Yet Another fast variant of Newton's method

Sadok Jerad[1,2] with Serge Gratton [1,2] and Philippe L.Toint[3]

[1]IRIT-APO

[2]Toulouse INP-ENSEEIHT

[3]University of Namur

EUROPT, 24 August 2023, Budapest

Université Fédérale Toulouse Midi-Pyrénées

ANITI ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE

IRIT Institut de Recherche en Informatique de Toulouse CNRS - INP - UT3 - UT1 - UT2J

Table of Contents

1. Intdroduction
2. Adaptive Newton Negative Curvature
3. Scalable AN2C
    - Definite step
    - Krylov Subspace
4. Numerical experiments
5. Conclusion

Problem motivation

Approximately solving the non convex problem

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

with a 'fast', adaptive convergent algorithm.

Problem motivation

Approximately solving the non convex problem

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

with a 'fast', adaptive convergent algorithm.
Gradient descent (Since Cauchy Lemarechal [2012])

$$x_{k+1} = x_k - \gamma_k \nabla^1 f(x_k),$$

where γ_k can be chosen adaptively (LineSearch : Armijo [1966], Adaptive : Duchi, Hazan, and Singer [2011]; McMahan and Streeter [2010])

Advantages

- Cheap cost.
- Workhorse of modern ML.

Problem motivation

Approximately solving the non convex problem

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

with a 'fast', adaptive convergent algorithm.
Gradient descent (Since Cauchy Lemarechal [2012])

$$x_{k+1} = x_k - \gamma_k \nabla^1 f(x_k),$$

where γ_k can be chosen adaptively (LineSearch : Armijo [1966], Adaptive : Duchi, Hazan, and Singer [2011]; McMahan and Streeter [2010])

Advantages

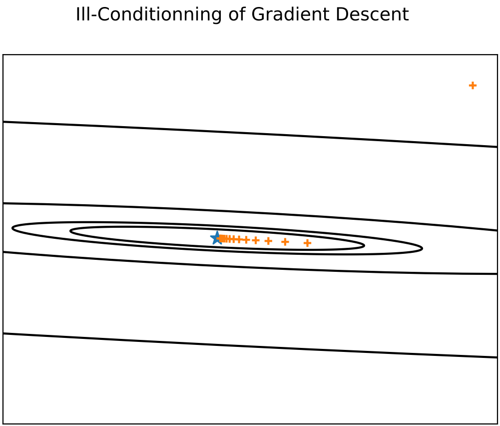
- Cheap cost.
- Workhorse of modern ML.

For deterministic optimization :

- Slow convergence rate (e.g : $\mathcal{O}(\epsilon^{-2})$ to $\|\nabla^1 f(x_\epsilon)\| \leq \epsilon$).
- Ill conditioning.

Illustrative example

Gradient Descent on a 2D quadratic



$\rightarrow$ Exploits second-order information

Newton's method

$$\text{Minimizes local quadratic : } s^\intercal \nabla^1 f(x_k) + \frac{1}{2} s^\intercal \nabla^2 f(x_k) s$$

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

- Super-fast convergence locally [Bertsekas, 1995].
- Only need to solve a linear system.
- Effecient approximation of the Hessian (Quasi-Newton) [Nocedal and Wright, 2006].

Newton's method

Minimizes local quadratic : $s^\intercal \nabla^1 f(x_k) + \frac{1}{2} s^\intercal \nabla^2 f(x_k) s$

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

- Super-fast convergence locally [Bertsekas, 1995].
- Only need to solve a linear system.
- Effecient approximation of the Hessian (Quasi-Newton) [Nocedal and Wright, 2006].

Still needs to be globalized :

- Saddle point (non-convexity).
- No guaranteed decrease.

Newton's method

Minimizes local quadratic : $s^\intercal \nabla^1 f(x_k) + \frac{1}{2} s^\intercal \nabla^2 f(x_k) s$

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

- Super-fast convergence locally [Bertsekas, 1995].
- Only need to solve a linear system.
- Effecient approximation of the Hessian (Quasi-Newton) [Nocedal and Wright, 2006].

Still needs to be globalized :

- Saddle point (non-convexity).
- No guaranteed decrease.

Measure of first-order stationarity :

$$\boxed{\|\nabla f(x_\epsilon)\| \leq \epsilon.}$$

Vanilla Nonconvex Second-Order

Newton LineSearch :

$$x_{k+1} = x_k - \gamma_k (\nabla^2 f(x_k))^{-1} \nabla^1 f(x_k)$$

γ_k is adapted via a Line-Search.
Only need to solve a system.
$\mathcal{O}(\epsilon^{-2})$ rate [Cartis, Gould, and Toint, 2022b].

Trust-region : Optimize for $\|s\| \leq \delta_k$

$$s^\intercal \nabla^1 f(x_k) + \frac{1}{2} s^\intercal \nabla^2 f(x_k) s$$

where δ_k is updated to ensure convergence.
Efficient Algorithms [Conn et al., 2000, 6] (e.g : Steihaug-Toint).
For standard TR, $\mathcal{O}(\epsilon^{-2})$ rate as shown [Cartis et al., 2022b].

Cubic Regularization

Griewank [1981] and Nesterov and Polyak [2006] propose to minimize (exactly) the following cubic model :

$$m_k(s) = s^\intercal \nabla^1 f(x_k) + \frac{1}{2} s^\intercal \nabla^2 f(x_k) s + \frac{\sigma_k}{6} \|s\|^3$$

with σ_k bigger than a specific constant and set $x_{k+1} = x_k + s_k$.
Enjoys $\mathcal{O}\left(\epsilon^{-3/2}\right)$ [Nesterov and Polyak, 2006] complexity rate.

Cubic Regularization

Griewank [1981] and Nesterov and Polyak [2006] propose to minimize (exactly) the following cubic model :

$$m_k(s) = s^\intercal \nabla^1 f(x_k) + \frac{1}{2} s^\intercal \nabla^2 f(x_k) s + \frac{\sigma_k}{6} \|s\|^3$$

with σ_k bigger than a specific constant and set $x_{k+1} = x_k + s_k$.
Enjoys $\mathcal{O}\left(\epsilon^{-3/2}\right)$ [Nesterov and Polyak, 2006] complexity rate. Extensions:

- Adaptive (ARC) Cartis et al. [2011a,b], adaptive without function values [Gratton, J, and Toint, 2023a].
- Inexact Xu et al. [2019], on manifold Agarwal et al. [2020], probabilistic Bellavia et al. [2022]; Cartis and Scheinberg [2017], ...

