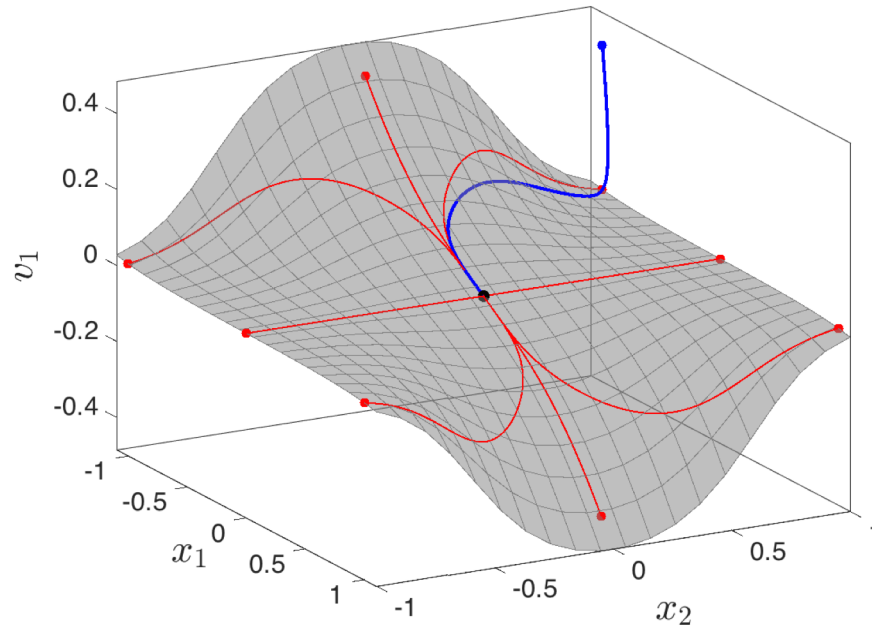


# Accelerated Gradient Optimization: A Multiscale Analysis



**Mohammad Farazmand**

Department of Mathematics

NC State University

# Problem Set-up

$$\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$$

$$f : \mathcal{X} \rightarrow \mathbb{R}$$

$$f(\mathbf{x}) \geq f(\mathbf{x}^*), \quad \mathbf{x}^* \in \mathcal{X}$$

## Assumptions:

- Convex set  $\mathcal{X} \subseteq \mathbb{R}^n$

- Convex functions:

$$f(\theta \mathbf{x} + (1 - \theta) \mathbf{y}) \leq \theta f(\mathbf{x}) + (1 - \theta) f(\mathbf{y}) \quad \begin{array}{l} \forall \mathbf{x}, \mathbf{y} \in \mathcal{X} \\ 0 \leq \theta \leq 1 \end{array}$$

- At least once continuously differentiable

## Recall:

- Strongly convex functions:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{m}{2} \|\mathbf{y} - \mathbf{x}\|^2 \quad \begin{array}{l} \forall \mathbf{x}, \mathbf{y} \in \mathcal{X} \\ m > 0 \end{array}$$

# Overview

Method	Discrete	Convergence	Continuous
<b>Gradient descent</b>	$\mathbf{x}_{k+1} = \mathbf{x}_k - s \nabla f(\mathbf{x}_k)$	$\mathcal{O}\left(\frac{1}{k}\right)$	$\dot{\mathbf{x}} = -\nabla f(\mathbf{x})$
<b>Polyak's heavy ball descent</b>	$\mathbf{y}_k = \mathbf{x}_k + \lambda(\mathbf{x}_k - \mathbf{x}_{k-1})$ $\mathbf{x}_{k+1} = \mathbf{y}_k - s \nabla f(\mathbf{x}_k)$ [Polyak (1964)]	$\mathcal{O}\left(\frac{1}{k}\right)$	$\ddot{\mathbf{x}} = -\mu\dot{\mathbf{x}} - \eta\nabla f(\mathbf{x})$ [Polyak (1964)]
<b>Nesterov's Accelerated gradient descent</b>	$\mathbf{y}_k = \mathbf{x}_k + \lambda_k(\mathbf{x}_k - \mathbf{x}_{k-1})$ $\mathbf{x}_{k+1} = \mathbf{y}_k - s_k \nabla f(\mathbf{y}_k)$ [Nesterov (1983)]	$\mathcal{O}\left(\frac{1}{k^2}\right)$	$\ddot{\mathbf{x}} = -\frac{\rho}{t}\dot{\mathbf{x}} - \nabla f(\mathbf{x})$ [Su et al. JMLR (2016)]

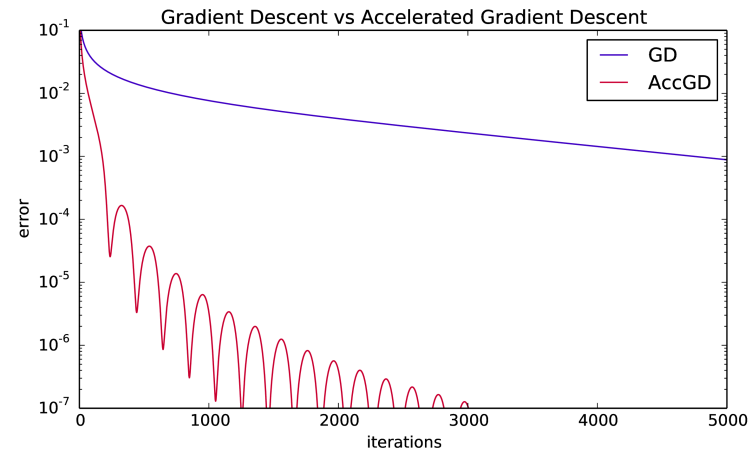
**Goal:** Understanding the phase space structure of the continuous systems

