Sub-Sampled Newton Methods

F. Roosta-Khorasani and M. W. Mahoney

ICSI and Dept of Statistics, UC Berkeley
February 2016

Outline

- Problem Statement
- Rough Overview of Existing Methods
  - First Order methods
  - Second order methods
- SSN:
  - Globally convergent algorithms
    - Hessian Sub-Sampling
    - Gradient & Hessian Sub-Sampling
  - Local convergent rates
    - Hessian Sub-Sampling
    - Gradient & Hessian Sub-Sampling
- Examples

- Problem Statement
- Rough Overview of Existing Methods
  - First Order methods
  - Second order methods
- SSN:
  - Globally convergent algorithms
    - Hessian Sub-Sampling
    - Gradient & Hessian Sub-Sampling
  - Local convergent rates
    - Hessian Sub-Sampling
    - Gradient & Hessian Sub-Sampling
- Examples

Problem Statement

Problem

$$\min_{\mathbf{x} \in \mathcal{D} \cap \mathcal{X}} F(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x})$$

- Convex constraint set: $\mathcal{X} \subseteq \mathbb{R}^p$
- Convex and open domain of F: $\mathcal{D} = \bigcap_{i=1}^{n} \text{dom}(f_i)$
- *Large Scale:*
    - $n \gg 1$
    - $p \gg 1$

Examples

- Machine Learning and Data fitting
  - Each f_i corresponds to an observation (or a measurement) which models the loss (or misfit)
  - Logistic regression
  - SVM
  - Neural Networks
  - Graphical Models
- Nonlinear inverse problems
  - e.g., PDE inverse problems

Iterative Scheme

$$x_{k+1} = \arg\min_{\mathbf{x} \in \mathcal{D} \cap \mathcal{X}} \left\{ F(\mathbf{x}^{(k)}) + (\mathbf{x} - \mathbf{x}^{(k)})^T \mathbf{g}(\mathbf{x}^{(k)}) + \frac{1}{2\alpha_k} (\mathbf{x} - \mathbf{x}^{(k)})^T H(\mathbf{x}^{(k)}) (\mathbf{x} - \mathbf{x}^{(k)}) \right\},$$

where

- $\mathbf{g}(\mathbf{x}) \approx \nabla F(\mathbf{x})$
- $H(\mathbf{x}) \approx \nabla^2 F(\mathbf{x})$

- Newton: $\mathbf{g}(\mathbf{x}^{(k)}) = \nabla F(\mathbf{x}^{(k)})$ and $H(\mathbf{x}^{(k)}) = \nabla^2 F(\mathbf{x}^{(k)})$
- Projected Gradient Descent: $\mathbf{g}(\mathbf{x}^{(k)}) = \nabla F(\mathbf{x}^{(k)})$ and $H(\mathbf{x}^{(k)}) = \mathbb{I}$
- Frank-Wolfe: $\mathbf{g}(\mathbf{x}^{(k)}) = \nabla F(\mathbf{x}^{(k)})$ and $H(\mathbf{x}^{(k)}) = 0$
- (mini-batch) SGD: $\mathbf{g}(\mathbf{x}^{(k)}) = 1/|\mathcal{S}_{\mathbf{g}}| \sum_{j \in \mathcal{S}_{\mathbf{g}}} \nabla f_j(\mathbf{x}^{(k)})$ and $H(\mathbf{x}^{(k)}) = \mathbb{I}$,
- SSN:
  - SSN w. Hessian Sub-Sampling:

$$\mathbf{g}(\mathbf{x}^{(k)}) = \nabla F(\mathbf{x}^{(k)})$$
$$H(\mathbf{x}^{(k)}) = 1/|\mathcal{S}_H| \sum_{j \in \mathcal{S}_H} \nabla^2 f_i(\mathbf{x}^{(k)})$$

  - SNN w. Gradient and Hessian Sub-Sampling:

$$\mathbf{g}(\mathbf{x}^{(k)}) = 1/|\mathcal{S}_{\mathbf{g}}| \sum_{j \in \mathcal{S}_{\mathbf{g}}} \nabla f_j(\mathbf{x}^{(k)})$$
$$H(\mathbf{x}^{(k)}) = 1/|\mathcal{S}_H| \sum_{j \in \mathcal{S}_H} \nabla^2 f_i(\mathbf{x}^{(k)})$$

Modern "Big-Data"

- $n \gg 1$
  - Evaluation $\nabla F(\mathbf{x})$ and $\nabla^2 F(\mathbf{x})$ scales "linearly" in n
- $p \gg 1$
  - Evaluation of each $\nabla f_i(\mathbf{x})$ and $\nabla^2 f_i(\mathbf{x})$ can be expensive
  - The "best" direction of descent $\Rightarrow$ computationally expensive
- Classical *deterministic* optimization algorithms $\rightarrow$ Inefficient
- Need to design *stochastic* variants
  - Should be efficient
  - Should preserve as much of original speed as possible

- Problem Statement
- Rough Overview of Existing Methods
  - First Order methods
  - Second order methods
- SSN:
  - Globally convergent algorithms
    - Hessian Sub-Sampling
    - Gradient & Hessian Sub-Sampling
  - Local convergent rates
    - Hessian Sub-Sampling
    - Gradient & Hessian Sub-Sampling
- Examples

First Order Methods

- Use only gradient information
  - E.g. : Gradient Descent:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \nabla F(\mathbf{x}^{(k)})$$

    - Smooth Convex $F \Rightarrow$ Sublinear, $\mathcal{O}(1/k)$
    - Smooth Strongly Convex $F \Rightarrow$ Linear, $\mathcal{O}(\rho^k), \ \rho < 1$
    - However, iteration cost scales linearly in n

First Order Methods

- Stochastic variants e.g., (mini-batch) SGD
  - $\mathcal{S} \subset \{1, 2, \cdots, n\}$ is chosen at random with $|\mathcal{S}| \ll n$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \sum_{j \in \mathcal{S}} \nabla f_j(\mathbf{x}^{(k)})$$

  - Cheap Per-Iteration costs!
  - However slower to converge:
    - Smooth Convex $F \Rightarrow \mathcal{O}(1/\sqrt{k})$
    - Smooth Strongly Convex $F \Rightarrow \mathcal{O}(1/k)$
    - Devise modifications to
      - achieve the convergence rate of the full GD
      - preserve the per-iteration cost of SGD
      - E.g.: SAG, SDCA, SVRG,...

