535520: Optimization Algorithms

Lecture 5 — Accelerated Gradient Methods

Ping-Chun Hsieh (謝秉均)

October 7, 2024

This Lecture

1. Heavy-Ball Momentum

2. Nesterov's Accelerated Gradient

- Reading Material:
  - W. Su, S. Boyd, and E. Candes, "A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights," NIPS 2014.
- Part of the slides are adapted from Prof. Suvrit Sra's and Prof. Yuxin Chen's lecture slides

Recall: Convergence Rates of GD

	Convergence Rate (under constant step sizes)
Quadratic problem	$\|x_t - x^*\|_2 = \left(\frac{\lambda_1(Q) - \lambda_n(Q)}{\lambda_1(Q) + \lambda_n(Q)}\right)^t \cdot \|x_0 - x^*\|_2$
Strongly convex and L-smooth	$\|x_t - x^*\| \leqslant \left(\frac{\kappa - 1}{\kappa + 1}\right)^t \cdot \|x_0 - x^*\|$
PL condition and L-smooth	$f\left(x_t\right) - f(x^*) \leqslant \left(1 - \frac{\mu}{L}\right)^t \cdot \left(f\left(x_0\right) - f(x^*)\right)$
Convex and L-smooth	$f(x_t) - f(x^*) \leqslant \frac{L}{2t} \cdot \|x_0 - x^*\|^2 = O(\frac{1}{t})$
Non-convex and L-smooth	$\min_{0 \leq k \leq T} \|\nabla f(x_k)\| \leq \sqrt{\frac{2L(f(x_0) - f(x^*))}{T}}, \quad \lim_{t \to \infty} \|\nabla f(x_t)\| = 0$

- Question: Can the above convergence rates be improved?

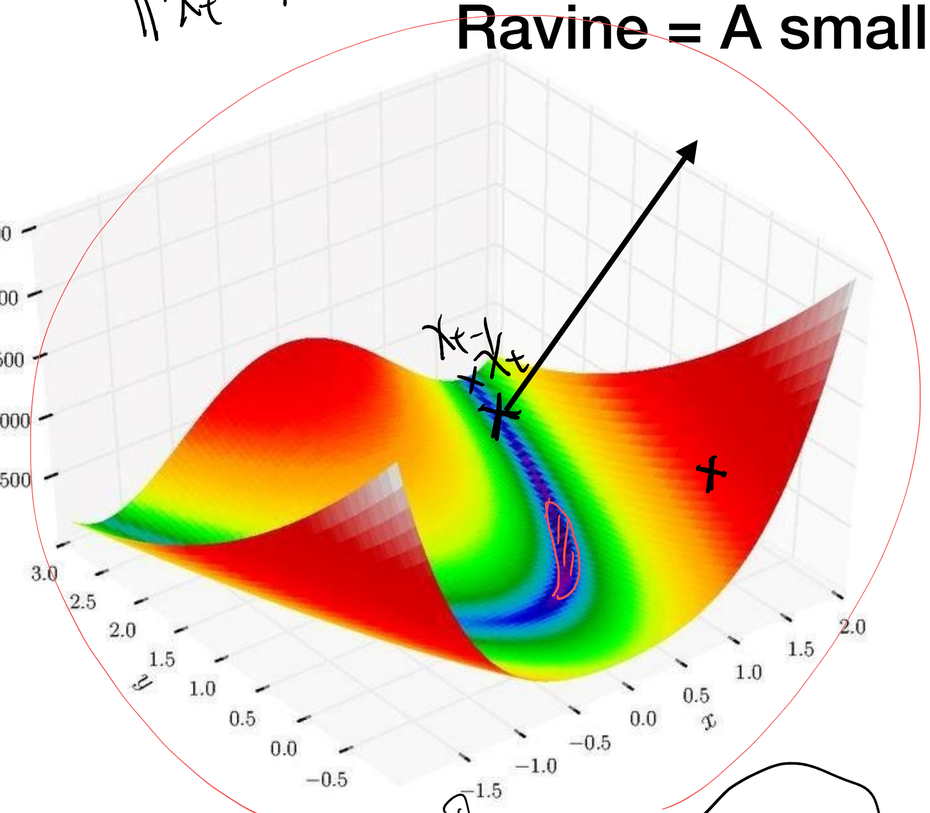
A Historical Account: Ravine Method

"Normalized gradient"

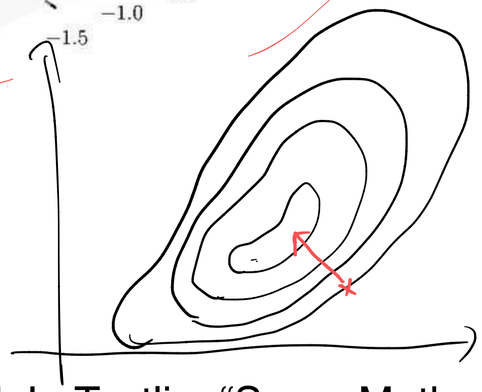
$x_{t+1} = x_t + \eta \cdot \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}$

$\|x_t - x_{t-1}\|$

Ravine = A small narrow steep-sided valley



Contour



Ravine method (Gel'fand & Tsetlin, 1962)

- **Observation**: Descent to the bottom of ravine is simple, but moving along the ravine is hard

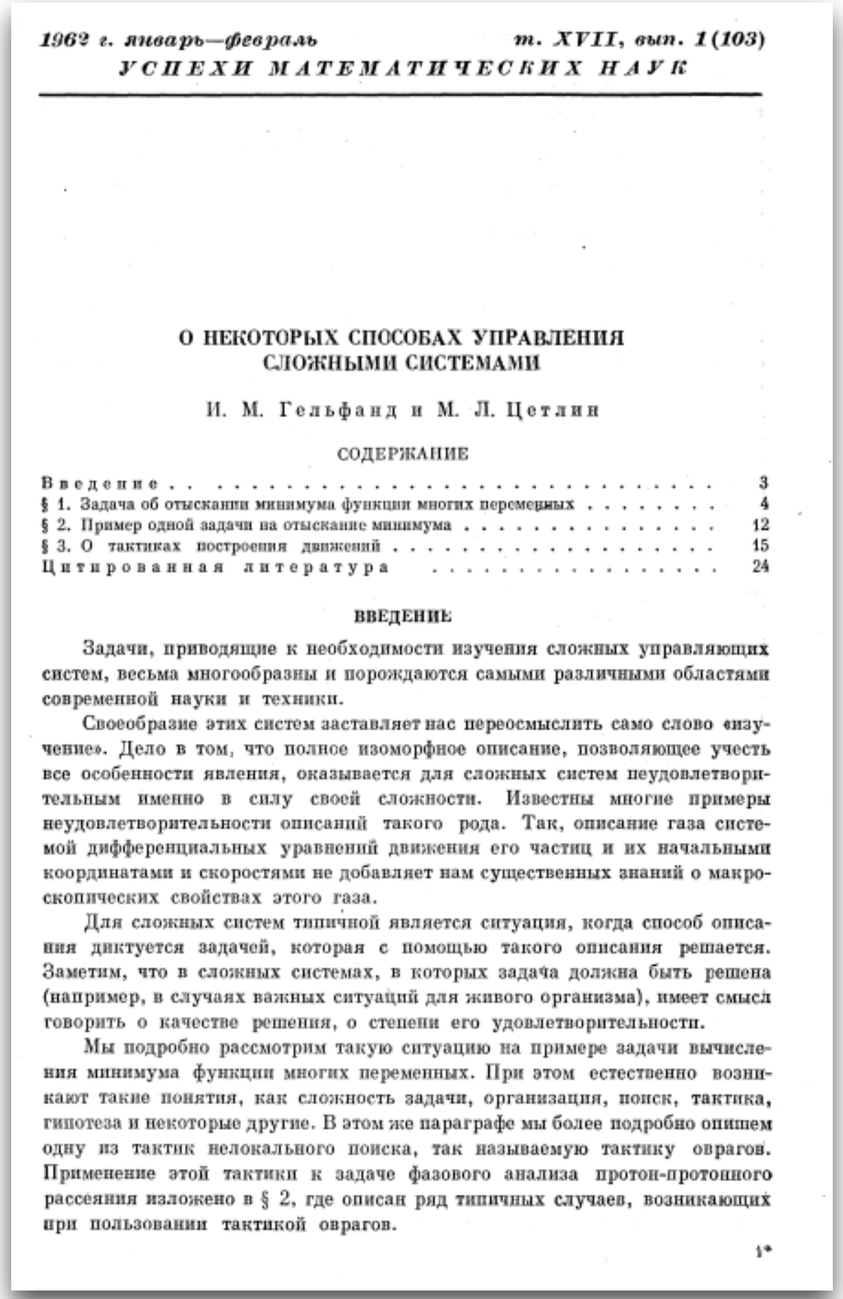
- **Main Idea**: Combine two types of updates:

  (1) Gradient step and (2) "Ravine step"

$$x_{t+1} = y_t - \eta \nabla f(y_t) \qquad y_t = x_t + \theta(x_t - x_{t-1})$$

Ravine step is also known as "momentum"

I. M. Gel'fand and M. L. Tsetlin, "Some Methods of Control for Complex Systems, Russian Mathematical Surveys, 1962



1962 г. январь—февраль т. XVII, вып. 1 (103)

УСПЕХИ МАТЕМАТИЧЕСКИХ НАУК

О НЕКОТОРЫХ СПОСОБАХ УПРАВЛЕНИЯ СЛОЖНЫМИ СИСТЕМАМИ

И. М. Гельфанд и М. Л. Цетлин

СОДЕРЖАНИЕ

Введение 3
§ 1. Задача об отыскании минимума функции многих переменных 4
§ 2. Пример одной задачи на отыскание минимума 12
§ 3. О тактиках построения движений 15
Цитированная литература 24

ВВЕДЕНИЕ

Задачи, приводящие к необходимости изучения сложных управляющих систем, весьма многообразны и порождаются самыми различными областями современной науки и техники.

Своеобразие этих систем заставляет нас переосмыслить само слово «изучение». Дело в том, что полное изоморфное описание, позволяющее учесть все особенности явления, оказывается для сложных систем неудовлетворительным именно в силу своей сложности. Известны многие примеры неудовлетворительности описаний такого рода. Так, описание газа системой дифференциальных уравнений движения его частиц и их начальными координатами и скоростями не добавляет нам существенных знаний о макроскопических свойствах этого газа.

Для сложных систем типичной является ситуация, когда способ описания диктуется задачей, которая с помощью такого описания решается. Заметим, что в сложных системах, в которых задача должна быть решена (например, в случаях важных ситуаций для живого организма), имеет смысл говорить о качестве решения, о степени его удовлетворительности.

Мы подробно рассмотрим такую ситуацию на примере задачи вычисления минимума функции многих переменных. При этом естественно возникают такие понятия, как сложность задачи, организация, поиск, тактика, гипотеза и некоторые другие. В этом же параграфе мы более подробно опишем одну из тактик нелокального поиска, так называемую тактику оврагов. Применение этой тактики к задаче фазового анализа протон-протонного рассеяния изложено в § 2, где описан ряд типичных случаев, возникающих при пользовании тактикой оврагов.

1*

- The original paper in Russian by Gel’fand and Tsetlin (but unfortunately not very well-known)
- Ravine method worked well and sparked numerous heuristics
- Ravine method inspired Polyak’s heavy-ball method, which seems to have inspired Nesterov’s method
- Convergence guarantee of Ravine-type method has remained open for a long time

Heavy-Ball Methods

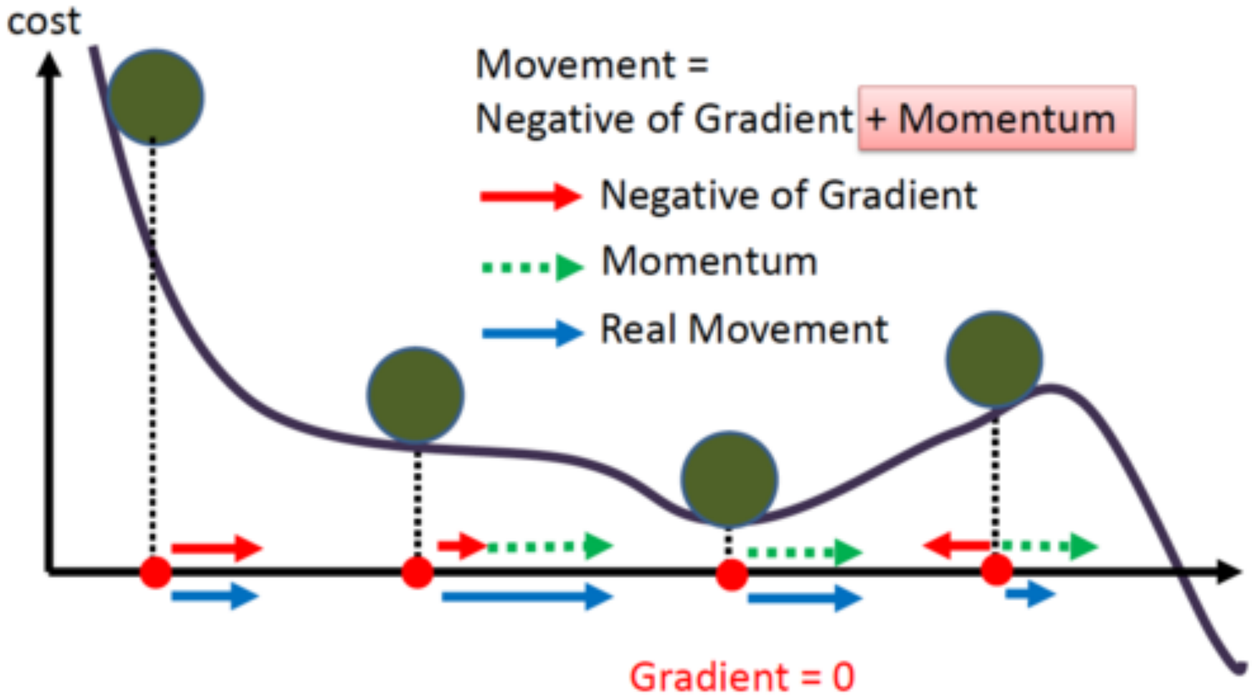
Heavy-Ball (HB) Method

- Consider an unconstrained problem as

$$\text{minimize}_x \quad f(x)$$

- Polyak's Momentum Method:

$$x_{t+1} = x_t - \eta_t \nabla f(x_t) + \underbrace{\theta_t (x_t - x_{t-1})}_{\text{Momentum term}}$$



Intuition: Add inertia to the ball via a momentum term to mitigate zigzagging

B. Polyak, "Some methods of speeding up the convergence of iteration methods," USSR Computational Mathematics and Mathematical Physics, 1964

