Stochastic Variance Reduced Optimization for Nonconvex Sparse Learning

Xingguo Li[*]

Joint work with Tuo Zhao[†], Raman Arora[†], Han Liu[‡], and Jarvis Haupt[*]

[*]University of Minnesota [†]Johns Hopkins University [†]Princeton University

JHU, July 20, 2016

UNIVERSITY OF MINNESOTA JOHNS HOPKINS UNIVERSITY PRINCETON UNIVERSITY

Background

Consider a Sparse Linear Model:

$$\boldsymbol{y} = \mathbf{A}\boldsymbol{\theta}^* + \boldsymbol{z},\ \mathbf{y} \in \mathbb{R}^n,\ \mathbf{A} \in \mathbb{R}^{n\times d}$$

$\|\boldsymbol{\theta}^*\|_0 \leq k^* < n \ll d$, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$

Background

Consider a Sparse Linear Model:

$$\boldsymbol{y} = \mathbf{A}\theta^* + \boldsymbol{z},\ \ \mathbf{y} \in \mathbb{R}^n,\ \ \mathbf{A} \in \mathbb{R}^{n\times d}$$

$\|\boldsymbol{\theta}^*\|_0 \leq k^* < n \ll d$, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$

Consider the nonconvex sparse learning problem:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{F}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{\theta}) \quad \text{subject to } ||\boldsymbol{\theta}||_0 \leq k,$$

$\mathcal{F}(\boldsymbol{\theta})$: Empirical risk – smooth and nonstrongly convex.

Example: $f_i(\boldsymbol{\theta}) = \frac{1}{b}\|\mathbf{y}_i - \mathbf{A}_i\theta\|_2^2$ for Sparse Linear Model

Background

Consider a Sparse Linear Model:

$$\boldsymbol{y} = \mathbf{A}\theta^* + \boldsymbol{z},\ \mathbf{y} \in \mathbb{R}^n,\ \mathbf{A} \in \mathbb{R}^{n\times d}$$

$\|\boldsymbol{\theta}^*\|_0 \leq k^* < n \ll d$, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$

Consider the nonconvex sparse learning problem:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{F}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{\theta}) \quad \text{subject to } ||\boldsymbol{\theta}||_0 \leq k,$$

$\mathcal{F}(\boldsymbol{\theta})$: Empirical risk – smooth and nonstrongly convex.

Example: $f_i(\boldsymbol{\theta}) = \frac{1}{b}\|\mathbf{y}_i - \mathbf{A}_i\theta\|_2^2$ for Sparse Linear Model

- Non-convex; NP-hard (in the worst case)
- Good empirical performance

Background

- **Restricted Strong Convexity**: for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_0 \leq s$,

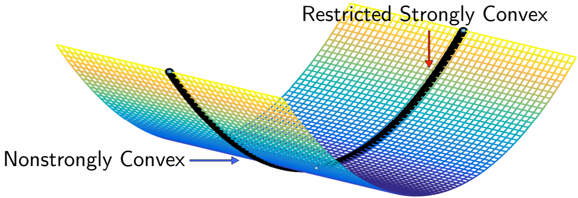
$$\mathcal{F}(\boldsymbol{\theta}) - \mathcal{F}(\boldsymbol{\theta}') - \langle \nabla \mathcal{F}(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle \geq \frac{\rho_s^-}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2.$$

- **Restricted Strong Smoothness**: For any $i \in [n]$, and any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_0 \leq s$,

$$f_i(\boldsymbol{\theta}) - f_i(\boldsymbol{\theta}') - \langle \nabla f_i(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle \leq \frac{\rho_s^+}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2.$$

Much weaker than Restricted Isometry Property (RIP): $\rho_s^+ < 2$

Hidden structure:



Motivation

Gradient Hard Thresholding (GHT, Jain et al., (2014)):

$$\boldsymbol{\theta}^{(t+1)} = \mathcal{H}_k\left(\boldsymbol{\theta}^{(t)} - \eta \nabla \mathcal{F}(\boldsymbol{\theta}^{(t)})\right),$$

- Computationally expensive – $\mathcal{O}(nd)$ each iteration

Motivation

Gradient Hard Thresholding (GHT, Jain et al., (2014)):

$$\boldsymbol{\theta}^{(t+1)} = \mathcal{H}_k\left(\boldsymbol{\theta}^{(t)} - \eta \nabla \mathcal{F}(\boldsymbol{\theta}^{(t)})\right),$$