# Comments on Accelerated Methods

- Accelerated methods are **not monotonic**

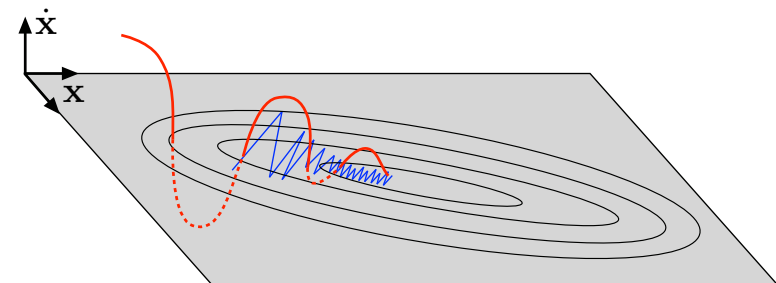
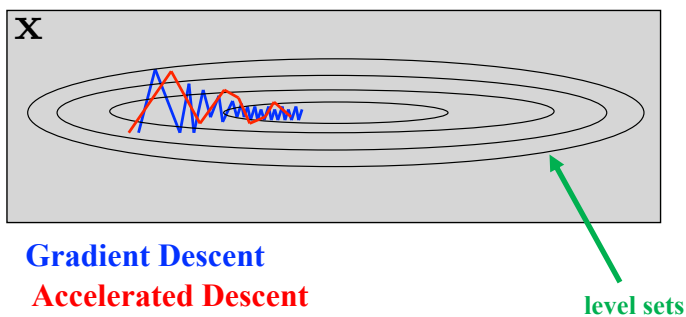
$$\ddot{\mathbf{x}} = -\mu\dot{\mathbf{x}} - \eta\nabla f(\mathbf{x})$$

$$\ddot{\mathbf{x}} = -\frac{\rho}{t}\dot{\mathbf{x}} - \nabla f(\mathbf{x})$$



[Source: Moritz Hardt, Berkeley]

- Higher dimensional phase space



Must be viewed in the phase space

$$\dot{\mathbf{x}} = -\nabla f(\mathbf{x})$$

$$\ddot{\mathbf{x}} = -\mu\dot{\mathbf{x}} - \eta\nabla f(\mathbf{x})$$

# Variational Formulation

Lagrangian Function: [\[Wibisono et al. PNAS \(2016\)\]](#)

$$\mathcal{L}(\mathbf{x}, \dot{\mathbf{x}}, t) = e^{\alpha t + \gamma t} \left( \frac{1}{2} \|e^{-\alpha t} \dot{\mathbf{x}}\|^2 - e^{\beta t} f(\mathbf{x}) \right)$$

Euler-Lagrange Eq:

$$\ddot{\mathbf{x}} + (e^{\alpha t} - \dot{\alpha}_t) \dot{\mathbf{x}} + e^{2\alpha t + \beta t} \nabla f(\mathbf{x}) = 0$$

Special Cases:

$$\left. \begin{aligned} \alpha_t &= \log \left[ \frac{\mu}{1 + (\mu e^{-\alpha_0} - 1) e^{\mu t}} \right] \\ \beta_t &= \log \eta - 2\alpha_t \end{aligned} \right\}$$

**Polyak's method**

$$\ddot{\mathbf{x}} = -\mu \dot{\mathbf{x}} - \eta \nabla f(\mathbf{x})$$

$$\left. \begin{aligned} \alpha_t &= \log((\rho - 1)/t) \\ \beta_t &= -2 \log((\rho - 1)/t) \end{aligned} \right\}$$

