The Physical Systems Behind Optimization Algorithms

Lin F. Yang, Raman Arora, Vladimir Braverman, and Tuo Zhao

1 Princeton University, 2 Johns Hopkins University, 3 GaTech

The Classic Optimization Problem:

\[
\begin{align*}
\text{argmin}_{x \in \mathbb{R}^n} f(x)
\end{align*}
\]

Simple first order algorithm:

\[
\begin{align*}
x_k & \leftarrow y_k - \eta \nabla f(y_k), \\
y_k & = x_k + \alpha x_k - x_{k-1},
\end{align*}
\]

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

Iteration complexity:

<table>
<thead>
<tr>
<th>Methods</th>
<th>Error After k Iters</th>
<th>Assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGD</td>
<td>( \exp(-\kappa/k) )</td>
<td>Strongly convex</td>
</tr>
<tr>
<td>AGD</td>
<td>( 1/k )</td>
<td>Convex</td>
</tr>
<tr>
<td>NAG</td>
<td>( \exp(-\kappa/k^2) )</td>
<td>Strongly Convex</td>
</tr>
<tr>
<td>NAG</td>
<td>( 1/k^2 )</td>
<td>Convex</td>
</tr>
</tbody>
</table>

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

How to connect discrete iterates to continuous iterates?

Set \( k(t) = \frac{\nu t}{\eta} \)

Taylor expansion the algorithm iterates:

\[
\begin{align*}
\{x^{(k(t))} - x^*(t)\} = \frac{\nu t}{\eta} \frac{\sigma}{2} h(t)^2 + o(h), \\
\{x^{(k(t))} - x^{(k(t-1))}\} = \frac{\nu t}{\eta} \frac{\sigma}{2} h(t)^2 + o(h),
\end{align*}
\]

and \( \eta \nabla f(x^{(k(t-1))}) = \eta \nabla f(x^{(k(t))}) + o(h) \).

Set \( k(t) = \frac{\nu t}{\eta} \) for all algorithms. Let the algorithm tell us what is \( h \)!

A unified ODE that describe a damped oscillator system:

\[
\begin{align*}
m \ddot{x}(t) + c \dot{x}(t) + \nabla f(x(t)) = 0.
\end{align*}
\]

A unified view of the choice of \( h \):

VGD: massless system

\[
\begin{align*}
\dot{X} + \nabla f(X) = 0, \\
\Rightarrow m = 0, c = 1, \\
\Rightarrow h = \Theta(\eta),
\end{align*}
\]

NAG: massive system

\[
\begin{align*}
m \ddot{X} + c \dot{X} + \nabla f(X) = 0, \\
\Rightarrow h = \Theta(\eta), 1 - \alpha = \Theta(\eta).
\end{align*}
\]

References:


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 \( f(x) = \frac{1}{2} x' M x \), which corresponds to NAG

Our framework naturally extends to other optimization methods:

- Randomized coordinate gradient descent (RCGD)
- Randomized accelerated coordinate gradient descent (ARCG)

Our framework naturally extends to other optimization methods:

- PL-condition: \( \frac{1}{\kappa} < \frac{\mu}{L^2} \)

Critical damping: \( c^2 = 4mK \), which corresponds to NAG

A unified ODE that describe a damped oscillator system:

A unified view of the choice of \( h \):

VGD: massless system

\[
\begin{align*}
\dot{X} + \nabla f(X) = 0, \\
\Rightarrow m = 0, c = 1, \\
\Rightarrow h = \Theta(\eta),
\end{align*}
\]

NAG: massive system

\[
\begin{align*}
m \ddot{X} + c \dot{X} + \nabla f(X) = 0, \\
\Rightarrow h = \Theta(\eta), 1 - \alpha = \Theta(\eta).
\end{align*}
\]