Newton-Krylov techniques for nonconvex optimization

Clément W. Royer

Computational Maths Seminar, Australian National University

October 25, 2021

French people in Australia, summer of 2021


- First victory in 31 years!
- My first Australian talk in 31 years!

Why I am here

Talk about complexity...

- As opposed to global/local convergence results;
- **Goal:** Equip popular practical schemes with such guarantees.

..and linear algebra...

- Key to high-performance implementation;
- Krylov methods+Randomization!

...to make a case for second-order methods.

- Newton+Conjugate Gradient;
- Nonconvex setting.

Outline

1. Complexity and nonconvexity
2. Conjugate gradient and nonconvex quadratics
3. Newton-CG framework
4. Numerics

Outline

Nonconvex optimization

Nonconvex ?

- Many data science problems are convex: linear classification, logistic regression,...
- **Nonconvex** instances: Deep learning, matrix/tensor optimization, robust statistics.

Nonconvex optimization

Nonconvex ?

- Many data science problems are convex: linear classification, logistic regression,...
- **Nonconvex** instances: Deep learning, matrix/tensor optimization, robust statistics.

Optimization ?

- Those problems often come with structure;
- Guarantees to find global optima using second-order conditions;
- **Are high-order methods suitable then?**

General problem and definitions

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

with $f \in \mathcal{C}^2(\mathbb{R}^n)$ bounded below and nonconvex.

General problem and definitions

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

with $f \in \mathcal{C}^2(\mathbb{R}^n)$ bounded below and nonconvex.

Definitions in smooth nonconvex minimization

- *First-order stationary point*: $\|\nabla f(x)\| = 0$;
- *Second-order stationary point*: $\|\nabla f(x)\| = 0, \nabla^2 f(x) \succeq 0$.

General problem and definitions

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

with $f \in \mathcal{C}^2(\mathbb{R}^n)$ bounded below and nonconvex.

Definitions in smooth nonconvex minimization

- *First-order stationary point*: $\|\nabla f(x)\| = 0$;
- *Second-order stationary point*: $\|\nabla f(x)\| = 0, \nabla^2 f(x) \succeq 0$.

If x does not satisfy these conditions, $\exists\ d$ such that

1. $d^\top \nabla f(x) < 0$: **gradient-related direction**.
   and/or
2. $d^\top \nabla^2 f(x) d < 0$: **negative curvature direction** $\Rightarrow$ **specific to nonconvex problems**.

The matrix completion example

Matrix completion

$$\min_{X \in \mathbb{R}^{n \times m}, \mathrm{rank}(X) \leq r} \sum_{(i,j) \in \Omega} (X_{ij} - M_{ij})^2 \quad M \in \mathbb{R}^{n \times m},\ \Omega \subset [n] \times [m].$$

- Data: observed entries of M.
- Assumption: The true matrix is of (low) rank $r \ll \min(m, n)$.

The matrix completion example

Matrix completion

$$\min_{X \in \mathbb{R}^{n \times m}, \mathrm{rank}(X) \leq r} \sum_{(i,j) \in \Omega} (X_{ij} - M_{ij})^2 \quad M \in \mathbb{R}^{n \times m},\ \Omega \subset [n] \times [m].$$

- Data: observed entries of M.
- Assumption: The true matrix is of (low) rank $r \ll \min(m, n)$.

Nonconvex factored reformulation (Burer & Monteiro, '03)

$$\min_{U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{m \times r}} \sum_{(i,j) \in \Omega} \left([U V^\top]_{ij} - M_{ij} \right)^2,$$

- $(n+m)r$ variables ($\ll nm$).
- **Nonconvex in U and V**...
- ..but global minima can be characterized.

A nice class of nonconvex problems

Nonconvex formulations for low-rank matrix problems (Ge et al. 2017)

$$\min_{U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{m \times r}} f(U V^\top) \quad f \text{ smooth.}$$

A nice class of nonconvex problems

Nonconvex formulations for low-rank matrix problems (Ge et al. 2017)

$$\min_{U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{m \times r}} f(U V^\top) \quad f \text{ smooth.}$$

- Second-order stationary points are global minima (or are close in function value);
- Strict saddle property: any first-order stationary point that is not a local minimum possesses negative curvature.

Nonconvex formulations for low-rank matrix problems (Ge et al. 2017)

$$\min_{U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{m \times r}} f(U V^\top) \quad f \text{ smooth.}$$

- Second-order stationary points are global minima (or are close in function value);
- Strict saddle property: any first-order stationary point that is not a local minimum possesses negative curvature.

- Obj: **efficient** algorithms to reach second-order stationary points;
- Efficiency measured by **complexity**.

Complexity in nonconvex optimization

Setup: Sequence of points $\{x_k\}$ generated by an algorithm applied to $\min_{x \in \mathbb{R}^n} f(x)$.

Complexity in nonconvex optimization

Setup: Sequence of points $\{x_k\}$ generated by an algorithm applied to $\min_{x \in \mathbb{R}^n} f(x)$.