**Nesterov's method**

$$\ddot{\mathbf{x}} = -\frac{\rho}{t} \dot{\mathbf{x}} - \nabla f(\mathbf{x})$$

# Motivation for Multiscale Analysis

Polyak's method

$$\ddot{\mathbf{x}} + \mu\dot{\mathbf{x}} + \eta\nabla f(\mathbf{x}) = 0$$

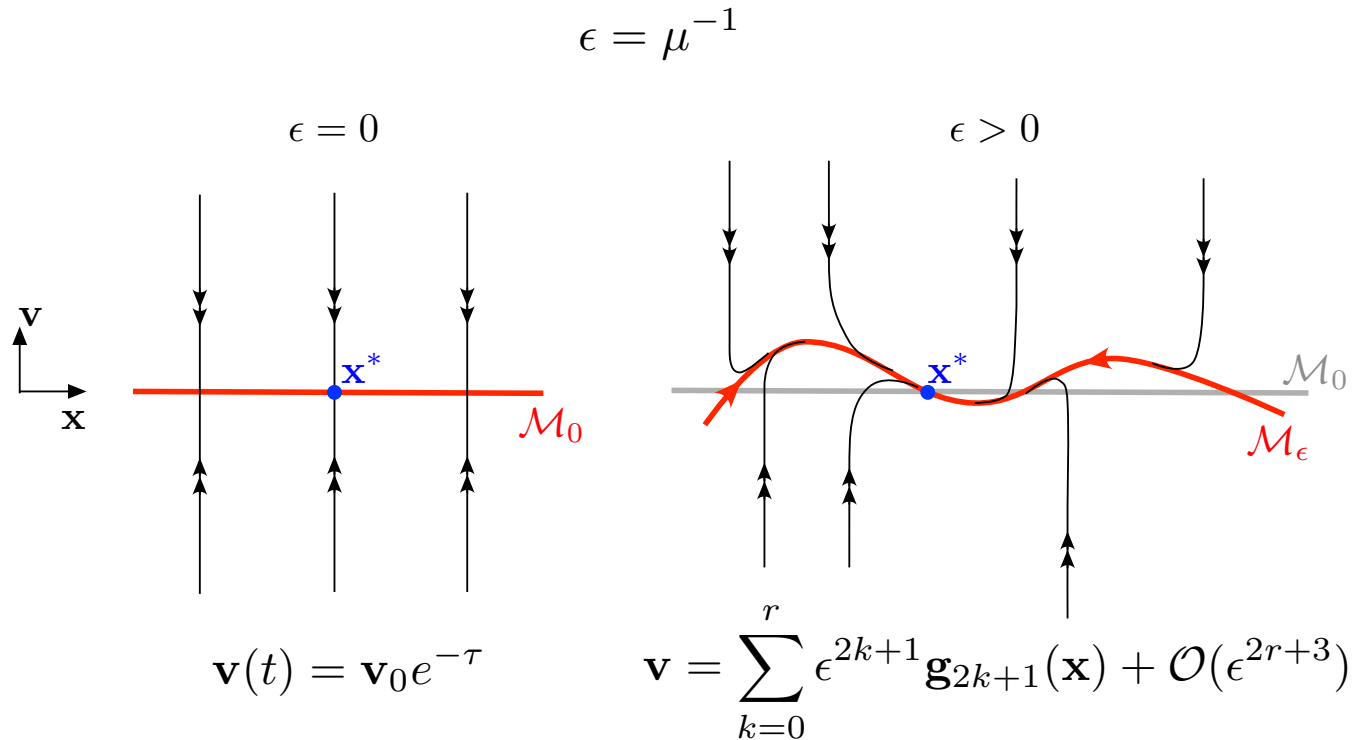
- Analogous to a damped mechanical system
- Do trajectories settle to an invariant manifold?
- If so, what is the shape of that manifold?
- What is the dynamics on the manifold?
- Fits in the framework of **geometric singular perturbation theory**

# Slow Manifold

## Polyak's method

$$\ddot{\mathbf{x}} = -\mu\dot{\mathbf{x}} - \eta\nabla f(\mathbf{x})$$

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{v} \\ \dot{\mathbf{v}} = -\mu\mathbf{v} - \eta\nabla f(\mathbf{x}) \end{cases} \xrightarrow{t = \epsilon\tau} \begin{cases} \mathbf{x}' = \epsilon\mathbf{v} \\ \mathbf{v}' = -\mathbf{v} - \epsilon\eta\nabla f(\mathbf{x}) \end{cases}$$



[M. Farazmand, Multiscale analysis of accelerated gradient methods (2019)]

# Slow Manifold

## Polyak's method

$$\ddot{\mathbf{x}} + \mu \dot{\mathbf{x}} + \eta \nabla f(\mathbf{x}) = 0$$

**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  $\mu$  and  $\eta$  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 fast 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

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

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

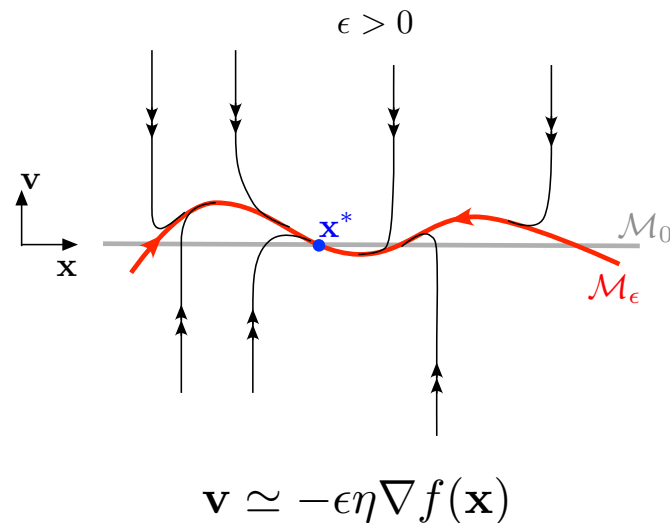
$$(2.4) \quad \begin{aligned} \mathbf{g}_1(\mathbf{x}) &= -\eta \nabla f(\mathbf{x}), \\ \mathbf{g}_{2k}(\mathbf{x}) &= \mathbf{0}, \\ \mathbf{g}_{2k+1}(\mathbf{x}) &= - \sum_{\ell=1}^{2k-1} \nabla \mathbf{g}_\ell(\mathbf{x}) \mathbf{g}_{2k-\ell}(\mathbf{x}), \quad k \geq 1. \end{aligned}$$

# First-order Approximation

Order  $\epsilon$  approximation:

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

(Gradient Descent)



