The fastest known globally convergent first-order method for minimizing strongly convex functions

Bryan Van Scoy

University of Wisconsin–Madison

Dec 12, 2017
Unconstrained optimization:

\[
\begin{align*}
&\text{minimize} & f(x) \\
&\text{subject to} & x \in \mathbb{R}^d
\end{align*}
\]

• Need methods which are fast and simple
• Use first-order methods
Unconstrained optimization:

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in \mathbb{R}^d
\end{align*}
\]

- Need methods which are \textit{fast} and \textit{simple}
- Use \textit{first-order} methods
- In this talk, we will design a first-order method for the case when \( f \) is smooth and strongly convex
Unconstrained optimization:

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in \mathbb{R}^d
\end{align*}
\]

- Need methods which are *fast* and *simple*
- Use *first-order* methods
- In this talk, we will design a first-order method for the case when \( f \) is smooth and strongly convex

**Main result**

Design and analyze a novel method which is both globally convergent and faster than Nesterov’s method

**Analysis**  Simple convergence proof (time domain)

**Design**  Intuition using IQCs (frequency domain)
Smooth strongly convex

A differentiable function \( f : \mathbb{R}^d \to \mathbb{R} \) is called \( L \)-smooth if

\[
\| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \| \quad \text{for all } x, y \in \mathbb{R}^d
\]

and \( m \)-strongly convex if

\[
f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \| x - y \|^2 \quad \text{for all } x, y \in \mathbb{R}^d.
\]
Smooth strongly convex

A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is called $L$-smooth if

$$
\| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \| \quad \text{for all } x, y \in \mathbb{R}^d
$$

and $m$-strongly convex if

$$
f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \| x - y \|^2 \quad \text{for all } x, y \in \mathbb{R}^d.
$$
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

\[ x_{k+1} = (1 + \beta)x_k - \beta x_{k-1} - \alpha \nabla f((1 + \beta)x_k - \beta x_{k-1}) \]
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>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM</td>
<td>(\alpha, 0, 0)</td>
</tr>
<tr>
<td>HBM (Polyak, 1964)</td>
<td>(\alpha, \beta, 0)</td>
</tr>
<tr>
<td>FGM (Nesterov, 2004)</td>
<td>(\alpha, \beta, \beta)</td>
</tr>
<tr>
<td>TMM (Van Scoy, Freeman, Lynch, 2017)</td>
<td>(\alpha, \beta, \gamma)</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 \( \kappa := \frac{L}{m} \)
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:
\[
\begin{align*}
\rho &= 1 - \frac{1}{\sqrt{\kappa}} \\
\alpha &= \frac{1 + \rho}{L} \\
\beta &= \frac{\rho^2}{2 - \rho} \\
\gamma &= \frac{\rho^2}{(1 + \rho)(2 - \rho)}
\end{align*}
\]

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

Theorem (Van Scoy, Freeman, Lynch, 2017)
Suppose \( f \) is \( L \)-smooth and \( m \)-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: \( \| x_k - x_\star \| \leq c \rho^k \)

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

- **HBM** does not converge if \( L/m \geq (2 + \sqrt{5})^2 \approx 17.94 \)
- For **FGM**, Nesterov proved the rate \( \sqrt{1 - \sqrt{m/L}} \) which is loose!
- **TMM** converges faster than Nesterov’s method!
Simulations

Objective function:

\[ f(x) = \sum_{i=1}^{p} g(a_i^T x - b_i) + \frac{m}{2} \|x\|^2, \quad x \in \mathbb{R}^d \]

where

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

with \( A = [a_1, \ldots, a_p] \in \mathbb{R}^{d \times p}, b \in \mathbb{R}^p, \) and \( \|A\| = \sqrt{L - m} \)

\( f \) is

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

Parameters: \( m = 1, \ L = 10^4, \ d = 100, \ p = 5, \ r = 10^{-6} \)
Robustness to $m$

**Parameters:** $m = 1$, $L = 10^4$, $d = 100$, $p = 5$, $r = 10^{-6}$

---

**Parameter used to tune TMM**
- $m = 10$
- $m = 5$
- $m = 2$
- $m = 1$
To prove the bound for TMM, use interpolation.
To prove the bound for TMM, use interpolation.

**Interpolation:** The set \( \{y, u, v\} \) is \( \mathcal{F} \)-interpolable if and only if 
\[ u_k = \nabla f(y_k) \quad \text{and} \quad v_k = f(y_k) \]
for some \( f \in \mathcal{F} \) and all \( k \).
To prove the bound for TMM, use interpolation.

**Interpolation:** The set \( \{y, u, v\} \) is \( \mathcal{F} \)-interpolable if and only if \( u_k = \nabla f(y_k) \) and \( v_k = f(y_k) \) for some \( f \in \mathcal{F} \) and all \( k \).

\[
\begin{align*}
    &\nabla f &\leftarrow u &\rightarrow y \\
    &f &\leftarrow v &\rightarrow u
\end{align*}
\]

**Theorem (Taylor, Hendrickx, Glineur, 2016)**

The set \( \{y, u, v\} \) is interpolable by an \( L \)-smooth \( m \)-strongly convex function if and only if \( q_{ij} \geq 0 \) for all \( i, j \) where

\[
q_{ij} := (L - m)(v_i - v_j) - \frac{1}{2}\|u_i - u_j\|^2 + (mu_i - Lu_j)^T(y_i - y_j) - \frac{mL}{2}\|y_i - y_j\|^2.
\]
Sketch of proof for TMM

1. Suppose $f$ is $L$-smooth and $m$-strongly convex. Then the interpolation conditions are satisfied; specifically, $q_{ij} \geq 0$ for all $i, j$. 
Sketch of proof for TMM

1. Suppose $f$ is $L$-smooth and $m$-strongly convex. Then the interpolation conditions are satisfied; specifically, $q_{ij} \geq 0$ for all $i,j$.

