First-Order Optimization Methods

Analysis and design

Bryan Van Scoy

University of Wisconsin–Madison

Nov 1, 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

Function class

- quadratic
- smooth strongly convex

Method

GM gradient method

HBM heavy ball method

FGM fast gradient method

Bound

- $f(x_k) - f(x_\star) \le c_1\, \rho^k$
- $\|x_k - x_\star\| \le c_2\, \rho^k$
- $\|\nabla f(x_k)\| \le c_3\, \rho^k$

function class $+$ method $\Longrightarrow$ bound

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:

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

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

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

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

Condition ratio $\kappa := L/m$

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:

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

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

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

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

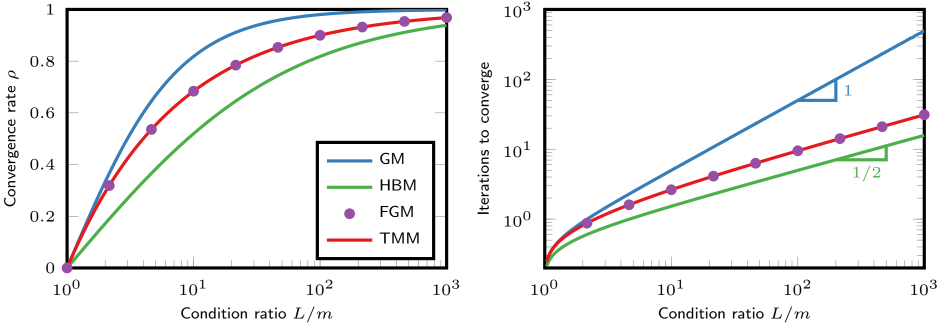
Condition ratio $\kappa := L/m$

Theorem (Van Scoy, Freeman, Lynch, 2017)

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

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

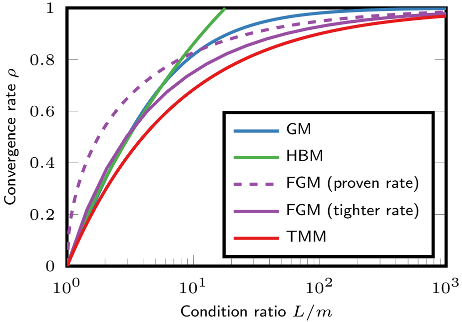
f quadratic

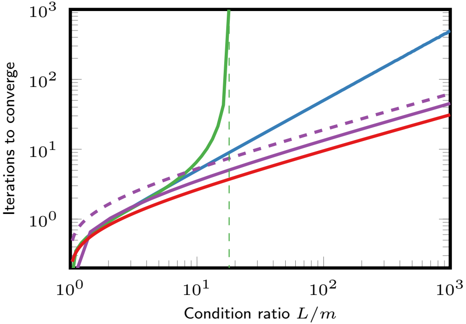


Convergence rate: $\|x_k - x_\star\| \le c\,\rho^k$

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:

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

where

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

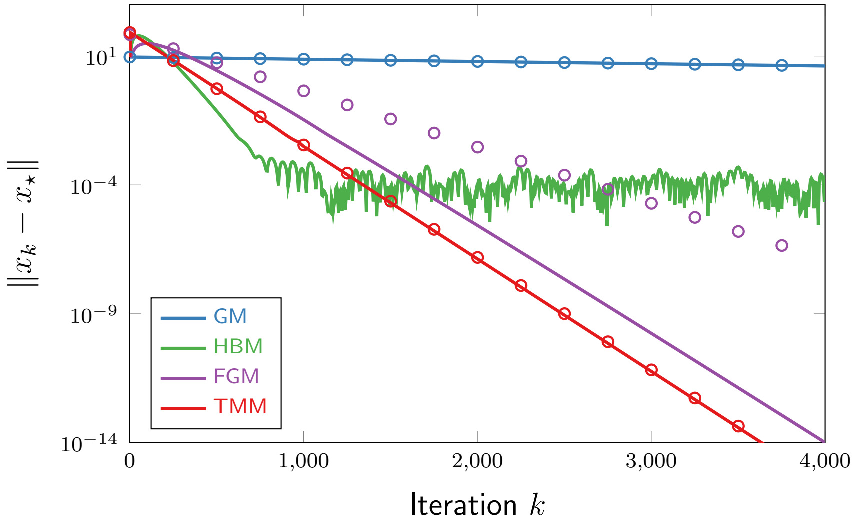
with $A = [a_1, \ldots, a_p] \in \mathbb{R}^{d \times p}$, $b \in \mathbb{R}^p$, and $\|A\| = \sqrt{L - m}$