Cubic Regularization

Griewank [1981] and Nesterov and Polyak [2006] propose to minimize (exactly) the following cubic model :

$$m_k(s) = s^\intercal \nabla^1 f(x_k) + \frac{1}{2} s^\intercal \nabla^2 f(x_k) s + \frac{\sigma_k}{6} \|s\|^3$$

with σ_k bigger than a specific constant and set $x_{k+1} = x_k + s_k$.
Enjoys $\mathcal{O}\left(\epsilon^{-3/2}\right)$ [Nesterov and Polyak, 2006] complexity rate. Extensions:

- Adaptive (ARC) Cartis et al. [2011a,b], adaptive without function values [Gratton, J, and Toint, 2023a].
- Inexact Xu et al. [2019], on manifold Agarwal et al. [2020], probabilistic Bellavia et al. [2022]; Cartis and Scheinberg [2017], ...

First-order stationarity condition on the model

$$\nabla^1 f(x_k) + \nabla^2 f(x_k) s_k + \frac{\sigma_k}{2} \|s_k\| s_k = 0$$

Implicit system:

$$s_k = -(\nabla^2 f(x_k) + \frac{\sigma_k}{2}\|s_k\| I_n)^{-1} \nabla^1 f(x_k).$$

Gradient Descent proposed in Carmon and Duchi [2019] or Lanczos process Cartis et al. [2011a].
Recently, Mishchenko [2023] and Doikov and Nesterov [2023] proposed for the **convex** case

$$s_k = -(\nabla^2 f(x_k) + \sqrt{\sigma_k \|\nabla^1 f(x_k)\|})^{-1} \nabla^1 f(x_k) \qquad (1),$$

with good initial numerical results Mishchenko [2023].

Implicit system:

$$s_k = -(\nabla^2 f(x_k) + \frac{\sigma_k}{2}\|s_k\| I_n)^{-1}\nabla^1 f(x_k).$$

Gradient Descent proposed in Carmon and Duchi [2019] or Lanczos process Cartis et al. [2011a].
Recently, Mishchenko [2023] and Doikov and Nesterov [2023] proposed for the **convex** case

$$s_k = -(\nabla^2 f(x_k) + \sqrt{\sigma_k\|\nabla^1 f(x_k)\|})^{-1}\nabla^1 f(x_k) \qquad (1),$$

with good initial numerical results Mishchenko [2023].

Can we extend (1) to the non-convex case? Devise an algorithm

- Uses (1) when convex.
- Convergence rate close to $\mathcal{O}\left(\epsilon^{-3/2}\right)$.
- Scalable implementation.

Table of Contents

1 Intdroduction

2 Adaptive Newton Negative Curvature

3 Scalable AN2C
- Definite step
- Krylov Subspace

4 Numerical experiments

5 Conclusion

Denote by $g_k \stackrel{\text{def}}{=} \nabla^1 f(x_k)$ and $H_k \stackrel{\text{def}}{=} \nabla^2 f(x_k)$. Recall for convex:

$$s_k = -(H_k + \sqrt{\sigma_k \|g_k\|} I_n)^{-1} g_k.$$

AN2C

Denote by $g_k \stackrel{\text{def}}{=} \nabla^1 f(x_k)$ and $H_k \stackrel{\text{def}}{=} \nabla^2 f(x_k)$. Recall for convex:

$$s_k = -(H_k + \sqrt{\sigma_k \|g_k\|} I_n)^{-1} g_k.$$

A natural extension for the nonconvex case,

$$s_k = -(H_k + (\sqrt{\sigma_k \|g_k\|} + \max[0, -\lambda_{\min}(H_k)]) I_n)^{-1} g_k$$

Denote by $g_k \stackrel{\text{def}}{=} \nabla^1 f(x_k)$ and $H_k \stackrel{\text{def}}{=} \nabla^2 f(x_k)$. Recall for convex:

$$s_k = -(H_k + \sqrt{\sigma_k \|g_k\|} I_n)^{-1} g_k.$$

A natural extension for the nonconvex case,

$$s_k = -(H_k + (\sqrt{\sigma_k \|g_k\|} + \max[0, -\lambda_{\min}(H_k)]) I_n)^{-1} g_k$$

Problem when $-\lambda_{\min}(H_k)$ is large $\rightarrow$ step too small.

AN2C

Denote by $g_k \stackrel{\text{def}}{=} \nabla^1 f(x_k)$ and $H_k \stackrel{\text{def}}{=} \nabla^2 f(x_k)$. Recall for convex:

$$s_k = -(H_k + \sqrt{\sigma_k \|g_k\|} I_n)^{-1} g_k.$$

A natural extension for the nonconvex case,

$$s_k = -(H_k + (\sqrt{\sigma_k \|g_k\|} + \max[0, -\lambda_{\min}(H_k)]) I_n)^{-1} g_k$$

Problem when $-\lambda_{\min}(H_k)$ is large $\rightarrow$ step too small.

$\rightarrow$ Use negative curvature instead.

Hence, **AN2C** (Adaptive Newton with Negative Curvature) denomination.

AN2C Algorithm presentation

Algorithm 1: AN2C Algorithm

Input: x_0, $\epsilon \in (0, 1]$, $\sigma_0, \sigma_{\min} > 0$, $0 < \eta_1 < \eta_2 < 1$, $0 < \gamma_1 < 1 < \gamma_2 \leq \gamma_3$, $\kappa_C > 0$ and $k = 0$.

Output: x_ϵ: an approximate first order point.

while $\|g_k\| > \epsilon$ **do**

Evaluate g_k and H_k.

Compute $\lambda_{\min}(H_k)$. If $\lambda_{\min}(H_k) \geq -\kappa_C \sqrt{\sigma_k \|g_k\|}$, compute s_k^{neig}

$$s_k^{neig} = -(H_k + (\sqrt{\sigma_k \|g_k\|} + \max[0, -\lambda_{\min}(H_k)]) I_n)^{-1} g_k.$$

AN2C Algorithm presentation

Algorithm 2: AN2C Algorithm

Input: x_0, $\epsilon \in (0, 1]$, $\sigma_0, \sigma_{\min} > 0$, $0 < \eta_1 < \eta_2 < 1$, $0 < \gamma_1 < 1 < \gamma_2 \leq \gamma_3$, $\kappa_C > 0$ and $k = 0$.

Output: x_ϵ: an approximate first order point.

while $\|g_k\| > \epsilon$ **do**

Evaluate g_k and H_k.

