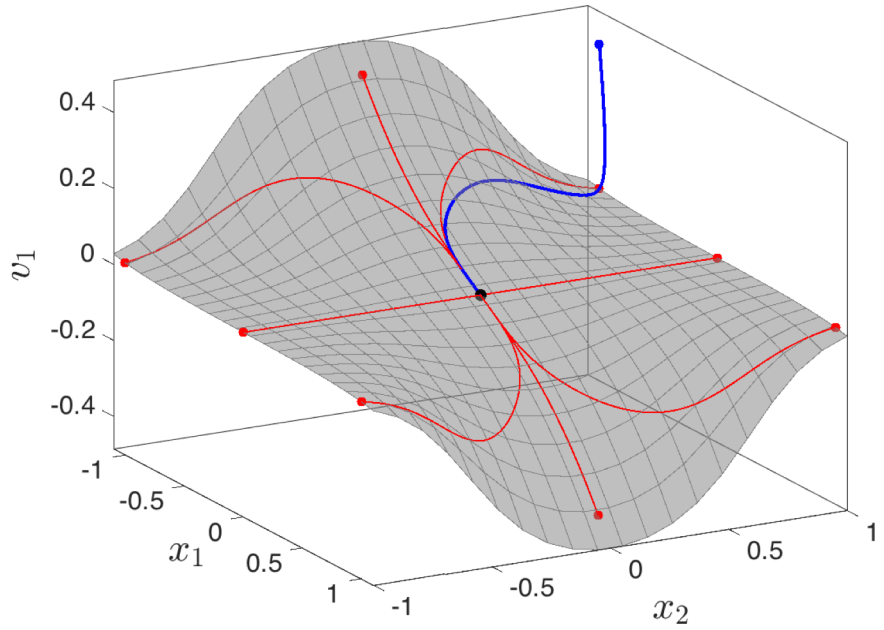
Accelerated Gradient Optimization: A Multiscale Analysis



Mohammad Farazmand
Department of Mathematics
NC State University

June 26, 2020

Outline

- Introduction (rather lengthy)
- Multiscale structure of AGD
  - Connection to **gradient descent** and **preconditioning**
  - Better choice of initial guess
- Examples
- Finally an open problem!

Problem Set-up

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

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

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

Assumptions:

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

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

- At least once continuously differentiable

Overview

Method	Discrete	Convergence	Continuous
Gradient descent	$\mathbf{x}_{k+1} = \mathbf{x}_k - s\nabla f(\mathbf{x}_k)$	$\mathcal{O}\left(\frac{1}{k}\right)$	$\dot{\mathbf{x}} = -\nabla f(\mathbf{x})$
Polyak's Accelerated gradient descent	$\mathbf{y}_k = \mathbf{x}_k + \lambda(\mathbf{x}_k - \mathbf{x}_{k-1})$ $\mathbf{x}_{k+1} = \mathbf{y}_k - s\nabla f(\mathbf{x}_k)$ [Polyak (1964)]	$\mathcal{O}\left(\frac{1}{k}\right)$	$\ddot{\mathbf{x}} = -\mu\dot{\mathbf{x}} - \eta\nabla f(\mathbf{x})$ [Polyak (1964)]
Nesterov's Accelerated gradient descent	$\mathbf{y}_k = \mathbf{x}_k + \lambda_k(\mathbf{x}_k - \mathbf{x}_{k-1})$ $\mathbf{x}_{k+1} = \mathbf{y}_k - s_k\nabla f(\mathbf{y}_k)$ [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)]