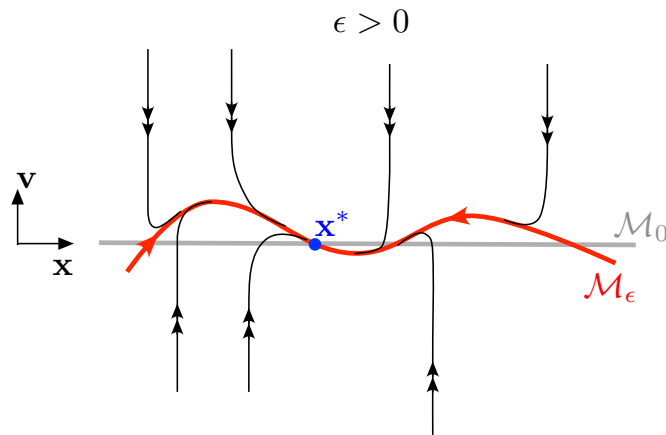
- The **minimizer** is its **asymptotically stable** fixed point

$$\mathcal{E}(\mathbf{x}) = \epsilon\eta (f(\mathbf{x}) - f(\mathbf{x}^*))$$

# Third-order Approximation

Order  $\epsilon^3$  approximation:

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



$$\mathbf{v} \simeq -\epsilon\eta\nabla f(\mathbf{x}) - \epsilon^3\eta^2\nabla^2 f(\mathbf{x})\nabla f(\mathbf{x})$$

- The **minimizer** is its **asymptotically stable** fixed point

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

[M. Farazmand, Multiscale analysis of accelerated gradient methods (2019)]

# Third-order Approximation

Order  $\epsilon^3$  approximation:

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

- **Preconditioned** gradient descent

$$\dot{\mathbf{x}} = -[P(\mathbf{x})]^{-1}\nabla f(\mathbf{x})$$

$$[P(\mathbf{x})]^{-1} = \epsilon\eta [I + \epsilon^2\eta\nabla^2 f(\mathbf{x})]$$

# Higher Order Approximation

Higher-order approximations:

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

- Smoothness  $f \in C^{r+1}(\mathcal{X})$
- Still the **minimizer** is its **asymptotically stable** fixed point

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

# Strong convexity for Polyak's method

Polyak's iteration vs. continuous system

- Polyak's iteration

$$\mathbf{y}_k = \mathbf{x}_k + \lambda(\mathbf{x}_k - \mathbf{x}_{k-1})$$

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

Requires **strong convexity** for convergence

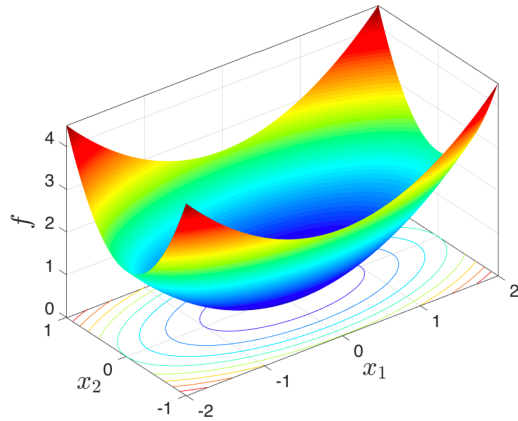
- Polyak's continuous system

$$\ddot{\mathbf{x}} = -\mu \dot{\mathbf{x}} - \eta \nabla f(\mathbf{x})$$

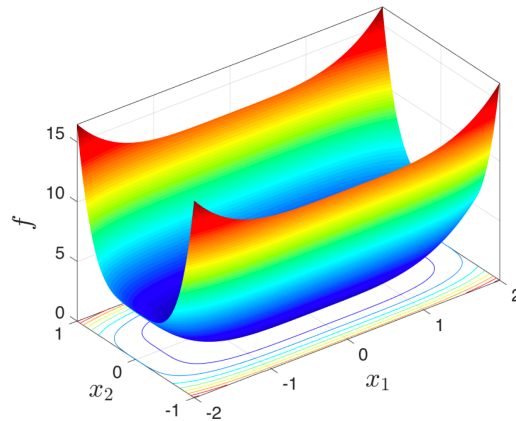
Only requires **convexity**

Lesson: Discretization is “tricky”

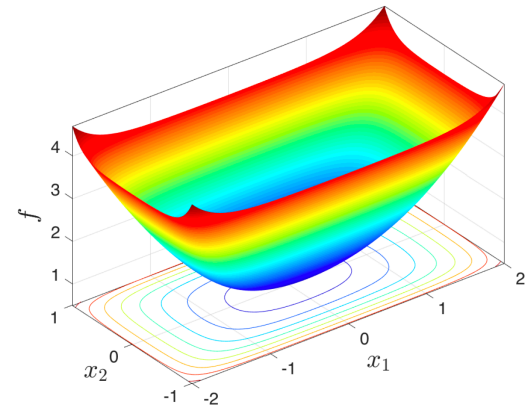
# Examples



(a) Example 1



(b) Example 2



(c) Example 3

$$\ddot{\mathbf{x}} = -\mu \dot{\mathbf{x}} - \eta \nabla f(\mathbf{x})$$

$$\ddot{\mathbf{x}} = -\frac{\rho}{t} \dot{\mathbf{x}} - \nabla f(\mathbf{x})$$

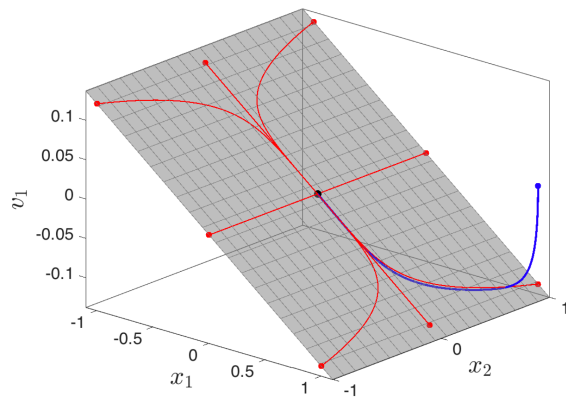
Example #	$f(\mathbf{x})$	Autonomous		Non-autonomous
		$\mu$	$\eta$	$\rho$
1	$\frac{1}{2} (x_1^2 + 5x_2^2)$	8	1	3
2	$\frac{1}{4} (x_1^4 + 50x_2^4)$	2	1	1.5
3	$\log(e^{x_1^2} + e^{4x_2^2})$	4	1	3