Second Order Methods

- Use both gradient and Hessian information
  - E.g. : Newton's method:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - [\nabla^2 F(\mathbf{x}^{(k)})]^{-1} \nabla F(\mathbf{x}^{(k)})$$

    - Smooth Convex $F \Rightarrow$ Locally Superlinear
    - Smooth Strongly Convex $F \Rightarrow$ Locally Quadratic
    - However, iteration cost is high!

Second Order Methods

- Approximating second order information cheaply
  - Quasi-Newton, e.g., BFGS and L-BFGS [Nocedal, 1980]
  - Sketching the Hessian [Pilanci et al., 2015]
  - Sub-Sampling the Hessian [Byrd et al., 2011, Erdogdu et al., 2015, Martens, 2010, RM-I & RM-II, 2016]
    - $\mathcal{S} \subset \{1, 2, \cdots, n\}$ is chosen at random with $|\mathcal{S}| \ll n$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \Big[\sum_{j \in \mathcal{S}} \nabla^2 f_j(\mathbf{x}^{(k)}) \Big]^{-1} \nabla F(\mathbf{x}^{(k)})$$

  - Sampling the Hessian and the gradient [RM-I & RM-II, 2016]
    - $\mathcal{S}_H, \mathcal{S}_\mathbf{g} \subset \{1, 2, \cdots, n\}$ is chosen at random with $|\mathcal{S}_H|, |\mathcal{S}_\mathbf{g}| \ll n$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \Big[\sum_{j \in \mathcal{S}_H} \nabla^2 f_j(\mathbf{x}^{(k)}) \Big]^{-1} \sum_{j \in \mathcal{S}_\mathbf{g}} \nabla f_j(\mathbf{x}^{(k)})$$

Outline

- Problem Statement
- Rough Overview of Existing Methods
    - First Order methods
    - Second order methods
- SSN:
    - Globally convergent algorithms
        - Hessian Sub-Sampling
        - Gradient & Hessian Sub-Sampling
    - Local convergent rates
        - Hessian Sub-Sampling
        - Gradient & Hessian Sub-Sampling
- Examples

SSN [RM-I & RM-II, 2016]

- Globally Convergent Algorithms [RM-I, 2016]
  - Approach the optimum, $\mathbf{x}^*$, from any $\mathbf{x}^{(0)}$
- Local Convergence Rate [RM-II, 2016]
  - Achieve fast rate, at least locally

Combine $\Rightarrow$ Globally convergent algorithms with fast local rates!!!

Assumptions

- Unconstrained, i.e., $\mathcal{X} = \mathcal{D} = \mathbb{R}^p$
- Each f_i is smooth and convex

$$\nabla^2 f_i(\mathbf{x}) \succeq 0, \quad \forall \mathbf{x} \in \mathbb{R}^p, \quad i = 1, 2, \ldots, n,$$
$$\nabla^2 f_i(\mathbf{x}) \preceq K < \infty, \quad \forall \mathbf{x} \in \mathbb{R}^p, \quad i = 1, 2, \ldots, n,$$
$$\|\nabla^2 f_i(\mathbf{x}) - \nabla^2 f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p, \quad i = 1, 2, \ldots, n.$$

- F is strongly convex

$$\nabla^2 F(\mathbf{x}) \succeq \gamma, \quad \forall \mathbf{x} \in \mathbb{R}^p.$$

- condition number: $\kappa := K/\gamma$

See [RM-II,2016] for general constraints and relaxed assumptions.

Outline

- Problem Statement
- Rough Overview of Existing Methods
  - First Order methods
  - Second order methods
- SSN:
  - Globally convergent algorithms
    - Hessian Sub-Sampling
    - Gradient & Hessian Sub-Sampling
  - Local convergent rates
    - Hessian Sub-Sampling
    - Gradient & Hessian Sub-Sampling
- Examples

Globally Convergent Algorithms

- Requirements:

(R.1) $|\mathcal{S}|$ must be independent of n, or at least smaller than n

(R.2) $H(\mathbf{x})$ must be, at least, invertible

(R.3) $\mathbf{g}(\mathbf{x})$ must be close to $\nabla F(\mathbf{x})$

(R.4) Global convergence guarantee

(R.5) For $p \gg 1$, allow for inexactness in solving the linear system

Outline

- Problem Statement
- Rough Overview of Existing Methods
  - First Order methods
  - Second order methods
- SSN:
  - Globally convergent algorithms
    - Hessian Sub-Sampling ⇒ SSN-H
    - Gradient & Hessian Sub-Sampling
  - Local convergent rates
    - Hessian Sub-Sampling
    - Gradient & Hessian Sub-Sampling
- Examples

Hessian Sub-Sampling

$$\mathbf{g}(\mathbf{x}) = \nabla F(\mathbf{x})$$

$$H(\mathbf{x}) = \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} \nabla^2 f_j(\mathbf{x})$$

Sub-Sampling Hessian

- Satisfying Requirements (R.1) and (R.2)

Lemma (Uniform Hessian Sub-Sampling)

Given any $0 < \epsilon < 1$, $0 < \delta < 1$, *and* $\mathbf{x} \in \mathbb{R}^p$, *if*

$$|\mathcal{S}| \geq \frac{2\kappa \ln(p/\delta)}{\epsilon^2},$$

then

$$\Pr\left((1-\epsilon)\gamma \leq \lambda_{\min}\left(H(\mathbf{x})\right)\right) \geq 1 - \delta.$$

