Smooth Strongly Convex Minimization
The Fastest-Known First-Order Method

Bryan Van Scoy

University of Wisconsin–Madison

International Symposium on Mathematical Programming
Bordeaux
July 5, 2018
\[
\text{minimize} \quad f(x) \\
\text{subject to} \quad x \in \mathbb{R}^d
\]

- \( f \) is \( L \)-smooth and \( \mu \)-strongly convex
- denote the optimizer as \( x^* \in \mathbb{R}^d \)
- \( \kappa := L/\mu \) is the condition ratio

**Main result**

We design a first-order method whose iterate sequence \( \{x_k\} \) satisfies

\[
\|x_k - x^*\| = \mathcal{O}(\rho^k)
\]

\[
f(x_k) - f(x^*) = \mathcal{O}(\rho^{2k})
\]

where \( \rho = 1 - 1/\sqrt{\kappa} \).

Compare with Nesterov's fast gradient method:

\[
\|x_k - x^*\| = \mathcal{O}(\rho^{k/2})
\]

\[
f(x_k) - f(x^*) = \mathcal{O}(\rho^k)
\]
Theorem (Nesterov, 2004)

The fast gradient method is “optimal” for the class of $L$-smooth and $\mu$-strongly convex functions.

Complexity: Number of iterations to obtain $\|x_k - x_\star\| \leq \varepsilon$

Rate of iterates: $\|x_k - x_\star\| = \mathcal{O}(\rho^k)$

<table>
<thead>
<tr>
<th>Method</th>
<th>Complexity</th>
<th>Rate of iterates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient method (stepsize $\frac{1}{L}$)</td>
<td>$\mathcal{O}(\kappa \ln(\frac{1}{\varepsilon}))$</td>
<td>$1 - \frac{1}{\kappa}$</td>
</tr>
<tr>
<td>Gradient method (stepsize $\frac{2}{L+\mu}$)</td>
<td>$\mathcal{O}(\kappa \ln(\frac{1}{\varepsilon}))$</td>
<td>$\frac{\kappa-1}{\kappa+1}$</td>
</tr>
<tr>
<td>Fast gradient method</td>
<td>$\mathcal{O}(\sqrt{\kappa} \ln(\frac{1}{\varepsilon}))$</td>
<td>$(1 - \frac{1}{\sqrt{\kappa}})^{k/2}$</td>
</tr>
<tr>
<td>Proposed method</td>
<td>$\mathcal{O}(\sqrt{\kappa} \ln(\frac{1}{\varepsilon}))$</td>
<td>$1 - \frac{1}{\sqrt{\kappa}}$</td>
</tr>
<tr>
<td>Lower bound</td>
<td>$\mathcal{O}(\sqrt{\kappa} \ln(\frac{1}{\varepsilon}))$</td>
<td>$\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$</td>
</tr>
</tbody>
</table>

Proposed method is twice as fast as Nesterov’s method

Method

gradient method
\[ x_{k+1} = x_k - \alpha \nabla f(x_k) \]

heavy ball method
\[ x_{k+1} = (1 + \beta)x_k - \beta x_{k-1} - \alpha \nabla f(x_k) \]

fast gradient method
\[ x_{k+1} = (1 + \beta)x_k - \beta x_{k-1} - \alpha \nabla f((1 + \beta)x_k - \beta x_{k-1}) \]

triple momentum method
\[ x_{k+1} = (1 + \beta)x_k - \beta x_{k-1} - \alpha \nabla f((1 + \gamma)x_k - \gamma x_{k-1}) \]

<table>
<thead>
<tr>
<th>Method</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM</td>
<td>$\frac{1}{L}$</td>
<td>$\frac{1}{4} \left( \sqrt{L} + \sqrt{\mu} \right)^2$</td>
<td>( \frac{1}{\sqrt{\kappa+1}} ) $\left( \sqrt{\kappa-1} \right)^2$</td>
</tr>
<tr>
<td>HBM</td>
<td>$\frac{1}{L}$</td>
<td>$\frac{1}{4} \left( \sqrt{L} + \sqrt{\mu} \right)^2$</td>
<td>( \frac{1}{\sqrt{\kappa+1}} ) $\left( \sqrt{\kappa-1} \right)^2$</td>
</tr>
<tr>
<td>FGM</td>
<td>$\frac{1}{L}$</td>
<td>$\frac{1}{\sqrt{\kappa+1}}$</td>
<td>( \frac{1}{\sqrt{\kappa+1}} ) $\left( \sqrt{\kappa-1} \right)^2$</td>
</tr>
<tr>
<td>TMM</td>
<td>$\frac{2\sqrt{L} - \sqrt{\mu}}{L\sqrt{L}}$</td>
<td>$\frac{1}{\sqrt{\kappa+1}}$</td>
<td>( \frac{1}{2\kappa+\sqrt{\kappa-1}} ) $\left( \sqrt{\kappa-1} \right)^2$</td>
</tr>
</tbody>
</table>
Triple momentum method

\[ x_{k+1} = (1 + \beta)x_k - \beta x_{k-1} - \alpha \nabla f((1 + \gamma)x_k - \gamma x_{k-1}) \]