Boris Polyak

SOME METHODS OF SPEEDING UP THE CONVERGENCE OF ITERATION METHODS*

B. T. POLYAK
(Moscow)

(*Received 26 November 1962*)

For the solution of the functional equation

$$P(x) = 0 \tag{1}$$

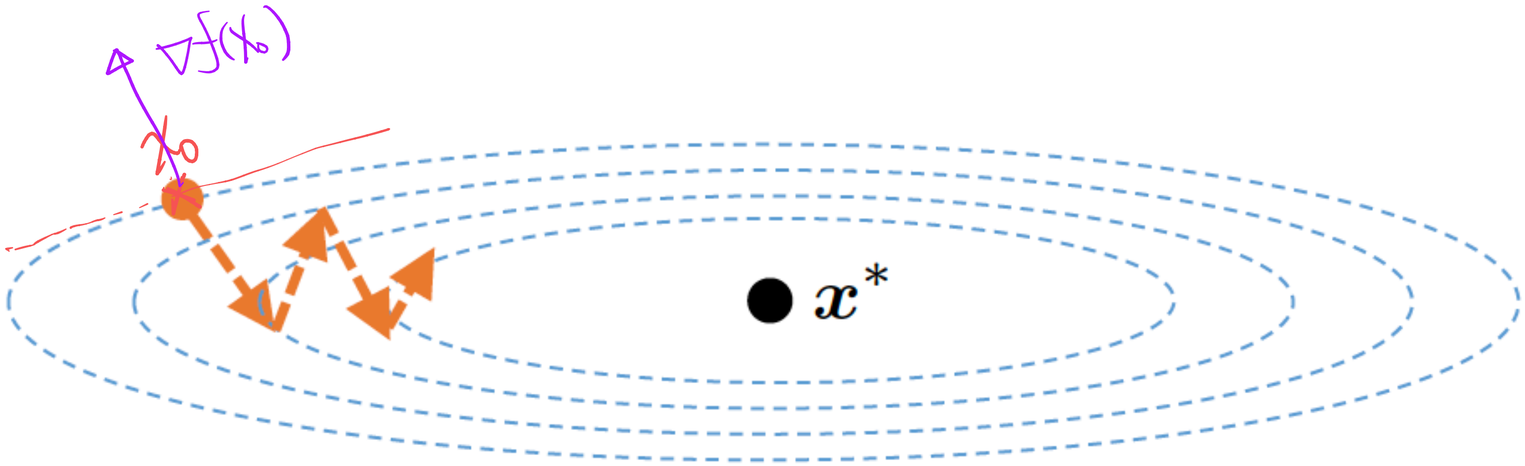
(where P is an operator, usually linear, from B into B, and B is a Banach space) iteration methods are generally used. These consist of the construction of a series $x^0, \ldots, x^n, \ldots$, which converges to the solution (see, for example [1]). Continuous analogues of these methods are also known, in which a trajectory $x(t)$, $0 \leqslant t \leqslant \infty$ is constructed, which satisfies the ordinary differential equation in B and is such that $x(t)$ approaches the solution of (1) as $t \to \infty$ (see [2]). We shall call the method a k-step method if for the construction of each successive iteration x^{n+1} we use k previous iterations $x^n, \ldots, x^{n-k+1}$. The same term will also be used for continuous methods if $x(t)$ satisfies a differential equation of the k-th order or k-th degree. Iteration methods which are more widely used are one-step (e.g. methods of successive approximations). They are generally simple from the calculation point of view but often converge very slowly. This is confirmed both by the evaluation of the speed of convergence and by calculation in practice (for more details see below). Therefore the question of the rate of convergence is most important. Some multistep methods, which we shall consider further, which are only slightly more complicated than the corresponding one-step methods, make it possible to speed up the convergence substantially. Note that all the methods mentioned below are applicable also to the problem of minimizing the differentiable functional $f(x)$ in Hilbert space, so long as this problem reduces to the solution of the equation grad $f(x) = 0$.

---

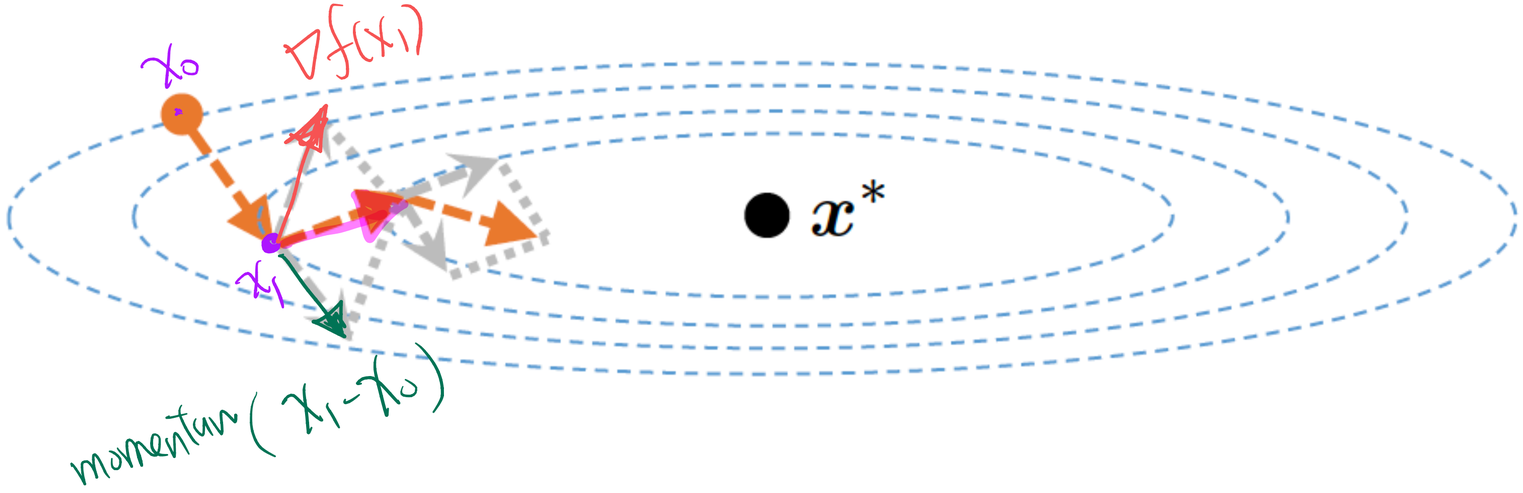
* *Zh. Vych. Mat.*, 4, No. 5, 791-803, 1964.

Visualization of Heavy-Ball Method

Gradient Descent

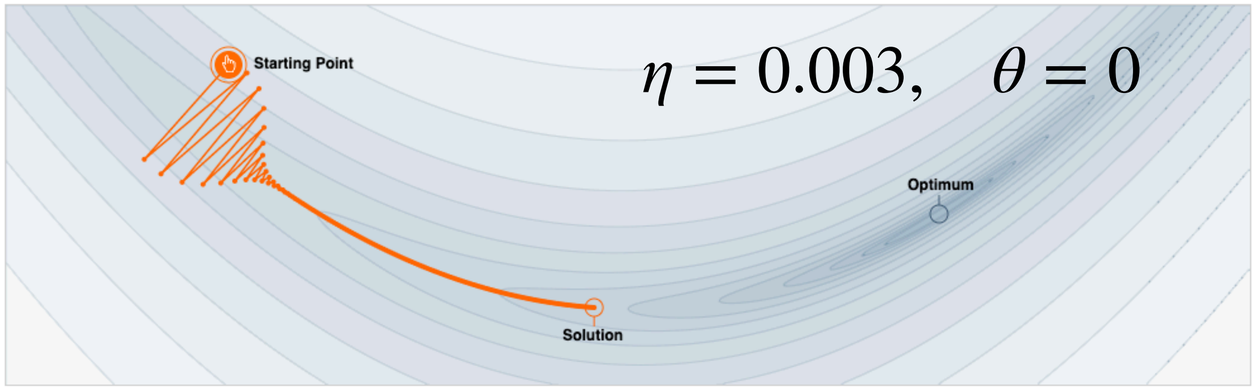


Heavy-Ball Method



(Figure Credit: Yuxin Chen)

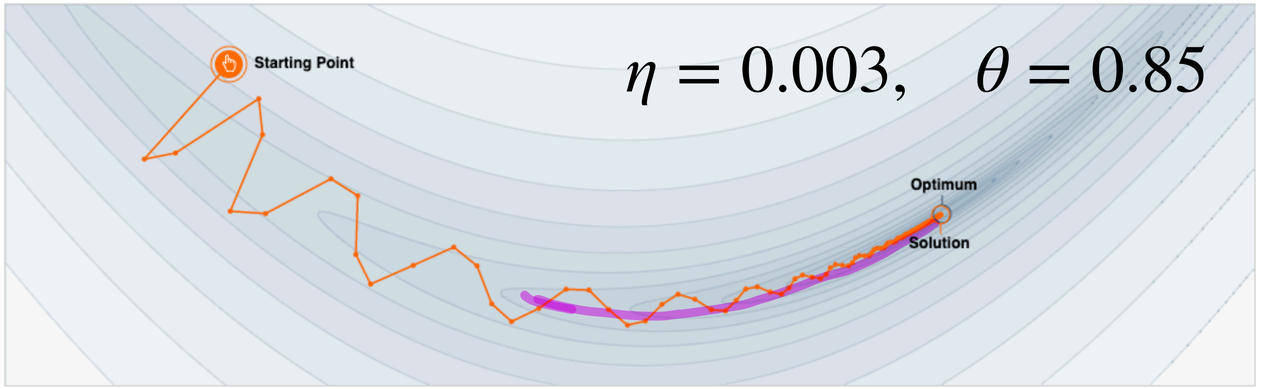
A Nice Interactive Visualization of Heavy-Ball Method



Step-size α = 0.0030 (0, 0.003, 0.006)

Momentum β = 0.0 (0.00, 0.500, 0.990)

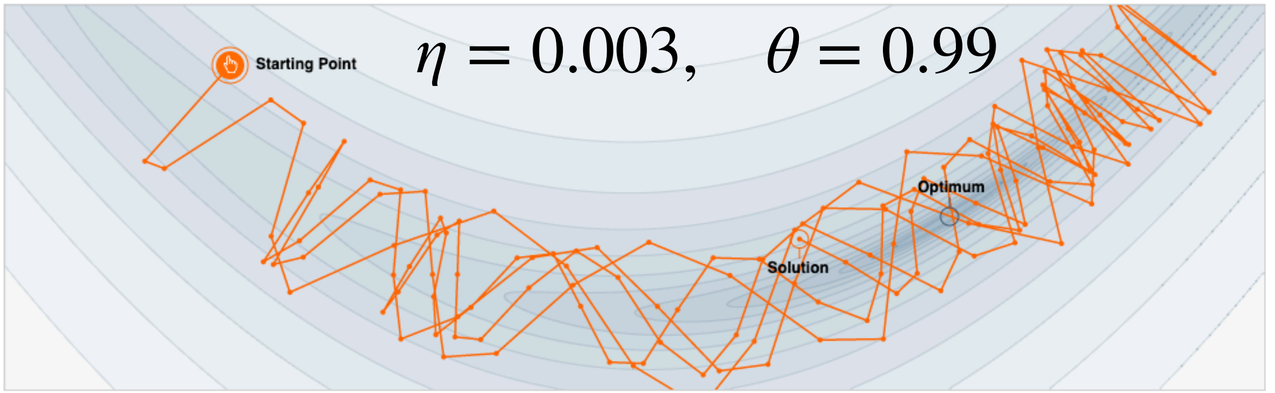
We often think of Momentum as a means of dampening oscillations and speeding up the iterations, leading to faster convergence. But it has other interesting behavior. It allows a larger range of step-sizes to be used, and creates its own oscillations. What is going on?



Step-size α = 0.0030 (0, 0.003, 0.006)

Momentum β = 0.85 (0.00, 0.500, 0.990)

We often think of Momentum as a means of dampening oscillations and speeding up the iterations, leading to faster convergence. But it has other interesting behavior. It allows a larger range of step-sizes to be used, and creates its own oscillations. What is going on?



Step-size α = 0.0030 (0, 0.003, 0.006)

Momentum β = 0.99 (0.00, 0.500, 0.990)

We often think of Momentum as a means of dampening oscillations and speeding up the iterations, leading to faster convergence. But it has other interesting behavior. It allows a larger range of step-sizes to be used, and creates its own oscillations. What is going on?

https://distill.pub/2017/momentum/

Heavy-Ball Method: An Interpretation via Dynamical Systems (1/1)

Let's take a quadratic problem as an example

$$\min_{x\in\mathbb{R}^n} f(x) := \frac{1}{2}(x-x^*)^\top Q(x-x^*),\ Q \text{ is pd} \qquad (\lambda_1(Q) \geq \cdots \geq \lambda_n(Q) > 0)$$

$f(x^*)=0$.

$L=\lambda_1(Q)$

$\mu=\lambda_n(Q)$

Heavy Ball: $x_{t+1} = x_t - \eta_t \nabla f(x_t) + \theta_t(x_t - x_{t-1}) = (1+\theta_t)x_t + (-\theta_t)x_{t-1}$

Let's rewrite the HB update in matrix form:

$$\begin{bmatrix} x_{t+1} \\ x_t \end{bmatrix} = \begin{bmatrix} (1+\theta_t)I & -\theta_t I \\ I & 0 \end{bmatrix} \begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix} - \begin{bmatrix} \eta_t \cdot \nabla f(x_t) \\ 0 \end{bmatrix}$$

$+(-\eta_t \nabla f(x_t))$

block matrix

p.d.

$x^T H x \geq 0$ for all $x \neq 0$

$\Updownarrow$

$H = V^T \Lambda V$

$(x^T V^T) \Lambda (V x) > 0$

Λ

$\begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$

Heavy-Ball Method: An Interpretation via Dynamical Systems (1/2)

$$\begin{bmatrix} x_{t+1} \\ x_t \end{bmatrix} = \begin{bmatrix} (1+\theta_t)I & -\theta_t I \\ I & 0 \end{bmatrix} \begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix} - \begin{bmatrix} \eta_t \cdot \nabla f(x_t) \\ 0 \end{bmatrix}$$

$$\Leftrightarrow \begin{bmatrix} x_{t+1} - x^* \\ x_t - x^* \end{bmatrix} = \begin{bmatrix} (1+\theta_t)I & -\theta_t I \\ I & 0 \end{bmatrix} \begin{bmatrix} x_t - x^* \\ x_{t-1} - x^* \end{bmatrix} - \begin{bmatrix} \eta_t \cdot \nabla f(x_t) \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} (1+\theta_t)I - \eta_t Q & -\theta_t I \\ I & 0 \end{bmatrix} \begin{bmatrix} x_t - x^* \\ x_{t-1} - x^* \end{bmatrix}$$

z_{t+1} H_t z_t

$$z_{t+1} = H_t \cdot z_t$$

$f(x) = \frac{1}{2}(x^* - x)^T Q (x^* - x)$

$\Downarrow$

$\nabla f(x) = Q(x^* - x)$

$\nabla f(x_t) = Q(x^* - x_t)$

This is essentially a linear dynamical system

Question: How to select η_t and θ_t?

