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:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathbb{R}^d \end{array}$$

- Need methods which are *fast* and *simple*
- Use *first-order* methods

Unconstrained optimization:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathbb{R}^d \end{array}$$

- 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

Unconstrained optimization:

$$\begin{array}{rl} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathbb{R}^d \end{array}$$

- 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

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

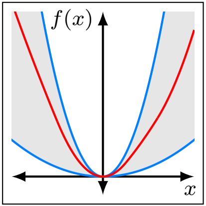
Smooth strongly convex

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

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

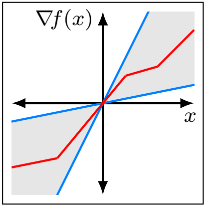
and m-strongly convex if

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



L-smooth m-strongly convex

$\iff$



slope restricted on $[m, L]$

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

$$x_{k+1} = (1+\beta)x_k - \beta x_{k-1} - \alpha \nabla f\big((1+\beta)x_k - \beta x_{k-1}\big)$$

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

$$x_{k+1} = (1+\beta)x_k - \beta x_{k-1} - \alpha \nabla f\big((1+\beta)x_k - \beta x_{k-1}\big)$$

triple momentum method

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

Method	Parameters
GM	$(\alpha, 0, 0)$
HBM (Polyak, 1964)	$(\alpha, \beta, 0)$
FGM (Nesterov, 2004)	(α, β, β)
TMM (Van Scoy, Freeman, Lynch, 2017)	(α, β, γ)

Triple momentum method

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

Parameters:

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

$$\alpha = \tfrac{1+\rho}{L}$$

$$\beta = \tfrac{\rho^2}{2-\rho}$$

$$\gamma = \tfrac{\rho^2}{(1+\rho)(2-\rho)}$$

Condition ratio $\kappa := L/m$

Triple momentum method

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

Parameters:

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

$\alpha = \frac{1+\rho}{L}$

$\beta = \frac{\rho^2}{2-\rho}$

$\gamma = \frac{\rho^2}{(1+\rho)(2-\rho)}$

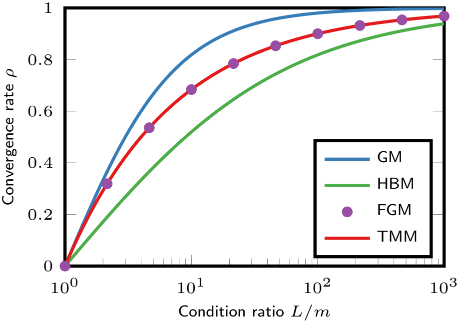
Condition ratio $\kappa := L/m$

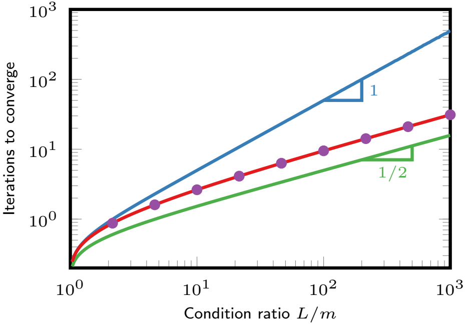
Theorem (Van Scoy, Freeman, Lynch, 2017)

