2.9}
$$

where prime denotes the derivative with respect to the fast time τ, e.g., $\mathbf{x}' = d\mathbf{x}/d\tau$.

In the singular limit, $\epsilon = 0$, (2.9) has the trivial solution $\mathbf{x}(\tau) = \mathbf{x}_0$ and $\mathbf{v}(\tau) = \mathbf{v}_0 e^{-\tau}$, where $(\mathbf{x}_0, \mathbf{v}_0)$ is the initial condition. The trajectories of the system converge exponentially quickly towards the *critical manifold*

$$
\mathcal{M}_0 = \{(\mathbf{x}, \mathbf{v}) \in \mathcal{X} \times \mathbb{R}^n : \mathbf{v} = \mathbf{0}\}.
\tag{2.10}
$$

By Fenichel's geometric singular perturbation theory [4], there exists $\epsilon_0 > 0$ such that for $0 < \epsilon < \epsilon_0$ the following hold: (i) The trajectories of (2.9) also converge to an n-dimensional invariant manifold $\mathcal{M}_\epsilon$. (ii) Furthermore, $\mathcal{M}_\epsilon$ is a C^r-smooth graph over $\mathcal{M}_0$ and $\mathcal{O}(\epsilon)$ close to it. More specifically, $\mathcal{M}_\epsilon$ can be expressed through the formal series expansion

$$
\mathcal{M}_\epsilon = \left\{(\mathbf{x}, \mathbf{v}) \in \mathcal{X} \times \mathbb{R}^n : \mathbf{v} = \sum_{k=1}^{\infty} \epsilon^k \mathbf{g}_k(\mathbf{x})\right\},
\tag{2.11}
$$

where $\mathbf{g}_k : \mathbb{R}^n \to \mathbb{R}^n$ are smooth functions to be determined. As a result, we have

$$
\begin{aligned}
\mathbf{v}' &= \sum_{k=1}^{\infty} \epsilon^k \nabla\mathbf{g}_k(\mathbf{x})\mathbf{x}' \\
&= \sum_{k=1}^{\infty} \epsilon^{k+1} \nabla\mathbf{g}_k(\mathbf{x})\mathbf{v} \\
&= \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} \epsilon^{k+j+1} \nabla\mathbf{g}_k(\mathbf{x})\mathbf{g}_j(\mathbf{x}).
\end{aligned}
\tag{2.12}
$$

On the other hand, (2.9) implies

$$
\begin{aligned}
\mathbf{v}' &= -\mathbf{v} - \epsilon\eta\nabla f(\mathbf{x}) \\
&= -\epsilon\left[\mathbf{g}_1(\mathbf{x}) + \eta\nabla f(\mathbf{x})\right] - \sum_{k=2}^{\infty} \epsilon^k \mathbf{g}_k(\mathbf{x}).
\end{aligned}
\tag{2.13}
$$

Matching the terms of the same order ϵ^k in (2.12) and (2.13), we obtain

$$\epsilon^1 : \mathbf{g}_1(\mathbf{x}) = -\eta\nabla f(\mathbf{x}),$$
$$\epsilon^2 : \mathbf{g}_2(\mathbf{x}) = \mathbf{0},$$
$$\epsilon^3 : \mathbf{g}_3(\mathbf{x}) = -\nabla\mathbf{g}_1(\mathbf{x})\mathbf{g}_1(\mathbf{x}),$$
$$\epsilon^4 : \mathbf{g}_4(\mathbf{x}) = \mathbf{0},$$
$$\epsilon^5 : \mathbf{g}_5(\mathbf{x}) = -\nabla\mathbf{g}_3(\mathbf{x})\mathbf{g}_1(\mathbf{x}) - \nabla\mathbf{g}_1(\mathbf{x})\mathbf{g}_3(\mathbf{x}),$$
$$\epsilon^6 : \mathbf{g}_6(\mathbf{x}) = \mathbf{0},$$
$$\epsilon^7 : \mathbf{g}_7(\mathbf{x}) = -\nabla\mathbf{g}_5(\mathbf{x})\mathbf{g}_1(\mathbf{x}) - \nabla\mathbf{g}_3(\mathbf{x})\mathbf{g}_3(\mathbf{x}) - \nabla\mathbf{g}_1(\mathbf{x})\mathbf{g}_5(\mathbf{x})$$
$$\vdots$$

$$(2.14) \qquad \epsilon^k : \mathbf{g}_k(\mathbf{x}) = -\sum_{\ell=1}^{k-2} \nabla\mathbf{g}_\ell(\mathbf{x})\mathbf{g}_{k-\ell-1}(\mathbf{x}), \quad \mathrm{mod}(k,2) = 1.$$

Note that all functions $\mathbf{g}_k$ with an even index vanish, i.e., $\mathbf{g}_{2k} = \mathbf{0}$ for all $k \in \mathbb{N}$. Equation (2.14) for the odd indices can be rearranged as

$$(2.15) \qquad \mathbf{g}_{2k+1}(\mathbf{x}) = -\sum_{\ell=1}^{2k-1} \nabla\mathbf{g}_\ell(\mathbf{x})\mathbf{g}_{2k-\ell}(\mathbf{x}), \quad k \geq 1,$$

where we have made the change of variables $k \mapsto 2k + 1$. Therefore, the flow on the slow manifold $\mathcal{M}_\epsilon$ is given by

$$(2.16) \qquad \mathbf{x}' = \epsilon\mathbf{v} = \epsilon\left[-\epsilon\eta\nabla f(\mathbf{x}) + \sum_{k=1}^{r}\epsilon^{2k+1}\mathbf{g}_{2k+1}(\mathbf{x}) + \mathcal{O}(\epsilon^{2r+3})\right].$$

Rescaling the fast time τ back to the original time t, we obtain (2.3). Note that the function $\mathbf{g}_{2r+1}$ involves derivatives of f up to and including order $r + 1$. Since we assumed $f \in C^{r+1}(\mathcal{X})$, the formal infinite series (2.11) must be truncated at this order. $\square$

Remark 5. Theorem 1 guarantees that the slow manifold persists for $\mu > \mu_0$. It would be very attractive to estimate the value of μ_0; however, there are two technical difficulties in the way of obtaining such estimates. First, μ_0 depends on the function f being minimized. Furthermore, even for a given function f, obtaining the parameter μ_0 is extremely difficult. The difficulty lies in the fact that, in singular perturbation theory, the invariant manifold arises as the fixed point of a certain integral equation [8, 19]. To prove the existence of the fixed point, one uses an infinite-dimensional contraction mapping argument. The contraction property of this map only holds for sufficiently large μ (or, equivalently, sufficiently small ϵ). Finding μ_0, such that for $\mu > \mu_0$ the contraction property holds, is a tedious task and has only been carried out for extremely simple cases (see, e.g., Example 1.3.2 in [19]). Nonetheless, in the present context, i.e., gradient flow of a convex function, we expect the threshold μ_0 to be small. The examples shown in section 4 confirm that our invariant slow manifolds persist for relatively small values of μ (see, in particular, the discussion on example 2).

As mentioned in Remark 4, to the leading order, the reduced flow on the slow manifold is a gradient flow. It is well known that the minimizer $\mathbf{x}^*$ is a locally asymptotically stable fixed point of the gradient flow for convex, continuously differentiable

functions f. The natural question is whether, for higher-order truncations, the minimizer $\mathbf{x}^*$ remains an asymptotically stable fixed point. It is straightforward to show this for the third-order truncations of the reduced flow.

