Big Data Sketching with Model Mismatch

Sundeep Prabhakar Chepuri*, Yu Zhang†, Geert Leus*, and G. B. Giannakis†
*Faculty of EEMCS, Delft University of Technology, The Netherlands.
†Dept. of ECE and the DTC, University of Minnesota, USA.
Emails: {s.p.chepuri; g.j.t.leus}@tudelft.nl; {zhan1220; georgios}@umn.edu.

***Abstract*—Data reduction for large-scale linear regression is one of the most important tasks in this era of data deluge. Exact model information is however not often available for big data analytics. Therefore, we propose a framework for big data sketching (i.e., a data reduction tool) that is robust to possible model mismatch. Such a sketching task is cast as a Boolean min-max optimization problem, and then equivalently reduced to a Boolean minimization program. Capitalizing on the block coordinate descent algorithm, a scalable solver is developed to yield an efficient sampler and a good estimate of the unknown regression coefficient.**

***Index Terms*—Big data, model mismatch, data reduction, linear regression, sketching.**

I. Introduction

Sensor networks, Internet, power grids, biological networks, and social media generate large volumes of data. Such massive datasets have to be processed to extract meaningful information, i.e., to solve an inference problem (e.g., learning or anomaly detection). The shear volume of the data makes this task challenging. Consequently, a significant portion of the data has to be discarded to arrive at a quick rough solution with reduced data processing costs. Instead of blindly discarding the data samples, it is of paramount importance to extract only the most informative data samples for further analysis while keeping the inference task in mind. We will specifically focus on regression in this paper.

Data reduction can be performed even before acquiring the data by solving a sensor selection problem (i.e., an optimal design of experiments) if the model information is known [1], [2]. Here, the samplers (or selection variables) are optimally designed to guarantee an average inference performance. In other words, the samplers are fixed and can be designed offline. On the other hand, if the data is already acquired and available, the data reduction is performed by *sketching* (or censoring) data samples to achieve a desired instantaneous inference performance. Here, the design of samplers are data driven, i.e., the samplers have to be designed for each data realization, which is more appropriate for handling model mismatch and outliers. Censoring is typically used to reduce the communications overheads in a distributed setup, where the data samples available at different locations have to be shipped to a central location for data processing while censored samples are not transmitted to reduce the transmission overheads [3], [4]. Censoring has also been proposed in the big data context to reduce the computational costs involved in solving the inference problem [5], [6]. In both sensor selection and censoring, the samplers are essentially deterministic and structured in nature. Different from the deterministic and structured sampler design, random sampling methods (e.g., random subset selection or random linear projection) have been used for data reduction for least squares problems [7], [8] and these are usually not tailored for a specific task at hand.

Existing works on sketching or censoring are limited to perfectly known data models. However, for massive datasets, such an assumption is often too ideal. That is, the datasets might not always follow the postulated model. For example, the data might be contaminated with outliers or the model information might not be completely known. Censoring with outlier rejection has recently been addressed in [9].

In this work, we propose a framework for big data sketching that is robust to a possible model mismatch. In particular, we are interested in scenarios where the regressors are known only up to a certain bounded perturbation. Without censoring, this is the well-known robust least squares problem [10]–[12]. We pose the problem of sketching that censors less informative samples as well as provides an estimate that performs well for any allowed perturbation. Mathematically, this is a Boolean min-max problem. We provide elegant and scalable algorithms to solve this problem, and apply the developed theory to synthetic as well as real datasets.

Notation. Boldface lower case letters represent column vectors; calligraphic letters stand for sets. The symbol $\mathbb{R}^d$ represents the real space of $d \times 1$ vectors; $\mathbf{a}^T$, $\|\mathbf{a}\|_0$ and $\|\mathbf{a}\|_2$ denote the transpose, ℓ_0 and ℓ_2 norm of $\mathbf{a}$, respectively; and $\mathtt{sgn}(x)$ extracts the sign of x; Finally, the expectation is denoted by $\mathbb{E}[\cdot]$.

