Sketching for Convex and Nonconvex Regularized Least Squares with Sharp Guarantees

Yingzhen Yang[1], Ping Li[2]
[1] Arizona State Uniersity
[2] Cognitive Computing Lab, Baidu Research

Introduction

- Randomized algorithms for efficient optimization are a critical area of research in machine learning and optimization, with wide-ranging applications in numerical linear algebra, data analysis, and scientific computing.
- These methods have been successfully applied to large-scale problems such as least squares regression, robust regression, low-rank approximation, singular value decomposition, and matrix factorization.

Introduction

- In this paper, we study efficient sketching algorithm for the optimization problem of regularized least squares with convex or nonconvex regularization, which is presented as follows:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} f(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + h_\lambda(\boldsymbol{\beta}). \tag{1}$$

  - Here $\mathbf{X}\in\mathbb{R}^{n\times d}$ is the data matrix or design matrix for regression problems.
  - $h_\lambda\colon \mathbb{R}^d \to \mathbb{R}$ is a regularizer function and λ is a positive regularization weight, and we emphasize that h_λ can be either convex or nonconvex (SCAD or MCP).
  - We focus on sparsity-inducing regularizers, and the solution to the sketched problem (1) can provably approximate the true parameter vector for sparse signal estimation to be introduced later.

Introduction

- Optimization for (1) is time consuming when n is large. To this end, we propose Sketching for Regularized Optimization (SRO) in this paper as an efficient randomized algorithm for problem (1).

- With $\tilde{n} < n$ where $\tilde{n}$ is the target row number of a sketch of the data matrix $\mathbf{X}$ which is also termed the sketch size, SRO first generates a sketched version of $\mathbf{X}$ by $\tilde{\mathbf{X}} = \mathbf{P}\mathbf{X}$, then solves the following sketched problem,

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \tilde{f}(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^\top \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \boldsymbol{\beta} - \langle \mathbf{y}, \mathbf{X}\boldsymbol{\beta} \rangle + h_\lambda(\boldsymbol{\beta}). \tag{2}$$

- One hopes that the optimization result of the sketched problem (2), denoted by $\tilde{\boldsymbol{\beta}}^*$, is a good approximation to that of the original problem (1), denoted by $\boldsymbol{\beta}^*$. The optimization and theoretical computer science literature prefers relative-error approximation to the solution of the original problem in the following form:
  - $\left\| \tilde{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^* \right\|_{\mathbf{X}} \le \rho \|\boldsymbol{\beta}^*\|_{\mathbf{X}}$, where $\|\mathbf{u}\|_{\mathbf{X}} := \|\mathbf{X}\mathbf{u}\|_2$ and $0 < \rho < 1$ is a positive constant.

Our Contributions

- We study the sparse signal estimation problem by sketching. For sparse convex or nonconvex learning problems where h_λ is convex or nonconvex, we obtain the minimax optimal rate of the order $\mathcal{O}\left(\sqrt{\bar{s}\log d/n}\right)$ for the parameter estimation error in sparse signal estimation by solving the sketched problem (2) where $\bar{s}$ is the support size of the unknown sparse parameter vector to be estimated. To the best of our knowledge, our analysis provides the first unified theoretical result for the minimax optimal error rate for sparse signal estimation using sketching based optimization method.

- In order to obtain such minimax optimal rate for sparse signal estimation by solving the sketched sparse convex learning problems, we propose an iterative sketching algorithm termed Iterative SRO, which provably reduces the approximation error $\left\|\tilde{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*\right\|_{\mathbf{X}}$ geometrically in an iterative manner.

Differences between Iterative SRO and Iterative Hessian Sketch (IHS)

- There are two key differences between Iterative SRO and Iterative Hessian Sketch (IHS) [1].
  - First, using the subspace embedding as the projection matrix $\mathbf{P}$, Iterative SRO does not need to sample $\mathbf{P}$ and compute the sketched matrix $\tilde{\mathbf{X}} = \mathbf{P}\mathbf{X}$ at each iteration, in contrast with IHS where a separate $\mathbf{P}$ is sampled and $\tilde{\mathbf{X}}$ is computed at each iteration. This advantage saves considerable computation and storage for large-scale problems.
  - Second, while IHS is restricted to constrained least-square problems with convex constraints, SRO and Iterative SRO are capable of handling all convex regularization and certain nonconvex regularization in a unified framework. For example, we show that Generalized Lasso can be efficiently and effectively solved by Iterative SRO in our experiments.

