A Statistical Perspective on Algorithmic Leveraging

Ping Ma PINGMA@UGA.EDU
Department of Statistics, University of Georgia, Athens, GA 30602

Michael W. Mahoney MMAHONEY@ICSI.BERKELEY.EDU
International Computer Science Institute and Dept. of Statistics, University of California at Berkeley, Berkeley, CA 94720

Bin Yu BINYU@STAT.BERKELEY.EDU
Departments of Statistics and EECS, University of California at Berkeley, Berkeley, CA 94720

Abstract

One popular method for dealing with large-scale data sets is sampling. Using the empirical statistical leverage scores as an importance sampling distribution, the method of *algorithmic leveraging* samples and rescales data matrices to reduce the data size before performing computations on the subproblem. Existing work has focused on algorithmic issues, but none of it addresses statistical aspects of this method. Here, we provide an effective framework to evaluate the statistical properties of algorithmic leveraging in the context of estimating parameters in a linear regression model. In particular, for several versions of leverage-based sampling, we derive results for the bias and variance. We show that from the statistical perspective of bias and variance, neither leverage-based sampling nor uniform sampling dominates the other. This result is particularly striking, given the well-known result that, from the algorithmic perspective of worst-case analysis, leverage-based sampling provides uniformly superior worst-case algorithmic results, when compared with uniform sampling. Based on these theoretical results, we propose and analyze two new leveraging algorithms: one constructs a smaller least-squares problem with “shrinked” leverage scores (SLEV), and the other solves a smaller and unweighted (or biased) least-squares problem (LEVUNW). The empirical results indicate that our theory is a good predictor of practical performance of existing and new leverage-based algorithms and that the new algorithms achieve improved performance.

1. Introduction

One popular method for dealing with large-scale data sets is sampling. In this approach, one first chooses a small portion of the full data, and then one uses this sample as a surrogate to carry out computations of interest for the full data. For example, one might randomly sample a small number of constraints or variables in a regression problem and then perform a regression computation on the subproblem thereby defined. For many problems, it is very easy to construct “worst-case” input for which *uniform* random sampling will perform very poorly. Motivated by this, there has been a great deal of work on developing algorithms for matrix-based machine learning and data analysis problems that construct the random sample in a *nonuniform* data-dependent fashion (Mahoney, 2011). Of particular interest here is when that data-dependent sampling process selects rows or columns from the input matrix according to a probability distribution that depends on the empirical statistical leverage scores of that matrix. This recently-developed approach of *algorithmic leveraging* has been applied to matrix-based problems that are of interest in large-scale data analysis, e.g., least-squares approximation (Drineas et al., 2006; 2010), least absolute deviations regression (Clarkson et al., 2013; Meng & Mahoney, 2013), and low-rank matrix approximation (Mahoney & Drineas, 2009; Clarkson & Woodruff, 2013). A detailed discussion of this approach can be found in (Mahoney, 2011). This algorithmic leveraging paradigm has already yielded impressive algorithmic benefits (Avron et al., 2010; Meng et al., 2014). In spite of these impressive *algorithmic* results, none of this recent work on leveraging or leverage-based sampling addresses *statistical* aspects of this approach. This is in spite of the central role of statistical leverage, a traditional concept from regression diagnostics (Hoaglin & Welsch, 1978; Chatterjee & Hadi, 1986; Velleman & Welsch, 1981).

In this paper, we bridge that gap by providing the first

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

statistical analysis of the algorithmic leveraging paradigm. We do so in the context of parameter estimation in fitting linear regression models for large-scale data—where, by "large-scale," we mean that the data define a high-dimensional problem in terms of sample size n, as opposed to the dimension p of the parameter space. Although $n \gg p$ is the classical regime in theoretical statistics, it is a relatively new phenomenon that in practice we routinely see a sample size n in the hundreds of thousands or millions or more. This is a size regime where sampling methods such as algorithmic leveraging are indispensable to meet computational constraints.

Our main theoretical contribution is to provide an analytic framework for evaluating the statistical properties of algorithmic leveraging under a linear regression model. This involves performing a Taylor series analysis around the ordinary least-squares solution to approximate the subsampling estimators as linear combinations of random sampling matrices. Within this framework, we consider biases and variances, both conditioned as well as not conditioned on the data, for several versions of the basic algorithmic leveraging procedure. We show that both leverage-based sampling and uniform sampling are unbiased to leading order; and that while leverage-based sampling improves the "size-scale" of the variance, relative to uniform sampling, the presence of very small leverage scores can inflate the variance considerably. It is well-known that, from the algorithmic perspective of worst-case analysis, leverage-based sampling provides uniformly superior worst-case algorithmic results, when compared with uniform sampling. However, our statistical analysis here reveals that from the statistical perspective of bias and variance, neither leverage-based sampling nor uniform sampling dominates the other.

Based on these theoretical results, we propose and analyze two new leveraging algorithms designed to improve upon vanilla leveraging and uniform sampling algorithms in terms of bias and variance. The first of these (denoted SLEV below) involves increasing the probability of low-leverage samples, and thus it also has the effect of "shrinking" the effect of large leverage scores. The second of these (denoted LEVUNW below) constructs an unweighted version of the leverage-subsampled problem; and thus for a given data set it involves solving a biased subproblem. In both cases, we obtain the algorithmic benefits of leverage-based sampling, while achieving improved statistical performance.

Our main empirical contribution is to provide a detailed evaluation of the statistical properties of these algorithmic leveraging estimators on both synthetic and real data sets. These empirical results indicate that our theory is a good predictor of practical performance for both existing algorithms and our two new leveraging algorithms as well as that our two new algorithms lead to improved performance. In addition, we show that using shrinked leverage scores typically leads to improved conditional and unconditional biases and variances; and that solving a biased subproblem typically yields improved unconditional biases and variances. Depending on whether one is interested in results unconditional on the data (which is more traditional from a statistical perspective) or conditional on the data (which is more natural from an algorithmic perspective), we recommend the use of SLEV or LEVUNW, respectively, in the future.

