Nesterov's Accelerated Gradient Descent as a Nash-Equilibrium Seeking Algorithm for Quadratic Games

Jay Paek

1 Notations and Preliminaries

Let $\mathbb{R}_{++}$ denote the set of real numbers strictly greater than 0. We denote the time variable by $t \in \mathbb{R}_{++}$, which represents the time step of the trajectory. For a differentiable function $x : \mathbb{R} \to \mathbb{R}^n$, its first and second time derivatives are denoted by $\dot{x}$ and $\ddot{x}$, respectively.

Consider the differential equation that simulates Nesterov Accelerated Gradient Descent (NAGD) [3]:

$$\ddot{x}(t) + \frac{3}{t}\,\dot{x}(t) + \nabla_x f(x(t)) = 0,$$

which can be equivalently rewritten as a first-order system:

$$\begin{cases} \dot{x}_1(t) = x_2(t), \\ \dot{x}_2(t) = -\frac{3}{t}\,x_2(t) - \nabla_{x_1} f(x_1(t)), \end{cases}$$

where $x_1(t)$, $x_2(t) \in \mathbb{R}^n$ represent the position and momentum of the dynamics, respectively. Here, $f(\cdot)$ denotes the objective function (for example, a cost function in a game or an error function in a model), and $\nabla_x f(\cdot)$ its gradient.

Next, consider a game with N players, where the state vector is given by

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \in \mathbb{R}^N,$$

with each component x_i representing the action of the ith player. Each player is associated with a quadratic cost function

$$J_i(x) = x^T Q_i x, \quad \text{for } i = 1, \ldots, N,$$

where $Q_i \in \mathbb{R}^{N \times N}$ is a suitable symmetric matrix.

We define the aggregate cost vector function $J : \mathbb{R}^N \to \mathbb{R}^N$ as

$$J(x) = \begin{bmatrix} J_1(x) \\ \vdots \\ J_N(x) \end{bmatrix} = \begin{bmatrix} x^T Q_1 x \\ \vdots \\ x^T Q_N x \end{bmatrix}.$$

Rather than using the full gradient of the joint cost, we employ a *pseudo-gradient* defined by

$$\nabla f(x) = \begin{bmatrix} \frac{\partial J_1(x)}{\partial x_1} \\ \vdots \\ \frac{\partial J_N(x)}{\partial x_N} \end{bmatrix}.$$

That is, each player's cost function is differentiated with respect to their own action. Since each entry of the pseudo-gradient is a linear combination of the components of x, we can write $\nabla f(x) = \mathcal{G}x$, tfor some matrix $\mathcal{G} \in \mathbb{R}^{N \times N}$.

For these games, the Nash equilibrium is characterized by the condition $\nabla f(x) = 0$; that is, the equilibrium resides in the null space of the pseudo-gradient matrix. In particular, if $\mathcal{G}$ is full-rank, then the unique Nash equilibrium is at $x = 0$.

Below, we present trajectories of the NAGD dynamics, where the horizontal axis corresponds to the first component and the vertical axis corresponds to the second component of the state. All simulations are generated using 10^5 iterations with $t \in [1, 10^3]$ and a time increment of 0.01.

Figure 1 displays two instances of the NAGD dynamics, each initialized at a random point satisfying $\|x\|_2 < 1$.

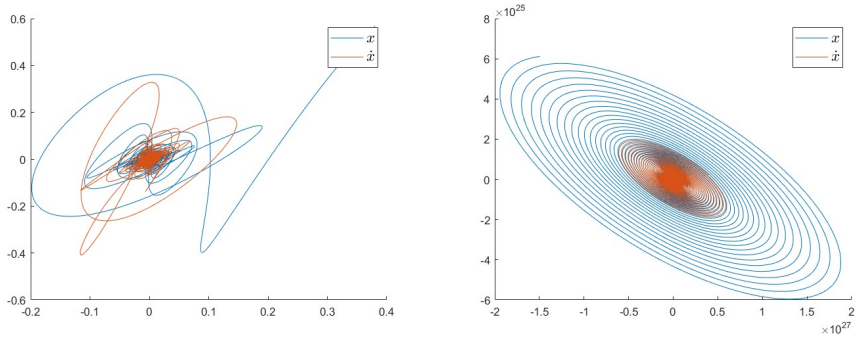

Figure 1: Nesterov dynamics with two different full-rank pseudo-gradient matrices. Left: $\mathcal{G} = \begin{bmatrix} 0.4 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$; Right: $\mathcal{G} = \begin{bmatrix} 0.1 & -0.1 \\ 0 & 0.1 \end{bmatrix}$.