The Basics of SRO

- In order to improve the efficiency of optimization for (1), we propose Regularized Optimization by Sketching (SRO) in this section. The key idea is to sketch matrix $\mathbf{X}$ in the quadratic term of (1) by random projection. It consisits of two steps:
  - Step 1. Project the matrix $\mathbf{X}$ onto a lower dimensional space by a linear transformation $\mathbf{P} \in \mathbb{R}^{\tilde{n} \times n}$ with $\tilde{n} < n$, i.e. $\tilde{\mathbf{X}} = \mathbf{P}\mathbf{X}$. $\tilde{n}$ is named the sketch size.
  - Step 2. Solve the sketched problem (2).

- The linear transformation $\mathbf{P}$ is required to be a subspace embedding [2] defined in Definition 3.1.

Definition 3.1

Suppose $\mathcal{P}$ is a distribution over $\tilde{n} \times n$ matrices, where $\tilde{n}$ is a function of n, d, ε, and δ. Suppose that with probability at least $1-\delta$, for any fixed $n \times d$ matrix $\mathbf{X}$, a matrix $\mathbf{P}$ drawn from distribution $\mathcal{P}$ has the property that $\mathbf{P}$ is a $(1 \pm \varepsilon)$ ℓ^2-subspace embedding for $\mathbf{X}$, that is,

$$(1-\varepsilon)\|\mathbf{X}\boldsymbol{\beta}\|_2^2 \leq \|\mathbf{P}\mathbf{X}\boldsymbol{\beta}\|_2^2 \leq (1+\varepsilon)\|\mathbf{X}\boldsymbol{\beta}\|_2^2 \tag{3}$$

holds for all $\boldsymbol{\beta} \in \mathbb{R}^d$. Then we call $\mathcal{P}$ an (ε, δ) oblivious ℓ^2-subspace embedding.

- We will use the following two types of subspace embedding, Gaussian Subspace Embedding and Sparse Subspace Embedding.

Definition 3.2 (Gaussian Subspace Embedding, [2, Theorem 2.3])

Let $0 < \varepsilon, \delta < 1$, $\mathbf{P} = \frac{\mathbf{P}'}{\sqrt{\tilde{n}}}$ where $\mathbf{P}^{'} \in \mathbb{R}^{\tilde{n} \times n}$ is a matrix whose elements are i.i.d. samples from the standard Gaussian distribution $\mathcal{N}(0, 1)$. Then if $\tilde{n} = \mathcal{O}((r + \log \frac{1}{\delta})\varepsilon^{-2})$, for any matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with $r = \operatorname{rank}(\mathbf{X})$, with probability $1 - \delta$, $\mathbf{P} = \frac{\mathbf{P}'}{\sqrt{\tilde{n}}}$ is a $(1 \pm \varepsilon)$ ℓ^2-subspace embedding for $\mathbf{X}$. $\mathbf{P}$ is named a Gaussian subspace embedding.

Definition 3.3 (Sparse Subspace Embedding)

Let $\mathbf{P} \in \mathbb{R}^{\tilde{n} \times n}$. For each $i \in [n]$, $h(i) \in [\tilde{n}]$ is uniformly chosen from $[\tilde{n}]$, and $\sigma(i)$ is a uniformly random element of $\{1, -1\}$. We then set $\mathbf{P}_{h(i)i} = \sigma(i)$ and set $\mathbf{P}_{ji} = 0$ for all $j \neq i$. As a result, $\mathbf{P}$ has only a single nonzero element per column, and it is called a sparse subspace embedding.

The Iterative SRO Algorithm and Its Guarantee

Algorithm 1 Iterative SRO

```
Input: Initialize $\boldsymbol{\beta}^{(0)} = \mathbf{0}$, iteration number $N > 0$, $t = 0$.
for $t \leftarrow 1$ to $N$
    Set
    $$\boldsymbol{\beta}^{(t)} = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{2}\left\|\tilde{\mathbf{X}}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^{(t-1)}\right)\right\|_2^2 - \left\langle \mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(t-1)}, \mathbf{X}\boldsymbol{\beta}\right\rangle + h_{\lambda}(\boldsymbol{\beta}) \qquad (4)$$
end for
Return $\boldsymbol{\beta}^{(N)}$
```

