Robust Partially-Compressed Least-Squares

Stephen Becker
University of Colorado Boulder
stephen.becker@colorado.edu

Ban Kawas
IBM T.J. Watson Research Center
bkawas@us.ibm.com

Marek Petrik
University of New Hampshire
mpetrik@cs.unh.edu

Abstract

Randomized matrix compression techniques, such as the Johnson-Lindenstrauss transform, have emerged as an effective and practical way for solving large-scale problems efficiently. With a focus on computational efficiency, however, forsaking solutions quality and accuracy becomes the trade-off. In this paper, we investigate compressed least-squares problems and propose new models and algorithms that address the issue of error and noise introduced by compression. While maintaining computational efficiency, our models provide robust solutions that are more accurate than those of classical compressed variants. We introduce tools from robust optimization together with a form of partial compression to improve the error-time trade-offs of compressed least-squares solvers. We develop an efficient solution algorithm for our *Robust Partially-Compressed* (RPC) model based on a reduction to a one-dimensional search.

Introduction

Random projection is a simple and effective dimensionality reduction technique that enables significant speedups in solving large-scale machine learning problems (Dasgupta 2000; Mahoney 2011; Woodruff 2014). It has been successfully used, for example, in classification (Pilanci and Wainwright 2014; Zhang et al. 2013), clustering (Boutsidis, Zouzias, and Drineas 2010; Fern and Brodley 2003; Urruty, Djeraba, and Simovici 2007), and least-squares problems (Drineas et al. 2011; Pilanci and Wainwright 2014). The focus of this paper will be on the latter. We consider the following canonical least-squares estimator, with $A \in \mathbb{R}^{M\times N}$:

$$x_{\text{LS}} \stackrel{\text{def}}{=} \operatorname*{argmin}_x \ \frac{1}{2}\|Ax-b\|^2 = (A^\mathsf{T}A)^{-1}A^\mathsf{T}b \qquad (1)$$

where $\|\cdot\|$, for vectors, denotes the Euclidean norm throughout the paper, and A has the full column rank. We assume that $M \gg N$ and refer to x_{LS} as the solution to the *uncompressed* problem.

When M is very large, solving the least-squares problem in (1) can be time-consuming and computationally expensive. To gain the necessary speedups, random projections are used. The standard approach to doing so proceeds as follows (Drineas et al. 2011). First, we construct a compression matrix $\Phi \in \mathbb{R}^{m\times M}$ from a random distribution such that $\mathbb{E}\left[\Phi^\mathsf{T}\Phi\right] = I$ and $m \ll M$. Then, we solve the *fully compressed* problem:

$$x_{\text{CLS}} \stackrel{\text{def}}{=} \operatorname*{argmin}_x \ \frac{1}{2}\|\Phi(Ax-b)\|^2 \qquad (2)$$

Numerous efficient methods for constructing the compression matrix Φ have been developed; surveys are provided in (Boutsidis, Drineas, and Magdon-Ismail 2013; Drineas et al. 2011; Mahoney 2011). We describe and use several common methods for constructing Φ in the empirical results section below.

When m in Φ is small, the fully compressed least-squares problem in (2) is much easier to solve than the uncompressed least-squares in (1). However, compression can introduce significant errors to the solution x_{CLS} when compared to the uncompressed solution, x_{LS} and one is forced to consider the trade-off between accuracy and efficiency. As our main contribution, we propose and analyze *two* new models that address this issue and provide a desirable trade-off; enabling robust solutions while preserving small computational complexity.

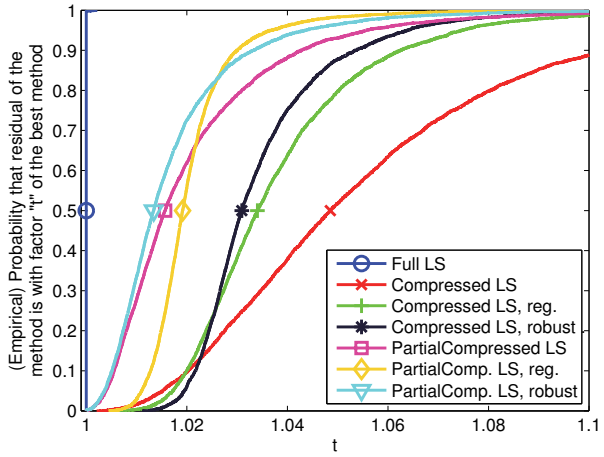

Figure 1: Cumulative probability distribution of residuals of various compressed least-squares methods. The horizontal axis represents the normalized residual compared to solving the full least squares problem.

