On Quadratic Convergence of DC Proximal Newton Algorithm in Nonconvex Sparse Learning

Xingguo Li*, Lin Yang◇, Jian Ge◇, Jarvis Haupt*, Tong Zhang⋆, Tuo Zhao†
*University of Minnesota ◇Princeton University ⋆Tencent Lab †Georgia Tech

Background

Consider the regularized optimization problem,

$$\overline{\theta} = \mathrm{argmin}_{\theta\in\mathbb{R}^d} \mathcal{L}(\theta) + \mathcal{R}_{\lambda_{\mathrm{tgt}}}(\theta), \quad (1)$$

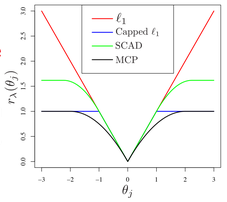
- $\mathcal{L}: \mathbb{R}^d \to \mathbb{R}$ is a twice differentiable convex loss function, e.g., negative log-likelihood for GLM,

$$\mathcal{L}(\theta) = \tfrac{1}{n}\textstyle\sum_{i=1}^n \ell_i(\theta),\ \ell_i(\theta) = \psi(x_i^\top\theta) - y_i x_i^\top\theta,$$

- $\mathcal{R}_{\lambda_{\mathrm{tgt}}}: \mathbb{R}^d \to \mathbb{R}$ is a sparsity-inducing decomposable regularizer, i.e., $\mathcal{R}_{\lambda_{\mathrm{tgt}}}(\theta) = \sum_{j=1}^d r_{\lambda_{\mathrm{tgt}}}(\theta_j)$ with $r_{\lambda_{\mathrm{tgt}}}: \mathbb{R}\to\mathbb{R}$, and $\lambda_{\mathrm{tgt}} > 0$ is the regularization parameter.

Convex vs. Nonconvex $\mathcal{R}_{\lambda_{\mathrm{tgt}}}$:

- Convex: easy optimization, large estimation bias (e.g., ℓ_1)
- Nonconvex: small estimation bias, hard optimization (e.g., SCAD, MCP, and Capped ℓ_1)

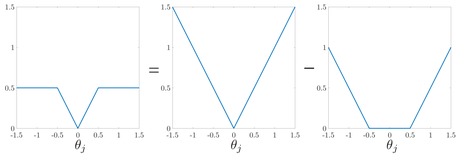


Consider the Capped ℓ_1 regularizer [2] defined as

$$\mathcal{R}_{\lambda_{\mathrm{tgt}}}(\theta) = \textstyle\sum_{j=1}^d r_{\mathrm{tgt}}(\theta_j) = \lambda_{\mathrm{tgt}}\sum_{j=1}^d \min\{|\theta_j|, \beta\lambda_{\mathrm{tgt}}\}, \quad (2)$$

where $\beta > 0$ is a tuning parameter. $r_{\lambda_{\mathrm{tgt}}}(\theta_j)$ can be decomposed as the **difference of convex functions**:

$$r_\lambda(\theta_j) = \underbrace{\lambda|\theta_j|}_{\text{convex}} - \underbrace{\max\{\lambda|\theta_j| - \beta\lambda^2, 0\}}_{\text{convex}}.$$

Capped ℓ_1: 

Our Approach

Algorithm 1 DC Proximal Newton Algorithm

Input: $\widehat{\theta}^{\{0\}}$, λ_{tgt}, β, ε
Warm Initialization: $\widehat{\theta}^{\{1\}} \leftarrow \mathrm{ProxNewton}(\widehat{\theta}^{\{0\}}, \lambda_{\mathrm{tgt}}, \varepsilon)$, $K \leftarrow 1$
Repeat:

$$\lambda_j^{\{K+1\}} \leftarrow \begin{cases} 0, & \text{if } |\widehat{\theta}_j^{\{K\}}| > \beta\lambda_{\mathrm{tgt}} \\ \lambda_{\mathrm{tgt}}, & \text{if } |\widehat{\theta}_j^{\{K\}}| \le \beta\lambda_{\mathrm{tgt}} \end{cases}$$

$t \leftarrow 0$, $\theta^{(0)} = \widehat{\theta}^{\{K\}}$
Repeat:
$\theta^{(t+1)} \leftarrow \mathrm{argmin}_\theta\ \mathcal{Q}(\theta; \theta^{(t)}, \lambda^{\{K+1\}})$
$t \leftarrow t+1$
Until $\omega_{\lambda^{\{K+1\}}}(\theta^{(t)}) \le \varepsilon$
$\widehat{\theta}^{\{K+1\}} \leftarrow \theta^{(t)}$
$K \leftarrow K+1$
Until Convergence
Return: $\widehat{\theta}^{\{K\}}$

