Krylov Cubic Regularized Newton: A Subspace Second-Order Method with Dimension-Free Convergence Rate

Ruichen Jiang
UT Austin

Parameswaran Raman
AWS

Shoham Sabach
Technion & AWS

Aryan Mokhtari
UT Austin

Mingyi Hong
UMN & AWS

Volkan Cevher
EPFL & AWS

January 12, 2024

Outline

1. Introduction
2. The Proposed Method
3. Convergence Analysis
4. Numerical Results

Convex Minimization

- Consider the unconstrained minimization problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x}),$$

  where f is convex and twice continuously differentiable

Convex Minimization

- Consider the unconstrained minimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}),$$

  where f is convex and twice continuously differentiable
- Popular methods: first-order methods such as gradient descent

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta\nabla f(\boldsymbol{x}_k)$$

  - Cheap to implement ✓
  - Slow convergence, esp. for ill-conditioned problems ✗

Second-Order Methods

- We focus on second-order methods that utilize the Hessian of f
- Newton's method

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \nabla^2 f(\boldsymbol{x}_k)^{-1} \nabla f(\boldsymbol{x}_k)$$

Second-Order Methods

- We focus on second-order methods that utilize the Hessian of f
- Newton's method

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \nabla^2 f(\boldsymbol{x}_k)^{-1} \nabla f(\boldsymbol{x}_k)$$

- Cubic regularized Newton (CRN) method [Griewank'81; Nesterov-Polyak'06]

$$\boldsymbol{x}_{k+1} = \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{R}^d} \left\{ f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^\top (\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_k)^\top \nabla^2 f(\boldsymbol{x}_k)(\boldsymbol{x} - \boldsymbol{x}_k) + \frac{M}{6}\|\boldsymbol{x} - \boldsymbol{x}_k\|^3 \right\}$$

Second-Order Methods

- We focus on second-order methods that utilize the Hessian of f
- Newton's method

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{x}_k - \nabla^2 f(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k) \\ &= \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^d} \left\{ f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top (\mathbf{x} - \mathbf{x}_k) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_k)^\top \nabla^2 f(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k) \right\} \end{aligned}$$

- Cubic regularized Newton (CRN) method [Griewank'81; Nesterov-Polyak'06]

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^d} \left\{ f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top (\mathbf{x} - \mathbf{x}_k) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_k)^\top \nabla^2 f(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k) + \frac{M}{6} \|\mathbf{x} - \mathbf{x}_k\|^3 \right\}$$

Second-Order Methods

- We focus on second-order methods that utilize the Hessian of f
- Newton's method

$$\begin{aligned}\boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - \nabla^2 f(\boldsymbol{x}_k)^{-1}\nabla f(\boldsymbol{x}_k) \\ &= \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^d}\left\{f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^\top(\boldsymbol{x}-\boldsymbol{x}_k) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_k)^\top\nabla^2 f(\boldsymbol{x}_k)(\boldsymbol{x}-\boldsymbol{x}_k)\right\}\end{aligned}$$

- Cubic regularized Newton (CRN) method [Griewank'81; Nesterov-Polyak'06]

$$\boldsymbol{x}_{k+1} = \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^d}\left\{f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^\top(\boldsymbol{x}-\boldsymbol{x}_k) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_k)^\top\nabla^2 f(\boldsymbol{x}_k)(\boldsymbol{x}-\boldsymbol{x}_k) + \frac{M}{6}\|\boldsymbol{x}-\boldsymbol{x}_k\|^3\right\}$$

  - When f is convex, $f(\boldsymbol{x}_k) - f^* = \mathcal{O}(1/k^2)$
  - When f is strongly convex, it achieves a superlinear convergence rate

The Curse of Dimensionality

- The main drawback of CRN is its substantial memory and computational costs
  - Computing & storing the Hessian $\nabla^2 f(\boldsymbol{x})$: $\mathcal{O}(d^2)$
  - Solving a cubic subproblem: $\mathcal{O}(d^3)$
- As a result, CRN becomes impractical for optimization problems with high dimensions

Subspace Second-Order Methods

- To reduce the cost, one approach is to execute 2nd-order updates in a subspace $\mathcal{V}_k$ of dimension $m \ll d$ [Doikov-Richtárik'18; Gower et al.'19; Hanzely et al.'20]

Subspace Second-Order Methods

- To reduce the cost, one approach is to execute 2nd-order updates in a subspace $\mathcal{V}_k$ of dimension $m \ll d$ [Doikov-Richtárik'18; Gower et al.'19; Hanzely et al.'20]
- Let $\boldsymbol{g}_k := \nabla f(\boldsymbol{x}_k)$ and $\mathbf{H}_k := \nabla^2 f(\boldsymbol{x}_k)$

CRN:
$$\boldsymbol{s}_k = \operatorname*{argmin}_{\boldsymbol{s}\in\mathbb{R}^d} \left\{\boldsymbol{g}_k^\top \boldsymbol{s} + \frac{1}{2}\boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s} + \frac{M}{6}\|\boldsymbol{s}\|^3\right\}$$
$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{s}_k$$

Subspace Second-Order Methods

- To reduce the cost, one approach is to execute 2nd-order updates in a subspace $\mathcal{V}_k$ of dimension $m \ll d$ [Doikov-Richtárik'18; Gower et al.'19; Hanzely et al.'20]
- Let $\boldsymbol{g}_k := \nabla f(\boldsymbol{x}_k)$ and $\mathbf{H}_k := \nabla^2 f(\boldsymbol{x}_k)$

Subspace CRN:

$$\boldsymbol{s}_k = \operatorname*{argmin}_{\boldsymbol{s} \in \mathcal{V}_k} \left\{ \boldsymbol{g}_k^\top \boldsymbol{s} + \frac{1}{2} \boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s} + \frac{M}{6} \|\boldsymbol{s}\|^3 \right\}$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{s}_k$$

Subspace Second-Order Methods

- To reduce the cost, one approach is to execute 2nd-order updates in a subspace $\mathcal{V}_k$ of dimension $m \ll d$ [Doikov-Richtárik'18; Gower et al.'19; Hanzely et al.'20]
- Let $\boldsymbol{g}_k := \nabla f(\boldsymbol{x}_k)$ and $\mathbf{H}_k := \nabla^2 f(\boldsymbol{x}_k)$

Subspace CRN:
$$\boldsymbol{s}_k = \operatorname*{argmin}_{\boldsymbol{s} \in \mathcal{V}_k} \left\{ \boldsymbol{g}_k^\top \boldsymbol{s} + \frac{1}{2} \boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s} + \frac{M}{6} \|\boldsymbol{s}\|^3 \right\}$$
$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{s}_k$$

- Equivalently, let $\mathbf{V}_k \in \mathbb{R}^{d \times m}$ whose columns form an orthonormal basis for $\mathcal{V}_k$

Subspace CRN:
$$\boldsymbol{z}_k = \operatorname*{argmin}_{\boldsymbol{z} \in \mathbb{R}^m} \left\{ \tilde{\boldsymbol{g}}_k^\top \boldsymbol{z} + \frac{1}{2} \boldsymbol{z}^\top \tilde{\mathbf{H}}_k \boldsymbol{z} + \frac{M}{6} \|\boldsymbol{z}\|^3 \right\}, \quad \boldsymbol{s}_k = \mathbf{V}_k \boldsymbol{z}_k,$$
$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{s}_k$$

where $\tilde{\boldsymbol{g}}_k = \mathbf{V}_k^\top \boldsymbol{g}_k \in \mathbb{R}^m$ and $\tilde{\mathbf{H}}_k = \mathbf{V}_k^\top \mathbf{H}_k \mathbf{V}_k \in \mathbb{R}^{m \times m}$

Stochastic Subspace Second-Order Methods

- How to choose the subspace $\mathcal{V}_k$?

Stochastic Subspace Second-Order Methods

- How to choose the subspace $\mathcal{V}_k$?
- The existing works choose a random subspace
  - Randomized Block Cubic Newton (RBCN) [Doikov-Richtárik'18]: Sampling a random block of m coordinates
  - Stochastic Subspace Cubic Newton (SSCN) [Hanzely et al.'20]: A general random subspace satisfying $\mathbb{E}[\mathbf{V}_k \mathbf{V}_k^\top] = \frac{m}{d}\mathbf{I}$

Stochastic Subspace Second-Order Methods

- How to choose the subspace $\mathcal{V}_k$?
- The existing works choose a random subspace
  - Randomized Block Cubic Newton (RBCN) [Doikov-Richtárik'18]: Sampling a random block of m coordinates
  - Stochastic Subspace Cubic Newton (SSCN) [Hanzely et al.'20]: A general random subspace satisfying $\mathbb{E}[\mathbf{V}_k\mathbf{V}_k^\top] = \frac{m}{d}\mathbf{I}$
- Reduced computational costs ...
  - Computing the subspace Hessian $\tilde{\mathbf{H}}_k$: $\mathcal{O}(m^2)$ in some special cases
  - Solving the cubic subproblem: $\mathcal{O}(m^3)$

Stochastic Subspace Second-Order Methods

- How to choose the subspace $\mathcal{V}_k$?
- The existing works choose a random subspace
  - Randomized Block Cubic Newton (RBCN) [Doikov-Richtárik'18]: Sampling a random block of m coordinates
  - Stochastic Subspace Cubic Newton (SSCN) [Hanzely et al.'20]: A general random subspace satisfying $\mathbb{E}[\mathbf{V}_k \mathbf{V}_k^\top] = \frac{m}{d}\mathbf{I}$
- Reduced computational costs ...
  - Computing the subspace Hessian $\tilde{\mathbf{H}}_k$: $\mathcal{O}(m^2)$ in some special cases
  - Solving the cubic subproblem: $\mathcal{O}(m^3)$
- ... but a much slower convergence rate

$$\mathcal{O}\left(\frac{d-m}{m} \cdot \frac{1}{k} + \left(\frac{d}{m}\right)^2 \cdot \frac{1}{k^2}\right)$$

Stochastic Subspace Second-Order Methods

- How to choose the subspace $\mathcal{V}_k$?
- The existing works choose a random subspace
  - Randomized Block Cubic Newton (RBCN) [Doikov-Richtárik'18]: Sampling a random block of m coordinates
  - Stochastic Subspace Cubic Newton (SSCN) [Hanzely et al.'20]: A general random subspace satisfying $\mathbb{E}[\mathbf{V}_k \mathbf{V}_k^\top] = \frac{m}{d}\mathbf{I}$
- Reduced computational costs ...
  - Computing the subspace Hessian $\tilde{\mathbf{H}}_k$: $\mathcal{O}(m^2)$ in some special cases
  - Solving the cubic subproblem: $\mathcal{O}(m^3)$
- ... but a much slower convergence rate

$$\mathcal{O}\left(\frac{d-m}{m} \cdot \frac{1}{k} + \left(\frac{d}{m}\right)^2 \cdot \frac{1}{k^2}\right)$$

Question: Can we improve the dimensional dependence of subspace second-order methods?