First-order complexity result

Given $\epsilon_g \in (0, 1)$:

- **Worst-case cost** to obtain an ϵ_g-point x_K such that $\|\nabla f(x_K)\| \leq \epsilon_g$.
- Focus: **Dependency on ϵ_g**.

Complexity in nonconvex optimization

Setup: Sequence of points $\{x_k\}$ generated by an algorithm applied to $\min_{x \in \mathbb{R}^n} f(x)$.

First-order complexity result

Given $\epsilon_g \in (0,1)$:

- **Worst-case cost** to obtain an ϵ_g-point x_K such that $\|\nabla f(x_K)\| \le \epsilon_g$.
- Focus: **Dependency on** ϵ_g.

Second-order complexity result

Given $\epsilon_g, \epsilon_H \in (0,1)$:

- **Worst-case cost** to obtain an (ϵ_g, ϵ_H)**-point** x_K such that

$$\|\nabla f(x_K)\| \le \epsilon_g, \qquad \lambda_{\min}(\nabla^2 f(x_K)) \ge -\epsilon_H.$$

- Focus: **Dependencies on** ϵ_g, ϵ_H.

Gradient descent

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

Gradient descent

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

- With appropriate stepsize choice,

$$f(x_k) - f(x_{k+1}) \geq \mathcal{O}\left(\|\nabla f(x_k)\|^2\right)$$

- $\|\nabla f(x_k)\| \leq \epsilon_g$ in at most $\mathcal{O}\left(\epsilon_g^{-2}\right)$ iterations;
- 1 iteration=1 gradient evaluation.

Gradient descent

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

- With appropriate stepsize choice,

$$f(x_k) - f(x_{k+1}) \geq \mathcal{O}\left(\|\nabla f(x_k)\|^2\right)$$

- $\|\nabla f(x_k)\| \leq \epsilon_g$ in at most $\mathcal{O}\left(\epsilon_g^{-2}\right)$ iterations;
- 1 iteration=1 gradient evaluation.

Sharp result

- Pathological examples (Cartis, Gould, Toint, 2010);
- Bound holds for several other methods.

Gradient descent+Negative curvature

1. If $\|\nabla f(x_k)\| > \epsilon_g$, set $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$ with $\alpha_k > 0$;
2. If $\|\nabla f(x_k)\| \leq \epsilon_g$ and $\lambda_k = \lambda_{\min}(\nabla^2 f(x_k)) < -\epsilon_H$, set $x_{k+1} = x_k + \alpha_k d_k$ where $\alpha_k > 0$ and

$$d_k^{\mathrm{T}} \nabla^2 f(x_k) d_k = -\lambda_k \|d_k\|^2, \quad d_k^{\mathrm{T}} \nabla f(x_k) \leq 0.$$

Gradient descent+Negative curvature

1. If $\|\nabla f(x_k)\| > \epsilon_g$, set $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$ with $\alpha_k > 0$;
2. If $\|\nabla f(x_k)\| \le \epsilon_g$ and $\lambda_k = \lambda_{\min}(\nabla^2 f(x_k)) < -\epsilon_H$, set $x_{k+1} = x_k + \alpha_k d_k$ where $\alpha_k > 0$ and

$$d_k^{\mathrm{T}} \nabla^2 f(x_k) d_k = -\lambda_k \|d_k\|^2, \quad d_k^{\mathrm{T}} \nabla f(x_k) \le 0.$$

- With appropriate stepsize choice,

$$f(x_k) - f(x_{k+1}) \ge \begin{cases} \mathcal{O}\left(\|\nabla f(x_k)\|^2\right) \\ \mathcal{O}\left(|\lambda_k|^3\right) \end{cases}$$

- $\|\nabla f(x_k)\| \le \epsilon_g$ **and** $\nabla^2 f(x_k) \succeq -\epsilon_H I$ in at most $\mathcal{O}\left(\max\{\epsilon_g^{-2}, \epsilon_H^{-3}\}\right)$ iterations;
- 1 iteration=1 gradient evaluation+1 eigenvalue/eigenvector calculation.

Complexity results

From nonconvex optimization (2006-)

- Cost measure: Number of iterations (but those may be expensive);
- Two types of guarantees:
  1. $\|\nabla f(x)\| \le \epsilon_g$;
  2. $\|\nabla f(x)\| \le \epsilon_g$ and $\nabla^2 f(x) \succeq -\epsilon_H I$.
- Best methods: Second-order methods, deterministic variations on Newton's iteration involving Hessians.

Complexity results

From nonconvex optimization (2006-)

- Cost measure: Number of iterations (but those may be expensive);
- Two types of guarantees:
  1. $\|\nabla f(x)\| \leq \epsilon_g$;
  2. $\|\nabla f(x)\| \leq \epsilon_g$ and $\nabla^2 f(x) \succeq -\epsilon_H I$.
- Best methods: Second-order methods, deterministic variations on Newton's iteration involving Hessians.