It is evident that the left system converges while the right system does not. Notably, the trajectory on the left exhibits large oscillations as it approaches the equilibrium at 0.

Figure 2 shows two additional trajectories of the NAGD dynamics with low-rank pseudo-gradient matrices (again, initialized at a random point with $\|x\|_2 < 1$). In these cases, both trajectories converge to the null space of $\mathcal{G}$. Specifically, the trajectory on the left follows a path along the line spanned by $\begin{bmatrix} 1 & -1 \end{bmatrix}^\top$, whereas the trajectory on the right aligns with the line spanned by $\begin{bmatrix} 2 & -1 \end{bmatrix}^\top$ (up to a scaling factor $\alpha \in \mathbb{R}$).

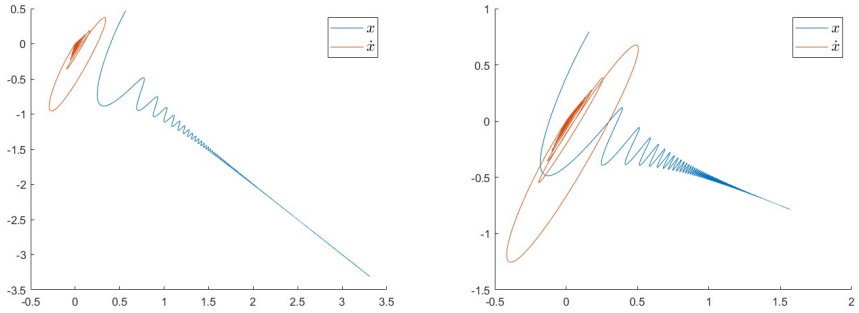

Figure 2: Nesterov dynamics with two different low-rank pseudo-gradient matrices. Left: $\mathcal{G} = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$; Right: $\mathcal{G} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$.

In this paper, we analyze the stability and instability of NAGD dynamics for N-player quadratic games with various pseudo-gradient structures by examining the components of the underlying differential equations.

2 Projected Dynamics (WIP)

In this section, we will introduce an intuitive way to analyze differential equations.

Proposition: Let the system

$$\dot{x}(t) = f(x(t))$$

be a linear ODE with $f(x) = Ax$, where A is an $n \times n$ matrix. Let $\{w_i\}_{i=1}^n \subset \mathbb{R}^n$ be an orthonormal basis, i.e., $\|w_i\|_2 = 1$ for each i. Suppose that for each i, there exists a function

$$g_i : \mathbb{R} \to \mathbb{R}$$

such that

$$w_i^\top f(x(t)) = g_i\big(w_i^\top x(t)\big)$$

for all $x(t) \in \mathbb{R}^n$. Then the following are equivalent:

1. The system $\dot{x}(t) = f(x(t))$ is Lyapunov stable.

2. The scalar dynamics

$$\dot{y}_i(t) = g_i\big(y_i(t)\big), \quad \text{with } y_i(t) = w_i^\top x(t),$$

have a stable equilibrium at $y_i = 0$ for every $i = 1, \ldots, n$.

Proof: We prove the proposition in two directions.

(Forward direction): Assume that the system

$$\dot{x}(t) = f(x(t))$$

is Lyapunov stable. Then, for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $\|x(0)\| < \delta$, then $\|x(t)\| < \varepsilon$ for all $t \geq 0$. Since the basis $\{w_i\}$ is orthonormal, we can write

$$\|x(t)\|^2 = \sum_{i=1}^{n} \big(w_i^\top x(t)\big)^2.$$

Thus, $\|x(t)\| < \varepsilon$ implies that

$$\big|w_i^\top x(t)\big| < \varepsilon \quad \text{for each } i.$$

Since the scalar dynamics are given by

$$\dot{y}_i(t) = g_i\big(y_i(t)\big) \quad \text{with } y_i(t) = w_i^\top x(t),$$

each scalar system inherits the Lyapunov stability property, so that the equilibrium $y_i = 0$ is stable for every i.