Where the Dimensional Dependence Comes from?

- Intuitively, it stems from the fact that the subspace is chosen uniformly random, oblivious to the objective function f

Where the Dimensional Dependence Comes from?

- Intuitively, it stems from the fact that the subspace is chosen uniformly random, oblivious to the objective function f

- Such a random subspace is unlikely to contain a "good" descent direction

Where the Dimensional Dependence Comes from?

- Intuitively, it stems from the fact that the subspace is chosen uniformly random, oblivious to the objective function f

- Such a random subspace is unlikely to contain a "good" descent direction

- It should be better to employ a subspace customized to the local geometry of f

Contributions

- We propose the Krylov CRN method, where we perform the CRN update over the Krylov subspace associated with $\mathbf{H}_k$ and $\boldsymbol{g}_k$

Contributions

- We propose the Krylov CRN method, where we perform the CRN update over the Krylov subspace associated with $\mathbf{H}_k$ and $\boldsymbol{g}_k$
- Can be implemented using the Lanczos method, with 1 gradient evaluation and m Hessian-vector products per iteration

Contributions

- We propose the Krylov CRN method, where we perform the CRN update over the Krylov subspace associated with $\mathbf{H}_k$ and $\boldsymbol{g}_k$
- Can be implemented using the Lanczos method, with 1 gradient evaluation and m Hessian-vector products per iteration
- In the convex case, we prove a dimension-free convergence rate

Contributions

- We propose the Krylov CRN method, where we perform the CRN update over the Krylov subspace associated with $\mathbf{H}_k$ and $\boldsymbol{g}_k$
- Can be implemented using the Lanczos method, with 1 gradient evaluation and m Hessian-vector products per iteration
- In the convex case, we prove a dimension-free convergence rate

Methods	Per-iteration cost	Convergence rate
CRN [Nesterov-Polyak'06]	$\mathcal{O}(d^3)$	$\mathcal{O}(\frac{1}{k^2})$
SSCN [Hanzely et al.'20]	$\mathcal{O}(m^3)^*$	$\mathcal{O}(\frac{d-m}{m} \cdot \frac{1}{k} + \frac{d^2}{m^2} \cdot \frac{1}{k^2})$
Krylov CRN (ours)	$\mathcal{O}(md)^{**}$	$\mathcal{O}(\frac{1}{mk} + \frac{1}{k^2})$

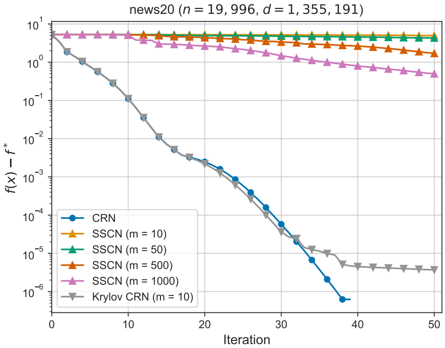
Contributions

- We propose the Krylov CRN method, where we perform the CRN update over the Krylov subspace associated with $\mathbf{H}_k$ and $\boldsymbol{g}_k$
- Can be implemented using the Lanczos method, with 1 gradient evaluation and m Hessian-vector products per iteration
- In the convex case, we prove a dimension-free convergence rate

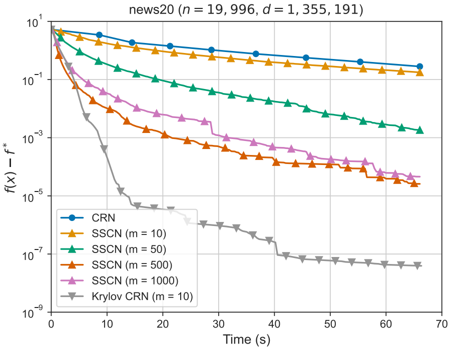
Methods	Per-iteration cost	Convergence rate
CRN [Nesterov-Polyak'06]	$\mathcal{O}(d^3)$	$\mathcal{O}(\frac{1}{k^2})$
SSCN [Hanzely et al.'20]	$\mathcal{O}(m^3)^*$	$\mathcal{O}(\frac{d-m}{m} \cdot \frac{1}{k} + \frac{d^2}{m^2} \cdot \frac{1}{k^2})$
Krylov CRN (ours)	$\mathcal{O}(md)^{**}$	$\mathcal{O}(\frac{1}{mk} + \frac{1}{k^2})$

- When the Hessian spectrum possesses certain structure, our method can achieve a faster rate

Contributions



(a) Loss curve by iteration



(b) Loss curve by time

Outline

1. Introduction
2. The Proposed Method
3. Convergence Analysis
4. Numerical Results

Review: Cubic Regularized Newton Method

- We assume that:
  - $f : \mathbb{R}^d \to \mathbb{R}$ is convex
  - f is bounded from below and has bounded level-sets
  - The Hessian of f is Lipchitz, i.e., $\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \leq L_2 \|\boldsymbol{x} - \boldsymbol{y}\|, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$

Review: Cubic Regularized Newton Method

- We assume that:
  - $f : \mathbb{R}^d \to \mathbb{R}$ is convex
  - f is bounded from below and has bounded level-sets
  - The Hessian of f is Lipchitz, i.e., $\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \leq L_2 \|\boldsymbol{x} - \boldsymbol{y}\|, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$
- An important property: upper bound on quadratic approximation error

$$\left| f(\boldsymbol{x}) - \left(f(\boldsymbol{x}_k) + \boldsymbol{g}_k^\top (\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{x}_k)^\top \mathbf{H}_k (\boldsymbol{x} - \boldsymbol{x}_k) \right) \right| \leq \frac{L_2}{6} \|\boldsymbol{x} - \boldsymbol{x}_k\|^3, \quad \forall \boldsymbol{x} \in \mathbb{R}^d$$

Review: Cubic Regularized Newton Method

- We assume that:
  - $f : \mathbb{R}^d \to \mathbb{R}$ is convex
  - f is bounded from below and has bounded level-sets
  - The Hessian of f is Lipchitz, i.e., $\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \le L_2 \|\boldsymbol{x} - \boldsymbol{y}\|, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$
- An important property: upper bound on quadratic approximation error

$$\left| f(\boldsymbol{x}) - \left(f(\boldsymbol{x}_k) + \boldsymbol{g}_k^\top (\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_k)^\top \mathbf{H}_k (\boldsymbol{x} - \boldsymbol{x}_k) \right) \right| \le \frac{L_2}{6} \|\boldsymbol{x} - \boldsymbol{x}_k\|^3, \quad \forall \boldsymbol{x} \in \mathbb{R}^d$$

- The CRN selects $\boldsymbol{x}_{k+1}$ as the minimizer of the cubic upper approximation

$$\boldsymbol{x}_{k+1} = \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{R}^d} \left\{ f(\boldsymbol{x}_k) + \boldsymbol{g}_k^\top (\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_k)^\top \mathbf{H}_k (\boldsymbol{x} - \boldsymbol{x}_k) + \frac{L_2}{6} \|\boldsymbol{x} - \boldsymbol{x}_k\|^3 \right\}$$

Review: Cubic Regularized Newton Method

Theorem

Let $\{\boldsymbol{x}_k\}_{k\geq 0}$ *be generated by CRN and define* $D := \sup\{\|\boldsymbol{x} - \boldsymbol{x}^*\| : \ \boldsymbol{x} \in \mathbb{R}^d, \ f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$*. Then we have*

$$f(\boldsymbol{x}_k) - f^* \leq \frac{9L_2 D^3}{(k+2)(k+1)}$$

Review: Cubic Regularized Newton Method

Theorem

Let $\{\boldsymbol{x}_k\}_{k\geq 0}$ *be generated by CRN and define* $D := \sup\{\|\boldsymbol{x}-\boldsymbol{x}^*\| : \boldsymbol{x} \in \mathbb{R}^d,\ f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$. *Then we have*

$$f(\boldsymbol{x}_k) - f^* \leq \frac{9L_2D^3}{(k+2)(k+1)}$$

- Using the cubic upper bound ($\boldsymbol{s}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$):

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) + \boldsymbol{g}_k^\top \boldsymbol{s}_k + \frac{1}{2}\boldsymbol{s}_k^\top \mathbf{H}_k \boldsymbol{s}_k + \frac{L_2}{6}\|\boldsymbol{s}_k\|^3$$

Review: Cubic Regularized Newton Method

Theorem

Let $\{\mathbf{x}_k\}_{k\geq 0}$ *be generated by CRN and define* $D := \sup\{\|\mathbf{x}-\mathbf{x}^*\| : \mathbf{x}\in\mathbb{R}^d,\ f(\mathbf{x})\leq f(\mathbf{x}_0)\}$. *Then we have*

$$f(\mathbf{x}_k) - f^* \leq \frac{9L_2D^3}{(k+2)(k+1)}$$

- Using the cubic upper bound ($\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$):

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \mathbf{g}_k^\top \mathbf{s}_k + \frac{1}{2}\mathbf{s}_k^\top \mathbf{H}_k \mathbf{s}_k + \frac{L_2}{6}\|\mathbf{s}_k\|^3$$

- Since $\mathbf{s}_k$ minimizes the cubic function:

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \mathbf{g}_k^\top \mathbf{s} + \frac{1}{2}\mathbf{s}^\top \mathbf{H}_k \mathbf{s} + \frac{L_2}{6}\|\mathbf{s}\|^3, \quad \forall \mathbf{s}\in\mathbb{R}^d$$

Review: Cubic Regularized Newton Method

Theorem

Let $\{\mathbf{x}_k\}_{k\geq 0}$ *be generated by CRN and define* $D := \sup\{\|\mathbf{x}-\mathbf{x}^*\| : \mathbf{x}\in\mathbb{R}^d,\ f(\mathbf{x})\leq f(\mathbf{x}_0)\}$. *Then we have*

$$f(\mathbf{x}_k)-f^*\leq \frac{9L_2D^3}{(k+2)(k+1)}$$

- Using the cubic upper bound ($\mathbf{s}_k=\mathbf{x}_{k+1}-\mathbf{x}_k$):

$$f(\mathbf{x}_{k+1})\leq f(\mathbf{x}_k)+\mathbf{g}_k^\top\mathbf{s}_k+\frac{1}{2}\mathbf{s}_k^\top\mathbf{H}_k\mathbf{s}_k+\frac{L_2}{6}\|\mathbf{s}_k\|^3$$

- Since $\mathbf{s}_k$ minimizes the cubic function:

$$f(\mathbf{x}_{k+1})\leq f(\mathbf{x}_k)+\mathbf{g}_k^\top\mathbf{s}+\frac{1}{2}\mathbf{s}^\top\mathbf{H}_k\mathbf{s}+\frac{L_2}{6}\|\mathbf{s}\|^3,\quad \forall\mathbf{s}\in\mathbb{R}^d$$

- Using the cubic lower bound:

$$f(\mathbf{x}_{k+1})\leq f(\mathbf{x}_k+\mathbf{s})+\frac{L_2}{3}\|\mathbf{s}\|^3,\quad \forall\mathbf{s}\in\mathbb{R}^d$$

Review: Cubic Regularized Newton Method (cont.)