Convergence of HB for Quadratic Problems

Theorem (Convergence of HB under strong convexity and smoothness):

Let f be μ-strongly convex and L-smooth quadratic problems. Let $\eta_t = 4/\left(\sqrt{L}+\sqrt{\mu}\right)^2$ and $\theta_t = \left(\max\left\{|1-\sqrt{\eta_t L}|, |1-\sqrt{\eta_t \mu}|\right\}\right)^2$.

Then, we have

$$\left\| \begin{bmatrix} x_{t+1} - x^* \\ x_t - x^* \end{bmatrix} \right\| \leqslant \left(\frac{\sqrt{k}-1}{\sqrt{k}+1} \right)^t \cdot \left\| \begin{bmatrix} x_1 - x^* \\ x_0 - x^* \end{bmatrix} \right\|$$

($\kappa := L/\mu$ is called the condition number)

Comparison: GD and Heavy-Ball Method

- For simple “quadratic” problems: $(\kappa := \lambda_1(Q)/\lambda_n(Q))$

	Convergence Rate
GD	$\|x_t - x^*\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^t \|x_0 - x^*\|$
Heavy Ball	$\|x_t - x^*\| \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^t \|x_0 - x^*\|$

This is a pretty good improvement over GD!

Definition: Iteration Complexity

Definition: Given any $\epsilon > 0$, the **iteration complexity** of an algorithm is defined as the number of iterations needed to reach an ϵ-optimal solution.

Let's quickly verify the following:

	Convergence Rate	Iteration Complexity
GD	$\|x_t - x^*\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^t \|x_0 - x^*\|$	$\kappa \log \frac{1}{\epsilon}$
Heavy Ball	$\|x_t - x^*\| \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^t \|x_0 - x^*\|$	$\sqrt{\kappa} \log \frac{1}{\epsilon}$

Proof: Convergence of HB for Quadratic Problems (1/3)

Recall:
$$\begin{bmatrix} x_{t+1} - x^* \\ x_t - x^* \end{bmatrix} = \underbrace{\begin{bmatrix} (1+\theta_t)I - \eta_t Q & -\theta_t I \\ I & 0 \end{bmatrix}}_{\boldsymbol{H}_t} \begin{bmatrix} x_t - x^* \\ x_{t-1} - x^* \end{bmatrix}$$

Eigenvalue decomposition

$Q = V^T \Lambda V$

orthonormal basis

$\Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$

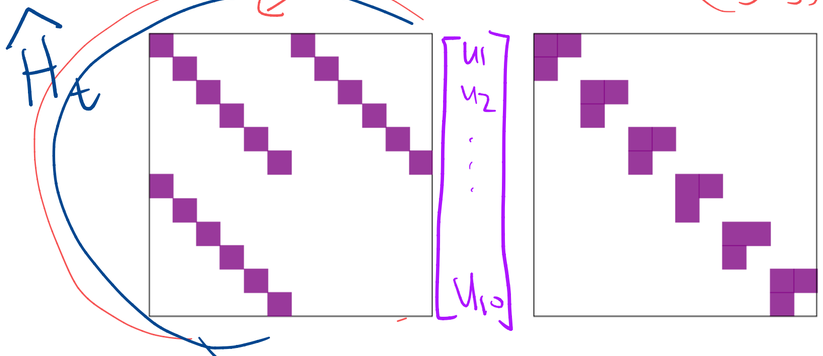
Step 1: Find the eigenvalues of $\boldsymbol{H}_t$

(Let $\lambda_i(A)$ denote the i-th eigenvalue of a matrix A)

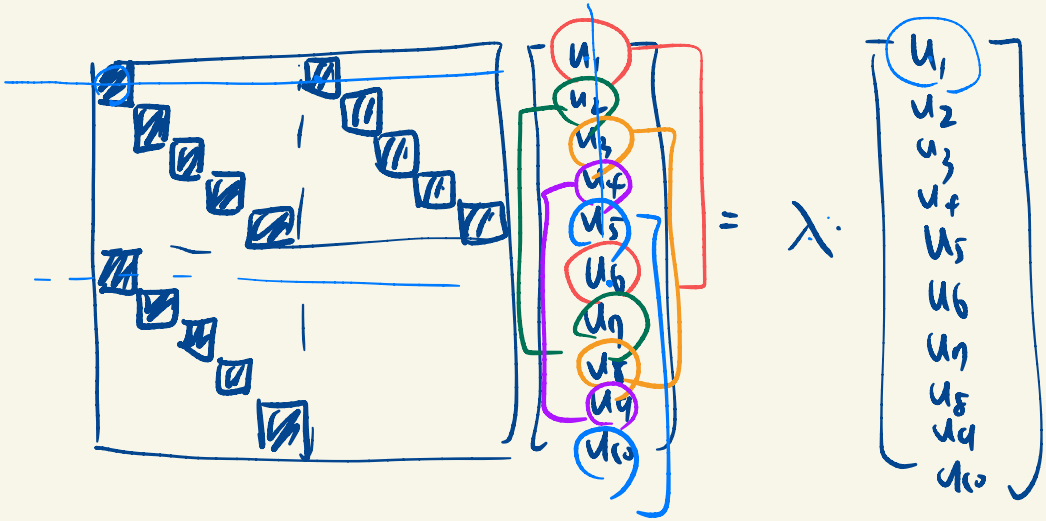
Define the spectral radius $\rho(\boldsymbol{H}_t) = \max\{\,|\lambda_1(\boldsymbol{H}_t)|, \cdots, |\lambda_n(\boldsymbol{H}_t)|\,\}$

$$\rho(\boldsymbol{H}_t) = \rho\left(\begin{bmatrix} (1+\theta_t)\boldsymbol{I} - \eta_t \boldsymbol{\Lambda} & -\theta_t \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix}\right) \le \max_{1 \le i \le n} \rho\left(\begin{bmatrix} 1 + \theta_t - \eta_t \lambda_i & -\theta_t \\ 1 & 0 \end{bmatrix}\right)$$

5×5 5×5 5×5 5×5



Find eigenvalues of $\hat{H}_t$: $\hat{H}_t \cdot u = \lambda \cdot u$


u_1
u_2
u_3
u_4
u_5
u_6
u_7
u_8
u_9
u_{10}
=
$\lambda \cdot$
u_1
u_2
u_3
u_4
u_5
u_6
u_7
u_8
u_9
u_{10}

Proof: Convergence of HB for Quadratic Problems (2/3)

Step 2: Let's show that $\max_{1\le i\le n} \rho\left(\begin{bmatrix} 1+\theta_t-\eta_t\lambda_i & -\theta_t \\ 1 & 0 \end{bmatrix}\right) \le \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$

$=: Z_i$

The eigenvalues of Z_i are the roots of $z^2-(1+\theta_t-\eta_t\lambda_i)z+\theta_t=0$

Notably, if $\left(1+\theta_t-\eta_t\lambda_i\right)^2 \le 4\theta_t$, then the two roots have the same magnitude $\sqrt{\theta_t}$ (either duplicate real roots or two imaginary roots)

Moreover, $\left(1+\theta_t-\eta_t\lambda_i\right)^2 \le 4\theta_t$ would hold for $\theta_t \in \left[\left(1-\sqrt{\eta_t\lambda_i}\right)^2, \left(1+\sqrt{\eta_t\lambda_i}\right)^2\right]$

Proof: Convergence of HB for Quadratic Problems (3/3)

Step 3: By Step 2, we can simply choose $\theta_t = \max\left\{\left(1-\sqrt{\eta_t L}\right)^2, \left(1-\sqrt{\eta_t \mu}\right)^2\right\}$

Step 4: Finally, by setting $\eta_t = 4/\left(\sqrt{L}+\sqrt{\mu}\right)^2$, we have

$$\theta_t = \max\left\{\left(1-\frac{2\sqrt{L}}{\sqrt{L}+\sqrt{\mu}}\right)^2, \left(1-\frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2\right\} = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$$

This implies that $\rho(\boldsymbol{H}_t) = \sqrt{\theta_t} \leq \dfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$

Nesterov’s Accelerated Gradient Methods (NAG)

How About General Convex Cases?

- For smooth quadratic problems, Heavy Ball improves the iteration complexity of GD from $O\left(\kappa \log(\frac{1}{\epsilon})\right)$ to $O\left(\sqrt{\kappa} \log(\frac{1}{\epsilon})\right)$

- A Natural Question: Can we achieve similar improvement for the *general convex problems*?

Recall: Convergence of GD for Convex and Smooth Problems

- Theorem (GD for General Convex Problems): Suppose f is convex and L-smooth.

  If we choose the step size $\eta_t \equiv \eta = \frac{1}{L}$, then GD achieves

$$f(x_t) - f(x^*) \leq \frac{2L\|x_0 - x^*\|^2}{t}$$

- In other words, the iteration complexity of GD is $O\left(\frac{1}{\epsilon}\right)$.

Recall: Convergence of GD for Convex and Smooth Problems

- Theorem (GD for General Convex Problems): Suppose f is convex and L-smooth.

  If we choose the step size $\eta_t \equiv \eta = \frac{1}{L}$, then GD achieves

$$f(x_t) - f(x^*) \leq \frac{2L\|x_0 - x^*\|^2}{t}$$

- In other words, the iteration complexity of GD is $O\left(\frac{1}{\epsilon}\right)$.

Nesterov's Method vs Heavy-Ball Method

GD ↓ "descent"

Nesterov's method

$$x_{t+1} = y_t - \eta_t \nabla f(y_t)$$

$$y_t = x_t + \frac{t-1}{t+2}(x_t - x_{t-1})$$

Extrapolation term

Heavy Ball's method

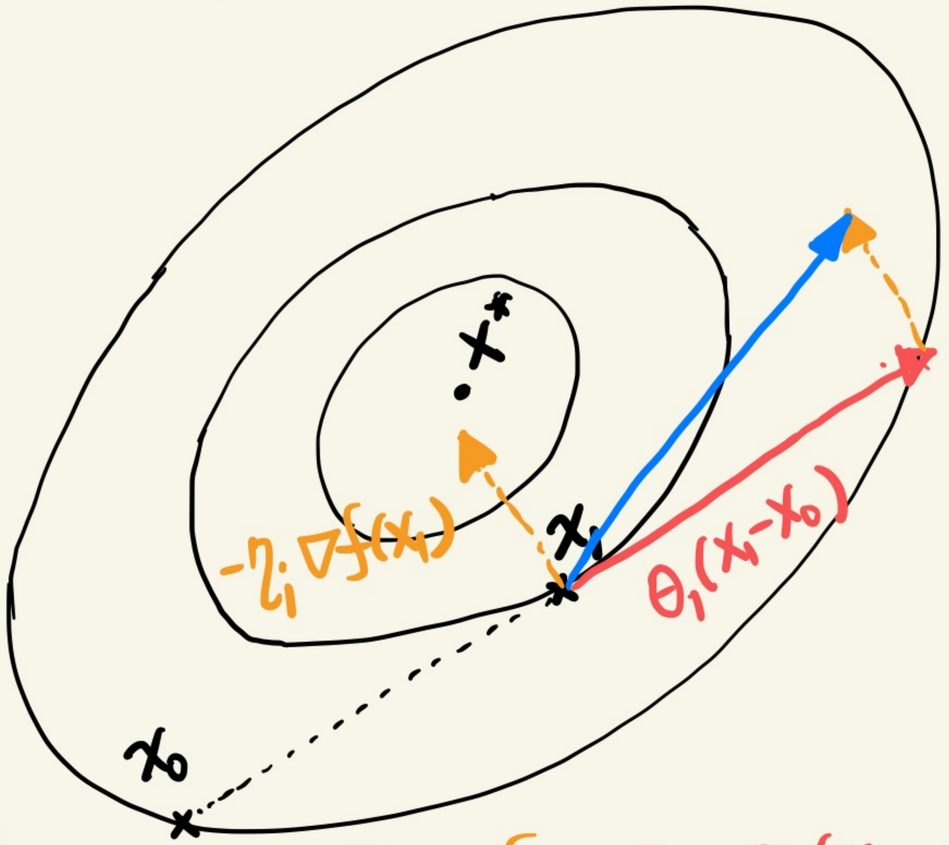
$$x_{t+1} = x_t - \eta_t \nabla f(x_t) + \theta_t(x_t - x_{t-1})$$

Momentum term

- Nesterov's method is essentially the same as the Ravine method!
- Nesterov's method is one of the most "mysterious" findings in optimization
  1. Nesterov's method alternatives between "gradient updates" and "extrapolation"
  2. Computation cost: Almost the same as GD
  3. Does not ensure "descent" in each iteration

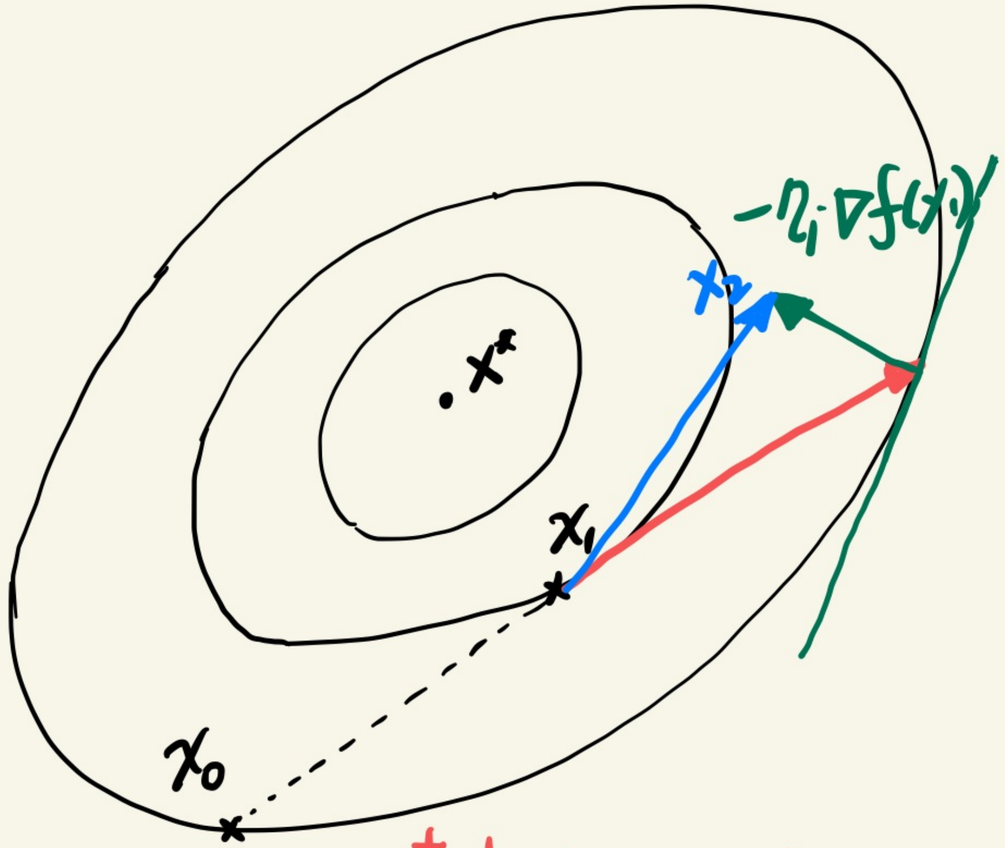
Visualization: Polyak's Momentum vs Nesterov's Method

Polyak's Momentum



$$x_{t+1} = x_t - \eta_t \nabla f(x_t) + \theta_t (x_t - x_{t-1})$$

Nesterov's Method



$$x_{t+1} = \underbrace{x_t + \frac{t-1}{t+2}(x_t - x_{t-1})}_{=y_t} - \eta_t \nabla f(y_t)$$

A Somewhat Surprising Result: Convergence of Nesterov's Method

- Theorem (Nesterov's Method for General Convex Problems): Suppose f is convex and L-smooth. If we choose the step size $\eta_t \equiv \eta = \frac{1}{L}$, then we have

$$f(x_t) - f(x^*) \leq \frac{2L\|x_0 - x^*\|^2}{(t+1)^2}$$

- Proof: HW1 Problem
- Question: What's the iteration complexity?