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

$$\|x_k - x_\star\| \le c\,\rho^k \quad \text{for all } k \ge 1.$$

f quadratic

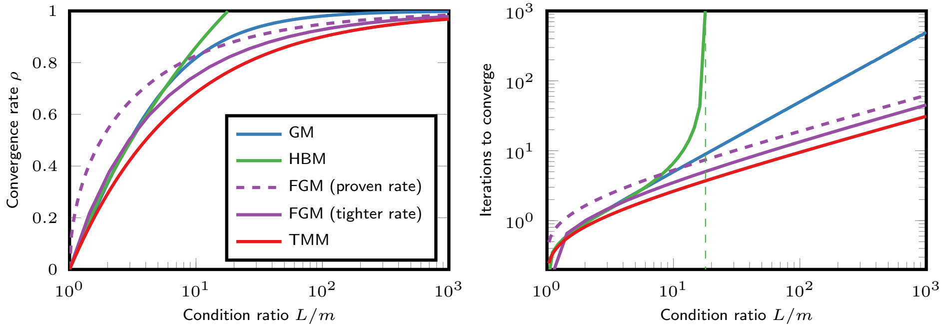




Convergence rate: $\|x_k - x_\star\| \le c\,\rho^k$

Iterations to converge $\propto -\dfrac{1}{\log \rho}$

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

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

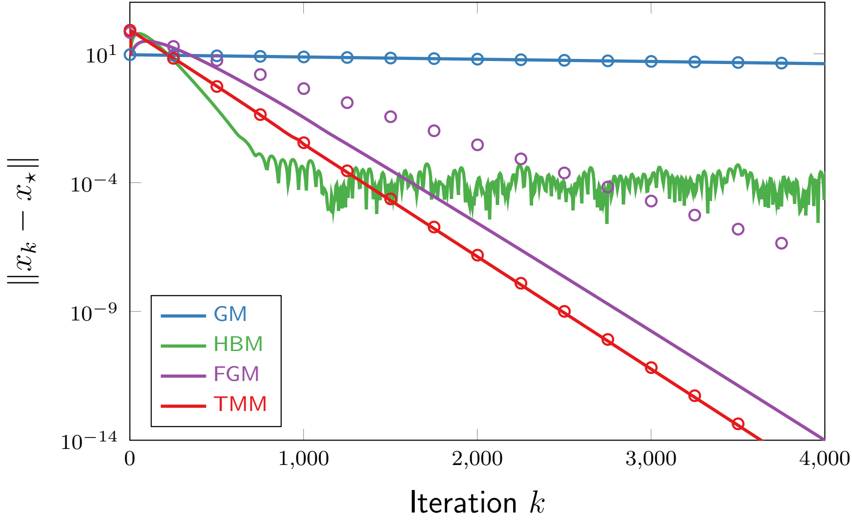
with $A = [a_1, \ldots, a_p] \in \mathbb{R}^{d \times p}$, $b \in \mathbb{R}^p$, and $\|A\| = \sqrt{L - m}$

f is

- m-smooth
- L-strongly convex
- infinitely differentiable (of class C^∞)

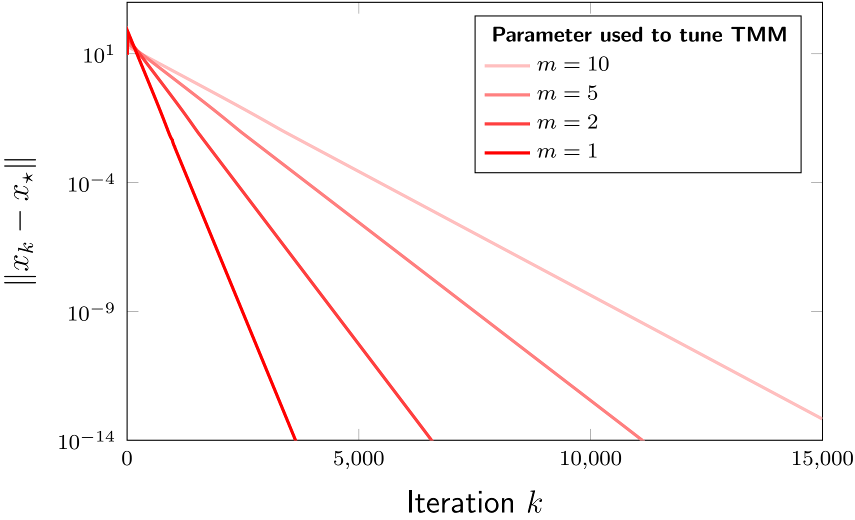
Simulations

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



Robustness to m

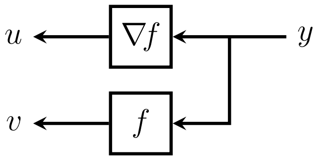
Parameters: $m = 1$, $L = 10^4$, $d = 100$, $p = 5$, $r = 10^{-6}$



To prove the bound for TMM, use *interpolation*.