(Reverse direction): Now assume that for each i, the scalar differential equation

$$\dot{y}_i(t) = g_i\big(y_i(t)\big)$$

has a stable equilibrium at $y_i = 0$. This means that for each i and for every $\varepsilon > 0$, there exists $\delta_i > 0$ such that if

$$\big|w_i^\top x(0)\big| < \delta_i,$$

then

$$\big|w_i^\top x(t)\big| < \varepsilon \quad \text{for all } t \geq 0.$$

Since $\{w_i\}_{i=1}^n$ is an orthonormal basis for $\mathbb{R}^n$, any state $x(t)$ may be uniquely decomposed as

$$x(t) = \sum_{i=1}^{n} \big(w_i^\top x(t)\big) w_i.$$

Using the linearity of f, we have:

$$f\left(\sum_{i=1}^{n} \big(w_i^\top x(t)\big) w_i\right) = \sum_{i=1}^{n} f\Big(\big(w_i^\top x(t)\big) w_i\Big) = \sum_{i=1}^{n} \Big[w_i^\top f\big(x(t)\big)\Big] w_i.$$

By the hypothesis, $w_i^\top f(x(t)) = g_i\big(w_i^\top x(t)\big)$; hence,

$$f\big(x(t)\big) = \sum_{i=1}^{n} g_i\big(w_i^\top x(t)\big) w_i.$$

Because the set $\{w_i\}$ is linearly independent, the equality

$$f\big(x(t)\big) = 0$$

holds if and only if

$$g_i\big(w_i^\top x(t)\big) = 0 \quad \text{for each } i.$$

Define $\delta = \min\{\delta_1, \delta_2, \ldots, \delta_n\}$. Then if $\|x(0)\| < \delta$, we have $|w_i^\top x(0)| < \delta$ for every i. By the stability of each scalar system, it follows that

$$\big|w_i^\top x(t)\big| < \varepsilon \quad \text{for all } t \geq 0,$$

and therefore,

$$\|x(t)\|^2 = \sum_{i=1}^{n} \big(w_i^\top x(t)\big)^2 < n\varepsilon^2.$$

Thus, by choosing ε sufficiently small, we see that the full state $x(t)$ remains arbitrarily close to the origin, implying that the equilibrium $x = 0$ is Lyapunov stable. □

3 Nesterov Dynamics for Quadratic Games

First, we will prove the stability of the dynamics with a low-rank pseudogradient. To incorporate Lyapunov stability for this dynamical system, it is crucial to map the entirety long run behavior of x to 0. Let $w \in \mathcal{N}(\mathcal{G})$ such that $w \neq \mathbf{0}$. The existence of this vector is guaranteed since the gradient is low-rank. Hence, $w^\top \mathcal{G} = \mathbf{0}$. We can project the entire system onto w:

$$w^\top \ddot{x} + \frac{3}{t} w^\top \dot{x} + \cancelto{0}{w^\top \mathcal{G} x_1} = 0.$$

A similar transformation as before can reformulate the second order system i

$$\begin{cases} w^\top \dot{x_1} = w^\top x_2 \\ w^\top \dot{x_2} = -\dfrac{3}{t} w^\top x_2 \end{cases}$$

With linearity, we obtain:

$$\begin{cases} \dfrac{d}{dt}(w^\top x_1) = w^\top x_2 \\ \dfrac{d}{dt}(w^\top x_2) = -\dfrac{3}{t} w^\top x_2 \end{cases} .$$

Now let $y_1 = w^\top x_1$ and $y_2 = w^\top x_2$. Note that both values are now scalars.

$$\begin{cases} \dfrac{d}{dt}(y_1) = y_2 \\ \dfrac{d}{dt}(y_2) = -\dfrac{3}{t} y_2 \end{cases}$$

In this scenario, we can proceed to solve the differential equation. First obtain the following for y_2:

$$y_2(t) = \frac{1}{t^3} y_2(t_0)$$

which allows the derivation for y_1:

$$y_1(t) = \int_{t_0}^{t} y_2(t)dt + y_1(t_0) = -\Big(\frac{2}{t^2} - \frac{2}{t_0^2}\Big) y_2(t_0) + y_1(t_0)$$

And thus we can model the the system in the form

$$\mathbf{y}(t) = \mathbf{\Phi}(t, t_0)\mathbf{y}(t_0)$$

where

$$\mathbf{\Phi}(t, t_0) = \begin{bmatrix} 1 & -\Big(\dfrac{2}{t^2} - \dfrac{2}{t_0^2}\Big) \\ 0 & \dfrac{1}{t^3} \end{bmatrix}$$

Fixing $t_0 \in \mathbb{R}$, it is clear that $\exists \gamma > 0$ such that $|\mathbf{\Phi}(t, t_0)| < \gamma, \forall t > t_0$, which proves uniform stability, but not asymptotic stability.