Y. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," Soviet Mathematics Doklady, 1983

Comparison: GD and Nesterov's Method

- For smooth and convex problems, the iteration complexity is:

	Convergence Rate	Iteration Complexity
GD	$f(x_t) - f(x^*) \leq \frac{2L\|x_0 - x^*\|^2}{t^1}$	$O(\frac{1}{\epsilon})$
Nesterov's	$f(x_t) - f(x^*) \leq \frac{2L\|x_0 - x^*\|^2}{(t+1)^2}$	$O(\frac{1}{\sqrt{\epsilon}})$

This is a remarkable improvement over GD!

Yurii Nesterov
(Professor @ University of Louvain)

Докл. Акад. Наук СССР
Том 269 (1983), № 3

Soviet Math. Dokl.
Vol. 27 (1983), No. 2

A METHOD OF SOLVING A CONVEX PROGRAMMING PROBLEM WITH CONVERGENCE RATE $O(1/k^2)$

UDC 51

YU. E. NESTEROV

1. In this note we propose a method of solving a convex programming problem in a Hilbert space E. Unlike the majority of convex programming methods proposed earlier, this method constructs a minimizing sequence of points $\{x_k\}_0^\infty$ that is not relaxational. This property allows us to reduce the amount of computation at each step to a minimum. At the same time, it is possible to obtain an estimate of convergence rate that cannot be improved for the class of problems under consideration (see [1]).

2. Consider first the problem of unconstrained minimization of a convex function $f(x)$. We will assume that $f(x)$ belongs to the class $C^{1,1}(E)$, i.e. that there exists a constant $L > 0$ such that for all $x, y \in E$

$$\|f'(x) - f'(y)\| \leqslant L\|x - y\|. \tag{1}$$

From (1) it follows that for all $x, y \in E$

$$f(y) \leqslant f(x) + \langle f'(x), y - x\rangle + 0.5L\|y - x\|^2. \tag{2}$$

To solve the problem $\min\{f(x) \mid x \in E\}$ with a nonempty set X^* of minima we propose the following method.

0) Select a point $y_0 \in E$. Put

$$k = 0, \quad a_0 = 1, \quad x_{-1} = y_0, \quad \alpha_{-1} = \|y_0 - z\|/\|f'(y_0) - f'(z)\|, \tag{3}$$

where z is an arbitrary point in E, $z \neq y_0$ and $f'(z) \neq f'(y_0)$.

1) *kth iteration*. a) Calculate the smallest index $i \geqslant 0$ for which

$$f(y_k) - f(y_k - 2^{-i}\alpha_{k-1}f'(y_k)) \geqslant 2^{-i-1}\alpha_{k-1}\|f'(y_k)\|^2. \tag{4}$$

b) Put

$$\begin{aligned} &\alpha_k = 2^{-i}\alpha_{k-1}, \quad x_k = y_k - \alpha_k f'(y_k), \\ &a_{k+1} = \left(1 + \sqrt{4a_k^2 + 1}\right)/2, \\ &y_{k+1} = x_k + (a_k - 1)(x_k - x_{k-1})/a_{k+1}. \end{aligned} \tag{5}$$

The way in which the one-dimensional search (4) is halted is similar to that proposed in [2]. The difference is only that in (4) the subdivision in the kth iteration is done starting with α_{k-1} (and not with 1 as in [2]). In view of this (see the proof of Theorem 1), when the sequence $\{x_k\}_0^\infty$ is constructed by method (3)–(5), no more than $O(\log_2 L)$ such subdivisions will be made. The recalculation of the points y_k in (5) is done using a "ravine" step.

Why is Nesterov's Method "Mysterious"?

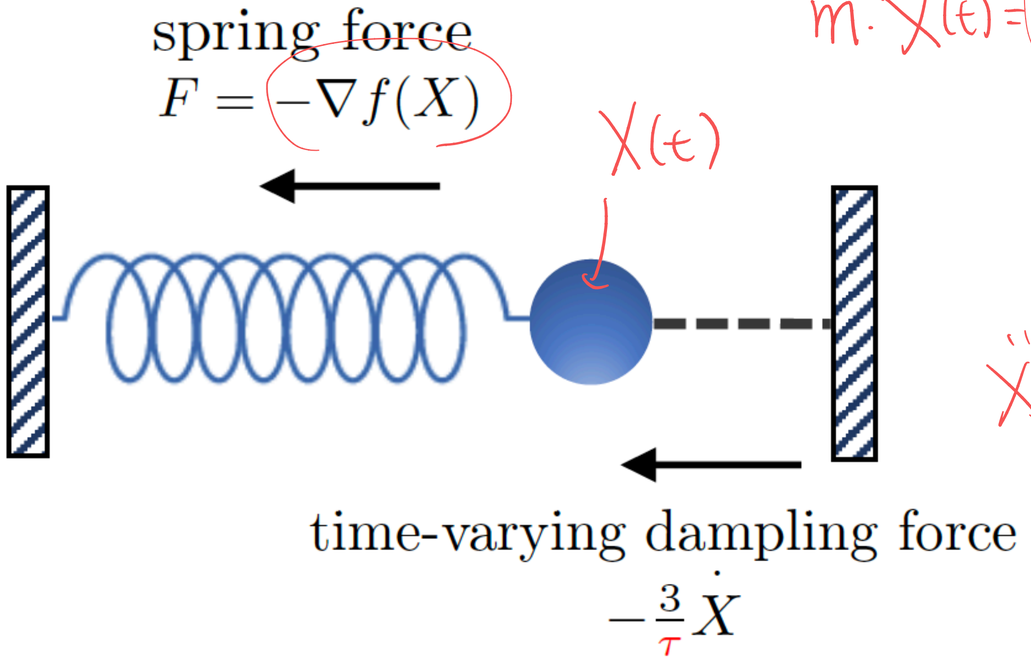
$$x_{t+1} = y_t - \eta_t \nabla f(y_t)$$

$$y_t = x_t + \underbrace{\frac{t-1}{t+2} \cdot (x_t - x_{t-1})}_{\text{Extrapolation term}}$$

$1 - \frac{3}{t+2}$

- The Nesterov's extrapolation coefficient is especially mysterious! (No clear intuition)
- Observation: For large t, we have $\frac{t-1}{t+2} \approx 1 - \frac{3}{t}$, 4, 2
- In 2014, Su, Boyd, and Candes offered a very interesting interpretation via dynamical systems

Interpreting Nesterov's Method via Differential Equations



$$\ddot{X}(\tau) + \frac{3}{\tau}\dot{X}(\tau) + \nabla f(X(\tau)) = 0, \quad \tau > 0$$

(With initial conditions $X(0) = x_0, \dot{X}(0) = 0$)

- A Profound Insight: Nesterov's method at its continuous limits ($\eta \to 0$) can be characterized by the above 2nd-order ordinary differential equation (ODE)

W. Su, S. Boyd, and E. Candes, "A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights", NIPS 2014.

ODE Perspective of NAG

A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights

Weijie Su SUW@WHARTON.UPENN.EDU
Department of Statistics
University of Pennsylvania
Philadelphia, PA 19104, USA

Stephen Boyd BOYD@STANFORD.EDU
Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA

Emmanuel J. Candès CANDES@STANFORD.EDU
Departments of Statistics and Mathematics
Stanford University
Stanford, CA 94305, USA

Editor: Yoram Singer

Abstract

We derive a second-order ordinary differential equation (ODE) which is the limit of Nesterov's accelerated gradient method. This ODE exhibits approximate equivalence to Nesterov's scheme and thus can serve as a tool for analysis. We show that the continuous time ODE allows for a better understanding of Nesterov's scheme. As a byproduct, we obtain a family of schemes with similar convergence rates. The ODE interpretation also suggests restarting Nesterov's scheme leading to an algorithm, which can be rigorously proven to converge at a linear rate whenever the objective is strongly convex.

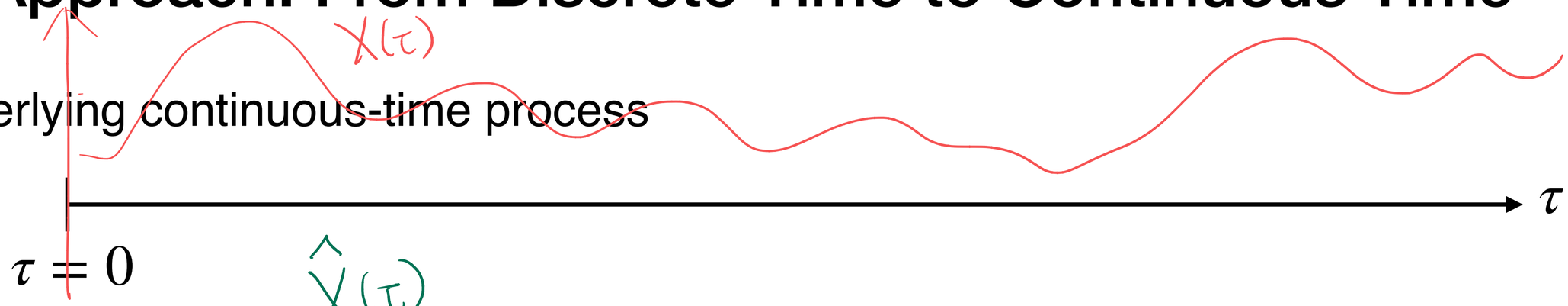
Keywords: Nesterov's accelerated scheme, convex optimization, first-order methods, differential equation, restarting

Stephen Boyd
@ Stanford

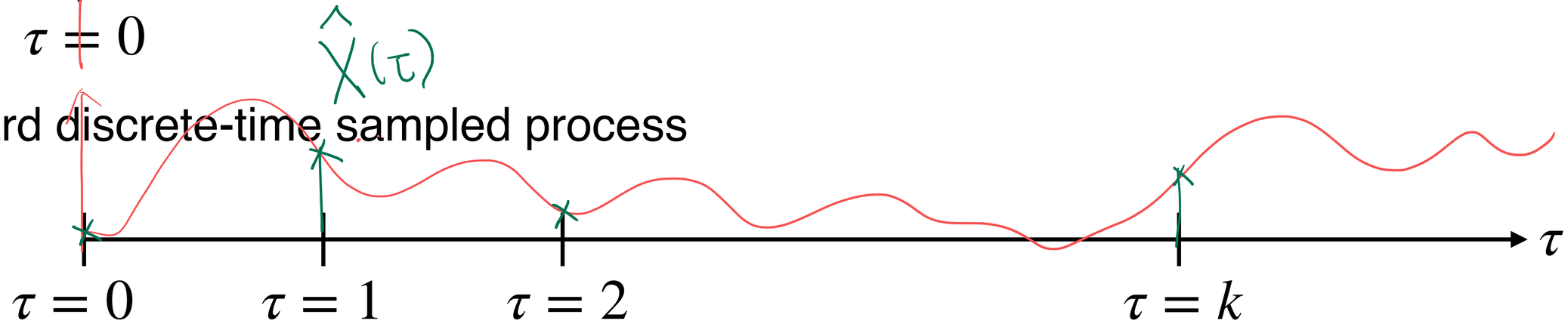
Emmanuel Candes
@ Stanford

ODE Approach: From Discrete Time to Continuous Time

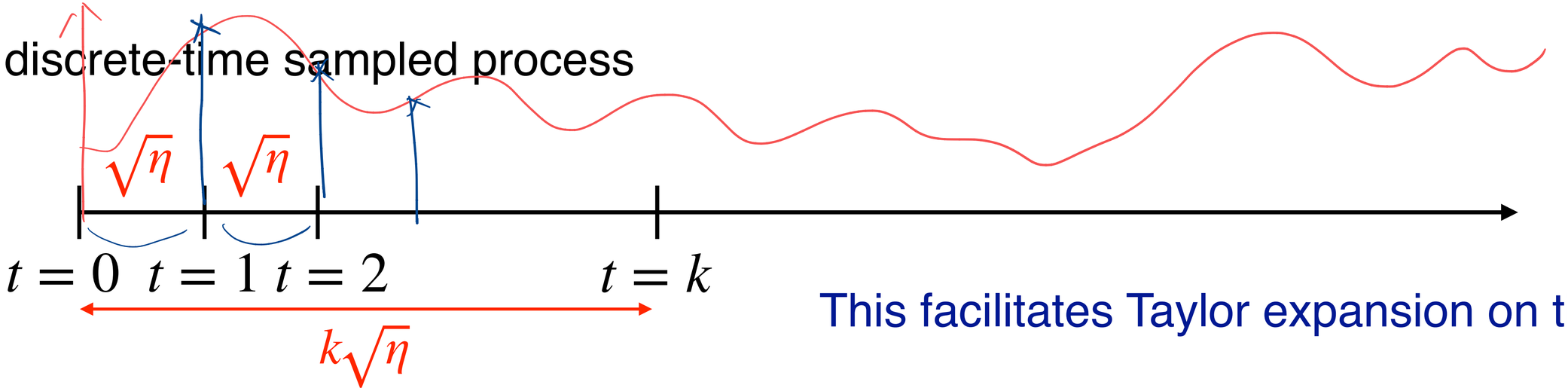
An underlying continuous-time process



Standard discrete-time sampled process



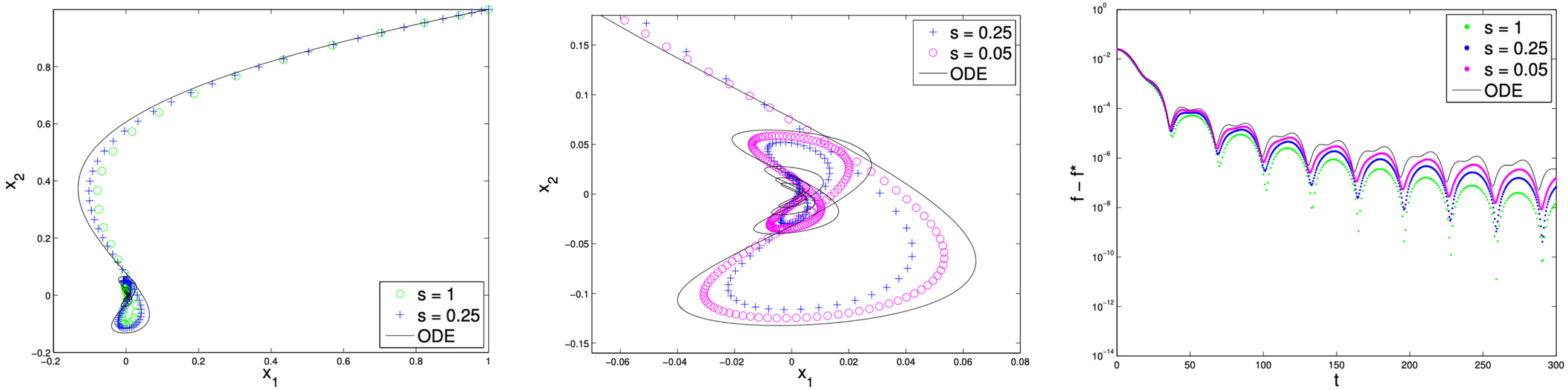
Scaled discrete-time sampled process



This facilitates Taylor expansion on time!