II. Problem Statement

Consider a linear regression setup, where an unknown vector $\boldsymbol{\theta} \in \mathbb{R}^p$ is to be estimated from the data collected in the vector $\mathbf{x} = [x_1, x_2, \ldots, x_D]^T \in \mathbb{R}^D$. We assume that $\mathbf{x}$ contains uninformative elements, where we interpret informative entries as data having a large likelihood, i.e., samples having smaller residuals for a linear regression problem with white Gaussian noise. In order to reduce the data processing costs and to obtain a quick rough solution (i.e., to reduce the inference time), inevitably, a significant portion of the data has to be discarded. The data reduction can be achieved either via random sampling (e.g., choosing entries of $\mathbf{x}$ uniformly at random), where the

This work was supported in part by the STW under the FASTCOM project (10551); by the NWO-STW under the VICI program (10382); by the U.S. NSF grants 1514056, 1500713, and NIH 1R01GM104975-01.

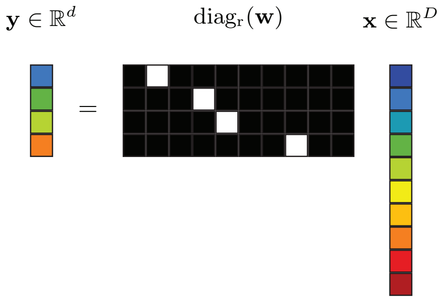
$\mathbf{y} \in \mathbb{R}^d$
$\mathrm{diag_r}(\mathbf{w})$
$\mathbf{x} \in \mathbb{R}^D$
=

Fig. 1: Sparse sketching scheme for data reduction. Here, a white (black) and colored square represents a one (zero) and an arbitrary value, respectively.

inference task is essentially ignored, or the samplers can be systematically designed taking the inference task into account, which is the approach we focus on in this paper. That is, we design samplers for data reduction that are tailored for linear regression problems with model uncertainties.

The dimensionality of the data is reduced from D to d, where $d \ll D$, and both d and D are assumed to be known. We achieve this through a *linear sketching operator* denoted by $\mathrm{diag_r}(\mathbf{w}) \in \{0,1\}^{d \times D}$ to obtain

$$\mathbf{y} = \mathrm{diag_r}(\mathbf{w})\mathbf{x}$$

where $\mathrm{diag_r}(\cdot)$ represents a diagonal matrix with the argument on its diagonal, but with the all-zero rows removed. Such a sketching scheme is illustrated in Fig. 1. The sketching operator is guided by a *censoring vector* denoted by $\mathbf{w} = [w_1, w_2, \dots, w_D]^T \in \{0,1\}^D$, where $w_m = 0$ indicates that x_m is censored (or discarded). The reduced dimension data vector $\mathbf{y}$ is subsequently used to solve the inference or learning problem.

We consider a linear regression problem where the data sample x_m is related to the unknown regression coefficients $\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_p]^T \in \mathbb{R}^p$ through the following linear model

$$x_m = \bar{\mathbf{a}}_m^T \boldsymbol{\theta} + n_m, \quad m = 1, 2, \dots, D \tag{1}$$

where the noise n_m is independent and identically distributed Gaussian variable having (without loss of generality) zero mean and unit variance. The regressors $\{\bar{\mathbf{a}}_m\}_{m=1}^D$ are assumed to be known up to a certain bounded uncertainty. Specifically, the assumption is that

$$\bar{\mathbf{a}}_m = \mathbf{a}_m + \mathbf{p}_m, \quad m = 1, 2, \dots, D \tag{2}$$

where $\{\mathbf{a}_m\}_{m=1}^D$ are assumed to be known a priori, and the perturbations $\{\mathbf{p}_m\}_{m=1}^D$ are unknown yet bounded by a known value η, i.e., $\|\mathbf{p}_m\|_2 \leq \eta$.

We pose the problem of finding a Boolean censoring vector $\mathbf{w}$ that chooses $d \ll D$ data samples censoring less informative samples as well as provides an estimate that performs well for any allowed perturbation. More formally, we state the problem as follows.