Gradient Descent + Negative Curvature	1	$\mathcal{O}\left(\epsilon_g^{-2}\right)$
	2	$\mathcal{O}\left(\max\{\epsilon_g^{-2}, \epsilon_H^{-3}\}\right)$
Trust Region	1	$\mathcal{O}\left(\epsilon_g^{-2}\right)$
	2	$\mathcal{O}\left(\max\{\epsilon_g^{-2}\epsilon_H^{-1}, \epsilon_H^{-3}\}\right)$
Cubic Regularization	1	$\mathcal{O}\left(\epsilon_g^{-3/2}\right)$
	2	$\mathcal{O}\left(\max\{\epsilon_g^{-3/2}, \epsilon_H^{-3}\}\right)$

Complexity results (2)

Influenced by convex optimization/learning (2016-)

- Cost measure: gradient evaluations+Hessian-vector products $\Rightarrow$ main iteration cost.
- Two types of guarantees:
  1. $\|\nabla f(x)\| \le \epsilon_g$
  2. $\|\nabla f(x)\| \le \epsilon_g$ and $\nabla^2 f(x) \succeq -\epsilon_g^{1/2} I$.
- Best methods: developed from accelerated gradient (better than gradient descent on convex problems).

Complexity results (2)

Influenced by convex optimization/learning (2016-)

- Cost measure: gradient evaluations+Hessian-vector products $\Rightarrow$ main iteration cost.
- Two types of guarantees:
  1. $\|\nabla f(x)\| \leq \epsilon_g$
  2. $\|\nabla f(x)\| \leq \epsilon_g$ and $\nabla^2 f(x) \succeq -\epsilon_g^{1/2} I$.
- Best methods: developed from accelerated gradient (better than gradient descent on convex problems).

Gradient descent + random perturbation	①,②	$\tilde{\mathcal{O}}\left(\epsilon_g^{-2}\right)$	(High probability)
Accelerated gradient + random perturbation	①,②	$\tilde{\mathcal{O}}(\epsilon_g^{-7/4})$	(High probability)
Accelerated gradient with nonconvexity detection	①	$\tilde{\mathcal{O}}(\epsilon_g^{-7/4})$	(Deterministic)

What we want to do

Newton's method

$$x_{k+1} = x_k + \alpha_k d_k, \quad \nabla^2 f(x_k) d_k = -\nabla f(x_k)$$

- α_k computed via line search for global convergence;
- Large-scale implementation: Conjugate Gradient (CG);
- Works well when $\nabla^2 f(x_k) \succ 0$.

What we want to do

Newton's method

$$x_{k+1} = x_k + \alpha_k d_k, \quad \nabla^2 f(x_k) d_k = -\nabla f(x_k)$$

- α_k computed via line search for global convergence;
- Large-scale implementation: Conjugate Gradient (CG);
- Works well when $\nabla^2 f(x_k) \succ 0$.

Newton's method in nonconvex case

Big issue: $\nabla^2 f(x_k) \not\succ 0$!

- Still used in practice $\Rightarrow$ **Can we explain it?**
- Efficient $\Rightarrow$ **Can we get complexity guarantees?**

Outline

Nonconvex quadratics

$$\min_{y \in \mathbb{R}^n} \frac{1}{2} y^{\mathrm{T}} H y + g^{\mathrm{T}} y$$

with $H = H^{\mathrm{T}} \in \mathbb{R}^{n \times n}$ **not necessarily positive definite**, $g \in \mathbb{R}^n$.

Regularized variants

Trust region: $\min_{y \in \mathbb{R}^n} \frac{1}{2} y^{\mathrm{T}} H y + g^{\mathrm{T}} y \quad \text{s.t.} \quad \|y\| \leq \delta$

Cubic regularization: $\min_{y \in \mathbb{R}^n} g^{\mathrm{T}} y + \frac{1}{2} y^{\mathrm{T}} H y + \frac{\sigma}{3} \|y\|^3$.

Nonconvex quadratics

$$\min_{y \in \mathbb{R}^n} \frac{1}{2} y^{\mathrm{T}} H y + g^{\mathrm{T}} y$$

with $H = H^{\mathrm{T}} \in \mathbb{R}^{n \times n}$ **not necessarily positive definite**, $g \in \mathbb{R}^n$.

Regularized variants

Trust region: $\min_{y \in \mathbb{R}^n} \frac{1}{2} y^{\mathrm{T}} H y + g^{\mathrm{T}} y \quad \text{s.t.} \quad \|y\| \leq \delta$

Cubic regularization: $\min_{y \in \mathbb{R}^n} g^{\mathrm{T}} y + \frac{1}{2} y^{\mathrm{T}} H y + \frac{\sigma}{3} \|y\|^3$.