Compute $\lambda_{\min}(H_k)$. If $\lambda_{\min}(H_k) \geq -\kappa_C \sqrt{\sigma_k \|g_k\|}$, compute s_k^{neig}

$$s_k^{neig} = -(H_k + (\sqrt{\sigma_k \|g_k\|} + \max[0, -\lambda_{\min}(H_k)]) I_n)^{-1} g_k.$$

Else $\lambda_{\min}(H_k) \leq -\kappa_C \sqrt{\sigma_k \|g_k\|}$,

$$g_k^\intercal u_k \leq 0, \|u_k\| = 1, H_k u_k = \lambda_{\min}(H_k) u_k \quad s_k^{curv} = \frac{\kappa_C \sqrt{\sigma_k \|g_k\|}}{\sigma_k} u_k.$$

Compute $\rho_k \leftarrow \frac{f(x_k) - f(x_k + s_k)}{-(g_k^\intercal s_k + \frac{1}{2} s_k^\intercal H_k s_k)}$ If $\rho_k \geq \eta_1$, $x_{k+1} \leftarrow x_k + s_k$. Else, $x_{k+1} \leftarrow x_k$.

Update σ_k with the values of η_1, η_2, γ_1, γ_2 and γ_3 as in [Cartis et al., 2022a, Algorithm 3.3.1].

$k \leftarrow k + 1$.

Comments on AN2C

- The $\sqrt{\sigma_k \|g_k\|} + \max[0, -\lambda_{\min}(H_k)]$ resembles the GQT method [Goldfeld, Quandt, and Trotter, 1966], $\mathcal{O}\left(\epsilon^{-2}\right)$ proven by Ueda and Yamashita [2014].
- Birgin and Martínez [2017] regularizes with $\max[0, -\lambda_{\min}(H_k)] + \mu$ and test multiple μ's to ensure $\mathcal{O}\left(\epsilon^{\frac{-3}{2}}\right)$ via 'cubic' descent.
- Negative curvature directions + gradient related ones. See Curtis and Robinson [2018]; Ferris et al. [1996]; Goldfarb [1980]; Gould et al. [2000].

Our algorithm enjoys a $\mathcal{O}\left(|\log \epsilon|\epsilon^{-\frac{3}{2}}\right)$ complexity rate.

Comments on AN2C

- The $\sqrt{\sigma_k \|g_k\|} + \max[0, -\lambda_{\min}(H_k)]$ resembles the GQT method [Goldfeld, Quandt, and Trotter, 1966], $\mathcal{O}\left(\epsilon^{-2}\right)$ proven by Ueda and Yamashita [2014].
- Birgin and Martínez [2017] regularizes with $\max[0, -\lambda_{\min}(H_k)] + \mu$ and test multiple μ's to ensure $\mathcal{O}\left(\epsilon^{\frac{-3}{2}}\right)$ via 'cubic' descent.
- Negative curvature directions + gradient related ones. See Curtis and Robinson [2018]; Ferris et al. [1996]; Goldfarb [1980]; Gould et al. [2000].

Our algorithm enjoys a $\mathcal{O}\left(|\log \epsilon|\epsilon^{-\frac{3}{2}}\right)$ complexity rate.

⚠ What about $\lambda_{\min}(H_k)$? Exact solve ?

Scalable variants will be discussed 😄.

Notations, bound on σ_k

We introduce

$$\mathcal{S}_k \stackrel{\text{def}}{=} \{0 \leq j \leq k |\, x_{j+1} = x_j + s_j, \}, \quad \mathcal{S}_k^{neig} \stackrel{\text{def}}{=} \left\{ j \in \mathcal{S}_k | s_j = s_j^{neig} \right\},$$

$$\mathcal{S}_k^{curv} \stackrel{\text{def}}{=} \left\{ j \in \mathcal{S}_k | s_j = s_j^{curv} \right\}.$$

As

$$\sigma_{k+1} \in \begin{cases} [\max(\sigma_{\min}, \gamma_1 \sigma_k), \sigma_k] & \text{if } \rho_k \geq \eta_2, \\ [\sigma_k, \gamma_2 \sigma_k] & \text{if } \rho_k \in [\eta_1, \eta_2), \\ [\gamma_2 \sigma_k, \gamma_3 \sigma_k] & \text{if } \rho_k < \eta_1. \end{cases}$$

From standard analysis [Cartis et al., 2022a, Lemma 2.4.1], only need a bound $\sigma_{\max}$ and $|\mathcal{S}_k|$.

AN2C

Assumptions

AS.1 f is a twice differentiable function.

AS.2 $\nabla^2 f$ is a Lipschitz function with constant L_H.

AS.3 f is lower bounded by f_{low}.

AS.4 $\lambda_{\min}(\nabla^2 f(x_k)) \geq -\kappa_H$.

Properties on when s_k^{neig}

$$\|s_k^{neig}\| \leq \sqrt{\frac{\|g_k\|}{\sigma_k}}, \quad -g_k^\intercal s_k^{neig} - \frac{1}{2} s_k^{neig} H_k s_k^{neig} \geq \sqrt{\sigma_k \|g_k\|} \|s_k^{neig}\|^2.$$

On s_k^{curv}

$$-g_k^\intercal s_k^{curv} - \frac{1}{2} s_k^{curv} H_k s_k^{curv} \geq \frac{1}{2} \sigma_k \|s_k^{curv}\|^3 = \frac{\eta_1 \kappa_C^3}{2\sqrt{\sigma_k}} \|g_k\|^{\frac{3}{2}}.$$

Decrease comes from step property, not from model as
in ARC Cartis et al. [2011a,b].

Bound on $\sigma +$ Elements

Bound on $\sigma_{\max}$.

From the local quadratic decrease $g_k^\intercal s_k + \frac{1}{2} s_k^\intercal H_k s_k$ in both cases + Lipschitz smoothness + σ_k update mechanism

$$\sigma_k \leq \sigma_{\max}.$$

Proof similar to ARp methods [Birgin, Gardenghi, Martínez, Santos, and Toint, 2016a].

Bound on $\sigma+$ Elements

Bound on $\sigma_{\max}$.
From the local quadratic decrease $g_k^\intercal s_k + \frac{1}{2} s_k^\intercal H_k s_k$ in both cases + Lipschitz smoothness + σ_k update mechanism

$$\sigma_k \leq \sigma_{\max}.$$

Proof similar to ARp methods [Birgin, Gardenghi, Martínez, Santos, and Toint, 2016a].
$k \in \mathcal{S}_k^{curv}$, guaranteed decrease

$$f(x_k) - f(x_{k+1}) \geq \eta_1(-g_k^\intercal s_k - \frac{1}{2} s_k^\intercal H_k s_k) \geq \frac{\eta_1 \kappa_C^3}{2\sqrt{\sigma_k}} \|g_k\|^{\frac{3}{2}} \geq \frac{\eta_1 \kappa_C^3}{2\sqrt{\sigma_{\min}}} \epsilon^{\frac{3}{2}}.$$