Problem statement. *Given the data vector $\mathbf{x} \in \mathbb{R}^D$ that is related to the unknown $\boldsymbol{\theta} \in \mathbb{R}^p$ through a known data model but with bounded uncertainties (i.e., the regressors $\{\boldsymbol{a}_m\}$ and the upper bound on the perturbation η, are perfectly known): (a) design $\mathbf{w}$ to censor the less-informative $D-d$ data samples; and (b) estimate $\boldsymbol{\theta}$ using d uncensored samples that performs well for any allowed perturbation $\{\mathbf{p}_m\}$, where $\|\mathbf{p}_m\|_2 \leq \eta$ for $m = 1, 2, \dots, D$.*

In other words, we seek a Boolean vector $\mathbf{w}$ and the unknown parameter vector $\boldsymbol{\theta}$ that minimizes the worst-case squared residual. This can be cast as the following optimization problem

$$\min_{\mathbf{w} \in \mathcal{W}, \boldsymbol{\theta}} \; \max_{\{\|\mathbf{p}_m\|_2 \leq \eta\}_{m=1}^D} \sum_{m=1}^{D} w_m \left(x_m - (\mathbf{a}_m + \mathbf{p}_m)^T \boldsymbol{\theta}\right)^2 \tag{3}$$

where we introduce the set $\mathcal{W} = \{\mathbf{w} \in \{0,1\}^D \,|\, \|\mathbf{w}\|_0 = d\}$. The optimization problem (3) is a Boolean optimization problem, which is generally hard to solve.

For a fixed d, the approach proposed here yields a solution to data censoring that is optimal in the maximum (worst-case) likelihood sense. In other words, we deem the data with smaller residual values as more informative as compared to the ones with higher residual values, thus it is also robust to possible outliers.

Note that for $\mathbf{p}_m = \mathbf{0}$, problem (3) simplifies to data sketching for linear regression, which can be used to censor less-informative data samples as well as to reject outliers, if any [9].

III. Proposed solvers

In this section, we will solve (3) suboptimally based on an alternating minimization technique. To begin with, we reformulate the min-max problem in (3) to a single minimization problem.

A. Reducing the min-max problem to a minimization problem

Since the perturbations $\{\mathbf{p}_m\}_{m=1}^D$ are independent across the data samples, problem (3) can be equivalently expressed as

$$\min_{\mathbf{w} \in \mathcal{W}, \boldsymbol{\theta}} \sum_{m=1}^{D} w_m \max_{\|\mathbf{p}_m\|_2 \leq \eta} \left(x_m - (\mathbf{a}_m + \mathbf{p}_m)^T \boldsymbol{\theta}\right)^2. \tag{4}$$

Let us define $b_m := \mathbf{a}_m^T \boldsymbol{\theta} - x_m$. Using the Cauchy-Schwarz inequality, we can derive the tightest upper bound for the objective of the inner maximization as

$$\begin{aligned}\left(x_m - (\mathbf{a}_m + \mathbf{p}_m)^T \boldsymbol{\theta}\right)^2 &= (\mathbf{p}_m^T \boldsymbol{\theta})^2 + 2 b_m \mathbf{p}_m^T \boldsymbol{\theta} + b_m^2 \\ &\leq \eta^2 \|\boldsymbol{\theta}\|_2^2 + 2\eta |b_m| \|\boldsymbol{\theta}\|_2 + b_m^2\end{aligned}$$

where the upper bound can be achieved by setting

$$\mathbf{p}_m^* = \eta \, \mathrm{sgn}(b_m) \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2}.$$

Clearly, the worst case perturbation $\mathbf{p}_m^*$ aligns with the regression coefficient $\boldsymbol{\theta}$, and has a length equal to the radius of the uncertainty region η. Plugging the optimal solution $\mathbf{p}_m^*$ back into (4), allows us to state the following lemma.