Lanczos-type approaches (Carmon & Duchi 2020, Gould & Simoncini 2020)

- Solve the problem over the Krylov subspace $\{g, Hg, H^2 g, \ldots, H^{j-1} g\}$;
- Can fail to compute the solution (hard case, occurs when $H \not\succeq 0$);
- But complexity guarantees hold **in probability!**

Nonconvex quadratics in nonconvex optimization

$$\min_{x \in \mathbb{R}^n} f(x) \;\Rightarrow\; \min_{y \in \mathbb{R}^n} \frac{1}{2} y^{\mathrm{T}} \nabla^2 f(x_k) y + y^{\mathrm{T}} \nabla f(x_k)$$

- Do we really want to solve the quadratic problem?
- We actually want to compute a step to go from x_k to x_{k+1}!
- If the quadratic is unbounded ($\nabla^2 f(x_k) \not\succeq 0$), negative curvature directions can be used.

Nonconvex quadratics in nonconvex optimization

$$\min_{x \in \mathbb{R}^n} f(x) \;\Rightarrow\; \min_{y \in \mathbb{R}^n} \frac{1}{2} y^{\mathrm{T}} \nabla^2 f(x_k) y + y^{\mathrm{T}} \nabla f(x_k)$$

- Do we really want to solve the quadratic problem?
- We actually want to compute a step to go from x_k to x_{k+1}!
- If the quadratic is unbounded ($\nabla^2 f(x_k) \not\succeq 0$), negative curvature directions can be used.

Our subproblem

Given a quadratic $q : y \in \mathbb{R}^n \mapsto \frac{1}{2} y^{\mathrm{T}} H y + g^T y$,

1. Find an approximate minimum of q...
2. OR compute a direction of negative curvature for H.

Can we do that using conjugate gradient?

The conjugate gradient method

Goal: Solve $Hy = -g$, where $H = H^{\mathrm{T}} \succ 0$.

Conjugate gradient method

Init: Set $y_0 = 0_{\mathbb{R}^n},\ r_0 = g,\ p_0 = -g,\ j = 0,\ \xi \geq 0$.

For $j = 0, 1, 2, \ldots$

- Compute $y_{j+1} = y_j + \frac{\|r_j\|^2}{p_j^{\mathrm{T}} H p_j} p_j$ and $r_{j+1} = H y_{j+1} + g$.
- Set $p_{j+1} = -r_{j+1} + \frac{\|r_{j+1}\|^2}{\|r_j\|^2} p_j$.
- Set $j = j + 1$; terminate if $\|r_j\| \leq \xi \|r_0\|$.

- Only requires $v \mapsto Hv$ ("matrix-free");
- Terminates in at most n iterations **when** $H \succ 0$.

Complexity of conjugate gradient

Recall: $r_j = Hy_j + g$.

Convergence rate of CG

If $\epsilon_H I \prec H \preceq MI$,

$$\|r_j\|^2 \leq 4\kappa \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^{2j} \|r_0\|^2, \quad \kappa = \frac{M}{\epsilon_H}.$$

Conjugate gradient for $Hy = -g$

If $\epsilon_H I \prec H \preceq MI$, $\|Hy_J + g\| \leq \xi \|g\|$ after at most

$$J = \min\left\{n, \mathcal{O}(\kappa^{1/2}\ln(\kappa/\xi))\right\} = \min\left\{n, \tilde{\mathcal{O}}(\epsilon_H^{-1/2})\right\}$$

iterations/matrix-vector products.

CG in nonconvex optimization

What can go wrong?

- We'll consider $Hy = -g$ with possibly $H \nsucceq 0$;
- Two issues:
  - Presence of negative curvature;
  - Loss of guarantees for CG steps.

How to make it right?

- **Regularization**;
- Use intrinsic nonconvexity detection properties of CG.

Conjugate gradient for $Hy = -g$

Algorithm

Init: Set $y_0 = 0_{\mathbb{R}^n},\ r_0 = g,\ p_0 = -g,\ j = 0, \xi \geq 0$.

For $j = 0, 1, \dots$

- Compute $y_{j+1} = y_j + \frac{\|r_j\|^2}{p_j^{\mathrm{T}} H p_j} p_j,\ r_{j+1} = H y_{j+1} + g$ and p_{j+1}.
- Set $j = j + 1$; terminate if $\|r_j\| \leq \xi \|r_0\|$.

Conjugate gradient for $Hy = -g$

Algorithm

Init: Set $y_0 = 0_{\mathbb{R}^n}, \ r_0 = g, \ p_0 = -g, \ j = 0, \xi \geq 0$.

For $j = 0, 1, \ldots$

- Compute $y_{j+1} = y_j + \frac{\|r_j\|^2}{p_j^{\mathrm{T}} H p_j} p_j, \ r_{j+1} = H y_{j+1} + g$ and p_{j+1}.
- Set $j = j + 1$; terminate if $\|r_j\| \leq \xi \|r_0\|$.