2. Background, Notation, and Related Work

We consider a Gaussian linear model $\boldsymbol{y} = X\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$, where $\boldsymbol{y}$ is an $n \times 1$ response vector, X is an $n \times p$ *fixed* predictor or design matrix, $\boldsymbol{\beta}_0$ is a $p \times 1$ coefficient vector, and the noise vector $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$. The unknown coefficient $\boldsymbol{\beta}_0$ can be estimated using least squares (LS) method,

$$\text{argmin}_{\beta \in \mathbb{R}^p} ||\boldsymbol{y} - X\boldsymbol{\beta}||^2, \tag{1}$$

where $|| \cdot ||$ represents the Euclidean norm on $\mathbb{R}^n$. The resulting estimate is

$$\hat{\boldsymbol{\beta}}_{ols} = \text{argmin}_{\boldsymbol{\beta}} ||\boldsymbol{y} - X\boldsymbol{\beta}||^2 = (X^T X)^{-1} X^T \boldsymbol{y}, \tag{2}$$

in which case the predicted response vector is $\hat{\boldsymbol{y}} = H\boldsymbol{y}$, where $H = X(X^T X)^{-1} X^T$ is the so-called Hat Matrix, which is of interest in classical regression diagnostics (Hoaglin & Welsch, 1978; Chatterjee & Hadi, 1986; Velleman & Welsch, 1981). The i^{th} diagonal element of H, $h_{ii} = \boldsymbol{x}_i^T (X^T X)^{-1} \boldsymbol{x}_i$, where $\boldsymbol{x}_i^T$ is the i^{th} row of X, is the *statistical leverage* of i^{th} observation or sample. The statistical leverage scores have been used historically to quantify the extent to which an observation is an outlier (Hoaglin & Welsch, 1978; Chatterjee & Hadi, 1986; Velleman & Welsch, 1981), and they will be important for our main results below.

2.1. Algorithmic Leveraging for Least-squares Approximation

A prototypical example of algorithmic leveraging is given by the following meta-algorithm (Drineas et al., 2006; Mahoney, 2011; Drineas et al., 2012), which we call `SubsampleLS`, and which takes as input an $n \times p$ matrix X, where $n \gg p$, a vector $\boldsymbol{y}$, and a probability distribution $\{\pi_i\}_{i=1}^n$, and which returns as output an approximate solution $\tilde{\boldsymbol{\beta}}_{ols}$, which is an estimate of $\hat{\boldsymbol{\beta}}_{ols}$ of Eqn. (2).

- Randomly sample $r > p$ constraints, i.e., rows of X and the corresponding elements of $\boldsymbol{y}$, using $\{\pi_i\}_{i=1}^n$ as an importance sampling distribution.
- Rescale each sampled row/element by $1/\sqrt{r\pi_i}$ to form a weighted LS subproblem.

- Solve the weighted LS subproblem, formally given in Eqn. (3) below, and then return the solution $\tilde{\boldsymbol{\beta}}_{ols}$.

It is convenient to describe SubsampleLS in terms of a random "sampling matrix" S_X^T and a random diagonal "rescaling matrix" D, in the following manner. If we draw r samples (rows or constraints or data points) with replacement, then define an $r \times n$ sampling matrix, S_X^T, where each of the r rows of S_X^T has one non-zero element indicating which row of X (and element of $\boldsymbol{y}$) is chosen in a given random trial. That is, if the k^{th} data unit (or observation) in the original data set is chosen in the i^{th} random trial, then the i^{th} row of S_X^T equals $\mathbf{e}_k$; and thus S_X^T is a random matrix that describes the process of sampling *with* replacement. Then, an $r \times r$ diagonal rescaling matrix D can be defined so that i^{th} diagonal element of D equals $1/\sqrt{r\pi_k}$ if the k^{th} data point is chosen in the i^{th} random trial. With this notation, SubsampleLS constructs and solves the *weighted LS estimator*:

$$\text{argmin}_{\beta \in \mathbb{R}^p} ||DS_X^T\boldsymbol{y} - DS_X^T X\boldsymbol{\beta}||^2. \qquad (3)$$

Since SubsampleLS samples constraints and not variables, the dimensionality of the vector $\tilde{\boldsymbol{\beta}}_{ols}$ that solves the (still overconstrained, but smaller) weighted LS subproblem is the same as that of the vector $\hat{\boldsymbol{\beta}}_{ols}$ that solves the original LS problem. The former may thus be taken as an approximation of the latter, where, of course, the quality of the approximation depends critically on the choice of $\{\pi_i\}_{i=1}^n$. There are several distributions that have been considered previously (Drineas et al., 2006; Mahoney, 2011; Drineas et al., 2012).

- **Uniform Subsampling.** Let $\pi_i = 1/n$, for all $i \in [n]$, i.e., draw the sample uniformly at random.
- **Leverage-based Subsampling.** Let $\pi_i = h_{ii}/\sum_{i=1}^n h_{ii} = h_{ii}/p$ be the normalized statistical leverage scores, i.e., draw the sample according to a sampling distribution that is proportional to the leverage scores of the data matrix X.

Although Uniform Subsampling (with or without replacement) is very simple to implement, it is easy to construct examples where it will perform very poorly (e.g., see (Drineas et al., 2006; Mahoney, 2011)).

Due to the crucial role of the statistical leverage scores, we refer to algorithms of the form of SubsampleLS as the *algorithmic leveraging* approach to approximating LS approximation. Several versions of the SubsampleLS algorithm are of particular interest to us in this paper. We start with two versions that have been studied in the past.