The computation of concludes that the position vector x_1 will converge uniformly to a vector in the null space of $\mathcal{G}$.

For the full-rank case, we need to borrow some ideas and invoke Lemma 5.15 from [2]. We assume the linear time-varying system takes the form of $\dot{x} = f(x, u(t))$, where $u(t) = \dfrac{3}{t}$. Since $u(t)$ is bounded and is continuously differential in the domain of t, it is possible to interpret this system to be slowly-varying. In other words, the lack of variation of the time-varying factor allow analysis at separate time steps in order to make conclusions regarding the entire system.

Due to the nature of the scenario, it is possible to perform a frozen-time analysis of the system. However, since it is not possible to generalize the analysis, let α be a specific control input $u(t)$ and $x = h(\alpha)$ be the equilibrium point for different values of α. If certain properties of x hold uniformly in α, then it is plausible to assume the same properties hold for the system $\dot{x} = f(x, u(t))$.

To analyze stability of the frozen equilibrium point $x = h(\alpha)$, we shift it to the origin via the change of variables $z = x - h(\alpha)$ and obtain

$$\dot{z} = f(z + h(\alpha), \alpha) := g(z, \alpha)$$

Note that $z = 0 \implies x = h(\alpha)$. For our system, if $\mathcal{G}$ is full-rank, then $A(t)$ must be full rank $\forall t$. Thus, the only equilibrium for any t is $x = 0$, which would

prove exponential stability for the NAGD system. We directly quote Lemma 5.15.

Lemma: Consider the system $\dot{z} = A(\alpha)z$, where a $\alpha \in \Gamma \subset \mathbb{R}^m$ and $A(\alpha)$ is continuously differentiable. Suppose the elements of A and their first partial derivatives with respect to α are uniformly bounded; that is,

$$\|A(t)\|_2 < c, \left\| \frac{\partial}{\partial \alpha_i} A(t) \right\|_2 < b_i, \forall \alpha \in \Gamma, \forall i = 1, \ldots, m$$

Suppose further that $A(\alpha)$ is Hurwitz uniformly in α; that is,

$$Re[\lambda(A(\alpha))] < -\sigma < 0, \forall \alpha \in \Gamma$$

Then, the Lyapunov equation $PA(\alpha) + A^\top(\alpha)P = -I$ has a unique positive definite solution $P(\alpha)$ for every $\alpha \in \Gamma$ such that $V(z, \alpha) = z^\top P(\alpha) z$ ensures that the origin of the new systems, $z = 0$, is exponentially stable.

Since our control input is simply in $\mathbb{R}$ and equilibrium is always $x = 0$, the renaming of variables is not needed. We revert the system back to the form $\dot{x} = A(t)x$.

We enforce a condition on $\lim_{t\to\infty} A(t) = A^*$. Let A^* be Hurwitz. Then there exists t_0 such that if $t > t_0$, then the all of the eigenvalues of $A(t)$ are sufficiently close to A^*, thus making $A(t)$ Hurwitz for $t \in (t_0, \infty)$.

Given this, we reformulate our problem to the desired format:

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \mathbf{0} & I \\ -\mathcal{G} & -\frac{3}{t}I \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Let $A(t) = \begin{bmatrix} \mathbf{0} & I \\ -\mathcal{G} & -\frac{3}{t}I \end{bmatrix}$, and it is evident that the $A^* = \begin{bmatrix} \mathbf{0} & I \\ -\mathcal{G} & \mathbf{0} \end{bmatrix}$. Solving for the eigenvalues of A^*, we obtain that they must satisfy the equation:

$$\det(-\lambda I)\det\Big(-\lambda I - \frac{1}{\lambda}\mathcal{G}\Big) = (-\lambda)^n \det\Big(\frac{1}{\lambda} I(-\mathcal{G} - \lambda^2 I)\Big) = \det(-\mathcal{G} - \lambda^2 I) = 0$$

An issue rises regarding this formula. λ^2 must take the form of eigenvalues of $-\mathcal{G}$. For $N \geq 2$, all possible values of λ will not have a strictly negative real part. Thus, another approach is required.