Results in Fig. 1—which we explain in detail in the experimental section—demonstrate the main issue with standard methods as well as how we alleviate it. Our new model is the following *partially-compressed* least-squares estimator:

$$\min_x \frac{1}{2} \|\Phi\, A\, x\|^2 - b^\mathsf{T} A\, x \tag{3}$$

which is denoted as "PartialCompressed LS" in Fig. 1 and its solution is given by:

$$x_{\text{PCLS}} \stackrel{\text{def}}{=} (A^\mathsf{T}\Phi^\mathsf{T}\Phi A)^{-1} A^\mathsf{T} b\,. \tag{4}$$

Note that only the computationally expensive parts of the ordinary least-squares estimator, which involve inverting $A^\mathsf{T}A$, are compressed. Also notice, in comparison, that the objective function of the fully compressed least-squares estimator is $\frac{1}{2}\|\Phi\, A\, x\|^2 - b^\mathsf{T}\,\Phi^\mathsf{T}\Phi\, A\, x$.

While not the focus of this paper—since our goal here is to introduce our new estimator in (3)—it is important to note that the prediction error $\|Ax_{\text{PCLS}} - Ax_{\text{LS}}\|$ of the partially compressed solution, x_{PCLS}, is not always smaller than the error of the fully compressed one, x_{CLS}. We describe when and why this is the case in the section that deals with the approximation error bounds.

To further reduce the residuals of partial least squares, we have derived our second model, the *robust partially-compressed* least-squares estimator (RPC). It is denoted as "PartialComp. LS, robust" in Fig. 1. RPC explicitly models errors introduced by compression and is closely related to robust least-squares regression (El Ghaoui and Lebret 1997). Leveraging robust optimization techniques makes it possible to reduce the solution error without excessively increasing computational complexity and is a data-driven approach that has been widely used in the last two decades (Ben-Tal, El Ghaoui, and Nemirovski 2009). In our numerical results, we have observed a similar effect to that when applying robust optimization to the fully compressed least-squares solution; increased accuracy and reduction in error.

While we show that RPC can be formulated as a second-order conic program (SOCP), generic off-the-shelf SOCP solvers may be slow for large problems. Therefore, as one of our contributions, we have developed a fast algorithm based on a *one-dimensional search* that can be significantly faster than CPLEX. Using this fast algorithm, the RPC model is asymptotically just as efficient as the non-robust model. Table 1 puts our results in context of prior related work.

Our empirical results, show that both *partially-compressed* and *robust partially-compressed* solutions can outperform models that use full compression in terms of quality of solutions. We also show that compressed variants are more computationally efficient than ordinary least-squares, especially as dimensions grow.

Robust Partially-Compressed Least-Squares

As described above, our objective is to enhance solution quality and increase robustness against noise and errors introduced by compression. One way of improving robustness is to use ridge regression, which when applied to our model (3), we obtain the following formulation:

$$\min_x \frac{1}{2} \|\Phi\, A\, x\|^2 - b^\mathsf{T} A\, x + \mu\|x\|^2, \tag{5}$$

for some regularization parameter μ. One caveat of using ridge regression is that it does not capture the error structure introduced by compression, which differs significantly from that present in the data of the original uncompressed ordinary least-squares problem. Robust optimization (Ben-Tal, El Ghaoui, and Nemirovski 2009), however, enables us to do exactly that and allows us to explicitly model the error structure. The following is our *Robust Partially-Compressed* (RPC) estimator:

$$x_{\text{RPC}} = \operatorname*{argmin}_x \max_{\|\Delta P\|_F \le \rho} \frac{1}{2}\|(P + \Delta P)x\|^2 - b^\mathsf{T} Ax \tag{6}$$

where $P = \Phi A$ and ΔP is a matrix variable of size $m \times N$. The general formulation of the problem allows for a more targeted model of the noise that captures the fact that $\|\Phi Ax\|$ is a random variable while $b^\mathsf{T}Ax$ is not. That is, the uncertainty is restricted to the data matrix P alone since the partial compression does not introduce any noise in the right-hand side.

Without compression, it is worth noting that applying robust optimization techniques to the *ordinary* least-squares problem yields the same solution as applying ridge regression with a data-dependent parameter (El Ghaoui and Lebret 1997). As we will show, this is not the case in our setting, as robust partially-compressed least-squares does not reduce to ridge regression. Empirically, we have also seen that robust partially-compressed least-squares is more likely to yield better results than ridge regression and has more intuition built behind it.

All of the above, motivated us to focus more on our RPC (6) model and to derive a corresponding efficient solution algorithm.

Optimal Solution of RPC

While the inner optimization in (6) is a non-convex optimization problem, we show in the following lemma that there exists a closed-form solution.

Lemma 1. *The inner maximization in* (6) *can be reformulated for any* x *as:*

$$\max_{\|\Delta P\|_F \le \rho} \|(P + \Delta P)x\|^2 = \left(\|Px\| + \rho\|x\|\right)^2\,. \tag{7}$$

In addition, the maximal value is achieved for $\Delta P = \frac{\rho}{\|Px\|\|x\|} Pxx^\mathsf{T}$.