Theorem 3.1

Suppose $\tilde{\boldsymbol{\beta}}^*$ is any critical point of the objective function in (2), and $\boldsymbol{\beta}^*$ is any critical point of the objective function in (1). Suppose $0 < \varepsilon < \varepsilon_0 < 1$ where ε_0 is a small positive constant, $0 < \delta < 1$, $\mathbf{P}$ is drawn from an (ε, δ) oblivious ℓ^2-subspace embedding over $\tilde{n} \times n$ matrices. Then with probability at least $1 - \delta$ with $\delta \in (0, 1)$, the output of Iterative SRO described by Algorithm 1 satisfies

$$\left\|\boldsymbol{\beta}^{(N)} - \boldsymbol{\beta}^*\right\|_{\mathbf{X}} \le \rho^N \left\|\boldsymbol{\beta}^*\right\|_{\mathbf{X}}$$

for a constant $0 < \rho < 1$ if h is convex, or the Frechet subdifferential of h is L_h-smooth and $\mathbf{X}$ has full column rank with $\frac{L_h}{\sigma^2_{\min}(\mathbf{X})} < (1 - \varepsilon)$. Frechet subdifferential of h is L_h-smooth if $\sup_{\mathbf{u}\in\tilde{\partial} h(\mathbf{x}), \mathbf{v}\in\tilde{\partial} h(\mathbf{y})} \|\mathbf{u}-\mathbf{v}\|_2 \le L_h \|\mathbf{x}-\mathbf{y}\|_2$ for a positive number L_h. In particular, if $\mathbf{P}$ is a Gaussian subspace embedding, then $\tilde{n} = \mathcal{O}\left(\left(r + \log\frac{1}{\delta}\right) \cdot (\rho+1)^2/\rho^2\right)$. If $\mathbf{P}$ is a sparse subspace embedding, then $\tilde{n} = \mathcal{O}\left(r^2/\delta \cdot (\rho+1)^2/\rho^2\right)$. Here $r = \text{rank}(\mathbf{X})$.

Sketching for Sparse Convex Learning

- We study the sparse signal estimation problem by solving the sketched Lasso problem with $h_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$ in the original problem (1) and the sketched problem (2) using our Iterative SRO algorithm.
- The following assumption is frequently used in the sparse signal estimation literature with the convex sparsity-inducing penalty, $\lambda\|\cdot\|_1$.

Assumption 1

(Assumption in [3, 4] for sparse signal estimation) $\rho_{\mathcal{L},+}(s) < \infty, \rho_{\mathcal{L},-}(s) > 0$ are positive constants. Moreover, for $\bar{s} = \left\|\bar{\boldsymbol{\beta}}\right\|_0$, there exists a $k^* \in \mathbb{N}$ such that $k^* \ge 2\bar{s}$ and

$$\rho_{\mathcal{L},+}(k^*)/\rho_{\mathcal{L},-}(2k^* + \bar{s}) \le 1 + 0.5k^*/\bar{s}. \tag{5}$$

Sketching for Sparse Convex Learning

- Suppose Assumption 1 holds, we show in the following theorem that the Iterative SRO described in Algorithm 1 achieves the minimax parameter estimation error of the order $\sqrt{\bar{s}\log d/n}$ where $\bar{s} = \left\|\bar{\boldsymbol{\beta}}\right\|_0$.

Theorem 4.1

Suppose Assumption 1 holds. Let $\lambda = c\sigma\sqrt{\log d/n}$ where c is a positive constant. Suppose Algorithm 1 returns $\tilde{\boldsymbol{\beta}}^* = \boldsymbol{\beta}^{(N)}$ with $\rho \in (0,1)$ in Theorem 3.1, and the iteration number N is chosen as $N = 1 + \log\left(\|\mathbf{X}\|_2\|\boldsymbol{\beta}^*\|_{\mathbf{X}}/(\lambda\mu)\right)/\log(1/\rho)$. Then with probability at least $1-\delta-2/d$ with $\delta \in (0,1)$,

$$\left\|\tilde{\boldsymbol{\beta}}^* - \bar{\boldsymbol{\beta}}\right\|_2 \le \frac{(1+\gamma)\,(c+\mu c+2)\,\sigma}{\rho_{\mathcal{L},-}(\bar{s}+k^*)\cdot\left(1-\gamma\sqrt{0.5}\right)}\sqrt{\frac{\bar{s}\log d}{n}}, \tag{6}$$

where μ is a positive constant, $\gamma = (1+\mu+2/c)/(1-\mu-2/c)$, and μ and c are chosen such that $\gamma\sqrt{0.5} < 1$. In particular, if $\mathbf{P}$ is a Gaussian subspace embedding, then $\tilde{n} = \mathcal{O}\left(\left(r+\log\frac{1}{\delta}\right)\cdot(\rho+1)^2/\rho^2\right)$. If $\mathbf{P}$ is a sparse subspace embedding, then $\tilde{n} = \mathcal{O}\left(r^2/\delta\cdot(\rho+1)^2/\rho^2\right)$. Here $r = \mathrm{rank}(\mathbf{X})$.