A Motivating Example

Continuous-time and discrete-time systems exhibit similar trajectories

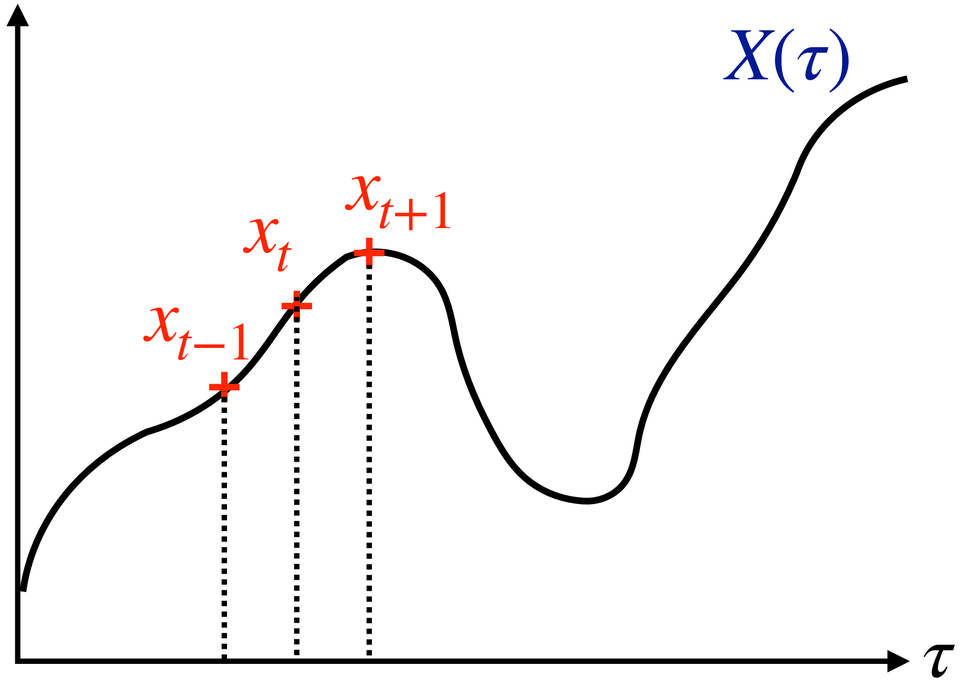


(a) Trajectories. (b) Zoomed trajectories. (c) Errors $f - f^\star$.

Figure 1: Minimizing $f = 2 \times 10^{-2} x_1^2 + 5 \times 10^{-3} x_2^2$, starting from $x_0 = (1,\ 1)$. The black and solid curves correspond to the solution to the ODE. In (c), for the x-axis we use the identification between time and iterations, $t = k\sqrt{s}$.

W. Su, S. Boyd, and E. Candes, "A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights," NIPS 2014.

Deriving the Equivalent ODE for NAG



By Taylor expansion w.r.t. time:

$$x_{t+1} = x_t + \sqrt{\eta} \cdot \dot{X}(t) + \frac{1}{2}\sqrt{\eta} \cdot \ddot{X}(t) \cdot \sqrt{\eta} + o(\sqrt{\eta})$$

$$x_t = x_{t-1} + \sqrt{\eta} \cdot \dot{X}(t) - \frac{1}{2}\sqrt{\eta} \cdot \ddot{X}(t) \cdot \sqrt{\eta} + o(\sqrt{\eta})$$

$$X_{t-1} = X_t + (-\sqrt{\eta}) \cdot \dot{X}(t) + \frac{1}{2}\sqrt{\eta} \cdot \ddot{X}(t) \cdot \sqrt{\eta} + o(\sqrt{\eta})$$

Derivation of ODE for Nesterov's Method

Nesterov's update

$$\begin{cases} x_{t+1} = y_t - \eta \cdot \nabla f(y_t) \\ y_t = x_t + \frac{t-1}{t+2}(x_t - x_{t-1}) \end{cases} \iff \frac{x_{t+1}-x_t}{\sqrt{\eta}} = \frac{t-1}{t+2}\cdot\frac{x_t - x_{t-1}}{\sqrt{\eta}} - \sqrt{\eta}\cdot\nabla f(y_t) \quad (A_1)$$

Step 1: Let $t = \frac{\tau}{\sqrt{\eta}}$ ($\tau = t\cdot\sqrt{\eta}$). For the process $X(\tau)$, suppose $x_t \approx X(t\cdot\sqrt{\eta}) \equiv X(\tau)$.

$x_{t+1} \approx X(t\cdot\sqrt{\eta} + \sqrt{\eta}) \equiv X(\tau + \sqrt{\eta})$

Step 2: By Taylor expansion w.r.t. time:

$$\frac{(x_{t+1} - x_t)}{\sqrt{\eta}} = \dot{X}(t) + \frac{1}{2}\ddot{X}(t)\cdot\sqrt{\eta} + o(\sqrt{\eta}) \quad -(A_2)$$

$$\frac{(x_t - x_{t-1})}{\sqrt{\eta}} = \dot{X}(t) - \frac{1}{2}\ddot{X}(t)\sqrt{\eta} + o(\sqrt{\eta}) \quad -(A_3)$$

(Cont.) By Combining (A_1)-(A_3), we have

$$\dot{X}(\tau) + \frac{1}{2}\ddot{X}(\tau)\cdot\sqrt{\eta} \approx \left(1 - \frac{3\sqrt{\eta}}{\tau}\right)\cdot\left(\dot{X}(\tau) - \frac{1}{2}\ddot{X}(\tau)\cdot\sqrt{\eta}\right) - \sqrt{\eta}\cdot\nabla f(X(\tau))$$

By (A_2) By (A_3)

$$\Rightarrow \quad \ddot{X}(\tau) + \frac{3}{\tau}\dot{X}(\tau) + \nabla f(X(\tau)) \approx 0$$

Solution to ODE and Convergence Rate

Nesterov's ODE

$$\ddot{X}(\tau) + \frac{3}{\tau}\dot{X}(\tau) + \nabla f(X(\tau)) = 0, \quad \tau > 0$$

- The above ODE can be solved via standard ODE theory as:

- Theorem: Let $X(\tau)$ be the unique solution to the above ODE with initial conditions $X(0) = x_0$ and $\dot{X}(0) = 0$. For any $\tau > 0$, we have

$$f(X(\tau)) - f(x^*) \leq \frac{2L\|x_0 - x^*\|^2}{\tau^2}$$

- Remark: This somewhat explains the $O\left(\frac{1}{t^2}\right)$ convergence rate of Nesterov's

W. Su, S. Boyd, and E. Candes, "A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights," NIPS 2014.

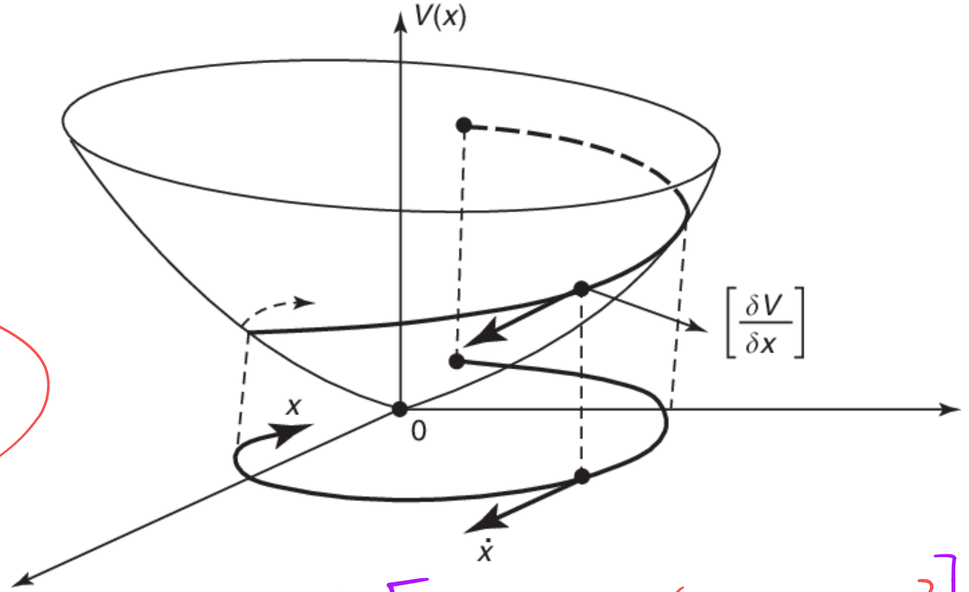
Lyapunov Function for Solving ODEs

To motivate the proof idea, let's take a slightly simpler ODE as

$$\dot{X}(\tau) + \nabla f(X(\tau)) = 0, \quad \tau > 0$$

Construct $\mathscr{E}$ Lyapunov function (or an energy function)

$$V(t) := t\big(f(X(t)) - f(x^*)\big) + \frac{\|X(t) - x^*\|^2}{2}$$



If we can show that $V(t)$ is decreasing with t, then we have a convergence rate

$$f(X(t)) - f(x^*) \le \frac{\|x_0 - x^*\|^2}{2t}$$

$$f(X(t)) - f(x^*) = \frac{1}{t}\left[V(t) - \left(\frac{\|X(t)-X^*\|^2}{2}\right)\right]$$
$$\le \frac{1}{t}\left[V(0) - \left(\frac{\|X(t)-X^*\|^2}{2}\right)\right]$$
$$\le \frac{1}{t}V(0)$$

Proof of ODE Solution

Step 1: Construct an energy function as

$$\mathcal{E}(t) := t^2\left(f(X(t)) - f(x^*)\right) + 2\left\| X(t) + \frac{t}{2}\dot{X}(t) - x^* \right\|^2$$

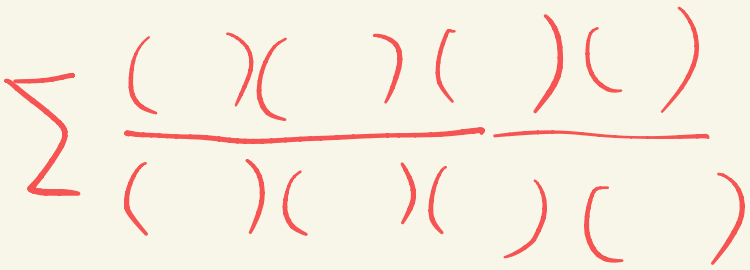
Then, the derivative of $\mathcal{E}(t)$ is

$$\dot{\mathcal{E}} = 2t\left(f(X(t)) - f(x^*)\right) + t^2 \cdot \nabla f^T \cdot \dot{X} + 4\left(X(t) + \frac{t}{2}\dot{X} - x^*\right)^T\left(\frac{3}{2}\dot{X} + \frac{t}{2}\ddot{X}\right)$$

Step 2: By using that $\ddot{X} + \frac{3}{t}\dot{X} + \nabla f(X) = 0$, we have

$$\dot{\mathcal{E}} = 2t\left(f(X) - f(x^*)\right) + 4\left(X - x^*\right)^T\left(-\frac{t}{2}\nabla f(X)\right)$$

$$= 2t\left(f(X) - f(x^*)\right) - 2t\left(X - x^*\right)^T \nabla f(X) \leq 0 \quad \text{(why?)}$$



(Cont.)

Step3: By the monotonicity of $\mathcal{E}$ and non-negativity of $2\cdot\|X+\frac{t\dot{X}}{2}-x^*\|^2$,

we have $$f(X(t))-f(x^*) \leq \frac{\mathcal{E}(t)}{t^2}$$

$$\leq \frac{\mathcal{E}(0)}{t^2}$$

$$= \frac{2\cdot\|X_0-x^*\|^2}{t^2}$$

The Mysterious Number 3

$$\ddot{X}(\tau) + \frac{3}{\tau}\dot{X}(\tau) + \nabla f(X(\tau)) = 0, \quad \tau > 0$$

3 appears to be the smallest number that can guarantee $O(\frac{1}{t^2})$ convergence

By (Su, Boyd, Candes, NIPS 2014), this "3" can be replaced by any value $r \geq 3$, i.e.,

$$\ddot{X}(\tau) + \frac{r}{\tau}\dot{X}(\tau) + \nabla f(X(\tau)) = 0, \quad \tau > 0$$

Similar convergence rate still hold under $r \geq 3$!