SSN-H with Exact Update

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{p}_k,$$

where

$$\begin{aligned}
\mathbf{p}_k &= -[H(\mathbf{x}^{(k)})]^{-1} \nabla F(\mathbf{x}^{(k)}), \quad \text{“exact solve”} \\
\alpha_k &= \arg\max \quad \alpha \\
&\quad \text{s.t.} \quad \alpha \leq \widehat{\alpha} \\
&\qquad\quad F(\mathbf{x}^{(k)} + \alpha \mathbf{p}_k) \leq F(\mathbf{x}^{(k)}) + \alpha \beta \mathbf{p}_k^T \nabla F(\mathbf{x}^{(k)})
\end{aligned}$$

$0 < \beta < 1$, $\widehat{\alpha} \geq 1$

SSN-H Algorithm: Exact Update

Algorithm 1 Globally Convergent SSN-H with exact solve

Input: $\mathbf{x}^{(0)}$, $0 < \delta < 1$, $0 < \epsilon < 1$, $0 < \beta < 1$, $\widehat{\alpha} \geq 1$
- Set the sample size, $|\mathcal{S}|$, with ϵ and δ
for $k = 0, 1, 2, \cdots$ until termination **do**
    - Select a sample set, $\mathcal{S}$, of size $|\mathcal{S}|$
    - Form $H(\mathbf{x}^{(k)})$
    - Update $\mathbf{x}^{(k+1)}$ with exact solve
end for

Gloabl Convergence of SSN-H: Exact Update

- Satisfying requirement (R.3)

Theorem (Global Convergence of Algorithm 1)

Using Algorithm 1 with any $\mathbf{x}^{(k)} \in \mathbb{R}^p$, with probability $1-\delta$, we have

$$F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^*) \leq (1-\rho)\big(F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*)\big),$$

where $\rho = 2\alpha_k \beta/\kappa$. Moreover, the step size is at least

$$\alpha_k \geq 2(1-\beta)(1-\epsilon)/\kappa.$$

SSN-H with Inexact Update

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{p}_k,$$

where

$$\|H(\mathbf{x}^{(k)})\mathbf{p}_k + \nabla F(\mathbf{x}^{(k)})\| \leq \theta_1 \|\nabla F(\mathbf{x}^{(k)})\|$$

$$\mathbf{p}_k^T \nabla F(\mathbf{x}^{(k)}) \leq -(1-\theta_2)\mathbf{p}_k^T H(\mathbf{x}^{(k)})\mathbf{p}_k$$

$$\begin{aligned} \alpha_k = \arg\max \quad & \alpha \\ \text{s.t.} \quad & \alpha \leq \widehat{\alpha} \\ & F(\mathbf{x}^{(k)} + \alpha \mathbf{p}_k) \leq F(\mathbf{x}^{(k)}) + \alpha\beta \mathbf{p}_k^T \nabla F(\mathbf{x}^{(k)}) \end{aligned}$$

$0 < \beta < 1$, $\widehat{\alpha} \geq 1$, $0 < \theta_1, \theta_2 < 1$

SSN-H Algorithm: Inexact Update

Algorithm 2 Globally Convergent SSN-H with inexact solve

1: **Input:** $\mathbf{x}^{(0)}$, $0 < \delta < 1$, $0 < \epsilon < 1$, $0 < \beta < 1$, $\widehat{\alpha} \geq 1$, $0 < \theta_1, \theta_2 < 1$
2: - Set the sample size, $|\mathcal{S}|$, with ϵ and δ
3: **for** $k = 0, 1, 2, \cdots$ until termination **do**
4: - Select a sample set, $\mathcal{S}$, of size $|\mathcal{S}|$
5: - Form $H(\mathbf{x}^{(k)})$
6: - Update $\mathbf{x}^{(k+1)}$ with inexact solve
7: **end for**

Gloabl Convergence SSN-H: Inexact Update

- Satisfying requirements (R.3) & (R.4)

Theorem (Global Convergence of Algorithm 2)

Using Algorithm 2 with any $\mathbf{x}^{(k)} \in \mathbb{R}^p$, *with probability* $1-\delta$, *we have*

$$F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^*) \leq (1-\rho)\big(F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*)\big),$$

where

(i) *if* $\theta_1 \leq \sqrt{\frac{(1-\epsilon)}{4\kappa}}$, *then* $\rho = \alpha_k \beta / \kappa$,

(ii) *otherwise* $\rho = 2(1-\theta_2)(1-\theta_1)^2(1-\epsilon)\alpha_k \beta / \kappa^2$.

Moreover, for both cases, the step size is at least $\alpha_k \geq \frac{2(1-\theta_2)(1-\beta)(1-\epsilon)}{\kappa}$.

Outline

- Problem Statement
- Rough Overview of Existing Methods
  - First Order methods
  - Second order methods
- SSN:
  - Globally convergent algorithms
    - Hessian Sub-Sampling
    - Gradient & Hessian Sub-Sampling $\Rightarrow$ SSN-GH
  - Local convergent rates
    - Hessian Sub-Sampling
    - Gradient & Hessian Sub-Sampling
- Examples

Gradient & Hessian Sub-Sampling

$$H(\mathbf{x}) := \frac{1}{|\mathcal{S}_H|} \sum_{j \in \mathcal{S}_H} \nabla^2 f_j(\mathbf{x})$$

$$\mathbf{g}(\mathbf{x}) := \frac{1}{|\mathcal{S}_{\mathbf{g}}|} \sum_{j \in \mathcal{S}_{\mathbf{g}}} \nabla f_j(\mathbf{x})$$

RandNLA

$$\nabla F(\mathbf{x}) = \begin{pmatrix} | & | & & | \\ \nabla f_1(\mathbf{x}) & \nabla f_2(\mathbf{x}) & \cdots & \nabla f_n(\mathbf{x}) \\ | & | & & | \end{pmatrix} \begin{pmatrix} 1/n \\ 1/n \\ \vdots \\ 1/n \end{pmatrix}$$

Gradient Sub-Sampling: RandNLA

Lemma (Uniform Gradient Sub-Sampling)

For a given $\mathbf{x} \in \mathbb{R}^p$*, let*

$$\|\nabla f_i(\mathbf{x})\| \leq G(\mathbf{x}), \quad i = 1, 2, \ldots, n.$$

For any $0 < \epsilon < 1$ *and* $0 < \delta < 1$*, if*

$$|\mathcal{S}| \geq \frac{G(\mathbf{x})^2}{\epsilon^2}\left(1 + \sqrt{8 \ln \frac{1}{\delta}}\right)^2,$$

then

$$\Pr\left(\|\nabla F(\mathbf{x}) - \mathbf{g}(\mathbf{x})\| \leq \epsilon\right) \geq 1 - \delta.$$