Parameters:

- \( \rho = 1 - \frac{1}{\sqrt{\kappa}} \)
- \( \alpha = \frac{1 + \rho}{L} \)
- \( \beta = \frac{\rho^2}{2 - \rho} \)
- \( \gamma = \frac{\rho^2}{(1 + \rho)(2 - \rho)} \)

Condition ratio \( \kappa := \frac{L}{\mu} \)

**Theorem (Van Scoy, Freeman, Lynch, 2017)**

Suppose \( f \) is \( L \)-smooth and \( \mu \)-strongly convex with minimizer \( x_\star \in \mathbb{R}^d \). Then for any initial conditions \( x_0, x_{-1} \in \mathbb{R}^d \), there exists a constant \( c > 0 \) such that

\[ \|x_k - x_\star\| \leq c \rho^k \quad \text{for all } k \geq 1. \]
Convergence rate: \[ \| x_k - x_* \| \leq c \rho^k \]

Iterations to converge \( \propto \frac{1}{\ln \rho} \)
$f$ smooth strongly convex

- **HBM** does not converge if $\kappa \geq (2 + \sqrt{5})^2 \approx 17.94$
- For **FGM**, Nesterov proved the rate $\sqrt{1 - \frac{1}{\sqrt{\kappa}}}$ which is loose
- **TMM** converges faster than **FGM**
Simulations

Objective function:

$$f(x) = \sum_{i=1}^{n} g(a_i^T x - b_i) + \frac{\mu}{2} \|x\|^2, \quad x \in \mathbb{R}^d$$

where

$$g(y) = \begin{cases} 
\frac{1}{2} y^2 e^{-r/y}, & y > 0 \\
0, & y \leq 0 
\end{cases}$$

with $A = [a_1, \ldots, a_p] \in \mathbb{R}^{d \times n}$, $b \in \mathbb{R}^n$, and $\|A\| = \sqrt{L - \mu}$

$f$ is

- $L$-smooth
- $\mu$-strongly convex
- infinitely differentiable (of class $C^\infty$)
Simulations

Parameters: $\mu = 1$, $L = 10^4$, $d = 100$, $n = 5$, $r = 10^{-6}$
Robustness to $\mu$

**Parameters:** $\mu = 1$, $L = 10^4$, $d = 100$, $n = 5$, $r = 10^{-6}$

![Graph showing iteration versus $\|x_k - x^*\|$ for different $\mu$ values.](image)
To prove the bound for TMM, use interpolation.

**Interpolation:** The set \( \{ y_k, f_k, g_k \} \) is \( \mathcal{F} \)-interpolable if and only if 
\[
f_k = f(y_k) \quad \text{and} \quad g_k = \nabla f(y_k)
\]
for some \( f \in \mathcal{F} \) and all \( k \).

\[
\begin{array}{ccc}
g_k & \nabla f & y_k \\
f_k & \uparrow & \\
\end{array}
\]

**Theorem (Taylor, Hendrickx, Glineur, 2017)**

The set \( \{ y_k, f_k, g_k \} \) is interpolable by an \( L \)-smooth \( \mu \)-strongly convex function if and only if \( \phi_{ij} \geq 0 \) for all \( i, j \) where
\[
\phi_{ij} := (L - \mu)(f_i - f_j) - \frac{1}{2}\|g_i - g_j\|^2 \\
+ (\mu g_i - L g_j)^T(y_i - y_j) - \frac{\mu L}{2}\|y_i - y_j\|^2
\]
Sketch of proof for TMM

1. Suppose $f$ is $L$-smooth and $\mu$-strongly convex. Then the interpolation conditions are satisfied, i.e., $\phi_{i,j} \geq 0$ for all $i, j$.

2. Define the Lyapunov function

$$V_k := \mu L \|z_k - x_*\|^2 + \phi_{k-1,*}$$

where $z_k := (1 + \delta)x_k - \delta x_{k-1}$ and $\delta := \frac{\rho^2}{1 - \rho^2}$.

3. Using the definition of TMM, it is straightforward to verify that

$$V_{k+1} - \rho^2 V_k + (1 - \rho^2)\phi_{*,k} + \rho^2 \phi_{k-1,k} = 0$$

for all $k \geq 1$, so $V_k$ decreases by at least $\rho^2$ at each iteration.