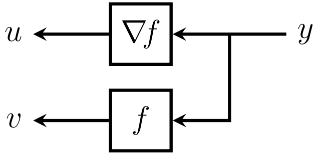
To prove the bound for TMM, use *interpolation*.

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



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

$$\begin{aligned} q_{ij} := (L-m)(v_i - v_j) - \frac{1}{2}\|u_i - u_j\|^2 \\ + (mu_i - Lu_j)^\mathsf{T}(y_i - y_j) - \frac{mL}{2}\|y_i - y_j\|^2. \end{aligned}$$

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**

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

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**

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

   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

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

   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**

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

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

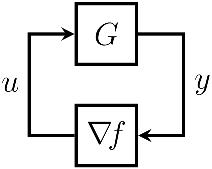
3. Using the definition of TMM, it is straighforward to verify that

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

   for all $k \geq 1$.

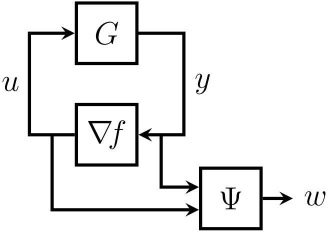
4. Iterating gives the **bound** $V_k \leq \rho^{2(k-1)} V_1$ for $k \geq 1$.

Integral Quadratic Constraints (IQCs)



$$G: \quad \begin{aligned} x_{k+1} &= (1+\beta)x_k - \beta x_{k-1} - \alpha u_k \\ y_k &= (1+\gamma)x_k - \gamma x_{k-1} \end{aligned}$$

Integral Quadratic Constraints (IQCs)

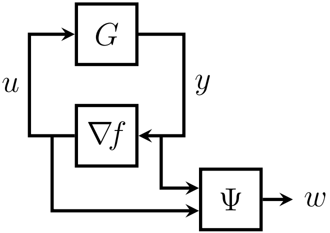


(Ψ, M) are chosen such that w satisfies

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

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

Integral Quadratic Constraints (IQCs)



(Ψ, M) are chosen such that w satisfies

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

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

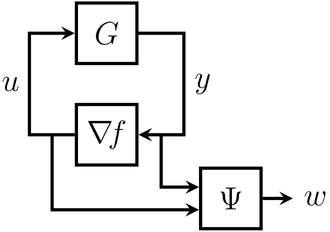
Theorem (Boczar, Lessard, Recht, 2015)

Define $\Pi(z) := \Psi(z)^* M \Psi(z)$. If there exists $\varepsilon > 0$ such that

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

then the state of G converges linearly with rate ρ.

Integral Quadratic Constraints (IQCs)



(Ψ, M) are chosen such that w satisfies

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

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

Theorem (Boczar, Lessard, Recht, 2015)

Define $\Pi(z) := \Psi(z)^* M \Psi(z)$. If there exists $\varepsilon > 0$ such that

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

then the state of G converges linearly with rate ρ.

The TMM parameters are the unique solution to

$$\begin{bmatrix} G(z) \\ I \end{bmatrix}^* \Pi(z) \begin{bmatrix} G(z) \\ I \end{bmatrix} = 0 \quad \text{for all } z \in \rho\mathbb{T}$$

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: $\|u - \nabla f(y)\|_2 \leq \delta \|\nabla f(y)\|_2$

S. Cyrus, B. Hu, B. Van Scoy, L. Lessard. "A Robust Accelerated Optimization Algorithm for Strongly Convex Functions". In ArXiv e-prints (Oct. 2017). arXiv: 170.04753 [math.OC].

Thanks!

