A Dynamical Systems Perspective on Nesterov Acceleration

Michael Muehlebach and Michael I. Jordan

UC Berkeley

Introduction

- Find $x^* \in \mathbb{R}^n$ such that $f(x^*) \leq f(x)$ for all $x \in \mathbb{R}^n$, where f is smooth and convex.
- Focus on the case where f is strongly convex, i.e. f is convex and satisfies, for any $\bar{x} \in \mathbb{R}^n$,

$$f(x) \geq f(\bar{x}) + \nabla f(\bar{x})(x - \bar{x}) + \frac{L}{2\kappa}|x - \bar{x}|^2, \quad \forall x \in \mathbb{R}^n.$$

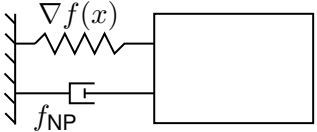
- $L > 0$ is the Lipschitz constant of the gradient.
- $\kappa \geq 1$ is the condition number.

Dynamical Systems Perspective

- Consider the ordinary differential equation (ODE)

$$\ddot{x}(t) + 2d\dot{x}(t) + \frac{1}{L}\nabla f(x(t) + \beta\dot{x}(t)) = 0, \quad \text{with}$$

$$d := \frac{1}{\sqrt{\kappa}+1}, \quad \beta := \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}.$$



- The ODE can be brought to the form

$$\dot{q}(t) = p(t), \quad \dot{p}(t) = -\frac{1}{L}\nabla f(q(t)) + f_{\mathsf{NP}}(q(t), p(t)),$$

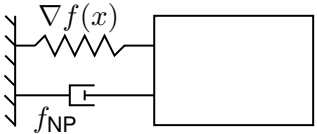
where

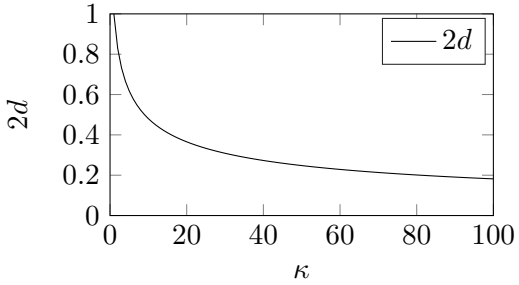
$$H(q,p) := \frac{1}{2}|p|^2 + \frac{1}{L}f(q), \quad f_{\mathsf{NP}}(q,p) := -2dp - \frac{1}{L}(\nabla f(q+\beta p) - \nabla f(q)).$$

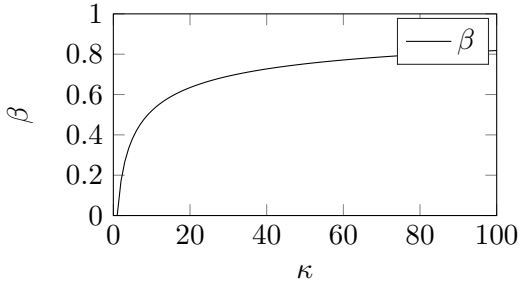
Damping

- The non-potential forces can be rewritten as

$$f_{\text{NP}}(q,p) = -2dp - \frac{1}{L}(\nabla f(q+\beta p) - \nabla f(q))$$
$$= \underbrace{-2dp}_{\substack{\text{isotropic} \\ \text{damping}}} \underbrace{- \frac{1}{L}\int_0^\beta \Delta f(q+\tau p)\mathrm{d}\tau\, p}_{\text{curv. dependent damping}}.$$







Convergence

- Asymptotic stability (through dissipation).
- Convergence rate (upper bound, stated for $p(0) = 0$)

$$f(q(t)) \leq 2(f(q(0)) - f^*) \exp(-1/(2\sqrt{\kappa})t), \quad \forall t \in [0, \infty).$$

- Convergence rate of $\mathcal{O}(1/t^2)$ in the non-strongly convex case.

- Derivation is based on the following Lyapunov-like function (stated for $x^* = f(x^*) = 0$)

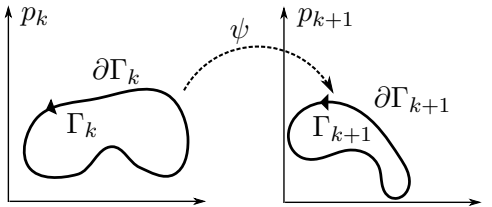
$$V(t) = \frac{1}{2}|aq(t) + p(t)|^2 + \frac{1}{L}f(q(t)).$$

Discretization

- Semi-implicit Euler discretization (with time step $T_s = 1$) leads to the accelerated gradient method

$$q_{k+1} = q_k + T_s p_{k+1}, \quad p_{k+1} = p_k + T_s(-\nabla f(q_k) - f_{\mathsf{NP}}(q_k, p_k)).$$

- What are the properties that are preserved through the discretization?
  - phase-space area contraction rate (contraction for $T_s \in (0, 2)$)
  - time-reversibility (for $T_s \in (0, 1)$)
    $\Rightarrow$ yields a worst-case bound on the convergence rate
  - convergence rate (for $T_s \in (0, 1]$)



Conclusion and Outlook

- We derived a dynamical system model for the accelerated gradient method.
  - The dynamics have an interpretation as mass-spring-damper system.
  - Discretization yields the accelerated gradient method.
  - Certain key properties are preserved through the discretization.
- Is a symplectic discretization the “right” discretization?
  - The behavior for large κ seems particularly important.

- Come to visit me at Poster 205.

The
Branco Weiss
Fellowship
Society in Science