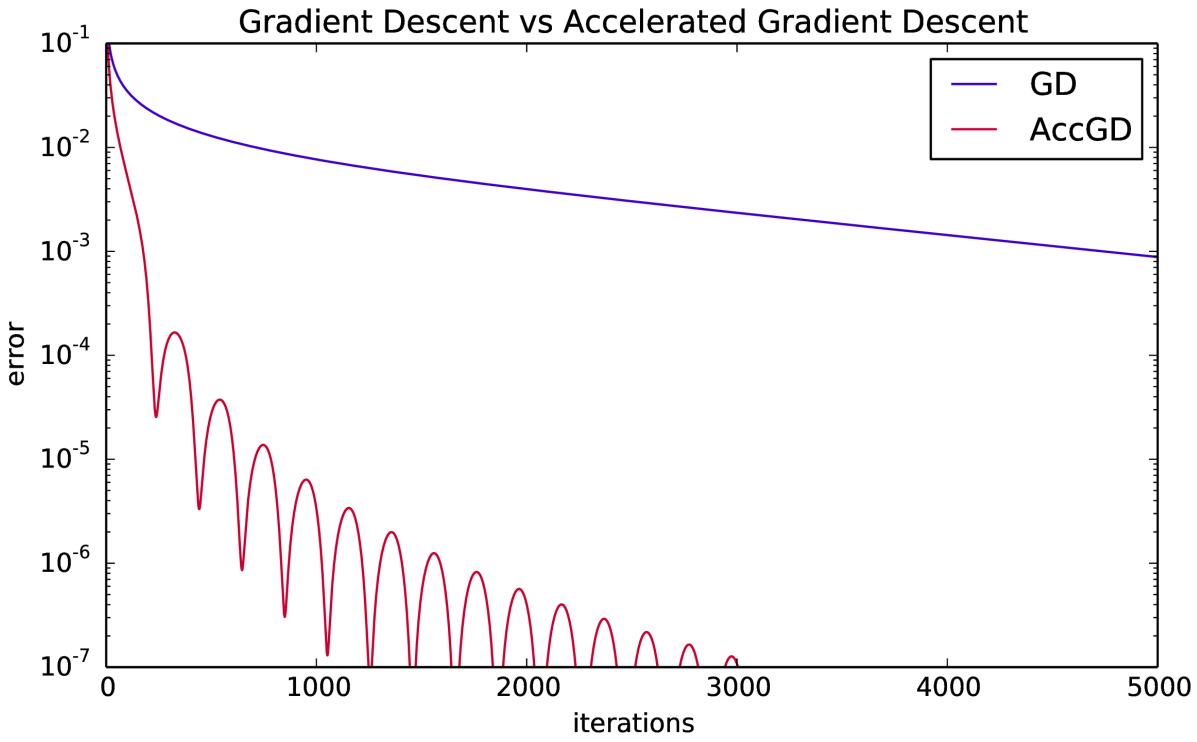
Goal: Understanding the phase space structure of the continuous systems

Comments on Accelerated Methods

Accelerated methods are not monotonic

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

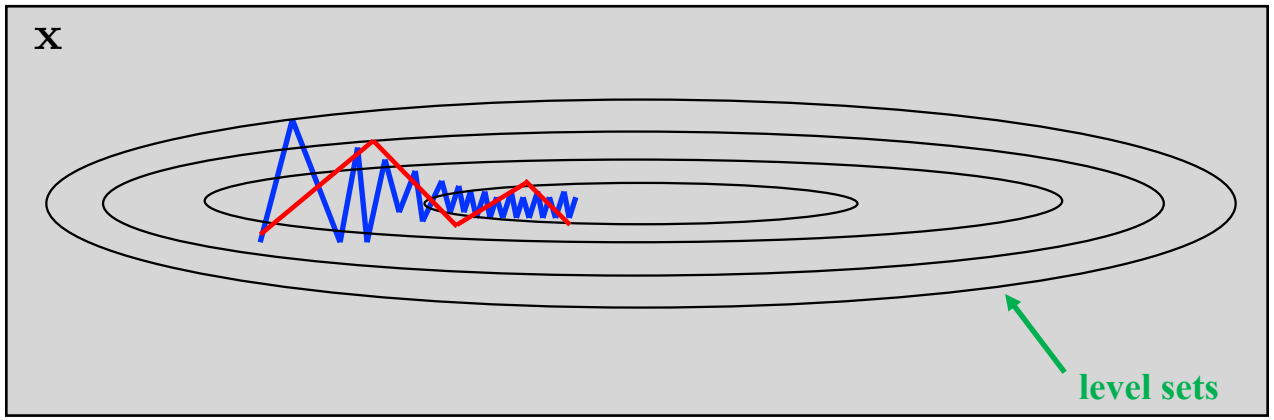
$$\ddot{\mathbf{x}} = -\frac{\rho}{t}\dot{\mathbf{x}} - \nabla f(\mathbf{x})$$



[Source: Moritz Hardt, Berkeley]

Comments on Accelerated Methods

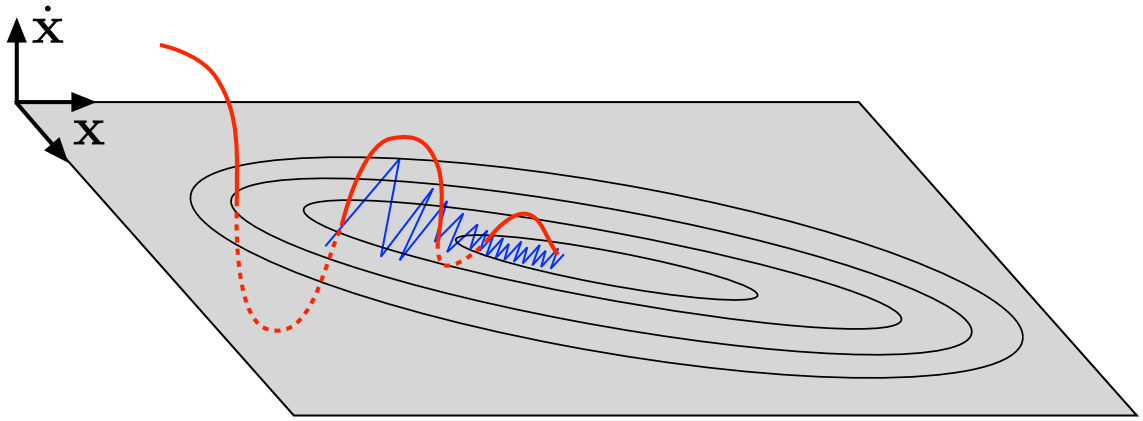
Higher dimensional phase space



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:

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

Special Cases:

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

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

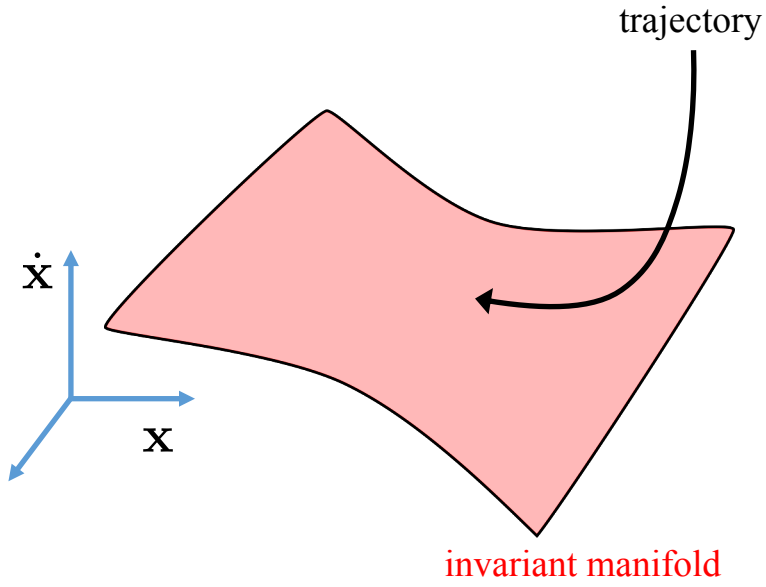
Polyak's method: Analogous to a damped mechanical system

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

acceleration

damping

potential force



- 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

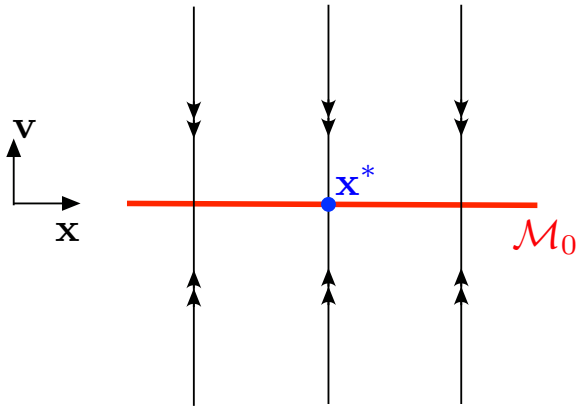
Polyak's method

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

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

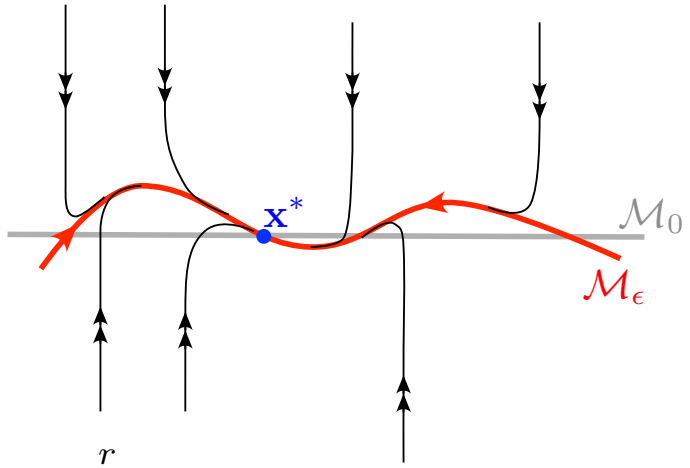
$$\epsilon = \mu^{-1}$$

$\epsilon = 0$



$$\mathbf{v}(t) = \mathbf{v}_0 e^{-\tau}$$

$\epsilon > 0$



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

[M. Farazmand, Multiscale analysis of accelerated gradient methods, SIAM J. Optimization (2020)]

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 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 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

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

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

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

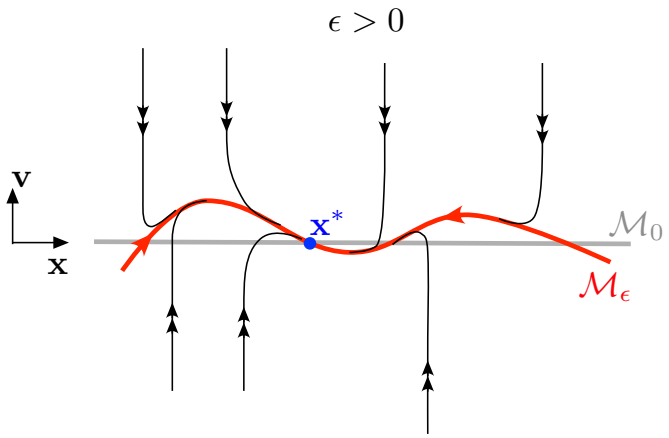
[M. Farazmand, Multiscale analysis of accelerated gradient methods, SIAM J. Optimization (2020)]