Proof. The objective function can be upper-bounded using the triangle inequality:

$$\begin{aligned}\max_{\|\Delta P\|_F \le \rho} \|(P + \Delta P)x\|^2 &\le \max_{\|\Delta P\|_F \le \rho} \left(\|Px\| + \|\Delta Px\|\right)^2 \\ &\le \left(\|Px\| + \rho\|x\|\right)^2\,.\end{aligned}$$

To show that this bound is tight, consider $\overline{\Delta P} = \frac{\rho}{\|Px\|\|x\|} Pxx^\mathsf{T}$. It can be readily seen that $\|\overline{\Delta P}\|_F = \rho$. Then by algebraic manipulation:

$$\begin{aligned}\max_{\|\Delta P\|_F \le \rho} \|(P + \Delta P)x\|^2 &\ge \|(P + \overline{\Delta P})\, x\|^2 \\ &= \left(\|Px\| + \rho\|x\|\right)^2\,.\end{aligned}$$

□

	Least Squares	Ridge Regression	Robust Least Squares
No Compression	many	e.g., (Boyd and Vandenberghe 2004)	(El Ghaoui and Lebret 1997)
Partial Compression	**new**: (3)	**new**: (5)	**new**: (6)
Full Compression	e.g., (Drineas et al. 2011)	e.g., (Boyd and Vandenberghe 2004)	**new** (but algo. via El Ghaoui)

Table 1: Our work in context of previous results. The equation numbers point to the objective functions for each method.

Using Lemma 1, the robust partially-compressed estimator x_{RPC} is the optimal solution to:

$$\min_x \ \frac{1}{2}\left(\|Px\| + \rho\|x\|\right)^2 - b^\mathsf{T} Ax \ . \tag{8}$$

We now analyze the structure of the optimal solution and point to connections and differences in comparison to results from ridge regression.

Theorem 1. *The optimal solution x_{RPC} to* (8) *must satisfy:*

$$x_{RPC} = \frac{1}{\alpha + \rho\,\beta}(\alpha^{-1}\,P^\mathsf{T}\,P + \rho\,\beta^{-1}\,I)^{-1}A^\mathsf{T}\,b\,, \tag{9}$$

such that $\alpha = \|Px_{RPC}\|$ and $\beta = \|x_{RPC}\|$, or $x_{RPC} = 0$ if $A^\mathsf{T}b = 0$.

Proof. The theorem follows the first-order optimality conditions. The function (8) is everywhere convex, and differentiable everywhere except at $x = 0$. We can show the solution $x = 0$ is only optimal if $A^\mathsf{T}b = 0$. The objective at $x = 0$ is 0. If $A^\mathsf{T}b \neq 0$, then for sufficiently small $t > 0$, the point $tA^\mathsf{T}b$ gives a strictly negative objective (since $t^2 = o(t)$ as $t \to 0$), hence $x = 0$ is not optimal. If $x \neq 0$, the following first-order conditions are necessary and sufficient:

$$0 = \left(\|Px\| + \rho\,\|x\|\right)\left(\frac{P^\mathsf{T}Px}{\|Px\|} + \rho\frac{x}{\|x\|}\right) - A^\mathsf{T}b,$$

from which we derive (9). The theorem follows directly from setting α and β to the required values. □

Theorem 1 shows that the optimal solution to the robust partially-compressed least-squares problem is structurally similar to a ridge regression solution. The two main differences are that there are two parameters, α and β, and these parameters are data-dependent. When setting ρ to 1—which is what we have done in our empirical study, one advantage over ridge regression would be that there is no need to fine-tune the regularization parameter, μ, and one can rely on only data-driven parameters α and β. Even when there is a need to fine-tune the free parameter ρ in RPC—which we have not done in our results and simply set ρ to be equal to 1—ρ has a structural meaning associated with it; ρ is the size of the uncertainty set in (6) and (7) and one can quickly build an intuition behind how to set its value, which is not the case for the regularization parameter μ. In a current investigation, which is out of the scope of this paper, we are building connections between ρ and the compression dimension m, which will enable us to appropriately set ρ as a function of m.

Note that Theorem 1 does not provide a method to calculate x_{RPC}, since α and β depend on x_{RPC}. However, given that (8) is a convex optimization problem, we are able to reformulate it as the following second-order conic program (SOCP) in standard form:

$$\begin{aligned} \min_{x,t,u,z} \quad & \frac{1}{2}z - b^T Ax \\ \text{s.t.} \quad & \|Px\| \le t\ ,\ \rho\,\|x\| \le u\ ,\ \left\|\begin{matrix} t+u \\ z-\frac{1}{4}\end{matrix}\right\| \le z + \frac{1}{4}\ . \end{aligned} \tag{10}$$

The last constraint in this program translates to $z \geq (t + u)^2$.

While efficient polynomial-time algorithms exists for solving SOCP problems, they are typically significantly slower than solving least-squares. Therefore, to achieve practical speedup, we need to derive a more efficient algorithm. In fact we propose a reduction to a one-dimensional optimization problem in the following section.

Efficient Computation of RPC

In this section, we describe a faster approach than solving the SOCP problem in (10) based on a reduction to a one-dimensional search problem.

