The Physical Systems Behind Optimization Algorithms

Lin F. Yang[1], Raman Arora[2], Vladimir Braverman[2], and Tuo Zhao[3]

[1] Princeton University, [2] Johns Hopkins University, [3] GaTech

The Classic Optimization Problem:

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

Simple first order algorithm:

$$x^k \leftarrow y^{k-1} + \eta\nabla f(y^{k-1}),\ y^k = x^k + \alpha(x^k - x^{k-1})$$

- ❑ Vanilla Gradient Descent (VGD): $\alpha = 0$
- ❑ Nesterov Accelerated Gradient Descent (NAG):
  - ➢ Strongly convex f: $\alpha = \frac{\sqrt{1/\mu\eta}-1}{\sqrt{1/\mu\eta}+1}$
  - ➢ General convex f: $\alpha = \frac{k-1}{k+1}$

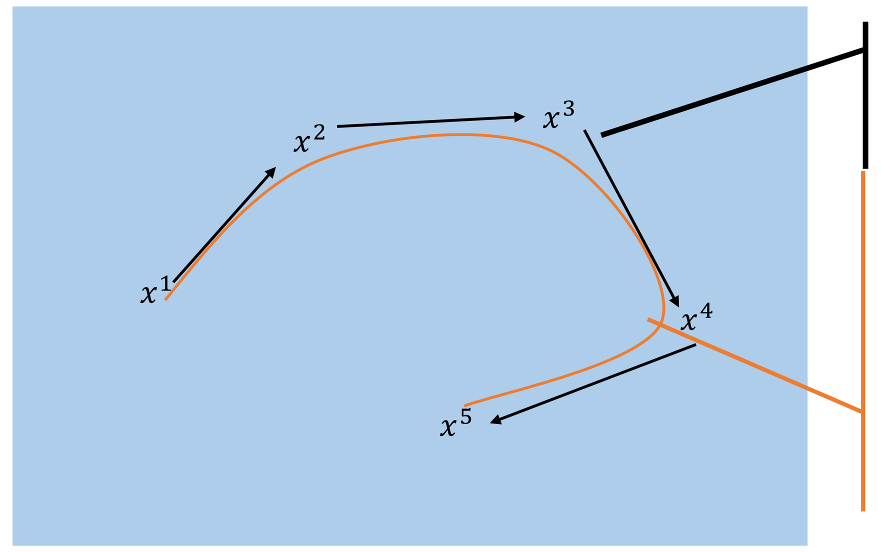
Iteration complexity:

Mthods	Error After k Iters	Assumptions
VGD	$\exp(-k/\kappa)$	Strongly convex
VGD	$1/k$	Convex
NAG	$\exp(-k/\sqrt{\kappa})$	Strongly Convex
NAG	$1/k^2$	Convex

*conditioning number $\kappa = L/\mu$, where μ is the strong convexity constant of f and L is the Lipchitz constant

How to understand these algorithms via an intuitive way?

Our answer: Study their continuous time limit and the physical systems corresponding to these algorithms!



The iterates of the algorithm

The continuous-time limit (step_size → 0) of iterates: a curve described by an ordinary differential equation (ODE)

How to connect discrete iterates to continuous iterates?

- ❑ Set (1) $k(t) := \left\lfloor \frac{t}{\sqrt{\eta}} \right\rfloor$ or (2) $k(t) := \left\lfloor \frac{t}{\eta} \right\rfloor$ and define $X(t) = x^{k(t)}$
- ❑ [1] chooses (1) for AGD and chooses (2) for NAG

Our Solution: a unified framework for time-scaling

Set $k(t) := \left\lfloor \frac{t}{h} \right\rfloor$ for all algorithms. Let the algorithm tell us what is h!