Decrease of required order. As $f(x) \geq f_{\text{low}}$, $|\mathcal{S}_k^{curv}| \leq \mathcal{O}\left(\epsilon^{-\frac{3}{2}}\right)$.
Lipschitz gradient error + negative curvature boundness and $\|s_k^{curv}\|$ bound

$$\|g_{k+1}\| \leq \left(\frac{L_H}{2\sigma_k} \kappa_C^2 + \frac{\kappa_H \kappa_C}{\sqrt{\epsilon \sigma_k}} + 1\right) \|g_k\|.$$

Newton decrease step

Lower bound required on $\|s_k^{neig}\|$. $k \in \mathcal{S}_k^{neig}$

$$
\begin{aligned}
\|g_{k+1}\| &\leq \|g_{k+1} - g_k - H_k s_k\| + \|g_k + H_k s_k\| \\
&= \|g_{k+1} - g_k - H_k s_k\| + \|(\sqrt{\sigma_k \|g_k\|} + \max[0, -\lambda_{\min}(H_k)]) s_k\| \\
&\leq \frac{L_H}{2}\|s_k\|^2 + (1 + \kappa_C)\sqrt{\sigma_k \|g_k\|}\|s_k\|,
\end{aligned}
$$

Newton decrease step

Lower bound required on $\|s_k^{neig}\|$. $k \in \mathcal{S}_k^{neig}$

$$\begin{aligned}
\|g_{k+1}\| &\leq \|g_{k+1} - g_k - H_k s_k\| + \|g_k + H_k s_k\| \\
&= \|g_{k+1} - g_k - H_k s_k\| + \|(\sqrt{\sigma_k \|g_k\|} + \max[0, -\lambda_{\min}(H_k)]) s_k\| \\
&\leq \frac{L_H}{2}\|s_k\|^2 + (1+\kappa_C)\sqrt{\sigma_k\|g_k\|}\|s_k\|,
\end{aligned}$$

$$\|s_k\| \geq \frac{-(1+\kappa_C)\sqrt{\sigma_k\|g_k\|} + \sqrt{(1+\kappa_C)^2\sigma_k\|g_k\| + 2L_H\|g_{k+1}\|}}{L_H}.$$

Newton decrease step

Lower bound required on $\|s_k^{neig}\|$. $k \in \mathcal{S}_k^{neig}$

$$
\begin{aligned}
\|g_{k+1}\| &\leq \|g_{k+1} - g_k - H_k s_k\| + \|g_k + H_k s_k\| \\
&= \|g_{k+1} - g_k - H_k s_k\| + \|(\sqrt{\sigma_k \|g_k\|} + \max[0, -\lambda_{\min}(H_k)])s_k\| \\
&\leq \frac{L_H}{2}\|s_k\|^2 + (1+\kappa_C)\sqrt{\sigma_k \|g_k\|}\|s_k\|,
\end{aligned}
$$

$$
\|s_k\| \geq \frac{-(1+\kappa_C)\sqrt{\sigma_k\|g_k\|} + \sqrt{(1+\kappa_C)^2\sigma_k\|g_k\| + 2L_H\|g_{k+1}\|}}{L_H}.
$$

When $\|g_{k+1}\| \ll \|g_k\|$, last bound becomes uninformative.

Newton steps division

Inspired by analysis of Mishchenko [2023], divide $\mathcal{S}_k^{neig}$ in two subsets.

$$\mathcal{S}_k^{divgrad} \stackrel{\text{def}}{=} \{k \in \mathcal{S}_k^{neig}, 2\kappa_m L_H \|g_{k+1}\| < \sigma_k \|g_k\|\}$$

where $\kappa_m = \frac{\sigma_{\max}}{L_H}$ depends only on problem constant.

$$\mathcal{S}_k^{decr} \stackrel{\text{def}}{=} \mathcal{S}_k^{neig} \setminus \mathcal{S}_k^{divgrad},$$

Newton steps division

Inspired by analysis of Mishchenko [2023], divide $\mathcal{S}_k^{neig}$ in two subsets.

$$\mathcal{S}_k^{divgrad} \stackrel{\text{def}}{=} \{k \in \mathcal{S}_k^{neig}, 2\kappa_m L_H \|g_{k+1}\| < \sigma_k \|g_k\|\}$$

where $\kappa_m = \frac{\sigma_{\max}}{L_H}$ depends only on problem constant.

$$\mathcal{S}_k^{decr} \stackrel{\text{def}}{=} \mathcal{S}_k^{neig} \setminus \mathcal{S}_k^{divgrad},$$

from $k \in \mathcal{S}_k^{decr}$ and $\|s_k^{neig}\|$ lower bound,

$$f(x_k) - f(x_{k+1}) \geq \kappa(\sigma_k \|g_k\|)^{\frac{3}{2}} \geq \mathcal{O}\left(\epsilon^{\frac{3}{2}}\right) \to |\mathcal{S}_k^{decr}| \leq \mathcal{O}\left(\epsilon^{\frac{-3}{2}}\right).$$

For $k \in \mathcal{S}_k^{divgrad}$,

$$\|g_{k+1}\| \leq \frac{\|g_k\|}{2}.$$

Bound on $|\mathcal{S}_k|$ and iterations number

$$|\mathcal{S}_k^{divgrad}| \leq \kappa_n |\mathcal{S}_k^{decr}| + \left(\frac{|\log \epsilon|}{2\log 2} + \kappa_{curv}\right)|\mathcal{S}_k^{curv}| + \frac{|\log(\epsilon)| + \log(\|g_0\|)}{\log 2} + 1$$

Idea : Since $\frac{\epsilon}{\|g_0\|} \leq \prod_{i\in\mathcal{S}_k} \frac{\|g_{i+1}\|}{\|g_i\|}$ with use bounds $\|g_{k+1}\|/\|g_k\|$ for three cases.

Bound on $|\mathcal{S}_k|$ and iterations number

$$|\mathcal{S}_k^{divgrad}| \leq \kappa_n |\mathcal{S}_k^{decr}| + \left(\frac{|\log \epsilon|}{2\log 2} + \kappa_{curv}\right)|\mathcal{S}_k^{curv}| + \frac{|\log(\epsilon)| + \log(\|g_0\|)}{\log 2} + 1$$

Idea : Since $\frac{\epsilon}{\|g_0\|} \leq \prod_{i\in\mathcal{S}_k} \frac{\|g_{i+1}\|}{\|g_i\|}$ with use bounds $\|g_{k+1}\|/\|g_k\|$ for three cases.
Since $|\mathcal{S}_k| = |\mathcal{S}_k^{divgrad}| + |\mathcal{S}_k^{curv}| + |\mathcal{S}_k^{decr}|$

$$|\mathcal{S}_k| \leq \mathcal{O}\left(|\log \epsilon|\epsilon^{-3/2}\right)$$

and required iterations k

$$k \leq |\mathcal{S}_k|\left(1 + \frac{|\log \gamma_1|}{\log \gamma_2}\right) + \frac{1}{\log \gamma_2}\log\left(\frac{\sigma_{\max}}{\sigma_0}\right).$$

Comments

- Up to a $|\log \epsilon|$ from the optimal rate of second-order methods as shown Carmon et al. [2019].
- $|\log \epsilon|$ comes from negative curvature steps. Rare in practice.
- Retrieve same rate as the Truncated Newton-CG of Curtis et al. [2021].
- Better than past hybrid algorithms of [Curtis and Robinson, 2018], [Liu et al., 2018] and [Goldfeld et al., 1966].