PROPOSITION 1. *Let* $f \in C^2(\mathcal{X})$ *be a convex function. Up to the third order (neglecting* $\mathcal{O}(\epsilon^5)$ *terms), the reduced equation* (2.3) *reads*

$$(2.17) \qquad \dot{\mathbf{x}} = -\epsilon\eta\nabla f(\mathbf{x}) - \epsilon^3\eta^2\nabla^2 f(\mathbf{x})\nabla f(\mathbf{x}).$$

The minimizer $\mathbf{x}^*$ *is a locally asymptotically stable fixed point of the above equation.*

Proof. We use the Lyapunov function

$$(2.18) \qquad \mathcal{E}(\mathbf{x}) = \epsilon\eta\left(f(\mathbf{x}) - f(\mathbf{x}^*)\right) + \frac{1}{2}\epsilon^3\eta^2\|\nabla f(\mathbf{x})\|^2.$$

Note that $\mathcal{E}(\mathbf{x}) > 0$ for all $x \in \mathcal{X}\backslash\{\mathbf{x}^*\}$ and that $\mathcal{E}(\mathbf{x}^*) = 0$. The gradient of this Lyapunov function is given by

$$(2.19) \qquad \nabla\mathcal{E}(\mathbf{x}) = \epsilon\eta\nabla f(\mathbf{x}) + \epsilon^3\eta^2\nabla^2 f(\mathbf{x})\nabla f(\mathbf{x}) = -\dot{\mathbf{x}},$$

and therefore $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(\mathbf{x}) = \langle\nabla\mathcal{E}(\mathbf{x}), \dot{\mathbf{x}}\rangle < 0$. This completes the proof. $\quad\square$

Remark 6. Examining (2.17), the reduced equation up to the third order is equivalent to a preconditioned gradient descent flow $\dot{\mathbf{x}} = -[P(\mathbf{x})]^{-1}\nabla f(\mathbf{x})$ with the preconditioner $[P(\mathbf{x})]^{-1} = \epsilon\eta\left[I + \epsilon^2\eta\nabla^2 f(\mathbf{x})\right]$. Since $I + \epsilon^2\eta\nabla^2 f(\mathbf{x})$ is a near-identity map, its inverse exists for sufficiently small ϵ.

It turns out that the minimum $\mathbf{x}^*$ is an asymptotically stable fixed point of the reduced flow at any finite order of truncation. This is stated in the following theorem. The proof, however, is much more involved and is presented in the appendix.

THEOREM 2. *Let* $f \in C^{r+1}(\mathcal{X})$ *be a convex function. The point* $\mathbf{x} = \mathbf{x}^*$ *is an asymptotically stable fixed point of the reduced flow* (2.3). *More precisely,* $\mathbf{x} = \mathbf{x}^*$ *is an asymptotically stable fixed point of the differential equation*

$$(2.20) \qquad \dot{\mathbf{x}} = \sum_{k=0}^{p}\epsilon^{2k+1}\mathbf{g}_{2k+1}(\mathbf{x})$$

for all $0 \leq p \leq r$ *where* $\mathbf{g}_k : \mathcal{X} \to \mathbb{R}^n$ *are defined in* (2.4).

Proof. See Appendix A. $\quad\square$

2.2. Non-Euclidean setting. In this section, we consider an important class of distance-generating functions h that appear in the Bregman divergence (1.5) and have the form

$$(2.21) \qquad h(\mathbf{x}) = \langle\mathbf{x}, \mathbf{x}\rangle_H \triangleq \frac{1}{2}\langle\mathbf{x}, H\mathbf{x}\rangle,$$

where $\langle\cdot, \cdot\rangle$ denotes the Euclidean inner product and H is an $n \times n$ symmetric, positive-definite matrix. Note than $\langle\cdot, \cdot\rangle_H$ defines a Riemannian metric on the space $\mathbb{R}^n$.

With the choice (2.21), the Bregman divergence is given by

$$(2.22) \qquad D_h(\mathbf{y}, \mathbf{x}) = \frac{1}{2}\langle\mathbf{y} - \mathbf{x}, H(\mathbf{y} - \mathbf{x})\rangle,$$

and the associated Euler–Lagrange equation (1.6) reduces to

$$(2.23) \qquad \ddot{\mathbf{x}} + (e^{\alpha_t} - \dot{\alpha}_t)\,\dot{\mathbf{x}} + e^{2\alpha_t + \beta_t} H^{-1} \nabla f(\mathbf{x}) = 0.$$

Note that, since H is symmetric and positive-definite, its inverse exists. We have the following result for the slow manifold reduction of (2.23), which is quite similar to Theorem 1.

THEOREM 3. *Let $f \in C^{r+1}(\mathcal{X})$ with $r \geq 0$. Define $\mu = e^{\alpha_t} - \dot{\alpha}_t$ and $\eta = e^{2\alpha_t + \beta_t}$. If μ and η are constant, then there exists $\mu_0 > 0$ such that for all $\mu > \mu_0$ the following are true.*

(i) *The trajectories of the Euler–Lagrange equation (2.23) converge exponentially quickly to an n-dimensional invariant manifold embedded in the 2n-dimensional phase space $(\mathbf{x}, \dot{\mathbf{x}})$. Furthermore, this invariant manifold is a graph over the $\mathbf{x}$ coordinates.*

(ii) *The flow of (2.23) on this invariant manifold is given by*

$$(2.24) \qquad \dot{\mathbf{x}} = \sum_{k=0}^{r} \epsilon^{2k+1} \mathbf{g}_{2k+1}(\mathbf{x}) + \mathcal{O}(\epsilon^{2r+3}),$$

where $\epsilon = \mu^{-1}$ and the maps $\mathbf{g}_k : \mathcal{X} \to \mathbb{R}^n$ are defined recursively by

$$\mathbf{g}_1(\mathbf{x}) = -\eta H^{-1} \nabla f(\mathbf{x}),$$
$$\mathbf{g}_{2k}(\mathbf{x}) = \mathbf{0},$$
$$(2.25) \qquad \mathbf{g}_{2k+1}(\mathbf{x}) = - \sum_{\ell=1}^{2k-1} \nabla \mathbf{g}_\ell(\mathbf{x}) \mathbf{g}_{2k-\ell}(\mathbf{x}), \quad k \geq 1.$$

Proof of Theorem 3. The proof is quite similar to the Euclidean case (Theorem 1) and therefore is omitted here for brevity. □

We point out that the only difference in the above non-Euclidean result compared to the Euclidean case is the appearance of H^{-1} in the definition of g_1 in (2.25) which trickles to the higher-order terms. For instance, we have

$$(2.26) \qquad \mathbf{g}_3(\mathbf{x}) = -\eta^2 \left[H^{-1} \nabla^2 f(\mathbf{x}) H^{-1} \right] \nabla f(\mathbf{x}).$$

Furthermore, Theorem 2 also holds in the non-Euclidean case. Namely, the minimizer $\mathbf{x}^*$ is an asymptotically stable fixed point of (2.24) truncated to any order $0 \leq p \leq r$. The proof is similar to the Euclidean case and therefore is omitted here.

3. Multiscale analysis: Nonautonomous case. The results in section 2 hold under the key assumptions that $\mu = e^{\alpha_t} - \dot{\alpha}_t$ and $\eta = e^{2\alpha_t + \beta_t}$ are independent of time. This implies that the accelerated gradient flow (1.6) is an autonomous ODE. While these assumptions hold for an important subclass of the accelerated gradient flows, they do not hold for the Nesterov flow (1.3). In this section, we treat the nonautonomous case of the Nesterov flow separately by applying a time-dependent extension of the singular perturbation theory [5, 6, 2, 15]. The results are similar in nature to those of section 2; however, in the nonautonomous case, the slow manifold is a graph over $(\mathbf{x}, t)$ in the extended phase space $(\mathbf{x}, t, \mathbf{v})$.