Let v be a right eigenvector of $\mathcal{G}$ with eigenvalue λ. Consider projecting the system to a right eigenvector w such that $v^\top w = 0$. Just like the low-rank case, we can simplify the problem to a second-order system in $\mathbb{R}$.

$$\begin{cases} \frac{d}{dt}(w^\top x_1) = w^\top x_2 \\ \frac{d}{dt}(w^\top x_2) = -\frac{3}{t} w^\top x_2 - w^\top \mathcal{G} x_1 = -\frac{3}{t} w^\top x_2 - \lambda w^\top x_1 \end{cases}$$

Now let $y_1 = w^\top x_1$ and $y_2 = w^\top x_2$. Note that both values are now scalars.

$$\begin{cases} \frac{d}{dt}(y_1) = y_2 \\ \frac{d}{dt}(y_2) = -\lambda y_1 - \frac{3}{t} y_2 \end{cases}$$

And such a system can be formulated as a single second-order nonlinear ODE. Thankfully, it is simplified to a single variable form $y : (t_0, \infty) \to \mathbb{R}$

$$\ddot{y} + \frac{3}{t}\dot{y} + \lambda y = 0$$

Such ODE yields the solution

$$y(t) = \frac{c_1}{t} J_1(\sqrt{\lambda}t) + \frac{c_2}{t} Y_1(\sqrt{\lambda}t), \text{where } \sqrt{\lambda} \in \mathbb{R}$$

otherwise

$$y(t) = \frac{c_1}{t} J_1(\sqrt{\lambda}t) + \frac{c_2}{t} Y_1(-\sqrt{\lambda}t)$$

Where $J_1(t)$ and $Y_1(t)$ denote Bessel functions of the first and second kind, respectively, with parameter $n = 1$, and c_1, c_2 are constants chosen depending on the initial condition $y(t_0), y'(t_0)$.

Consider the long-run behavior of this solution. The asymptotic form of each Bessel function for $|\sqrt{\lambda}t|$ very large is as follows [1].

$$J_1(\sqrt{\lambda}t) \sim \sqrt{\frac{2}{\pi\sqrt{\lambda}t}}\Big(\cos\Big(\sqrt{\lambda}t - \frac{3\pi}{4}\Big) + e^{|\Im(\sqrt{\lambda}t)|}\mathcal{O}\Big(\big|\sqrt{\lambda}t\big|^{-1}\Big)\Big)$$

$$Y_1(\sqrt{\lambda}t) \sim \sqrt{\frac{2}{\pi\sqrt{\lambda}t}}\Big(\sin\Big(\sqrt{\lambda}t - \frac{3\pi}{4}\Big) + e^{|\Im(\sqrt{\lambda}t)|}\mathcal{O}\Big(\big|\sqrt{\lambda}t\big|^{-1}\Big)\Big)$$

where the phase $|\arg\sqrt{\lambda}| < \pi$ and $\Im$ is the imaginary part of the complex number.

The exponential term in each asymptotic form is activated if there is a complex part, which forces λ to be strictly positive and real in order to prevent the trajectory from diverging.

If these conditions are met, then $\lim_{t\to\infty} y(t) = 0$. The asymptotics imply that $y(t)$ does not decay exponentially, rather $\mathcal{O}(t^{-1/2})$ at best.

Now, why did we invoke Lemma 5.15? Well, the requirements for stability in a projected domain are similar to that of a similar, intuitive approach.

Reconsider the projected system,

$$\begin{cases} \dfrac{d}{dt}(y_1) = y_2 \\ \dfrac{d}{dt}(y_2) = -\lambda y_1 - \dfrac{3}{t} y_2 \end{cases}$$

and reformat it to a state-space model $\dot{y} = A(t)y$. Then $A(t) = \begin{bmatrix} 0 & 1 \\ -\lambda & -\frac{3}{t} \end{bmatrix}$ If we apply lemma 5.15 for this system, then it is impossible to prove exponential stability since the characteristic is $\gamma^2 + \lambda = 0$. Strictly positive values of λ would lead to two complex root, and otherwise there will be a real part in at least one of the roots.