Sketching for Sparse Nonconvex Learning

- We now study sparse signal estimation by sparse nonconvex learning where the regularizer h_λ is nonconvex in the original problem (1) and the sketched problem (2), that is, $h_\lambda(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_1 + Q_\lambda(\boldsymbol{\beta})$.
  - $Q_\lambda(\boldsymbol{\beta}) \coloneqq \sum_{j=1}^{d} q_\lambda(\beta_j)$, q_λ is a concave function and β_j is the j-th element of $\boldsymbol{\beta}$. We have $h_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{d} (\lambda |\beta_j| + q_\lambda(\beta_j))$.
  - Following the analysis of sparse parameter vector recovery in [5], $\lambda |\cdot| + q_\lambda(\cdot)$ is a nonconvex function which can be either smoothly clipped absolute deviation (SCAD) [6] or minimax concave penalty (MCP) [7].

Sketching for Sparse Nonconvex Learning

- The following regularity conditions on the concave function q_λ are used in [5].

Assumption 2 (Regularity Conditions on Nonconvex Penalty in [5] for sparse signal recovery)

(a) $q'_\lambda(\beta_j)$ is monotone and Lipschitz continuous. For $\beta'_j > \beta_j$, there exist two constants $\zeta_- \geq 0, \zeta_+ \geq 0$ such that $-\zeta_- \leq \frac{q'_\lambda(\beta'_j) - q'_\lambda(\beta_j)}{\beta'_j - \beta_j} \leq -\zeta_+$.

(b) $q_\lambda(-\beta_j) = q_\lambda(\beta_j)$ for all $\beta_j \in \mathbb{R}$. Also, $q_\lambda(0) = q'_\lambda(0) = 0$.

(c) $q'_\lambda(\beta_j) \leq \lambda$ for all $\beta_j \in \mathbb{R}$, and $\left| q'_{\lambda_1}(\beta_j) - q'_{\lambda_2}(\beta_j) \right| \leq |\lambda_1 - \lambda_2|$ for all $\lambda_1 > 0, \lambda_2 > 0$.

- The following assumption is the standard assumption in [5] for sparse signal estimation with the minimax error rate, that is, $\left\| \boldsymbol{\beta}^* - \bar{\boldsymbol{\beta}} \right\|_2 \leq \mathcal{O}\left(\sqrt{\bar{s} \log d / n} \right)$.

Assumption 3 (Assumption in [5] for sparse signal estimation)

Let $\bar{s} = ||\bar{\boldsymbol{\beta}}||_0$. There exist an integer $\tilde{s}$ such that $\tilde{s} > C\bar{s}$ such that $\rho_{\mathcal{L},+}(\bar{s} + 2\tilde{s}) < \infty$, $\rho_{\mathcal{L},-}(\bar{s} + 2\tilde{s}) > 0$ are two absolute constants. The concavity parameter ζ_- in Assumption 2 satisfies $\zeta_- \leq C' \rho_{\mathcal{L},-}(\bar{s} + 2\tilde{s})$ with constant $C' \in (0, 1)$. Here $C = 144\kappa^2 + 250\kappa$ with $\kappa = (\rho_{\mathcal{L},+}(\bar{s} + 2\tilde{s}) - \zeta_+)/(\rho_{\mathcal{L},-}(\bar{s} + 2\tilde{s}) - \zeta_-)$.

Sketching for Sparse Nonconvex Learning

- The following assumption, Assumption 4, is necessary to achieve the minimax parameter estimation error by sketching, and the subsequent Remark 4.3 in our paper explains that Assumption 4 is mild. That is, if Assumption 3 holds, then Assumption 4 also holds under mild conditions.