- Using the cubic lower bound:
$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k + \boldsymbol{s}) + \frac{L_2}{3}\|\boldsymbol{s}\|^3, \quad \forall \boldsymbol{s} \in \mathbb{R}^d$$

Review: Cubic Regularized Newton Method (cont.)

- Using the cubic lower bound:

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k + \boldsymbol{s}) + \frac{L_2}{3}\|\boldsymbol{s}\|^3, \quad \forall \boldsymbol{s} \in \mathbb{R}^d$$

- Choosing $\boldsymbol{s} = \frac{3}{k+3}(\boldsymbol{x}^* - \boldsymbol{x}_k)$ and using convexity of f:

$$f(\boldsymbol{x}_{k+1}) \leq f\left(\frac{k}{k+3}\boldsymbol{x}_k + \frac{3}{k+3}\boldsymbol{x}^*\right) + \frac{9L_2\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^3}{(k+3)^3} \leq \frac{k}{k+3}f(\boldsymbol{x}_k) + \frac{3}{k+3}f^* + \frac{9L_2 D^3}{(k+3)^3}$$

Review: Cubic Regularized Newton Method (cont.)

- Using the cubic lower bound:

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k + \boldsymbol{s}) + \frac{L_2}{3}\|\boldsymbol{s}\|^3, \quad \forall \boldsymbol{s} \in \mathbb{R}^d$$

- Choosing $\boldsymbol{s} = \frac{3}{k+3}(\boldsymbol{x}^* - \boldsymbol{x}_k)$ and using convexity of f:

$$f(\boldsymbol{x}_{k+1}) \leq f\left(\frac{k}{k+3}\boldsymbol{x}_k + \frac{3}{k+3}\boldsymbol{x}^*\right) + \frac{9L_2\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^3}{(k+3)^3} \leq \frac{k}{k+3}f(\boldsymbol{x}_k) + \frac{3}{k+3}f^* + \frac{9L_2D^3}{(k+3)^3}$$

- Define $A_k = k(k+1)(k+2)$:

$$f(\boldsymbol{x}_{k+1}) - f^* \leq \frac{k}{k+3}(f(\boldsymbol{x}_k) - f^*) + \frac{9L_2D^3}{(k+3)^3}$$

$$\Rightarrow \quad A_{k+1}(f(\boldsymbol{x}_{k+1}) - f^*) \leq A_k(f(\boldsymbol{x}_k) - f^*) + 9L_2D^3$$

$$\Rightarrow \quad A_{k+1}(f(\boldsymbol{x}_{k+1}) - f^*) \leq 9L_2D^3(k+1) \quad \Rightarrow \quad f(\boldsymbol{x}_{k+1}) - f^* \leq \frac{9L_2D^3}{(k+2)(k+3)}$$

Subspace Cubic Regularized Newton

- Subspace CRN:

$$\begin{aligned} \boldsymbol{s}_k &= \operatorname*{argmin}_{\boldsymbol{s} \in {\color{red}\mathcal{V}_k}} \left\{ \boldsymbol{g}_k^\top \boldsymbol{s} + \frac{1}{2} \boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s} + \frac{L_2}{6} \|\boldsymbol{s}\|^3 \right\} \\ \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k + \boldsymbol{s}_k \end{aligned}$$

- Where the analysis would break when introducing the subspace $\mathcal{V}_k$?

Subspace Cubic Regularized Newton

- Subspace CRN:

$$
\begin{aligned}
\boldsymbol{s}_k &= \operatorname*{argmin}_{\boldsymbol{s}\in\mathcal{V}_k} \left\{ \boldsymbol{g}_k^\top \boldsymbol{s} + \frac{1}{2}\boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s} + \frac{L_2}{6}\|\boldsymbol{s}\|^3 \right\} \\
\boldsymbol{x}_{k+1} &= \boldsymbol{x}_k + \boldsymbol{s}_k
\end{aligned}
$$

- Where the analysis would break when introducing the subspace $\mathcal{V}_k$?
- Recall the crucial inequality

$$
f(\boldsymbol{x}_{k+1}) \le f(\boldsymbol{x}_k) + \boldsymbol{g}_k^\top \boldsymbol{s} + \frac{1}{2}\boldsymbol{s}_k^\top \mathbf{H}_k \boldsymbol{s} + \frac{L_2}{6}\|\boldsymbol{s}\|^3, \quad \forall \boldsymbol{s} \in \mathbb{R}^d
$$

Subspace Cubic Regularized Newton

- Subspace CRN:

$$\mathbf{s}_k = \operatorname*{argmin}_{\mathbf{s} \in \mathcal{V}_k} \left\{ \mathbf{g}_k^\top \mathbf{s} + \frac{1}{2} \mathbf{s}^\top \mathbf{H}_k \mathbf{s} + \frac{L_2}{6} \|\mathbf{s}\|^3 \right\}$$
$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_k$$

- Where the analysis would break when introducing the subspace $\mathcal{V}_k$?
- Recall the crucial inequality

$$f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k) + \mathbf{g}_k^\top \mathbf{s} + \frac{1}{2} \mathbf{s}_k^\top \mathbf{H}_k \mathbf{s} + \frac{L_2}{6} \|\mathbf{s}\|^3, \quad \cancel{\forall \mathbf{s} \in \mathbb{R}^d} \quad \forall \mathbf{s} \in \mathcal{V}_k$$

Subspace Cubic Regularized Newton

- Subspace CRN:

$$\mathbf{s}_k = \operatorname*{argmin}_{\mathbf{s}\in\mathcal{V}_k} \left\{ \mathbf{g}_k^\top \mathbf{s} + \frac{1}{2}\mathbf{s}^\top \mathbf{H}_k \mathbf{s} + \frac{L_2}{6}\|\mathbf{s}\|^3 \right\}$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_k$$

- Where the analysis would break when introducing the subspace $\mathcal{V}_k$?
- Recall the crucial inequality

$$f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k) + \mathbf{g}_k^\top \mathbf{s} + \frac{1}{2}\mathbf{s}_k^\top \mathbf{H}_k \mathbf{s} + \frac{L_2}{6}\|\mathbf{s}\|^3, \quad \cancel{\forall \mathbf{s} \in \mathbb{R}^d} \quad \forall \mathbf{s} \in \mathcal{V}_k$$

- Let $\mathbf{P}_k := \mathbf{V}_k \mathbf{V}_k^\top$ be the orthogonal projection matrix of $\mathcal{V}_k$. Since $\mathbf{P}_k \mathbf{s} \in \mathcal{V}_k$, $\forall \mathbf{s} \in \mathbb{R}^d$:

$$f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k) + \mathbf{g}_k^\top \mathbf{P}_k \mathbf{s} + \frac{1}{2}(\mathbf{P}_k \mathbf{s})^\top \mathbf{H}_k (\mathbf{P}_k \mathbf{s}) + \frac{L_2}{6}\|\mathbf{P}_k \mathbf{s}\|^3, \quad \forall \mathbf{s} \in \mathbb{R}^d$$

Subspace Cubic Regularized Newton

- The exactly same analysis would apply if, for any $\boldsymbol{s} \in \mathbb{R}^d$,
  (A) $\boldsymbol{g}_k^\top \mathbf{P}_k \boldsymbol{s} \le \boldsymbol{g}_k^\top \boldsymbol{s}$
  (B) $(\mathbf{P}_k \boldsymbol{s})^\top \mathbf{H}_k (\mathbf{P}_k \boldsymbol{s}) \le \boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s}$
  (C) $\|\mathbf{P}_k \boldsymbol{s}\| \le \|\boldsymbol{s}\|$

Subspace Cubic Regularized Newton

- The exactly same analysis would apply if, for any $\boldsymbol{s} \in \mathbb{R}^d$,

  (A) $\boldsymbol{g}_k^\top \mathbf{P}_k \boldsymbol{s} \le \boldsymbol{g}_k^\top \boldsymbol{s} \quad \Leftrightarrow \quad \boldsymbol{g}_k = \mathbf{P}_k \boldsymbol{g}_k \quad \Leftrightarrow \quad \boldsymbol{g}_k \in \mathcal{V}_k$

  (B) $(\mathbf{P}_k \boldsymbol{s})^\top \mathbf{H}_k (\mathbf{P}_k \boldsymbol{s}) \le \boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s}$

  (C) $\|\mathbf{P}_k \boldsymbol{s}\| \le \|\boldsymbol{s}\|$

Subspace Cubic Regularized Newton

- The exactly same analysis would apply if, for any $\boldsymbol{s} \in \mathbb{R}^d$,

  (A) $\boldsymbol{g}_k^\top \mathbf{P}_k \boldsymbol{s} \leq \boldsymbol{g}_k^\top \boldsymbol{s} \quad \Leftrightarrow \quad \boldsymbol{g}_k = \mathbf{P}_k \boldsymbol{g}_k \quad \Leftrightarrow \quad \boldsymbol{g}_k \in \mathcal{V}_k$

  (B) $(\mathbf{P}_k \boldsymbol{s})^\top \mathbf{H}_k (\mathbf{P}_k \boldsymbol{s}) \leq \boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s} \quad \Leftrightarrow \quad \mathbf{P}_k \mathbf{H}_k \mathbf{P}_k \preceq \mathbf{H}_k \quad \Leftrightarrow \quad \mathbf{H}_k \boldsymbol{v} \in \mathcal{V}_k, \forall \boldsymbol{v} \in \mathcal{V}_k$

  (C) $\|\mathbf{P}_k \boldsymbol{s}\| \leq \|\boldsymbol{s}\|$

Subspace Cubic Regularized Newton

- The exactly same analysis would apply if, for any $\boldsymbol{s} \in \mathbb{R}^d$,

(A) $\boldsymbol{g}_k^\top \mathbf{P}_k \boldsymbol{s} \leq \boldsymbol{g}_k^\top \boldsymbol{s} \quad \Leftrightarrow \quad \boldsymbol{g}_k = \mathbf{P}_k \boldsymbol{g}_k \quad \Leftrightarrow \quad \boldsymbol{g}_k \in \mathcal{V}_k$