Input: A, b, Φ, $P = \Phi A$, ρ
Output: x
$U\,\Sigma\,V^\mathsf{T} \leftarrow \text{SVD}(P)$;
$\tau \leftarrow \rho\,\|b\|_2/2$; `// Initialization`
`// Solve` $x \leftarrow \text{argmin}_x\, h_{\tau_k}(x)$
while $|\,\|\Sigma y\|\gamma_k - 1| \leq \epsilon$ **do**
  $\gamma_k \leftarrow \arg\min_\gamma \phi(\gamma) = \sum_{i=1}^N \frac{\bar{b}_i^2}{\left(\gamma\sigma_i^2+\rho\right)^2} - 1$
  $y_k \leftarrow \frac{1}{\tau}V^\mathsf{T}(P^\mathsf{T}P + \gamma_k I)^{-1}A^\mathsf{T}b$;
  `// When` $\tau = \tau^\star$ `then` $\alpha = \|\Sigma y\|$
  $\tau_{k+1} \leftarrow \tau_k\|\Sigma y_k\|\,\gamma_k$;
end
`// Recover the solution`
$\alpha \leftarrow \frac{\tau}{1+\rho\gamma^\star}$; `// Using:` $\alpha + \rho\,\beta = \tau$
$\beta \leftarrow \frac{\tau-\alpha}{\rho}$;
$x \leftarrow \frac{1}{\beta}Vy$;

Algorithm 1: Efficient Algorithm for Solving RPC

First, we reformulate the optimization problem (8) as:

$$\min_{x,t} \ \frac{1}{2}t^2 - b^\mathsf{T}Ax \quad \text{s.t.} \quad \|Px\| + \rho\|x\| \le t \tag{11}$$

Our goal is to derive and then solve the dual problem. The Lagrangian of (11) is

$$\mathcal{L}(x,t,\tau) = \frac{1}{2}t^2 - b^\mathsf{T}Ax + \tau\left(\|Px\| + \rho\|x\| - t\right)$$

Since strong duality conditions hold, we solve the one-dimensional dual maximization problem $\max_{\tau\geq 0} g(\tau)$

where $g(\tau)$ is given as

$$\min_t \left(\frac{1}{2}t^2 - \tau t\right) + \min_x \tau\left(\|Px\| + \rho\|x\|\right) - b^\mathsf{T}Ax$$
$$= -\frac{1}{2}\tau^2 + \min_x \underbrace{\tau\left(\|Px\| + \rho\|x\|\right) - b^\mathsf{T}Ax}_{h_\tau(x)} . \quad (12)$$

The second equality follows since $\|Px\| + \rho\|x\| = t = \tau$ for the optimal primal and dual solution. Observe that $h_\tau(x)$ is positive homogeneous in x and therefore:

$$\min_x \; h_\tau(x) = \begin{cases} -\infty & \tau < \tau^\star \quad \text{(Case 1)} \\ 0 & \tau = \tau^\star \quad \text{(Case 2)} \\ 0 & \tau > \tau^\star \quad \text{(Case 3)} \end{cases} \quad (13)$$

where $\tau^\star \geq 0$ is the optimal dual value.

Intuitively, to solve for the optimal solution, we need to find the maximal value of τ such that $h_\tau(x) = 0$. Appendix derives the approach that is summarized in Algorithm 1. Observe that the function $h_\tau(x)$ is convex. The main idea is to reduce the optimization to a single-dimensional minimization and solve it using Newton method. We also use the SVD decomposition of P to make the search more efficient so that only a single $\mathcal{O}(N^3)$ step is needed.

In terms of the computational complexity, Algorithm 1 requires $\mathcal{O}(mN^2 + N^3)$ operations. All operations inside of the loop are dominated by $\mathcal{O}(N^3)$. The number of iteration that is needed depends on the desired precision. Table 2 compares the asymptotic computational complexity of the proposed robust partial compression with the complexity of computing the full least-squares solution.

Approximation Error Bounds

Analysis of solution quality is known for the fully compressed least-squares problem (e.g. (Pilanci and Wainwright 2014)). In this section, we derive bounds for the partially-compressed least-squares regression.

First, the following simple analysis elucidates the relative trade-offs in computing full or partial projection solutions. Let $x^\star$ be the solution to the full least-squares problem (3) and $z^\star = b - Ax^\star$ be the residual. Recall that $A^\mathsf{T}z^\star = 0$. Now when x_{CLS} is the solution to (2), then:

$$x_{\text{CLS}} = (A^\mathsf{T}\Phi^\mathsf{T}\Phi A)^{-1}A^\mathsf{T}\Phi^\mathsf{T}\Phi b$$
$$= x^\star + (A^\mathsf{T}\Phi^\mathsf{T}\Phi A)^{-1}A^\mathsf{T}\Phi^\mathsf{T}\Phi z^\star$$

On the other hand, the solution x_{PCLS} to (4) satisfies:

$$x_{\text{PCLS}} = (A^\mathsf{T}\Phi^\mathsf{T}\Phi A)^{-1}A^\mathsf{T}b = (A^\mathsf{T}\Phi^\mathsf{T}\Phi A)^{-1}A^\mathsf{T}Ax^\star$$

The error in x_{CLS} is additive and is a function of the remainder $z^\star$. The error in x_{PCLS} is, on the other hand, multiplicative and is independent of $z^\star$. As a result, a small $z^\star$ will favor the standard fully compressed least-squares formulation, and a large $z^\star$ will favor the new partial compressed one.