Comments

- Up to a $|\log \epsilon|$ from the optimal rate of second-order methods as shown Carmon et al. [2019].
- $|\log \epsilon|$ comes from negative curvature steps. Rare in practice.
- Retrieve same rate as the Truncated Newton-CG of Curtis et al. [2021].
- Better than past hybrid algorithms of [Curtis and Robinson, 2018], [Liu et al., 2018] and [Goldfeld et al., 1966].

How to devise scalable variant ?

Table of Contents

1 Intdroduction

2 Adaptive Newton Negative Curvature

3 Scalable AN2C
- Definite step
- Krylov Subspace

4 Numerical experiments

5 Conclusion

Improvements

Goals : Avoid the computation of $\lambda_{\min}(H_k)$ and inexact resolution.

Improvements

Goals : Avoid the computation of $\lambda_{\min}(H_k)$ and inexact resolution.

Denote by

$$r_k^{neig} = (H_k + (\sqrt{\sigma_k \|g_k\|} + \max[0, -\lambda_{\min}(H_k)])I_n)s_k^{neig} + g_k$$

To compute s_k^{neig} only needs

$$\|r_k^{neig}\| \leq \min(\varsigma\sqrt{\sigma_k\|g_k\|}\|s_k^{neig}\|, \kappa_\theta\|g_k\|),$$

with $\varsigma < 1$.

Goals : Avoid the computation of $\lambda_{\min}(H_k)$ and inexact resolution.
Denote by

$$r_k^{neig} = (H_k + (\sqrt{\sigma_k \|g_k\|} + \max[0, -\lambda_{\min}(H_k)])I_n)s_k^{neig} + g_k$$

To compute s_k^{neig} only needs

$$\|r_k^{neig}\| \leq \min(\varsigma\sqrt{\sigma_k\|g_k\|}\|s_k^{neig}\|, \kappa_\theta\|g_k\|),$$

with $\varsigma < 1$.

Comments

- $\kappa_\theta\|g_k\|$ standard for CG methods.
- $\varsigma\sqrt{\sigma_k\|g_k\|}\|s_k^{neig}\|$ mimics the condition on the cubic model gradient $\|\nabla_s^1 m_k(s_k)\| \leq \mathcal{O}\left(\|s_k\|^2\right)$, see [Curtis, Robinson, Royer, and Wright, 2021].

Definite step

Use only the gradient to "compensate" for the negative curvature. Consider $\kappa_a \geq 1$ and attempt to solve the following linear system

$$(H_k + \sqrt{\kappa_a \sigma_k \|g_k\|}) s_k^{def} = -g_k$$

to find a solution such that

$$(s_k^{def})^\intercal (H_k + \sqrt{\kappa_a \sigma_k \|g_k\|}) s_k^{def} > 0,$$

$$\|s_k^{def}\| \leq \kappa_\theta \sqrt{\frac{\|g_k\|}{\kappa_a \sigma_k}}.$$

Definite step

Use only the gradient to "compensate" for the negative curvature. Consider $\kappa_a \geq 1$ and attempt to solve the following linear system

$$(H_k + \sqrt{\kappa_a \sigma_k \|g_k\|})s_k^{def} = -g_k$$

to find a solution such that

$$(s_k^{def})^\intercal (H_k + \sqrt{\kappa_a \sigma_k \|g_k\|})s_k^{def} > 0,$$

$$\|s_k^{def}\| \leq \kappa_\theta \sqrt{\frac{\|g_k\|}{\kappa_a \sigma_k}}.$$

Comments

- Adds a new kind of step.
- First condition ensures that we have "sufficiently" regularized.
- Second condition ensures that the step is bounded w.r.t the gradient.
- Inexact resolution is possible.

Comments on the definite step

- Can use a factorization or iterative conjugate gradient to check the definiteness. See 'Capped-CG' subroutine of Royer, O'Neill, and Wright [2019]
- May slow numerical efficiency. For convex subregions, $\sqrt{\sigma_k \|g_k\|}$ is the "right" regularization term [Doikov and Nesterov, 2023; Mishchenko, 2023].
- Still needs to compute $\lambda_{\min}(H_k)$ if the step fails.

Propose subspace version of the exact AN2C.

1. Intdroduction
2. Adaptive Newton Negative Curvature
3. Scalable AN2C
   - Definite step
   - Krylov Subspace
4. Numerical experiments
5. Conclusion

Subspaces Methods

Project the current derivatives into a "well-behaved" subspace. **We extend our AN2C to propose a AN2CK**

Subspaces Methods

Project the current derivatives into a "well-behaved" subspace. **We extend our AN2C to propose a AN2CK**

Krylov methods.

Structured nested sequence of subspaces. See [Conn, Gould, and Toint, 2000; Nocedal and Wright, 2006] for trust-region variant and Newton-CG algorithm.

Condition on the subspace

Consider $p \in \{1, \ldots, n\}$ and

$$V_p \in \mathbb{R}^{n \times p}, \quad \widehat{g}_k \stackrel{\text{def}}{=} V_p^\intercal g_k \in \mathbb{R}^p, \quad \widehat{H}_k \stackrel{\text{def}}{=} V_p^\intercal H_k V_p \in \mathbb{R}^{p \times p}.$$

where

$$\|V_p\| \leq V_{\max} \text{ for all } p \in \{1, \ldots, n\}.$$

Condition on the subspace

Consider $p \in \{1, \dots, n\}$ and

$$V_p \in \mathbb{R}^{n \times p}, \quad \widehat{g}_k \stackrel{\text{def}}{=} V_p^\intercal g_k \in \mathbb{R}^p, \quad \widehat{H}_k \stackrel{\text{def}}{=} V_p^\intercal H_k V_p \in \mathbb{R}^{p \times p}.$$

where

$$\|V_p\| \leq V_{\max} \text{ for all } p \in \{1, \dots, n\}.$$

Let $\kappa_B \geq 1$. If $\lambda_{\min}(\widehat{H}_k) \geq -\kappa_C \sqrt{\sigma_k \|g_k\|}$,

$$(\widehat{H}_k + (\sqrt{\sigma_k \|g_k\|} + \max[0, -\lambda_{\min}(\widehat{H}_k)]) I_p) y_k^{neig} = -\widehat{g}_k.$$