# Examples

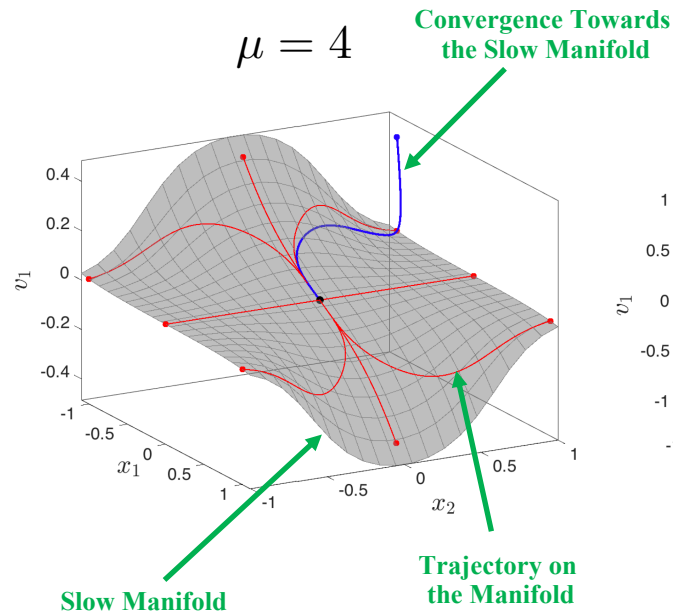
## Polyak's Eq.

$$\ddot{\mathbf{x}} + \mu \dot{\mathbf{x}} + \eta \nabla f(\mathbf{x}) = 0 \quad \eta = 1$$

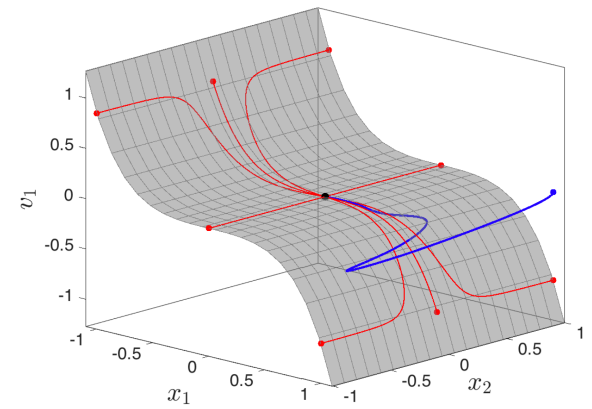
$\mu = 8$



$\mu = 4$



$\mu = 2$

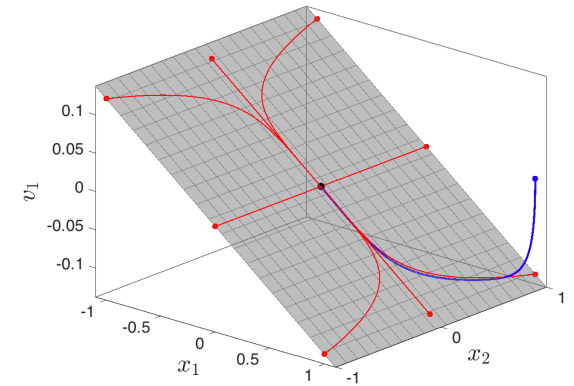


Projections from 4-D Phase Space

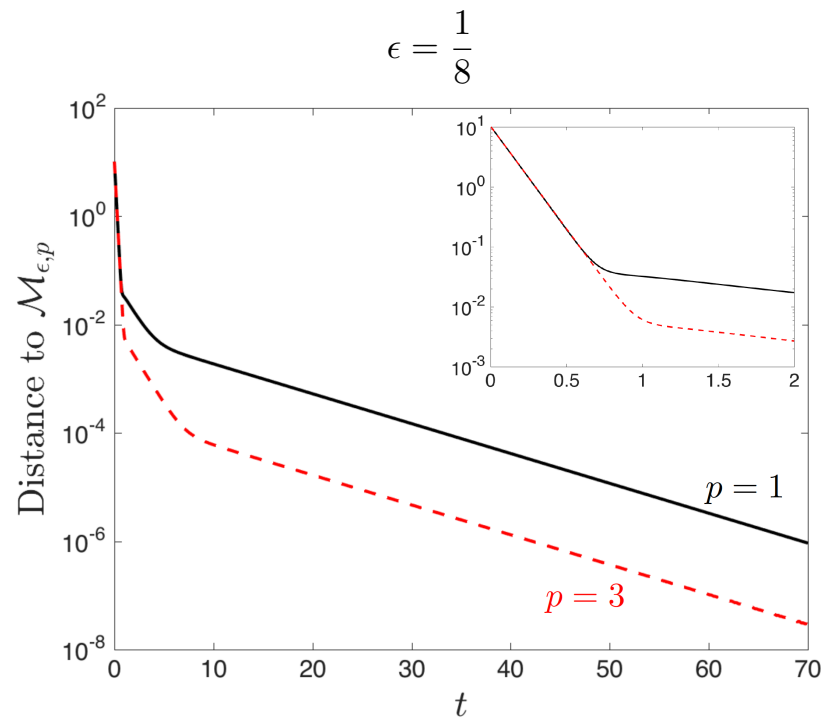
# Convergence to Slow Manifold

Approximate slow manifold:

$$\mathcal{M}_{\epsilon,p} : \mathbf{v}^{\epsilon,p}(\mathbf{x}) \triangleq \sum_{k=1}^p \epsilon^k \mathbf{g}_k(\mathbf{x})$$