DC Proximal Newton Algorithm

Our DC proximal Newton algorithm contains three components:

(I) The multistage convex relaxation. At the $\{K+1\}$-th stage, we solve a convex relaxation of (1) at $\theta = \widehat{\theta}^{\{K\}}$ as

$$\overline{\theta}^{\{K+1\}} = \mathrm{argmin}_{\theta\in\mathbb{R}^d} \mathcal{L}(\theta) + ||\lambda^{\{K+1\}} \odot \theta||_1, \quad (3)$$

where $\lambda^{\{K+1\}} = \left(\lambda_1^{\{K+1\}}, ..., \lambda_d^{\{K+1\}}\right)^\top$, $\lambda_j^{\{K+1\}} = \lambda_{\mathrm{tgt}} \cdot \mathbb{1}\left(|\widehat{\theta}_j^{\{K\}}| \le \beta\lambda_{\mathrm{tgt}}\right)$ for all $j = 1, \ldots, d$.

(II) The warm initialization. In the first stage of DC programming, we solve the ℓ_1-regularized counterpart of (1) as

$$\overline{\theta}^{\{1\}} = \mathrm{argmin}_{\theta\in\mathbb{R}^d} \mathcal{L}(\theta) + \lambda_{\mathrm{tgt}}||\theta||_1. \quad (4)$$

(III) The proximal Newton algorithm. At t-th iteration of the proximal Newton algorithm, we consider a quadratic approximation of (3) at $\theta^{(t)}$ as

$$\mathcal{Q}(\theta; \theta^{(t)}, \lambda^{\{K\}}) = \mathcal{L}(\theta^{(t)}) + (\theta - \theta^{(t)})^\top \nabla\mathcal{L}(\theta^{(t)}) + \tfrac{1}{2}||\theta - \theta^{(t)}||^2_{\nabla^2\mathcal{L}(\theta^{(t)})} + ||\lambda^{\{K\}} \odot \theta||_1, \quad (5)$$

where $||\theta - \theta^{(t)}||^2_{\nabla^2\mathcal{L}(\theta^{(t)})} = (\theta - \theta^{(t)})^\top \nabla^2\mathcal{L}(\theta^{(t)})(\theta - \theta^{(t)})$. Then

$$\theta^{(t+1)} = \mathrm{argmin}_\theta\ \mathcal{Q}(\theta; \theta^{(t)}, \lambda^{\{K\}}).$$

Backtracking line search is used in the warm initialization. We terminate the iterations when the *approximate KKT condition* holds:

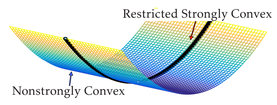
$$\omega_{\lambda^{\{1\}}}\left(\theta^{(t)}\right) = \min_{\xi\in\partial||\theta^{(t)}||_1} ||\nabla\mathcal{L}(\theta^{(t)}) + \lambda^{\{1\}} \odot \xi||_\infty \le \varepsilon,$$

Assumptions

Assumption 1 (**Local Restricted Sparse Eigenvalue** (SE)). *Given $\theta \in \mathcal{B}(\theta^*, R) = \left\{\phi \in \mathbb{R}^d \mid ||\phi - \theta^*||_2 \le R\right\}$ for a generic constant R, $\exists$ a constant C_0 such that $\nabla^2\mathcal{L}(\theta)$ satisfies the SE properties with parameters $\rho^-_{s^*+2\widetilde{s}}$ and $\rho^+_{s^*+2\widetilde{s}}$ satisfying*

$$0 < \rho^-_{s^*+2\widetilde{s}} = \inf_v \tfrac{v^\top\nabla^2\mathcal{L}(\theta)v}{v^\top v} < \rho^+_{s^*+2\widetilde{s}} = \sup_v \tfrac{v^\top\nabla^2\mathcal{L}(\theta)v}{v^\top v} < +\infty$$

where $\|v\|_0 \le s^* + 2\widetilde{s}$, $\widetilde{s} \ge C_0\kappa^2_{s^*+2\widetilde{s}}\, s^*$, and $\kappa_{s^*+2\widetilde{s}} = \rho^+_{s^*+2\widetilde{s}}/\rho^-_{s^*+2\widetilde{s}}$.