2. Define the **Lyapunov function**

$$V_k := mL \|z_k - x_*\|^2 + q_{k-1,*}$$

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

1. Suppose $f$ is $L$-smooth and $m$-strongly convex. Then the interpolation conditions are satisfied; specifically, $q_{ij} \geq 0$ for all $i, j$.

2. Define the Lyapunov function

$$V_k := mL \|z_k - x_*\|^2 + q_{k-1,*}$$

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

$$V_{k+1} - \rho^2 V_k = -[(1 - \rho^2)q_{*,k} + \rho^2 q_{k-1,k}] \leq 0$$

for all $k \geq 1$. 

Sketch of proof for TMM

1. Suppose $f$ is $L$-smooth and $m$-strongly convex. Then the **interpolation conditions** are satisfied; specifically, $q_{ij} \geq 0$ for all $i, j$.

2. Define the **Lyapunov function**

$$V_k := mL \|z_k - x_*\|^2 + q_{k-1,*}$$

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 straighforward to verify that

$$V_{k+1} - \rho^2 V_k = -[(1 - \rho^2)q_{*,k} + \rho^2 q_{k-1,k}] \leq 0$$

for all $k \geq 1$.

4. Iterating gives the **bound** $V_k \leq \rho^2(k-1)V_1$ for $k \geq 1$. 
Integral Quadratic Constraints (IQC)s

\[
G : \begin{align*}
x_{k+1} &= (1 + \beta)x_k - \beta x_{k-1} - \alpha u_k \\
y_k &= (1 + \gamma)x_k - \gamma x_{k-1}
\end{align*}
\]

Suppose \( f \) satisfies the IQC defined by \((\Pi, M)\). If there exists \( \varepsilon > 0 \) with
\[
G(z) \Psi(z) \ast M \Psi(z) \ast G(z) \preceq -\varepsilon I
\]
for all \( z \in \rho \), then the state of \( G \) converges linearly with rate \( \rho \).

The TMM parameters are the unique solution to
\[
G(z) \ast \Psi(z) \ast M \ast \Psi(z) \ast G(z) = 0
\]
for all \( z \in \rho \).
Integral Quadratic Constraints (IQCṣ)

\[(\Psi, M) \text{ are chosen such that } w \text{ satisfies} \]
\[0 \leq \sum_{j=0}^{k} \rho^{-2j} (w_j - w_\star)^T M (w_j - w_\star)\]
when \(f\) is \(L\)-smooth and \(m\)-strongly convex.
Integral Quadratic Constraints (IQC)

$$(\Psi, M)$$ are chosen such that $w$ satisfies

$$0 \leq \sum_{j=0}^{k} \rho^{-2j} (w_j - w_\star)^T M (w_j - w_\star)$$

when $f$ is $L$-smooth and $m$-strongly convex.

**Theorem (Boczar, Lessard, Recht, 2015)**

Suppose $f$ satisfies the IQC defined by $(\Pi, M)$. If there exists $\varepsilon > 0$ with

$$\begin{bmatrix} G(z) \\ I \end{bmatrix}^* \Psi(z)^* M \Psi(z) \begin{bmatrix} G(z) \\ I \end{bmatrix} \preceq -\varepsilon I \quad \text{for all } z \in \rho \mathbb{T}$$

then the state of $G$ converges linearly with rate $\rho$. 
Integral Quadratic Constraints (IQC)

\[ G \] \[ \nabla f \] \[ \Psi \] \[ w \]
\[ u \] \[ y \]

\((\Psi, M)\) are chosen such that \(w\) satisfies

\[
0 \leq \sum_{j=0}^{k} \rho^{-2j} (w_j - w_\star)^T M (w_j - w_\star)
\]

when \(f\) is \(L\)-smooth and \(m\)-strongly convex.

**Theorem (Boczar, Lessard, Recht, 2015)**

Suppose \(f\) satisfies the IQC defined by \((\Pi, M)\). If there exists \(\varepsilon > 0\) with

\[
\begin{bmatrix}
  G(z) \\
  I
\end{bmatrix}^* \Psi(z)^* M \Psi(z) \begin{bmatrix}
  G(z) \\
  I
\end{bmatrix} \preceq -\varepsilon I \quad \text{for all} \quad z \in \rho \mathbb{T}
\]

then the state of \(G\) converges linearly with rate \(\rho\).

The TMM parameters are the unique solution to

\[
\begin{bmatrix}
  G(z) \\
  I
\end{bmatrix}^* \Psi(z)^* M \Psi(z) \begin{bmatrix}
  G(z) \\
  I
\end{bmatrix} = 0 \quad \text{for all} \quad z \in \rho \mathbb{T}
\]
Summary

**Triple momentum method**: globally convergent with rate

\[1 - \sqrt{m/L}\] when \(f\) is \(L\)-smooth and \(m\)-strongly convex

**Analysis** Simple convergence proof (time domain)

**Design** Intuition using IQCs (frequency domain)
**Summary**

**Triple momentum method:** globally convergent with rate $1 - \sqrt{m/L}$ when $f$ is $L$-smooth and $m$-strongly convex

**Analysis**  Simple convergence proof (time domain)

**Design**  Intuition using IQCs (frequency domain)

**Extension: gradient noise**

$$x_{k+1} = (1 + \beta)x_k - \beta x_{k-1} - \alpha u_k$$

$$y_k = (1 + \gamma)x_k - \gamma x_{k-1}$$

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

**Relative gradient noise:** $\|u - \nabla f(y)\|_2 \leq \delta \|\nabla f(y)\|_2$

---