- Computationally expensive – $\mathcal{O}(nd)$ each iteration

Stochastic Gradient Hard Thresholding (SGHT, Nguyen et al., (2014)):

$$\boldsymbol{\theta}^{(t+1)} = \mathcal{H}_k\left(\boldsymbol{\theta}^{(t)} - \eta \nabla f_i(\boldsymbol{\theta}^{(t)})\right),$$

- Large variance $\Rightarrow$ Large statistical error

Motivation

Gradient Hard Thresholding (GHT, Jain et al., (2014)):

$$\boldsymbol{\theta}^{(t+1)} = \mathcal{H}_k \left(\boldsymbol{\theta}^{(t)} - \eta \nabla \mathcal{F}(\boldsymbol{\theta}^{(t)}) \right),$$

- Computationally expensive – $\mathcal{O}(nd)$ each iteration

Stochastic Gradient Hard Thresholding (SGHT, Nguyen et al., (2014)):

$$\boldsymbol{\theta}^{(t+1)} = \mathcal{H}_k \left(\boldsymbol{\theta}^{(t)} - \eta \nabla f_i(\boldsymbol{\theta}^{(t)}) \right),$$

- Large variance ⇒ Large statistical error

We propose **Stochastic Variance Reduced Gradient Hard Thresholding** (SVR-GHT):

- Computationally efficient – $\mathcal{O}(d)$ each iteration
- Reduced variance ⇒ Small statistical error

Intuition

Main Idea of *SVR-GHT*: $\boldsymbol{\theta}^*$ – true sparse model parameter

- "Ideal gradient": $\nabla f_i(\boldsymbol{\theta}^{(t)}) - \nabla f_i(\boldsymbol{\theta}^*)$

Intuition

Main Idea of *SVR-GHT*: $\boldsymbol{\theta}^*$ – true sparse model parameter

- "Ideal gradient": $\nabla f_i(\boldsymbol{\theta}^{(t)}) - \nabla f_i(\boldsymbol{\theta}^*)$
- Reduce variance: $\mathbb{E}\|\nabla f_i(\boldsymbol{\theta}^{(t)}) - \nabla f_i(\boldsymbol{\theta}^*)\|_2 \leq \rho_s^+ \|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\|_2 \downarrow$

Intuition

Main Idea of *SVR-GHT*: $\boldsymbol{\theta}^*$ – true sparse model parameter

- "Ideal gradient": $\nabla f_i(\boldsymbol{\theta}^{(t)}) - \nabla f_i(\boldsymbol{\theta}^*)$
- Reduce variance: $\mathbb{E}\|\nabla f_i(\boldsymbol{\theta}^{(t)}) - \nabla f_i(\boldsymbol{\theta}^*)\|_2 \leq \rho_s^+ \|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\|_2 \downarrow$
- Unbiased estimator: $\mathbb{E}\nabla f_i(\boldsymbol{\theta}^{(t)}) - \nabla f_i(\boldsymbol{\theta}^*) + \nabla \mathcal{F}(\boldsymbol{\theta}^*) = \nabla \mathcal{F}(\boldsymbol{\theta}^{(t)})$

Intuition

Main Idea of *SVR-GHT*: $\boldsymbol{\theta}^*$ – true sparse model parameter

- "Ideal gradient": $\nabla f_i(\boldsymbol{\theta}^{(t)}) - \nabla f_i(\boldsymbol{\theta}^*)$
- Reduce variance: $\mathbb{E}\|\nabla f_i(\boldsymbol{\theta}^{(t)}) - \nabla f_i(\boldsymbol{\theta}^*)\|_2 \leq \rho_s^+ \|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\|_2 \downarrow$
- Unbiased estimator: $\mathbb{E}\nabla f_i(\boldsymbol{\theta}^{(t)}) - \nabla f_i(\boldsymbol{\theta}^*) + \nabla \mathcal{F}(\boldsymbol{\theta}^*) = \nabla \mathcal{F}(\boldsymbol{\theta}^{(t)})$
- $\boldsymbol{\theta}^*$ is unknown $\Longrightarrow \widetilde{\boldsymbol{\theta}}$: **Variance Reduced Stochastic Gradient**

$$\nabla f_{i_t}(\boldsymbol{\theta}^{(t)}) - \nabla f_{i_t}(\widetilde{\boldsymbol{\theta}}) + \nabla \mathcal{F}(\widetilde{\boldsymbol{\theta}})$$

