Line-search methods with restarting for nonconvex optimization

Clément W. Royer

Séminaire de Statistique, LPSM

December 2, 2025

Co-authors and funding



1. Rémi Chan–Renous-Legoubin
   Dauphine→IDEMIA
2. Albert S. Berahas
   University of Michigan
3. Michael J. O'Neill
   UNC Chapel Hill.

My research: Nonconvex optimization

$$\min_{x \in \mathbb{R}^n} f(x)$$

with $f \in \mathcal{C}^1$ with Lipschitz gradient, bounded below and nonconvex.

Motivation Robust regression formulations in data analysis (see later).

My research: Nonconvex optimization

$$\min_{x \in \mathbb{R}^n} f(x)$$

with $f \in \mathcal{C}^1$ with Lipschitz gradient, bounded below and nonconvex.
Motivation Robust regression formulations in data analysis (see later).

What do we want to do?

- Finding global/local minima is NP-hard.
- Finding x such that $\|\nabla f(x)\| = 0$ is also NP-hard.
- We settle for x such that $\|\nabla f(x)\|$ is small (PLS-complete).

Problem $\text{minimize}_{x \in \mathbb{R}^n} f(x)$.

Solving the problem

- **Algorithm** Iterative process $\{x_0, x_1, \dots\}$.
- **Cost** Iterations, evaluations of f and ∇f.

Problem $\text{minimize}_{x \in \mathbb{R}^n} f(x)$.

Solving the problem

- **Algorithm** Iterative process $\{x_0, x_1, \ldots\}$.
- **Cost** Iterations, evaluations of f and ∇f.

Complexity

Given $\epsilon \in (0,1)$ and an algorithm $\{x_k\}_k$, find worst-case cost of the algorithm to reach x_k such that

$$\|\nabla f(x_k)\| \le \epsilon.$$

Gradient descent (with line search)

Algorithm Starting from $x_0 \in \mathbb{R}^n$, iterate

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \qquad \alpha_k > 0.$$

Gradient descent (with line search)

Algorithm Starting from $x_0 \in \mathbb{R}^n$, iterate

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \qquad \alpha_k > 0.$$

Armijo backtracking line search

Set α_k as the largest value $\alpha \in \{1, \theta, \theta^2, \dots\}$ such that

$$f(x_k - \alpha \nabla f(x_k)) \leq f(x_k) - \eta \alpha \|\nabla f(x_k)\|^2$$

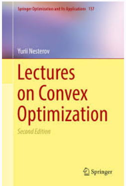
for $\theta \in (0,1)$ and $\eta \in (0,1)$.

Complexity of gradient descent

Theorem

Gradient descent finds $\|\nabla f(x_k)\| \leq \epsilon$ in at most

- $\mathcal{O}(\epsilon^{-2})$ iterations/gradient evaluations.
- $\mathcal{O}(\epsilon^{-2})$ function evaluations.



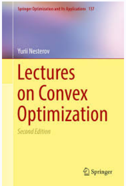
Complexity of gradient descent

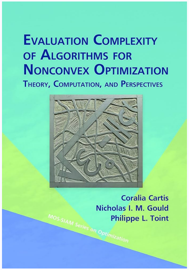
Theorem

Gradient descent finds $\|\nabla f(x_k)\| \leq \epsilon$ in at most

- $\mathcal{O}(\epsilon^{-2})$ iterations/gradient evaluations.
- $\mathcal{O}(\epsilon^{-2})$ function evaluations.

- Sharp There exist functions on which gradient descent takes $\Omega(\epsilon^{-2})$ iterations.
- Better bounds possible when f $\mathcal{C}^2$.
- Many variants developed with complexity in mind since 2009!





A puzzling example (COLT, 2017)

"Convex Until Proven Guilty": Dimension-Free Acceleration of Gradient Descent on Non-Convex Functions

Yair Carmon John C. Duchi Oliver Hinder Aaron Sidford [1]

A puzzling example (COLT, 2017)

"Convex Until Proven Guilty": Dimension-Free Acceleration of Gradient Descent on Non-Convex Functions

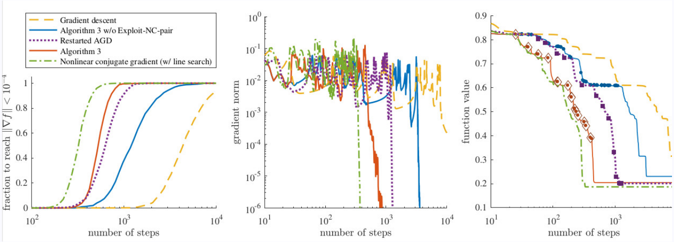
Yair Carmon John C. Duchi Oliver Hinder Aaron Sidford [1]

- New algo: $\tilde{\mathcal{O}}(\epsilon_g^{-7/4})$ complexity (better than $\mathcal{O}(\epsilon_g^{-2})$)!