Condition on the subspace

Consider $p \in \{1, \ldots, n\}$ and

$$V_p \in \mathbb{R}^{n \times p}, \quad \widehat{g}_k \stackrel{\text{def}}{=} V_p^\intercal g_k \in \mathbb{R}^p, \quad \widehat{H}_k \stackrel{\text{def}}{=} V_p^\intercal H_k V_p \in \mathbb{R}^{p \times p}.$$

where

$$\|V_p\| \leq V_{\max} \text{ for all } p \in \{1, \ldots, n\}.$$

Let $\kappa_B \geq 1$. If $\lambda_{\min}(\widehat{H}_k) \geq -\kappa_C \sqrt{\sigma_k \|g_k\|}$,

$$(\widehat{H}_k + (\sqrt{\sigma_k \|g_k\|} + \max[0, -\lambda_{\min}(\widehat{H}_k)]) I_p) y_k^{neig} = -\widehat{g}_k.$$

If

$$\|H_k V_p y_k^{neig} + g_k\|_n \leq \kappa_B \|\widehat{H}_k y_k^{neig} + \widehat{g}_k\|_p,$$

set $s_k^{neig} \stackrel{\text{def}}{=} V_p y_k^{neig}$.

Else, consider a new subspace V_p.

Condition on the subspace

Consider $p \in \{1, \dots, n\}$ and

$$V_p \in \mathbb{R}^{n\times p}, \quad \widehat{g}_k \stackrel{\text{def}}{=} V_p^\intercal g_k \in \mathbb{R}^p, \quad \widehat{H}_k \stackrel{\text{def}}{=} V_p^\intercal H_k V_p \in \mathbb{R}^{p\times p}.$$

where

$$\|V_p\| \le V_{\max} \text{ for all } p \in \{1, \dots, n\}.$$

Let $\kappa_B \ge 1$. If $\lambda_{\min}(\widehat{H}_k) \ge -\kappa_C\sqrt{\sigma_k\|g_k\|}$,

$$(\widehat{H}_k + (\sqrt{\sigma_k\|g_k\|} + \max[0, -\lambda_{\min}(\widehat{H}_k)])I_p)y_k^{neig} = -\widehat{g}_k.$$

If

$$\|H_k V_p y_k^{neig} + g_k\|_n \le \kappa_B\|\widehat{H}_k y_k^{neig} + \widehat{g}_k\|_p,$$

set $s_k^{neig} \stackrel{\text{def}}{=} V_p y_k^{neig}$.

Else, consider a new subspace V_p.

Else $\lambda_{\min}(\widehat{H}_k) \le -\kappa_C\sqrt{\sigma_k\|g_k\|}$,

$$\widehat{g}_k^\intercal u_k \le 0, \|u_k\| = 1, \widehat{H_k} u_k = \lambda_{\min}(\widehat{H_k})u_k \quad s_k^{curv} = \frac{\kappa_C\sqrt{\sigma_k\|g_k\|}}{\sigma_k} V_p u_k.$$

Kryolov implementation details

Krylov via Lanczos

We use the Lanczos procedure to generate a sequence of orthonormal bases of the Krylov subspace. $V_{\max} = 1$ and $p = n$ gives a valid step.

Kryolov implementation details

Krylov via Lanczos

We use the Lanczos procedure to generate a sequence of orthonormal bases of the Krylov subspace. $V_{\max} = 1$ and $p = n$ gives a valid step. Exploit

the structure of the specific subproblem

- $\widehat{H}_k$ is a tridiagonal matrix.
- Computation of $\lambda_{\min}(\widehat{H}_k)$ can be performed easily Coakley and Rokhlin [2013].
- V_p can be regenerated at the end as in Gould et al. [2003].
- Preconditionner can be employed.

For more details on Krylov methods, See [Conn et al., 2000, Subsection 5.2]

Recap + Other Results

- Complexity rate $\mathcal{O}\left(|\log\epsilon|\epsilon^{-3/2}\right)$ still valid for all the introduced AN2C variants. See [Gratton, J, and Toint, 2023b].
- Second-order algorithm. $\mathcal{O}\left(|\log\epsilon_1|\max(\epsilon_1^{-3/2},\epsilon_2^{-3})\right)$ iterations to reach (ϵ_1,ϵ_2) second order point.
- Approximate negative curvature can be employed.

Table of Contents

1. Intdroduction
2. Adaptive Newton Negative Curvature
3. Scalable AN2C
   - Definite step
   - Krylov Subspace
4. Numerical experiments
5. Conclusion

Numerical experiments framework

- A first set of 117 small dimensional problems, second set of 74 medium problems and third set of 59 "largish" problems from CUTEst problems (OPM : Matlab library) Gratton and Toint [2021].
- Baselines : standard AR2 of Birgin et al. [2016b] and trust-region methods Conn et al. [2000] with factorized and iterative solver. with $\epsilon = 1\mathrm{e}{-6}$.
- Max number of iteration set out to 5000 or timeout if cputime > 1 hour.
- Results are reported using performance profiles Dolan et al. [2006].

Numerical experiments framework

- A first set of 117 small dimensional problems, second set of 74 medium problems and third set of 59 "largish" problems from CUTEst problems (OPM : Matlab library) Gratton and Toint [2021].
- Baselines : standard AR2 of Birgin et al. [2016b] and trust-region methods Conn et al. [2000] with factorized and iterative solver. with $\epsilon = 1\mathrm{e}{-6}$.
- Max number of iteration set out to 5000 or timeout if cputime $>$ 1 hour.
- Results are reported using performance profiles Dolan et al. [2006].

See [Gratton, J, and Toint, 2023b] for more details.

Variants of AN2C

- $\mathrm{AN2CE}$ for exact computation and $\mathrm{AN2CER}$ where the trial definite step is performed (regularization $\propto 10 * \sqrt{\sigma_k \|g_k\|}$).

- $\mathrm{AN2CKU}$ and $\mathrm{AN2CKYU}$ two variants of the Krylov subspace. First variant : aligned with $\lambda_{\min}(\widehat{H_k})$, second variant : uses past information and fraction of $\lambda_{\min}(\widehat{H_k})$.