(B) $(\mathbf{P}_k \boldsymbol{s})^\top \mathbf{H}_k (\mathbf{P}_k \boldsymbol{s}) \leq \boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s} \quad \Leftrightarrow \quad \mathbf{P}_k \mathbf{H}_k \mathbf{P}_k \preceq \mathbf{H}_k \quad \Leftrightarrow \quad \mathbf{H}_k \boldsymbol{v} \in \mathcal{V}_k, \forall \boldsymbol{v} \in \mathcal{V}_k$

(C) $\|\mathbf{P}_k \boldsymbol{s}\| \leq \|\boldsymbol{s}\|$ ✓ since $\mathbf{P}_k$ is an orthogonal projection matrix

Subspace Cubic Regularized Newton

- The exactly same analysis would apply if, for any $\boldsymbol{s} \in \mathbb{R}^d$,

  (A) $\boldsymbol{g}_k^\top \mathbf{P}_k \boldsymbol{s} \le \boldsymbol{g}_k^\top \boldsymbol{s} \quad \Leftrightarrow \quad \boldsymbol{g}_k = \mathbf{P}_k \boldsymbol{g}_k \quad \Leftrightarrow \quad \boldsymbol{g}_k \in \mathcal{V}_k$

  (B) $(\mathbf{P}_k \boldsymbol{s})^\top \mathbf{H}_k (\mathbf{P}_k \boldsymbol{s}) \le \boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s} \quad \Leftrightarrow \quad \mathbf{P}_k \mathbf{H}_k \mathbf{P}_k \preceq \mathbf{H}_k \quad \Leftrightarrow \quad \mathbf{H}_k \boldsymbol{v} \in \mathcal{V}_k, \forall \boldsymbol{v} \in \mathcal{V}_k$

  (C) $\|\mathbf{P}_k \boldsymbol{s}\| \le \|\boldsymbol{s}\|$ ✓ since $\mathbf{P}_k$ is an orthogonal projection matrix

- Combining (A) and (B), we obtain that

$$\boldsymbol{g}_k \in \mathcal{V}_k,\ \mathbf{H}_k \boldsymbol{g}_k \in \mathcal{V}_k,\ \mathbf{H}_k^2 \boldsymbol{g}_k \in \mathcal{V}_k,\ \ldots, \quad \Rightarrow \quad \mathrm{span}\{\mathbf{H}_k^i \boldsymbol{g}_k \mid i = 0, 1, \ldots\} \subset \mathcal{V}_k$$

Subspace Cubic Regularized Newton

- The exactly same analysis would apply if, for any $\boldsymbol{s} \in \mathbb{R}^d$,
  - (A) $\boldsymbol{g}_k^\top \mathbf{P}_k \boldsymbol{s} \leq \boldsymbol{g}_k^\top \boldsymbol{s} \quad \Leftrightarrow \quad \boldsymbol{g}_k = \mathbf{P}_k \boldsymbol{g}_k \quad \Leftrightarrow \quad \boldsymbol{g}_k \in \mathcal{V}_k$
  - (B) $(\mathbf{P}_k \boldsymbol{s})^\top \mathbf{H}_k (\mathbf{P}_k \boldsymbol{s}) \leq \boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s} \quad \Leftrightarrow \quad \mathbf{P}_k \mathbf{H}_k \mathbf{P}_k \preceq \mathbf{H}_k \quad \Leftrightarrow \quad \mathbf{H}_k \boldsymbol{v} \in \mathcal{V}_k, \forall \boldsymbol{v} \in \mathcal{V}_k$
  - (C) $\|\mathbf{P}_k \boldsymbol{s}\| \leq \|\boldsymbol{s}\|$ ✓ since $\mathbf{P}_k$ is an orthogonal projection matrix
- Combining (A) and (B), we obtain that

$$\boldsymbol{g}_k \in \mathcal{V}_k,\ \mathbf{H}_k \boldsymbol{g}_k \in \mathcal{V}_k,\ \mathbf{H}_k^2 \boldsymbol{g}_k \in \mathcal{V}_k,\ \ldots, \quad \Rightarrow \quad \mathrm{span}\{\mathbf{H}_k^i \boldsymbol{g}_k \mid i = 0, 1, \ldots\} \subset \mathcal{V}_k$$

- This is exactly the maximal Krylov subspace generated by $\mathbf{H}_k$ and $\boldsymbol{g}_k$!

Krylov Subspace

- Formally, the j-th Krylov subspace generated by $\mathbf{A} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{b} \in \mathbb{R}^{d}$ is defined as

$$\mathcal{K}_j(\mathbf{A}, \boldsymbol{b}) = \mathrm{span}\{\boldsymbol{b}, \mathbf{A}\boldsymbol{b}, \ldots, \mathbf{A}^{j-1}\boldsymbol{b}\}.$$

Krylov Subspace

- Formally, the j-th Krylov subspace generated by $\mathbf{A} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{b} \in \mathbb{R}^d$ is defined as

$$\mathcal{K}_j(\mathbf{A}, \boldsymbol{b}) = \mathrm{span}\{\boldsymbol{b}, \mathbf{A}\boldsymbol{b}, \ldots, \mathbf{A}^{j-1}\boldsymbol{b}\}.$$

- Moreover, there exists an integer $r_0 \leq d$ such that:

$$\mathcal{K}_1(\mathbf{A}, \boldsymbol{b}) \subset \mathcal{K}_2(\mathbf{A}, \boldsymbol{b}) \subset \cdots \subset \underbrace{\mathcal{K}_{r_0}(\mathbf{A}, \boldsymbol{b})}_{\dim = r_0} = \mathcal{K}_{r_0+1}(\mathbf{A}, \boldsymbol{b}) = \mathcal{K}_{r_0+2}(\mathbf{A}, \boldsymbol{b}) = \cdots$$

We call $\mathcal{K}_{r_0}(\mathbf{A}, \boldsymbol{b})$ the maximal Krylov subspace

Krylov Subspace

- Formally, the j-th Krylov subspace generated by $\mathbf{A} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{b} \in \mathbb{R}^{d}$ is defined as

$$\mathcal{K}_j(\mathbf{A}, \boldsymbol{b}) = \operatorname{span}\{\boldsymbol{b}, \mathbf{A}\boldsymbol{b}, \ldots, \mathbf{A}^{j-1}\boldsymbol{b}\}.$$

- Moreover, there exists an integer $r_0 \leq d$ such that:

$$\mathcal{K}_1(\mathbf{A}, \boldsymbol{b}) \subset \mathcal{K}_2(\mathbf{A}, \boldsymbol{b}) \subset \cdots \subset \underbrace{\mathcal{K}_{r_0}(\mathbf{A}, \boldsymbol{b})}_{\dim = r_0} = \mathcal{K}_{r_0+1}(\mathbf{A}, \boldsymbol{b}) = \mathcal{K}_{r_0+2}(\mathbf{A}, \boldsymbol{b}) = \cdots$$

  We call $\mathcal{K}_{r_0}(\mathbf{A}, \boldsymbol{b})$ the maximal Krylov subspace
- To sum up: if we let $\mathcal{V}_k = \mathcal{K}_{r_0}(\mathbf{H}_k, \boldsymbol{g}_k)$, the subspace CRN retains the same convergence rate of CRN

Krylov Subspace

- Formally, the j-th Krylov subspace generated by $\mathbf{A} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{b} \in \mathbb{R}^d$ is defined as

$$\mathcal{K}_j(\mathbf{A}, \boldsymbol{b}) = \mathrm{span}\{\boldsymbol{b}, \mathbf{A}\boldsymbol{b}, \ldots, \mathbf{A}^{j-1}\boldsymbol{b}\}.$$

- Moreover, there exists an integer $r_0 \leq d$ such that:

$$\mathcal{K}_1(\mathbf{A}, \boldsymbol{b}) \subset \mathcal{K}_2(\mathbf{A}, \boldsymbol{b}) \subset \cdots \subset \underbrace{\mathcal{K}_{r_0}(\mathbf{A}, \boldsymbol{b})}_{\dim = r_0} = \mathcal{K}_{r_0+1}(\mathbf{A}, \boldsymbol{b}) = \mathcal{K}_{r_0+2}(\mathbf{A}, \boldsymbol{b}) = \cdots$$

  We call $\mathcal{K}_{r_0}(\mathbf{A}, \boldsymbol{b})$ the maximal Krylov subspace
- To sum up: if we let $\mathcal{V}_k = \mathcal{K}_{r_0}(\mathbf{H}_k, \boldsymbol{g}_k)$, the subspace CRN retains the same convergence rate of CRN
- However, r_0 can be as large as d $\quad \Rightarrow \quad$ we use the Krylov subspace up to dim m

Krylov Subspace

- Let us check if $\mathcal{V}_k = \mathcal{K}_m(\mathbf{H}_k, \boldsymbol{g}_k)$ satisfies the three conditions:
  - (A) $\boldsymbol{g}_k^\top \mathbf{P}_k \boldsymbol{s} \leq \boldsymbol{g}_k^\top \boldsymbol{s} \quad \Leftrightarrow \quad \boldsymbol{g}_k \in \mathcal{V}_k$
  - (B) $(\mathbf{P}_k \boldsymbol{s})^\top \mathbf{H}_k (\mathbf{P}_k \boldsymbol{s}) \leq \boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s} \quad \Leftrightarrow \quad \mathbf{H}_k \boldsymbol{v} \in \mathcal{V}_k, \forall \boldsymbol{v} \in \mathcal{V}_k$
  - (C) $\|\mathbf{P}_k \boldsymbol{s}\| \leq \|\boldsymbol{s}\|$

Krylov Subspace

- Let us check if $\mathcal{V}_k = \mathcal{K}_m(\mathbf{H}_k, \boldsymbol{g}_k)$ satisfies the three conditions:

  (A) $\boldsymbol{g}_k^\top \mathbf{P}_k \boldsymbol{s} \leq \boldsymbol{g}_k^\top \boldsymbol{s} \quad \Leftrightarrow \quad \boldsymbol{g}_k \in \mathcal{V}_k \quad \checkmark$

  (B) $(\mathbf{P}_k \boldsymbol{s})^\top \mathbf{H}_k (\mathbf{P}_k \boldsymbol{s}) \leq \boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s} \quad \Leftrightarrow \quad \mathbf{H}_k \boldsymbol{v} \in \mathcal{V}_k, \forall \boldsymbol{v} \in \mathcal{V}_k$