4. Iterating gives the bound $V_k \leq \rho^2(k-1)V_1$ for $k \geq 1$. 
Gradient noise

What if the measured gradient is *not* the actual gradient?

\[ x_{k+1} = (1 + \beta)x_k - \beta x_{k-1} - \alpha u_k \]
\[ y_k = (1 + \gamma)x_k - \gamma x_{k-1} \]

No noise: \( u = \nabla f(y) \)

Relative gradient noise: \( \|u - \nabla f(y)\|_2 \leq \delta \|\nabla f(y)\|_2 \)
Robust momentum method

\[ x_{k+1} = (1 + \beta)x_k - \beta x_{k-1} - \alpha \nabla f((1 + \gamma)x_k - \gamma x_{k-1}) \]

Parameters:

- \( \rho \in \left[ 1 - \frac{1}{\sqrt{\kappa}}, 1 - \frac{1}{\kappa} \right] \)
- \( \alpha = \frac{\kappa(1-\rho)^2(1+\rho)}{L} \)
- \( \beta = \frac{\kappa \rho^3}{\kappa - 1} \)
- \( \gamma = \frac{\rho^3}{(\kappa - 1)(1-\rho)^2(1+\rho)} \)

Fast

TMM

GM (\( \alpha = \frac{1}{L} \))

Slow

Fragile

Robustness


Suppose \( f \) is \( L \)-smooth and \( \mu \)-strongly convex with minimizer \( x_\star \in \mathbb{R}^d \), and there is no gradient noise (i.e., \( \delta = 0 \)). Then for any initial conditions \( x_0, x_{-1} \in \mathbb{R}^d \), there exists a constant \( c > 0 \) such that

\[ \| x_k - x_\star \| \leq c \rho^k \quad \text{for all} \quad k \geq 1. \]
Sketch of proof for RMM

1. Suppose \( f \) is \( L \)-smooth and \( \mu \)-strongly convex. Then the interpolation conditions are satisfied, i.e., \( \phi_{ij} \geq 0 \) for all \( i, j \).

2. Define the Lyapunov function

\[
V_k := \mu L \| z_k - x^* \|^2 + \phi_{k-1,*}
\]

where \( z_k := (1 + \delta)x_k - \delta x_{k-1} \) and \( \delta := \frac{\rho^2}{1 - \rho^2} \).

3. Using the definition of RMM, it is straightforward to verify that

\[
V_{k+1} - \rho^2 V_k + (1 - \rho^2)\phi_{*,k} + \rho^2 \phi_{k-1,k} + \frac{(1 + \rho)(1 - \kappa + 2\kappa \rho - \kappa \rho^2)}{2\rho} \| \nabla f(y_k) - \mu (y_k - y^*) \|^2 = 0
\]

for all \( k \geq 1 \), so \( V_k \) decreases by at least \( \rho^2 \) at each iteration.

4. Iterating gives the bound \( V_k \leq \rho^2(k-1)V_1 \) for \( k \geq 1 \).
Trade-off: Speed vs. Robustness

\[\kappa = 10\]

- Noise strength ($\delta$)
- Convergence rate ($\rho$)
- GM ($\alpha = \frac{1}{L}$)
- GM ($\alpha = \frac{2}{L+\mu}$)
- GM (min $\alpha \in [0, \frac{2}{L}]$)
- FGM
- TMM
- RMM ($\rho = 1 - \frac{1}{\sqrt{\kappa}}$)
- RMM ($\rho = \frac{\kappa - 1}{\kappa + 1}$)
- RMM ($\rho = 1 - \frac{1}{\kappa}$)
- RMM (min $\rho \in [1 - \frac{1}{\kappa}, 1 - \frac{1}{\sqrt{\kappa}}]$)
For TMM, we can analyze the convergence rate in closed-form.

What can we say when a closed-form expression for the convergence rate is unknown (e.g., when there is gradient noise)?

Calculate an upper bound on the convergence rate numerically using:

- Integral Quadratic Constraints
  - Megretzki, Rantzer, 1997
  - Lessard, Recht, Packard, 2016

- Performance Estimation Problem
  - Drori, Teboulle, 2014
  - Taylor, Hendrickx, Glineur, 2017

- Quadratic Lyapunov functions
  - Taylor, Van Scoy, Lessard, 2018 (ICML)
Conclusion

Triple momentum method

- Iterates converge linearly with rate $\rho = 1 - 1/\sqrt{\kappa}$
- This is the fastest known convergence rate for first-order methods on smooth strongly convex functions (twice as fast as FGM)

Robust momentum method

- Interpolates TMM and GM (with stepsize $\frac{1}{L}$) to exploit the trade-off between convergence rate and robustness to gradient noise

TMM

$\rho = 1 - \frac{1}{\sqrt{\kappa}}$

fast, fragile

GM ($\alpha = \frac{1}{L}$)

$\rho = 1 - \frac{1}{\kappa}$

slow, robust
Collaborators

Laurent Lessard  Saman Cyrus  Bin Hu
Randy Freeman  Kevin Lynch  Adrien Taylor

Papers

- Taylor, Van Scoy, Lessard, *ICML*, 2018
- Available on my website: vanscoy.github.io