First, we write the Nesterov flow in the form of an autonomous ODE in terms of the extended phase space variables $(\mathbf{x}, \mathbf{v}, t)$ where $\mathbf{v} = \dot{\mathbf{x}} := d\mathbf{x}/d\theta$. Here, $\theta = t - t_0$
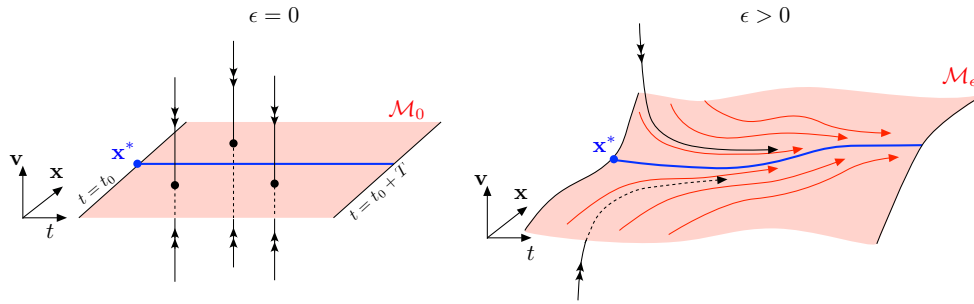
FIG. 3.1. *A sketch of the geometric singular perturbation theory for the nonautonomous acceler-ated gradient flow. In the singular limit, $\epsilon = 0$, the set $\mathcal{M}_0 = \{(\mathbf{x}, t, \mathbf{v}) \in \mathcal{X} \times [t_0, t_0+T] \times \mathbb{R}^n : \mathbf{v} = 0\}$ is filled with fixed points. The trajectories $(\mathbf{x}(\tau), t(\tau), \mathbf{v}(\tau)) = (\mathbf{x}_0, t_0, \mathbf{v}_0 e^{-\tau/t_0})$ converge exponen-tially quickly to $(\mathbf{x}_0, t_0, \mathbf{0}) \in \mathcal{M}_0$ from any initial condition $(\mathbf{x}_0, t_0, \mathbf{v}_0)$. In other words, the set $\mathcal{M}_0$ is invariant and the global attractor of the system. For $0 < \epsilon \ll 1$, the manifold $\mathcal{M}_0$ deforms into a nearby manifold $\mathcal{M}_\epsilon$ which is also invariant and globally attracting. However, $\mathcal{M}_\epsilon$ is not necessarily a collection of fixed points.*

is a reparameterrization of time. In the extended phase space, (1.3) can be written as a system of first-order differential equations

$$(3.1) \qquad \dot{\mathbf{x}} = \mathbf{v}, \quad \dot{\mathbf{v}} = -\frac{\rho}{t}\mathbf{v} - \nabla f(\mathbf{x}), \quad \dot{t} = 1,$$

with some initial conditions $\mathbf{x}(0) = \mathbf{x}_0 \in \mathcal{X}$, $\mathbf{v}(0) = \mathbf{v}_0 \in \mathbb{R}^n$, and $t(0) = t_0 > 0$. Note that $(\mathbf{x}, \mathbf{v}, t)$ are functions of the time-like variable θ and $(\dot{\mathbf{x}}, \dot{\mathbf{v}}, \dot{t})$ denotes their derivatives with respect to θ.

Next, we introduce the rescaled time $\tau = \theta/\epsilon$ where $\epsilon = \rho^{-1}$. Denoting the derivatives with respect to the fast time τ with a prime, we can recast (3.1) as

$$(3.2) \qquad \mathbf{x}' = \epsilon\mathbf{v}, \quad \mathbf{v}' = -\frac{1}{t}\mathbf{v} - \epsilon\nabla f(\mathbf{x}), \quad t' = \epsilon.$$

In the singular limit, $\epsilon = 0$, we have

$$(3.3) \qquad \mathbf{x}(\tau) = \mathbf{x}_0, \quad \mathbf{v}(\tau) = e^{-\tau/t_0}\mathbf{v}_0, \quad t(\tau) = t_0,$$

for all $\tau \geq 0$. This implies that the plane

$$(3.4) \qquad \mathcal{M}_0 = \{(\mathbf{x}, \mathbf{v}, t) \in \mathcal{X} \times \mathbb{R}^n \times [t_0, t_0 + T] : \mathbf{v} = \mathbf{0}\}$$

is the critical manifold towards which all trajectories converge with the exponential rate $e^{-\tau/t_0}$ (see Figure 3.1).

The nonautonomous singular perturbation theory [5, 6] guarantees the following. For sufficiently small $\epsilon > 0$, the manifold $\mathcal{M}_0$ deforms into a nearby normally at-tracting invariant manifold $\mathcal{M}_\epsilon$ which is $\mathcal{O}(\epsilon)$-close to the critical manifold $\mathcal{M}_0$. The trajectories of the Nesterov flow converge exponentially quickly towards the invariant manifold $\mathcal{M}_\epsilon$. Furthermore, $\mathcal{M}_\epsilon$ is a graph over the slow variables $(\mathbf{x}, t)$. We write this graph as a formal functional series expansion in ϵ,

$$(3.5) \qquad \mathcal{M}_\epsilon = \left\{(\mathbf{x}, \mathbf{v}, t) \in \mathcal{X} \times \mathbb{R}^n \times [t_0, t_0 + T] : \mathbf{v} = \sum_{k=1}^{\infty} \epsilon^k \mathbf{g}_k(\mathbf{x}, t)\right\},$$

where $\mathbf{g}_k : \mathcal{X} \times [t_0, t_0 + T] \to \mathbb{R}^n$. By differentiating this expression with respect to τ, we obtain

$$
\begin{aligned}
\mathbf{v}' &= \sum_{k=1}^{\infty} \epsilon^k \left(\nabla \mathbf{g}_k(\mathbf{x}, t) \mathbf{x}' + \left. \frac{\partial \mathbf{g}_k}{\partial t} \right|_{(\mathbf{x},t)} t' \right) \\
&= \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} \epsilon^{k+j+1} \nabla \mathbf{g}_k \mathbf{g}_j + \sum_{k=1}^{\infty} \epsilon^{k+1} \frac{\partial \mathbf{g}_k}{\partial t}
\end{aligned}
$$

$$
(3.6) \qquad = \epsilon^2 \frac{\partial \mathbf{g}_1}{\partial t} + \sum_{k=3}^{\infty} \epsilon^k \left(\frac{\partial \mathbf{g}_{k-1}}{\partial t} + \sum_{j=1}^{k-2} \nabla \mathbf{g}_j \mathbf{g}_{k-j-1} \right).
$$

On the other hand, using the fact that $\mathbf{v}' = -t^{-1} \mathbf{v} - \epsilon \nabla f(\mathbf{x})$, we obtain

$$
(3.7) \qquad \mathbf{v}' = \epsilon \left(-\frac{1}{t} \mathbf{g}_1(\mathbf{x}, t) - \nabla f(\mathbf{x}) \right) - \sum_{k=2}^{\infty} \epsilon^k \frac{\mathbf{g}_k(\mathbf{x}, t)}{t}.
$$

Equating these two expressions, we obtain

$$
(3.8a) \qquad \mathbf{g}_1(\mathbf{x}, t) = -t \nabla f(\mathbf{x}), \quad \mathbf{g}_2(\mathbf{x}, t) = t \nabla f(\mathbf{x}),
$$

$$
(3.8b) \qquad \mathbf{g}_k(\mathbf{x}, t) = -t \left[\frac{\partial \mathbf{g}_{k-1}}{\partial t} + \sum_{j=1}^{k-2} \nabla \mathbf{g}_j \mathbf{g}_{k-j-1} \right]_{(\mathbf{x},t)}, \quad k \geq 3.
$$

Note that functions $\mathbf{g}_k$ include derivatives of f up to order $\lceil k/2 \rceil$. The above results are summarized in the following theorem.