  (C) $\|\mathbf{P}_k \boldsymbol{s}\| \leq \|\boldsymbol{s}\|$

Krylov Subspace

- Let us check if $\mathcal{V}_k = \mathcal{K}_m(\mathbf{H}_k, \boldsymbol{g}_k)$ satisfies the three conditions:
  (A) $\boldsymbol{g}_k^\top \mathbf{P}_k \boldsymbol{s} \leq \boldsymbol{g}_k^\top \boldsymbol{s} \quad \Leftrightarrow \quad \boldsymbol{g}_k \in \mathcal{V}_k$ ✓

  (B) $(\mathbf{P}_k \boldsymbol{s})^\top \mathbf{H}_k (\mathbf{P}_k \boldsymbol{s}) \leq \boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s} \quad \Leftrightarrow \quad \mathbf{H}_k \boldsymbol{v} \in \mathcal{V}_k, \forall \boldsymbol{v} \in \mathcal{V}_k$

  (C) $\|\mathbf{P}_k \boldsymbol{s}\| \leq \|\boldsymbol{s}\|$ ✓ since $\mathbf{P}_k$ is an orthogonal projection matrix

Krylov Subspace

- Let us check if $\mathcal{V}_k = \mathcal{K}_m(\mathbf{H}_k, \boldsymbol{g}_k)$ satisfies the three conditions:
  - (A) $\boldsymbol{g}_k^\top \mathbf{P}_k \boldsymbol{s} \leq \boldsymbol{g}_k^\top \boldsymbol{s} \quad \Leftrightarrow \quad \boldsymbol{g}_k \in \mathcal{V}_k$ ✓
  - (B) $(\mathbf{P}_k \boldsymbol{s})^\top \mathbf{H}_k (\mathbf{P}_k \boldsymbol{s}) \leq \boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s} \quad \Leftrightarrow \quad \mathbf{H}_k \boldsymbol{v} \in \mathcal{V}_k, \forall \boldsymbol{v} \in \mathcal{V}_k$ ✗ $\mathbf{H}_k^{m-1} \boldsymbol{g}_k \in \mathcal{V}_k$ but $\mathbf{H}_k^m \boldsymbol{g}_k \notin \mathcal{V}_k$
  - (C) $\|\mathbf{P}_k \boldsymbol{s}\| \leq \|\boldsymbol{s}\|$ ✓ since $\mathbf{P}_k$ is an orthogonal projection matrix

Krylov Subspace

- Let us check if $\mathcal{V}_k = \mathcal{K}_m(\mathbf{H}_k, \boldsymbol{g}_k)$ satisfies the three conditions:
  - (A) $\boldsymbol{g}_k^\top \mathbf{P}_k \boldsymbol{s} \leq \boldsymbol{g}_k^\top \boldsymbol{s} \quad \Leftrightarrow \quad \boldsymbol{g}_k \in \mathcal{V}_k$ ✓
  - (B) $(\mathbf{P}_k \boldsymbol{s})^\top \mathbf{H}_k (\mathbf{P}_k \boldsymbol{s}) \leq \boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s} \quad \Leftrightarrow \quad \mathbf{H}_k \boldsymbol{v} \in \mathcal{V}_k, \forall \boldsymbol{v} \in \mathcal{V}_k$ ✗ $\mathbf{H}_k^{m-1} \boldsymbol{g}_k \in \mathcal{V}_k$ but $\mathbf{H}_k^m \boldsymbol{g}_k \notin \mathcal{V}_k$
  - (C) $\|\mathbf{P}_k \boldsymbol{s}\| \leq \|\boldsymbol{s}\|$ ✓ since $\mathbf{P}_k$ is an orthogonal projection matrix
- Only need to control the error from Condition (B)

Krylov Subspace

- Let us check if $\mathcal{V}_k = \mathcal{K}_m(\mathbf{H}_k, \boldsymbol{g}_k)$ satisfies the three conditions:
  - (A) $\boldsymbol{g}_k^\top \mathbf{P}_k \boldsymbol{s} \le \boldsymbol{g}_k^\top \boldsymbol{s} \quad \Leftrightarrow \quad \boldsymbol{g}_k \in \mathcal{V}_k$ ✓
  - (B) $(\mathbf{P}_k \boldsymbol{s})^\top \mathbf{H}_k (\mathbf{P}_k \boldsymbol{s}) \le \boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s} \quad \Leftrightarrow \quad \mathbf{H}_k \boldsymbol{v} \in \mathcal{V}_k, \forall \boldsymbol{v} \in \mathcal{V}_k$ ✗ $\mathbf{H}_k^{m-1} \boldsymbol{g}_k \in \mathcal{V}_k$ but $\mathbf{H}_k^m \boldsymbol{g}_k \notin \mathcal{V}_k$
  - (C) $\|\mathbf{P}_k \boldsymbol{s}\| \le \|\boldsymbol{s}\|$ ✓ since $\mathbf{P}_k$ is an orthogonal projection matrix
- Only need to control the error from Condition (B)
- In comparison, when $\mathcal{V}_k$ is chosen randomly, both (A) and (B) only hold approximately and the induced errors depend on m/d

Krylov Cubic Regularized Newton

- **Input:** Initial point $\boldsymbol{x}_0 \in \mathbb{R}^d$, subspace dimension m, regularization parameter $M > 0$
- **for** $k = 0, 1, \ldots,$ **do**
  - $\mathbf{V}_k \leftarrow$ the orthnormal basis of $\mathcal{K}_m(\mathbf{H}_k, \boldsymbol{g}_k)$, $\tilde{\boldsymbol{g}}_k \leftarrow \mathbf{V}_k^\top \boldsymbol{g}_k$, $\tilde{\mathbf{H}}_k \leftarrow \mathbf{V}_k^\top \mathbf{H}_k \mathbf{V}_k$
  - Solve the cubic subproblem

$$\boldsymbol{z}_k = \operatorname*{argmin}_{\boldsymbol{z} \in \mathbb{R}^m} \left\{ \tilde{\boldsymbol{g}}_k^\top \boldsymbol{z} + \frac{1}{2} \boldsymbol{z}^\top \tilde{\mathbf{H}}_k \boldsymbol{z} + \frac{M}{6} \|\boldsymbol{z}\|^3 \right\},$$

  - Update $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \mathbf{V}_k \boldsymbol{z}_k$
- **end for**

Krylov Cubic Regularized Newton

- **Input:** Initial point $\boldsymbol{x}_0 \in \mathbb{R}^d$, subspace dimension m, regularization parameter $M > 0$
- **for** $k = 0, 1, \ldots,$ **do**
  - $\mathbf{V}_k \leftarrow$ the orthnormal basis of $\mathcal{K}_m(\mathbf{H}_k, \boldsymbol{g}_k)$, $\tilde{\boldsymbol{g}}_k \leftarrow \mathbf{V}_k^\top \boldsymbol{g}_k$, $\tilde{\mathbf{H}}_k \leftarrow \mathbf{V}_k^\top \mathbf{H}_k \mathbf{V}_k$
  - Solve the cubic subproblem

$$\boldsymbol{z}_k = \operatorname*{argmin}_{\boldsymbol{z} \in \mathbb{R}^m} \left\{ \tilde{\boldsymbol{g}}_k^\top \boldsymbol{z} + \frac{1}{2} \boldsymbol{z}^\top \tilde{\mathbf{H}}_k \boldsymbol{z} + \frac{M}{6} \|\boldsymbol{z}\|^3 \right\},$$

  - Update $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \mathbf{V}_k \boldsymbol{z}_k$
- **end for**

- The remaining question: How to compute $\mathbf{V}_k$, $\tilde{\boldsymbol{g}}_k$, and $\tilde{\mathbf{H}}_k$?

Krylov Cubic Regularized Newton

- **Input:** Initial point $\boldsymbol{x}_0 \in \mathbb{R}^d$, subspace dimension m, regularization parameter $M > 0$
- **for** $k = 0, 1, \ldots,$ **do**
  - $\mathbf{V}_k \leftarrow$ the orthnormal basis of $\mathcal{K}_m(\mathbf{H}_k, \boldsymbol{g}_k)$, $\tilde{\boldsymbol{g}}_k \leftarrow \mathbf{V}_k^\top \boldsymbol{g}_k$, $\tilde{\mathbf{H}}_k \leftarrow \mathbf{V}_k^\top \mathbf{H}_k \mathbf{V}_k$
  - Solve the cubic subproblem

$$\boldsymbol{z}_k = \operatorname*{argmin}_{\boldsymbol{z} \in \mathbb{R}^m} \left\{ \tilde{\boldsymbol{g}}_k^\top \boldsymbol{z} + \frac{1}{2} \boldsymbol{z}^\top \tilde{\mathbf{H}}_k \boldsymbol{z} + \frac{M}{6} \|\boldsymbol{z}\|^3 \right\},$$

  - Update $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \mathbf{V}_k \boldsymbol{z}_k$
- **end for**

- The remaining question: How to compute $\mathbf{V}_k$, $\tilde{\boldsymbol{g}}_k$, and $\tilde{\mathbf{H}}_k$? $\Leftarrow$ Lanczos method!

Lanczos Method

- **Input:** $\mathbf{H} \in \mathbb{R}^{d\times d}$, $\boldsymbol{g} \in \mathbb{R}^d$, and the dimension m
- **Initialize:** $\boldsymbol{v}_1 = \boldsymbol{g}/\|\boldsymbol{g}\|$, $\beta_1 = 0$, $\boldsymbol{v}_0 = 0$
- **for** $j = 1, 2, \ldots, m$ **do**
  - $\boldsymbol{w}_j \leftarrow \mathbf{H}\boldsymbol{v}_j - \beta_j \boldsymbol{v}_{j-1}$ // one Hessian-vector product (HVP)
  - $\alpha_j \leftarrow \boldsymbol{w}_j^\top \boldsymbol{v}_j$
  - $\boldsymbol{w}_j \leftarrow \boldsymbol{w}_j - \alpha_j \boldsymbol{v}_j$
  - $\beta_{j+1} \leftarrow \|\boldsymbol{w}_j\|_2$
  - $\boldsymbol{v}_{j+1} \leftarrow \boldsymbol{w}_j / \beta_{j+1}$
- **end for**
- **Output:** $\mathbf{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m]$, $\tilde{\boldsymbol{g}} = \|\boldsymbol{g}\|\boldsymbol{e}_1$, and $\tilde{\mathbf{H}} = \begin{bmatrix} \alpha_1 & \beta_2 & & & & \\ \beta_2 & \alpha_2 & \beta_3 & & & \\ & \beta_3 & \ddots & \ddots & & \\ & & \ddots & \ddots & \beta_{m-1} & \\ & & & \beta_{m-1} & \alpha_{m-1} & \beta_m \\ & & & & \beta_m & \alpha_m \end{bmatrix}$