We will now show that, in the sense of the following definition, the residual of the optimal solution of the partial projection problem is close to the residual of the true solution of the least-squares problem.

Definition 1 (ϵ-optimal solution)**.** *We say that a solution $\hat{x}$ is ϵ-optimal if it satisfies*

$$\frac{\|A(\hat{x} - x_{LS})\|}{\|Ax_{LS}\|} \leq \epsilon, \quad \epsilon \in (0,1) \quad (14)$$

where x_{LS} is an optimal solution of the original high-dimensional system (1)*.*

For sub-Gaussian and ROS sketches, we can show that results in (Pilanci and Wainwright 2014) can be extended to bound approximation errors for partially-compressed least-squares based on the definition of ϵ-optimal above. These results are nearly independent of the number of rows M in the data matrix (except for how these affect $\|Ax_{\text{LS}}\|$). The main guarantees for unconstrained least-squares are given in the following theorem (proof in appendix) which provides an exponential tail bound:

Theorem 2 (Approximation Guarantee)**.** *Given a normalized sketching matrix $\Phi \in \mathbb{R}^{m\times M}$, and universal constants c_0, c_0', c_1, c_2, the sketched solution x_{PCLS}* (4) *is ϵ-optimal* (14) *with probability at least $1 - c_1\exp(-c_2\, m\, \epsilon^2)$, for any tolerance parameter $\epsilon \in (0,1)$, when the sketch or compression size m is bounded below by*

(i) $m > c_0 \frac{rank(A)}{\epsilon^2}$, if Φ is a scaled sub-Gaussian sketch

(ii) $m > c_0' \frac{rank(A)}{\epsilon^2}\log^4(N)$, if Φ is a scaled randomized orthogonal systems (ROS) sketch

By "scaled" sketch, we mean $\mathbb{E}(\Phi^\mathsf{T}\Phi) = I$, since for partial compression, scaling Φ does affect the answer, unlike full compression. For example, in the Gaussian case, we draw the entries of Φ from $\mathcal{N}(0, \frac{1}{m})$ instead of $\mathcal{N}(0,1)$.

Empirical Results

Our focus in this section is on the improvement of the solution error in comparison with the non-compressed least squares solution as well as the improvement over regular full compression.

We also investigate the computational speed of the algorithms and show that partial compression is just as fast as full compression (and hence sometimes faster than standard least-squares), and that robust partial compression is only roughly twice as slow (and asymptotically it is the same cost).

For completeness, we compare with (ridge-)regularized and robust versions of the standard compressed LS problem (2). The robust version is solved following the algorithm outlined in (El Ghaoui and Lebret 1997) since this can be treated as a robust ordinary least squares problem.

We first use a data set from the National Health Interview Survey from 1992, containing 44085 rows and only 9 columns; since it is highly overcomplete and contains potentially sensitive data, it is a good candidate for sketching. To test over this, we do 100 realizations of 5000 randomly chosen training data and 10000 testing data, and for each realization draw 50 random Walsh-Hadamard sketches with $m = 10N$.

The residual on the testing data (median over all 5000 realizations) is shown in Fig. 1. For robust variants, we set μ

	Least Squares	Robust Partial Compression		
Compression		*Gaussian*	*Walsh-Hadamard*	*Counting*
Comp. Time		$\mathcal{O}(m\,M\,N)$	$\mathcal{O}(M\,\log MN)$	$\mathcal{O}(nnz)$
Solution Time	$\mathcal{O}(M\,N^2)$	$\mathcal{O}(mN^2+N^3)$	$\mathcal{O}(mN^2+N^3)$	$\mathcal{O}(mN^2+N^3)$
Total Time	$\mathcal{O}(M\,N^2)$	$\mathcal{O}(m\,M\,N+mN^2)$	$\mathcal{O}(M\,\log MN+mN^2)$	$\mathcal{O}(nnz+mN^2)$

Table 2: Asymptotic computational complexity of various compression methods. Symbol nnz denotes the number of non-zero elements in A we are assuming that $m \gg N$ and $M \gg N$.