THEOREM 4. *Let $f \in C^r(\mathcal{X})$ with $r \geq 1$. There exists $\rho_0 > 0$ such that for all $\rho > \rho_0$ the following hold.*

(i) *The trajectories of the Nesterov equation (1.3) converge exponentially quickly to an $(n + 1)$-dimensional invariant manifold embedded in the $(2n + 1)$-dimensional extended phase space $(\mathbf{x}, \dot{\mathbf{x}}, t) \in \mathcal{X} \times \mathbb{R}^n \times [t_0, t_0 + T]$ for $0 < t_0, T < \infty$. Furthermore, this invariant manifold is a graph over the $(\mathbf{x}, t)$ coordinates (see Figure 3.1 for an illustration).*

(ii) *The flow of (1.3) on the slow manifold $\mathcal{M}_\epsilon$ is given by*

$$
(3.9) \qquad \dot{\mathbf{x}} = \sum_{k=1}^{2r} \epsilon^k \mathbf{g}_k(\mathbf{x}, t) + \mathcal{O}(\epsilon^{2r+1}),
$$

where $\epsilon = \rho^{-1}$ and the maps $\mathbf{g}_k : \mathcal{X} \times [t_0, t_0 + T] \to \mathbb{R}^n$ are defined recursively by (3.8).

Remark 7. Here, we consider the critical manifold $\mathcal{M}_0$ over the finite time interval $[t_0, t_0 + T]$ for some finite $T > 0$. As a result, the critical manifold is a graph over the compact domain $\mathcal{X} \times [t_0, t_0 + T]$. It is tempting to extend the slow manifold over the infinite time interval $[t_0, \infty)$. However, the fixed-point argument that guarantees the persistence of the slow manifold $\mathcal{M}_\epsilon$ is not generally valid over noncompact sets [8].

Note that to the leading order, we have

$$
(3.10) \qquad \frac{d\mathbf{x}}{dt} = \mathbf{v} = -\epsilon t \nabla f(\mathbf{x}).
$$

Reparameterizing time, by defining $\hat{t} = \epsilon\, t^2/2$, we obtain the usual gradient descent

$$\tag{3.11} \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\hat{t}} = -\nabla f(\mathbf{x}).$$

The trajectories of the two systems (3.10) and (3.11) are identical since one is a reparameterization of the other. Therefore, to the first order, the Nesterov flow reduced to the slow manifold coincides with the usual gradient descent. Recall that the same conclusion held for the autonomous case (cf. Remark 4).

Furthermore, examining the terms $\mathbf{g}_k(\mathbf{x}, t)$ with $k \geq 2$, we see that the nonautonomous slow manifold approximation contains higher-order terms with $t\nabla f(\mathbf{x})$. Collecting all such terms, we obtain

$$\tag{3.12} \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{v} = t\nabla f(\mathbf{x})\sum_{k=1}^{2r}(-\epsilon)^k = -S_r(\epsilon)t\nabla f(\mathbf{x}),$$

where $S_r(\epsilon) = \epsilon\left(1 - \epsilon^{2r}\right)/(1 + \epsilon)$, using the geometric series formula. If $f \in C^\infty(\mathcal{X})$ and $\epsilon < 1$, we have $S_r(\epsilon) \to \epsilon/(1 + \epsilon)$ as $r \to \infty$. Again, reparameterizing time by defining $\hat{t} = S_r(\epsilon)t^2/2$, we obtain the usual gradient flow (3.11).

Recall from Remark 7 that the existence of the slow manifold $\mathcal{M}_\epsilon$ is guaranteed over the finite time interval $t \in [t_0, t_0 + T]$. This prohibits asymptotic analysis of the reduced flow trajectories in the limit $t \to \infty$. However, in optimization, it suffices to ensure that the reduced flow (3.12) reaches a small neighborhood $B_\delta(\mathbf{x}^*)$ of the minimizer $\mathbf{x}^*$, where $0 < \delta \ll 1$ is the prescribed optimization tolerance (here, $B_\delta(\mathbf{x}^*)$ denotes the ball of radius δ in $\mathcal{X}$ centered at $\mathbf{x}^*$). The set $B_\delta(\mathbf{x}^*)$ can be reached in finite time, avoiding the need for the existence of the slow manifold for infinite times.

4. Examples. In this section, we demonstrate our results on three functions as listed in Table 4.1 and plotted in Figure 4.1. These functions are convex with their global minima at the origin, $\mathbf{x}^* = \mathbf{0}$. We divide our numerical results into two parts: section 4.1 contains the results for the autonomous case discussed in section 2; section 4.2 corresponds to the nonautonomous Nesterov flow discussed in section 3.

TABLE 4.1
Three functions used as examples here to demonstrate the results. The parameters μ and η refer to the parameters defined in Theorem 1. The parameter ρ denotes the constant in the Nesterov flow (1.3), and $\mathbf{x} = (x_1, x_2)$.

		Autonomous		Nonautonomous
Example #	$f(\mathbf{x})$	μ	η	ρ
1	$\frac{1}{2}\left(x_1^2 + 5x_2^2\right)$	8	1	3
2	$\frac{1}{4}\left(x_1^4 + 50x_2^4\right)$	2	1	1.5
3	$\log(e^{x_1^2} + e^{4x_2^2})$	4	1	3

4.1. Autonomous case. The numerical results presented in this section correspond to the autonomous case discussed in section 2. Recall that the autonomous case assumes that the coefficients $\mu = e^{\alpha_t} - \dot{\alpha}_t$ and $\eta = e^{2\alpha_t + \beta_t}$ are time-independent. For each example, we choose a different combination of these constant coefficients as listed in Table 4.1. While the results are valid for all $\mu > \mu_0$, the parameter μ_0 is not a priori known (see Remark 5). In example 2, we set $\epsilon = \mu^{-1} = 0.5$ to demonstrate that the slow manifolds may persist even for relatively large values of the perturbation parameter ϵ. We also note that for such larger values of ϵ (such as the one in example

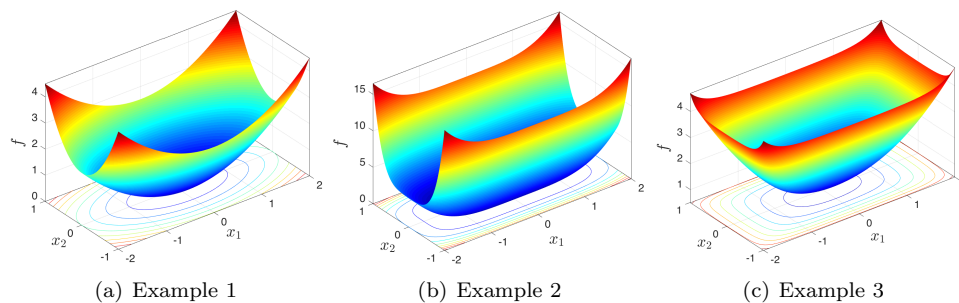(a) Example 1   (b) Example 2   (c) Example 3

FIG. 4.1. *The functions f corresponding to the examples listed in Table* 4.1.