Lanczos Method

- **Input:** $\mathbf{H} \in \mathbb{R}^{d\times d}$, $\boldsymbol{g} \in \mathbb{R}^d$, and the dimension m
- **Initialize:** $\boldsymbol{v}_1 = \boldsymbol{g}/\|\boldsymbol{g}\|$, $\beta_1 = 0$, $\boldsymbol{v}_0 = 0$
- **for** $j = 1, 2, \dots, m$ **do**
  - $\boldsymbol{w}_j \leftarrow \mathbf{H}\boldsymbol{v}_j - \beta_j \boldsymbol{v}_{j-1}$ // one Hessian-vector product (HVP)
  - $\alpha_j \leftarrow \boldsymbol{w}_j^\top \boldsymbol{v}_j$
  - $\boldsymbol{w}_j \leftarrow \boldsymbol{w}_j - \alpha_j \boldsymbol{v}_j$
  - $\beta_{j+1} \leftarrow \|\boldsymbol{w}_j\|_2$
  - $\boldsymbol{v}_{j+1} \leftarrow \boldsymbol{w}_j / \beta_{j+1}$
- **end for**
- **Output:** $\mathbf{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_m]$, $\tilde{\boldsymbol{g}} = \|\boldsymbol{g}\|\boldsymbol{e}_1$, and $\tilde{\mathbf{H}} = \begin{bmatrix} \alpha_1 & \beta_2 & & & & \\ \beta_2 & \alpha_2 & \beta_3 & & & \\ & \beta_3 & \ddots & \ddots & & \\ & & \ddots & \ddots & \beta_{m-1} & \\ & & & \beta_{m-1} & \alpha_{m-1} & \beta_m \\ & & & & \beta_m & \alpha_m \end{bmatrix}$

- No need to compute & store $\mathbf{H}$ explicitly; only requires m HVPs

  $\Rightarrow$ Can be done efficiently via back-propagation [Pearlmutter'94]

Lanczos Method

- **Input:** $\mathbf{H} \in \mathbb{R}^{d\times d}$, $\boldsymbol{g} \in \mathbb{R}^d$, and the dimension m
- **Initialize:** $\boldsymbol{v}_1 = \boldsymbol{g}/\|\boldsymbol{g}\|$, $\beta_1 = 0$, $\boldsymbol{v}_0 = 0$
- **for** $j = 1, 2, \ldots, m$ **do**
  - $\boldsymbol{w}_j \leftarrow \mathbf{H}\boldsymbol{v}_j - \beta_j \boldsymbol{v}_{j-1}$ // one Hessian-vector product (HVP)
  - $\alpha_j \leftarrow \boldsymbol{w}_j^\top \boldsymbol{v}_j$
  - $\boldsymbol{w}_j \leftarrow \boldsymbol{w}_j - \alpha_j \boldsymbol{v}_j$
  - $\beta_{j+1} \leftarrow \|\boldsymbol{w}_j\|_2$
  - $\boldsymbol{v}_{j+1} \leftarrow \boldsymbol{w}_j / \beta_{j+1}$
- **end for**
- **Output:** $\mathbf{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m]$, $\tilde{\boldsymbol{g}} = \|\boldsymbol{g}\|\boldsymbol{e}_1$, and $\tilde{\mathbf{H}} = \begin{bmatrix} \alpha_1 & \beta_2 & & & & \\ \beta_2 & \alpha_2 & \beta_3 & & & \\ & \beta_3 & \ddots & \ddots & & \\ & & \ddots & \ddots & \beta_{m-1} & \\ & & & \beta_{m-1} & \alpha_{m-1} & \beta_m \\ & & & & \beta_m & \alpha_m \end{bmatrix}$

- No need to compute & store $\mathbf{H}$ explicitly; only requires m HVPs

  $\Rightarrow$ Can be done efficiently via back-propagation [Pearlmutter'94]
- Bonus: $\tilde{\mathbf{H}}$ is a sparse tridiagonal matrix

Krylov Cubic Regularized Newton

- **Input:** Initial point $\boldsymbol{x}_0 \in \mathbb{R}^d$, subspace dimension m, regularization parameter $M > 0$
- **for** $k = 0, 1, \ldots,$ **do**
  - $(\mathbf{V}_k, \tilde{\boldsymbol{g}}_k, \tilde{\mathbf{H}}_k) \leftarrow \textsc{Lanczos}(\mathbf{H}_k, \boldsymbol{g}_k; m)$
  - Solve the cubic subproblem

$$\boldsymbol{z}_k = \operatorname*{argmin}_{\boldsymbol{z} \in \mathbb{R}^m} \left\{ \tilde{\boldsymbol{g}}_k^\top \boldsymbol{z} + \frac{1}{2} \boldsymbol{z}^\top \tilde{\mathbf{H}}_k \boldsymbol{z} + \frac{M}{6} \|\boldsymbol{z}\|^3 \right\},$$

  - Update $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \mathbf{V}_k \boldsymbol{z}_k$
- **end for**

- Per-iteration computational cost
  - Performing Lanczos iteration: m HVPs $\Rightarrow$ $\mathcal{O}(md)$
  - Solving the cubic subproblem: $\mathcal{O}(m)$

Outline

1. Introduction
2. The Proposed Method
3. Convergence Analysis
4. Numerical Results

Convergence Analysis

- For simplicity, we further assume that $\nabla^2 f(\mathbf{x}) \preceq L_1 \mathbf{I}$ as in [Hanzely et al.'20]

- For simplicity, we further assume that $\nabla^2 f(\boldsymbol{x}) \preceq L_1 \mathbf{I}$ as in [Hanzely et al.'20]

Lemma

Let $\{\boldsymbol{x}_k\}$ be generated by Krylov CRN with subspace dim m. We have

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) + \boldsymbol{g}_k^\top \boldsymbol{s} + \frac{1}{2}\boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s} + \frac{L_2}{6}\|\boldsymbol{s}\|^3 + \frac{L_1}{2m}\|\boldsymbol{s}\|^2, \quad \forall \boldsymbol{s} \in \mathbb{R}^d.$$

- For simplicity, we further assume that $\nabla^2 f(\boldsymbol{x}) \preceq L_1 \mathbf{I}$ as in [Hanzely et al.'20]

Lemma

Let $\{\boldsymbol{x}_k\}$ be generated by Krylov CRN with subspace dim m. We have

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) + \boldsymbol{g}_k^\top \boldsymbol{s} + \frac{1}{2}\boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s} + \frac{L_2}{6}\|\boldsymbol{s}\|^3 + \frac{L_1}{2m}\|\boldsymbol{s}\|^2, \quad \forall \boldsymbol{s} \in \mathbb{R}^d.$$

- The additional error term is independent of d

- For simplicity, we further assume that $\nabla^2 f(\boldsymbol{x}) \preceq L_1 \mathbf{I}$ as in [Hanzely et al.'20]

Lemma

Let $\{\boldsymbol{x}_k\}$ be generated by Krylov CRN with subspace dim m. We have

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) + \boldsymbol{g}_k^\top \boldsymbol{s} + \frac{1}{2}\boldsymbol{s}^\top \mathbf{H}_k \boldsymbol{s} + \frac{L_2}{6}\|\boldsymbol{s}\|^3 + \frac{L_1}{2m}\|\boldsymbol{s}\|^2, \quad \forall \boldsymbol{s} \in \mathbb{R}^d.$$

- The additional error term is independent of d
- It diminishes as m increases

Proof Sketch

- For simplicity, ignore the subscript k
- Let $\mathbf{P}^{(m)}$ be the orthogonal projection matrix associated with $\mathcal{K}_m(\mathbf{H}, \boldsymbol{g})$

Proof Sketch

- For simplicity, ignore the subscript k
- Let $\mathbf{P}^{(m)}$ be the orthogonal projection matrix associated with $\mathcal{K}_m(\mathbf{H}, \boldsymbol{g})$
- Recall the key inequality: for any $\boldsymbol{s} \in \mathbb{R}^d$,

$$f(\boldsymbol{x}_+) \leq f(\boldsymbol{x}) + \boldsymbol{g}^\top \mathbf{P}^{(m)} \boldsymbol{s} + \frac{1}{2}(\mathbf{P}^{(m)} \boldsymbol{s})^\top \mathbf{H} (\mathbf{P}^{(m)} \boldsymbol{s}) + \frac{L_2}{6} \|\mathbf{P}^{(m)} \boldsymbol{s}\|^3$$

Proof Sketch

- For simplicity, ignore the subscript k
- Let $\mathbf{P}^{(m)}$ be the orthogonal projection matrix associated with $\mathcal{K}_m(\mathbf{H}, \boldsymbol{g})$
- Recall the key inequality: for any $\boldsymbol{s} \in \mathbb{R}^d$,

$$f(\boldsymbol{x}_+) \leq f(\boldsymbol{x}) + \boldsymbol{g}^\top \mathbf{P}^{(m)} \boldsymbol{s} + \frac{1}{2} (\mathbf{P}^{(m)} \boldsymbol{s})^\top \mathbf{H} (\mathbf{P}^{(m)} \boldsymbol{s}) + \frac{L_2}{6} \|\mathbf{P}^{(m)} \boldsymbol{s}\|^3$$

- Since $\boldsymbol{g} \in \mathcal{K}_m(\mathbf{H}, \boldsymbol{g}) \;\Leftrightarrow\; \mathbf{P}^{(m)} \boldsymbol{g} = \boldsymbol{g}$ and $\|\mathbf{P}^{(m)} \boldsymbol{s}\| \leq \|\boldsymbol{s}\|$:

$$f(\boldsymbol{x}_+) \leq f(\boldsymbol{x}) + \boldsymbol{g}^\top \boldsymbol{s} + \frac{1}{2} \boldsymbol{s}^\top \mathbf{H} \boldsymbol{s} + \frac{L_2}{6} \|\boldsymbol{s}\|^3 + \frac{1}{2} \boldsymbol{s}^\top \left(\mathbf{P}^{(m)} \mathbf{H} \mathbf{P}^{(m)} - \mathbf{H} \right) \boldsymbol{s}$$

Proof Sketch

- For simplicity, ignore the subscript k
- Let $\mathbf{P}^{(m)}$ be the orthogonal projection matrix associated with $\mathcal{K}_m(\mathbf{H}, \boldsymbol{g})$
- Recall the key inequality: for any $\boldsymbol{s} \in \mathbb{R}^d$,

$$f(\boldsymbol{x}_+) \leq f(\boldsymbol{x}) + \boldsymbol{g}^\top \mathbf{P}^{(m)} \boldsymbol{s} + \frac{1}{2} (\mathbf{P}^{(m)} \boldsymbol{s})^\top \mathbf{H} (\mathbf{P}^{(m)} \boldsymbol{s}) + \frac{L_2}{6} \|\mathbf{P}^{(m)} \boldsymbol{s}\|^3$$

- Since $\boldsymbol{g} \in \mathcal{K}_m(\mathbf{H}, \boldsymbol{g}) \;\Leftrightarrow\; \mathbf{P}^{(m)} \boldsymbol{g} = \boldsymbol{g}$ and $\|\mathbf{P}^{(m)} \boldsymbol{s}\| \leq \|\boldsymbol{s}\|$:

$$f(\boldsymbol{x}_+) \leq f(\boldsymbol{x}) + \boldsymbol{g}^\top \boldsymbol{s} + \frac{1}{2} \boldsymbol{s}^\top \mathbf{H} \boldsymbol{s} + \frac{L_2}{6} \|\boldsymbol{s}\|^3 + \frac{1}{2} \boldsymbol{s}^\top \left(\mathbf{P}^{(m)} \mathbf{H} \mathbf{P}^{(m)} - \mathbf{H} \right) \boldsymbol{s}$$

- Recall the Lanczos algorithm generates $\{\boldsymbol{v}_j\}_{j=1}^m$ and $\{\beta_{j+1}\}_{j=1}^m$:

$$\frac{1}{2} \boldsymbol{s}^\top \left(\mathbf{P}^{(m)} \mathbf{H} \mathbf{P}^{(m)} - \mathbf{H} \right) \boldsymbol{s} \leq \beta_{m+1} |\boldsymbol{v}_m^\top \boldsymbol{s}| |\boldsymbol{v}_{m+1}^\top \boldsymbol{s}|$$

Proof Sketch (cont.)