Taylor expansion the algorithm iterates:

$$\left(x^{(k+1)} - x^{(k)}\right) = \dot{X}(t)h + \frac{1}{2}\ddot{X}(t)h^2 + o(h),$$
$$\left(x^{(k)} - x^{(k-1)}\right) = \dot{X}(t)h - \frac{1}{2}\ddot{X}(t)h^2 + o(h),$$
$$\text{and } \eta\nabla f\left[x^{(k)} + \alpha\left(x^{(k)} - x^{(k-1)}\right)\right] = \eta\nabla f(X(t)) + O(\eta h).$$

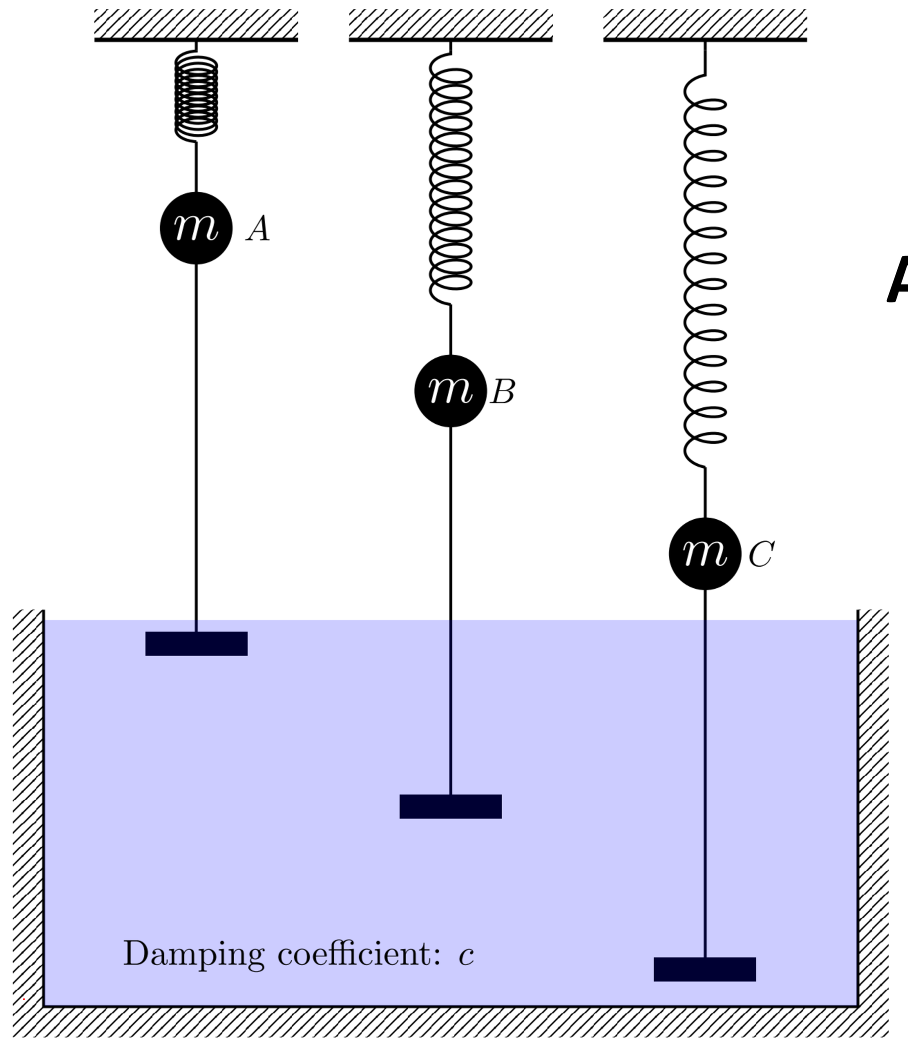
A unified ODE that describe a damped oscillator system:

$$m\ddot{X}(t) + c\dot{X}(t) + \nabla f(X(t)) = 0.$$

$m := \frac{1+\alpha}{2}\frac{h^2}{\eta}$ as *the particle mass,*

$c := \frac{(1-\alpha)h}{\eta}$ as *the damping coefficient,*

$f(x)$ as *the potential field.*



A damped harmonic oscillator system

A unified view of the choice of h:

VGD: massless system

$$\dot{X} + \nabla f(X) = 0$$
$$\Rightarrow m = 0, c = 1$$
$$\Rightarrow h = \Theta(\eta)$$

NAG: massive system

$$m\ddot{X} + c\dot{X} + \nabla f(X) = 0$$
$$\Rightarrow h = \Theta(\sqrt{\eta}),\ 1 - \alpha = \Theta(h)$$

