Minimax Linear Regression under Measurement Constraints

Yining Wang YININGWA@CS.CMU.EDU
Aarti Singh AARTISINGH@CMU.EDU
Machine Learning Department, Carnegie Mellon University, Pittsburgh PA, USA

Abstract

We consider the problem of linear regression under measurement constraints and derive computationally feasible subsampling strategies to sample a small portion of design (data) points in a linear regression model $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The derived subsampling algorithms are minimax optimal for estimating the regression coefficients $\boldsymbol{\beta}$ under the fixed design setting, up to a small $(1+\epsilon)$ relative factor. Experiments on real-world data confirmed the effectiveness of our subsampling based linear regression algorithm with comparison to several other popular competitors. A longer technical report for this work can be found in (Wang & Singh, 2016).

1. Introduction

We consider the linear regression model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a fixed *design matrix* or *data matrix*, $\boldsymbol{y} \in \mathbb{R}^n$ is the response, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ are i.i.d. white Gaussian noise with variance σ^2 and $\boldsymbol{\beta}$ is a fixed but unknown p-dimensional coefficient vector. We are interested in the setting when no distributional assumptions are made on the data $\mathbf{X}$. If there are more samples than variables ($n > p$) i.e. the "low-dimensional" setting and $\mathbf{X}$ has full column rank, *ordinary least squares* (OLS) estimator

$$\hat{\boldsymbol{\beta}}^{\text{ols}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y} \tag{2}$$

is known to be optimal for estimating both $\boldsymbol{\beta}$ and $\mathbf{X}\boldsymbol{\beta}$.

Despite the optimality of OLS, in practice it may not be feasible to obtain the full n-dimensional response vector $\boldsymbol{y}$ due to sampling or measurement constraints. For example, in some geographical or genetic applications the number of data points n might be equal to the entire population of a region or all genes on human chromosomes. Acquiring response variables (labels) for all data points can then be very expensive or even infeasible. It is then an important question to subsample a small set of "representative" data points to regress on so that the resulting estimation or prediction is as accurate as possible.

In this paper, we present a systematic approach for data subsampling in low-dimensional linear regression models. This problem is known as *experimental design* in the statistics literature (Pukelsheim, 1993), and leads to a combinatorial optimization problem. Our main idea is to consider a convex relaxation of this otherwise computationally intractable problem and perform sampling with respect to the optimal solution of the relaxed convex problem. Our main results are polynomial-time near-optimal minimax subsampling strategies for linear regression with finite-sample guarantees, which greatly generalizes prior attempts at deriving statistically optimal subsampling strategies (Zhu et al., 2015; Chen et al., 2015; Ma et al., 2015).

2. A minimax framework

We consider the following subsampling model:

Definition 2.1 (Subsampling model)**.** *Let $\mathbf{X}$ be a fixed $n\times p$ design matrix with full column rank and k be the subsampling budget, with $p \le k \le n$. An algorithm A first observes $\mathbf{X}$ in full and produces, either deterministically or randomly, a matrix $\widetilde{\mathbf{X}} \in \mathbb{R}^{k\times p}$ such that each row of $\widetilde{\mathbf{X}}$ is equal to a particular row in $\mathbf{X}$ (duplicates allowed). A then observes $\tilde{\boldsymbol{y}} = \widetilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}_k(\mathbf{0}, \sigma^2\mathbf{I}_k)$ and attempts to estimate the underlying model $\boldsymbol{\beta}$. We use $\mathcal{A}(k)$ to denote the set of all such subsampling algorithms.*

The main goal of this paper is to characterize the minimax performance of subsampled linear regression defined as

$$\inf_{A\in\mathcal{A}(k)} \sup_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\right], \tag{3}$$

where the expectation is taken over the noise variables $\boldsymbol{\varepsilon}$ and also the inherent randomness in the algorithm A.

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

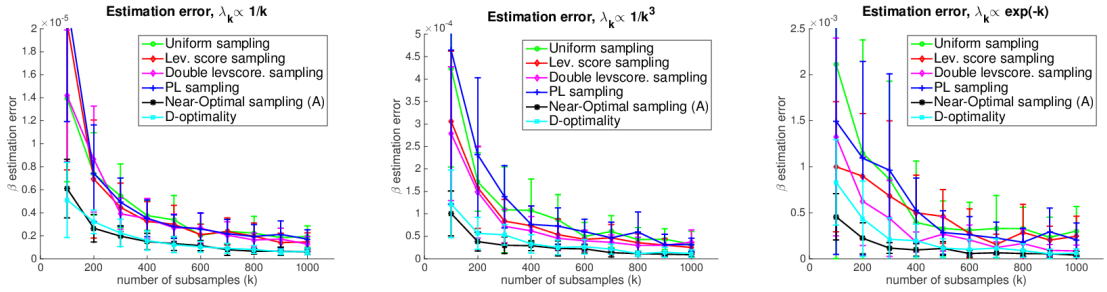

Figure 1. $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$ against number of subsamples k under different spectral decay regimes of $\mathbf{X}$.

3. Near-optimal subsampling

We present computationally efficient algorithms for estimating regression coefficients $\boldsymbol{\beta}$ in the subsampled linear regression framework.

Combinatorial A-optimality Because the variance of $\hat{\boldsymbol{\beta}}^{\text{ols}}$ regressed upon subset $(\mathbf{X}_S, \boldsymbol{y}_S)$ (where $\mathbf{X}_S, \boldsymbol{y}_S$ indicate rows of $\mathbf{X}, \boldsymbol{y}$ corresponding to indices in S) is $\sigma^2 \text{tr}((\mathbf{X}_S^\top \mathbf{X}_S)^{-1})$, a natural formulation is the "A-optimality" criterion:

$$\textbf{A-optimality:} \qquad \min_{|S| \le k} \text{tr}\left((\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\right). \tag{4}$$

Unfortunately, Eq. (4) is a combinatorial optimization and is in general computationally intractable. Approximation methods exist (Avron & Boutsidis, 2013) but their analysis do not reveal improved or near minimax statistical rates.