A Family of Generalized Nesterov's Scheme

$$\ddot{X}(\tau) + \frac{r}{\tau}\dot{X}(\tau) + \nabla f(X(\tau)) = 0, \quad \tau > 0$$

- Theorem: Let $X(\tau)$ be the unique solution to the above ODE with initial conditions $X(0) = x_0$ and $\dot{X}(0) = 0$. For any $\tau > 0$, we have

$$f(X(\tau)) - f(x^*) \leq \frac{(r-1)^2 \|x_0 - x^*\|^2}{2\tau^2}$$

W. Su, S. Boyd, and E. Candes, "A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights," NIPS 2014.

Proof of Generalized Scheme

Step1: Construct a "modified" energy function

$$\mathcal{E}(t) = \frac{2t^2}{r-1}\left(f(X(t)) - f(x^*)\right) + (r-1)\left\| X(t) + \frac{t}{r-1}\dot{X}(t) - x^* \right\|^2$$

Step2:

$$\dot{\mathcal{E}} = \frac{4t}{r-1}\left(f(X) - f(x^*)\right) + \frac{2t^2}{r-1}\cdot \nabla f^T \dot{X} + 2\cdot\left(X + \frac{t}{r-1}\dot{X} - x^*\right)^T \left(r\dot{X} + t\ddot{X}\right)$$

$$= \frac{4t}{r-1}\left(f(X) - f(x^*)\right) - 2t\cdot (X - x^*)^T \nabla f(X)$$

↑ by that $r\dot{X} + t\ddot{X} = -t\cdot\nabla f(X)$

Step3: Therefore,

$$\dot{\mathcal{E}} = \frac{4t}{r-1}\left(f(X)-f(x^*)\right) - 2t\cdot(X-x^*)^T\nabla f(X)$$

$$= \underbrace{2t\cdot\left(f(X)-f(x^*)\right) - 2t\cdot(X-x^*)^T\nabla f(X)}_{\leq 0} - \frac{2(r-3)t}{r-1}\underbrace{\left(f(X)-f(x^*)\right)}_{\geq 0}$$

(This implies that $\mathcal{E}(t)$ is non-increasing)

Step4: Hence, $\frac{2t^2}{r-1}\left(f(X(t))-f(x^*)\right) \leq \mathcal{E}(t) \leq \mathcal{E}(0) = (r-1)\cdot\|X_0-x^*\|^2$

Remark: The Extrapolation Coefficient

- Another widely-used extrapolation coefficient is

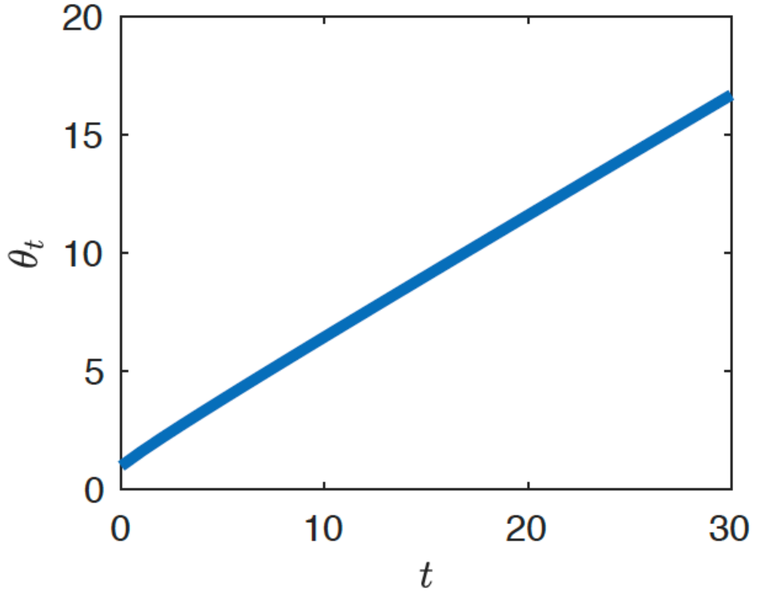
$$\theta_{t+1} = \frac{1 + \sqrt{1 + 4\theta_t^2}}{2} \quad \text{(With initial condition } \theta_0 = 1\text{)}$$

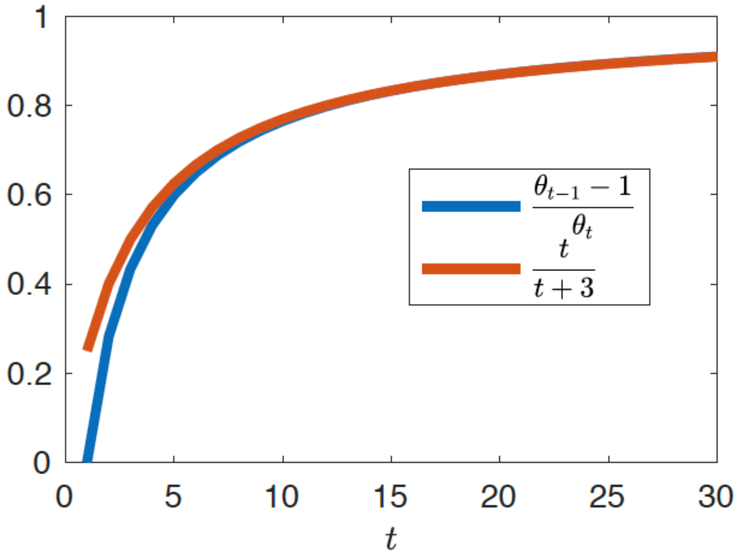
Moreover, the corresponding Nesterov's update is:

$$x_{t+1} = y_t - \eta_t \nabla f(y_t)$$

$$y_{t+1} = x_{t+1} + \underbrace{\frac{\theta_t - 1}{\theta_{t+1}}}_{\text{Extrapolation term}} \cdot (x_{t+1} - x_t)$$

(This is the original form proposed by Nesterov)





- Fact: The two coefficients are asymptotically equivalent

Remark: Monotonicity

- In general, NAG is not monotone

- To impose monotonicity, there are two common approaches
  - Restarting when $f(x_{t+1}) > f(x_t)$
  - Restarting when ∇f

A Natural Question: Is NAG optimal?

A Lower Bound

- An Interesting Fact: No first-order methods can improve over Nesterov's method in general (in other words, Nesterov's method is optimal in terms of "first-order oracle complexity")

- Theorem: There exists a convex and L-smooth function such

$$f(x_t) - f(x^*) \geq \frac{3L\|x_0 - x^*\|^2}{32(t+1)^2}$$

where $x_\tau \in x_0 + \text{span}\{\nabla f(x_0), \cdots, \nabla f(x_{\tau-1})\}$, for all $1 \leq \tau \leq t$

- Question: What's the implication of the above condition of x_τ?

Nemirovsky and Yudin, "Problem complexity and method efficiency in optimization", 1983

Constructing an Example for the Lower Bound (1/3)

$$\min_{x \in \mathbb{R}^{2t+1}} f(x) = \frac{L}{4}\left(\frac{1}{2}x^\top A x - e_1^\top x\right)$$

$$A \in \mathbb{R}^{(2t+1)\times(2t+1)} \qquad e_1 \in \mathbb{R}^{2t+1}$$

$$A = \begin{bmatrix} 2 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & -1 & 2 & -1 \\ 0 & \cdots & 0 & 0 & -1 & 2 \end{bmatrix} \qquad e_1 = \begin{bmatrix} 1 \\ 0 \\ \cdots \\ \cdots \\ \cdots \\ 0 \end{bmatrix}$$

Fact 1: f is convex and L-smooth

Fact 2: $\nabla f(x) = \frac{L}{4}Ax - \frac{L}{4}e_1$

Fact 3: Optimal solution $x^* = (x_1^*, \cdots, x_n^*)$ is given by $x_i^* = 1 - \frac{i}{2t+2}$

Fact 4: $\|x^*\|^2 \leq \frac{2t+2}{3}$ and $f(x^*) = \frac{L}{8}\left(\frac{1}{(2t+1)+1} - 1\right)$

Constructing an Example for the Lower Bound (2/3)

$$\min_{x \in \mathbb{R}^{2t+1}} f(x) = \frac{L}{4}\left(\frac{1}{2}x^\top A x - e_1^\top x\right)$$

$$A \in \mathbb{R}^{(2t+1)\times(2t+1)} \quad e_1 \in \mathbb{R}^{2t+1}$$

$$A = \begin{bmatrix} 2 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & -1 & 2 & -1 \\ 0 & \cdots & 0 & 0 & -1 & 2 \end{bmatrix} \quad e_1 = \begin{bmatrix} 1 \\ 0 \\ \cdots \\ \cdots \\ \cdots \\ 0 \end{bmatrix}$$

Fact 5: $\underbrace{\text{Span}\{\nabla f(x_0), \cdots \nabla f(x_{t-1})\}}_{=:C_t} = \text{Span}\{e_1, \cdots, e_t\}$ if $x_0 = 0$

Under **any** first-order method, in each iteration, one can expand the search space by **at most 1 dimension**

Constructing an Example for the Lower Bound (3/3)

Let's derive the lower bound!

Step 1: Given that $x_0 = 0$, we have

$$f(x_t) \geqslant \inf_{x \in C_t} f(x) = \frac{L}{8}\left(\frac{1}{t+1} - 1\right)$$

Step 2: Therefore, we have

$$\frac{f(x_t) - f(x^*)}{\|x_0 - x^*\|^2} \geqslant \frac{\frac{L}{8}\left(\frac{1}{t+1} - \frac{1}{2t+2}\right)}{\frac{1}{3}(2t+2)} = \frac{3L}{32(t+1)^2}$$

Another Interpretation of NAG

Aleksandar Botev, Guy Lever, and David Barber, “Nesterov’s Accelerated Gradient and Momentum as approximations to Regularised Update Descent,” IJCNN 2017

Remark: An Alternative Expression of NAG Updates

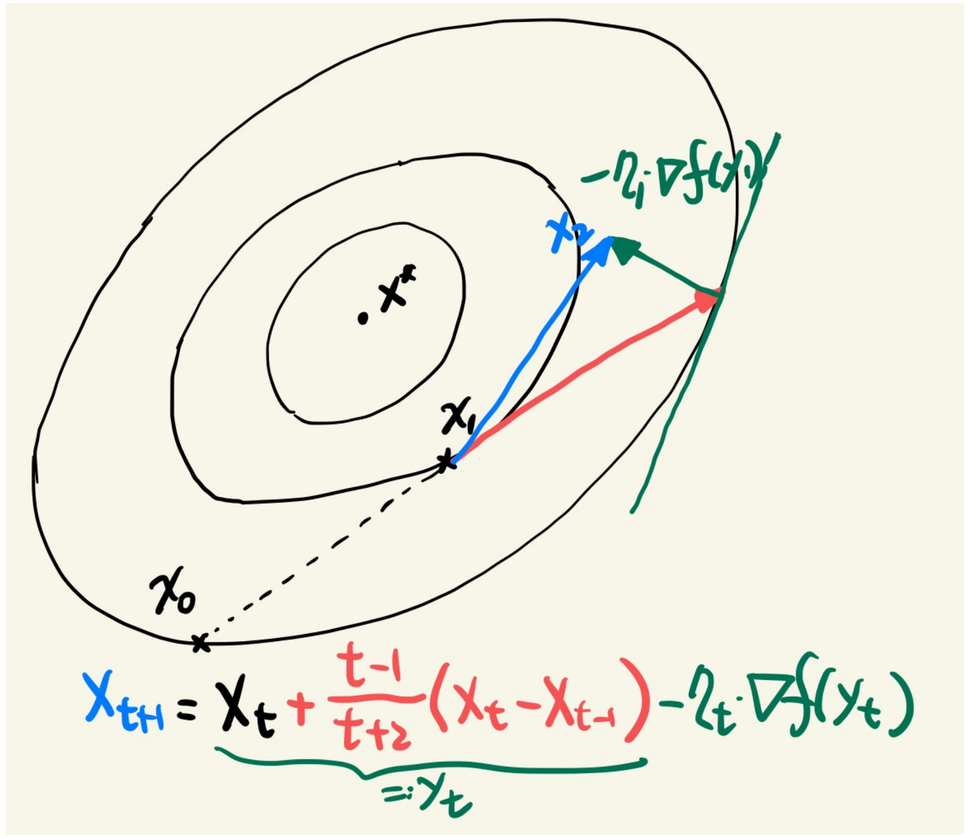
$$x_{t+1} = y_t - \eta_t \nabla f(y_t)$$

$$y_t = x_t + \beta_t (x_t - x_{t-1})$$

$$v_{t+1} = x_{t+1} - x_t \quad \longleftrightarrow$$

$$x_{t+1} = x_t + v_{t+1}$$

$$v_{t+1} = \beta_t v_t - \eta_t \nabla f(x_t + \beta_t v_t)$$



Interpreting NAG as GD on Regularized Function

Define $F(x_t, v_t) := f(x_t + \beta_t v_t) + \frac{\gamma}{2}\|v_t\|^2$

(Clearly, the x^* that minimizes F also minimizes f)

Suppose we take a gradient step of F:

$$v_{t+1} = v_t - \eta_t\big(\nabla f(x_t + \beta_t v_t) + \gamma_t v_t\big) = \underbrace{(1 - \eta_t\gamma_t)}_{=:\beta_t} v_t - \eta_t \nabla f(x_t + \beta_t v_t)$$

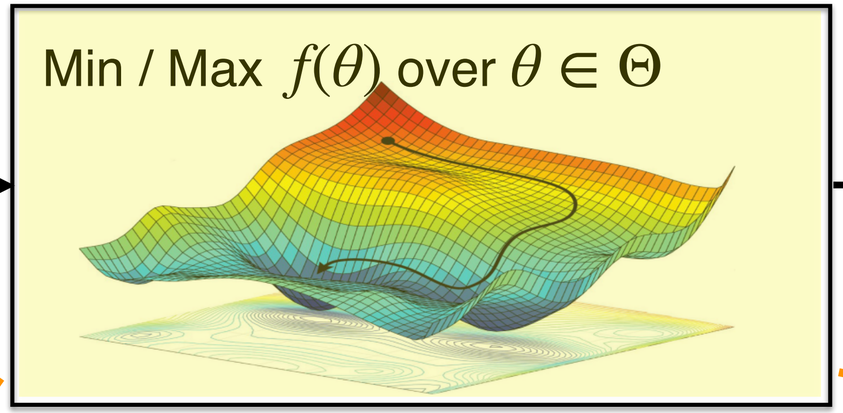
Then, we arrive at the NAG update

$$x_{t+1} = x_t + v_{t+1}$$

$$v_{t+1} = \beta_t v_t - \eta_t \nabla f(x_t + \beta_t v_t)$$

Case Study: Reinforcement Learning

Gradient Descent in Reinforcement Learning (RL)