"Convex Until Proven Guilty": Dimension-Free Acceleration of Gradient Descent on Non-Convex Functions

Yair Carmon John C. Duchi Oliver Hinder Aaron Sidford [1]

- New algo: $\tilde{\mathcal{O}}(\epsilon_g^{-7/4})$ complexity (better than $\mathcal{O}(\epsilon_g^{-2})$)!
- **Outperformed by a textbook method without complexity guarantees (nonlinear CG)!**



Outline

Outline

Standard nonlinear CG scheme

Init: $x_0 \in \mathbb{R}^n$, $(\eta, \theta) \in (0,1)^2$, $d_0 = -\nabla f(x_0)$.

For k=$0, 1, 2, \dots$

1. Compute $\alpha_k = \theta^{j_k}$ as the largest stepsize such that

$$f(x_k + \alpha_k d_k) \le f(x_k) + \eta \alpha_k \nabla f(x_k)^{\mathrm{T}} d_k.$$

2. Set $x_{k+1} = x_k + \alpha_k d_k$, and $d_{k+1} = -\nabla f(x_{k+1}) + \beta_{k+1} d_k$.

Standard nonlinear CG scheme

Init: $x_0 \in \mathbb{R}^n$, $(\eta, \theta) \in (0,1)^2$, $d_0 = -\nabla f(x_0)$.

For k=$0, 1, 2, \dots$

1. Compute $\alpha_k = \theta^{j_k}$ as the largest stepsize such that

$$f(x_k + \alpha_k d_k) \le f(x_k) + \eta \alpha_k \nabla f(x_k)^{\mathrm{T}} d_k.$$

2. Set $x_{k+1} = x_k + \alpha_k d_k$, and $d_{k+1} = -\nabla f(x_{k+1}) + \beta_{k+1} d_k$.

- Momentum-based method (but not heavy ball!)
- One formula for β_{k+1}:

$$\beta_{k+1} = \max\left\{ \frac{\nabla f(x_{k+1})^{\mathrm{T}}(\nabla f(x_{k+1}) - \nabla f(x_k))}{\|\nabla f(x_k)\|^2}, 0 \right\}.$$

- Reduces to linear conjugate gradient for quadratics.

Standard nonlinear CG scheme

Init: $x_0 \in \mathbb{R}^n$, $(\eta, \theta) \in (0,1)^2$, $d_0 = -\nabla f(x_0)$.

For k=$0, 1, 2, \dots$

1. Compute $\alpha_k = \theta^{j_k}$ as the largest stepsize such that

$$f(x_k + \alpha_k d_k) \le f(x_k) + \eta \alpha_k \nabla f(x_k)^{\mathrm{T}} d_k.$$

2. Set $x_{k+1} = x_k + \alpha_k d_k$, and $d_{k+1} = -\nabla f(x_{k+1}) + \beta_{k+1} d_k$.
3. If $\nabla f(x_{k+1})^{\mathrm{T}} d_{k+1} \ge 0$, set $d_{k+1} = -\nabla f(x_{k+1})$.

- Momentum-based method (but not heavy ball!)
- One formula for β_{k+1}:

$$\beta_{k+1} = \max \left\{ \frac{\nabla f(x_{k+1})^{\mathrm{T}}(\nabla f(x_{k+1}) - \nabla f(x_k))}{\|\nabla f(x_k)\|^2}, 0 \right\}.$$

- Reduces to linear conjugate gradient for quadratics.
- In practice: Use a restart condition.

What about the restart?

Key divide Restarted VS Non-restarted iterations.

Iteration k is restarted

- $d_k = -\nabla f(x_k) \Rightarrow$ gradient descent step.
- Line search guarantees

$$f(x_{k+1}) \le f(x_k) - \mathcal{O}(\|\nabla f(x_k)\|^2).$$

$\rightarrow$ Good for complexity!

What about the restart?

Key divide Restarted VS Non-restarted iterations.

Iteration k is restarted

- $d_k = -\nabla f(x_k) \Rightarrow$ gradient descent step.
- Line search guarantees

$$f(x_{k+1}) \le f(x_k) - \mathcal{O}(\|\nabla f(x_k)\|^2).$$

$\rightarrow$ Good for complexity!