Assumption 2 (**Local Restricted Hessian Smoothness**). *$\exists$ generic constants $L_{s^*+2\widetilde{s}}$ and R such that for any $\theta, \theta' \in \mathcal{B}(\theta^*, R)$ with $||\theta_{\mathcal{S}_\perp}||_0 \le \widetilde{s}$ and $||\theta'_{\mathcal{S}_\perp}||_0 \le \widetilde{s}$, we have*

$$||\nabla^2\mathcal{L}(\theta) - \nabla^2\mathcal{L}(\theta')||_2 \le L_{s^*+2\widetilde{s}}||\theta - \theta'||_2.$$

Assumption 3. *Given the true parameter θ^*, $\exists$ generic constant C_1 such that $\lambda_{\mathrm{tgt}} = C_1\sqrt{\frac{\log d}{n}} \ge 4||\nabla\mathcal{L}(\theta^*)||_\infty$. Moreover, for large enough n, we have $\sqrt{s^*}\lambda_{\mathrm{tgt}} \le C_2 R\rho^-_{s^*+2\widetilde{s}}$ with $R = \frac{\rho^-_{s^*+2\widetilde{s}}}{2L_{s^*+2\widetilde{s}}}$.*

Assumption 4. *For each stage of solving the convex relaxed subproblem (3) for all $K \ge 1$, we set $\varepsilon = \frac{C_3}{\sqrt{n}} \le \frac{\lambda_{\mathrm{tgt}}}{8}$ for some generic small constant C_3.*

Computational Theory

For K-th stage of convex relaxation, we denote
$\widehat{\theta}^{\{K\}}$:output solution of (3) $\Rightarrow \omega_{\lambda^{\{K\}}}(\widehat{\theta}^{\{K\}}) \le \varepsilon$,
$\overline{\theta}^{\{K\}}$:uniq. global min. of (3) $\Rightarrow \omega_{\lambda^{\{K\}}}(\overline{\theta}^{\{K\}}) = 0$, $||\overline{\theta}^{\{K\}}_{\mathcal{S}_\perp}||_0 \le \widetilde{s}$.

Theorem 1 (Warm Initialization, $K = 1$). *Suppose Assumptions 1 $\sim$ 4 hold. For large enough $T < \infty$, we have for all $t \ge T$,*

$$||\theta^{(t)}_{\mathcal{S}_\perp}||_0 \le \widetilde{s},\ \ \textit{and}\ \ ||\theta^{(t+1)} - \overline{\theta}^{\{1\}}||_2 \le \tfrac{L_{s^*+2\widetilde{s}}}{2\rho^-_{s^*+2\widetilde{s}}}||\theta^{(t)} - \overline{\theta}^{\{1\}}||_2^2.$$

Moreover, we need $\le T + \log\log\left(3\rho^+_{s^*+2\widetilde{s}}/\varepsilon\right)$ iterations to terminate the prox. Newton update for (4), where $\widehat{\theta}^{\{1\}}$ satisfies

$$||\widehat{\theta}^{\{1\}}_{\mathcal{S}_\perp}||_0 \le \widetilde{s}\ \ \textit{and}\ \ ||\widehat{\theta}^{\{1\}} - \theta^*||_2 \le \tfrac{18\lambda_{\mathrm{tgt}}\sqrt{s^*}}{\rho^-_{s^*+2\widetilde{s}}}.$$

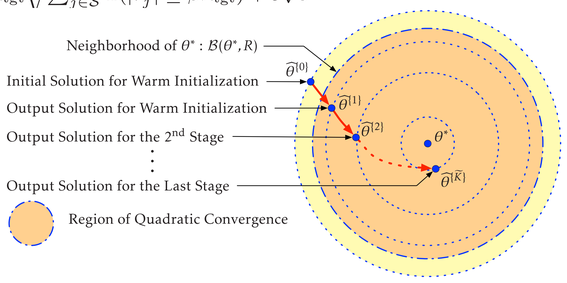
Theorem 2 (Stage K, $K \ge 2$). *Suppose Assumptions 1 $\sim$ 4 hold. Then within each stage $K \ge 2$, for all iterations $t \ge 1$, we have*

$$||\theta^{(t)}_{\mathcal{S}_\perp}||_0 \le \widetilde{s}\ \ \textit{and}\ \ ||\theta^{(t+1)} - \overline{\theta}^{\{K\}}||_2 \le \tfrac{L_{s^*+2\widetilde{s}}}{2\rho^-_{s^*+2\widetilde{s}}}||\theta^{(t)} - \overline{\theta}^{\{K\}}||_2^2,$$

Moreover, we need $\le \log\log\left(3\rho^+_{s^*+2\widetilde{s}}/\varepsilon\right)$ iterations to terminate the prox. Newton update for (3), where $\widehat{\theta}^{\{K\}}$ satisfies

$$||\widehat{\theta}^{\{K\}}_{\mathcal{S}_\perp}||_0 \le \widetilde{s}\ \textit{and}\ ||\widehat{\theta}^{\{K\}} - \theta^*||_2 \le C_2 0.7^{K-1}||\widehat{\theta}^{\{1\}} - \theta^*||_2 + C_3\alpha$$

for some constants C_2 and C_3, where $\alpha = \|\nabla\mathcal{L}(\theta^*)_{\mathcal{S}}\|_2 + \lambda_{\mathrm{tgt}}\sqrt{\sum_{j\in\mathcal{S}}\mathbb{1}(|\theta^*_j| \le \beta\lambda_{\mathrm{tgt}})} + \varepsilon\sqrt{s^*}$.