- It can be shown that $\beta_{j+1} \leq L_1/2$ for all $j \geq 1$:

$$\frac{1}{2}\boldsymbol{s}^\top \left(\mathbf{P}^{(m)}\mathbf{H}\mathbf{P}^{(m)} - \mathbf{H}\right)\boldsymbol{s} \leq \beta_{m+1}|\boldsymbol{v}_m^\top \boldsymbol{s}||\boldsymbol{v}_{m+1}^\top \boldsymbol{s}| \leq \frac{L_1}{2}\|\boldsymbol{s}\|^2$$

Proof Sketch (cont.)

- It can be shown that $\beta_{j+1} \leq L_1/2$ for all $j \geq 1$:

$$\frac{1}{2}\boldsymbol{s}^\top \left(\mathbf{P}^{(m)}\mathbf{H}\mathbf{P}^{(m)} - \mathbf{H}\right)\boldsymbol{s} \leq \beta_{m+1}|\boldsymbol{v}_m^\top \boldsymbol{s}||\boldsymbol{v}_{m+1}^\top \boldsymbol{s}| \leq \frac{L_1}{2}\|\boldsymbol{s}\|^2$$

- However, the error term does not diminish when m increases

Proof Sketch (cont.)

- It can be shown that $\beta_{j+1} \leq L_1/2$ for all $j \geq 1$:

$$\frac{1}{2}\boldsymbol{s}^\top \left(\mathbf{P}^{(m)}\mathbf{H}\mathbf{P}^{(m)} - \mathbf{H}\right)\boldsymbol{s} \leq \beta_{m+1}|\boldsymbol{v}_m^\top \boldsymbol{s}||\boldsymbol{v}_{m+1}^\top \boldsymbol{s}| \leq \frac{L_1}{2}\|\boldsymbol{s}\|^2$$

- However, the error term does not diminish when m increases
- Turns out we can strengthen the bound to

$$f(\boldsymbol{x}_+) \leq f(\boldsymbol{x}) + \boldsymbol{g}^\top \boldsymbol{s} + \frac{1}{2}\boldsymbol{s}^\top \mathbf{H}\boldsymbol{s} + \frac{L_2}{6}\|\boldsymbol{s}\|^3 + \frac{1}{2}\min_{j\in\{1,\ldots,m\}}\left\{\boldsymbol{s}^\top \left(\mathbf{P}^{(j)}\mathbf{H}\mathbf{P}^{(j)} - \mathbf{H}\right)\boldsymbol{s}\right\},$$

where $\mathbf{P}^{(j)}$ be the orthogonal projection matrix associated with $\mathcal{K}_j(\mathbf{H}, \boldsymbol{g})$

Proof Sketch (cont.)

- It can be shown that $\beta_{j+1} \le L_1/2$ for all $j \ge 1$:

$$\frac{1}{2}\boldsymbol{s}^\top \left(\mathbf{P}^{(m)}\mathbf{H}\mathbf{P}^{(m)} - \mathbf{H}\right)\boldsymbol{s} \le \beta_{m+1}|\boldsymbol{v}_m^\top \boldsymbol{s}||\boldsymbol{v}_{m+1}^\top \boldsymbol{s}| \le \frac{L_1}{2}\|\boldsymbol{s}\|^2$$

- However, the error term does not diminish when m increases
- Turns out we can strengthen the bound to

$$f(\boldsymbol{x}_+) \le f(\boldsymbol{x}) + \boldsymbol{g}^\top \boldsymbol{s} + \frac{1}{2}\boldsymbol{s}^\top \mathbf{H}\boldsymbol{s} + \frac{L_2}{6}\|\boldsymbol{s}\|^3 + \frac{1}{2}\min_{j\in\{1,\ldots,m\}}\left\{\boldsymbol{s}^\top \left(\mathbf{P}^{(j)}\mathbf{H}\mathbf{P}^{(j)} - \mathbf{H}\right)\boldsymbol{s}\right\},$$

  where $\mathbf{P}^{(j)}$ be the orthogonal projection matrix associated with $\mathcal{K}_j(\mathbf{H}, \boldsymbol{g})$
- Hence,

$$\frac{1}{2}\min_{j\in\{1,\ldots,m\}}\left\{\boldsymbol{s}^\top \left(\mathbf{P}^{(j)}\mathbf{H}\mathbf{P}^{(j)} - \mathbf{H}\right)\boldsymbol{s}\right\} \le \frac{L_1}{2}\min_{j\in\{1,\ldots,m\}}|\boldsymbol{v}_j^\top \boldsymbol{s}||\boldsymbol{v}_{j+1}^\top \boldsymbol{s}|$$

Proof Sketch (cont.)

- It can be shown that $\beta_{j+1} \leq L_1/2$ for all $j \geq 1$:

$$\frac{1}{2}\boldsymbol{s}^\top \left(\mathbf{P}^{(m)}\mathbf{H}\mathbf{P}^{(m)} - \mathbf{H}\right)\boldsymbol{s} \leq \beta_{m+1}|\boldsymbol{v}_m^\top \boldsymbol{s}||\boldsymbol{v}_{m+1}^\top \boldsymbol{s}| \leq \frac{L_1}{2}\|\boldsymbol{s}\|^2$$

- However, the error term does not diminish when m increases
- Turns out we can strengthen the bound to

$$f(\boldsymbol{x}_+) \leq f(\boldsymbol{x}) + \boldsymbol{g}^\top \boldsymbol{s} + \frac{1}{2}\boldsymbol{s}^\top \mathbf{H}\boldsymbol{s} + \frac{L_2}{6}\|\boldsymbol{s}\|^3 + \frac{1}{2}\min_{j\in\{1,\dots,m\}}\left\{\boldsymbol{s}^\top \left(\mathbf{P}^{(j)}\mathbf{H}\mathbf{P}^{(j)} - \mathbf{H}\right)\boldsymbol{s}\right\},$$

  where $\mathbf{P}^{(j)}$ be the orthogonal projection matrix associated with $\mathcal{K}_j(\mathbf{H}, \boldsymbol{g})$
- Hence,

$$\frac{1}{2}\min_{j\in\{1,\dots,m\}}\left\{\boldsymbol{s}^\top \left(\mathbf{P}^{(j)}\mathbf{H}\mathbf{P}^{(j)} - \mathbf{H}\right)\boldsymbol{s}\right\} \leq \frac{L_1}{2}\min_{j\in\{1,\dots,m\}}|\boldsymbol{v}_j^\top \boldsymbol{s}||\boldsymbol{v}_{j+1}^\top \boldsymbol{s}|$$

- Since $\{\boldsymbol{v}_j\}$ are orthonormal, we further have $\min_{j\in\{1,\dots,m\}}|\boldsymbol{v}_j^\top \boldsymbol{s}||\boldsymbol{v}_{j+1}^\top \boldsymbol{s}| \leq \frac{1}{m}\|\boldsymbol{s}\|^2$

Main Result: Convex Setting

Theorem

Let $\{\mathbf{x}_k\}$ be generated by the Krylov CRN method. Then we have

$$f(\mathbf{x}_k) - f^* \leq \frac{9L_1D^2}{2mk} + \frac{9L_2D^3}{k^2}$$

- The convergence rate is independent of d

Main Result: Convex Setting

Theorem

Let $\{\boldsymbol{x}_k\}$ be generated by the Krylov CRN method. Then we have

$$f(\boldsymbol{x}_k) - f^* \leq \frac{9L_1D^2}{2mk} + \frac{9L_2D^3}{k^2}$$

- The convergence rate is independent of d
- To achieve accuracy ϵ, the number of iterations required is $\mathcal{O}\left(\frac{1}{m\epsilon} + \frac{1}{\sqrt{\epsilon}}\right)$

Main Result: Convex Setting

Theorem

Let $\{\mathbf{x}_k\}$ be generated by the Krylov CRN method. Then we have

$$f(\mathbf{x}_k) - f^* \leq \frac{9L_1D^2}{2mk} + \frac{9L_2D^3}{k^2}$$

- The convergence rate is independent of d
- To achieve accuracy ϵ, the number of iterations required is $\mathcal{O}\left(\frac{1}{m\epsilon} + \frac{1}{\sqrt{\epsilon}}\right)$
- In comparison, SSCN [Hanzely et al.'20] requires $\mathcal{O}(\frac{d-m}{m} \cdot \frac{1}{\epsilon} + \frac{d}{m} \cdot \frac{1}{\sqrt{\epsilon}})$

Main Result: Convex Setting

Theorem

Let $\{\mathbf{x}_k\}$ be generated by the Krylov CRN method. Then we have

$$f(\mathbf{x}_k) - f^* \leq \frac{9L_1 D^2}{2mk} + \frac{9L_2 D^3}{k^2}$$

- The convergence rate is independent of d
- To achieve accuracy ϵ, the number of iterations required is $\mathcal{O}\left(\frac{1}{m\epsilon} + \frac{1}{\sqrt{\epsilon}}\right)$
- In comparison, SSCN [Hanzely et al.'20] requires $\mathcal{O}(\frac{d-m}{m} \cdot \frac{1}{\epsilon} + \frac{d}{m} \cdot \frac{1}{\sqrt{\epsilon}})$
- When $m \ll d$, our complexity is lower by a factor of d