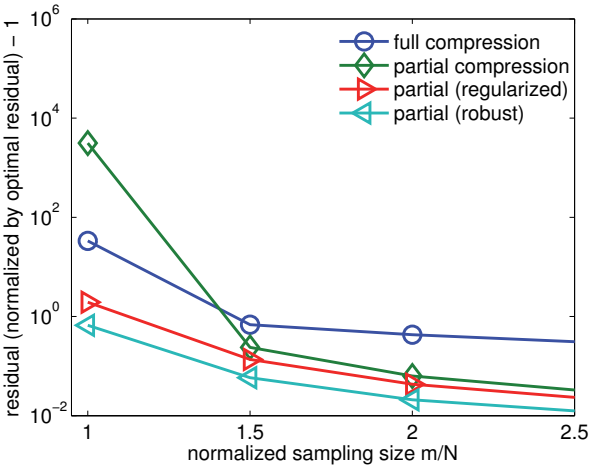

Figure 2: For very high compression (m/M very small), with even $m = N$, robustness/regularization is beneficial.

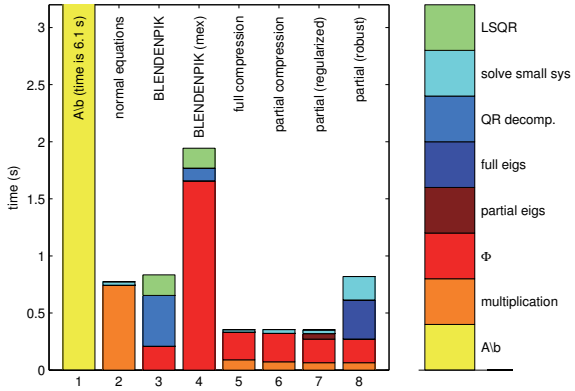

Figure 3: Breakdown of the timing of the parts of the algorithms. Note that the time for $A\backslash b$ continues off the chart.

to be 5 times the minimum eigenvalue of $A^{\mathsf{T}}\Phi^{\mathsf{T}}\Phi A$, and for robust variants we set $\rho = 1$.

Figure 1 presents the results similar to a CDF or a "performance profile" as used in benchmarking software; a smaller area above the curve indicates better performance. A point such as $(0.5, 1.02)$ means that on half the simulations, the method achieved a residual within a factor of 1.02 of the least-square residual.

There are two clear observations from the Fig. 1: partial compression gives lower residuals than full compression, and the regularized and robust variants may do slightly worse in the lower-left (i.e., more bias) but better in the worst-case upper-right (i.e., less variance). Put another way, the robust and regularized versions have stronger tail bounds than the standard versions. We also see a slight benefit of robustness over regularization, though the effect depends on how μ and ρ are chosen.

Figure 3 shows the breakdown of timing for the individual parts of each of the algorithms that we consider. The compression method used the counting sketch in all compressed methods with the exception of Blendenpik (mex) which used the Walsh-Hadamard random matrix vis the Spiral WHT Package (Püschel et al. 2005), and both Blendenpik (Avron, Maymounkov, and Toledo 2010) versions are set to use low-accuracy for the LSQR step. The matrix A is $5 \cdot 10^4 \times 500$ random matrix with condition number 10^6.

Fig. 2 investigates the effect of the number of rows m of the compressed matrix. As expected, the residual quickly deteriorates when m is small (note the logarithmic scale). Employing regularization decreases the residual significantly. Finally, the robust partially-compressed least squares leads to a further improvement over regularization. This is expected since regularization can be seen as a form of an approximation to the robust solution.

We also evaluated our new methods on other synthetic and realistic problems. The additional results also show that robust partial least squares can lead to significant improvements in comparison with the standard projected least squares. Please see the extended version of the paper for these results (Becker et al. 2015).

Conclusion

We developed two new models to address the issue of noise and errors introduced to solutions of compressed least-squares problems; the *partially-compressed* and the *robust partially-compressed* least-squares models. Our models reduce the error introduced by random projection, or sketching, while retaining the computational benefits of matrix compression. The robust model specifically captures the error structure introduced by partial compression, unlike ridge regression with the partially compressed model. Our experimental results indicate that the robust partial compression model outperforms both the partially-compressed model (with or without regularization) as well as the fully compressed one. Partial compression alone can also significantly improve the solution quality over full compression. We also derived the first approximation-error bounds for the

partially-compressed least-squares model.

While the partially-compressed least-squares retains the same computational complexity as full compression, the robust approach introduces an additional difficulty in solving the convex optimization problem. By introducing an algorithm based on one-dimensional parameter search, even the *robust partially-compressed* least-squares can be faster than ordinary least-squares.

Appendix

Minimization of h_τ in Algorithm 1

In this section, we describe how to efficiently minimize $h_\tau(x)$. Recall that h_τ is defined in (12). Now consider the problem of computing $\operatorname{argmin} h_\tau(x)$ for a fixed τ. In case 2 of (13), there exists a solution $x \neq 0$ and therefore the function is differentiable and the optimality conditions read

$$\tau\left(\frac{P^{\mathsf{T}}P}{\alpha} + \rho\frac{I}{\beta}\right)x = A^{\mathsf{T}}b, \quad \alpha = \|Px\|,\ \beta = \|x\|\,. \tag{15}$$

The optimality conditions are scale invariant for $x \neq 0$ and therefore we can construct a solution such that $\beta = \|x\| = 1$.