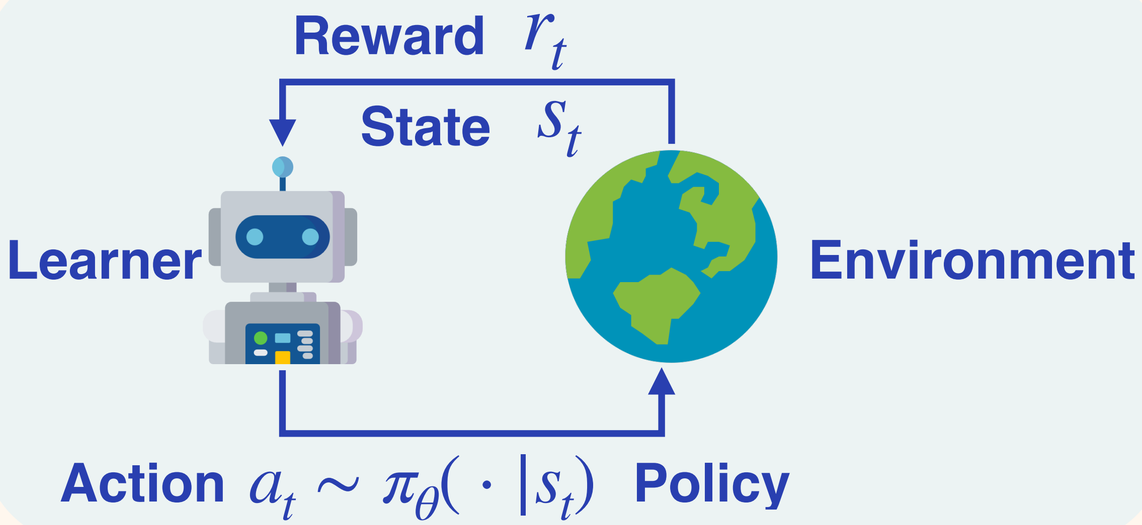
Next iterate θ_t

Information about f at θ_t

Policy Gradient Algorithm

μ : initial state distribution

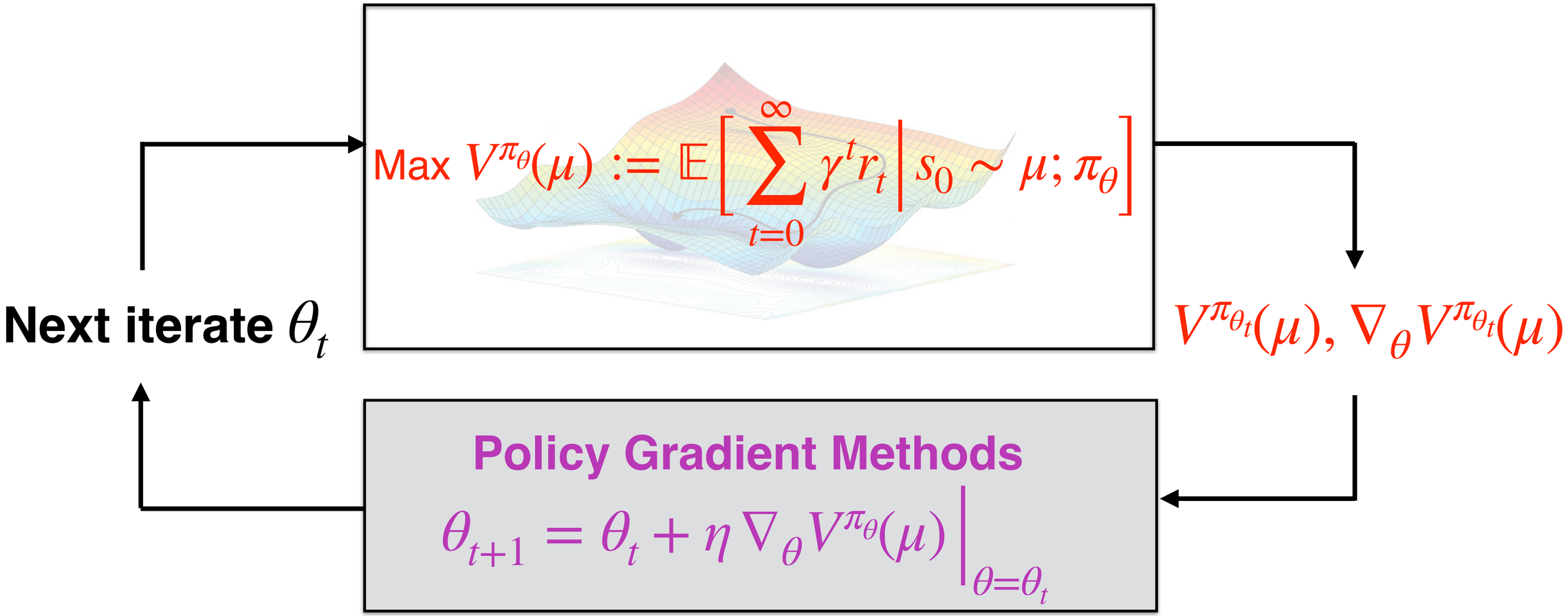
γ : discount factor



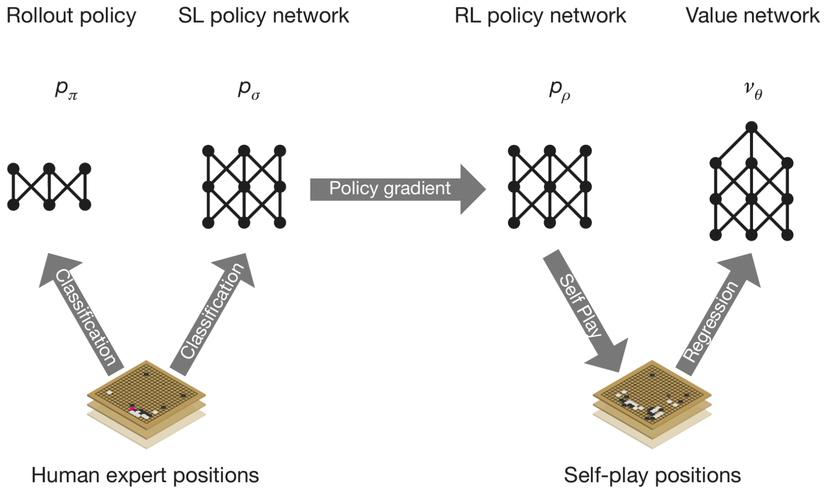
RL objective function

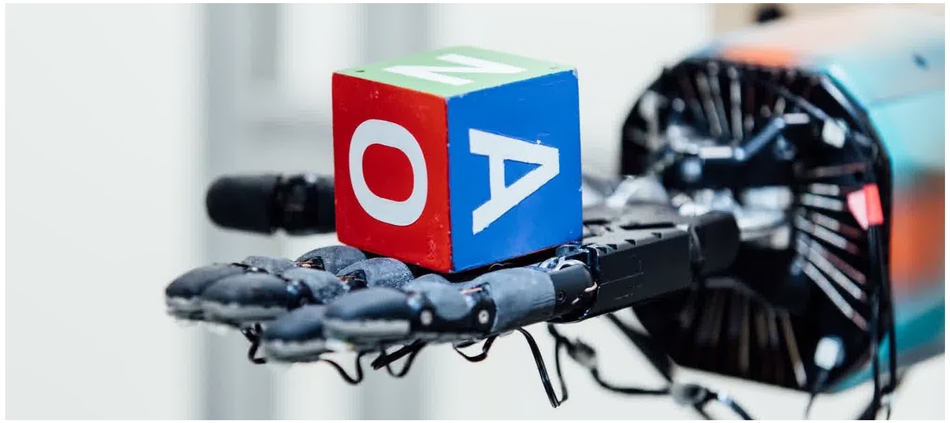
$$f(\theta) \equiv V^{\pi_\theta}(\mu) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 \sim \mu; \pi_\theta\right]$$

Policy Gradient (PG): A Conceptually Simple but Fundamental Algorithm



Policy Gradient (PG): Empirical Success





AlphaGo (Silver at al., Nature 2016)

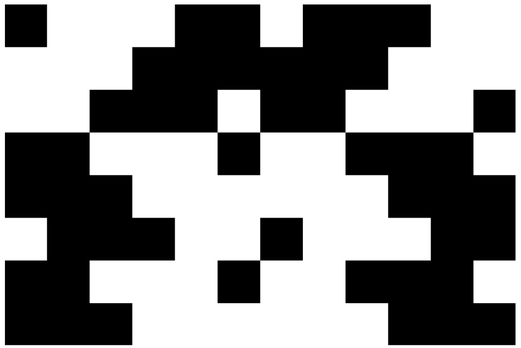
Use PG for improving upon a pre-trained policy

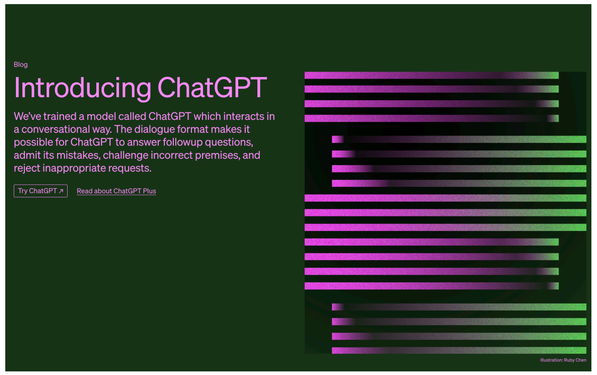
OpenAI Five (OpenAI, 2018)

Use PPO (a variant of PG) for policy training

Research

Learning to summarize with human feedback





ChatGPT & RLHF (OpenAI)

Use PPO (a variant of PG) for policy fine-tuning

Silver et al., Mastering the game of Go with deep neural networks and tree search, Nature 2016

OpenAI et al., Learning Dexterous In-Hand Manipulation, arXiv 2018

Stiennon et al., Learning to summarize with human feedback, NeurIPS 2020

Convergence of PG in Reinforcement Learning

Theorem [Mei et al., 2020]

Under softmax policies and $\eta = \dfrac{(1-\gamma)^3}{8}$, PG achieves the convergence rate as

$$V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \leq C \cdot \frac{1}{t}$$

where C is a problem-dependent constant (depending on S, A, μ, γ)

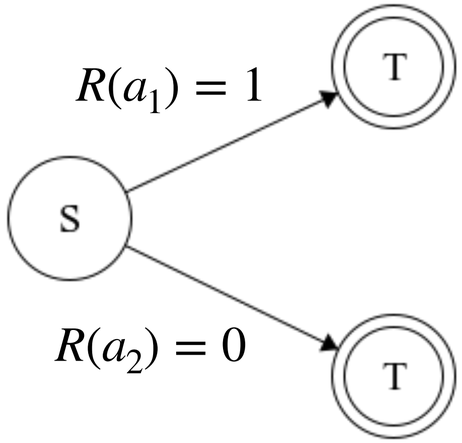
- (Mei et al., 2020) proves this by discovering an analytical condition
- In (Chen et al., 2024), we recover this result by discovering a structural property

Mei et al., On the Global Convergence Rates of Softmax Policy Gradient Methods, ICML 2020

Chen et al., “Accelerated Policy Gradient: On the Convergence Rates of Nesterov Momentum for Reinforcement Learning”

Our Structural Finding: “Local Concavity” in RL

- Use 1-state, 2-action MDPs as an example



Policy parameters: $\theta \equiv [\theta_{a_1}, \theta_{a_2}]$

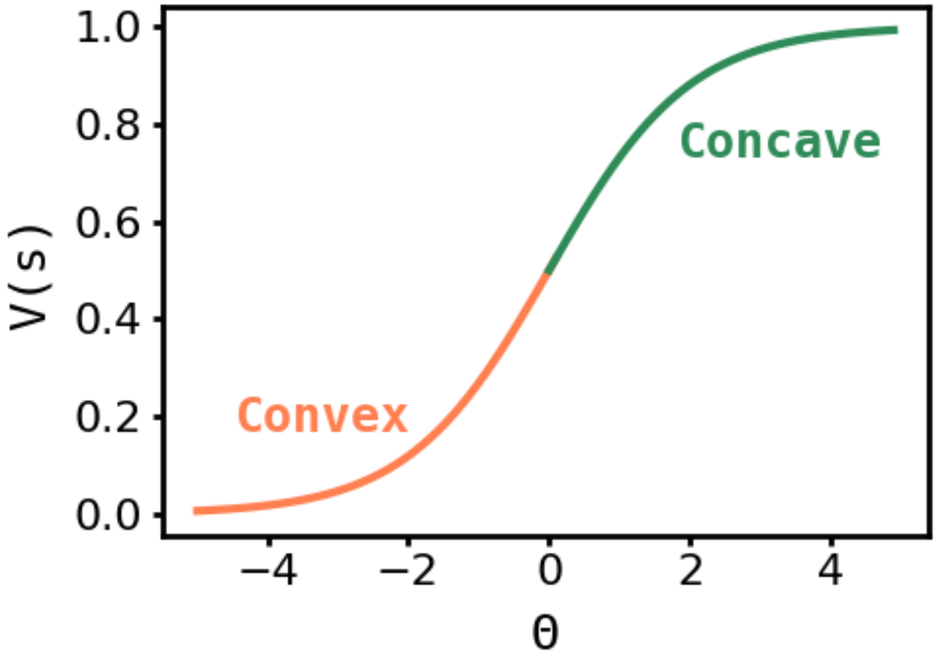
Reward function: $R \equiv [R(a_1), R(a_2)]$

$$V^{\pi_\theta}(s) = \pi_\theta(a_1) \cdot R(a_1) + \pi_\theta(a_2) \cdot R(a_2)$$

$$= \frac{1}{1 + \exp(\theta_{a_2} - \theta_{a_1})}$$

$$= \frac{1}{1 + \exp(-\Theta)}$$

(where $\Theta := \theta_{a_1} - \theta_{a_2}$)