Iteration k not restarted

- $\nabla f(x_k)^{\mathrm{T}} d_k < 0 \Rightarrow$ Descent direction.
- Line search only guarantees

$$f(x_{k+1}) \le f(x_k)$$

$\rightarrow$ Need more for complexity!

Nonlinear CG with modified restart

Init: $x_0 \in \mathbb{R}^n$, $(\eta, \theta) \in (0,1)^2$, $d_0 = -\nabla f(x_0)$, $p \geq 0$, $\kappa \geq 1$.

For k=$0, 1, 2, \dots$

1. Compute $\alpha_k = \theta^{j_k}$ as the largest stepsize such that

$$f(x_k + \alpha_k d_k) < f(x_k) + \eta \alpha_k \nabla f(x_k)^{\mathrm{T}} d_k.$$

2. Set $x_{k+1} = x_k + \alpha_k d_k$, and $d_{k+1} = -\nabla f(x_{k+1}) + \beta_{k+1} d_k$.
3. If

$$\nabla f(x_{k+1})^{\mathrm{T}} d_{k+1} \geq -\kappa^{-1} \|\nabla f(x_{k+1})\|^{1+p}$$

or

$$\|d_{k+1}\| \geq \kappa \|\nabla f(x_{k+1})\|^{\frac{1+p}{2}},$$

set $d_{k+1} = -\nabla f(x_{k+1})$.

About the restart condition

$$\nabla f(x_{k+1})^{\mathrm{T}} d_{k+1} \geq -\kappa^{-1} \|\nabla f(x_{k+1})\|^{1+p} \quad \text{or} \quad \|d_{k+1}\| \geq \kappa \|\nabla f(x_{k+1})\|^{\frac{1+p}{2}}.$$

1. Restart if d_{k+1} not a good enough descent direction...
2. ...or if $\|d_k\|$ is too large.

About the restart condition

$$\nabla f(x_{k+1})^{\mathrm{T}} d_{k+1} \geq -\kappa^{-1} \|\nabla f(x_{k+1})\|^{1+p} \quad \text{or} \quad \|d_{k+1}\| \geq \kappa \|\nabla f(x_{k+1})\|^{\frac{1+p}{2}}.$$

1. Restart if d_{k+1} not a good enough descent direction...
2. ...or if $\|d_k\|$ is too large.

Special cases

- Classical restart ($\kappa \to \infty$):

$$\nabla f(x_{k+1})^{\mathrm{T}} d_{k+1} \geq 0.$$

- Assumptions for proving convergence of nonlinear CG ($p = 1$):

$$\nabla f(x_{k+1})^{\mathrm{T}} d_{k+1} < -\kappa^{-1} \|\nabla f(x_{k+1})\|^{2} \quad \text{and} \quad \|d_{k+1}\| < \kappa \|\nabla f(x_{k+1})\|.$$

Main result

Theorem (Chan–Renous-Legoubin & R. '22)

NCG with modified restart finds x_k such that $\|\nabla f(x_k)\| \le \epsilon$ in at most

$$\begin{cases} \mathcal{O}(\epsilon^{-2}) & \text{restarted iterations } (d_k = -\nabla f(x_k)) \\ \mathcal{O}(\epsilon^{-(1+p)}) & \text{non-restarted iterations.} \end{cases}$$

Main result

Theorem (Chan–Renous-Legoubin & R. '22)

NCG with modified restart finds x_k such that $\|\nabla f(x_k)\| \leq \epsilon$ in at most

$$\begin{cases} \mathcal{O}(\epsilon^{-2}) & \text{restarted iterations } (d_k = -\nabla f(x_k)) \\ \mathcal{O}(\epsilon^{-(1+p)}) & \text{non-restarted iterations.} \end{cases}$$

- $p = 1$: same complexity as gradient descent.
- $p \in [0, 1)$: better complexity?

Complexity table

Complexity bound

$$\underbrace{\mathcal{O}(\epsilon^{-2})}_{\text{restarted}} + \underbrace{\mathcal{O}(\epsilon^{-(1+p)})}_{\text{non restarted}}$$

- For $p \in [0,1]$, the bound is $\mathcal{O}(\epsilon^{-2})$;
- But better results for **non-restarted iterations**.

p	1	0.75	0.5	0
Order $\epsilon^{-(1+p)}$	ϵ^{-2}	$\epsilon^{-7/4}$	$\epsilon^{-3/2}$	ϵ^{-1}.