Insights:

- ❑ The energy of the physical system is a Lyapunov function of the ODE
- ❑ Energy decreases fastest when the system is under critical damping

The energy decreasing of the system:

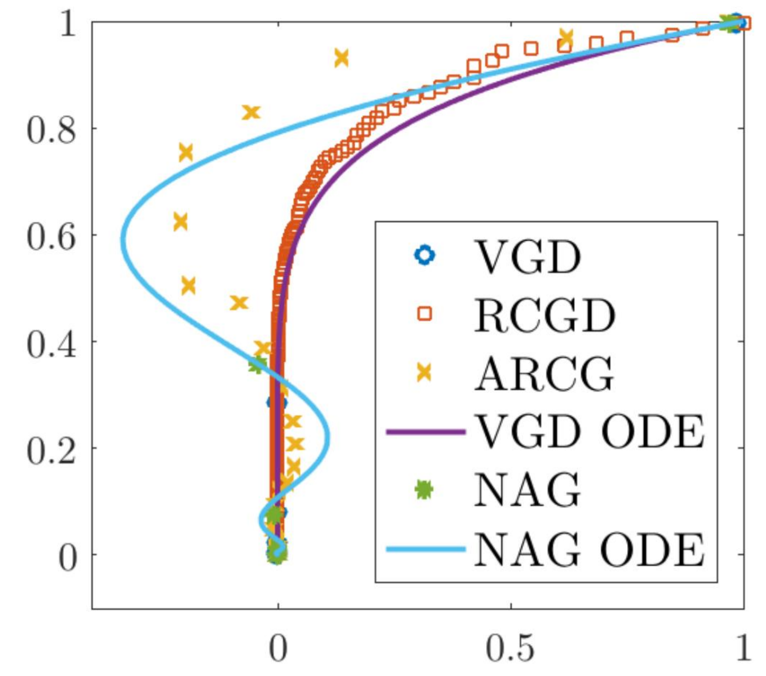
Consider quadratic function $f(x) = K\,||x - x^*||^2/2$

$$\mathcal{E}(t) \propto \exp\left(-\frac{1}{2}\left[\frac{c}{m} - \sqrt{\frac{c^2}{m^2} - \frac{4K}{m}}\right]t\right).$$

Critical damping: $c^2 = 4mK$, which corresponds to NAG

Our framework naturally extends to the PL-condition f:

- ❑ PL-condition: $0 < \frac{f(x)}{\|\nabla f(x)\|^2} \le \frac{1}{2\mu}$
- ❑ $K = L/\mu$



Our framework naturally extends to other optimization methods:

- ❑ Randomized coordinate gradient descent (RCGD)

$$x_j^{(k)} = x_j^{(k-1)} - \eta\nabla_j f(x^{(k-1)}) \quad \text{and} \quad x_{\setminus j}^{(k)} = x_{\setminus j}^{(k-1)}$$

- ❑ Randomized accelerated coordinate gradient descent (ARCG)

$$x_j^{(k)} = y_j^{(k-1)} - \eta\nabla_j f(y^{(k-1)}),\quad x_{\setminus j}^{(k)} = y_{\setminus j}^{(k-1)},\quad \text{and}\quad y^{(k)} = x^{(k)} + \alpha\left(x^{(k)} - x^{(k-1)}\right)$$

[1] / [2] /Ours	VGD	NAG	RCGD	ARCG	Newton
General Convex	--/--/☑	☑/☑/☑	--/--/☑	--/--/☑	--/☑/--
Strongly Convex	--/--/☑	--/--/☑	--/--/☑	--/--/☑	--/--/☑
Proximal Variants	--/--/☑	☑/--/☑	--/--/☑	--/--/☑	--/--/☑
PL Condition	--/--/☑	--/--/☑	--/--/☑	--/--/☑	--/--/--
Physical Systems	--/--/☑	--/--/☑	--/--/☑	--/--/☑	--/--/☑

References:

- [1] Su, W., Boyd, S. and Candes, E (2014); [2] Wibisono, A., Wilson, A. C. and Jordan, M. I. (2016)