Assume the former case holds our system and analyze which direction the real part of the root converges to 0. Our characteristic equation is $\gamma^2 + \frac{3}{t}\gamma + \lambda = 0$, where γ is the eigenvalue of the projected system and λ is eigenvalue constant provided from the initial projection. The roots, $r(t)$, change over t:

$$r(t) = \frac{-\frac{3}{t} \pm \sqrt{\frac{9}{t^2} - 4\lambda}}{2}$$

λ is a finite constant, so $\exists t_1 \in (t_0, \infty)$ such that $\sqrt{\frac{9}{t^2} - 4\lambda}, \forall t > t_1$. This means that for $t \in (t_1, \infty)$, the real part of $r(t)$ is negative, but converges to 0.

In essence, we obtain that $A(t)$ is Hurwitz $\forall t \in (t_1, \infty)$. Intuitively, of course, this shouldn't ensure exponential stability, but it still offers a weaker sense of stability: the same stability as the explicit solution analysis.

The selection of $\frac{3}{t}$ as the momentum factor allows continual convergence over time, while any function such that $\int_{t_0}^{\infty} f(t)dt < \infty$ converges to the solution of $\ddot{y} + \lambda y = 0$ because the $\dot{y}$ decays too fast.

4 Simulated Examples

The previous section summarized the behavior for the dynamics of Nesterov's accelerated gradient descent as a partial differential equation when applied for 2 player quadratic games. Below are simulated trajectories on MATLAB via the Runge-Kutta method with step sizes of 0.01 and from $t \in [1, 1000]$. The blue and orange curves represent the position and momentum vectors, respectively.

Figure 3 and Figure 4 are trajectories are for low rank $\mathcal{G}$. Clearly, the position vectors converge towards the null space of $\mathcal{G}$. The vector is seen to converge to 0 just as (1) predicts. This establishes Lyapunov stability for the momentum.

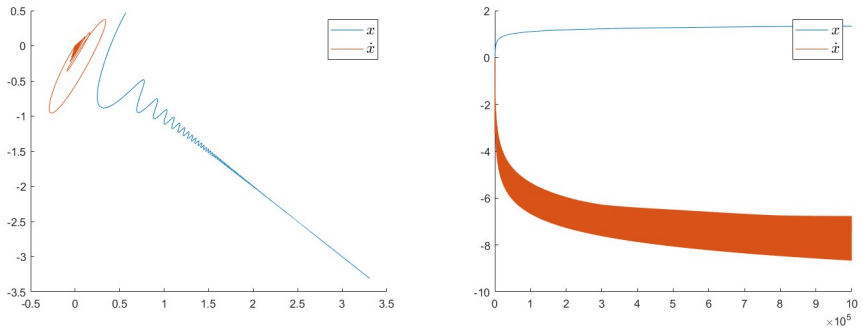

Figure 3: Nesterov dynamics with $\mathcal{G} = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$ and log10-norm of trajectory

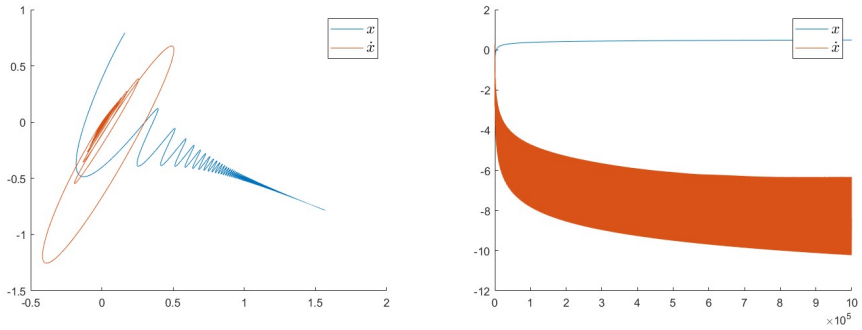

Figure 4: Nesterov dynamics with $\mathcal{G} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$ and log10-norm of trajectory

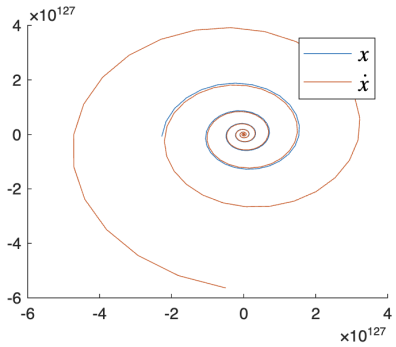

Figure 5: $\mathcal{G} = \begin{bmatrix} 6 & 1.5 \\ -1.5 & 6 \end{bmatrix}$ and $\lambda = 6 + 1.5i, 6 - 1.5i$