2), the trajectories of the accelerated gradient flow exhibit a more oscillatory behavior than smaller ϵ, where the oscillations are mostly damped.

Figure 4.2 shows the corresponding truncated slow manifolds $\mathcal{M}_{\epsilon,p}$ for each example. Here, $\mathcal{M}_{\epsilon,p}$ denotes the slow manifold $\mathcal{M}_\epsilon$ truncated to the pth order. More specifically, the truncated slow manifold $\mathcal{M}_{\epsilon,p}$ is a graph over the $\mathbf{x}$-plane. This graph, denoted by $\mathbf{v}^{\epsilon,p} : \mathcal{X} \to \mathbb{R}^n$, is defined by

$$(4.1) \qquad \mathbf{v}^{\epsilon,p}(\mathbf{x}) \triangleq \sum_{k=1}^{p} \epsilon^k \mathbf{g}_k(\mathbf{x}),$$

where $\mathbf{g}_k$'s are given in (2.4). For example 1, we plot the first-order truncation ($p = 1$), and, for the other two examples, we plot the third-order truncations ($p = 3$). Recall that the even terms in the series vanish so that the $p = 3$ truncation only contains two terms. In all three examples, the difference between the first- and third-order truncations is insignificant and visually unnoticeable.

In each panel of Figure 4.2, two types of trajectories are shown. The blue curves mark the trajectories of the second-order Euler–Lagrange equation (2.2), while the red curves mark the trajectories of the first-order reduced equation (2.3) plotted on the slow manifold. (Color for all figures is available online.)

The trajectories of the Euler–Lagrange equation (blue curves) undergo two stages. First they approach the slow manifold exponentially fast. In the second stage, they closely follow the slow manifold towards the minimizer $\mathbf{x}^*$. These two stages are demonstrated in Figure 4.3, which shows the distance from the slow manifold $\mathcal{M}_{\epsilon,p}$ along the trajectory $(\mathbf{x}(t), \mathbf{v}(t))$ of the Euler–Lagrange equation corresponding to example 1. This distance is computed as

$$(4.2) \qquad d(t) = \|\mathbf{v}(t) - \mathbf{v}^{\epsilon,p}(\mathbf{x}(t))\|.$$

The quantity $d(t)$ shows two exponential slopes. Within the first time unit (see the inset of Figure 4.3), the distance drops sharply, which corresponds to the convergence towards the slow manifold. The rate of convergence seems identical for both truncations $p = 1$ and $p = 3$. However, the distance corresponding to $p = 3$ decreases by a greater amount since this higher-order truncation more accurately approximates the true slow manifold $\mathcal{M}_\epsilon$. Later ($t > 10$), the distance $d(t)$ continues to decrease but at a lower rate. We attribute this second decaying stage to the asymptotic convergence of the Euler–Lagrange solutions to the minimizer $\mathbf{x}^*$ within the slow manifold.

A practical implication of our slow manifold reduction is that one can reduce the computational cost of accelerated gradient flow by skipping the first stage of their
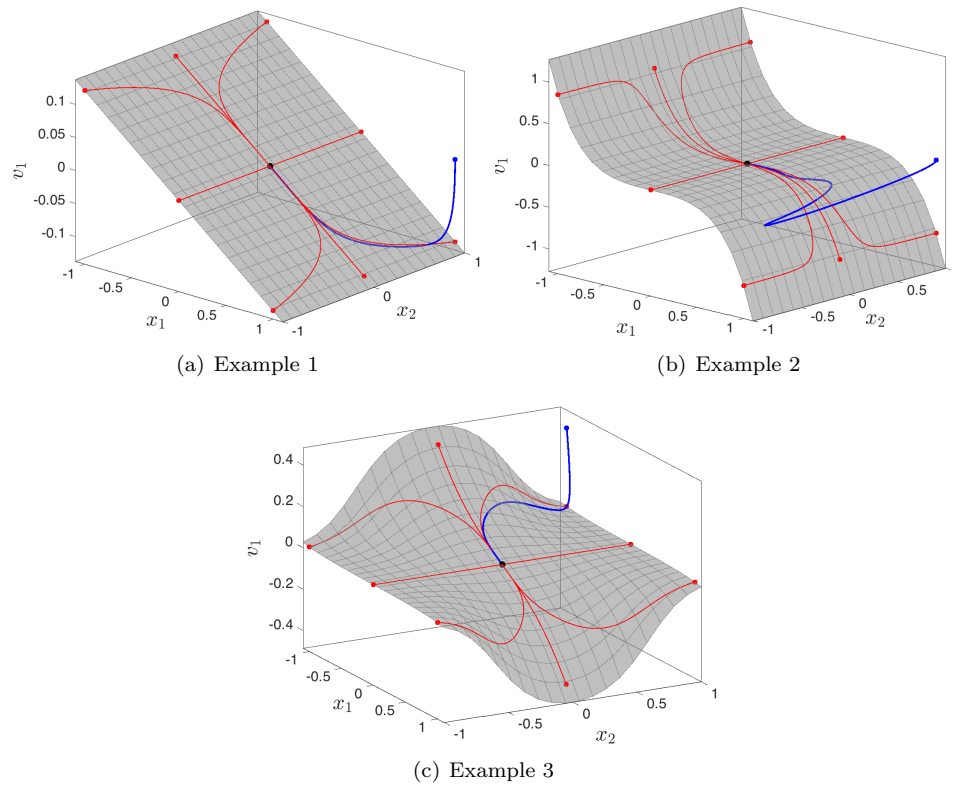
(a) Example 1

(b) Example 2

(c) Example 3

FIG. 4.2. *The projection of the slow manifolds $\mathcal{M}_\epsilon$ (gray surfaces) onto the (x_1, x_2, v_1) subspace. One trajectory of the Euler–Lagrange equation (2.2) (blue) and several trajectories of the reduced system (2.3) (red) are also shown. The initial conditions for each trajectory are marked by blue and red dots, respectively. The black dot marks the minimum $\mathbf{x}^*$ (which coincides with the origin). (Color available online.)*
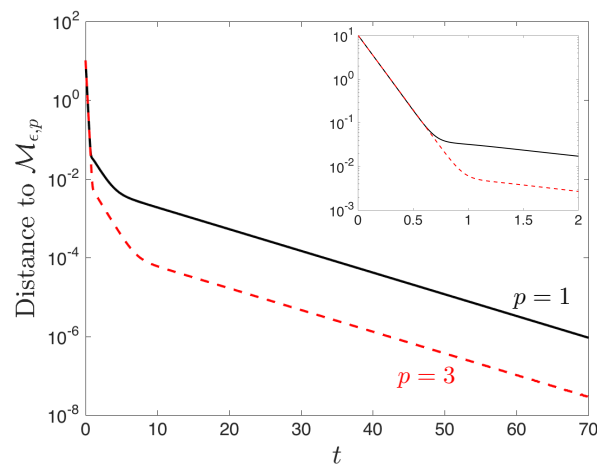