- **Uniform Sampling Estimator (UNIF)** is the estimator resulting from *uniform subsampling* and *weighted LS estimation*, i.e., where Eqn. (3) is solved, where both the sampling and rescaling/reweighting are done with the uniform sampling probabilities. This version corresponds to vanilla uniform sampling, and it's solution will be denoted by $\tilde{\boldsymbol{\beta}}_{UNIF}$.
- **Basic Leveraging Estimator (LEV)** is the estimator resulting from *exact leverage-based sampling* and *weighted LS estimation*, i.e., where Eqn. (3) is solved, where both the sampling and rescaling/reweighting are done with the leverage-based sampling probabilities. This is the basic algorithmic leveraging algorithm that was originally proposed in (Drineas et al., 2006), and it's solution will be denoted by $\tilde{\boldsymbol{\beta}}_{LEV}$.

Motivated by our statistical analysis (to come later in the paper), we will introduce two variants of SubsampleLS; since these are new to this paper, we also describe them here.

- **Shrinked Leveraging Estimator (SLEV)** is the estimator resulting from a *shrinked leverage-based sampling* and *weighted LS estimation*. By shrinked leverage-based sampling, we mean that we will sample according to a distribution that is a convex combination of a leverage score distribution and the uniform distribution, thereby obtaining the benefits of each; and the rescaling/reweighting is done according to the same distribution. Thus, with SLEV, Eqn. (3) is solved, where both the sampling and rescaling/reweighting are done with the above probabilities. This estimator will be denoted by $\tilde{\boldsymbol{\beta}}_{SLEV}$, and to our knowledge it has not been explicitly considered previously.
- **Unweighted Leveraging Estimator (LEVUNW)** is the estimator resulting from a *leverage-based sampling* and *unweighted LS estimation*. That is, after the samples have been selected with leverage-based sampling probabilities, rather than solving the unweighted LS estimator of (3), we will compute the solution of the *unweighted LS estimator*:

$$\text{argmin}_{\beta \in \mathbb{R}^p} ||S_X^T\boldsymbol{y} - S_X^T X\boldsymbol{\beta}||^2. \qquad (4)$$

Whereas the previous estimators all follow the basic framework of sampling and rescaling/reweighting according to the same distribution, with LEVUNW they are essentially done according to two different distributions—the reason being that not rescaling leads to the same solution as rescaling with the uniform distribution. This estimator will be denoted by $\tilde{\boldsymbol{\beta}}_{LEVUNW}$, and to our knowledge it has not been considered previously.

These methods can all be used to estimate the coefficient vector $\boldsymbol{\beta}$, and we will analyze—both theoretically and empirically—their statistical properties in terms of bias and variance.

A naïve algorithm involves using a QR decomposition or the thin SVD of X to obtain the exact leverage scores. Unfortunately, this exact algorithm takes $O(np^2)$ time and is thus no faster than solving the original LS problem exactly. However, (Drineas et al., 2012) developed an algorithm that computes relative-error approximations to all of the leverage scores of X in $O(np\log(p)/\epsilon)$ time, where the error parameter $\epsilon \in (0,1)$. Thus, it provides a way to implement BELV, SLEV, or LEVUNW in $o(np^2)$ time.

Our leverage-based methods for estimating $\boldsymbol{\beta}$ are related to resampling methods (Efron, 1979; Wu, 1986; Miller, 1974b;a; Jaeckel, 1972; Efron & Gong, 1983; Politis et al., 1999). They usually produce resamples at a similar size to that of the full data, whereas algorithmic leveraging is primarily interested in constructing subproblems that are much smaller than the full data. In addition, the goal of resampling is traditionally to perform statistical inference and not to improve the running time of an algorithm, except in the very recent work (Kleiner et al., 2012).

3. Bias and Variance Analysis of Subsampling Estimators

Analyzing the subsampling methods is challenging for at least the following two reasons: first, there are two layers of randomness in the estimators, i.e., the randomness inherent in the linear regression model as well as random subsampling of a particular sample from the linear model; and second, the estimators depends on random subsampling through the inverse of random sampling matrix, which is a nonlinear function.

3.1. Traditional Weighted Sampling Estimators

We start with the bias and variance of the traditional weighted sampling estimator $\tilde{\boldsymbol{\beta}}_W$, given in Eqn. (5) below. The estimate obtained by solving the weighted LS problem of (3) can be represented as

$$\tilde{\boldsymbol{\beta}}_W = (X^TWX)^{-1}X^TW\boldsymbol{y}, \tag{5}$$

where $W = S_X D^2 S_X^T$ is an $r \times r$ diagonal random matrix, i.e., all off-diagonal elements are zeros, and where both S_X and D are defined in terms of the sampling/rescaling probabilities. Clearly, the vector $\tilde{\boldsymbol{\beta}}_W$ can be regarded as a function of the random weight vector $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)^T$, denoted as $\tilde{\boldsymbol{\beta}}_W(\boldsymbol{w})$, where $(w_1, w_2, \ldots, w_n)$ are diagonal entries of W. By setting $\boldsymbol{w}_0$, the vector around which we will perform our Taylor series expansion, to be the all-ones vector, i.e., $\boldsymbol{w}_0 = \mathbf{1}$, then $\tilde{\boldsymbol{\beta}}(\boldsymbol{w})$ can be expanded around the full sample ordinary LS estimate $\hat{\boldsymbol{\beta}}_{ols}$, i.e., $\tilde{\boldsymbol{\beta}}_W(\mathbf{1}) = \hat{\boldsymbol{\beta}}_{ols}$.

Lemma 1 *Let $\tilde{\boldsymbol{\beta}}_W$ be the output of the `SubsampleLS` Algorithm, obtained by solving the weighted LS problem of (3). Then, a Taylor expansion of $\tilde{\boldsymbol{\beta}}_W$ around the point $\boldsymbol{w}_0 = \mathbf{1}$ yields*

$$\tilde{\boldsymbol{\beta}}_W = \hat{\boldsymbol{\beta}}_{ols} + (X^TX)^{-1}X^T Diag\left\{\hat{\mathbf{e}}\right\}(\boldsymbol{w} - \mathbf{1}) + R_W,$$

where $\hat{\mathbf{e}} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}_{ols}$ is the LS residual vector, and where R_W is the Taylor expansion remainder.