Example 1

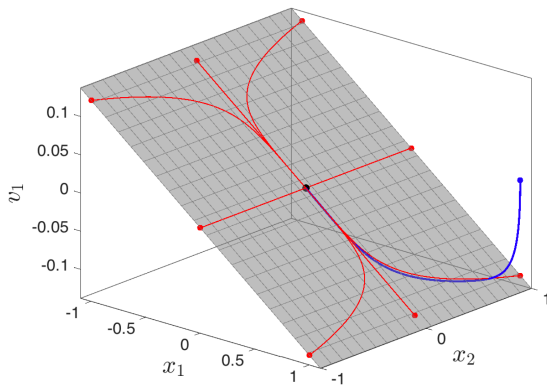


# Choice of Initial Conditions

## Polyak's method

Equation	$\ddot{\mathbf{x}} = -\mu\dot{\mathbf{x}} - \eta\nabla f(\mathbf{x})$	$\mathbf{y}_k = \mathbf{x}_k + \lambda(\mathbf{x}_k - \mathbf{x}_{k-1})$ $\mathbf{x}_{k+1} = \mathbf{y}_k - s\nabla f(\mathbf{x}_k)$
Initial conditions	$\mathbf{x}(0) = \mathbf{x}_0, \quad \dot{\mathbf{x}}(0) = \mathbf{v}_0$	$\mathbf{x}_0, \quad \mathbf{x}_1$

$$\mathcal{M}_{\epsilon,p} : \quad \mathbf{v}^{\epsilon,p}(\mathbf{x}) \triangleq \sum_{k=1}^p \epsilon^k \mathbf{g}_k(\mathbf{x})$$



Skip the convergence towards the slow manifold

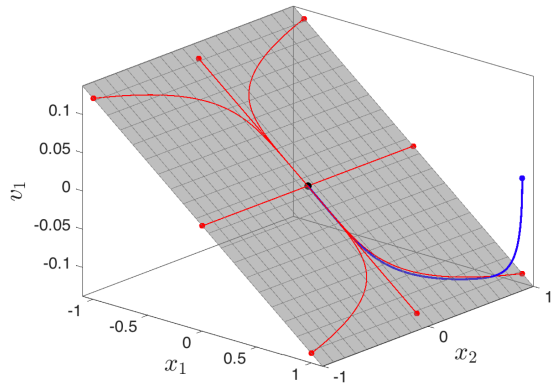
$$\mathbf{v}_0 = \mathbf{v}^{\epsilon,p}(\mathbf{x}_0)$$

$p = 1 \Rightarrow$  Requires only **one** gradient evaluation

$$\mathbf{v}_0 = -\epsilon\eta\nabla f(\mathbf{x}_0)$$

# Choice of Initial Conditions

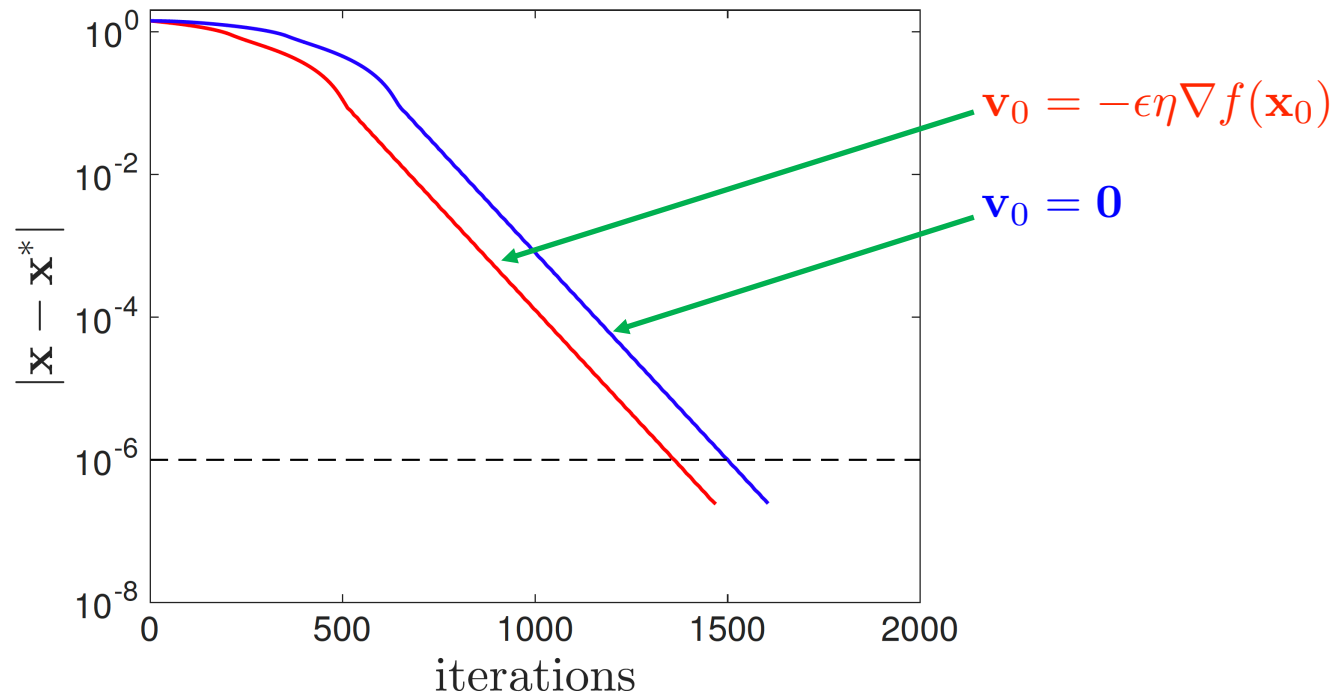
$$\mathcal{M}_{\epsilon,p} : \mathbf{v}^{\epsilon,p}(\mathbf{x}) \triangleq \sum_{k=1}^p \epsilon^k \mathbf{g}_k(\mathbf{x})$$