- Need to efficiently estimate $G(\mathbf{x})$ at every iteration...see examples in [RM-I,2016]

SSN-GH with Exact Update

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{p}_k,$$

where

$$\mathbf{p}_k = -[H(\mathbf{x}^{(k)})]^{-1}\mathbf{g}(\mathbf{x}^{(k)}), \qquad \text{“exact solve”}$$

$$\begin{aligned} \alpha_k = \arg\max \quad & \alpha \\ \text{s.t.} \quad & \alpha \leq \widehat{\alpha} \\ & F(\mathbf{x}^{(k)} + \alpha \mathbf{p}_k) \leq F(\mathbf{x}^{(k)}) + \alpha\beta \mathbf{p}_k^T \mathbf{g}(\mathbf{x}^{(k)}) \end{aligned}$$

$0 < \beta < 1$, $\widehat{\alpha} \geq 1$

SSN-GH with Exact Update

Algorithm 3 Globally Convergent SSN-GH with exact solve

Input: $\mathbf{x}^{(0)}$, $0 < \delta < 1$, $0 < \epsilon_1 < 1$, $0 < \epsilon_2 < 1$, $0 < \beta < 1$, $\widehat{\alpha} \geq 1$ and $\sigma \geq 0$

- Set the sample size, $|\mathcal{S}_H|$, with ϵ_1 and δ

for $k = 0, 1, 2, \cdots$ until termination **do**

- Select a sample set, $\mathcal{S}_H$, of size $|\mathcal{S}_H|$ and form $H(\mathbf{x}^{(k)})$

- Set the sample size, $|\mathcal{S}_\mathbf{g}|$, with ϵ_2, δ and $\mathbf{x}^{(k)}$

- Select a sample set, $\mathcal{S}_\mathbf{g}$ of size $|\mathcal{S}_\mathbf{g}|$ and form $\mathbf{g}(\mathbf{x}^{(k)})$

if $\|\mathbf{g}(\mathbf{x}^{(k)})\| < \sigma\epsilon_2$ **then**

- STOP

end if

- Update $\mathbf{x}^{(k+1)}$ with exact solve

end for

Global Convergence SSN-GH: Exact Update

Theorem (Global Convergence of Algorithm 3)

Using Algorithm 3 with any $\mathbf{x}^{(k)} \in \mathbb{R}^p$, $\epsilon_1 \leq 1/2$ *and* $\sigma \geq 4\kappa/(1-\beta)$, *we have the following with probability* $(1-\delta)^2$*:*

(i) *if "STOP", then*

$$\|\nabla F(\mathbf{x}^{(k)})\| < (1+\sigma)\,\epsilon_2,$$

(ii) *otherwise, we have*

$$F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^*) \leq (1-\rho)\big(F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*)\big),$$

with $\rho = 8\alpha_k\beta/(9\kappa)$ *with the step size of at least*

$$\alpha_k \geq (1-\beta)(1-\epsilon_1)/\kappa.$$

SSN-GH with Inexact Update

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{p}_k,$$

where

$$\|H(\mathbf{x}^{(k)})\mathbf{p}_k + \mathbf{g}(\mathbf{x}^{(k)})\| \leq \theta_1 \|\mathbf{g}(\mathbf{x}^{(k)})\|$$

$$\mathbf{p}_k^T \mathbf{g}(\mathbf{x}^{(k)}) \leq -(1-\theta_2)\mathbf{p}_k^T H(\mathbf{x}^{(k)})\mathbf{p}_k$$

$$\begin{aligned} \alpha_k = \arg\max \quad & \alpha \\ \text{s.t.} \quad & \alpha \leq \widehat{\alpha} \\ & F(\mathbf{x}^{(k)} + \alpha \mathbf{p}_k) \leq F(\mathbf{x}^{(k)}) + \alpha\beta \mathbf{p}_k^T \mathbf{g}(\mathbf{x}^{(k)}) \end{aligned}$$

$0 < \beta < 1$, $\widehat{\alpha} \geq 1$, $0 < \theta_1, \theta_2 < 1$

SSN-GH with Inexact Update

Algorithm 4 Globally Convergent SSN-GH with exact solve

```
Input: x^(0), 0 < δ < 1, 0 < ε1 < 1, 0 < ε2 < 1, 0 < β < 1, α̂ ≥ 1,
    σ ≥ 0, 0 < θ1, θ2 < 1
- Set the sample size, |S_H|, with ε1 and δ
for k = 0, 1, 2, ··· until termination do
    - Select a sample set, S_H, of size |S_H| and form H(x^(k))
    - Set the sample size, |S_g|, with ε2, δ and x^(k)
    - Select a sample set, S_g of size |S_g| and form g(x^(k))
    if ||g(x^(k))|| < σε2 then
        - STOP
    end if
    - Update x^(k+1) with inexact solve
end for
```

Gloabl Convergence of SSN-GH: Inexact Update

Theorem (Global Convergence of Algorithm 4)

Using Algorithm 4 with $\epsilon_1 \leq 1/2$, *any* $\mathbf{x}^{(k)} \in \mathbb{R}^p$, *and* $\sigma \geq \frac{4\kappa}{(1-\theta_1)(1-\theta_2)(1-\beta)}$, *we have the following with probability* $1-\delta$:

(I) *if "STOP", then* $\|\nabla F(\mathbf{x}^{(k)})\| < (1+\sigma)\,\epsilon_2$,

(II) *otherwise* $F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^*) \leq (1-\rho)\big(F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*)\big)$, *where*

(1) *if* $\theta_1 \leq \sqrt{(1-\epsilon_1)/(4\kappa)}$, *then* $\rho = 4\alpha_k\beta/(9\kappa)$,

(2) *otherwise* $\rho = 8\alpha_k\beta(1-\theta_2)(1-\theta_1)^2(1-\epsilon_1)/(9\kappa^2)$.