Main Result: Strongly Convex Setting

- We also consider the strongly convex setting, i.e., $\nabla^2 f(\boldsymbol{x}) \succeq \mu \mathbf{I}, \forall \boldsymbol{x} \in \mathbb{R}^d$

Theorem

Let $\{\boldsymbol{x}_k\}$ be generated by the Krylov CRN method. Then, the number of iterations required to reach $\delta_k := f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \epsilon$ can be upper bounded by

$$k = \mathcal{O}\left(\left(\frac{L_1}{m\mu} + 1\right) \log \frac{\delta_0}{\epsilon} + \frac{\sqrt{L_2}\delta_0^{0.25}}{\mu^{0.75}}\right)$$

- In comparison, SSCN requires $\mathcal{O}\left(\left(\frac{d-m}{m}\frac{L_1}{\mu} + \frac{d}{m}\right) \log \frac{1}{\epsilon}\right)$
- Again, when $m \ll d$, we shave a factor of d

Hessian Eigenspectrum

- The convergence bound depends on L_1, i.e., the largest eigenvalue of the Hessian

Hessian Eigenspectrum

- The convergence bound depends on L_1, i.e., the largest eigenvalue of the Hessian
- With a more refined analysis, we can replace L_1 by

$$\rho_{\max}^{(m)} := \max_{i \in \{0,1,\ldots,k-1\}} \{\rho^{(m)}(\mathbf{H}_i, \boldsymbol{g}_i)\},$$

Hessian Eigenspectrum

- The convergence bound depends on L_1, i.e., the largest eigenvalue of the Hessian
- With a more refined analysis, we can replace L_1 by

$$\rho^{(m)}_{\max} := \max_{i \in \{0,1,\ldots,k-1\}} \{\rho^{(m)}(\mathbf{H}_i, \boldsymbol{g}_i)\},$$

where $\rho^{(m)}(\mathbf{H}, \boldsymbol{g})$ is defined by

$$\rho^{(m)}(\mathbf{H}, \boldsymbol{g}) = \min_{c_0,\ldots,c_{m-1} \in \mathbb{R}} \left\| \mathbf{H}^m \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|} - \sum_{i=0}^{m-1} c_i \mathbf{H}^i \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|} \right\|^{\frac{1}{m}}$$

Hessian Eigenspectrum

- The convergence bound depends on L_1, i.e., the largest eigenvalue of the Hessian
- With a more refined analysis, we can replace L_1 by

$$\rho_{\max}^{(m)} := \max_{i \in \{0,1,\ldots,k-1\}} \{\rho^{(m)}(\mathbf{H}_i, \boldsymbol{g}_i)\},$$

where $\rho^{(m)}(\mathbf{H}, \boldsymbol{g})$ is defined by

$$\rho^{(m)}(\mathbf{H}, \boldsymbol{g}) = \min_{c_0,\ldots,c_{m-1} \in \mathbb{R}} \left\| \mathbf{H}^m \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|} - \sum_{i=0}^{m-1} c_i \mathbf{H}^i \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|} \right\|^{\frac{1}{m}}$$

- For a polynomial $p(x) = x^m - \sum_{i=0}^{m-1} c_i x^i$, define $p(\mathbf{H}) = \mathbf{H}^m - \sum_{i=0}^{m-1} c_i \mathbf{H}^i$

Hessian Eigenspectrum

- The convergence bound depends on L_1, i.e., the largest eigenvalue of the Hessian
- With a more refined analysis, we can replace L_1 by

$$\rho^{(m)}_{\max} := \max_{i \in \{0,1,\ldots,k-1\}} \{\rho^{(m)}(\mathbf{H}_i, \boldsymbol{g}_i)\},$$

where $\rho^{(m)}(\mathbf{H}, \boldsymbol{g})$ is defined by

$$\rho^{(m)}(\mathbf{H}, \boldsymbol{g}) = \min_{c_0,\ldots,c_{m-1} \in \mathbb{R}} \left\| \mathbf{H}^m \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|} - \sum_{i=0}^{m-1} c_i \mathbf{H}^i \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|} \right\|^{\frac{1}{m}}$$

- For a polynomial $p(x) = x^m - \sum_{i=0}^{m-1} c_i x^i$, define $p(\mathbf{H}) = \mathbf{H}^m - \sum_{i=0}^{m-1} c_i \mathbf{H}^i$
- Let $\mathcal{M}_m$ be the set of monic polynomials of degree m. Then we have

$$\rho^{(m)}(\mathbf{H}, \boldsymbol{g}) = \min_{p \in \mathcal{M}_m} \left\| p(\mathbf{H}) \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|} \right\|^{\frac{1}{m}} \le \min_{p \in \mathcal{M}_m} \|p(\mathbf{H})\|^{\frac{1}{m}}$$

Hessian Eigenspectrum

- Assume that the Hessian $\mathbf{H}$ has r distinct eigenvalues in decreasing order: $\lambda_1 > \lambda_2 > \cdots > \lambda_r$

Hessian Eigenspectrum

- Assume that the Hessian $\mathbf{H}$ has r distinct eigenvalues in decreasing order: $\lambda_1 > \lambda_2 > \cdots > \lambda_r$
- Then
$$\rho^{(m)}(\mathbf{H}, \boldsymbol{g}) \leq \min_{p \in \mathcal{M}_m} \|p(\mathbf{H})\|^{\frac{1}{m}} \leq \min_{p \in \mathcal{M}_m} \max_{i \in \{1,2,\ldots,r\}} |p(\lambda_i)|^{\frac{1}{m}}.$$

Hessian Eigenspectrum

- Assume that the Hessian $\mathbf{H}$ has r distinct eigenvalues in decreasing order: $\lambda_1 > \lambda_2 > \cdots > \lambda_r$
- Then

$$\rho^{(m)}(\mathbf{H}, \boldsymbol{g}) \leq \min_{p \in \mathcal{M}_m} \|p(\mathbf{H})\|^{\frac{1}{m}} \leq \min_{p \in \mathcal{M}_m} \max_{i \in \{1,2,\ldots,r\}} |p(\lambda_i)|^{\frac{1}{m}}.$$

- Example I: $\rho^{(m)}(\mathbf{H}, \boldsymbol{g}) \leq (\prod_{i=1}^{m} \lambda_i)^{\frac{1}{m}}$ when $m < r$, $\rho^{(m)}(\mathbf{H}, \boldsymbol{g}) = 0$ when $m \geq r$
  - If the rank of Hessian is at most $m-1$, then $\rho^{(m)}(\mathbf{H}, \boldsymbol{g}) = 0$

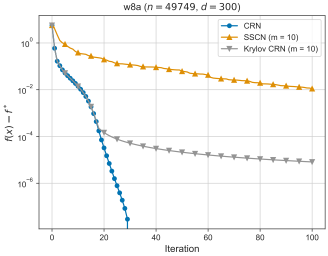
Hessian Eigenspectrum

- Assume that the Hessian $\mathbf{H}$ has r distinct eigenvalues in decreasing order: $\lambda_1 > \lambda_2 > \cdots > \lambda_r$
- Then
$$\rho^{(m)}(\mathbf{H}, \boldsymbol{g}) \leq \min_{p \in \mathcal{M}_m} \|p(\mathbf{H})\|^{\frac{1}{m}} \leq \min_{p \in \mathcal{M}_m} \max_{i \in \{1,2,\ldots,r\}} |p(\lambda_i)|^{\frac{1}{m}}.$$
- Example I: $\rho^{(m)}(\mathbf{H}, \boldsymbol{g}) \leq (\prod_{i=1}^{m} \lambda_i)^{\frac{1}{m}}$ when $m < r$, $\rho^{(m)}(\mathbf{H}, \boldsymbol{g}) = 0$ when $m \geq r$
  - If the rank of Hessian is at most $m-1$, then $\rho^{(m)}(\mathbf{H}, \boldsymbol{g}) = 0$
- Example II: Assume that all the eigenvalues of $\mathbf{H}$ lie within $[0, \Delta] \cup [L_1 - \Delta, L_1]$ for some $L_1 > \Delta > 0$. Then, when m is even, we have
$$\rho^{(m)}(\mathbf{H}, \boldsymbol{g}) \leq 2^{1/m} \sqrt{\Delta(L_1 - \Delta)}/2$$

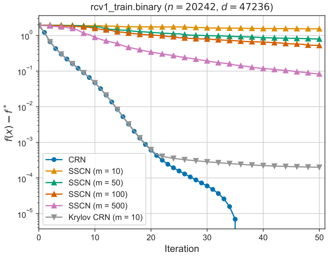
Outline

1. Introduction
2. The Proposed Method
3. Convergence Analysis
4. Numerical Results

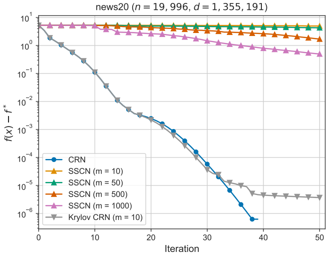
Logistic Regression Problems



(a) w8a ($d = 300$)



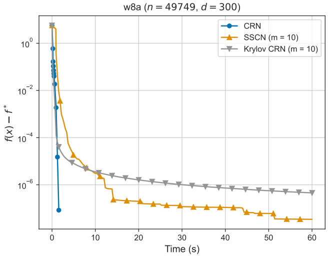
(b) rcv1 ($d = 47,236$)



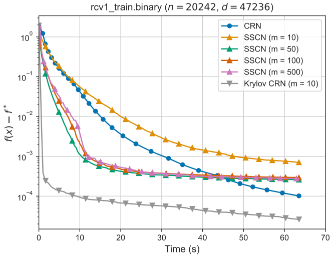
(c) news20 ($d = 1,355,191$)

- The convergence path of Krylov-CRN remains almost unchanged as d increases
- SSCN converges slower as d increases

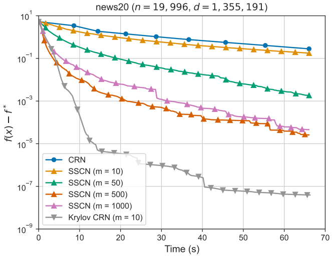
Logistic Regression Problems



(a) w8a ($d = 300$)



(b) rcv1 ($d = 47,236$)



(c) news20 ($d = 1,355,191$)

- When d is large, Krylov CRN converges much faster than the others

Reference

Ruichen Jiang, **Parameswaran Raman**, **Shoham Sabach**, **Aryan Mokhtari**, **Mingyi Hong**, and **Volkan Cevher**. Krylov Cubic Regularized Newton: A Subspace Second-Order Method with Dimension-Free Convergence Rate, 2024. arXiv: 2401.03058 [math.OC]

Thank you!