Complexity table

Complexity bound

$$\underbrace{\mathcal{O}(\epsilon^{-2})}_{\text{restarted}} + \underbrace{\mathcal{O}(\epsilon^{-(1+p)})}_{\text{non restarted}}$$

- For $p \in [0,1]$, the bound is $\mathcal{O}(\epsilon^{-2})$;
- But better results for **non-restarted iterations**.

p	1	0.75	0.5	0
Order $\epsilon^{-(1+p)}$	ϵ^{-2}	$\epsilon^{-7/4}$	$\epsilon^{-3/2}$	ϵ^{-1}.

Are those values relevant in practice?

Outline

Nonconvex robust regression (from Carmon et al. '17)

Problem data ($n = 30, m = 60$)

- $a_i \sim \mathcal{N}(0, I_n)$;
- $b = Az + 3u_1 + u_2$ with $A = [a_1 \cdots a_m]$, z, u_1 Gaussian and u_2 Bernoulli.

Nonconvex robust regression (from Carmon et al. '17)

Problem data ($n = 30, m = 60$)

- $a_i \sim \mathcal{N}(0, I_n)$;
- $b = Az + 3u_1 + u_2$ with $A = [a_1 \cdots a_m]$, z, u_1 Gaussian and u_2 Bernoulli.

Objective ($\mathcal{C}^2$, Lipschitz gradient)

$$\min_{x \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} \phi(a_i^{\mathrm{T}} x - b_i), \quad \phi(t) = \tfrac{t^2}{1+t^2}.$$

ϕ smoothed version of Tukey loss

$$\rho(t) = \begin{cases} \frac{t^6}{6c^4} - \frac{t^4}{2c^2} + \frac{t^2}{2} & \text{if } |t| \leq c = \frac{1}{\sqrt{6}} \\ \frac{c^2}{6} & \text{otherwise.} \end{cases}$$

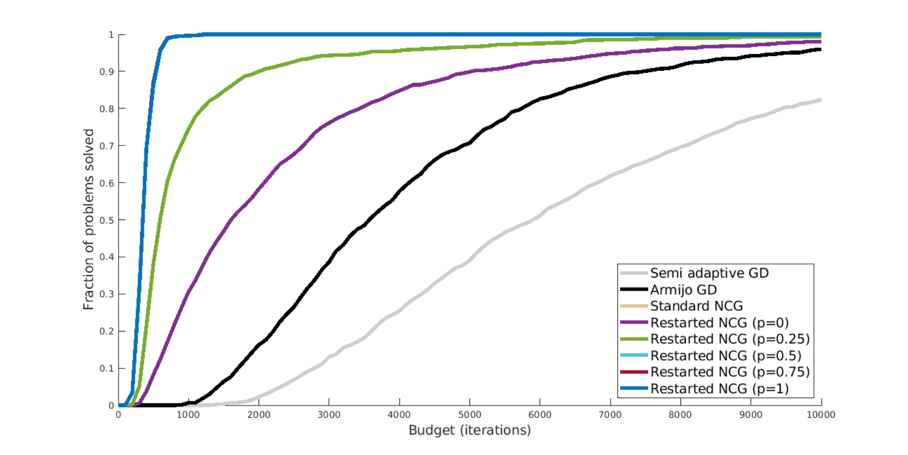
- Semi-adaptive GD (GD+Armijo line search).
- Gradient descent+enhanced Armijo line search.
- Standard nonlinear CG ($\kappa = \infty$);
- Nonlinear CG with several values of p ($\kappa = 100$).

Parameters

- Line search (from Carmon et al. '17): $\theta = \eta = 0.5$.
- Enhancement: For every k, start the line search at iteration $k+1$ with $\theta^{-1}\alpha_k$.
- Nonlinear CG parameter:

$$\beta_{k+1} = \max\left\{ \frac{\nabla f(x_{k+1})^{\mathrm{T}}(\nabla f(x_{k+1}) - \nabla f(x_k))}{\|\nabla f(x_k)\|^2}, 0 \right\}.$$

Data profiles (over 1000 instances)



- Iteration plots for $p \geq 0.5$ overlap with standard NCG!
- But are those iterations similar?