Intuition

Main Idea of *SVR-GHT*: $\boldsymbol{\theta}^*$ – true sparse model parameter

- "Ideal gradient": $\nabla f_i(\boldsymbol{\theta}^{(t)}) - \nabla f_i(\boldsymbol{\theta}^*)$
- Reduce variance: $\mathbb{E}\|\nabla f_i(\boldsymbol{\theta}^{(t)}) - \nabla f_i(\boldsymbol{\theta}^*)\|_2 \le \rho_s^+ \|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\|_2 \downarrow$
- Unbiased estimator: $\mathbb{E}\nabla f_i(\boldsymbol{\theta}^{(t)}) - \nabla f_i(\boldsymbol{\theta}^*) + \nabla \mathcal{F}(\boldsymbol{\theta}^*) = \nabla \mathcal{F}(\boldsymbol{\theta}^{(t)})$
- $\boldsymbol{\theta}^*$ is unknown $\Longrightarrow \widetilde{\boldsymbol{\theta}}$: **Variance Reduced Stochastic Gradient**

$$\nabla f_{i_t}(\boldsymbol{\theta}^{(t)}) - \nabla f_{i_t}(\widetilde{\boldsymbol{\theta}}) + \nabla \mathcal{F}(\widetilde{\boldsymbol{\theta}})$$

For $r = 1, 2, \ldots$ (outer loop)

  $\boldsymbol{\theta}^{(0)} = \widetilde{\boldsymbol{\theta}}^{(r-1)}$; $\tilde{\mathbf{u}} = \nabla \mathcal{F}(\widetilde{\boldsymbol{\theta}}^{(r-1)})$

  For $t = 0, 1, \ldots, m-1$ (inner loop)

    Randomly sample i_t from $\{1, \ldots, n\}$

    $\bar{\boldsymbol{\theta}}^{(t)} = \boldsymbol{\theta}^{(t)} - \eta \cdot \left(\nabla f_{i_t}(\boldsymbol{\theta}^{(t)}) - \nabla f_{i_t}(\widetilde{\boldsymbol{\theta}}^{(r-1)}) + \tilde{\mathbf{u}}\right)$

    $\boldsymbol{\theta}^{(t+1)} = \mathcal{H}_k(\bar{\boldsymbol{\theta}}^{(t)})$

  $\widetilde{\boldsymbol{\theta}}^{(r)} = \boldsymbol{\theta}^{(m)}$

Why Variance Reduction

SVR-GHT:

- SVRG: Sufficient contraction, e.g.,
$$\|\overline{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\|_2 \cong 0.8\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\|_2$$
- Hard Thresholding: Slight expansion, e.g.,
$$\|\mathcal{H}_k(\overline{\boldsymbol{\theta}}^{(t)}) - \boldsymbol{\theta}^*\| \cong 1.1\|\overline{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\|_2$$
- Overall linear rate, e.g., $\|\mathcal{H}_k(\overline{\boldsymbol{\theta}}^{(t)}) - \boldsymbol{\theta}^*\| \cong 0.88\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\|_2$

Why Variance Reduction

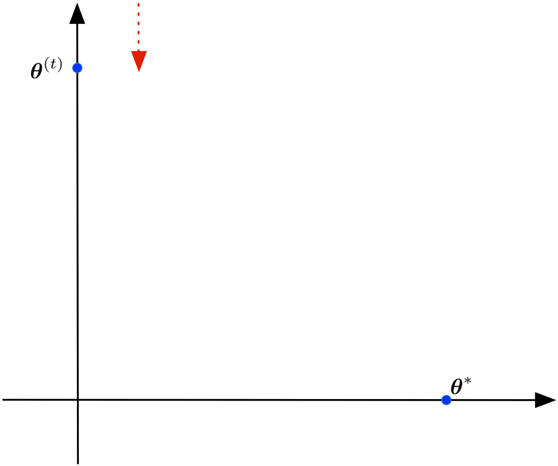
SVR-GHT:

- SVRG: Sufficient contraction, e.g.,
$$\|\overline{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\|_2 \cong 0.8\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\|_2$$
- Hard Thresholding: Slight expansion, e.g.,
$$\|\mathcal{H}_k(\overline{\boldsymbol{\theta}}^{(t)}) - \boldsymbol{\theta}^*\| \cong 1.1\|\overline{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\|_2$$
- Overall linear rate, e.g., $\|\mathcal{H}_k(\overline{\boldsymbol{\theta}}^{(t)}) - \boldsymbol{\theta}^*\| \cong 0.88\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\|_2$