Let $VDV^{\mathsf{T}} = P^{\mathsf{T}}P$ be an eigenvalue decomposition of $P^{\mathsf{T}}P$, i.e., $D_{ii} = d_i = \sigma_i^2$ are the squared singular values of P, and V are the right singular vectors of $P = U\Sigma V^{\mathsf{T}}$. We make the change-of-variables to $y = V^{\mathsf{T}}x$ (hence $\|y\| = \|x\|$) and define $\bar{b} = V^{\mathsf{T}}A^{\mathsf{T}}b$, which gives an equation for y which is separable if α is fixed. We thus need to solve

$$\tau(\gamma D + \rho I)y = \bar{b} \tag{16}$$

$$1 = \beta = \|y\| \tag{17}$$

$$1/\gamma = \alpha = \|\Sigma y\| \tag{18}$$

Since $d_i \geq 0$ the solution of (16) is unique for a given γ. Therefore, the equations (16)-(18) are satisfied if and only if there exists a γ such that the solution to (16) satisfies both (17) and (18).

We use Newton's method to compute γ that satisfies (16). Define

$$\phi(\gamma) = \tau^{-2}\sum_{i=1}^{N}\frac{\bar{b}_i^2}{\left(\gamma\sigma_i^2 + \rho\right)^2} - 1$$

so (16) and (17) are satisfied if $\phi(\gamma) = 0$ for $\gamma \geq 0$. We note that

$$\phi'(\gamma) = -2\tau^{-2}\sum_{i=1}^{N}\frac{\sigma_i^2\bar{b}_i^2}{\left(\gamma\sigma_i^2 + \rho\right)^3}$$

which is always negative when $\gamma \geq 0$, hence ϕ is monotonic and we are guaranteed that there is a unique root (i.e., it is analogous to convex minimization) in the region $\gamma \geq 0$. We can apply any variant of safe-guarded Newton style methods to solve for the root.

Let $\overline{x} = V^{\mathsf{T}}\overline{y}$ for $\overline{y}$ the optimal solution of the Newton method optimization. We now check if (18) is satisfied for this particular value of $\gamma \equiv \alpha^{-1}$ to determine which case of (13) we are in. If (18) is satisfied and $h_\tau(\overline{x}) = 0$ that means that we are in Case 2 and $\tau = \tau^\star$. That is, the complementary slackness conditions are satisfied and the minimum of $h_\tau(x)$ is 0. If, on the other hand, $h_\tau(\overline{x}) < 0$ then we are in Case 1 and scaling $\overline{x}$ yields an arbitrarily small value. Finally, if $\overline{y}$ does not satisfy (18), then the optimal solution is $y = 0$ and we are in Case 3. Note that $h_\tau(x)$ is not differentiable at $x = 0$.

Finally, if we are in Case 2, then $\overline{x}$ is a scaled optimal solution. To recover the optimal solution, we use $t = \tau$ to appropriately scale $\overline{x}$. Specifically, since we took $\beta = 1$ and worked with $\gamma \equiv \alpha^{-1}$, this was equivalent to working with $\gamma = \beta/\alpha$ so we can recover the properly scaled $\beta^\star = \alpha^\star\gamma$ and hence $\alpha^\star = (1 + \rho\gamma)^{-1}\tau^\star$.

Proof of Theorem 2

Proof. The proof uses the stochastic arguments of (Pilanci and Wainwright 2014) directly, and modifies their deterministic argument (Lemma 1). For brevity, write $\widehat{x} = x_{\text{PCLS}}$ and $x^\star = x_{\text{LS}}$. From the optimality of $\widehat{x}$ to the partial-compressed least squares problem (3), we have:

$$\|\Phi A\widehat{x}\|^2 \leq \|\Phi Ax\|^2 + 2\langle A\left(\widehat{x} - x\right), b\rangle. \tag{19}$$

for all x, and in particular $x = x^\star$. From the optimality of $x^\star$ to equation (1), the gradient at $x^\star$ is zero so we have $\langle A\,x, Ax^\star - b\rangle = 0$ for any x, and hence, using $x = \widehat{x} - x^\star$, re-arranging this gives

$$\langle A(\widehat{x} - x^\star), b\rangle = \langle A(\widehat{x} - x^\star), Ax^\star\rangle \tag{20}$$

Thus

$$\begin{aligned}
&\frac{1}{2}\|\Phi A\left(\widehat{x} - x^\star\right)\|^2 \\
&\quad= \frac{1}{2}\|\Phi A\widehat{x}\|^2 + \frac{1}{2}\|\Phi Ax^\star\|^2 - \langle\Phi A\widehat{x}, \Phi Ax^\star\rangle \\
&\quad\leq \|\Phi Ax^\star\|^2 + \langle A\left(\widehat{x} - x^\star\right), b\rangle - \langle\Phi A\widehat{x}, \Phi Ax^\star\rangle \\
&\quad= \langle A\left(\widehat{x} - x^\star\right), b\rangle - \langle\Phi A\left(\widehat{x} - x^\star\right), \Phi Ax^\star\rangle \\
&\quad= \langle A\left(\widehat{x} - x^\star\right), (I - \Phi^{\mathsf{T}}\Phi)Ax^\star\rangle
\end{aligned}$$

where the first inequality follows from (19) and the final equality follows from (20).