A convex relaxation A convex relaxation of Eq. (4) is

$$\min_{\pi_1, \cdots, \pi_n} \quad \text{tr}\left((\mathbf{X}^\top \mathbf{diag}(\boldsymbol{\pi}) \mathbf{X})^{-1}\right) \tag{5}$$
$$s.t. \quad \sum_{i=1}^n \pi_i \le 1; \quad \pi_1, \cdots, \pi_n \ge 0.$$

Note that we have substituted 1 for k in the "signal level" $\sum_{i=1}^n \pi_i$ for normalization purposes.

The subsampling algorithm Let $\boldsymbol{\pi}^*$ be the optimal solution and f_{opt} be the optimal value of the objective in Eq. (5), which can be computed in polynomial time via SDP or any conventional convex optimization techniques. The algorithm then obtains k i.i.d. sampled rows of $\mathbf{X}$, where row $\boldsymbol{x}_i$ is sampled with probability $\boldsymbol{\pi}_i^*$. This step is repeated $\Theta(\log n)$ times and $\mathbf{X}_S$ with the smallest $\text{tr}((\mathbf{X}_S^\top \mathbf{X}_S)^{-1})$ is used for subsampled regression. The following theorem shows that such algorithm achieves near-optimal rates for estimating regression coefficients.

Theorem 3.1. *Fix $\mathbf{X} \in \mathbb{R}^{n \times p}$ with full column rank and error tolerance parameter $\epsilon \in (0, 1/2)$. Suppose $\sup_{1 \le i \le n} \|\boldsymbol{x}_i\|_2 \le B < \infty$. Let $\boldsymbol{\Sigma}^* = \sum_{i=1}^n \pi_i^* \boldsymbol{x}_i \boldsymbol{x}_i^\top$. If $k = \Omega(\epsilon^{-2} B^2 \|(\boldsymbol{\Sigma}^*)^{-1}\|_2 \log(n/\epsilon))$ then*

$$\frac{\sigma^2}{k} f_{\text{opt}} \le \inf_{A \in \mathcal{A}(k)} \sup_{\boldsymbol{\beta}} \mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\right] \le \frac{(1+\epsilon)\sigma^2}{k} f_{\text{opt}}.$$

In addition, the algorithm described before achieves the upper bound above with $\text{poly}(n, \log(1/\epsilon))$ running time.

In (Wang & Singh, 2016), we also give interpretable subsampling probabilities for the random design setting and demonstrate explicit gaps in statistical rates between optimal and baseline (e.g., uniform) subsampling methods.

4. Simulation results

We compare our methods on synthetic datasets with existing subsampling strategies in prior literature, which include *uniform sampling* ($\pi_i = 1/n$), *leverage score sampling* ($\pi_i \propto \boldsymbol{x}_i^\top \boldsymbol{\Sigma}_X^{-1} \boldsymbol{x}_i$, (Ma et al., 2015)), *double leverage score sampling* ($\pi_i \propto \|\boldsymbol{\Sigma}_X^{-1} \boldsymbol{x}_i\|_2^2$) Though not a subsampling method, we also compare our algorithm with the popular *D-optimality* criterion [1] which finds a subset S of size k that maximizes $\det(\mathbf{X}_S^\top \mathbf{X}_S)$. For synthetic datasets, we use $n = 10000$ data points with $p = 10$ variables and generate $\boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$ and each row of $\mathbf{X}$ i.i.d. from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_X)$, where $\boldsymbol{\Sigma}_X = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ for some random orthonormal basis $\mathbf{U}$ and $\boldsymbol{\Lambda} = \mathbf{diag}(\lambda_1, \cdots, \lambda_p)$. We set $\sigma^2 = 0.01$ throughout the synthetic experiments. We adopt the sampling with replacement setting, where fresh noise are imposed on the same data point $\boldsymbol{x}_i$ if it is sampled more than once.

Figure 1 depicts the average estimation error against number of subsamples k (ranging from $0.01n$ to $0.1n$) under different spectral decay regimes of $\mathbf{X}$. We observe that the near-optimal sampling strategy (depicted in black lines) outperforms the other subsampling methods, including the approximate D-optimality designs. The performance gap is even larger when the design matrix $\mathbf{X}$ is closer to singular (e.g., exponential spectral decay $\lambda_k \propto e^{-k}$), which is consistent with out theoretical findings.

[1] Implemented using Matlab's `candexch` routine.

References

Avron, Haim and Boutsidis, Christos. Faster subset selection for matrices and applications. *SIAM Journal on Matrix Analysis and Applications*, 34(4):1464–1499, 2013.

Chen, Siheng, Varma, Rohan, Singh, Aarti, and Kovačević, Jelena. Signal recovery on graphs: Random versus experimentally designed sampling. In *SAMPTA*, 2015.

Ma, Ping, Mahoney, Michael W, and Yu, Bin. A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, 16(1):861–911, 2015.

Pukelsheim, Friedrich. *Optimal design of experiments*, volume 50. SIAM, 1993.

Wang, Yining and Singh, Aarti. Minimax subsamping for estimation and prediction in low-dimensional linear regression. *arXiv:1601.02068*, 2016.

Zhu, Rong, Ma, Ping, Mahoney, Michael, and Yu, Bin. Optimal subsampling approaches for large sample linear regression. *arXiv:1509.05111*, 2015.