Convergence rate of CG (gives complexity)

If $\epsilon_H I \prec H \preceq MI$,

$$\|r_j\|^2 \leq 4\kappa \left(1 - \frac{2}{\sqrt{\kappa} + 1}\right)^{2j} \|r_0\|^2, \quad \kappa = \frac{M}{\epsilon_H}.$$

Conjugate gradient for $Hy = -g$

Algorithm assuming $\epsilon_H I \prec H \preceq MI$

Init: Set $y_0 = 0_{\mathbb{R}^n}$, $r_0 = g$, $p_0 = -g$, $j = 0, \xi \geq 0$.

While $p_j^\top H p_j > \epsilon_H \|p_j\|^2$ **and** $\|r_j\|^2 \leq 4\kappa \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^{2j} \|r_0\|^2$

- Compute $y_{j+1} = y_j + \frac{\|r_j\|^2}{p_j^\mathrm{T} H p_j} p_j$, $r_{j+1} = Hy_{j+1} + g$ and p_{j+1}.
- Set $j = j + 1$; terminate if $\|r_j\| \leq \xi \|r_0\|$.

Convergence rate of CG (gives complexity)

If $\epsilon_H I \prec H \preceq MI$,

$$\|r_j\|^2 \leq 4\kappa \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^{2j} \|r_0\|^2, \quad \kappa = \frac{M}{\epsilon_H}.$$

Conjugate gradient for $Hy = -g$

Algorithm assuming $\epsilon_H I \prec H \preceq MI$

Init: Set $y_0 = 0_{\mathbb{R}^n}$, $r_0 = g$, $p_0 = -g$, $j = 0, \xi \geq 0$.

While $p_j^\top H p_j > \epsilon_H \|p_j\|^2$ **and** $\|r_j\|^2 \leq 4\kappa \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^{2j} \|r_0\|^2$

- Compute $y_{j+1} = y_j + \frac{\|r_j\|^2}{p_j^\top H p_j} p_j$, $r_{j+1} = Hy_{j+1} + g$ and p_{j+1}.
- Set $j = j+1$; terminate if $\|r_j\| \leq \xi \|r_0\|$.

Convergence rate of CG (gives complexity)

If $\epsilon_H I \prec H \preceq MI$,

$$\|r_j\|^2 \leq 4\kappa \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^{2j} \|r_0\|^2, \quad \kappa = \frac{M}{\epsilon_H}.$$

What if $H \not\succ \epsilon_H I$?

Conjugate gradient for possibly indefinite systems

Capped Conjugate Gradient

Init: Set $y_0 = 0_{\mathbb{R}^n}$, $r_0 = g$, $p_0 = -g$, $j = 0, \xi \geq 0$.

While $p_j^\top H p_j > \epsilon_H \|p_j\|^2$ **and** $\|r_j\|^2 \leq T \tau^j \|r_0\|^2$

- Compute $y_{j+1} = y_j + \frac{\|r_j\|^2}{p_j^\top H p_j} p_j$, $r_{j+1} = H y_{j+1} + g$ and p_{j+1}.
- Set $j = j + 1$; terminate if $\|r_j\| \leq \xi \|r_0\|$.

Conjugate gradient for possibly indefinite systems

Capped Conjugate Gradient

Init: *Set* $y_0 = 0_{\mathbb{R}^n},\ r_0 = g,\ p_0 = -g,\ j = 0, \xi \geq 0.$

While $p_j^\top H p_j > \epsilon_H \|p_j\|^2$ **and** $\|r_j\|^2 \leq T\tau^j \|r_0\|^2$

- Compute $y_{j+1} = y_j + \frac{\|r_j\|^2}{p_j^\top H p_j} p_j,\ r_{j+1} = H y_{j+1} + g$ and p_{j+1}.
- Set $j = j+1$; terminate if $\|r_j\| \leq \xi \|r_0\|$.

Properties of Capped CG

If $H \preceq MI$:

- As long as r_j is computed:

$$\|r_j\|^2 \leq T\tau^j \|r_0\|^2, \qquad T = 16\kappa^5,\ \tau = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}, \kappa = \frac{M}{\epsilon_H}.$$

- The method runs at most $\hat{J} = \min\left\{n, \tilde{\mathcal{O}}\left(\epsilon_H^{-1/2}\right)\right\}$ iterations ("cap") before terminating or **violating one condition**.