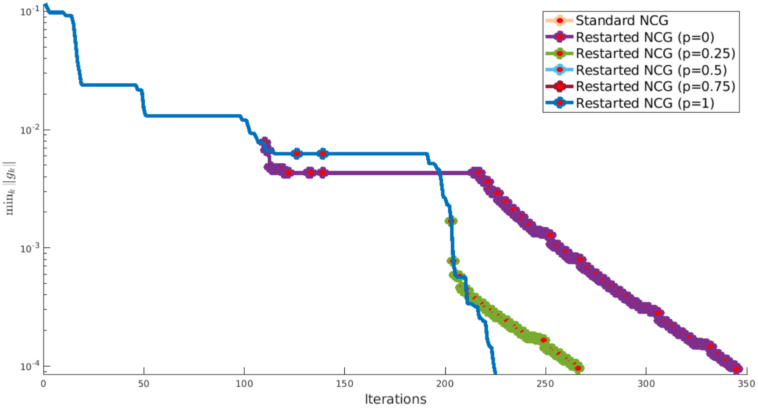
Restart numbers for nonlinear CG

Method (p)	Problems solved	Avg. restart (%)
Standard NCG	1000	0.74%
NCG(0)	1000	83.5%
NCG(0.25)	1000	53.2%
NCG(0.5)	1000	0.89%
NCG(0.75)	1000	0.76%
NCG(1)	1000	0.76%

Percentage of restarted iterations for standard and modified Nonlinear CG (PRP+).

- Small amount of restarted iterations for $p \geq 0.5$;
- Practical evidence of better decrease guarantees!

A representative run



- Plots for $p \geq 0.5$ overlap with standard NCG.
- Restarts triggered for $p < 0.5$ lead to performing gradient descent!

Recap: the nonconvex regression example

Nonlinear CG with modified restarts

- Tracks properties of directions.
- Reverts to gradient descent if needed.
- Complexity guarantees, never worse than GD!

Practical impact of restarting

- For good values of p, most iterations do not restart ($<1\%$).
- Behavior close to original nonlinear CG.

Recap: the nonconvex regression example

Nonlinear CG with modified restarts

- Tracks properties of directions.
- Reverts to gradient descent if needed.
- Complexity guarantees, never worse than GD!

Practical impact of restarting

- For good values of p, most iterations do not restart (<1%).
- Behavior close to original nonlinear CG.

Are we lucky with our setup?

- What about more algorithms/problems?
- What about noisy problems?

Outline

Beyond nonlinear CG

Theory for restarting NCG

$$d_k = -\nabla f(x_k) + \beta_k d_{k-1}.$$

- Does not use β_k.
- Does not really use the definition of d_k.

Beyond nonlinear CG

Theory for restarting NCG

$$d_k = -\nabla f(x_k) + \beta_k d_{k-1}.$$

- Does not use β_k.
- Does not really use the definition of d_k.

One possible extension: L-BFGS

$$d_k = -H_k \nabla f(x_k), \qquad H_k \text{ built from } \{\nabla f(x_{k-i})\}_{i=0,\dots,m}.$$

- One example of the quasi-Newton family.
- Meant for $\mathcal{C}^2$ functions but only requires gradients!

Noisy setup

Problem $\min_{x \in \mathbb{R}^n} f(x)$.

- f and ∇f only available through **noisy oracles $\tilde{f}$ and $\tilde{g}$**.

Noisy setup

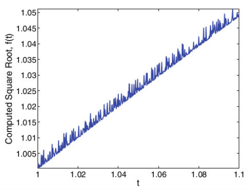
Problem $\min_{x \in \mathbb{R}^n} f(x)$.

- f and ∇f only available through **noisy oracles $\tilde{f}$ and $\tilde{g}$**.
- Noise is more computational than stochastic
  $\Rightarrow$ Can get accurate estimates if needed!

SIAM J. Sci. Comput.
Vol. 33, No. 3, pp. 1292–1314

ESTIMATING COMPUTATIONAL NOISE*

JORGE J. MORÉ† and STEFAN M. WILD†



A simple noise model

Bounded noise model For every $x \in \mathbb{R}^n$,

$$|f(x) - \tilde{f}(x)| \leq \epsilon_f \qquad \|\nabla f(x) - \tilde{g}(x)\| \leq \epsilon_g.$$

A simple noise model

Bounded noise model For every $x \in \mathbb{R}^n$,

$$|f(x) - \tilde{f}(x)| \le \epsilon_f \qquad \|\nabla f(x) - \tilde{g}(x)\| \le \epsilon_g.$$

- Assumes ϵ_f available for algorithmic use.
- Multiple variants to be used together with line search.

SIAM J. OPTIM.
Vol. 31, No. 2, pp. 1489–1518

GLOBAL CONVERGENCE RATE ANALYSIS OF A GENERIC LINE SEARCH ALGORITHM WITH NOISE*

A. S. BERAHAS†, L. CAO‡, AND K. SCHEINBERG§

Line-search algorithm

Init $x_0 \in \mathbb{R}^n$, $(\eta, \theta) \in (0,1)^2$, $d_0 = -\tilde{g}(x_0)$, $\epsilon_f \geq 0$.