FIG. 4.3. *The distance from the truncated slow manifold $\mathcal{M}_{\epsilon,p}$ along the trajectories of the Euler–Lagrange equation (2.2) for example 1. The inset shows a closeup view.*
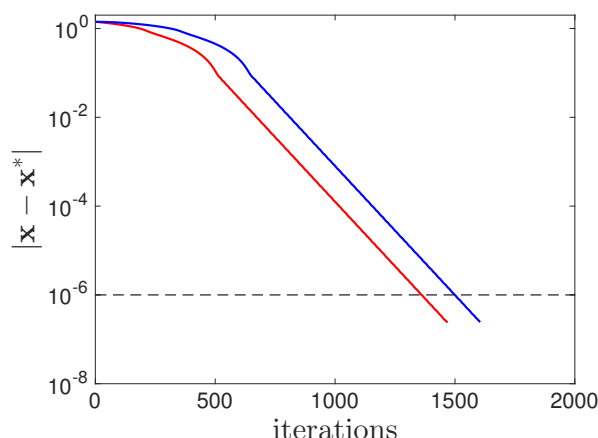
FIG. 4.4. *Alternative initial conditions. The function in example 1 is minimized using the accelerated gradient descent (2.2). The blue curve corresponds to the initial conditions* $\mathbf{x}_0 = (1,1)$ *and* $\mathbf{v}_0 = (0,0)$. *The red curve corresponds to the alternative initial conditions* $\tilde{\mathbf{x}}_0 = \mathbf{x}_0 = (1,1)$ *and* $\tilde{\mathbf{v}}_0 = \mathbf{v}^{\epsilon,1}(\mathbf{x}_0) \simeq (-0.125, -0.625)$. *It takes fewer iterations for the red curve to reach the error tolerance* 10^{-6} *marked by the dashed black curve. (Color available online.)*

evolution (i.e., the convergence towards the slow manifold). Since the accelerated gradient flows are second-order differential equations, they require an initial guess $(\mathbf{x}_0, \mathbf{v}_0)$ as their initial condition. Using the slow manifold, it is advantageous to choose the alternative initial velocity $\tilde{\mathbf{v}}_0 = \mathbf{v}^{\epsilon,p}(\tilde{\mathbf{x}}_0)$, where

$$(4.3) \qquad \tilde{\mathbf{x}}_0 = \arg\min_{\mathbf{x} \in \mathcal{X}} \left[|\mathbf{x} - \mathbf{x}_0|^2 + |\mathbf{v}^{\epsilon,p}(\mathbf{x}) - \mathbf{v}_0|^2 \right].$$

The point $(\tilde{\mathbf{x}}_0, \tilde{\mathbf{v}}_0)$ is the closest point on the slow manifold $\mathcal{M}_{\epsilon,p}$ to the initial guess $(\mathbf{x}_0, \mathbf{v}_0)$. This alternative initial guess $(\tilde{\mathbf{x}}_0, \tilde{\mathbf{v}}_0)$ avoids the initial decay phase of the flow (i.e., convergence towards the slow manifold) by placing the trajectory $\mathcal{O}(\epsilon^p)$-close to this manifold at the initial time. As a result, several evaluations of the function f and its derivatives during the decay phase are dispensed with.

However, carrying out the minimization (4.3) can be costly itself, outweighing the saved computational cost from skipping the decay phase. Here, we propose a cheaper approach by choosing the alternative initial guess $\tilde{\mathbf{x}}_0 = \mathbf{x}_0$ and $\tilde{\mathbf{v}}_0 = \mathbf{v}^{\epsilon,p}(\mathbf{x}_0)$ for a given $\mathbf{x}_0 \in \mathcal{X}$. For $p = 1$, for instance, this only requires one computation of the gradient of f since $\mathbf{v}^{\epsilon,1}(\mathbf{x}_0) = -\epsilon\eta\nabla f(\mathbf{x}_0)$.

Figure 4.4 shows an implementation of this alternative initial condition on example 1. We solve the accelerated gradient flow (2.2) from two different initial conditions. First, we set $\mathbf{x}_0 = (1,1)$ and $\mathbf{v}_0 = (0,0)$. This corresponds to the blue curve in Figure 4.4. Then we choose the alternative initial conditions $\tilde{\mathbf{x}}_0 = \mathbf{x}_0 = (1,1)$ and $\tilde{\mathbf{v}}_0 = \mathbf{v}^{\epsilon,1}(\mathbf{x}_0) = -\epsilon\eta\nabla f(\mathbf{x}_0) \simeq (-0.125, -0.625)$. This is marked by the red curve. The alternative initial guess $(\tilde{\mathbf{x}}_0, \tilde{\mathbf{v}}_0)$ takes fewer iterations to reach a given error $|\mathbf{x} - \mathbf{x}^*|$.

For instance, it takes 1499 iterations to reach the error $|\mathbf{x} - \mathbf{x}^*| < 10^{-6}$ if we start from the initial guess $(\mathbf{x}_0, \mathbf{v}_0)$. To reach the same error tolerance, it takes 1361 iterations if we start from the alternative initial guess $(\tilde{\mathbf{x}}_0, \tilde{\mathbf{v}}_0)$. Accounting for the gradient evaluation required to compute $\tilde{\mathbf{v}}_0$, the alternative initial guess takes 137 fewer gradient evaluations to reach the error tolerance 10^{-6}. In high dimensions this can amount to a noticeable reduction in the computational time.
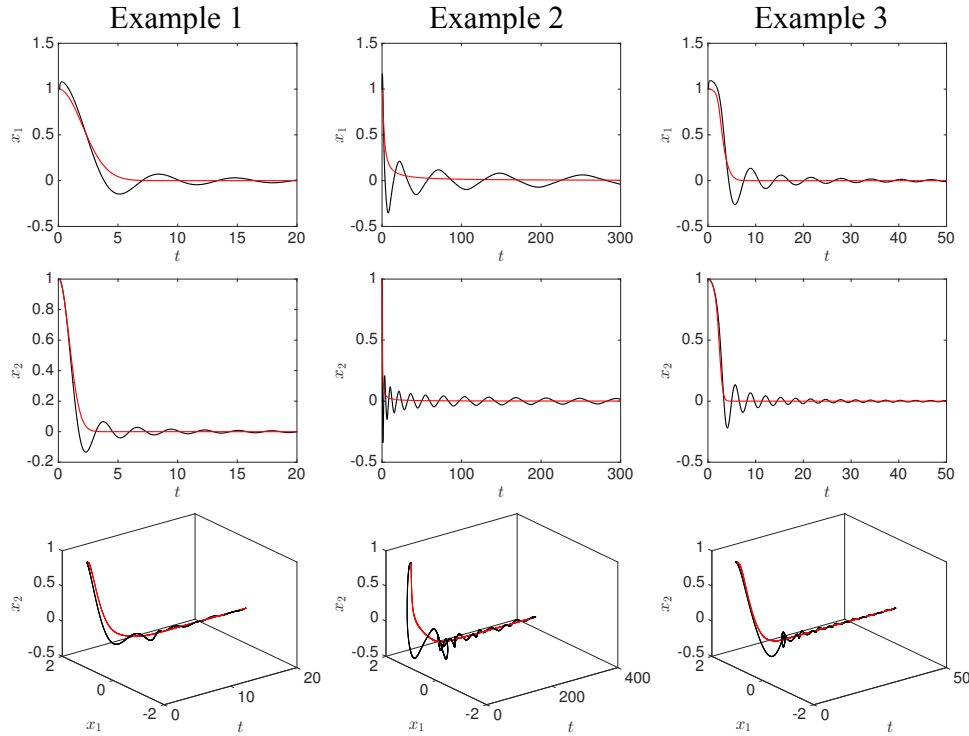
FIG. 4.5. *Trajectories of the accelerated gradient flows. Example numbers refer to the functions listed in Table* 4.1. *The black curves correspond to the trajectories of the Nesterov flow* (1.3), *and the red curves correspond to the reduced-order flow* (3.12). *(Color available online.)*

4.2. Nonautonomous case. The numerical results presented in this section correspond to the nonautonomous case discussed in section 3. The parameter $\rho = \epsilon^{-1}$ is reported in the last column of Table 4.1. For examples 1 and 3, we use the standard value $\rho = 3$. To demonstrate that the slow manifold reduction may be valid for even smaller values of ρ (or, equivalently, larger values of ϵ), we set $\rho = 1.5$ in example 2.