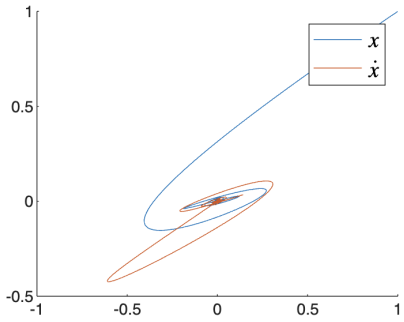

Figure 6: $\mathcal{G} = \begin{bmatrix} 0.6557 & 0.8491 \\ 0.0357 & 0.9340 \end{bmatrix}$ and $\lambda = 0.5720, 1.0178$

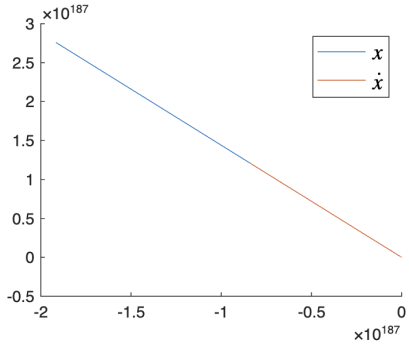

Figure 7: $\mathcal{G} = \begin{bmatrix} 0.9572 & 0.8003 \\ 0.4854 & 0.1419 \end{bmatrix}$ and $\lambda = 1.2942, -0.1952$

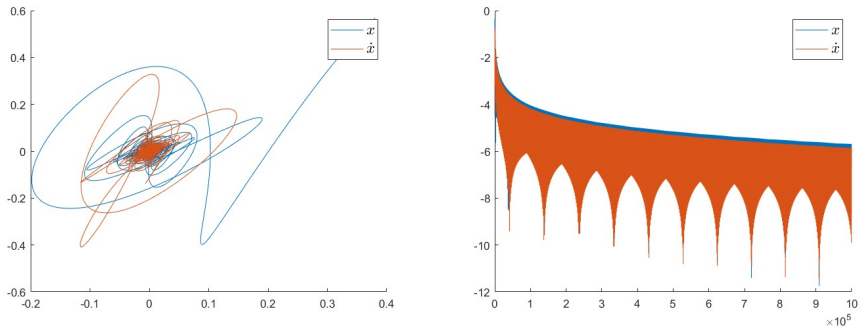

Figure 8: Nesterov dynamics with $\mathcal{G} = \begin{bmatrix} 0.4 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$ and log10-norm of trajectory

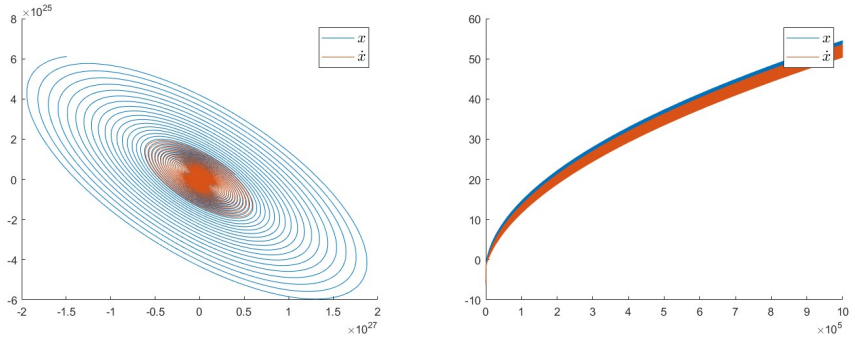

Figure 9: Nesterov dynamics with $\mathcal{G} = \begin{bmatrix} 0.1 & -0.1 \\ 0 & 0.1 \end{bmatrix}$ and log10-norm of trajectory

References

[1] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth dover printing, tenth gpo printing edition, 1964.

[2] Hassan K Khalil. *Nonlinear systems; 3rd ed.* Prentice-Hall, Upper Saddle River, NJ, 2002. The book can be consulted by contacting: PH-AID: Wallet, Lionel.

[3] Weijie Su, Stephen Boyd, and Emmanuel J. Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights, 2015.