f is

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

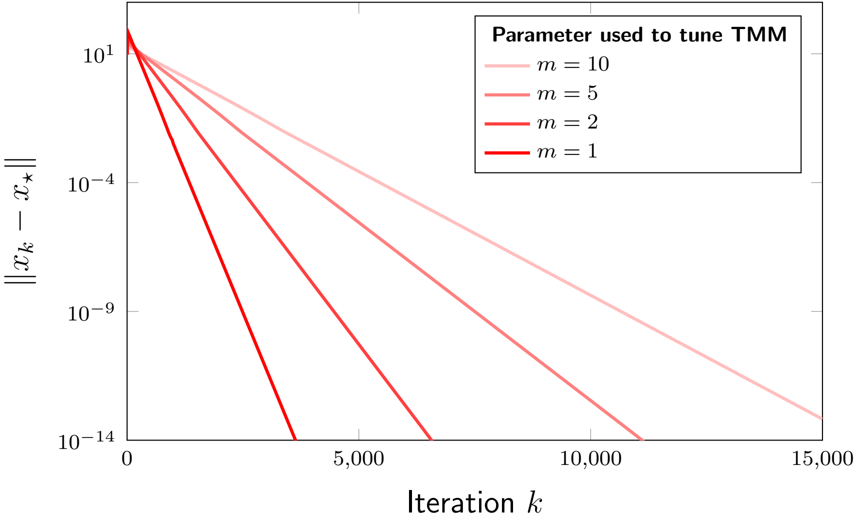
Simulations

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



Robustness to m

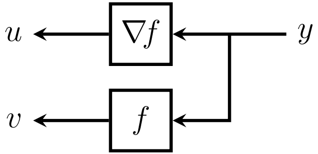
Parameters: $m = 1$, $L = 10^4$, $d = 100$, $p = 5$, $r = 10^{-6}$



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

To prove the bound for TMM, using *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, using *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.

$$
\begin{array}{ccccc}
u & \longleftarrow & \boxed{\nabla f} & \longleftarrow & y \\
v & \longleftarrow & \boxed{f} & \longleftarrow & y
\end{array}
$$

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**

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

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 $k \geq 1$.

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 $k \geq 1$.

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

Numerics

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?

Numerics

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?

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, 2016

Gradient noise

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

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

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

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

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 $x_\star$. Then for any initial conditions $x_0, x_{-1} \in \mathbb{R}^n$, 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:

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

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

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

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

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

$$1 - \frac{1}{\sqrt{\kappa}} \quad \overset{\rho}{\overline{\qquad\qquad\qquad}} \quad 1 - \frac{1}{\kappa}$$

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

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

Suppose f is L-smooth and m-strongly convex with minimizer $x_\star$. Then for any initial conditions $x_0, x_{-1} \in \mathbb{R}^n$, 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:

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

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

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

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

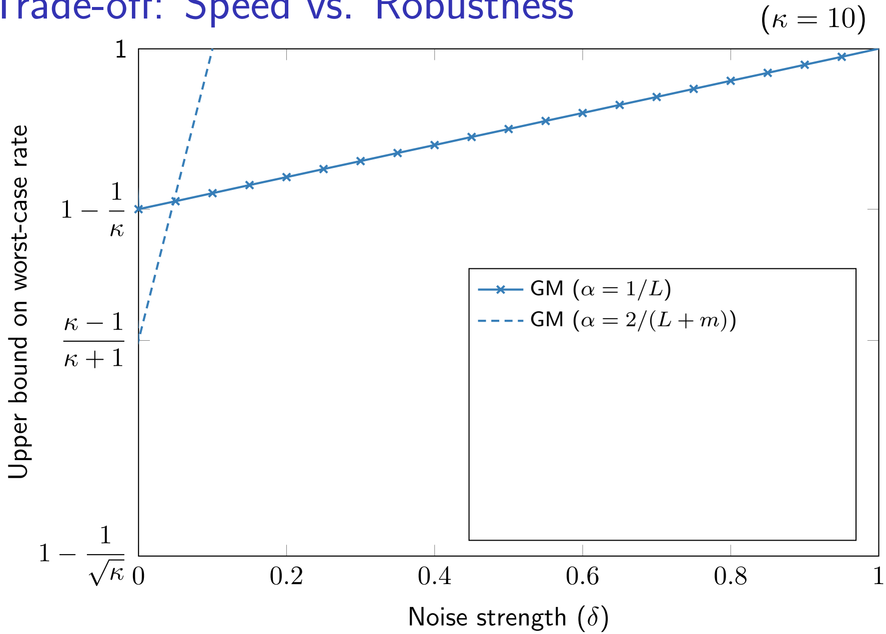
TMM		GM ($\alpha = 1/L$)
Fast	$\xleftarrow{\text{Speed}}$	Slow
$1 - \frac{1}{\sqrt{\kappa}}$	ρ	$1 - \frac{1}{\kappa}$
Fragile	$\xrightarrow{\text{Robustness}}$	Robust