For k=$0, 1, 2, \dots$

1. Compute $\alpha_k = \theta^{j_k}$ as the largest stepsize such that

$$\tilde{f}(x_k + \alpha_k d_k) < \tilde{f}(x_k) + \eta \alpha_k \tilde{g}(x_k)^{\mathrm{T}} d_k + 4\epsilon_f.$$

2. Set $x_{k+1} = x_k + \alpha_k d_k$, and compute d_{k+1}.
3. If

$$\tilde{g}(x_{k+1})^{\mathrm{T}} d_{k+1} \geq -\kappa^{-1} \|\tilde{g}(x_{k+1})\|^{1+p}$$

or

$$\|d_{k+1}\| \geq \kappa \|\tilde{g}(x_{k+1})\|^{\frac{1+p}{2}},$$

set $d_{k+1} = -\tilde{g}(x_{k+1})$.

Complexity theory

Assume

- Bounded noise $|f(x) - \tilde{f}(x)| \leq \epsilon_f$, $\|\nabla f(x) - \tilde{g}(x)\| \leq \epsilon_g$.
- $\epsilon_f < \min(\epsilon_g^2, {\epsilon_g}^{1+p})$.

Complexity theory

Assume

- Bounded noise $|f(x) - \tilde{f}(x)| \leq \epsilon_f$, $\|\nabla f(x) - \tilde{g}(x)\| \leq \epsilon_g$.
- $\epsilon_f < \min(\epsilon_g^2, \epsilon_g^{1+p})$.

Theorem (Berahas, O'Neill, R. '25)

For any $\epsilon \geq \max\left\{\epsilon_g, \epsilon_g^{\frac{2}{1+p}}\right\}$, the method finds x_k such that $\|\nabla f(x_k)\| \leq \epsilon$ in at most

$$\begin{cases} \mathcal{O}(\epsilon^{-2}) & \text{restarted iterations } (d_k = -\nabla f(x_k)) \\ \mathcal{O}(\epsilon^{-(1+p)}) & \text{non-restarted iterations.} \end{cases}$$

Assume

- Bounded noise $|f(x) - \tilde{f}(x)| \leq \epsilon_f$, $\|\nabla f(x) - \tilde{g}(x)\| \leq \epsilon_g$.
- $\epsilon_f < \min(\epsilon_g^2, \epsilon_g^{1+p})$.

Theorem (Berahas, O'Neill, R. '25)

For any $\epsilon \geq \max\left\{\epsilon_g, \epsilon_g^{\frac{2}{1+p}}\right\}$, the method finds x_k such that $\|\nabla f(x_k)\| \leq \epsilon$ in at most

$$\begin{cases} \mathcal{O}(\epsilon^{-2}) & \text{restarted iterations } (d_k = -\nabla f(x_k)) \\ \mathcal{O}(\epsilon^{-(1+p)}) & \text{non-restarted iterations.} \end{cases}$$

- Matches deterministic setting and noisy analysis of gradient descent.
- Holds with more generic noise assumptions on $\tilde{g}(x)$.

Outline

Test problems

- 234 problems from CUTEst, $1 \leq n \leq 1000$.
- S2PMJ interface.

Test problems

- 234 problems from CUTEst, $1 \leq n \leq 1000$.
- S2PMJ interface.

- Add uniform noise, five noise levels per problem

$$\epsilon_f \in \left\{0, 10^{-8}, 10^{-4}, 10^{-2}, 10^{-1}\right\}, \qquad \epsilon_g = \sqrt{\epsilon_f}.$$

- 10 runs for each pair (problem,noise).
- Runs on UNC's Longleaf cluster.

Methods

- Gradient descent (GD)
- Nonlinear CG (NLCG), LBFGS
- Variants with restarts (NLCGr, LBFGSr).

Methods

- Gradient descent (GD)
- Nonlinear CG (NLCG), LBFGS
- Variants with restarts (NLCGr, LBFGSr).

- ϵ_f known to solvers, same line-search parameters.
- Goal: Reach $\|\tilde{g}(x_k)\|_\infty \le \max\{2\epsilon_g, 10^{-8}\}$.
- True success: $\|\nabla f(x_k)\|_\infty \le \epsilon_g + \max\{2\epsilon_g, 10^{-8}\}$.

$\Rightarrow$ Convergence easier with larger noise.