Remark. The significance of Lemma 1 is that, to leading order, the vector $\boldsymbol{w}$ that encodes information about the sampling process and subproblem construction enters the estimator of $\tilde{\boldsymbol{\beta}}_W$ linearly. The additional error, R_W depends strongly on the details of the sampling process, and in particular will be very different for UNIF, LEV, and SLEV.

Remark. Our approximations hold when the Taylor series expansion is valid, i.e., when R_W is "small," e.g., $R_W = o_p(||\boldsymbol{w} - \boldsymbol{w}_0||)$, where $o_p(\cdot)$ means "little o" with high probability over the randomness in the random vector $\boldsymbol{w}$. Here, we simply make two observations. First, this expression will fail to hold if rank is lost in the sampling process. This is because in general there will be a bias due to failing to capture information in the dimensions that are not represented in the sample. (Recall that one may use the Moore-Penrose generalized inverse for inverting rank-deficient matrices.) Second, this expression will tend to hold better as the subsample size r is increased. However, for a fixed value of r, the linear approximation regime will be larger when the sample is constructed using information in the leverage scores—since, among other things, using leverage scores in the sampling process is designed to preserve the rank of the subsampled problem (Drineas et al., 2006; Mahoney, 2011; Drineas et al., 2012).

Given Lemma 1, we can establish the following lemma,

Lemma 2 *The conditional expectation and conditional variance for the traditional algorithmic leveraging procedure, i.e., when the subproblem solved is a weighted LS problem of the form (3), are given by:*

$$\mathbf{E_w}\left[\tilde{\boldsymbol{\beta}}_W|\boldsymbol{y}\right] = \hat{\boldsymbol{\beta}}_{ols} + \mathbf{E_w}\left[R_W\right]; \tag{6}$$

$$\begin{aligned}\mathbf{Var_w}\left[\tilde{\boldsymbol{\beta}}_W|\boldsymbol{y}\right] = {} & (X^TX)^{-1}X^T\left[Diag\left\{\hat{\mathbf{e}}\right\}Diag\left\{\frac{1}{r\boldsymbol{\pi}}\right\}\right. \\ & \left. Diag\left\{\hat{\mathbf{e}}\right\}\right]X(X^TX)^{-1} + \mathbf{Var_w}\left[R_W\right],\end{aligned} \tag{7}$$

where W specifies the probability distribution used in the sampling and rescaling steps. The unconditional expectation and unconditional variance for the traditional algorithmic leveraging procedure are given by:

$$\mathbf{E}\left[\tilde{\boldsymbol{\beta}}_W\right] = \boldsymbol{\beta}_0; \tag{8}$$

$$\begin{aligned}\mathbf{Var}\left[\tilde{\boldsymbol{\beta}}_W\right] = {} & \sigma^2(X^TX)^{-1} + \frac{\sigma^2}{r}(X^TX)^{-1}X^T \\ & Diag\left\{\frac{(1-h_{ii})^2}{\pi_i}\right\}X(X^TX)^{-1} + \mathbf{Var}\left[R_W\right].\end{aligned} \tag{9}$$

Remark. Eqn. (6) states that, when the $\mathbf{E}_{\mathbf{w}}\left[R_W\right]$ term is negligible, i.e., when the linear approximation is valid, then, conditioning on the observed data $\boldsymbol{y}$, the estimate $\tilde{\boldsymbol{\beta}}_W$ is approximately unbiased, relative to the full sample ordinarily LS estimate $\hat{\boldsymbol{\beta}}_{ols}$; and Eqn. (8) states that the estimate $\tilde{\boldsymbol{\beta}}_W$ is unbiased, relative to the "true" value $\boldsymbol{\beta}_0$ of the parameter vector $\boldsymbol{\beta}$. That is, given a particular data set $(X, \boldsymbol{y})$, the conditional expectation result of Eqn. (6) states that the leveraging estimators can approximate well $\hat{\boldsymbol{\beta}}_{ols}$; and, as a statistical inference procedure for arbitrary data sets, the unconditional expectation result of Eqn. (8) states that the leveraging estimators can infer well $\boldsymbol{\beta}_0$.

Remark. Both the conditional variance of Eqn. (7) and the (second term of the) unconditional variance of Eqn. (9) are inversely proportional to the subsample size r; and both contain a sandwich-type expression, the middle of which depends on how the leverage scores interact with the sampling probabilities. Moreover, the first term of the unconditional variance, $\sigma^2(X^TX)^{-1}$, equals the variance of the ordinary LS estimator; this implies, e.g., that the unconditional variance of Eqn. (9) is larger than the variance of the ordinary LS estimator, which is consistent with the Gauss-Markov theorem.

3.2. Leverage-based Sampling and Uniform Sampling Estimators

Here, we specialize Lemma 2 by stating two lemmas. A key conclusion from the lemmas is that, with respect to their variance or MSE, neither LEV nor UNIF is uniformly superior for all input.

We start with the bias and variance of the leverage subsampling estimator $\tilde{\boldsymbol{\beta}}_{LEV}$.