First-order Approximation

Order ϵ approximation:

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

(Gradient Descent)



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

- The minimizer is its asymptotically stable fixed point

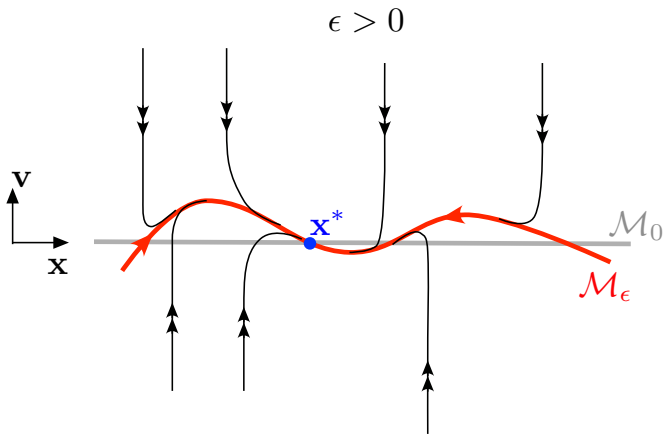
$$\mathcal{E}(\mathbf{x}) = \epsilon\eta\left(f(\mathbf{x}) - f(\mathbf{x}^*)\right)$$

[M. Farazmand, Multiscale analysis of accelerated gradient methods, SIAM J. Optimization (2020)]

Third-order Approximation

Order ϵ^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

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

[M. Farazmand, Multiscale analysis of accelerated gradient methods, SIAM J. Optimization (2020)]

Third-order Approximation

Order ϵ^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

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

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

[M. Farazmand, Multiscale analysis of accelerated gradient methods, SIAM J. Optimization (2020)]

Higher Order Approximation

Higher-order approximations:

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

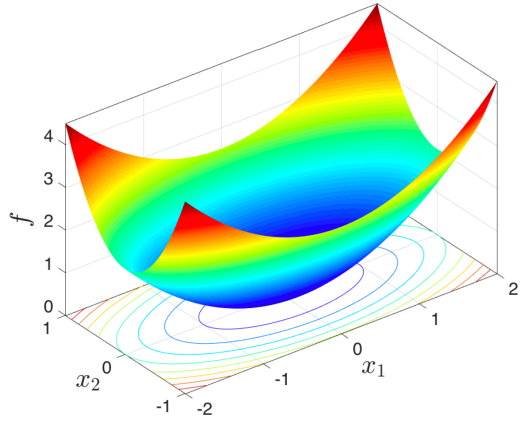
- Smoothness $f \in C^{r+1}(\mathcal{X})$

- Still the minimizer is its asymptotically stable fixed point

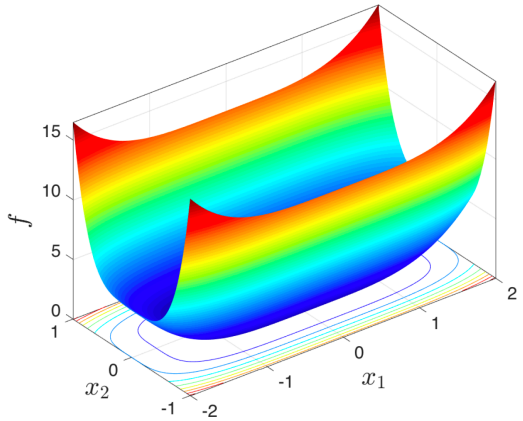
$$\mathcal{E}(\mathbf{x}) = \epsilon\eta \left(f(\mathbf{x}) - f(\mathbf{x}^*)\right) + \sum_{k=1}^{r} \frac{\epsilon^{2k+1}}{2} \sum_{\ell=1}^{2k-1} \langle \mathbf{g}_\ell(\mathbf{x}), \mathbf{g}_{2k-\ell}(\mathbf{x}) \rangle$$

[M. Farazmand, Multiscale analysis of accelerated gradient methods, SIAM J. Optimization (2020)]