Numerical performances small

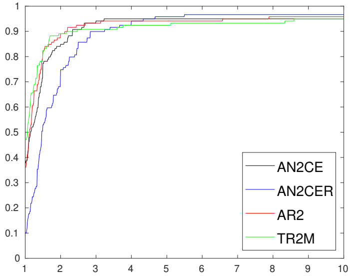


Figure: Exact and Factorized AN2C algorithms and Baselines

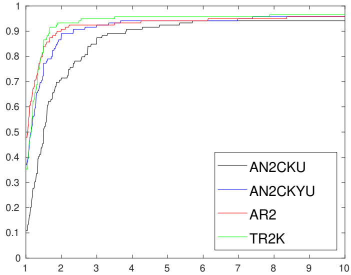


Figure: Two variants of Krylov AN2C and Baselines

Numerical performance medium

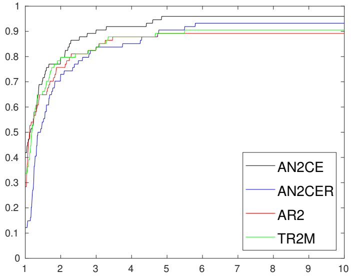


Figure: Exact and Factorized AN2C algorithms and Baselines

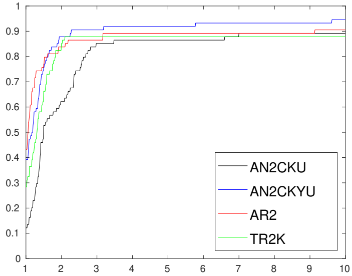


Figure: Two variants of Krylov AN2C and Baselines

Numerical performance large

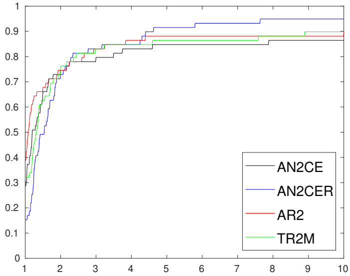


Figure: Exact and Factorized AN2C algorithms and Baselines

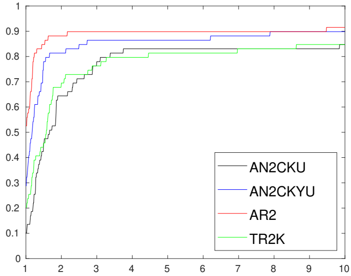


Figure: Two variants of Krylov AN2C and Baselines

Comments

- AN2C methods on par with standard methods. Best variant depends on the problems.
- AN2CE best for small datasets while AN2CER is the 'fastest' for large problems.
- Iterative methods: AN2CKU significantly trails the other variants $\rightarrow$ Does not exploit past informations

Comments

- AN2C methods on par with standard methods. Best variant depends on the problems.
- AN2CE best for small datasets while AN2CER is the 'fastest' for large problems.
- Iterative methods: AN2CKU significantly trails the other variants $\rightarrow$ Does not exploit past informations .
- Usage of negative curvature happens rarely (at most 0.25%). Basically a Newton method.
- When trial definite step is tested, used in at most 93% of case. one solution of a linear system required

More experiments are required to better asses these methods: cpu-time, preconditionner,...

Table of Contents

1 Intdroduction

2 Adaptive Newton Negative Curvature

3 Scalable AN2C
- Definite step
- Krylov Subspace

4 Numerical experiments

5 Conclusion

Conclusion

Recap

- Second order method alternating between Newton and negative curvature (In practice, mostly Newton).
- May require costly negative curvature information.
- Inexact solution + trial step that uses only gradient as a regularization + subspace implementation allow practical variants.
- Optimal complexity up to a log factor.
- On par with standard methods numerically.

Extensions

- β Hölder Hessian for $\beta \in (0, 1]$.

Perspectives

Other subspaces techniques : Random sketching methods [Mahoney, 2011; Woodruff, 2014] ?

Non-Euclidean norm [Doikov and Nesterov, 2023; Gratton and Toint, 2022] ?

For more details

Thank for your attention.

References I

N. Agarwal, N. Boumal, B. Bullins, and C. Cartis. Adaptive regularization with cubics on manifolds. *Mathematical Programming*, 188(1):85–134, May 2020. doi: 10.1007/s10107-020-01505-1. URL https://doi.org/10.1007/s10107-020-01505-1.

L. Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, Jan. 1966.

S. Bellavia, G. Gurioli, B. Morini, and P. L. Toint. Adaptive regularization for nonconvex optimization using inexact function values and randomly perturbed derivatives. *Journal of Complexity*, 68:101591, Feb. 2022. doi: 10.1016/j.jco.2021.101591. URL https://doi.org/10.1016/j.jco.2021.101591.

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, Massachussetts, USA, 1995.

E. G. Birgin and J. M. Martínez. The use of quadratic regularization with a cubic descent condition for unconstrained optimization. *SIAM Journal on Optimization*, 27(2):1049–1074, Jan. 2017. doi: 10.1137/16m110280x. URL https://doi.org/10.1137/16m110280x.

References II

E. G. Birgin, J. L. Gardenghi, J. M. Martínez, S. A. Santos, and P. L. Toint. Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Mathematical Programming*, 163(1-2): 359–368, Aug. 2016a. doi: 10.1007/s10107-016-1065-8.

E. G. Birgin, J. L. Gardenghi, J. M. Martínez, S. A. Santos, and P. L. Toint. Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Mathematical Programming*, 163(1-2): 359–368, Aug. 2016b. doi: 10.1007/s10107-016-1065-8.

Y. Carmon and J. Duchi. Gradient descent finds the cubic-regularized nonconvex newton step. *SIAM Journal on Optimization*, 29(3):2146–2178, Jan. 2019. doi: 10.1137/17m1113898. URL `https://doi.org/10.1137/17m1113898`.

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1-2):71–120, June 2019. doi: 10.1007/s10107-019-01406-y. URL `https://doi.org/10.1007/s10107-019-01406-y`.

References III

C. Cartis and K. Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming*, 169(2):337–375, Apr. 2017. doi: 10.1007/s10107-017-1137-4. URL `https://doi.org/10.1007/s10107-017-1137-4`.

C. Cartis, N. I. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. part i: Motivation, convergence and numerical results. *Math. Program.*, 127(2):245–295, apr 2011a. ISSN 0025-5610.

C. Cartis, N. I. M. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. part II: worst-case function- and derivative-evaluation complexity. *Math. Program.*, 130(2):295–319, 2011b. doi: 10.1007/s10107-009-0337-y. URL `https://doi.org/10.1007/s10107-009-0337-y`.

C. Cartis, N. I. M. Gould, and P. L. Toint. *Evaluation Complexity of Algorithms for Nonconvex Optimization: Theory, Computation and Perspectives*. Society for Industrial and Applied Mathematics, Jan. 2022a. doi: 10.1137/1.9781611976991. URL `https://doi.org/10.1137/1.9781611976991`.

