A Differential Equation for Modeling Nesterov’s Accelerated Gradient Method

WeiJie Su, Stephen Boyd, and Emmanuel J. Candès

**Nesterov’s Scheme**

Nesterov’s scheme:

\[
\begin{align*}
&x_{k+1} = \frac{k-1}{k+2} x_k - \frac{1}{k+2} s_k, \\
&x_{k+1} = x_k - \frac{1}{k+2} s_k,
\end{align*}
\]

with \(y_k = x_k\). For step size \(s \leq 1/L\),

\[
f(x_k) - f(x^*) \leq \frac{2\|x_0 - x^*\|^2}{k+1}
\]

Rate optimal for this class.

**Mysteries**

Momentum coefficient: why \(\frac{k-1}{k+2}\)?

Convergence rate: why \(O\left(\frac{1}{t^2}\right)\)?

Gradient descent:

\[
x_k = x_{k-1} - \eta \nabla f(x_{k-1})
\]

with \(\eta \leq 1/L\), has convergence rate

\[
f(x_k) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2\eta}
\]

**Property**

a) If \(f(x) = ||x||^2/2\), the solution to the ODE:

\[
X(t) = \frac{2x(t)}{t} x_0
\]

\(J_1()\) Bessel function of first kind with order 1. For large \(t\),

\[
f(X(t)) = O(\|x_0\|^2/t^3)
\]

b) Finite scheme of ODE:

\[
\frac{X(t_{k+1}) - X(t_k)}{\Delta t} = \left(1 - \frac{3}{t} \right) X(t_k) - \Delta t \nabla f(X(t_k))
\]

The characteristic equation is approximately

\[
\det \left( \lambda^2 - 2 - \Delta t \lambda + \frac{3}{t} \right) = \lambda^3 + \frac{3}{t} \lambda = 0
\]

Numerical stability approx. equiv. to all the roots lie in the unit circle approx. equiv. to \(\Delta t < 2/t\).

c) Let \(\varepsilon(t) = t^2 (f(X(t)) - f(x^*)) + 2\|X(t) - x^*\|^2\) and the simple fact \(\varepsilon'(t) \leq 0\) leads to

\[
f(X(t)) - f(x^*) \leq \frac{2\|x_0 - x^*\|^2}{2t}
\]

**Generalization**

Replace \(3\) by a constant \(r > 3\),

\[
\dot{X} + \frac{1}{r} X + \nabla f(X) = 0
\]

From studying

\[
\varepsilon(t) = \frac{2t^2}{r} (f(X(t)) - f^*) + (r - 1)\|X(t) + \frac{t}{r} X(t) - x^*\|^2
\]

we get

\[
f(X(t)) - f(x^*) \leq \frac{(r-1)^2 \|x_0 - x^*\|^2}{2(r-3)}
\]

Suggests a family of generalized Nesterov’s schemes:

\[
x_k = y_{k-1} - \lambda \nabla f(y_{k-1})
\]

\(y_k = x_k - \lambda \left(1 - \frac{r-1}{r+1} \right) (x_k - x_{k-1})\)

Convergence rate with \(s \leq 1/L\),

\[
f(x_k) - f(x^*) \leq \frac{(r-1)^2 \|x_0 - x^*\|^2}{2(r+1)(r-2)}
\]

\[
\sum_{k=1}^t (k+1) (f(x_k) - f(x^*)) \leq \frac{(r-1)^2 \|x_0 - x^*\|^2}{2(r-3)}
\]

Works for \(r\) smooth + non-smooth.

\[\sqrt{s} \text{ VS } s \text{ IN EACH ITERATION}\]

**Speed Restart**

\(\lambda_x\) is set to zero whenever \(\|X(t)\|/\|x^*\| = 0\) and then \(\lambda_x\) grows just as \(t\),

\[
X(t) + \frac{1}{\lambda_x} X(t) + \nabla f(X(t)) = 0
\]

Prevent long-term recession of \(\|X\|\). If \(f\) is \(\nu\)-strongly convex,

\[
f(x^*(t)) - f(x(t)) \leq \epsilon \lambda_x^t \|x_0 - x^*\|^2 - \nu\tau
\]

\(\nu\) is not too small.

**Small Step Size**

As step size \(\to 0\), trajectory approaches a continuous curve.

<table>
<thead>
<tr>
<th>(s)</th>
<th>(t = 0.5)</th>
<th>(t = 0.1)</th>
<th>(t = 0.01)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed</td>
<td>(\sqrt{s})</td>
<td>(\sqrt{s})</td>
<td>(\sqrt{s})</td>
</tr>
<tr>
<td>(\sqrt{s}) VS (s) IN EACH ITERATION</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Gradient descent:** moves \(s\) on curve of \(X + \nabla f(X) = 0\), whose numerical integration step size is approx. \(2/\sqrt{s}\).

**Nesterov’s Scheme:** moves \(\sqrt{s}\) on the curve of \(X + \frac{1}{r} X + \nabla f(X) = 0\), whose numerical integration step size is approx. \(2/r\).

<table>
<thead>
<tr>
<th>(s)</th>
<th>(t = 0.5)</th>
<th>(t = 0.1)</th>
<th>(t = 0.01)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed</td>
<td>(\sqrt{s})</td>
<td>(\sqrt{s})</td>
<td>(\sqrt{s})</td>
</tr>
<tr>
<td>(\sqrt{s}) VS (s) IN EACH ITERATION</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>