Faster convergence if

$$\mathbf{v}_0 = \mathbf{v}^{\epsilon,p}(\mathbf{x}_0)$$

Same initial position  $\mathbf{x}_0$



# Slow Manifold for Nesterov

**Nesterov's method**       $\ddot{\mathbf{x}} = -\frac{\rho}{t}\dot{\mathbf{x}} - \nabla f(\mathbf{x})$        $\mathbf{x}(\theta), \mathbf{v}(\theta), t(\theta)$

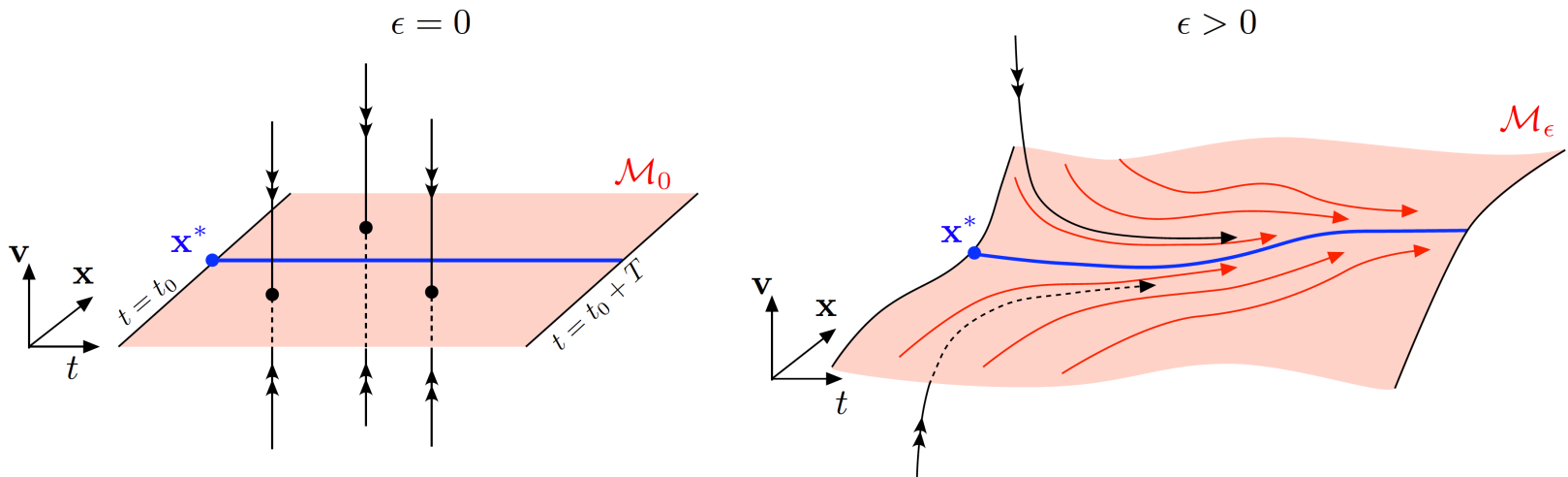
- Non-autonomous system
- Must be viewed in the **extended phase space**

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{v} \\ \dot{\mathbf{v}} = -\frac{\rho}{t}\mathbf{v} - \nabla f(\mathbf{x}) \\ \dot{t} = 1 \end{cases}$$

Rescaled time

$$\theta = \epsilon\tau$$

$$\epsilon = \rho^{-1}$$



[M. Farazmand, Multiscale analysis of accelerated gradient methods (2019)]

# Slow Manifold for Nesterov

Nesterov's method  $\ddot{\mathbf{x}} = -\frac{\rho}{t}\dot{\mathbf{x}} - \nabla f(\mathbf{x})$

**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 holds.

- (i) The trajectories of the Nesterov equation (1.3) converge exponentially fast 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

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

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

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

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

# Slow Manifold for Nesterov

Nesterov's method  $\ddot{\mathbf{x}} = -\frac{\rho}{t}\dot{\mathbf{x}} - \nabla f(\mathbf{x})$

First-order approximation:

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

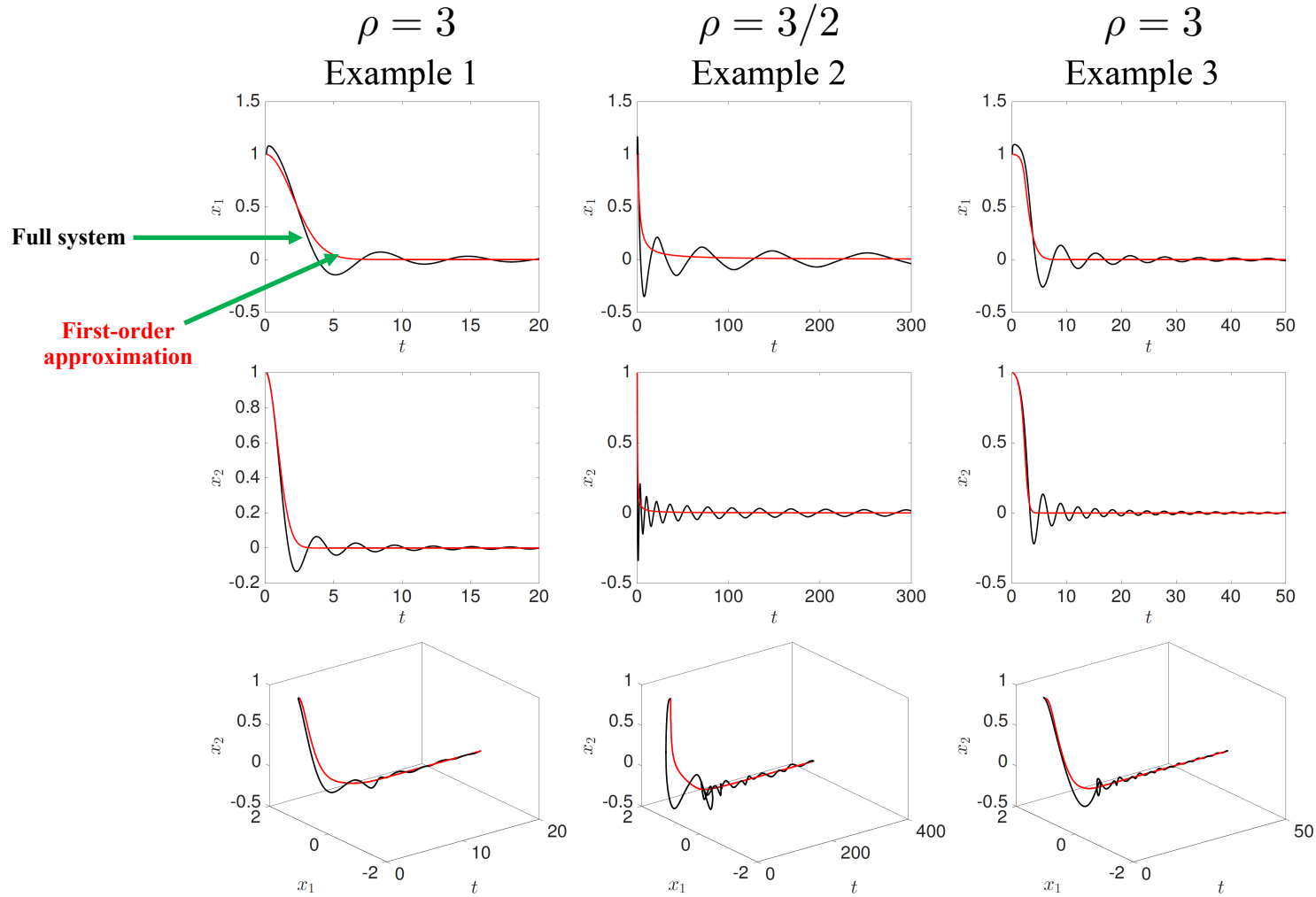
Rescaled time  $\hat{t} = \epsilon t^2/2$

In rescaled time:

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

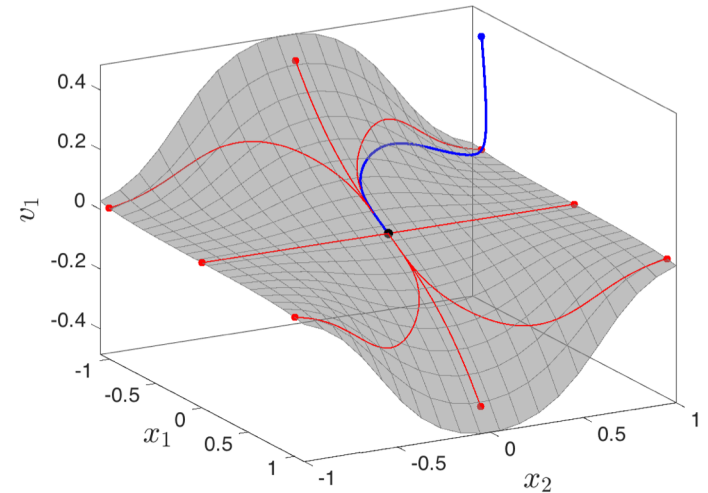
# Examples

Nestrov's method  $\ddot{\mathbf{x}} = -\frac{\rho}{t}\dot{\mathbf{x}} - \nabla f(\mathbf{x})$



# Summary

- Accelerated Gradient Descent (AGD) has a **multiscale structure**
- AGD converges to an **invariant slow manifold**
- Two stages:
  1. Convergence towards the slow manifold
  2. Reduced-order evolution on the slow manifold
- 
- Recursive formula for the slow manifold
  - First-order approximation: GD
  - Second-order approximation: Preconditioned GD
- Speed up AGD by judicious choice of the **initial data**



# Thank You

M. Farazmand, Multiscale analysis of accelerated gradient methods, arXiv:1807.11354v2 (2019)

[www.mfarazmand.com](http://www.mfarazmand.com)

Email: [farazmand@ncsu.edu](mailto:farazmand@ncsu.edu)