Lemma 3 *The conditional expectation and conditional variance for the LEV procedure are given by:*

$$\mathbf{E}_{\mathbf{w}}\left[\tilde{\boldsymbol{\beta}}_{LEV}|\boldsymbol{y}\right]=\hat{\boldsymbol{\beta}}_{ols}+\mathbf{E}_{\mathbf{w}}\left[R_{LEV}\right];$$
$$\mathbf{Var}_{\mathbf{w}}\left[\tilde{\boldsymbol{\beta}}_{LEV}|\boldsymbol{y}\right]=\frac{p}{r}(X^TX)^{-1}X^T\left[Diag\left\{\hat{\mathbf{e}}\right\}Diag\left\{\frac{1}{h_{ii}}\right\}\right.$$
$$\left.Diag\left\{\hat{\mathbf{e}}\right\}\right]X(X^TX)^{-1}+\mathbf{Var}_{\mathbf{w}}\left[R_{LEV}\right].$$

The unconditional expectation and unconditional variance for the LEV procedure are given by:

$$\mathbf{E}\left[\tilde{\boldsymbol{\beta}}_{LEV}\right]=\boldsymbol{\beta}_0;$$
$$\mathbf{Var}\left[\tilde{\boldsymbol{\beta}}_{LEV}\right]=\sigma^2(X^TX)^{-1}+\frac{p\sigma^2}{r}(X^TX)^{-1}X^T$$
$$Diag\left\{\frac{(1-h_{ii})^2}{h_{ii}}\right\}X(X^TX)^{-1}+\mathbf{Var}\left[R_{LEV}\right]. \quad (10)$$

Remark. Two points are worth making. First, the variance expressions for LEV depend on the size (i.e., the number of columns and rows) of the $n \times p$ matrix X and the number of samples r as p/r. This variance size-scale many be made to be very small if $p \ll r \ll n$. Second, the sandwich-type expression depends on the leverage scores as $1/h_{ii}$, implying that the variances could be inflated to arbitrarily large values by very small leverage scores. Both of these observations will be confirmed empirically in Section 4.

We next turn to the bias and variance of the uniform subsampling estimator $\tilde{\boldsymbol{\beta}}_{UNIF}$.

Lemma 4 *The conditional expectation and conditional variance for the UNIF procedure are given by:*

$$\mathbf{E}_{\mathbf{w}}\left[\tilde{\boldsymbol{\beta}}_{UNIF}|\boldsymbol{y}\right]=\hat{\boldsymbol{\beta}}_{ols}+\mathbf{E}_{\mathbf{w}}\left[R_{UNIF}\right]$$
$$\mathbf{Var}_{\mathbf{w}}\left[\tilde{\boldsymbol{\beta}}_{UNIF}|\boldsymbol{y}\right]=\frac{n}{r}(X^TX)^{-1}X^T\left[Diag\left\{\hat{\mathbf{e}}\right\}Diag\left\{\hat{\mathbf{e}}\right\}\right]$$
$$X(X^TX)^{-1}+\mathbf{Var}_{\mathbf{w}}\left[R_{UNIF}\right]. \quad (11)$$

The unconditional expectation and unconditional variance for the UNIF procedure are given by:

$$\mathbf{E}\left[\tilde{\boldsymbol{\beta}}_{UNIF}\right]=\boldsymbol{\beta}_0;$$
$$\mathbf{Var}\left[\tilde{\boldsymbol{\beta}}_{UNIF}\right]=\sigma^2(X^TX)^{-1}+\frac{n}{r}\sigma^2(X^TX)^{-1}X^T$$
$$Diag\left\{(1-h_{ii})^2\right\}X(X^TX)^{-1}+\mathbf{Var}\left[R_{UNIF}\right].(12)$$

Remark. Two points are worth making. First, the variance expressions for UNIF depend on the size (i.e., the number of columns and rows) of the $n \times p$ matrix X and the number of samples r as n/r. Since this variance size-scale is very large, e.g., compared to the p/r from LEV, these variance expressions will be large unless r is nearly equal to n. Second, the sandwich-type expression is not inflated by very small leverage scores.

Remark. Apart from a factor n/r, the conditional variance for UNIF, as given in Eqn. (11), is the same as Hinkley's weighted jackknife variance estimator (Hinkley, 1977).

3.3. Novel Leveraging Estimators

In view of Lemmas 3 and 4, we consider several ways to take advantage of the complementary strengths of the LEV and UNIF procedures. Recall that we would like to sample with respect to probabilities that are "near" those defined by the empirical statistical leverage scores. We at least want to identify large leverage scores to preserve rank. This helps ensure that the linear regime of the Taylor expansion is large, and it also helps ensure that the scale of the variance is p/r and not n/r. But we would like to avoid rescaling by $1/h_{ii}$ when certain leverage scores are extremely small, thereby avoiding inflated variance estimates.

3.3.1. The Shrinked Leveraging (SLEV) Estimator

Consider first the SLEV procedure. As described in Section 2.1, this involves sampling and reweighting with respect to a distribution that is a convex combination of the empirical leverage score distribution and the uniform distribution. That is, let π^{Lev} denote a distribution defined by the normalized leverage scores (i.e., $\pi_i^{Lev} = h_{ii}/p$), and let π^{Unif} denote the uniform distribution (i.e., $\pi_i^{Unif} = 1/n$, for all $i \in [n]$); then the sampling probabilities for the SLEV procedure are of the form

$$\pi_i = \alpha\pi_i^{Lev} + (1-\alpha)\pi_i^{Unif}, \tag{13}$$

where $\alpha \in (0,1)$.