Normalizing both sides of the last inequality appropriately, we obtain:

$$\begin{aligned}
&\frac{1}{2}\underbrace{\frac{\|\Phi A\left(\widehat{x} - x^\star\right)\|^2}{\|A\left(\widehat{x} - x^\star\right)\|^2}}_{U_1}\|A\left(\widehat{x} - x^\star\right)\| \\
&\leq \|Ax^\star\|\underbrace{\left\langle\frac{A\left(\widehat{x} - x^\star\right)}{\|A\left(\widehat{x} - x^\star\right)\|}, (I - \Phi^{\mathsf{T}}\Phi)\frac{Ax^\star}{\|Ax^\star\|}\right\rangle}_{U_2}
\end{aligned}$$

To complete the proof, we need to show that $2\,\frac{U_2}{U_1}$ is bounded above by $\epsilon \in (0, 1)$ for both the sub-Gaussian sketch and the ROS sketch. Define $Z_1(A) = \inf_{v\in\text{range}(A),\ \|v\|=1}\ \|\Phi v\|^2$ and $Z_2(A) = \sup_{v\in\text{range}(A),\ \|v\|=1}\ \left|\left\langle u, \left(\Phi^T\Phi - I\right)v\right\rangle\right|$ where u is any fixed vector of norm 1. Then $U_2/U_1 \leq Z_2/Z_1$.

Taking the scaling of Φ into account, then $Z_2/Z_1 < \epsilon$ if: (a) Φ is a scaled sub-Gaussian sketch and condition *(i)* of the theorem holds, since we apply Lemmas 2 and 3 of (Pilanci and Wainwright 2014); or (b) Φ is a scaled ROS sketch and condition *(ii)* of the theorem holds, since we apply Lemmas 4 and 5 of (Pilanci and Wainwright 2014). □

References

Avron, H.; Maymounkov, P.; and Toledo, S. 2010. Blendenpik: Supercharging lapack's least-squares solver. *SIAM Journal on Scientific Computing* 32(3):1217–1236.

Becker, S.; Kawas, B.; Petrik, M.; and Ramamurthy, K. N. 2015. Robust partially-compressed least-squares. *arXiv:1510.04905*.

Ben-Tal, A.; El Ghaoui, L.; and Nemirovski, A. 2009. *Robust optimization*. Princeton University Press.

Boutsidis, C.; Drineas, P.; and Magdon-Ismail, M. 2013. Near-optimal coresets for least-squares regression. *IEEE Transactions on Information Theory* 59(10):6880–6892.

Boutsidis, C.; Zouzias, A.; and Drineas, P. 2010. Random projections for k-means clustering. In *Advances in Neural Information Processing Systems (NIPS)*, 298–306.

Boyd, S., and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge: Cambridge University Press.

Dasgupta, S. 2000. Experiments with random projection. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 143–151.

Drineas, P.; Mahoney, M. W.; Muthukrishnan, S.; and Sarlós, T. 2011. Faster least squares approximation. *Numerische Mathematik* 117(2):219–249.

El Ghaoui, L., and Lebret, H. 1997. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications* 18(4):1035–1064.

Fern, X. Z., and Brodley, C. E. 2003. Random projection for high dimensional data clustering: A cluster ensemble approach. In *International Conference on Machine Learning (ICML)*, 186–193.

Mahoney, M. W. 2011. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning* 3.

Pilanci, M., and Wainwright, M. J. 2014. Randomized sketches of convex programs with sharp guarantees. *IEEE Transactions on Information Theory* 61:5096–5115.

Püschel, M.; Moura, J. M. F.; Johnson, J.; Padua, D.; Veloso, M.; Singer, B.; Xiong, J.; Franchetti, F.; Gacic, A.; Voronenko, Y.; Chen, K.; Johnson, R. W.; and Rizzolo, N. 2005. SPIRAL: Code generation for DSP transforms. *Proceedings of the IEEE, special issue on "Program Generation, Optimization, and Adaptation"* 93(2):232– 275.

Urruty, T.; Djeraba, C.; and Simovici, D. A. 2007. Clustering by random projections. In *Advances in Data Mining. Theoretical Aspects and Applications*. Springer. 107–119.

Woodruff, D. P. 2014. Sketching as a Tool for Numerical Linear Algebra. *Foundations and Trends in Theoretical Computer Science* 10(1-2):1–157.

Zhang, L.; Mahdavi, M.; Jin, R.; Yang, T.; and Zhu, S. 2013. Recovering optimal solution by dual random projection. In *International Conference on Learning Theory (COLT)*, 135–157.