Moreover, for both cases, the step size is at least $\alpha_k \geq (1-\theta_2)(1-\beta)(1-\epsilon_1)/\kappa$.

Outline

- Problem Statement
- Rough Overview of Existing Methods
    - First Order methods
    - Second order methods
- SSN:
    - Globally convergent algorithms
        - Hessian Sub-Sampling
        - Gradient & Hessian Sub-Sampling
    - Local convergent rates
        - Hessian Sub-Sampling
        - Gradient & Hessian Sub-Sampling
- Examples

Local Convergence Rates

- Requirements:

(R.1) $|\mathcal{S}|$ must be independent of n, or at least smaller than n

(R.2) $H(\mathbf{x})$ must preserve the spectrum of $\nabla^2 F(\mathbf{x})$ as much as possible

(R.3) $\mathbf{g}(\mathbf{x})$ must be close to $\nabla F(\mathbf{x})$

(R.4) Fast local convergence rate, close to that of Newton!

(R.5) For $p \gg 1$, allow for inexactness $\Rightarrow$ future work

Outline

- Problem Statement
- Rough Overview of Existing Methods
  - First Order methods
  - Second order methods
- SSN:
  - Globally convergent algorithms
    - Hessian Sub-Sampling
    - Gradient & Hessian Sub-Sampling
  - Local convergent rates
    - Hessian Sub-Sampling
    - Gradient & Hessian Sub-Sampling
- Examples

Hessian Sub-Sampling

$$\mathbf{g}(\mathbf{x}) = \nabla F(\mathbf{x})$$

$$H(\mathbf{x}) = \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} \nabla^2 f_j(\mathbf{x})$$

Sub-sampling Hessian

- Satisfying Requirements (R.1) and (R.2)

Lemma (Uniform Hessian Sub-Sampling)

Given any $0 < \epsilon < 1$, $0 < \delta < 1$ *and* $\mathbf{x} \in \mathbb{R}^p$, *if*

$$|\mathcal{S}| \geq \frac{2\kappa^2 \ln(2p/\delta)}{\epsilon^2},$$

then

$$\Pr\left(\left|\lambda_i\left(\nabla^2 F(\mathbf{x})\right) - \lambda_i\left(H(\mathbf{x})\right)\right| \leq \epsilon \lambda_i\left(\nabla^2 F(\mathbf{x})\right);\ i = 1, 2, \cdots, p \right) \geq 1 - \delta.$$

- Difference between (R.2) here and (R.2) before $\Rightarrow$ κ^2 here vs. κ before!

Error Recursion: Hessian Sub-Sampling

Theorem (Error Recursion)

Let $0 < \delta < 1$ *and* $0 < \epsilon < 1$ *be given. Using* $\alpha_k = 1$ *, with probability* $1 - \delta$*, we have*

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\| \leq \rho_0 \|\mathbf{x}^{(k)} - \mathbf{x}^*\| + \xi \|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2,$$

where

$$\rho_0 = \frac{\epsilon}{(1-\epsilon)}, \quad \textit{and} \quad \xi = \frac{L}{2(1-\epsilon)\gamma}.$$

- ρ_0 is problem-independent! $\Rightarrow$ Can be made arbitrarily small!

SSN-H Algorithm

Algorithm 5 Locally Convergent SSN-H with exact solve

Input: $\mathbf{x}^{(0)}$, $0 < \delta < 1$, $0 < \epsilon < 1$
- Set the sample size, $|\mathcal{S}|$, with ϵ and δ
for $k = 0, 1, 2, \cdots$ until termination **do**
    - Select a sample set, $\mathcal{S}$, of size $|\mathcal{S}|$ and $H(\mathbf{x}^{(k)})$
    - Update $\mathbf{x}^{(k+1)}$ with $H(\mathbf{x}^{(k)})$ and $\alpha_k = 1$
end for

SSN-H: Q-Linear Convergence

Theorem (Q-Linear Convergence)

Consider any $0 < \rho_0 < \rho < 1$ *and* $\epsilon \leq \rho_0/(1+\rho_0)$. *If*

$$\|\mathbf{x}^{(0)} - \mathbf{x}^*\| \leq \frac{\rho - \rho_0}{\xi},$$

using Algorithm 5, we get locally Q-linear convergence

$$\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \leq \rho \|\mathbf{x}^{(k-1)} - \mathbf{x}^*\|, \quad k = 1, \ldots, k_0$$

with probability $(1-\delta)^{k_0}$.

SSN-H: Q-Superinear Convergence

Theorem (Q-Superlinear Convergence: Geometric Growth)

Using Algorithm 5, with

$$\epsilon^{(k)} = \rho^k \epsilon, \quad k = 0, 1, \ldots, k_0,$$

if $\mathbf{x}^{(0)}$ is close-enough, we get locally Q-superlinear convergence

$$\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \leq \rho^k \|\mathbf{x}^{(k-1)} - \mathbf{x}^*\|, \quad k = 1, \ldots, k_0$$

with probability $(1-\delta)^{k_0}$.

Hessian Sub-Sampling: Q-Superinear Convergence

Theorem (Q-Superlinear Convergence: Slow Growth)

Using Algorithm 5 with

$$\epsilon^{(k)} = \frac{1}{1 + 2\ln(4+k)}, \quad k = 0, 1, \ldots, k_0,$$

if $\mathbf{x}^{(0)}$ *is close-enough, we get locally Q-superlinear convergence*

$$\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \leq \frac{1}{\ln(3+k)} \|\mathbf{x}^{(k-1)} - \mathbf{x}^*\|, \quad k = 1, \ldots, k_0,$$

with probability $(1-\delta)^{k_0}$.

Local vs. Global

How do we connect the global and local results?