Assumption 4 (Assumption in [5] for sparse signal estimation)

Let $\tilde{s}$, C' be the parameters specified in Assumption 3 such that Assumption 3 holds. Then it is assumed that $\rho_{\tilde{\mathcal{L}},+}(\bar{s}+2\tilde{s}) < \infty$, $\rho_{\tilde{\mathcal{L}},-}(\bar{s}+2\tilde{s}) > 0$ are two absolute constants. In addition, ζ_- satisfies $\zeta_- \le C'\rho_{\tilde{\mathcal{L}},-}(\bar{s}+2\tilde{s})$, and $\tilde{s} > \tilde{C}\bar{s}$ where $\tilde{C} = 144\tilde{\kappa}^2 + 250\tilde{\kappa}$ with $\tilde{\kappa} := (\rho_{\tilde{\mathcal{L}},+}(\bar{s}+2\tilde{s}) - \zeta_+)/(\rho_{\tilde{\mathcal{L}},-}(\bar{s}+2\tilde{s}) - \zeta_-)$.

- We have the following sharp bound for the parameter estimation error with sparse nonconvex learning by sketching in Theorem 4.2. We note that the approximate path following method described in [5, Algorithm 1] is used to solve the original problem (1) and the sketched problem (2) to obtain $\tilde{\boldsymbol{\beta}}^*$ and $\tilde{\boldsymbol{\beta}}^*$ such that $\boldsymbol{\beta}^*$ is an critical point of problem (1) and $\tilde{\boldsymbol{\beta}}^*$ is an critical point of (2). We use $\lambda = \Theta\left(\sqrt{\bar{s}\log d/n}\right)$ for both (1) and (2) at the final stage of the path following method [5, Algorithm 1].

Sketching for Sparse Nonconvex Learning

- Minimax optimal parameter estimation error with sparse nonconvex learning by sketching:

Theorem 4.2

Let $\delta \in (0,1)$ and $\mathbf{P} \in \mathbb{R}^{\tilde{n}\times n}$ be a Gaussian subspace embedding defined in Definition 3.2, and $\varepsilon = \min\left\{C_1\sqrt{\log d/n}, \varepsilon_0\right\}$ with C_1, ε_0 being positive constants and $\varepsilon \in (0, (1-C')/2)$. Suppose $\tilde{\boldsymbol{\beta}}^*$ is the optimization result of the sketched problem (2) with sketch size $\tilde{n} \geq n/C_3^2$ for $n \geq \Theta(1)$, Assumption 3 and Assumption 4 hold, $d \geq 5$, and let $s_0 = \bar{s} + 2\tilde{s}$. Then with probability $1-\delta$,, with probability at least $1 - 4/d$,

$$\left\|\tilde{\boldsymbol{\beta}}^* - \bar{\boldsymbol{\beta}}\right\|_2 \leq \frac{C_1\left\|\bar{\boldsymbol{\beta}}\right\|_2\sqrt{(1+\tilde{s}/\bar{s})\rho_{\mathcal{L},+}(s_0)}}{(1+C')/2\cdot\rho_{\mathcal{L},-}(s_0)-\zeta_-}\sqrt{\frac{\bar{s}\log d}{n}} + \frac{22\left(C_1\sqrt{\bar{s}}\left\|\bar{\boldsymbol{\beta}}\right\|_2 + 2\sigma\right)}{\rho_{\mathcal{L},-}(s_0)-\zeta_-}\sqrt{\frac{\bar{s}\log d}{n}}, \qquad (7)$$

where $C_1 = \sqrt{c_0 792 s_0/789 C_3}$ with $C_3 > 1$ being a positive constant.

More Details in Our Paper

- Please refer to our paper for more detailed results, discussions, and simulation results.

Thank you!

M. Pilanci and M. J. Wainwright, "Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1842–1879, Jan. 2016.

D. P. Woodruff, "Sketching as a tool for numerical linear algebra," *Foundations and Trends® in Theoretical Computer Science*, vol. 10, no. 1–2, pp. 1–157, 2014.

Z. Yang, Z. Wang, H. Liu, Y. C. Eldar, and T. Zhang, "Sparse nonlinear regression: Parameter estimation under nonconvexity," in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, ser. JMLR Workshop and Conference Proceedings, M. Balcan and K. Q. Weinberger, Eds., vol. 48. JMLR.org, 2016, pp. 2472–2481.

T. Zhang, "Analysis of multi-stage convex relaxation for sparse regularization," *J. Mach. Learn. Res.*, vol. 11, pp. 1081–1107, 2010.

Z. Wang, H. Liu, and T. Zhang, "Optimal computational and statistical rates of convergence for sparse nonconvex learning problems," *The Annals of Statistics*, vol. 42, no. 6, pp. 2164–2201, 2014.

J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.