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

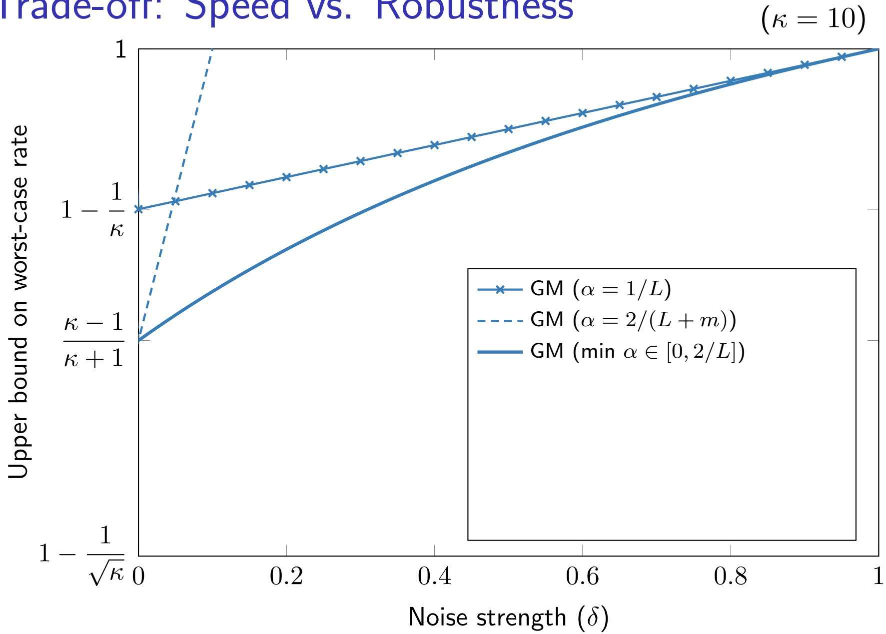
Suppose f is L-smooth and m-strongly convex with minimizer $x_\star$. Then for any initial conditions $x_0, x_{-1} \in \mathbb{R}^n$, there exists a constant $c > 0$ such that