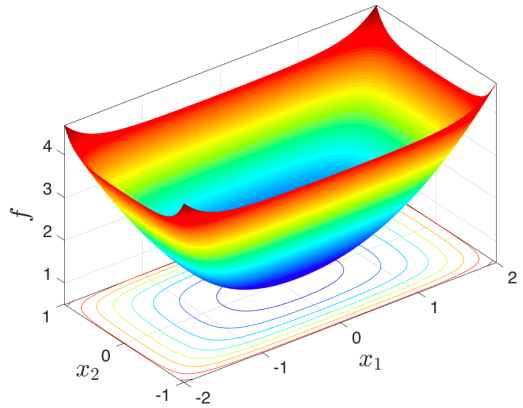
Examples



(a) Example 1



(b) Example 2



(c) Example 3

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

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

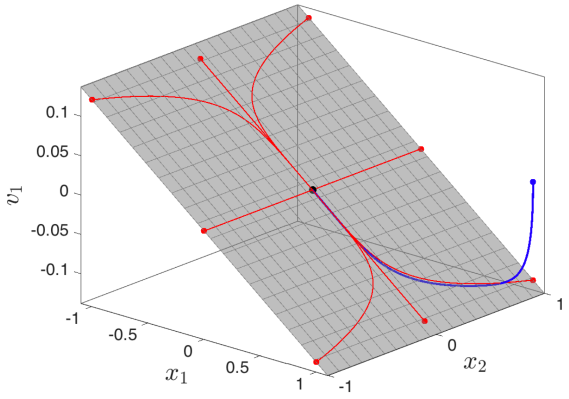
		Autonomous		Non-autonomous
Example #	$f(\mathbf{x})$	μ	η	ρ
1	$\frac{1}{2}\left(x_1^2 + 5x_2^2\right)$	8	1	3
2	$\frac{1}{4}\left(x_1^4 + 50x_2^4\right)$	2	1	1.5
3	$\log(e^{x_1^2} + e^{4x_2^2})$	4	1	3

Examples

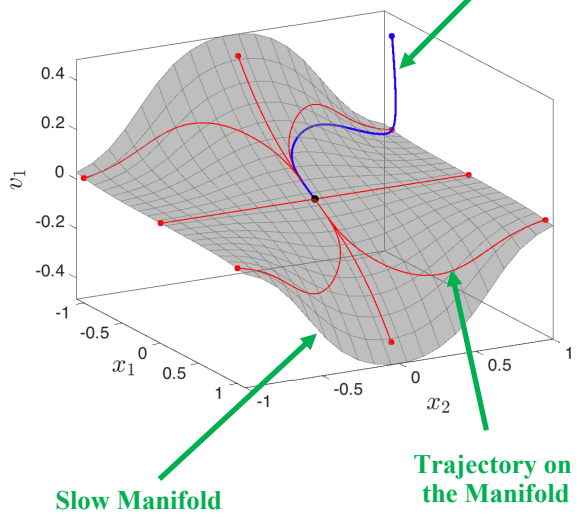
Polyak's Eq.