Since SLEV involves solving a weighted LS problem of the form of Eqn. (3), expressions of the form provided by Lemma 2 hold immediately. In particular, SLEV enjoys approximate unbiasedness, in the same sense that the LEV and UNIF procedures do. The particular expressions for the higher order terms can be easily derived, but they are much messier and less transparent than the bounds provided by Lemmas 3 and 4 for LEV and UNIF, respectively. Thus, rather than presenting them, we simply point out several aspects of the SLEV procedure that should be immediate, given our earlier theoretical discussion. First, note that $\min_i \pi_i \geq (1-\alpha)/n$, with equality obtained when $h_{ii} = 0$. Thus, assuming that $1-\alpha$ is not extremely small, e.g., $1-\alpha = 0.1$, then none of the SLEV sampling probabilities is too small, and thus the variance of the SLEV estimator does not get inflated too much, as it could with the LEV estimator. Second, assuming that $1-\alpha$ is not too large, e.g., $1-\alpha = 0.1$, then the amount of oversampling that is required, relative to the LEV procedure, is not much, e.g., 10%. In this case, the variance of the SLEV procedure has a scale of p/r, as opposed to n/r scale of UNIF, assuming that r is increased by that 10%. Third, since Eqn. (13) is still required to be a probability distribution, combining the leverage score distribution with the uniform distribution has the effect of not only increasing the very small scores, but it also has the effect of performing shrinkage on the very large scores. Finally, all of these observations also hold if, rather that using the exact leverage score distribution (which recall takes $O(np^2)$ time to compute), we instead use approximate leverage scores, as computed with the fast algorithm of (Drineas et al., 2012). For this reason, this approximate version of the SLEV procedure is the most promising for very large-scale applications.

3.3.2. The Unweighted Leveraging (LEVUNW) Estimator

Consider next the LEVUNW procedure. As described in Section 2.1, this estimator is different than the previous estimators, in that the sampling and reweighting are done according to different distributions. For this reason, we have examined the bias and variance of the unweighted leveraging estimator $\tilde{\boldsymbol{\beta}}_{LEVUNW}$. Rather than presenting these lemmas in detail, we mention three remarks.

Remark. Since the sampling and reweighting are performed according to different distributions, the point about which the Taylor expansion is performed, as well as the prefactors of the linear term, are somewhat different than in Section 3.1

Remark. The two expectation results state: (i), when $\mathbf{E}_{\mathbf{w}}[R_{LEVUNW}]$ is negligible, then, conditioning on the observed data $\boldsymbol{y}$, the estimator $\tilde{\boldsymbol{\beta}}_{LEVUNW}$ is approximately unbiased, relative to the full sample *weighted* LS estimator $\hat{\boldsymbol{\beta}}_{wls}$; and (ii) the estimator $\tilde{\boldsymbol{\beta}}_{LEVUNW}$ is unbiased, relative to the "true" value $\boldsymbol{\beta}_0$ of the parameter vector $\boldsymbol{\beta}$. That is, if we apply LEVUNW to a given data set N times, then the average of the N LEVUNW estimates are *not* centered at the LS estimate, but instead are centered roughly at the weighted least squares estimate; while if we generate many data sets from the true model and apply LEVUNW to these data sets, then the average of these estimates is centered around true value $\boldsymbol{\beta}_0$.

Remark. As expected, when the leverage scores are all the same, the variance is the same as the variance of uniform random sampling. This is expected since, when reweighting with respect to the uniform distribution, one does not change the problem being solved, and thus the solutions to the weighted and unweighted LS problems are identical. More generally, the variance is not inflated by very small leverage scores, as it is with LEV. For example, the conditional variance expression is also a sandwich-type expression, the center of which is $W_0 = Diag\{rh_{ii}/n\}$, which is not inflated by very small leverage scores.

4. Main Empirical Evaluation

We consider synthetic data of 1000 runs generated from $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(0, 9I_n)$, where several different values of n and p, leading to both "very rectangular" and "moderately rectangular" matrices X, are considered. The design matrix X is generated from one of three different classes of distributions introduced below.

- **Nearly uniform leverage scores (GA).** We generated an $n \times p$ matrix X from multivariate normal $N(\mathbf{1}_p, \Sigma)$, where the (i,j)th element of $\Sigma_{ij} = 2 \times 0.5^{|i-j|}$, and where we set $\boldsymbol{\beta} = (\mathbf{1}_{10}, 0.1\mathbf{1}_{p-20}, \mathbf{1}_{10})^T$. (Referred to as GA data.)
- **Moderately nonuniform leverage scores (T_3).** We generated X from multivariate t-distribution with 3 degree of freedom and covariance matrix Σ as before. (Referred to as T_3 data.)

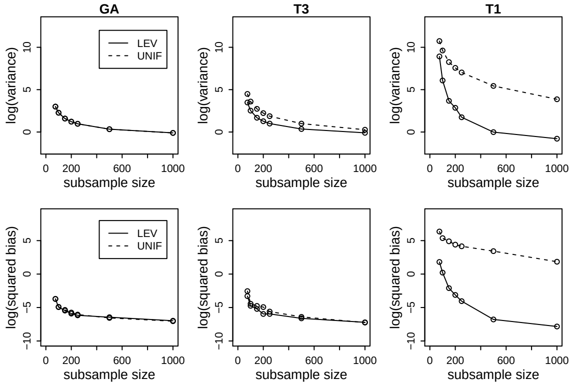

Figure 1. Comparison of variances and squared biases of the LEV and UNIF estimators in three data sets (GA, T_3, and T_1) for $n = 1000$ and $p = 50$. Left panels are GA data; Middle panels are T_3 data; Right panels are T_1 data. Upper panels are Logarithm of Variances; Lower panels are Logarithm of Squared bias. Black lines are LEV; Dash lines are UNIF.

- **Very nonuniform leverage scores (T_1).** We generated X from multivariate t-distribution with 1 degree of freedom and covariance matrix Σ as before. (Referred to as T_1 data.)

4.1. Leveraging Versus Uniform Sampling on Synthetic Data

Here, we will describe the properties of LEV versus UNIF for synthetic data. See Figure 1 for the results on data matrices with $n = 1000$ and $p = 50$. (The results for data matrices for other values of n are similar.)