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

Trade-off: Speed vs. Robustness

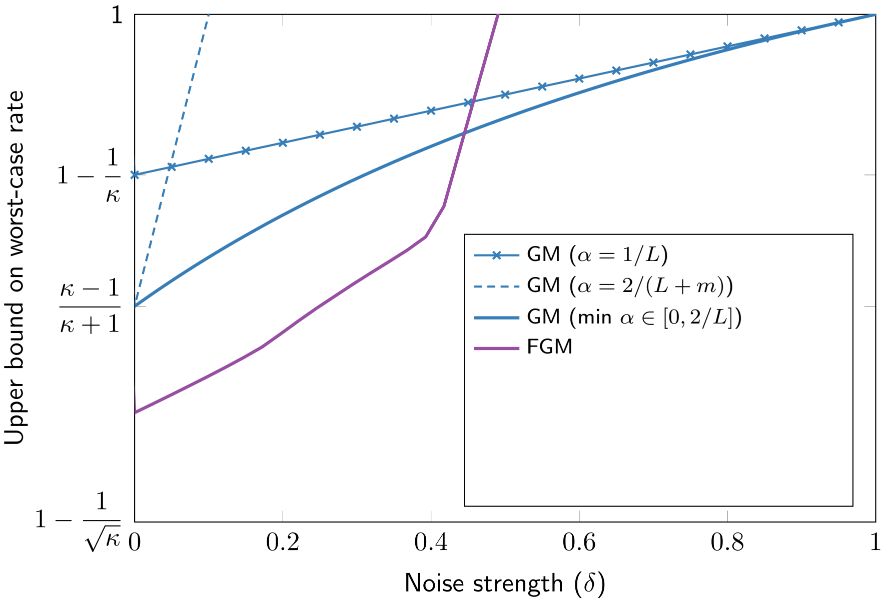


Trade-off: Speed vs. Robustness



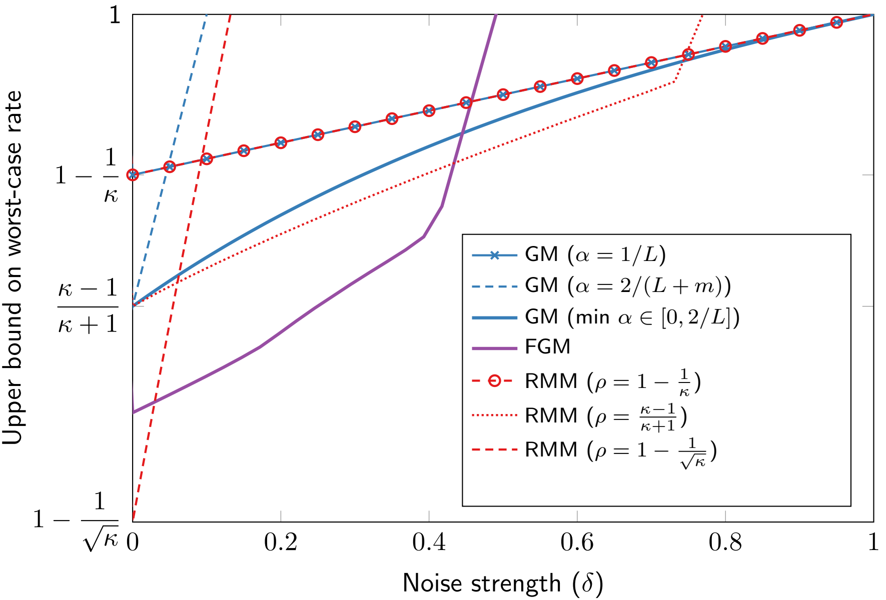
Trade-off: Speed vs. Robustness

$(\kappa = 10)$

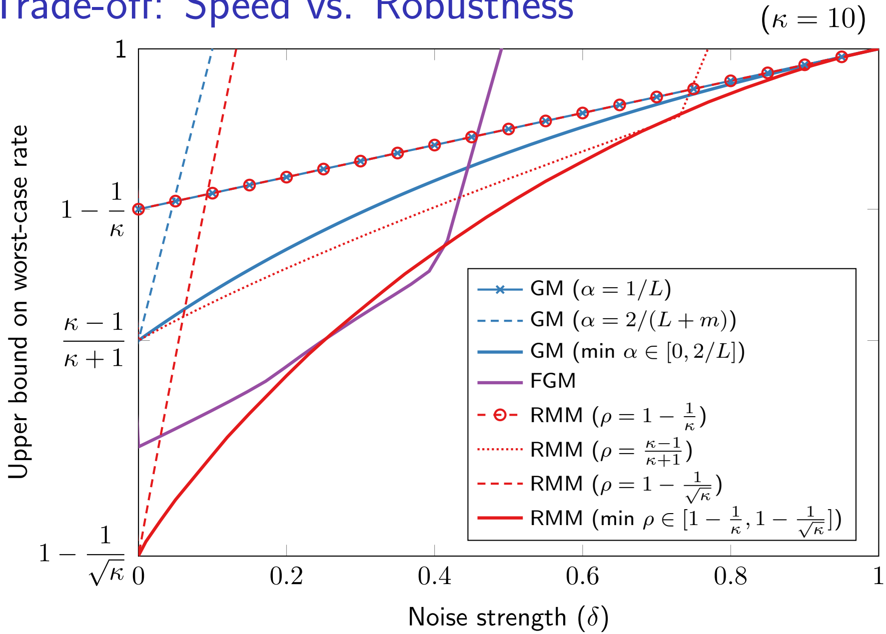


Trade-off: Speed vs. Robustness

$(\kappa = 10)$



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