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

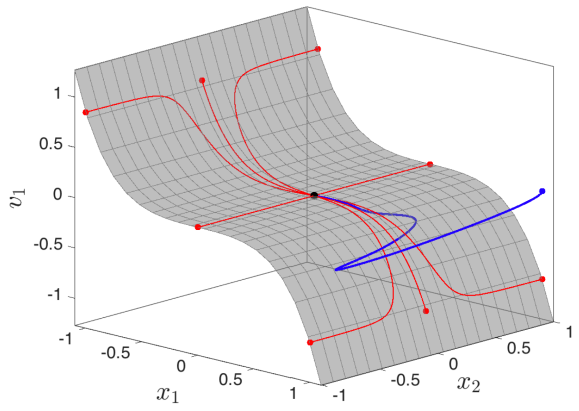
$\mu = 8$



$\mu = 4$



$\mu = 2$

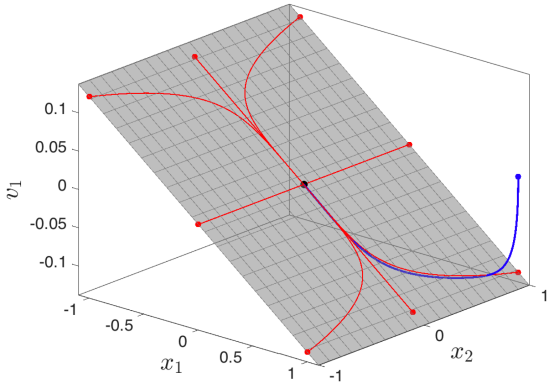


Projections from 4-D Phase Space

Convergence to Slow Manifold

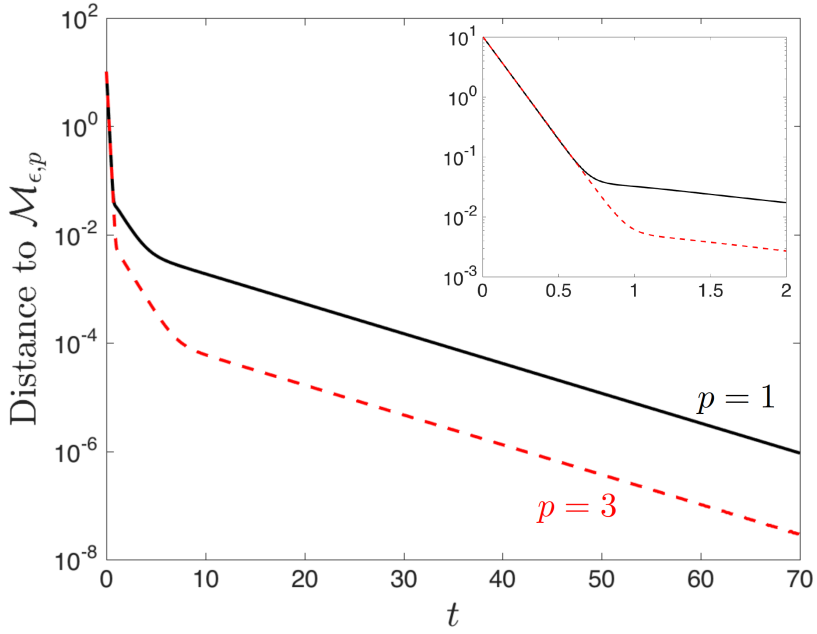
Approximate slow manifold:

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



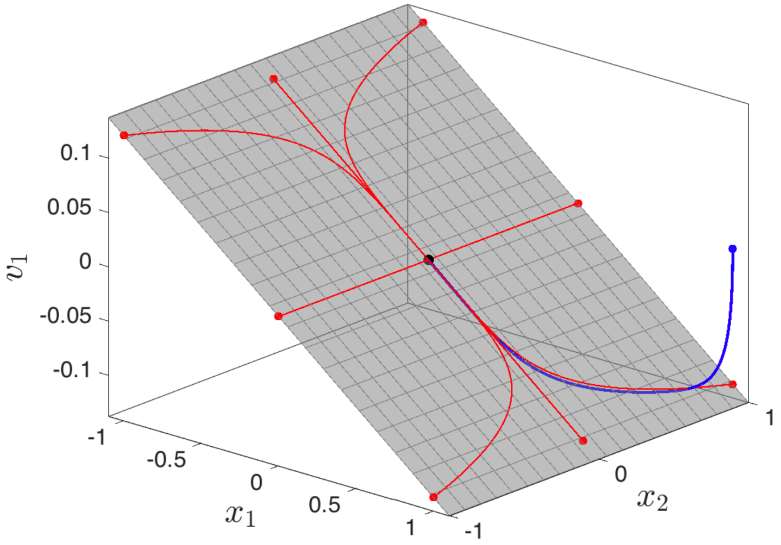
Example 1

$$\epsilon = \frac{1}{8}$$



Choice of Initial Conditions

	Continuous	Discrete
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	$\mathbf{x}(0) = \mathbf{x}_0, \quad \dot{\mathbf{x}}(0) = \mathbf{v}_0$	$\mathbf{x}_0, \quad \mathbf{x}_1$



Skip the convergence towards the slow manifold

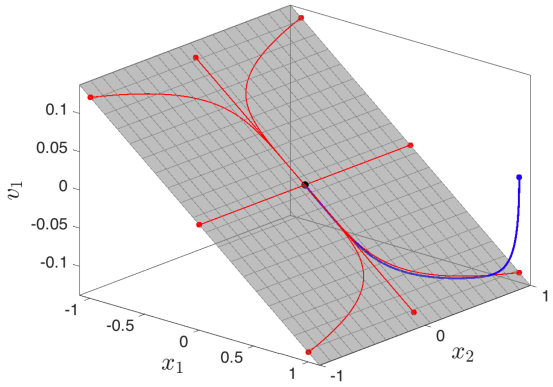
$$\mathbf{v}_0 = \mathbf{v}^{\epsilon,p}(\mathbf{x}_0)$$

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

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

Choice of Initial Conditions

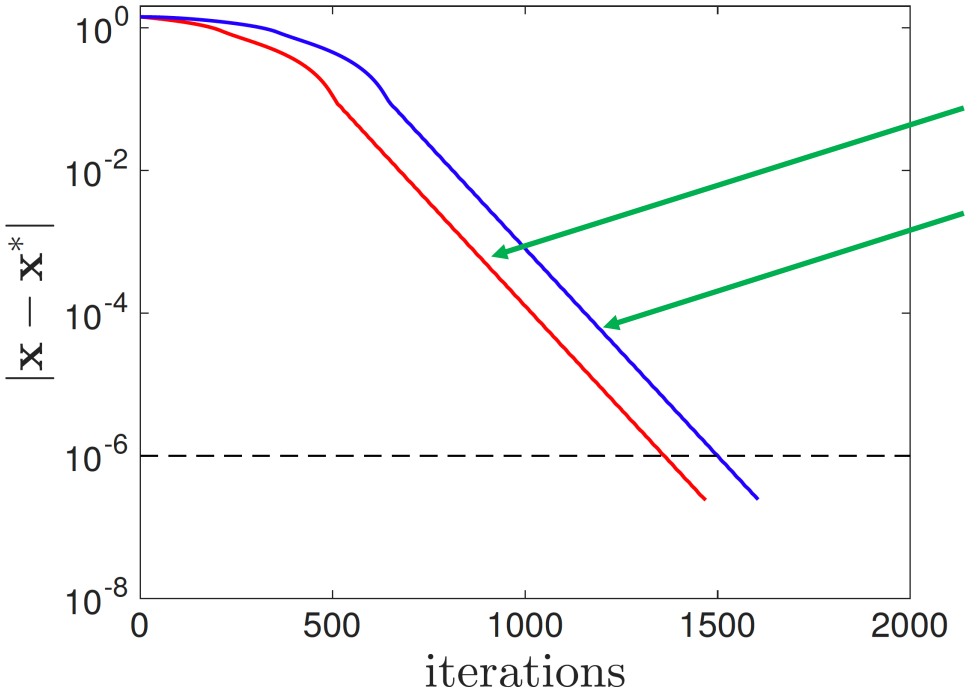
$$\mathcal{M}_{\epsilon,p}: \quad \mathbf{v}^{\epsilon,p}(\mathbf{x}) \triangleq \sum_{k=1}^{p} \epsilon^k \mathbf{g}_k(\mathbf{x})$$