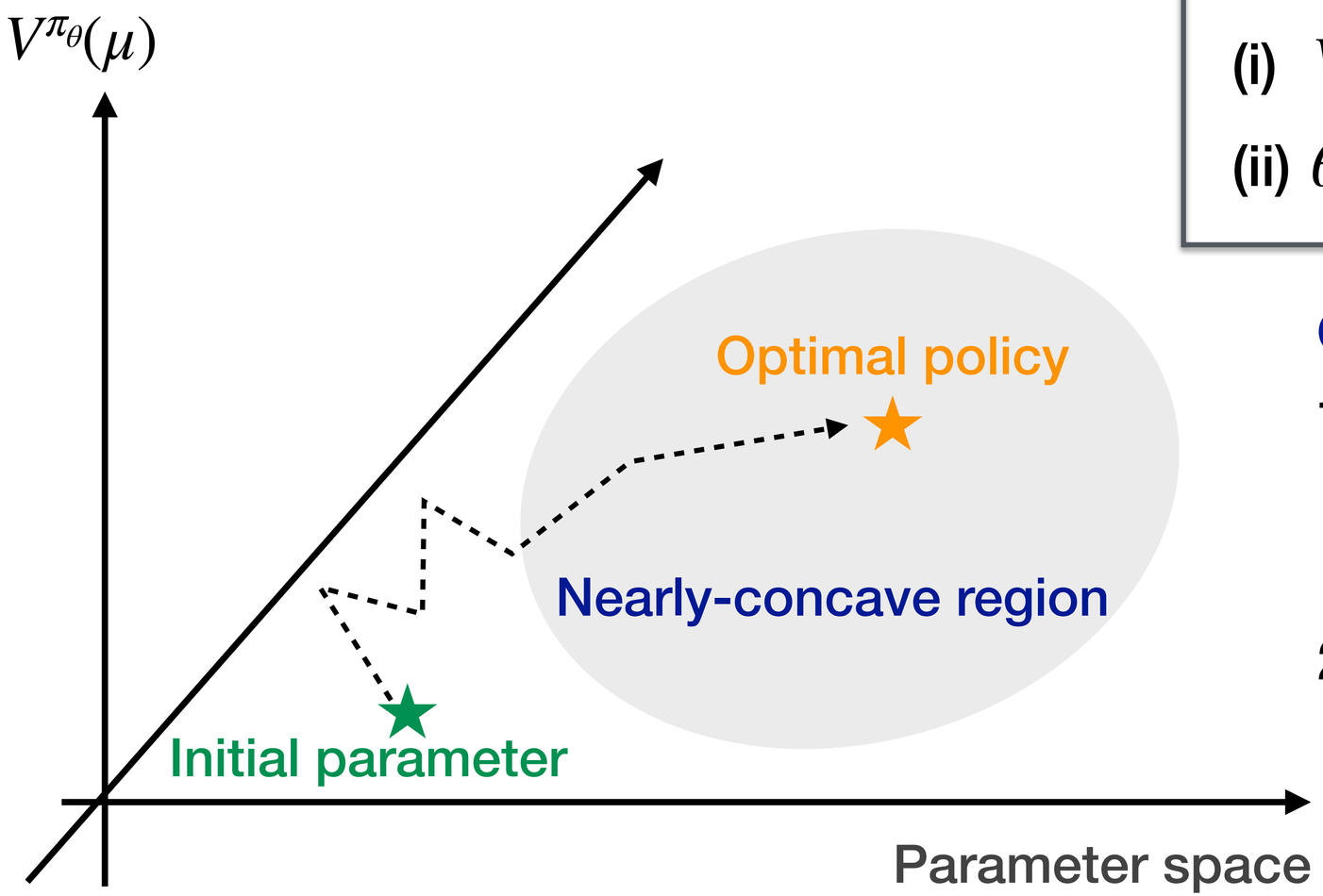


Observation 1: RL objective function is non-concave

Observation 2: However, RL objective function is “locally concave”

Observation 3: PG would work if it always enter the “concave” region after finite time

"Local Near-Concavity" Under General MDPs



Local nearly-concave region: This region contains θs that satisfy

(i) $V^{\pi_\theta}(s) > Q^*(s, a_2(s))$

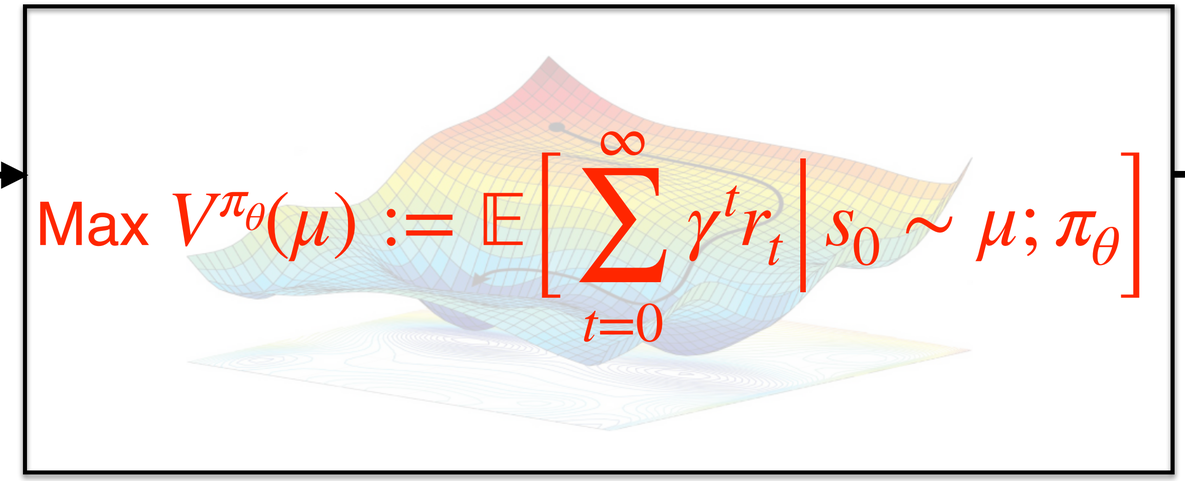
(ii) $\theta_{s,a^*(s)} - \theta_{s,a} > M$, for some $M > 0$

Our Proof Idea:

1. Regardless of initialization, PG always enters a nearly-concave region in finite T and stays there
2. We can then directly use the standard convergence rate of GD, which is $O(1/t)$

Next question: Could we leverage the "local near-concavity" and improve the $O(1/t)$ rate?

Accelerated Policy Gradient (APG): Nesterov's Momentum for RL



$V^{\pi_{\theta_t}}(\mu), \nabla_\theta V^{\pi_{\theta_t}}(\mu)$

Next iterate θ_t

Accelerated Policy Gradient

$$\theta^{(t)} \leftarrow \omega^{(t-1)} + \eta^{(t)} \nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\omega^{(t-1)}}$$

$$\varphi^{(t)} \leftarrow \theta^{(t)} + \frac{t-1}{t+2}(\theta^{(t)} - \theta^{(t-1)})$$

$$\omega^{(t)} \leftarrow \begin{cases} \varphi^{(t)}, & \text{if } V^{\pi_\varphi^{(t)}}(\mu) \geq V^{\pi_\theta^{(t)}}(\mu), \\ \theta^{(t)}, & \text{otherwise.} \end{cases}$$

Accelerated PG: Convergence Rate

Theorem (Chen et al., 2024; Informal)

Under Accelerated PG with learning rate $\eta_t = \frac{t}{t+1} \cdot \frac{(1-\gamma)^3}{16}$, we have

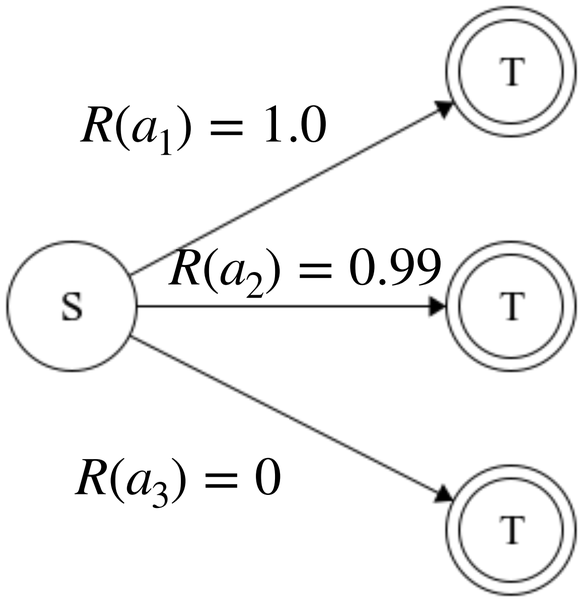
$$V^*(\mu) - V^{\pi_\theta^{(t)}}(\mu) = \tilde{O}\left(\frac{1}{t^2}\right)$$

Despite the RL non-concavity, APG improves the convergence rate over PG (with nearly-constant step sizes)

Yen-Ju Chen, Nai-Chieh Huang, Ching-pei Lee, and Ping-Chun Hsieh, "Accelerated Policy Gradient: On the Convergence Rates of Nesterov Momentum for Reinforcement Learning," ICML 2024 (available at https://arxiv.org/abs/2310.11897)

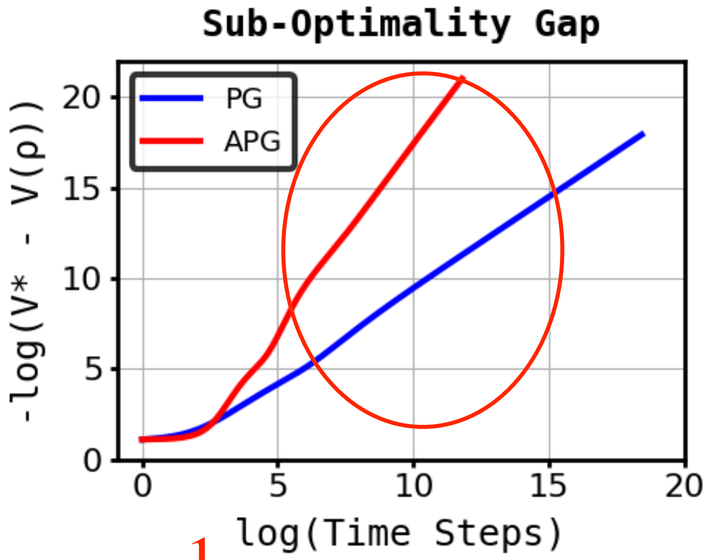
Accelerated PG: Small-Scale Experiments

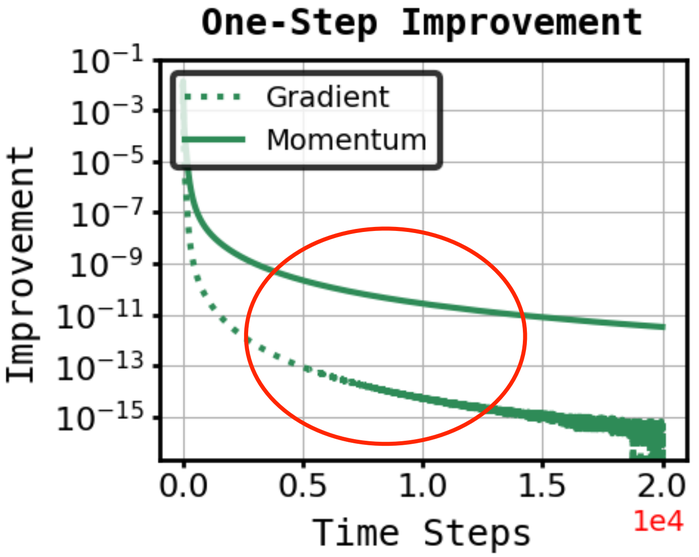
- Use 1-state, 3-action MDPs to empirically evaluate the rate of APG vs PG



$$V^* - V^{(t)}(s) \leq \frac{c}{t^2}$$

$$\Rightarrow -\log(V^* - V^{(t)}(s)) \geq -\log(\frac{c}{t^2}) = 2\log(t) - \log(c)$$



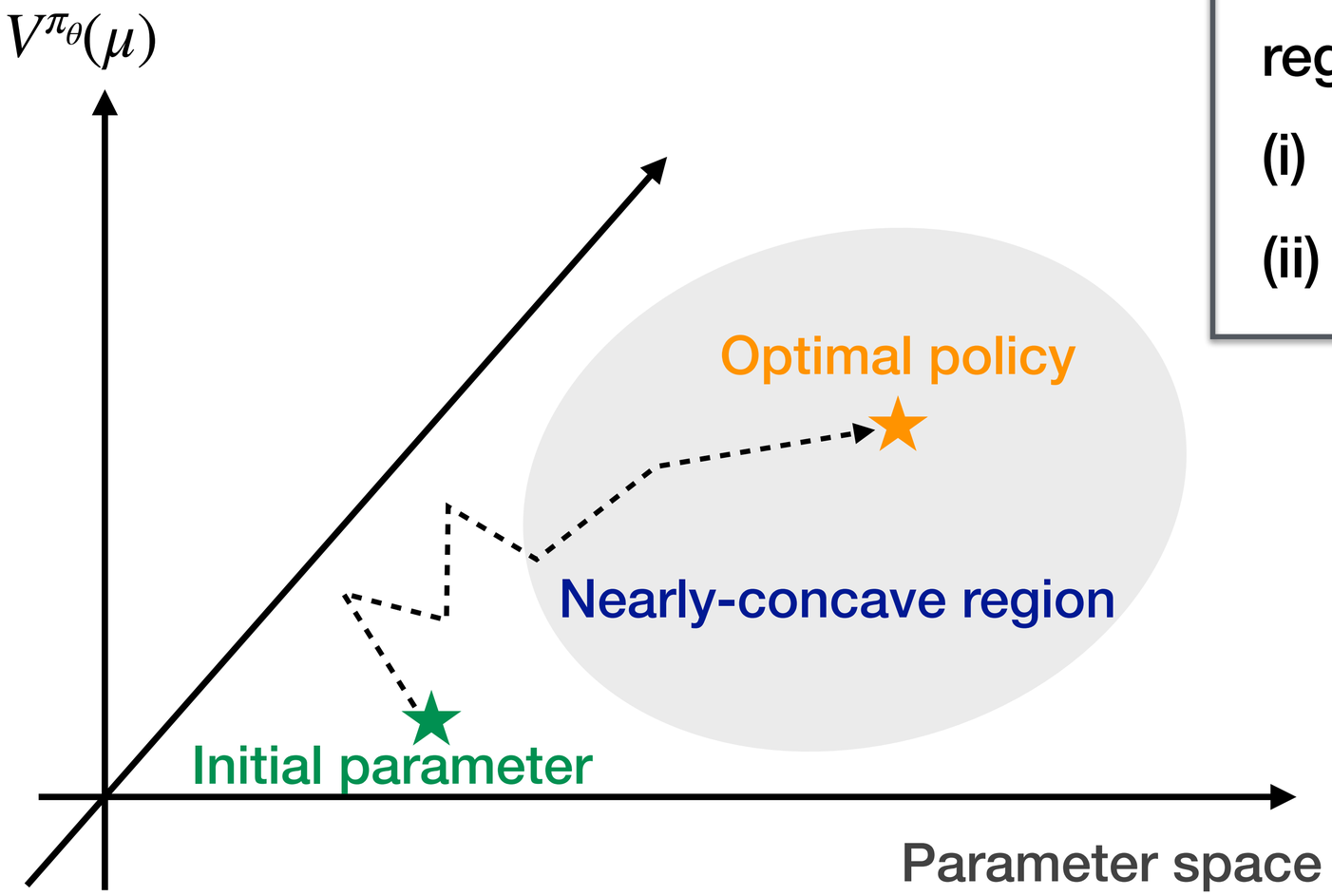


APG: slope is 2 in log-log plot → $O(\frac{1}{t^2})$ rate

PG: slope is 1 in log-log plot → $O(\frac{1}{t})$ rate

Momentum contributes most of the improvement!

Proof Idea: Use Local Near-Concavity



Local nearly-concave region: This region contains θs that satisfy

(i) $V^{\pi_\theta}(s) > Q^*(s, a_2(s))$

(ii) $\theta_{s,a^*(s)} - \theta_{s,a} > M$, for some $M > 0$

Our Proof Idea:

1. Regardless of initialization, APG always enters a nearly-concave region in finite T and stays there
2. We can then directly use the standard convergence rate of NAG, which is $O(1/t^2)$

The same proof procedure still works for APG!