The simulation results corroborate what we have learned from our theoretical analysis, and there are several things worth noting. First, in general the squared bias is much less than the variance, even for the T_1 data, suggesting that the solution is unbiased in the sense quantified in Lemmas 3 and 4. Second, LEV and UNIF perform very similarly for GA, somewhat less similarly for T_3, and quite differently for T_1, indicating that the leverage scores are very uniform for GA and very nonuniform for T_1. In addition, when they are different, LEV tends to perform better than UNIF, i.e., have a lower MSE for a fixed sampling complexity. Third, as the subsample size increases, the squared bias and variance tend to decrease monotonically. In particular, the variance tends to decrease roughly as $1/r$, where r is the size of the subsample, in agreement with Lemmas 3 and 4. Moreover, the decrease for UNIF is much slower, in a manner more consistent with the leading term of n/r in Eqn. (12), than is the decrease for LEV, which by Eqn. (10) has leading term p/r. All in all, LEV is comparable to or outperforms UNIF, especially when the leverage scores are nonuniform.

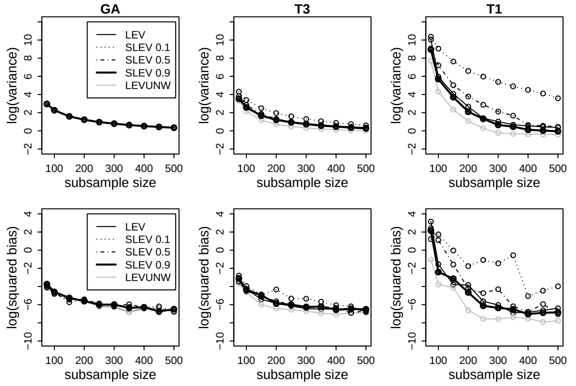

Figure 2. Comparison of variances and squared biases of the LEV, SLEV, and LEVUNW estimators in three data sets (GA, T_3, and T_1) for $n = 1000$ and $p = 50$. Left panels are GA data; Middle panels are T_3 data; Right panels are T_1 data. Grey lines are LEVUNW; black lines are LEV; dotted lines are SLEV with $\alpha = 0.1$; dotdashed lines are SLEV with $\alpha = 0.5$; thick black lines are SLEV with $\alpha = 0.9$.

4.2. Improvements from Shrinked Leveraging and Unweighted Leveraging

Consider Figure 2, which present the variance and bias for synthetic data matrices (for GA, T_3, and T_1 data) of size $n \times p$, where $n = 1000$ and $p = 50$. In each case, LEV, SLEV for three different values of the convex combination parameter α, and LEVUNW were considered. Several observations are worth making. First of all, for GA data , all the results tend to be quite similar; but for T_3 data and even more so for T_1 data, differences appear. Second, SLEV with $\alpha \simeq 0.1$, i.e., when SLEV consists mostly of the uniform distribution, is notably worse in a manner similarly as with UNIF. Moreover, there is a gradual decrease in both bias and variance for our proposed SLEV as α is increased; and when $\alpha \simeq 0.9$ SLEV is slightly better than LEV. Finally, our proposed LEVUNW often has the smallest bias and variance over a wide range of subsample sizes for both T_3 and T_1, although the effect is not major. All in all, these observations are consistent with our main theoretical results.

Consider next Figure 3. This figure examines the optimal convex combination choice for α in SLEV, and α is the x-axis in all the plots. Different column panels in Figure 3 correspond to different subsample sizes r. Recall that there are two conflicting goals for SLEV: adding $(1 - \alpha)/n$ to the small leverage scores will avoid substantially inflating the variance of the resulting estimate by samples with extremely small leverage scores; and doing so will lead to larger sample size r. Figure 3 plots the variance and bias for T_1 data for a range of parameter values and for a range of subsample sizes. In general, one sees that using SLEV to increase the probability of choosing small leverage components with α around $0.8 - 0.9$ (and relatedly shrinking

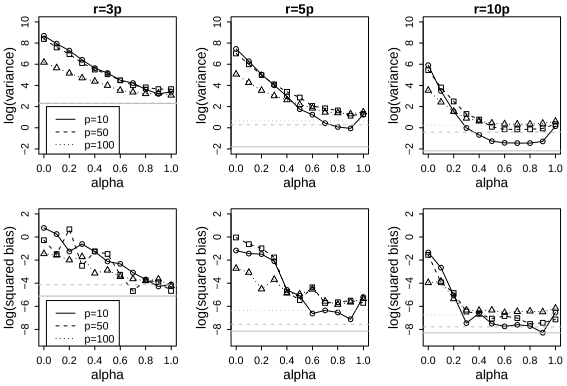

Figure 3. Varying α in SLEV. Comparison of variances and squared biases of the SLEV estimator in data generated from T_1 with $n = 1000$ and variable p. Left panels are subsample size $r = 3p$; Middle panels are $r = 5p$; Right panels are $r = 10p$. Circles connected by black lines are $p = 10$; squares connected by dash lines are $p = 50$; triangles connected by dotted lines are $p = 100$. Grey corresponds to the LEVUNW estimator.

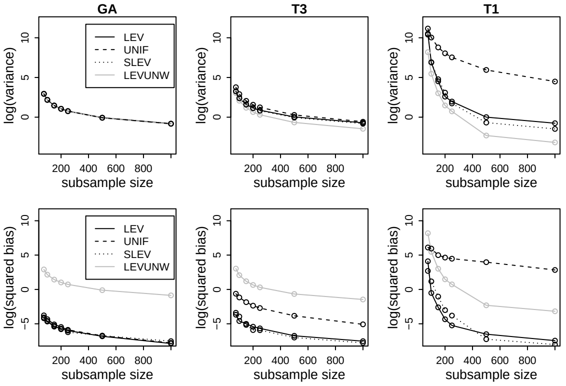

Figure 4. Comparison of *conditional* variances and squared biases of the LEV and UNIF estimators in three data sets (GA, T_3, and T_1) for $n = 1000$ and $p = 50$. Left panels are GA data; Middle panels are T_3 data; Right panels are T_1 data. Upper panels are Variances; Lower panels are Squared Bias. Black lines for LEV estimate; dash lines for UNIF estimate; grey lines for LEVUNW estimate; dotted lines for SLEV estimate with $\alpha = 0.9$.