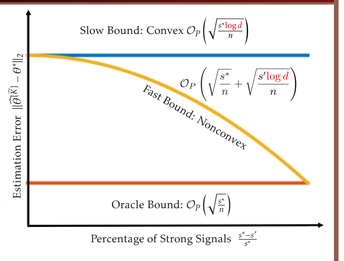


Statistical Theory

Suppose x_i's are i.i.d. samples from a *sub-Gaussian* distribution with:

- bounded eigenvalues of covariance, and
- other mild conditions.

Then Assumptions 1 $\sim$ 3 hold w.h.p.



Theorem 3. *Suppose $\{x_i, y_i\}_{i=1}^n$ are generated from GLM satisfying mild conditions with large enough n such that $n \ge C_4 s^*\log d$ and $\beta = C_5/c_{\min}$ is defined in (2), then w.h.p., $\widehat{\theta}^{\{K\}}$ satisfies*

$$||\widehat{\theta}^{\{K\}} - \theta^*||_2 \le C_6\left(\sqrt{\frac{s^*}{n}} + \sqrt{\frac{s'\log d}{n}}\right) + C_7 0.7^K\left(\sqrt{\frac{s^*\log d}{n}}\right)$$

for constants $C_4 \sim C_7$, where $s' = \sum_{j\in\mathcal{S}}\mathbb{1}(|\theta^*_j| \le \beta\lambda_{\mathrm{tgt}})$.

Experiments

Competing algorithms:

- DC+PN: our proposed DC proximal Newton algorithm
- DC+ACD: DC coordinate descent algorithm combined with the active set strategy
- DC+APG: DC accelerated proximal gradient algorithm

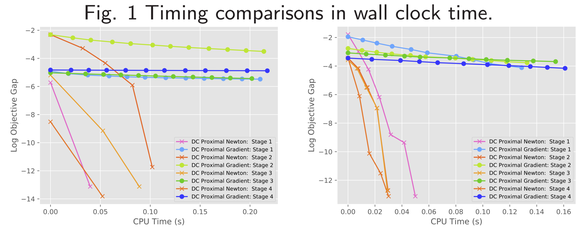
Datasets:

- "madelon" ($n = 2000$, $d = 500$, [1])
- "gisette" ($n = 2000$, $d = 5000$, [1])
- simulated datasets: "sim_1k", "sim_5k", and "sim_10k" ($n = 1000$, $d =$1k, 5k, 10k)

Tab. 1 Quantitive timing comparisons for on nonconvex-regularized sparse logistic regression. The average values and standard errors (in parentheses) of timing performance (in seconds) are presented.

	DC+PN	DC+ACD	DC+APG
madelon	**1.51**($\pm$0.01) obj value: 0.52	**5.83**($\pm$0.03) obj value: 0.52	**1.60**($\pm$0.03) obj value: 0.52
gisette	**5.35**($\pm$0.11) obj value: 0.01	**18.92**($\pm$2.25) obj value: 0.01	**207**($\pm$2.25) obj value: 0.01
sim_1k	**1.07**($\pm$0.02) obj value: 0.01	**9.46**($\pm$0.09) obj value: 0.01	**17.8**($\pm$1.23) obj value: 0.01
sim_5k	**4.53**($\pm$0.06) obj value: 0.01	**16.20**($\pm$0.24) obj value: 0.01	**111**($\pm$1.28) obj value: 0.01
sim_10k	**8.82**($\pm$0.04) obj value: 0.01	**19.1**($\pm$0.56) obj value: 0.01	**222**($\pm$5.79) obj value: 0.01

Fig. 1 Timing comparisons in wall clock time.



References

[1] I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror. Result analysis of the nips 2003 feature selection challenge. In *Advances in neural information processing systems*, pages 545–552, 2005.

[2] T. Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11:1081–1107, 2010.

Acknowledgment

The authors acknowledge support from DARPA YFA N66001-14-1-4047 and NSF Grant IIS-1447639.

University of Minnesota · Princeton University · Tencent 腾讯 · Georgia Tech