Faster convergence if

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

Same initial position $\mathbf{x}_0$



Slow Manifold for Nesterov

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

$\mathbf{x}(\theta), \mathbf{v}(\theta), t(\theta)$

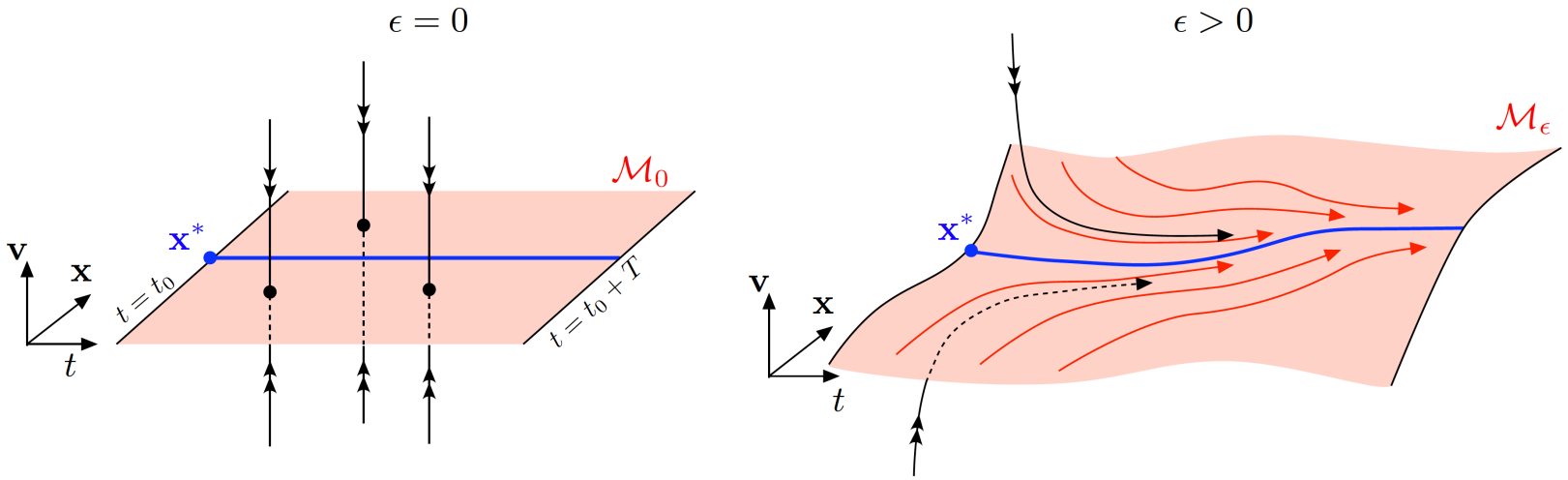
- Non-autonomous system
- Must be viewed in the extended phase space $(\mathbf{x}, \mathbf{v}, t)$

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

Rescaled time

$\theta = \epsilon\tau$

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



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

Slow Manifold for Nesterov

Nestrov'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

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

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).

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

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

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

Slow Manifold for Nesterov

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

First-order approximation:

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

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

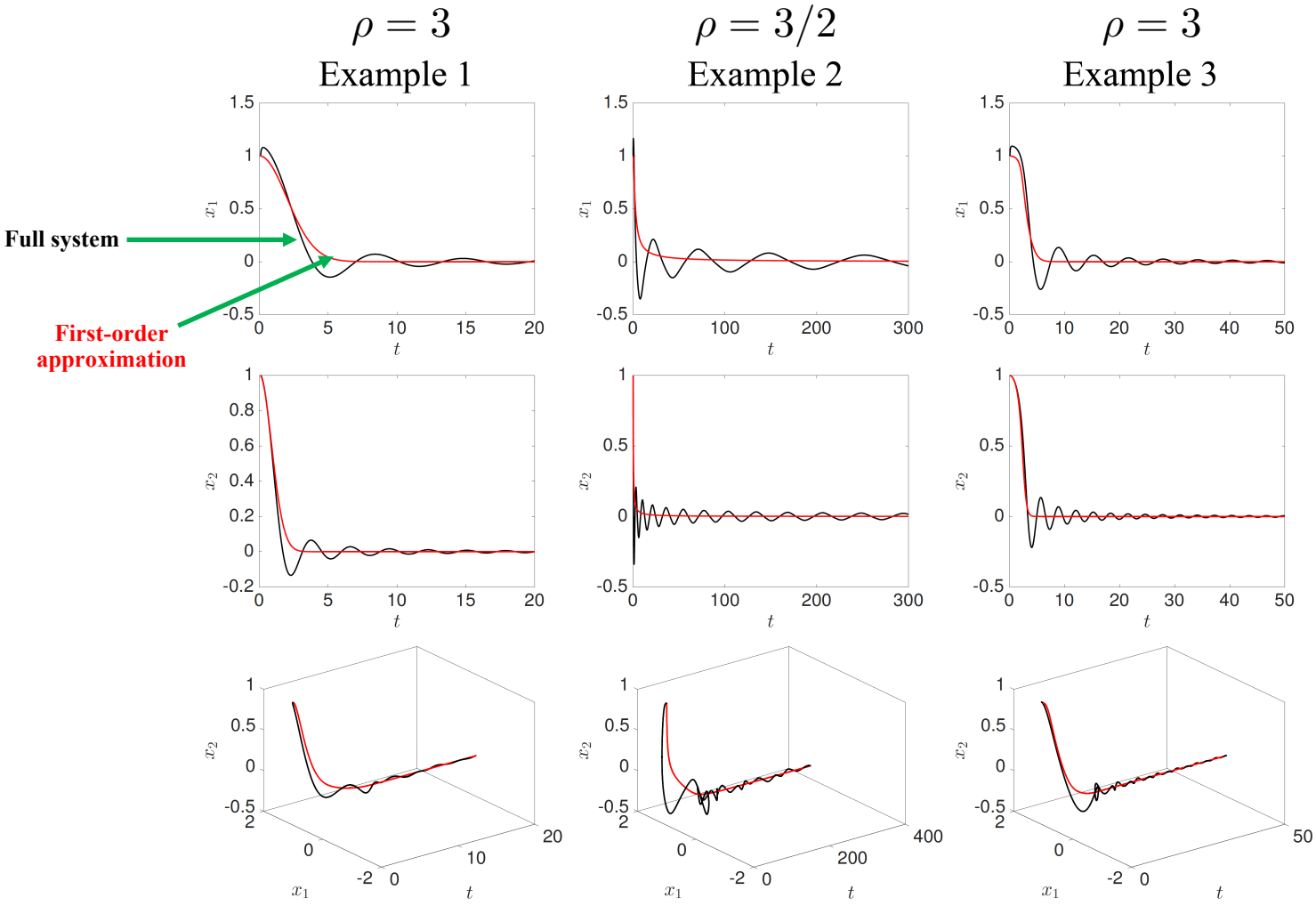
In rescaled time:

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

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

Examples

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



Summary

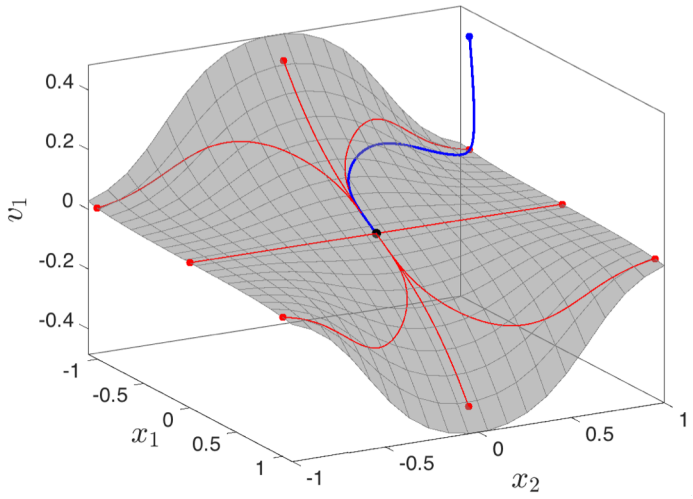
- Accelerated Gradient Descent (AGD) has a multiscale structure

- 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



A Problem with AGD

AGD is slow if the gradient is inexact!

$$\nabla f + \epsilon$$

error/noise

[O. Devolder et al., Math. Program., Ser. A (2014)]

Creates issues for

- Stochastic AGD
- Application in ML

Thank You

[M. Farazmand, Multiscale analysis of accelerated gradient methods, SIAM J. Optimization (2020)]

www.mfarazmand.com
Email: farazmand@ncsu.edu