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

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

Parameters:

$\rho \in \left[1 - \frac{1}{\sqrt{\kappa}}, 1 - \frac{1}{\kappa}\right]$

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

$\beta = \frac{\kappa\rho^3}{\kappa-1}$

$\gamma = \frac{\rho^3}{(\kappa-1)(1-\rho)^2(1+\rho)}$

Theorem (Cyrus, Hu, Van Scoy, Lessard, 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

$$\|x_k - x_\star\| \le c\,\rho^k \quad \text{for all } k \ge 1.$$

Robust momentum method

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

Parameters:

$$\rho \in \left[1 - \tfrac{1}{\sqrt{\kappa}}, 1 - \tfrac{1}{\kappa}\right]$$

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

$$\beta = \frac{\kappa\rho^3}{\kappa-1}$$

$$\gamma = \frac{\rho^3}{(\kappa-1)(1-\rho)^2(1+\rho)}$$

$$\text{Fast} \xleftarrow{\quad\text{Speed}\quad} \text{Slow}$$

$$1 - \tfrac{1}{\sqrt{\kappa}} \quad \overset{\rho}{\text{———}} \quad 1 - \tfrac{1}{\kappa}$$

$$\text{Fragile} \xrightarrow{\quad\text{Robustness}\quad} \text{Robust}$$

Theorem (Cyrus, Hu, Van Scoy, Lessard, 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

$$\|x_k - x_\star\| \le c\,\rho^k \quad \text{for all } k \ge 1.$$

Robust momentum method

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

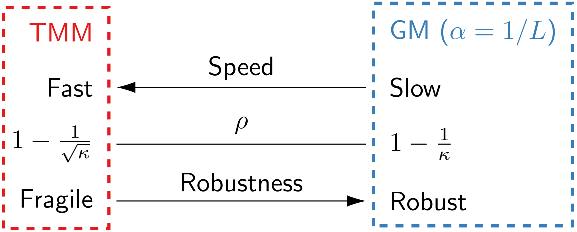
Parameters:

$\rho \in \left[1 - \frac{1}{\sqrt{\kappa}}, 1 - \frac{1}{\kappa}\right]$

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

$\beta = \frac{\kappa \rho^3}{\kappa - 1}$

$\gamma = \frac{\rho^3}{(\kappa-1)(1-\rho)^2(1+\rho)}$



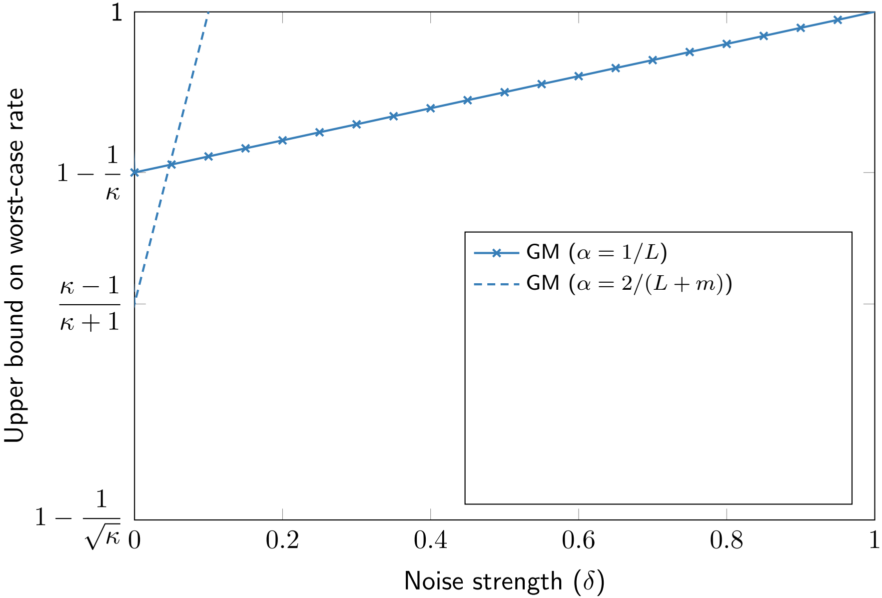
Theorem (Cyrus, Hu, Van Scoy, Lessard, 2017)