The numerical results are reported in Figure 4.5. In all examples, the initial conditions are $\mathbf{x}_0 = (1,1)$ and $t_0 = 0.1$. The black curves mark the trajectories of the full Nesterov flow (1.3) with the initial velocity $\mathbf{v}_0 = \dot{\mathbf{x}}(t_0) = (2,0)$. Note that the phase space $\mathcal{X} \times \mathbb{R}^2 \times [t_0, t_0 + T]$ of the Nesterov flow is five-dimensional. Therefore, the bottom panel in Figure 4.5 shows the projections of the trajectories on the three-dimensional subspace $\mathcal{X} \times [t_0, t_0 + T]$.

The red curves mark the trajectories of the slow manifold reduced flow (3.12). The trajectories of both systems (the reduced flow and the full Nesterov flow) converge to the global minimum $\mathbf{x}^* = \mathbf{0}$. As expected, the trajectories of the Nesterov flow oscillate around the trajectories of the reduced flow. As time increases, the amplitude of these oscillations decay, and the Nesterov flow trajectories converge onto the slow manifold trajectories.

We have repeated the simulations for various initial conditions $(\mathbf{x}_0, \mathbf{v}_0)$ and observed very similar behavior (namely, the convergence of the oscillatory Nesterov trajectories onto the slow manifold trajectories), ensuring that the results are not sensitive to the initial conditions.

5. Conclusions. It has recently been shown that the continuous-time limit of accelerated gradient descent methods are the Euler–Lagrange equations corresponding to a single Lagrangian. This Euler–Lagrange equation takes the form of a second-order ODE that we refer to as the accelerated gradient flow.

Here we showed that, under certain assumptions, the accelerated gradient flows possess an attracting, invariant, slow manifold. The trajectories of accelerated gradient flow undergo two stages. First, they converge exponentially quickly towards the slow manifold. Then they closely follow the flow within the slow manifold. We derived a general explicit formula that approximates the slow manifold to any arbitrary order of accuracy.

To the leading order, the flow within the slow manifold coincides with the usual gradient (or steepest) descent. Higher-order approximations of the slow manifold, however, involve higher-order derivatives of the objective function.

We divided our analysis into two parts. In the first part (section 2), it is assumed that the accelerated gradient flow is autonomous (i.e., it does not have any explicit dependence on time). In this setting, the classical geometric singular perturbation theory was applied to prove the existence of the slow manifold as an n-dimensional submanifold of the phase space. The second part (section 3) concerns the Nesterov accelerated gradient flow which is a nonautonomous differential equation. More recent singular perturbation results are applicable to this flow. While the conclusions are similar to the autonomous case, the Nesterov flow has an $(n + 1)$-dimensional slow manifold in the extended phase space.

A practical implication of our results is the reduced computational cost of accelerated gradient flows by initializing it close to the slow manifold, which is now known explicitly. This initialization avoids the initial stage of the flow, which involves converging towards the slow manifold and hence reducing the computational cost.

Since accelerated gradient iterations are temporal discretizations of the accelerated gradient flow [18], we expect similar slow manifold reductions to hold for the discrete accelerated methods. However, a rigorous singular perturbation analysis of these iterations is desirable and will be pursued in future work.

Appendix A. Proof of Theorem 2. To prove this theorem, we will need the following lemma.

LEMMA 1. *The matrices $\nabla \mathbf{g}_k(\mathbf{x}) \in \mathbb{R}^{n \times n}$ are symmetric for any $\mathbf{x} \in \mathcal{X}$.*

Proof. To show that $\nabla \mathbf{g}_k(\mathbf{x})$ are symmetric, we first show that $\mathbf{g}_k = \nabla \phi_k$ for twice continuously differentiable functions $\phi_k : \mathcal{X} \to \mathbb{R}$. And therefore $\nabla \mathbf{g}_k = \nabla^2 \phi_k$ is symmetric. First, note that $\mathbf{g}_{2k} = \mathbf{0}$, and therefore this assertion is trivially correct for even indices with $\phi_{2k} = 0$. For odd indices, we proceed with induction.

Note that $\mathbf{g}_1 = -\eta \nabla f$, and therefore we have $\phi_1(\mathbf{x}) = -\eta f(\mathbf{x})$. Now assume that $\mathbf{g}_\ell = \nabla \phi_\ell$ for all $1 \leq \ell < 2k + 1$. This implies

$$
\begin{aligned}
\mathbf{g}_{2k+1} &= -\sum_{\ell=1}^{2k-1} \nabla^2 \phi_\ell \nabla \phi_{2k-\ell} \\
&= \nabla \left[-\frac{1}{2} \sum_{\ell=1}^{2k-1} \langle \nabla \phi_\ell, \nabla \phi_{2k-\ell} \rangle \right] \\
&= \nabla \left[-\frac{1}{2} \sum_{\ell=1}^{2k-1} \langle \mathbf{g}_\ell, \mathbf{g}_{2k-\ell} \rangle \right].
\end{aligned}
$$

(A.1)

The second line in the above equation follows from the series of identities

$$
\begin{aligned}
\nabla\left[-\frac{1}{2}\sum_{\ell=1}^{2k-1}\langle\nabla\phi_\ell,\nabla\phi_{2k-\ell}\rangle\right] &= -\frac{1}{2}\sum_{\ell=1}^{2k-1}\left[\nabla^2\phi_\ell\nabla\phi_{2k-\ell}+\nabla^2\phi_{2k-\ell}\nabla\phi_\ell\right] \\
&= -\frac{1}{2}\sum_{\ell=1}^{2k-1}\nabla^2\phi_\ell\nabla\phi_{2k-\ell}-\frac{1}{2}\sum_{\ell=1}^{2k-1}\nabla^2\phi_{2k-\ell}\nabla\phi_\ell \\
&= -\frac{1}{2}\sum_{\ell=1}^{2k-1}\nabla^2\phi_\ell\nabla\phi_{2k-\ell}-\frac{1}{2}\sum_{j=1}^{2k-1}\nabla^2\phi_j\nabla\phi_{2k-j} \\
&= -\sum_{\ell=1}^{2k-1}\nabla^2\phi_\ell\nabla\phi_{2k-\ell},
\end{aligned}
$$

(A.2)

where in the third line we used the change of indices $2k-\ell\mapsto j$. Therefore, $\mathbf{g}_k=\nabla\phi_k$ for all $k\geq 1$, where the scalar functions $\phi_k:\mathcal{X}\to\mathbb{R}$ are defined recursively by $\phi_1=-\eta f$, $\phi_{2k}=0$, and

$$
(A.3)\qquad \phi_{2k+1}=-\frac{1}{2}\sum_{\ell=1}^{2k-1}\langle\nabla\phi_\ell,\nabla\phi_{2k-\ell}\rangle,\quad k\geq 1. \qquad\square
$$