Local + Global

Theorem (Global Conv. of Alg. 1 with Problem-Independent Local Rate)

Consider any $0 < \rho_0 < \rho_1 < 1/\sqrt{\kappa}$*. Using Algorithm 1 with any* $\mathbf{x}^{(0)} \in \mathbb{R}^p$, $\widehat{\alpha} = 1$, $\epsilon \leq \rho_0/((1+\rho_0)\sqrt{2\kappa})$ *and* $\beta \leq (1-\epsilon)(1-\kappa\rho_1^2)/(2\kappa)$*, after*

$$k \geq 2\ln\left(\frac{2(\rho_1-\rho_0)(1-\epsilon)\gamma}{\|\mathbf{x}^{(0)}-\mathbf{x}^*\|\sqrt{\kappa}L}\right) / \ln(1-2\beta/\kappa)$$

iterations, with probability $(1-\delta)^k$ *we get "problem-independent" linear convergence, i.e.,*

$$\|\mathbf{x}^{(k+1)}-\mathbf{x}^*\| \leq \rho_1\|\mathbf{x}^{(k)}-\mathbf{x}^*\|.$$

Moreover, the step size of $\alpha_k = 1$ *passes Armijo rule for all subsequent iterations.*

MODIFYING $H(\mathbf{x})$

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\| \leq \rho_0 \|\mathbf{x}^{(k)} - \mathbf{x}^*\| + \xi \|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2$$

$$\rho_0 = \frac{\epsilon}{(1-\epsilon)} \qquad \xi = \frac{L}{2(1-\epsilon)\gamma}$$

$$\gamma \downarrow \Rightarrow \xi \uparrow$$

Spectral Regularization

$$\hat{H} := \mathcal{P}(\lambda; H) = \sum_{i=1}^{p} \max\{\lambda_i, \lambda\} \mathbf{v}_i \mathbf{v}_i^T$$

- $(\lambda_i, \mathbf{v}_i)$: eigen-pair of $H(\mathbf{x})$

Spectral Regularization

Theorem (Error Recursion)

Let $0 < \delta < 1$ *and* $0 < \epsilon < 1$ *be given. Set* $|\mathcal{S}|$ *and form* $H(\mathbf{x}^{(k)})$*. For some* $\lambda > 0$*, let*

$$\hat{H}(\mathbf{x}^{(k)}) \quad = \quad \mathcal{P}\Big(\lambda; H(\mathbf{x}^{(k)})\Big).$$

Then, with $\alpha_k = 1$*, if* $\lambda \geq (1-\epsilon)\gamma$*, it follows that, with probability* $1-\delta$

$$\rho_0 = \frac{\lambda - (1-\epsilon)\gamma + \gamma\epsilon}{\lambda}, \quad \text{and} \quad \xi = \frac{L}{2\lambda}.$$

Spectral Regularization

Algorithm 6 SSN-H and Spectral Regularization

Input: $\mathbf{x}^{(0)}$, $0 < \delta < 1$, $0 < \epsilon < 1$, and $0 < \epsilon_0 < 1$
- Set the sample size, $|\mathcal{S}_0|$, with ϵ_0 and δ
- Set the sample size, $|\mathcal{S}|$, with ϵ and δ
for $k = 0, 1, 2, \cdots$ until termination **do**
    - Select a sample set, $\mathcal{S}_0$, of size $|\mathcal{S}_0|$ and form $H_0(\mathbf{x}^{(k)})$
    - Compute $\lambda_{\min}\left(H_0(\mathbf{x}^{(k)})\right)$
    - Set the threshold, $\lambda^{(k)}$
    - Select a sample set, $\mathcal{S}$, of size $|\mathcal{S}|$ and form $H(\mathbf{x}^{(k)})$
    - Form $\hat{H}(\mathbf{x}^{(k)})$ with $H(\mathbf{x}^{(k)})$ and $\lambda^{(k)}$
    - Update $\mathbf{x}^{(k+1)}$ with $\hat{H}(\mathbf{x}^{(k)})$ and $\alpha_k = 1$
end for

Spectral Regularization

Theorem (Q-Linear Convergence of Algorithm 6)

Using Algorithm 6, if $\|\mathbf{x}^{(0)} - \mathbf{x}^*\| < \gamma/(3L)$, $\epsilon \leq 1/6$ *and at every iteration the threshold is chosen as*

$$\lambda^{(k)} \geq \left(\frac{1-\epsilon}{1-\epsilon_0}\right) \lambda_{\min}\left(H_0(\mathbf{x}^{(k)})\right),$$

we get locally Q-linear convergence with variable rates

$$\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \leq \rho^{(k-1)} \|\mathbf{x}^{(k-1)} - \mathbf{x}^*\|, \quad k = 1, \ldots, k_0$$

with probability $(1-\delta)^{2k_0}$*, where*

$$\rho^{(k)} = 1 - \frac{\gamma}{2\lambda^{(k)}}.$$

Ridge Regularization

$$\hat{H}(\mathbf{x}) := H(\mathbf{x}) + \lambda \mathbb{I}$$

Ridge Regularization

Theorem (Error Recursion)

Let $0 < \delta < 1$ *and* $0 < \epsilon < 1$ *be given. Set* $|\mathcal{S}|$ *and form* $H(\mathbf{x}^{(k)})$*. For some* $\lambda \geq 0$*, form* $\hat{H}(\mathbf{x}^{(k)})$*. Then, with* $\alpha_k = 1$*, it follows that, with probability* $1 - \delta$*,*

$$\rho_0 = \frac{\lambda + \gamma\epsilon}{(1-\epsilon)\gamma + \lambda}, \quad \xi = \frac{L}{2(1-\epsilon)\gamma + 2\lambda}.$$

Ridge Regularization

Algorithm 7 SSN-H and Ridge Regularization

Input: $\mathbf{x}^{(0)}$, $0 < \delta < 1$, $0 < \epsilon < 1$, $\lambda \geq 0$
- Set the sample size, $|\mathcal{S}|$, with ϵ and δ
for $k = 0, 1, 2, \cdots$ until termination **do**
    - Select a sample set, $\mathcal{S}$, of size $|\mathcal{S}|$ and form $H(\mathbf{x}^{(k)})$
    - Form $\hat{H}(\mathbf{x}^{(k)})$ with $H(\mathbf{x}^{(k)})$ and λ
    - Update $\mathbf{x}^{(k+1)}$ with $\hat{H}(\mathbf{x}^{(k)})$ and $\alpha_k = 1$
end for

Ridge Regularization

Theorem (Q-Linear Convergence of Algorithm 7)

For any $\lambda \geq 0$, consider ρ_0 and ρ such that

$$1 - \frac{\gamma}{\gamma + \lambda} < \rho_0 < \rho < 1.$$

Using Algorithm 7 with

$$\epsilon \leq \frac{\rho_0 \gamma + (\rho_0 - 1)\lambda}{(1 + \rho_0)\gamma},$$

if $\mathbf{x}^{(0)}$ is close enough, then with probability $(1-\delta)^{k_0}$, we get locally Q-linear convergence with the rate ρ.