Numerics in a nutshell

Choosing good restart parameters (p, κ)

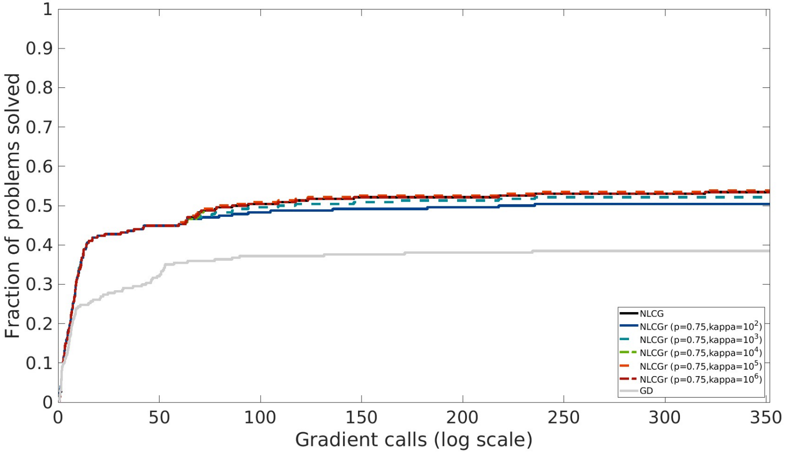
- NLCGr: Use $p = 0.75$ and $\kappa \geq 10^2$.
- LBFGSr: Use $\kappa = 10^6$, best p depends on the noise level.

Comparing best restarted variants with standard ones

- NLCGr usually overlaps with NLCG for $p \geq 0.5$.
- LBFGSr close to non-restarted variants with good (p, κ).
- Standard LBFGS best overall!

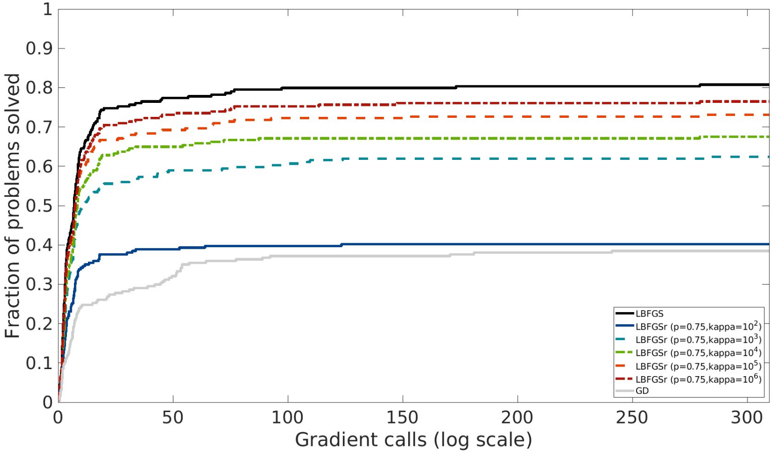
Sensitivity of restarting conditions (1/2)

$$\tilde{g}(x_{k+1})^{\mathrm{T}} d_{k+1} \geq -\kappa^{-1} \|\tilde{g}(x_{k+1})\|^{1+p} \quad \text{or} \quad \|d_{k+1}\| \geq \kappa \|\tilde{g}(x_{k+1})\|^{\frac{1+p}{2}}.$$

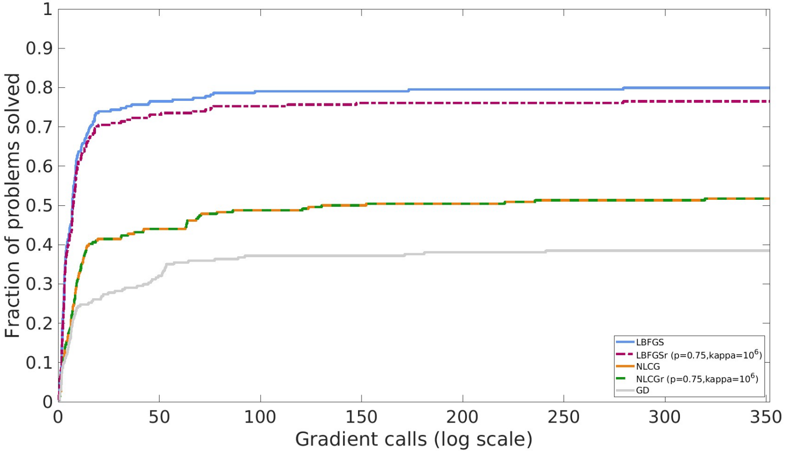


Sensitivity of restarting conditions (2/2)