Proof of Theorem 2. Consider the Lyapunov function

$$
(A.4)\qquad \mathcal{E}(\mathbf{x})=\epsilon\eta\left(f(\mathbf{x})-f(\mathbf{x}^*)\right)+\sum_{k=1}^{r}\frac{\epsilon^{2k+1}}{2}\sum_{\ell=1}^{2k-1}\langle\mathbf{g}_\ell(\mathbf{x}),\mathbf{g}_{2k-\ell}(\mathbf{x})\rangle.
$$

Note that $\mathcal{E}(\mathbf{x}^*)=0$ and that the first term $\epsilon\eta(f(\mathbf{x})-f(\mathbf{x}^*))$ is positive for all $\mathbf{x}\in\mathcal{X}\backslash\{\mathbf{x}^*\}$. The second term (corresponding to $k=1$) is $\epsilon^3\langle\mathbf{g}_1(\mathbf{x}),\mathbf{g}_1(\mathbf{x})\rangle=\epsilon^3\eta^2\|\nabla f(\mathbf{x})\|^2$, which is also positive for all $\mathbf{x}\in\mathcal{X}\backslash\{\mathbf{x}^*\}$. It is straightforward to show that the higher-order terms in the series are also positive at least in a neighborhood of $\mathbf{x}^*$.

The gradient of the Lyapunov function is given by

$$
\begin{aligned}
\nabla\mathcal{E}(\mathbf{x}) &= \epsilon\eta\nabla f(\mathbf{x})+\sum_{k=1}^{r}\frac{\epsilon^{2k+1}}{2}\sum_{\ell=1}^{2k-1}\left[\nabla\mathbf{g}_\ell(\mathbf{x})^\top\mathbf{g}_{2k-\ell}(\mathbf{x})+\nabla\mathbf{g}_{2k-\ell}(\mathbf{x})^\top\mathbf{g}_\ell(\mathbf{x})\right] \\
&= \epsilon\eta\nabla f(\mathbf{x})+\sum_{k=1}^{r}\epsilon^{2k+1}\sum_{\ell=1}^{2k-1}\nabla\mathbf{g}_\ell(\mathbf{x})\mathbf{g}_{2k-\ell}(\mathbf{x}) \\
&= -\epsilon\mathbf{g}_1(\mathbf{x})-\sum_{k=1}^{r}\epsilon^{2k+1}\mathbf{g}_{2k+1}(\mathbf{x})=-\sum_{k=0}^{r}\epsilon^{2k+1}\mathbf{g}_{2k+1}(\mathbf{x})=-\dot{\mathbf{x}}.
\end{aligned}
$$

(A.5)

In these series of identities, we used the fact that $\nabla\mathbf{g}_\ell(\mathbf{x})$ are symmetric matrices (see Lemma 1) and that

$$
(A.6)\qquad \sum_{\ell=1}^{2k-1}\nabla\mathbf{g}_{2k-\ell}(\mathbf{x})\mathbf{g}_\ell(\mathbf{x})=\sum_{j=1}^{2k-1}\nabla\mathbf{g}_j(\mathbf{x})\mathbf{g}_{2k-j}(\mathbf{x}).
$$

Therefore, we have $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(\mathbf{x})=\langle\nabla\mathcal{E}(\mathbf{x}),\dot{\mathbf{x}}\rangle=-\|\dot{\mathbf{x}}\|^2\leq 0$ with the identity attained only at the minimum $\mathbf{x}^*$. This completes the proof. $\qquad\square$

REFERENCES

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004, https://doi.org/10.1017/CBO9780511804441.

[2] A. Carvalho, J. A. Langa, and J. Robinson, *Attractors for Infinite-Dimensional Non-Autonomous Dynamical Systems*, Appl. Math. Sci. 182, Springer, New York, 2013, https://doi.org/10.1007/978-1-4614-4581-4.

[3] M. Farazmand, *An adjoint-based approach for finding invariant solutions of Navier-Stokes equations*, J. Fluid Mech., 795 (2016), pp. 278–312, https://doi.org/10.1017/jfm.2016.203.

[4] N. Fenichel, *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53–98, https://doi.org/10.1016/0022-0396(79)90152-9.

[5] G. Haller and T. Sapsis, *Where do inertial particles go in fluid flows?*, Phys. D, 237 (2008), pp. 573–583, https://doi.org/10.1016/j.physd.2007.09.027.

[6] M. Haragus and G. Iooss, *Local Bifurcations, Center Manifolds, and Normal Forms in Infinite-Dimensional Dynamical Systems*, Springer, London, 2011, https://doi.org/10.1007/978-0-85729-112-7.

[7] C. Jin, P. Netrapalli, and M. I. Jordan, *Accelerated gradient descent escapes saddle points faster than gradient descent*, in Proceedings of the 31st Conference on Learning Theory, S. Bubeck, V. Perchet, and P. Rigollet, eds., Proc. Mach. Learn. Res. 75, PMLR, 2018, pp. 1042–1085, http://proceedings.mlr.press/v75/jin18a.html.

[8] C. K. Jones, *Geometric Singular Perturbation Theory*, in Dynamical Systems, R. Johnson, ed., Springer, Berlin, 1995, pp. 44–118, https://doi.org/10.1007/BFb0095239.

[9] C. Kuehn, *Multiple Time Scale Dynamics*, Springer, Cham, 2015, https://doi.org/10.1007/978-3-319-12316-5.

[10] J. LaSalle, *Some extensions of Liapunov's second method*, IRE Trans. Circuit Theory, 7 (1960), pp. 520–527, https://doi.org/10.1109/TCT.1960.1086720.

[11] L. Lessard, B. Recht, and A. Packard, *Analysis and design of optimization algorithms via integral quadratic constraints*, SIAM J. Optim., 26 (2016), pp. 57–95, https://doi.org/10.1137/15M1009597.

[12] Y. Nesterov, *A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$*, Dokl. Akad. Nauk SSSR, 269 (1983), pp. 543–547 (in Russian).

[13] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Appl. Optim. 87, Springer, Boston, MA, 2004, https://doi.org/10.1007/978-1-4419-8853-9.

[14] B. Polyak, *Some methods of speeding up the convergence of iteration methods*, USSR Comput. Math. Math. Phys., 4 (1964), pp. 1–17, https://doi.org/10.1016/0041-5553(64)90137-5.

[15] A. J. Roberts, *Backwards Theory Supports Modelling via Invariant Manifolds for Non-Autonomous Dynamical Systems*, preprint, https://arxiv.org/abs/1804.06998, 2019.

[16] S. Smale, *The fundamental theorem of algebra and complexity theory*, Bull. Amer. Math. Soc. (N.S.), 4 (1981), pp. 1–36, https://doi.org/10.1090/S0273-0979-1981-14858-8.

[17] W. Su, S. Boyd, and E. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, J. Mach. Learn. Res., 17 (2016), pp. 1–43, http://jmlr.org/papers/v17/15-084.html.

[18] A. Wibisono, A. C. Wilson, and M. I. Jordan, *A variational perspective on accelerated methods in optimization*, Proc. Natl. Acad. Sci. USA, 113 (2016), pp. E7351–E7358, https://doi.org/10.1073/pnas.1614734113.

[19] S. Wiggins, *Normally Hyperbolic Invariant Manifolds in Dynamical Systems*, Appl. Math. Sci. 105, Springer, New York, 1994, https://doi.org/10.1007/978-1-4612-4312-0.

[20] A. C. Wilson, B. Recht, and M. I. Jordan, *A Lyapunov Analysis of Momentum Methods in Optimization*, preprint, https://arxiv.org/abs/1611.02635, 2016.