Main result - Violating conditions in Capped CG

Theorem (Royer, O'Neill, Wright - 2020)

If Capped CG applied to $Hd = -g$ stops after $J \leq \hat{J}$ iterations with $\|r_j\| > \xi \|r_0\|$, then

1. Either $p_J^\top H p_J \leq \epsilon_H \|p_J\|^2$,
2. Or $\|r_J\|^2 > T\tau^J \|r_0\|^2$,

Main result - Violating conditions in Capped CG

Theorem (Royer, O'Neill, Wright - 2020)

If Capped CG applied to $Hd = -g$ stops after $J \le \hat{J}$ iterations with $\|r_j\| > \xi \|r_0\|$, then

1. Either $p_J^\top H p_J \le \epsilon_H \|p_J\|^2$,
2. Or $\|r_J\|^2 > T\tau^J \|r_0\|^2$, y_{J+1} can be computed and there exists $j \in \{0, \dots, J-1\}$ such

$$(y_{J+1} - y_j)^\top H (y_{J+1} - y_j) \le \epsilon_H \|y_{J+1} - y_j\|^2.$$

Main result - Violating conditions in Capped CG

Theorem (Royer, O'Neill, Wright - 2020)

If Capped CG applied to $Hd = -g$ stops after $J \leq \hat{J}$ iterations with $\|r_j\| > \xi \|r_0\|$, then

1. Either $p_J^\top H p_J \leq \epsilon_H \|p_J\|^2$,
2. Or $\|r_J\|^2 > T\tau^J \|r_0\|^2$, y_{J+1} can be computed and there exists $j \in \{0, \ldots, J-1\}$ such

$$(y_{J+1} - y_j)^\top H (y_{J+1} - y_j) \leq \epsilon_H \|y_{J+1} - y_j\|^2.$$

What it means

- Can run (Capped) CG without computing $\lambda_{\min}(H)$ first!
- Either we converge **as if we had $H \succ \epsilon_H I$**...
- ...or we find a direction of curvature $\leq \epsilon_H$!

CG and minimum eigenvalues

Estimating eigenvalues

Task: Given $H = H^\top$, find d such that $d^\top H d \le 0$ if $H \nsucceq -\epsilon_H I$.

CG and minimum eigenvalues

Estimating eigenvalues

Task: Given $H = H^\top$, find d such that $d^\top H d \leq 0$ if $H \not\succeq -\epsilon_H I$.

- Even Capped CG does not necessarily detect negative curvature!
- We would like to know whether $\lambda_{\min}(\nabla^2 f(x)) > -\epsilon_H$ (for complexity).

CG and minimum eigenvalues

Estimating eigenvalues

Task: Given $H = H^\top$, find d such that $d^\top H d \le 0$ if $H \nsucceq -\epsilon_H I$.

- Even Capped CG does not necessarily detect negative curvature!
- We would like to know whether $\lambda_{\min}(\nabla^2 f(x)) > -\epsilon_H$ (for complexity).

Approach

Run CG on a linear system with a random right-hand side **uniformly distributed on the unit sphere**.

- Guarantees approximation of $\lambda_{\min}(H)$ with high probability (Kuczyński and Woźniakowski 1992) for Lanczos' method;
- Lanczos and CG generate the same Krylov subspaces!

CG and minimum eigenvalues (2)

Theorem (Royer, O'Neill, Wright 2020)

Let $H \in \mathbb{R}^{n \times n}$ symmetric with $\|H\| \leq M$, $\delta \in [0,1)$, and CG be applied to

$$\left(H + \frac{\epsilon_H}{2} I\right) y = b \quad \text{with} \quad b \sim \mathcal{U}(\mathbb{S}^{n-1}).$$

Then, after

$$J = \min\left\{n, \left\lceil \frac{\ln(3n/\delta^2)}{2} \sqrt{\frac{M}{\epsilon_H}} \right\rceil \right\} = \min\left\{n, \tilde{\mathcal{O}}(\epsilon_H^{-1/2})\right\}.$$

iterations,

- Either CG finds negative curvature explicitly: $p_J^{\mathrm{T}} \left(H + \frac{\epsilon_H}{2} I\right) p_J \leq 0$;
- Or it certifies with probability at least $1 - \delta$ that $H \succ -\epsilon_H I$.

Outline

Line-Search Newton-Capped CG

Inputs: $x_0 \in \mathbb{R}^n$, $\theta, \xi \in (0,1)$, $\eta > 0, \epsilon_g, \epsilon_H \in (0,1), \delta \in [0,1)$.

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

1. If $\|\nabla f(x_k)\| > \epsilon_g$, compute d_k via **Capped CG** applied to

$$\left(\nabla^2 f(x_k) + 2\epsilon_H I\right) d = -\nabla f(x_k).$$

2. Otherwise, use CG as an eigenvalue oracle with probability δ. If it certifies that $\nabla^2 f(x_k) \succ -\epsilon_H I$ terminate, otherwise use its output as d_k.
3. Perform a backtracking line search to compute $\alpha_k = \theta^{j_k}$ such that

$$f(x_k + \alpha_k d_k) \ < \ f(x_k) - \frac{\eta}{6}\alpha_k^3 \|d_k\|^3.$$

4. Set $x_{k+1} = x_k + \alpha_k\, d_k$.

Capped Conjugate Gradient for Newton steps

Key result

Apply Capped CG to

$$\left(\nabla^2 f(x_k) + 2\epsilon_H I\right) d \;=\; -\nabla f(x_k).$$

Then, after **at most** $\min\left\{n, \tilde{\mathcal{O}}(\epsilon_H^{-1/2})\right\}$ **iterations/Hessian-vector products**, the methods outputs