References IV

C. Cartis, N. I. M. Gould, and P. L. Toint. *Evaluation complexity of algorithms for nonconvex optimization*. Number 30 in MOS-SIAM Series on Optimization. SIAM, Philadelphia, USA, June 2022b.

E. S. Coakley and V. Rokhlin. A fast divide-and-conquer algorithm for computing the spectra of real symmetric tridiagonal matrices. *Applied and Computational Harmonic Analysis*, 34(3):379–414, May 2013. doi: 10.1016/j.acha.2012.06.003. URL https://doi.org/10.1016/j.acha.2012.06.003.

A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust-Region Methods*. Number 1 in MOS-SIAM Optimization Series. SIAM, Philadelphia, USA, 2000.

F. E. Curtis and D. P. Robinson. Exploiting negative curvature in deterministic and stochastic optimization. *Mathematical Programming*, 176(1-2):69–94, Oct. 2018. doi: 10.1007/s10107-018-1335-8. URL https://doi.org/10.1007/s10107-018-1335-8.

References V

F. E. Curtis, D. P. Robinson, and M. Samadi. An inexact regularized newton framework with a worst-case iteration complexity of \$ {\mathscr o}(\varepsilon^{-3/2}) \$ for nonconvex optimization. *IMA Journal of Numerical Analysis*, 39(3):1296–1327, May 2018. doi: 10.1093/imanum/dry022. URL https://doi.org/10.1093/imanum/dry022.

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*, 31(1):518–544, Jan. 2021. doi: 10.1137/19m130563x. URL https://doi.org/10.1137/19m130563x.

N. Doikov and Y. Nesterov. Gradient regularization of newton method with bregman distances. *Mathematical Programming*, Mar. 2023. doi: 10.1007/s10107-023-01943-7. URL https://doi.org/10.1007/s10107-023-01943-7.

E. D. Dolan, J. J. Moré, and T. S. Munson. Optimality measures for performance profiles. *SIAM Journal on Optimization*, 16(3):891–909, Jan. 2006. doi: 10.1137/040608015. URL https://doi.org/10.1137/040608015.

References VI

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12, July 2011.

M. C. Ferris, S. Lucid, and M. Roma. Nonmonotone curvilinear line search methods for unconstrained optimization. *Computational Optimization and Applications*, 6(2):117–136, Sept. 1996. doi: 10.1007/bf00249642. URL https://doi.org/10.1007/bf00249642.

D. Goldfarb. Curvilinear path steplength algorithms for minimization which use directions of negative curvature. *Mathematical Programming*, 18(1):31–40, Dec. 1980. doi: 10.1007/bf01588294. URL https://doi.org/10.1007/bf01588294.

S. M. Goldfeld, R. E. Quandt, and H. F. Trotter. Maximization by quadratic hill-climbing. *Econometrica*, 34(3):541, July 1966. doi: 10.2307/1909768. URL https://doi.org/10.2307/1909768.

References VII

N. I. M. Gould, S. Lucidi, M. Roma, and P. L. Toint. Exploiting negative curvature directions in linesearch methods for unconstrained optimization. *Optimization Methods and Software*, 14(1-2):75–98, Jan. 2000. doi: 10.1080/10556780008805794. URL https://doi.org/10.1080/10556780008805794.

N. I. M. Gould, D. Orban, and P. L. Toint. GALAHAD, a library of thread-safe fortran 90 packages for large-scale nonlinear optimization. *ACM Transactions on Mathematical Software*, 29(4):353–372, Dec. 2003.

S. Gratton and P. L. Toint. Opm, a collection of optimization problems in matlab, 2021.

S. Gratton and P. L. Toint. Adaptive regularization minimization algorithms with nonsmooth norms. *IMA Journal of Numerical Analysis*, 03 2022. ISSN 0272-4979. doi: 10.1093/imanum/drac005. URL https://doi.org/10.1093/imanum/drac005. drac005.

References VIII

S. Gratton, J, and P. L. Toint. Convergence properties of an objective-function-free optimization regularization algorithm, including an $\wr(\epsilon^{-3/2})$ complexity bound. *SIAM Journal on Optimization*, 33(3):1621–1646, 2023a.

S. Gratton, J, and P. L. Toint. Yet another fast variant of newton's method for nonconvex optimization, 2023b.

A. Griewank. The modification of Newton's method for unconstrained optimization by bounding cubic terms, 1981.

C. Lemarechal. Cauchy and the gradient method. *Doc Math Extra*, pages 251–254, 2012.

M. Liu, Z. Li, X. Wang, J. Yi, and T. Yang. Adaptive negative curvature descent with applications in non-convex optimization. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/file/f52854cc99ae1c1966b0a21d0127975b-Paper.pdf`.

M. W. Mahoney. Randomized algorithms for matrices and data. *Foundational Trends in Machine Learning*, 2011.

B. McMahan and M. Streeter. Adaptive bound optimization for online convex optimization. In *Conference on Learning Theory*, page 244sq, 2010.

K. Mishchenko. Regularized newton method with global \({\boldsymbol{\mathcal{o}}(1/{k}^2)}\) convergence. *SIAM Journal on Optimization*, 33(3):1440–1462, July 2023. doi: 10.1137/22m1488752. URL https://doi.org/10.1137/22m1488752.

Y. Nesterov and B. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, Apr 2006. ISSN 1436-4646. doi: 10.1007/s10107-006-0706-8. URL http://dx.doi.org/10.1007/s10107-006-0706-8.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, NY, USA, 2e edition, 2006.

C. W. Royer and S. J. Wright. Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1448–1477, 2018. doi: 10.1137/17M1134329. URL https://doi.org/10.1137/17M1134329.

References X

C. W. Royer, M. O'Neill, and S. J. Wright. A newton-CG algorithm with complexity guarantees for smooth unconstrained optimization. *Mathematical Programming*, 180(1-2):451–488, Jan. 2019. doi: 10.1007/s10107-019-01362-7. URL `https://doi.org/10.1007/s10107-019-01362-7`.

K. Ueda and N. Yamashita. A regularized newton method without line search for unconstrained optimization. *Computational Optimization and Applications*, 59 (1-2):321–351, Apr. 2014. doi: 10.1007/s10589-014-9656-x. URL `https://doi.org/10.1007/s10589-014-9656-x`.

D. P. Woodruff. Computational advertising: Techniques for targeting relevant ads. *Foundations and Trends® in Theoretical Computer Science*, 10(1-2):1–157, 2014. doi: 10.1561/0400000060.

P. Xu, F. Roosta, and M. W. Mahoney. Newton-type methods for non-convex optimization under inexact hessian information. *Mathematical Programming*, 184(1-2):35–70, May 2019.