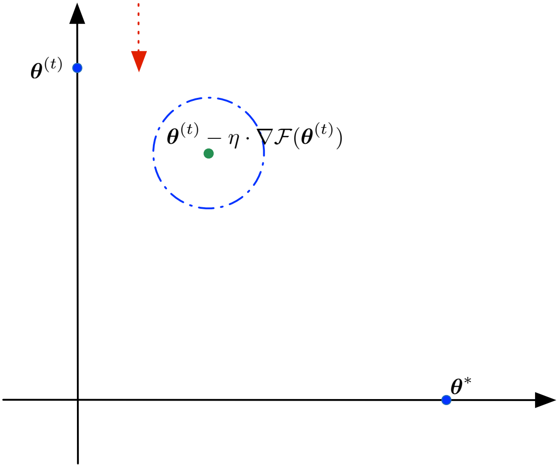
SGHT:

- Overall linear rate only for well condition problem
- No sufficient contraction otherwise

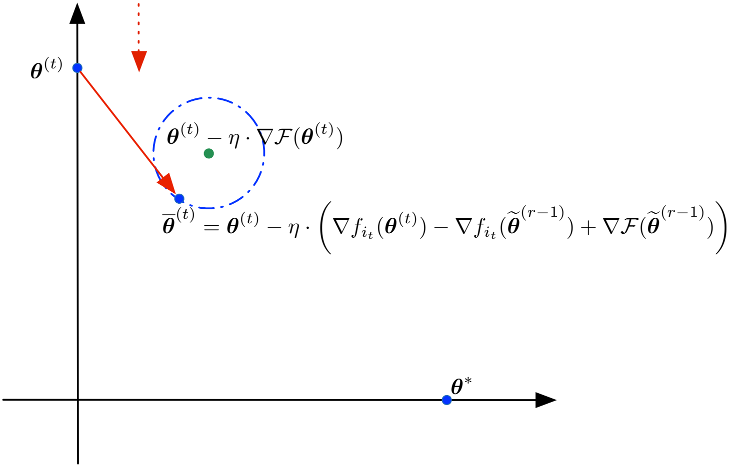
Geometric Intuition



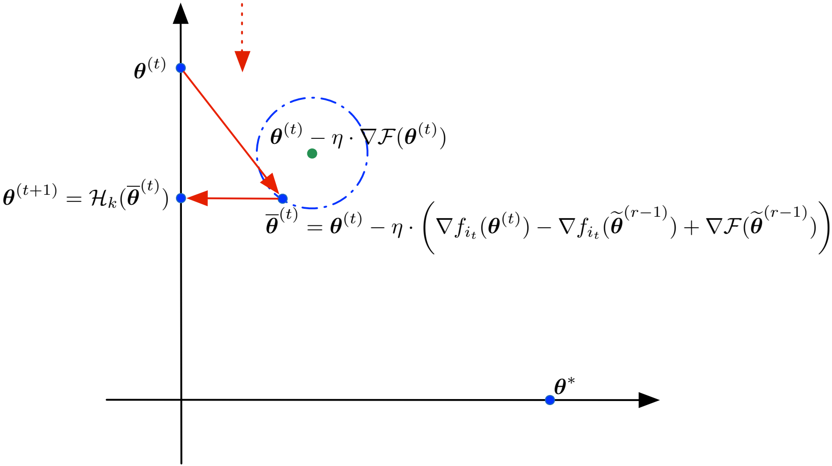
Geometric Intuition



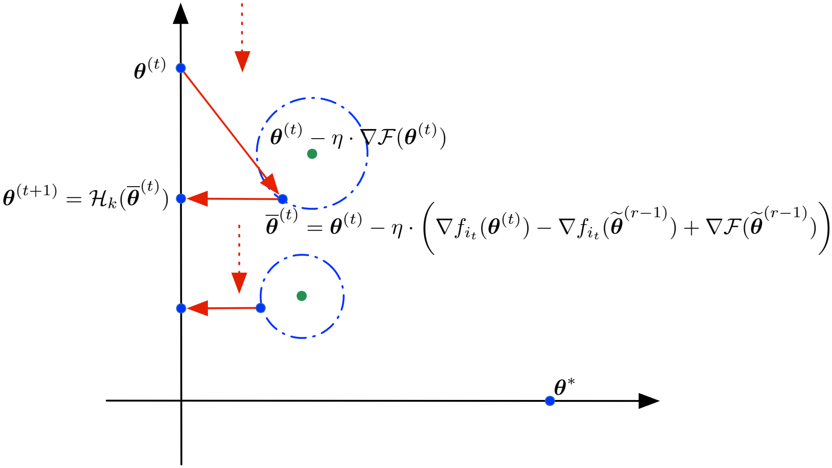
Geometric Intuition



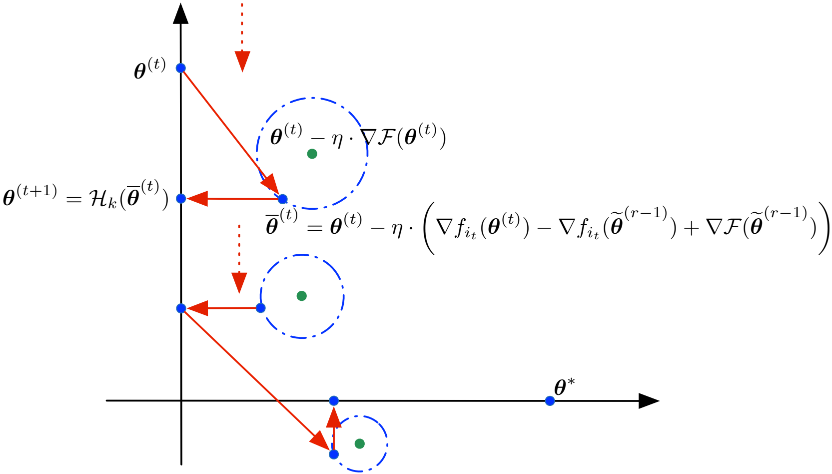
Geometric Intuition



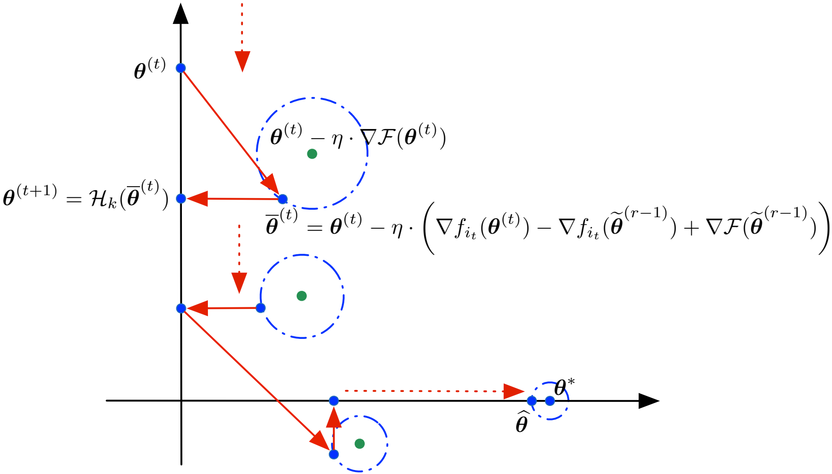
Geometric Intuition



Geometric Intuition



Geometric Intuition



Computational Theory

$\boldsymbol{\theta}^*$: True sparse model parameter. Suppose

- $\mathcal{F}(\boldsymbol{\theta})$ satisfies RSC and $\{f_i(\boldsymbol{\theta})\}_{i=1}^n$ satisfy RSS with $s = 2k + k^*$
- $\|\boldsymbol{\theta}^*\|_0 \le k^*$, $k \gtrsim \kappa_s^2 k^*$, $\eta\rho_s^+ \backsimeq 1$, $m \gtrsim \kappa_s$ and $r \lesssim \log\left(\frac{\mathcal{F}(\tilde{\boldsymbol{\theta}}^{(0)}) - \mathcal{F}(\boldsymbol{\theta}^*)}{\varepsilon\delta}\right)$

Then with probability at least $1 - \delta$,

$$\|\widetilde{\boldsymbol{\theta}}^{(r)} - \boldsymbol{\theta}^*\|_2 \le \sqrt{\frac{2\varepsilon}{\rho_s^-}} + g_2(\boldsymbol{\theta}^*),$$

$g_2(\boldsymbol{\theta}^*) = \mathcal{O}\left(\sqrt{s}\|\nabla\mathcal{F}(\boldsymbol{\theta}^*)\|_\infty\right)$: the statistical error.