1. a regularized Newton step d_k with

$$\left\|(\nabla^2 f(x_k) + 2\epsilon_H I)d_k + \nabla f(x_k)\right\| \;\leq\; \xi\|\nabla f(x_k)\|;$$

2. Or a direction of curvature $\leq \epsilon_H$ for $\nabla^2 f(x_k) + 2\epsilon_H I$.

Capped Conjugate Gradient for Newton steps

Key result

Apply Capped CG to

$$\left(\nabla^2 f(x_k) + 2\epsilon_H I\right) d \;=\; -\nabla f(x_k).$$

Then, after **at most** $\min\left\{n, \tilde{\mathcal{O}}(\epsilon_H^{-1/2})\right\}$ **iterations/Hessian-vector products**, the methods outputs

1. a regularized Newton step d_k with
$$\left\|(\nabla^2 f(x_k) + 2\epsilon_H I)d_k + \nabla f(x_k)\right\| \;\leq\; \xi\|\nabla f(x_k)\|;$$
2. Or a direction of negative curvature $\leq -\epsilon_H$ for $\nabla^2 f(x_k)$!

CG as minimum eigenvalue oracle

For the matrix $\nabla^2 f(x_k)$, consider CG applied to

$$\left(\nabla^2 f(x_k) + \frac{\epsilon_H}{2} I\right) d \;=\; b, \quad \text{with} \quad b \sim \mathbb{S}^{n-1}.$$

Then, for every $\delta \in [0,1)$, we obtain one of the two outcomes below:

1. a direction of negative curvature $\leq -\epsilon_H/2$,
2. a certificate that $\nabla^2 f(x_k) \succeq -\epsilon_H I$,

using at most $\tilde{\mathcal{O}}\left(\min\{n, \epsilon_H^{-1/2}\}\right)$ gradients/Hessian-vector products, with probability at least $1-\delta$.

Complexity results

First-order deterministic complexity

With $\epsilon_H = \epsilon_g^{1/2}$, reaches x_k such that $\|\nabla f(x_k)\| \leq \epsilon_g$ in at most

- $\mathcal{O}(\epsilon_g^{-3/2})$ iterations;
- $\tilde{\mathcal{O}}\left(\min\{n\epsilon_g^{-3/2}, \epsilon_g^{-7/4}\}\right)$ gradients/Hessian-vector products.

Complexity results

First-order deterministic complexity

With $\epsilon_H = \epsilon_g^{1/2}$, reaches x_k such that $\|\nabla f(x_k)\| \le \epsilon_g$ in at most

- $\mathcal{O}(\epsilon_g^{-3/2})$ iterations;
- $\tilde{\mathcal{O}}\left(\min\{n\epsilon_g^{-3/2}, \epsilon_g^{-7/4}\}\right)$ gradients/Hessian-vector products.

Second-order high probability result

In addition to the results above, we also have

$$\lambda_{\min}(\nabla^2 f(x_k)) \ge -\epsilon_g^{1/2}$$

with probability at least $(1-\delta)^{\mathcal{O}(\epsilon_g^{-3/2})}$.

Complexity results

First-order deterministic complexity

With $\epsilon_H = \epsilon_g^{1/2}$, reaches x_k such that $\|\nabla f(x_k)\| \le \epsilon_g$ in at most

- $\mathcal{O}(\epsilon_g^{-3/2})$ iterations;
- $\tilde{\mathcal{O}}\left(\min\{n\epsilon_g^{-3/2}, \epsilon_g^{-7/4}\}\right)$ gradients/Hessian-vector products.

Second-order high probability result

In addition to the results above, we also have

$$\lambda_{\min}(\nabla^2 f(x_k)) \ge -\epsilon_g^{1/2}$$

with probability at least $(1-\delta)^{\mathcal{O}(\epsilon_g^{-3/2})}$.

- Sharp in terms of iteration complexity (Cartis, Gould, Toint 2018);
- Best know computational complexity for second-order methods.

Numerical illustration

Back to our low-rank matrix problem

$$\min_{U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r}} \frac{1}{2} \left\| P_{\Omega}(UV^{\top} - M) \right\|_F^2,$$

