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

Ruichen Jiang	Parameswaran Raman	Shoham Sabach	Aryan Mokhtari	Mingyi Hong	Volkan Cevher
UT Austin	AWS	Technion & AWS	UT Austin	UMN & AWS	EPFL & AWS

TL;DR

- We propose a subspace cubic regularized Newton method for convex minimization
- The first dimension-independent global rate of $\mathcal{O}(\frac{1}{mk}+\frac{1}{k^2})$ among subspace methods, where m is the subspace dimension
- Key idea: use the Krylov subspace instead of a random subspace
- Empirically, our method with $m=10$ outperforms existing subspace methods in logistic regression problems with $d=10^6$

Cubic Regularized Newton Method

- The unconstrained minimization problem

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

  - $f:\mathbb{R}^d\to\mathbb{R}$ is convex
  - f is bounded from below and has bounded level-sets
  - The Hessian of f is Lipchitz, i.e., $\|\nabla^2 f(\boldsymbol{x})-\nabla^2 f(\boldsymbol{y})\|\le L_2\|\boldsymbol{x}-\boldsymbol{y}\|$, $\forall \boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d$
- Let $\boldsymbol{g}_k:=\nabla f(\boldsymbol{x}_k)$ and $\mathbf{H}_k:=\nabla^2 f(\boldsymbol{x}_k)$
- Cubic regularized Newton (CRN) [Griewank'81; Nesterov-Polyak'06]

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

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

Prior Works: Stochastic Subspace CRN

- The drawback of CRN: high memory and computational costs
  - Computing & storing the Hessian: $\mathcal{O}(d^2)$
  - Solving a cubic subproblem: $\mathcal{O}(d^3)$
- Subspace methods: executing 2nd-order updates in a subspace $\mathcal{V}_k$ of dim $m\ll d$ [Doikov-Richtárik'18; Gower et al.'19; Hanzely et al.'20]

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

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

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

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

- The existing works choose a random subspace $\mathcal{V}_k$
  - 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: $\mathcal{O}(m^2)$ in some special cases
  - Solving a 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?

Our Contributions: Krylov CRN

- We propose the Krylov CRN method, where $\mathcal{V}_k$ is chosen as the Krylov subspace $\operatorname{span}\{\boldsymbol{g}_k,\mathbf{H}_k\boldsymbol{g}_k,\dots,\mathbf{H}_k^{m-1}\boldsymbol{g}_k\}$
- Can be implemented using the Lanczos method, with 1 gradient evaluation and m Hessian-vector products (HVPs) 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)	$\mathcal{O}(md)^{**}$	$\mathcal{O}(\frac{1}{mk}+\frac{1}{k^2})$

*Assume computing the subspace gradient is $\mathcal{O}(m)$ and the subspace Hessian is $\mathcal{O}(m^2)$

**Assume the cost of Hessian-vector product evaluations is $\mathcal{O}(d)$

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

Krylov CRN

- **Input:** Initial point $\boldsymbol{x}_0\in\mathbb{R}^d$, subspace dimension m, regularization parameter $M>0$
- **for** $k=0,1,\dots,$ **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
  - Lanczos iteration: m HVPs $\Rightarrow$ $\mathcal{O}(md)$
  - Solving the cubic subproblem: $\mathcal{O}(m)$

$(\mathbf{V},\tilde{\boldsymbol{g}},\tilde{\mathbf{H}})=\textsc{Lanczos}(\mathbf{H},\boldsymbol{g};m)$

- **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:** $\mathbf{V}=[\boldsymbol{v}_1,\dots,\boldsymbol{v}_m]$, $\tilde{\boldsymbol{g}}=\|\boldsymbol{g}\|\boldsymbol{e}_1$, $\tilde{\mathbf{H}}=\begin{bmatrix}\alpha_1&\beta_2&&&\\\beta_2&\alpha_2&\beta_3&&\\&\beta_3&\ddots&\ddots&\\&&\ddots&\ddots&\beta_{m-1}\\&&&\beta_{m-1}&\alpha_{m-1}&\beta_m\\&&&&\beta_m&\alpha_m\end{bmatrix}$

Convergence Analysis

- Key inequality in the analysis of classical CRN:

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

- For subspace CRN, we instead have

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

Key Lemma

Suppose $\mathbf{H}_k\preceq L_1\mathbf{I}$ and Let $\{\boldsymbol{x}_k\}$ be generated by Krylov CRN with subspace dim m. We have

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

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

Main Results

Theorem (Convex Setting)

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

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

- Achieving accuracy ϵ requires $\mathcal{O}\left(\frac{1}{m\epsilon}+\frac{1}{\sqrt{\epsilon}}\right)$ iterations
- 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

Theorem (Strongly Convex Setting)

Let $\{\boldsymbol{x}_k\}$ be generated by Krylov CRN. Then the number of iterations needed to reach $\delta_k:=f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)\le\epsilon$ is 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 presented bounds rely on L_1, the Hessian's largest eigenvalue
- A more refined analysis: we can replace L_1 by

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

where

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

- For a monic polynomial $p(x)=x^m-\sum_{i=0}^{m-1}c_ix^i$, 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