Computational Theory

$\boldsymbol{\theta}^*$: True sparse model parameter. Suppose

- $\mathcal{F}(\boldsymbol{\theta})$ satisfies RSC and $\{f_i(\boldsymbol{\theta})\}_{i=1}^n$ satisfy RSS with $s = 2k + k^*$
- $\|\boldsymbol{\theta}^*\|_0 \leq k^*$, $k \gtrsim \kappa_s^2 k^*$, $\eta\rho_s^+ \backsimeq 1$, $m \gtrsim \kappa_s$ and $r \lesssim \log\left(\frac{\mathcal{F}(\tilde{\boldsymbol{\theta}}^{(0)}) - \mathcal{F}(\boldsymbol{\theta}^*)}{\varepsilon\delta}\right)$

Then with probability at least $1 - \delta$,

$$\|\widetilde{\boldsymbol{\theta}}^{(r)} - \boldsymbol{\theta}^*\|_2 \leq \sqrt{\frac{2\varepsilon}{\rho_s^-}} + g_2(\boldsymbol{\theta}^*),$$

$g_2(\boldsymbol{\theta}^*) = \mathcal{O}\left(\sqrt{s}\|\nabla\mathcal{F}(\boldsymbol{\theta}^*)\|_\infty\right)$: the statistical error.

Linear Convergence to $\boldsymbol{\theta}^*$ within optimal statistical error.

Computational Theory

$\boldsymbol{\theta}^*$: True sparse model parameter. Suppose

- $\mathcal{F}(\boldsymbol{\theta})$ satisfies RSC and $\{f_i(\boldsymbol{\theta})\}_{i=1}^n$ satisfy RSS with $s = 2k + k^*$
- $\|\boldsymbol{\theta}^*\|_0 \leq k^*$, $k \gtrsim \kappa_s^2 k^*$, $\eta\rho_s^+ \backsimeq 1$, $m \gtrsim \kappa_s$ and $r \lesssim \log\left(\frac{\mathcal{F}(\tilde{\boldsymbol{\theta}}^{(0)}) - \mathcal{F}(\boldsymbol{\theta}^*)}{\varepsilon\delta}\right)$

Then with probability at least $1 - \delta$,

$$\|\widetilde{\boldsymbol{\theta}}^{(r)} - \boldsymbol{\theta}^*\|_2 \leq \sqrt{\frac{2\varepsilon}{\rho_s^-}} + g_2(\boldsymbol{\theta}^*),$$

$g_2(\boldsymbol{\theta}^*) = \mathcal{O}\left(\sqrt{s}\|\nabla\mathcal{F}(\boldsymbol{\theta}^*)\|_\infty\right)$: the statistical error.

Linear Convergence to $\boldsymbol{\theta}^*$ within optimal statistical error.

Extensions: SAGA-GHT, Asynchronous SVR-GHT

Statistical Rate of Convergence

(I) Sparse Linear Regression:

- $\boldsymbol{y} = \mathbf{A}\boldsymbol{\theta}^* + \boldsymbol{z}$, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, $\|\boldsymbol{\theta}^*\|_0 \leq k^*$

$$\|\tilde{\boldsymbol{\theta}}^{(r)} - \boldsymbol{\theta}^*\|_2 = \mathcal{O}_p\left(\sigma\sqrt{\frac{k^* \log d}{nb}}\right)$$

Statistical Rate of Convergence

(I) Sparse Linear Regression:

- $\boldsymbol{y} = \mathbf{A}\boldsymbol{\theta}^* + \boldsymbol{z}$, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, $\|\boldsymbol{\theta}^*\|_0 \leq k^*$

$$\|\tilde{\boldsymbol{\theta}}^{(r)} - \boldsymbol{\theta}^*\|_2 = \mathcal{O}_p\left(\sigma\sqrt{\frac{k^* \log d}{nb}}\right)$$

(II) Generalized Linear Models:

- $\mathbb{P}\left(y_i | \mathbf{A}_{i*}, \boldsymbol{\theta}^*, \sigma\right) \propto \exp\left\{y_i \mathbf{A}_{i*}\boldsymbol{\theta}^* - h(\mathbf{A}_{i*}\boldsymbol{\theta}^*)\right\}$, $\|\boldsymbol{\theta}^*\|_0 \leq k^*$

$$\|\tilde{\boldsymbol{\theta}}^{(r)} - \boldsymbol{\theta}^*\|_2 = \mathcal{O}_p\left(\sqrt{\frac{k^* \log d}{nb}}\right)$$

Statistical Rate of Convergence

(I) Sparse Linear Regression:

- $\boldsymbol{y} = \mathbf{A}\boldsymbol{\theta}^* + \boldsymbol{z}$, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, $\|\boldsymbol{\theta}^*\|_0 \leq k^*$

$$\|\tilde{\boldsymbol{\theta}}^{(r)} - \boldsymbol{\theta}^*\|_2 = \mathcal{O}_p\left(\sigma\sqrt{\frac{k^* \log d}{nb}}\right)$$

(II) Generalized Linear Models:

- $\mathbb{P}\left(y_i | \mathbf{A}_{i*}, \boldsymbol{\theta}^*, \sigma\right) \propto \exp\left\{y_i \mathbf{A}_{i*}\boldsymbol{\theta}^* - h(\mathbf{A}_{i*}\boldsymbol{\theta}^*)\right\}$, $\|\boldsymbol{\theta}^*\|_0 \leq k^*$

$$\|\tilde{\boldsymbol{\theta}}^{(r)} - \boldsymbol{\theta}^*\|_2 = \mathcal{O}_p\left(\sqrt{\frac{k^* \log d}{nb}}\right)$$

(III) Low-rank Matrix Regression;

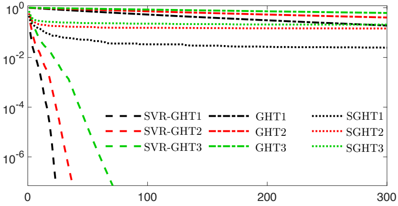
- $\mathbf{y} = \mathcal{A}(\boldsymbol{\Theta}^*) + \mathbf{z}$, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, $\boldsymbol{\Theta}^* \in \mathbb{R}^{d \times p}$, $\mathrm{rank}(\boldsymbol{\Theta}^*) \leq k^*$

$$\|\tilde{\boldsymbol{\Theta}}^{(r)} - \boldsymbol{\Theta}^*\|_{\mathrm{F}} = \mathcal{O}_p\left(\sigma\sqrt{\frac{k^*(d+p)}{nb}}\right)$$

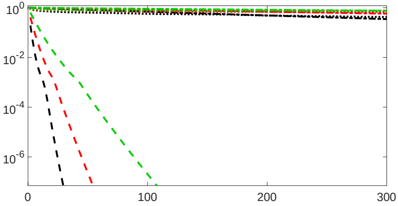
Synthetic Data

Sparse linear model:

- Methods: GHT, SGHT, SVR-GHT
- Settings: $k^* = 3$, $k = 500$, $nb = 10000$, $d = 25000$
- Horizontal-axis: # passes of data
- Vertical-axis: $\frac{\mathcal{F}(\widetilde{\boldsymbol{\theta}}^{(r)})}{\mathcal{F}(\mathbf{0})}$



Low Correlation

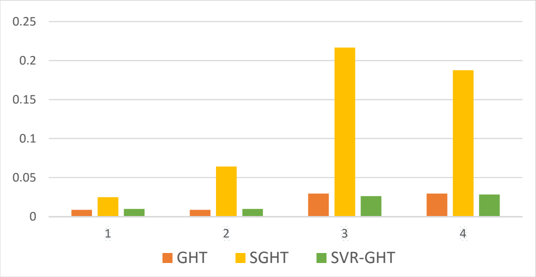


High Correlation

Synthetic Data

Sparse linear model:

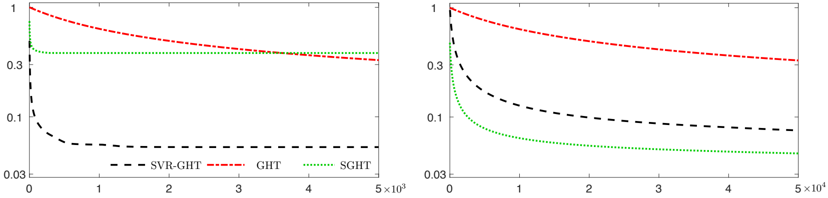
- Methods: GHT, SGHT, SVR-GHT
- Settings: $k^* = 3$, $k = 500$, $nb = 10000$, $d = 25000$
- Low/High correlation; Small/Large mini-batch size b
- Criterion: $\|\widetilde{\boldsymbol{\theta}}^{(10^6)} - \boldsymbol{\theta}^*\|_2 / \|\boldsymbol{\theta}^*\|_2$



Real Data

Logistic regression for binary classification:

- RCV1 dataset: d =29992, nb = 5000 for training / 4625 for testing
- Small/Large mini-batch size b
- Horizontal-axis: # passes of data
- Vertical-axis: Training error



Small b Large b

Real Data

Logistic regression for binary classification:

- RCV1 dataset: d =29992, nb = 5000 for training / 4625 for testing
- Small/Large mini-batch size b
- Criterion: Test classification error

Summary

- Linear Convergence to $\boldsymbol{\theta}^*$ within statistical error
- Optimal statistical rate of convergence

Table 1. Comparison of GHT, SGHT and SVR-GHT.