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

$$\|x_k - x_\star\| \le c\,\rho^k \quad \text{for all } k \ge 1.$$

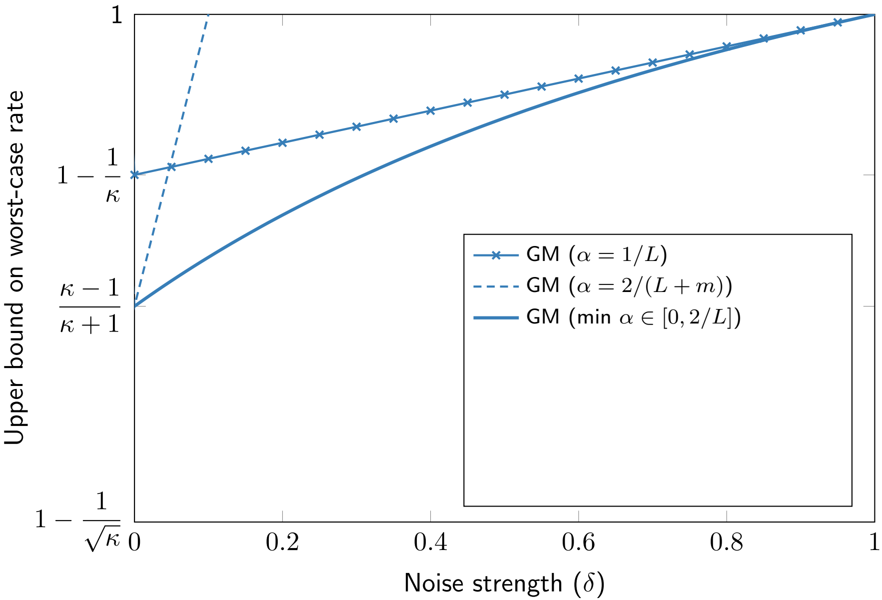
Trade-off: Speed vs. Robustness

$(\kappa = 10)$



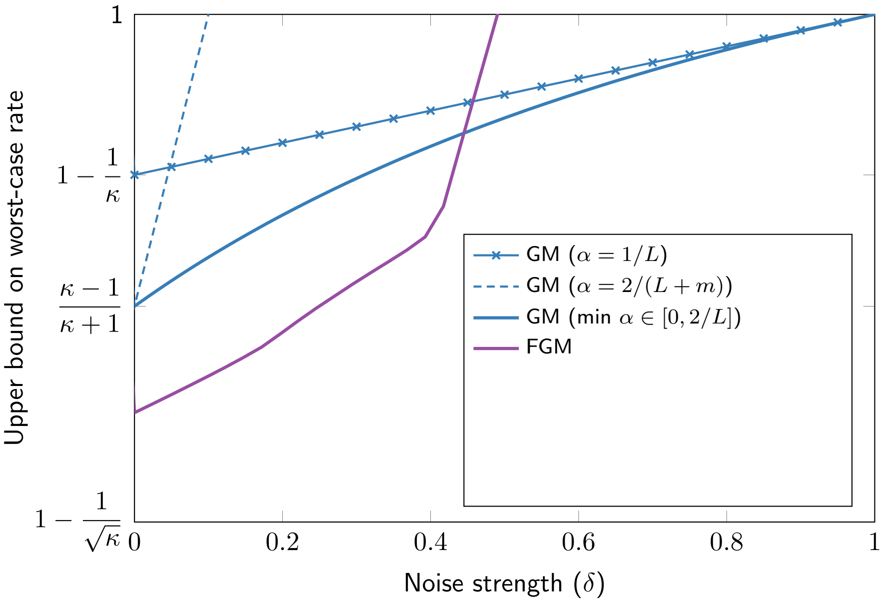
Trade-off: Speed vs. Robustness

$(\kappa = 10)$



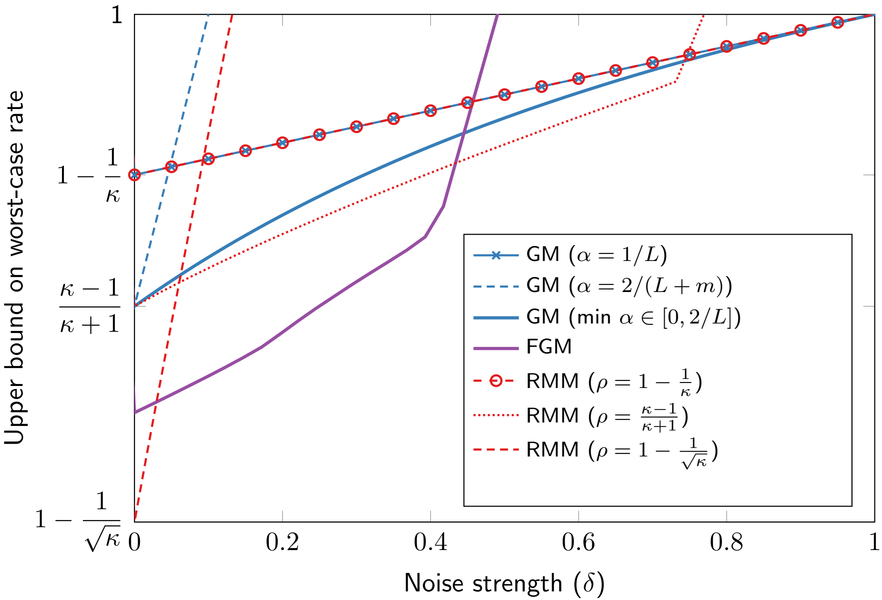
Trade-off: Speed vs. Robustness

$(\kappa = 10)$

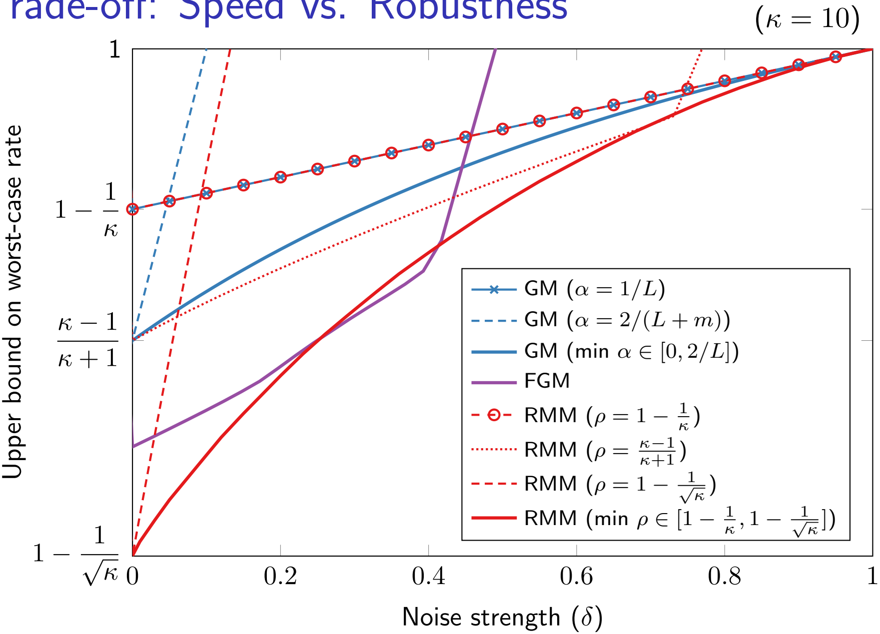


Trade-off: Speed vs. Robustness

$(\kappa = 10)$



Trade-off: Speed vs. Robustness



Conclusion

Analysis

- **Numerical:** solve SDP to calculate upper bound on convergence rate
- **Closed-form:** have expressions for convergence rate for some methods and functions classes (such as TMM on smooth strongly convex functions)

Design

- **Triple momentum method** - Fastest known convergence rate for first-order methods on smooth strongly convex functions
- **Robust momentum method** - Interpolates TMM and GM (with $\alpha = 1/L$) to exploit the trade-off between convergence rate and robustness to gradient noise