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

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

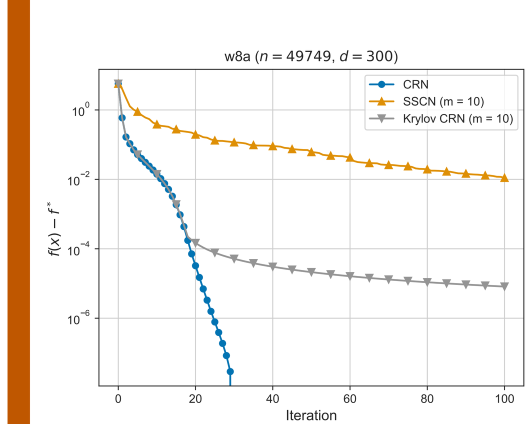
$$\rho^{(m)}(\mathbf{H},\boldsymbol{g})\le\min_{p\in\mathcal{M}_m}\max_{i\in\{1,2,\dots,r\}}|p(\lambda_i)|^{\frac{1}{m}}$$

- Example I: $\rho^{(m)}(\mathbf{H},\boldsymbol{g})\le\left(\prod_{i=1}^m\lambda_i\right)^{\frac{1}{m}}$ when $m\le r$
  - If the rank of Hessian $\le m-1$, 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 $L_1>\Delta>0$. When m is even,

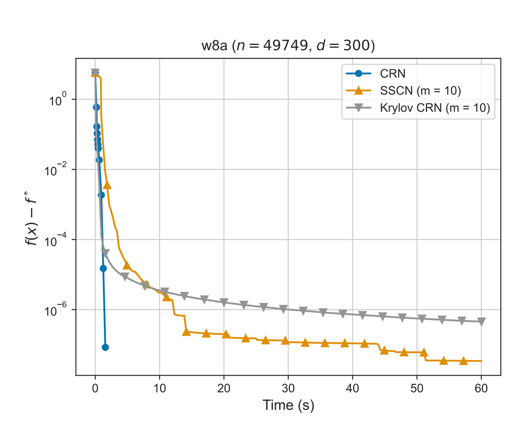
$$\rho^{(m)}(\mathbf{H},\boldsymbol{g})\le2^{1/m}\sqrt{\Delta(L_1-\Delta)}/2$$

Logistic Regression Problems

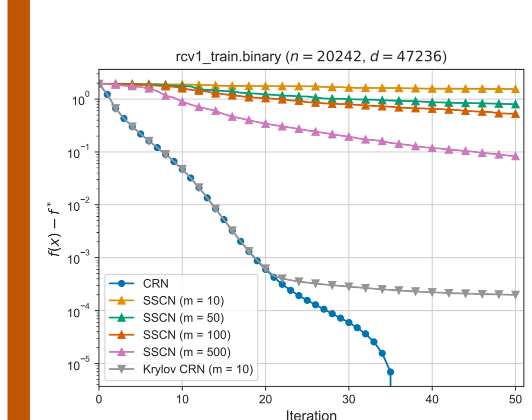
Plot by iteration: | Plot by time:



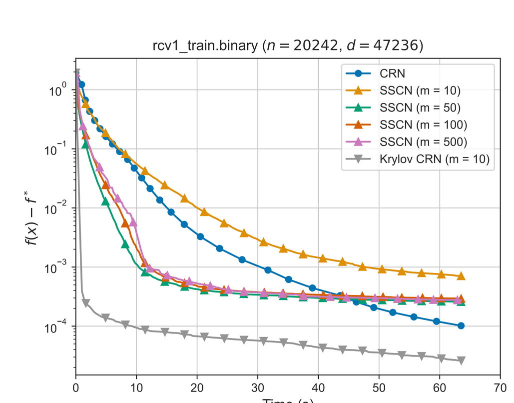
w8a ($d=300$)



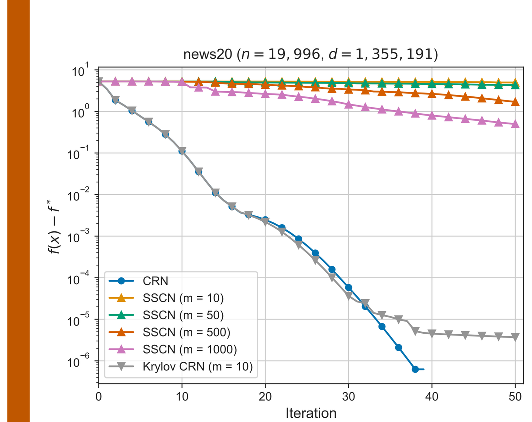
w8a ($d=300$)



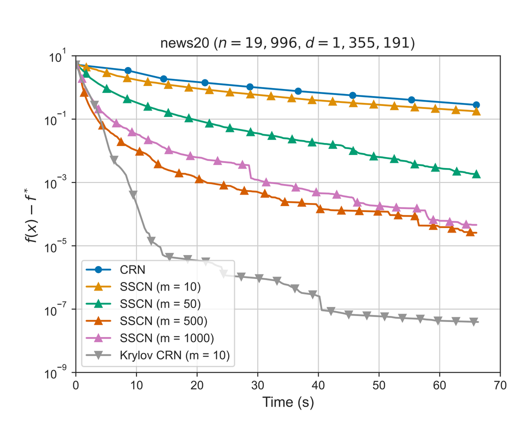
rcv1 ($d=47,236$)



rcv1 ($d=47,236$)



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



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

- In terms of iterations:
  - The convergence path of Krylov CRN remains almost unchanged as d increases from 300 to 10^6
  - SSCN converges slower as d increases
- In terms of time:
  - As d increases, the gap between Krylov CRN and other methods becomes wider

Take a picture to download the full paper