Method	GHT	SGHT	SVR-GHT
Assumption on κ_s	κ_s bounded	$\boldsymbol{\kappa_s \leq \frac{4}{3}}$	κ_s **bounded**
Comput. Complx.	$\mathcal{O}\left(n\kappa_s \cdot \log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left([n+\kappa_s] \cdot \log\left(\frac{1}{\epsilon}\right)\right)$
Statistical Err.	$\mathcal{O}\left(\sigma\sqrt{\frac{k^* \log d}{nb}}\right)$	$\mathcal{O}\left(\sigma\sqrt{\frac{k^* \log d}{b}}\right)$	$\mathcal{O}\left(\sigma\sqrt{\frac{k^* \log d}{nb}}\right)$

$\kappa_s = \frac{\rho_s^+}{\rho_s^-}$; b: mini-batch size; ϵ: optimization accuracy; $\|\boldsymbol{\theta}^*\|_0 = k^*$

Summary

- Linear Convergence to $\boldsymbol{\theta}^*$ within statistical error
- Optimal statistical rate of convergence

Table 1. Comparison of GHT, SGHT and SVR-GHT.

Method	GHT	SGHT	SVR-GHT
Assumption on κ_s	κ_s bounded	$\kappa_s \leq \frac{4}{3}$	κ_s bounded
Comput. Complx.	$\mathcal{O}\left(\boldsymbol{n\kappa_s} \cdot \log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left([\boldsymbol{n} + \boldsymbol{\kappa_s}] \cdot \log\left(\frac{1}{\epsilon}\right)\right)$
Statistical Err.	$\mathcal{O}\left(\sigma\sqrt{\frac{k^* \log d}{nb}}\right)$	$\mathcal{O}\left(\sigma\sqrt{\frac{k^* \log d}{b}}\right)$	$\mathcal{O}\left(\sigma\sqrt{\frac{k^* \log d}{nb}}\right)$

$\kappa_s = \frac{\rho_s^+}{\rho_s^-}$; b: mini-batch size; ϵ: optimization accuracy; $\|\boldsymbol{\theta}^*\|_0 = k^*$

Summary

- Linear Convergence to $\boldsymbol{\theta}^*$ within statistical error
- Optimal statistical rate of convergence

Table 1. Comparison of GHT, SGHT and SVR-GHT.

Method	GHT	SGHT	SVR-GHT
Assumption on κ_s	κ_s bounded	$\kappa_s \leq \frac{4}{3}$	κ_s bounded
Comput. Complx.	$\mathcal{O}\left(n\kappa_s \cdot \log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left([n+\kappa_s] \cdot \log\left(\frac{1}{\epsilon}\right)\right)$
Statistical Err.	$\mathcal{O}\left(\sigma\sqrt{\frac{k^* \log d}{nb}}\right)$	$\mathcal{O}\left(\sigma\sqrt{\frac{k^* \log d}{\boldsymbol{b}}}\right)$	$\mathcal{O}\left(\sigma\sqrt{\frac{k^* \log d}{\boldsymbol{nb}}}\right)$

$\kappa_s = \frac{\rho_s^+}{\rho_s^-}$; b: mini-batch size; ϵ: optimization accuracy; $\|\boldsymbol{\theta}^*\|_0 = k^*$

Discussion

Existing nonconvex optimization method (Loh & Wainwright, 2013):

$$\min_{\boldsymbol{\theta}} \; \mathcal{F}(\boldsymbol{\theta}) + \mathcal{P}_{\lambda,\gamma}(\boldsymbol{\theta}) \quad \text{subject to } ||\boldsymbol{\theta}||_1 \leq R,$$

- $\mathcal{P}_{\lambda,\gamma}(\boldsymbol{\theta})$: a nonconvex regularization function, such as MCP, SCAD
- More tuning efforts: $\lambda, \gamma, R \Longleftrightarrow$ SVR-GHT: k

Thank you !