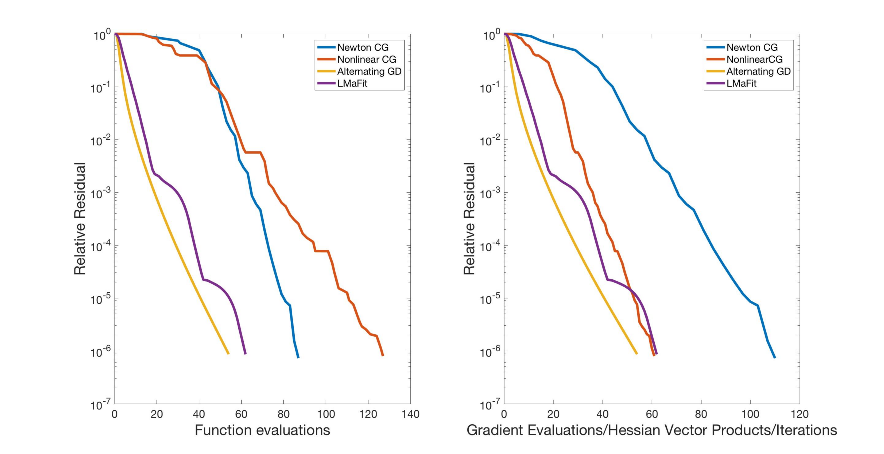
with $M \in \mathbb{R}^{m \times n}$, $|\Omega| \approx \{5\%, 15\%\} \times mn$.

- Synthetic data: $(n, m) = (500, 499)$.

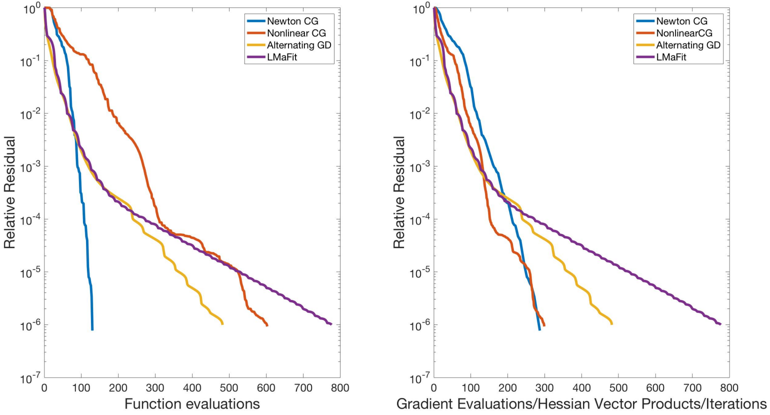
Comparison

- First-order Newton-Capped CG;
- Nonlinear CG (Polak-Ribière);
- Dedicated solvers (Alternating methods):
  - Alternated gradient descent (Tanner and Wei 2016);
  - LMaFit (Wen et al. 2012).

Matrix completion (synthetic data, rank 5)



Matrix completion (synthetic data, rank 15)



Conclusion

CG and nonconvex quadratics

- Can detect negative curvature in probability;
- Can detect nonconvexity!
- **Keys:** Regularization+Extra checks.

Newton-CG methods

- Best known complexity guarantees;
- Works with line search/trust region framework.

\+ Extensions to constraints.

\+ Specialization to matrix problems.

Follow-ups

Other practical variants

- Nonlinear CG;
- Other linear algebra routines (Newton-MR);
- Key: Dealing with negative curvature.

Better algorithms

- Can we do even better than $\epsilon^{-7/4}$?
- With something that we can implement?

Some references

- N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan and T. Ma, *Finding approximate local minima faster than gradient descent*, ACM-SIGACT Symposium on Theory of Computing (STOC), 2017.
- S. Burer and R. D. C. Monteiro, *A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization*, Mathematical Programming, 2003.
- Y. Carmon and J. C. Duchi, *First-order methods for nonconvex quadratic minimization*, SIAM Review, 2020.
- Y. Carmon, J. C. Duchi, O. Hinder and A. Sidford, *Convex until proven guilty: dimension-free acceleration of gradient descent on non-convex functions*, International Conference on Machine Learning (ICML), 2017.
- C. Cartis, N. I. M. Gould and Ph. L. Toint, *On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization*, SIAM Journal on Optimization, 2010.
- C. Cartis, N. I. M. Gould and Ph. L. Toint,*Worst-case evaluation complexity and optimality of second-order methods for nonconvex smooth optimization*, International Congress of Mathematicians (ICM), 2018.
- F. E. Curtis, D. P. Robinson, C. W. Royer and S. J. Wright, *Trust-region Newton-CG with strong second-order complexity guarantees for nonconvex optimization*, SIAM Journal on Optimization, 2021.
- R. Ge, C. Jin and Y. Zheng, *No spurious local minima in nonconvex low rank problems: A unified geometric analysis*, International Conference on Machine Learning (ICML), 2017.
- N.I.M. Gould and V. Simoncini, *Error estimates for iterative algorithms for minimizing regularized quadratic subproblems*, Optimization Methods and Software, 2019.
- J. Kuczyński and H. Woźniakowski, *Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start*, SIAM Journal on Matrix Analysis and Applications, 1992.
- C. W. Royer, M. O'Neill and S. J. Wright, *A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization*, Mathematical Programming, 2020.
- C. W. Royer and S. J. Wright, *Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization*, SIAM Journal on Optimization, 2018.

That's all!

Thank you for your attention!

`clement.royer@dauphine.psl.eu`

https://www.lamsade.dauphine.fr/∼croyer/