the effect of large leverage components) has a beneficial effect on bias as well as variance. As a rule of thumb, these plots suggest that choosing $\alpha = 0.9$, and thus using $\pi_i = \alpha\pi_i^{Lev} + (1-\alpha)/n$ as the importance sampling probabilities, strikes a balance between needing more samples and avoiding variance inflation. One can also see in Figure 3 the grey lines, dots, and dashes, which correspond to LEVUNW for the corresponding values of p, that LEVUNW consistently has smaller variances than SLEV for all values of α. We should emphasize, though, that these are *unconditional* biases and variances. Since LEVUNW is approximately unbiased relative to the full sample *weighted* LS estimate $\hat{\boldsymbol{\beta}}_{wls}$, however, there is a large bias away from the full sample *unweighted* LS estimate $\hat{\boldsymbol{\beta}}_{ols}$. This suggests that LEVUNW may be used when the primary goal is to infer the true $\boldsymbol{\beta}_0$; but that when the primary goal is rather to approximate the full sample unweighted LS estimate, or when *conditional* biases and variances are of interest, then SLEV may be more appropriate.

4.3. Conditional Bias and Variance

Consider Figure 4, which presents our main empirical results for conditional biases and variances. As before, matrices were generated from GA, T_3 and T_1; and we calculated the empirical bias and variance of UNIF, LEV, SLEV with $\alpha = 0.9$, and LEVUNW—in all cases, conditional on the empirical data $\boldsymbol{y}$. Several observations are worth making. First, for GA the variances are all very similar; and the biases are also similar, with the exception of LEVUNW. This is expected, since LEVUNW is approximately unbiased, relative to the full sample *weighted* LS estimate $\hat{\boldsymbol{\beta}}_{wls}$—and thus there should be a large bias away from the full sample unweighted LS estimate. Second, for T_3 and even more prominently for T_1, the variance of LEVUNW is less than that for the other estimators. Third, when the leverage scores are very nonuniform, as with T_1, the relative merits of UNIF versus LEVUNW depend on the subsample size r. In particular, the bias of LEVUNW is larger than that of even UNIF for very aggressive downsampling; but it is substantially less than UNIF for moderate to large sample sizes.

Based on these and our other results, our default recommendation is to use SLEV (with either exact or approximate leverage scores) with $\alpha \approx 0.9$: it is no more than slightly worse than LEVUNW when considering unconditional biases and variances, and it can be much better than LEVUNW when considering conditional biases and variances.

5. Discussion and Conclusion

In this paper, we have adopted a *statistical* perspective on algorithmic leveraging, and we have demonstrated how this leads to improved performance of this paradigm on synthetic data. We should note that, while our results are straightforward and intuitive, obtaining them was not easy, in large part due to seemingly-minor differences between problem formulations in statistics, computer science, machine learning, and numerical linear algebra. Now that we have "bridged the gap" by providing a statistical perspective on a recently-popular algorithmic framework, we expect that one can ask even more refined statistical questions of this and other related algorithmic frameworks for large-scale computation.

Acknowledgments: This work was funded in part by NSF DMS-1055815, 1228288, 1228246, and 1228155.

References

Avron, H., Maymounkov, P., and Toledo, S. Blendenpik: Supercharging LAPACK's least-squares solver. *SIAM Journal on Scientific Computing*, 32:1217–1236, 2010.

Chatterjee, S. and Hadi, A. S. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1(3):379–393, 1986.

Clarkson, K. L. and Woodruff, D. P. Low rank approximation and regression in input sparsity time. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing*, pp. 81–90, 2013.

Clarkson, K. L., Drineas, P., Magdon-Ismail, M., Mahoney, M. W., Meng, X., and Woodruff, D. P. The Fast Cauchy Transform and faster robust linear regression. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 466–477, 2013.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1127–1136, 2006.

Drineas, P., Mahoney, M. W., Muthukrishnan, S., and Sarlós, T. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2010.

Drineas, P., Magdon-Ismail, M., Mahoney, M. W., and Woodruff, D. P. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506, 2012.

Efron, B. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

Efron, B. and Gong, G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48, 1983.

Hinkley, D. V. Jackknifing in unbalanced situations. *Technometrics*, 19(3):285–292, 1977.

Hoaglin, D. C. and Welsch, R. E. The hat matrix in regression and ANOVA. *The American Statistician*, 32(1): 17–22, 1978.

Jaeckel, L. The infinitesimal jackknife. *Bell Laboratories Memorandum*, MM:72–1215–11, 1972.

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. The big data bootstrap. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Mahoney, M. W. *Randomized algorithms for matrices and data*. Foundations and Trends in Machine Learning. NOW Publishers, Boston, 2011. Also available at: arXiv:1104.5557.

Mahoney, M. W. and Drineas, P. CUR matrix decompositions for improved data analysis. *Proceedings of National Academy of Sciences*, 106:697–702, 2009.

Meng, X. and Mahoney, M. W. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing*, pp. 91–100, 2013.

Meng, X., Saunders, M. A., and Mahoney, M. W. LSRN: A parallel iterative solver for strongly over- or underdetermined systems. *To appear in: SIAM Journal on Scientific Computing*, 2014.

Miller, R. G. An unbalanced jackknife. *The Annals of Statistics*, 2(5):880–891, 1974a.

Miller, R. G. The jackknife–a review. *Biometrika*, 61(1): 1–15, 1974b.

Politis, D. N., Romano, J. P., and Wolf, M. *Subsampling*. Springer-Verlag, New York, 1999.

Velleman, P. F. and Welsch, R. E. Efficient computing of regression diagnostics. *The American Statistician*, 35(4): 234–242, 1981.

Wu, C. F. J. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295, 1986.