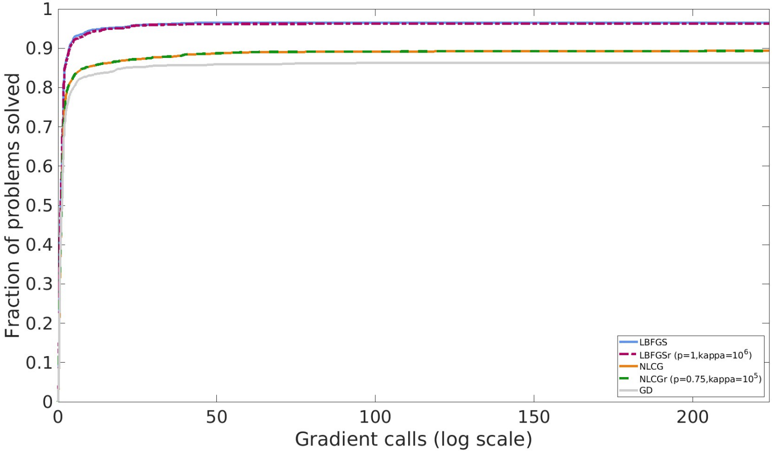
$$\tilde{g}(x_{k+1})^{\mathrm{T}} d_{k+1} \geq -\kappa^{-1}\|\tilde{g}(x_{k+1})\|^{1+p} \quad \text{or} \quad \|d_{k+1}\| \geq \kappa \|\tilde{g}(x_{k+1})\|^{\frac{1+p}{2}}.$$



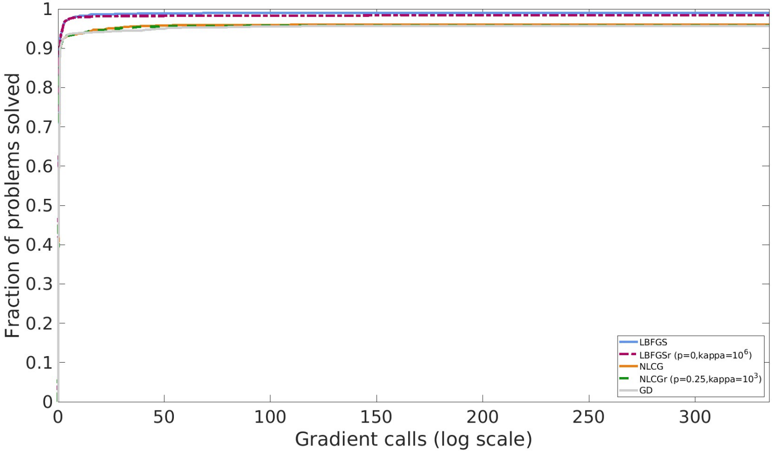
Data profiles in noiseless case ($\epsilon_f = 0$)



Data profiles with moderate noise ($\epsilon_f = 10^{-4}$)



Data profiles with large noise ($\epsilon_f = 10^{-1}$)



Conclusion

What we started with

- A puzzling nonconvex regression example.
- No complexity theory for nonlinear CG.

Conclusion

What we started with

- A puzzling nonconvex regression example.
- No complexity theory for nonlinear CG.

What we got

- A theory that applies to nonlinear CG and L-BFGS thanks to **restarts**.
- Implementable techniques that are close to original methods in practice.

Conclusion

What we started with

- A puzzling nonconvex regression example.
- No complexity theory for nonlinear CG.

What we got

- A theory that applies to nonlinear CG and L-BFGS thanks to **restarts**.
- Implementable techniques that are close to original methods in practice.

What we still have to do

- Use other line searches.
- Understand nonconvex regression case?

EURO Journal on Computational Optimization
Volume 10, 2022, 100044

A nonlinear conjugate gradient method with complexity guarantees and its application to nonconvex regression

Rémi Chan–Renous-Legoubin [a], Clément W. Royer [b, 1]

arXiv > math > arXiv:2506.03358

Mathematics > Optimization and Control

[Submitted on 3 Jun 2025 (v1), last revised 18 Oct 2025 (this version, v2)]

A line search framework with restarting for noisy optimization problems

Albert S. Berahas, Michael J. O'Neill, Clément W. Royer

For more details

EURO Journal on Computational Optimization
Volume 10, 2022, 100044

A nonlinear conjugate gradient method with complexity guarantees and its application to nonconvex regression

Rémi Chan–Renous-Legoubin [a], Clément W. Royer [b 1]

arXiv > math > arXiv:2506.03358

Mathematics > Optimization and Control

[Submitted on 3 Jun 2025 (v1), last revised 18 Oct 2025 (this version, v2)]

A line search framework with restarting for noisy optimization problems

Albert S. Berahas, Michael J. O'Neill, Clément W. Royer

Thank you!

`clement.royer@lamsade.dauphine.fr`