Outline

- Problem Statement
- Rough Overview of Existing Methods
  - First Order methods
  - Second order methods
- SSN:
  - Globally convergent algorithms
    - Hessian Sub-Sampling
    - Gradient & Hessian Sub-Sampling
  - Local convergent rates
    - Hessian Sub-Sampling
    - Gradient & Hessian Sub-Sampling
- Examples

Gradient & Hessian Sub-Sampling

$$H(\mathbf{x}) := \frac{1}{|\mathcal{S}_H|} \sum_{j \in \mathcal{S}_H} \nabla^2 f_j(\mathbf{x})$$

$$\mathbf{g}(\mathbf{x}) := \frac{1}{|\mathcal{S}_\mathbf{g}|} \sum_{j \in \mathcal{S}_\mathbf{g}} \nabla f_j(\mathbf{x})$$

Theorem (Error Recursion: Independent Sub-Sampling)

Let $0 < \delta < 1$, $0 < \epsilon_1 < 1$, *and* $0 < \epsilon_2 < 1$ *be given. Set* $|\mathcal{S}_H|$ *with* (ϵ_1, δ) *and* $|\mathcal{S}_{\mathbf{g}}|$ *with* (ϵ_2, δ). *Independently, choose* $\mathcal{S}_H$ *and* $\mathcal{S}_{\mathbf{g}}$, *and form* $H(\mathbf{x}^{(k)})$ *and* $\mathbf{g}(\mathbf{x}^{(k)})$. *Then using* $\alpha_k = 1$, *with probability* $(1-\delta)^2$, *we have*

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\| \leq \eta + \rho_0 \|\mathbf{x}^{(k)} - \mathbf{x}^*\| + \xi \|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2,$$

where

$$\eta = \frac{\epsilon_2}{(1-\epsilon_1)\gamma}, \quad \rho_0 = \frac{\epsilon_1}{(1-\epsilon_1)}, \quad \text{and} \quad \xi = \frac{L}{2(1-\epsilon_1)\gamma}.$$

See [RM-II,2016] for simultaneous sampling.

SSN-GH Algorithm

Algorithm 8 Locally Convergent SSN-GH

1: **Input:** $\mathbf{x}^{(0)}$, $0 < \delta < 1$, $0 < \epsilon_1 < 1$, $0 < \epsilon_2 < 1$ and $0 < \rho < 1$
2: - Set the sample size, $|\mathcal{S}_H|$, with ϵ_1 and δ
3: **for** $k = 0, 1, 2, \cdots$ until termination **do**
4: - Select a sample set, $\mathcal{S}_H$, of size $|\mathcal{S}_H|$ and form $H(\mathbf{x}^{(k)})$
5: - Set $\epsilon_2^{(k)} = \rho^k \epsilon_2$
6: - Set the sample size, $|\mathcal{S}_{\mathbf{g}}^{(k)}|$, with $\epsilon_2^{(k)}$, δ and $\mathbf{x}^{(k)}$
7: - Select a sample set, $\mathcal{S}_{\mathbf{g}}^{(k)}$ of size $|\mathcal{S}_{\mathbf{g}}^{(k)}|$ and form $\mathbf{g}(\mathbf{x}^{(k)})$
8: - Update $\mathbf{x}^{(k+1)}$ with $H(\mathbf{x}^{(k)})$, $\mathbf{g}(\mathbf{x}^{(k)})$ and $\alpha_k = 1$
9: **end for**

R-Linear convergence of SSN-GH

Theorem (R-Linear Convergence)

Consider any $0 < \rho < 1$, $0 < \rho_0 < 1$, and $0 < \rho_1 < 1$ such that $\rho_0 + \rho_1 < \rho$. Let $\epsilon_1 \leq \rho_0/(1+\rho_0)$, and define $\sigma := 2(\rho - (\rho_0 + \rho_1))(1-\epsilon_1)\gamma/L$. Using Algorithm 8 with $\epsilon_2 \leq (1-\epsilon_1)\gamma\rho_1\sigma$, if the initial iterate, $\mathbf{x}^{(0)}$, satisfies $\|\mathbf{x}^{(0)} - \mathbf{x}^*\| \leq \sigma$, we get locally R-linear convergence

$$\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \leq \rho^k \sigma,$$

with probability $(1-\delta)^{2k}$.

Local vs. Global

How do we connect the global and local results?

Theorem (Global Conv. of Alg. 3 with Problem-Independent Local Rate)

Consider any $0 < \rho_0, \rho_1, \rho_2 < 1/\sqrt{\kappa}$ *such that* $\rho_0 + \rho_1 < \rho_2$, *set* $\epsilon_1 \leq \rho_0/((1+\rho_0)\sqrt{2\kappa})$. *Using Algorithm 3 with any* $\mathbf{x}^{(0)} \in \mathbb{R}^p$ *and*

$$\widehat{\alpha} = 1, \quad \beta \leq \frac{(1-\epsilon_1)(1-\kappa\rho_2^2)}{8\kappa}$$

$$\epsilon_2^{(0)} \leq (1-\epsilon_1)\gamma\rho_1\sigma, \quad \epsilon_2^{(k)} = \rho_2\epsilon_2^{(k-1)}, \quad k = 1, 2, \cdots,$$

after

$$k \geq 2\ln\left(\frac{\sigma}{\|\mathbf{x}^{(0)} - \mathbf{x}^*\|\sqrt{\kappa}}\right) / \ln\left(1 - 8\beta/(9\kappa)\right)$$

iterations, we have the following with probability $(1-\delta)^{2k}$:

1. *if "STOP", then* $\|\nabla F(\mathbf{x}^{(k)})\| < (1+\sigma)\,\rho_2^k\epsilon_2^{(0)}$
2. *otherwise, we get "problem-independent" linear convergence, i.e.,* $\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\| \leq \rho_2\|\mathbf{x}^{(k)} - \mathbf{x}^*\|$. *Moreover, the step size of* $\alpha_k = 1$ *passes Armiju rule for all subsequent iterations.*

Outline

- Problem Statement
- Rough Overview of Existing Methods
  - First Order methods
  - Second order methods
- SSN:
  - Globally convergent algorithms
    - Hessian Sub-Sampling
    - Gradient & Hessian Sub-Sampling
  - Local convergent rates
    - Hessian Sub-Sampling
    - Gradient & Hessian Sub-Sampling
- Examples

GLM With Gaussian Prior

$$F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\left(\Phi(\mathbf{a}_i^T\mathbf{x}) - b_i\mathbf{a}_i^T\mathbf{x}\right) + \frac{\lambda}{2}\|\mathbf{x}\|^2.$$

- Φ: cumulant generating function
  - $\Phi(t) = 0.5t^2 \Rightarrow$ Ridge Regression (RR)
  - $\Phi(t) = \ln(1 + \exp(t)) \Rightarrow \ell_2$ regularized Logistic Regression (LR)
  - $\Phi(t) = \exp(t) \Rightarrow \ell_2$ regularized Poisson Regression (PR)

ℓ_2 regularized Logistic Regression

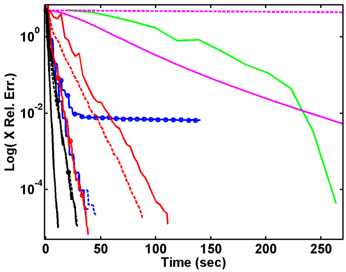
	LR
$\nabla f_i(\mathbf{x})$	$\left(\frac{1}{1+e^{-\mathbf{a}_i^T \mathbf{x}}} - b_i\right) \mathbf{a}_i + \lambda \mathbf{x}$
$\nabla^2 f_i(\mathbf{x})$	$\frac{e^{\mathbf{a}_i^T \mathbf{x}}}{(e^{\mathbf{a}_i^T \mathbf{x}}+1)^2} \mathbf{a}_i \mathbf{a}_i^T + \lambda \mathbb{I}$
K	$0.25 \max_i \\|\mathbf{a}_i\\|^2 + \lambda$
γ	λ
$G(\mathbf{x})$	$\lambda \\|\mathbf{x}\\| + \max_i (1 + \vert b_i \vert) \\|\mathbf{a}_i\\|$

Numerical Examples

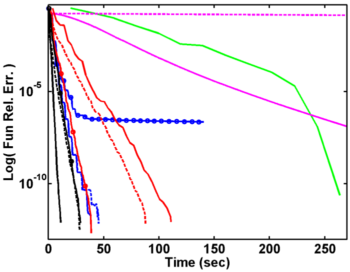
- ℓ_2 regularized Logistic Regression

Data	n	p	NNZ	κ
D_1	10^6	10^4	0.02%	$\approx 10^4$
D_2	5×10^4	5×10^3	Dense	$\approx 10^6$
D_3	10^7	2×10^4	0.006%	$\approx 10^{10}$

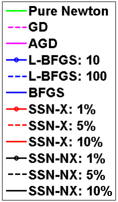
$D_1, n = 10^6, p = 10^4$, SPARSITY : 0.02%, $\kappa \approx 10^4$



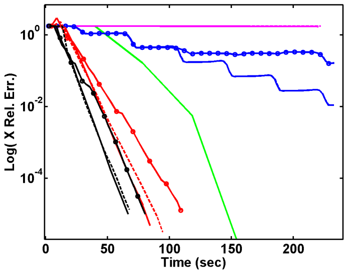
(a) Iterate Relative Error



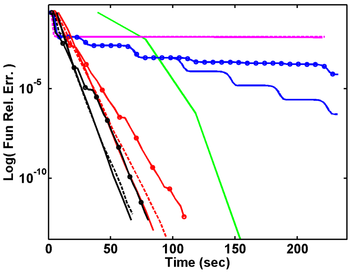
(b) Function Relative Error



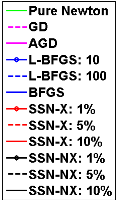
$D_2, n = 5 \times 10^4, p = 5 \times 10^3$, Sparsity : Dense, $\kappa \approx 10^6$



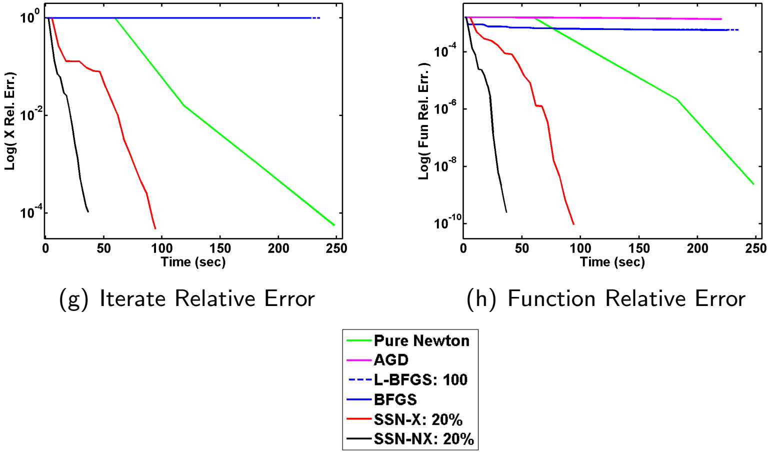
(d) Iterate Relative Error



(e) Function Relative Error



D_3, $n = 10^7$, $p = 2 \times 10^4$, SPARSITY : 0.006%, $\kappa \approx 10^{10}$



(g) Iterate Relative Error

(h) Function Relative Error

THANK YOU!