Lemma 1. *The optimization problem* (4) *is equivalent to the following minimization problem*

$$\min_{\mathbf{w}\in\mathcal{W},\boldsymbol{\theta}} \sum_{m=1}^{D} w_m \left(|x_m - \mathbf{a}_m^T\boldsymbol{\theta}| + \eta\|\boldsymbol{\theta}\|_2\right)^2 \tag{5}$$

which determines the optimal $(\mathbf{w}, \boldsymbol{\theta})$ *with given* $\{\boldsymbol{a}_m\}_{m=1}^D$, $\mathbf{x}$, *and* η.

It can be seen that due to the worst-case model mismatch, the considered problem (5) can be equivalently recast as the following *regularized least trimmed squares* problem

$$\min_{\boldsymbol{\theta}} \sum_{m=1}^{d} r_{[m]}^2(\boldsymbol{\theta}), \tag{6}$$

with the regularized residuals $r_m(\boldsymbol{\theta})$ defined as

$$r_m(\boldsymbol{\theta}) = |x_m - \mathbf{a}_m^T\boldsymbol{\theta}| + \eta\|\boldsymbol{\theta}\|_2$$

and $r_{[m]}^2(\boldsymbol{\theta})$ denoting the squared regularized residuals in ascending order. The value d determines the breakdown point of the regularized least trimmed squares estimator because the $D - d$ largest regularized residuals will not affect the performance. The optimization problem (6) incurs combinatorial complexity, where an exhaustive search would yield choosing the optimal $\boldsymbol{\theta}$ with the smallest cost among all the $\binom{D}{d}$ candidate regularized least squares estimators. Note that solving (6) yields the well-known *least trimmed squares* estimator when $\eta = 0$ [13].

B. Alternating descent minimization

The optimization problem (5) involves the binary variables $\mathbf{w}$, and hence is non-convex and generally hard to solve. However, by exploiting the structure, we propose to use the block coordinate descent (BCD) algorithm for updating $\mathbf{w}$ and $\boldsymbol{\theta}$ alternatingly. That is, whenever the binary variable $\mathbf{w}$ is fixed, (5) boils down to a reduced-ordered robust least squares problem, which is a convex program [10], [11]. Given the regression coefficient $\boldsymbol{\theta}$, the problem becomes a Boolean linear program that admits a closed-form solution with respect to $\mathbf{w}$. Thus, the iterative BCD algorithm can be used to yield successive estimates of $\boldsymbol{\theta}$ with fixed $\mathbf{w}$, and vice versa.

Specifically, the algorithm proceeds as follows. With $\boldsymbol{\theta}[k]$ available, $\mathbf{w}[k+1]$ is obtained as

$$\mathbf{w}[k+1] = \arg\min_{\mathbf{w}\in\mathcal{W}} \sum_{m=1}^{D} w_m r_m^2(\boldsymbol{\theta}[k])\,.$$

Although the above linear programming problem has Boolean and cardinality constraints, there exists a simple analytical solution for $\mathbf{w}[k+1]$ based on ordering the squared regularized residuals $\{r_m^2(\boldsymbol{\theta}[k])\}$. In other words, the optimal $\mathbf{w}$ is obtained by setting $w_m = 1$ corresponding to the d smallest regularized residuals, and setting $w_m = 0$ otherwise.

Given $\mathbf{w}[k]$ and by introducing the auxiliary variables $\{t_m\}$, we solve for $\boldsymbol{\theta}[k+1]$ using a reduced-ordered robust least squares as [12]

$$\boldsymbol{\theta}[k+1] = \arg\min_{\boldsymbol{\theta},\{t_m\}_{m\in\mathcal{S}_k}} \sum_{m\in\mathcal{S}_k} t_m^2 \tag{7a}$$

$$\text{s.t. } |x_m - \mathbf{a}_m^T\boldsymbol{\theta}| + \eta\|\boldsymbol{\theta}\|_2 \leq t_m, \forall m \in \mathcal{S}_k \tag{7b}$$

where $\mathcal{S}_k := \{m | w_m[k] = 1, m = 1, 2, \ldots, D\}$ with $|\mathcal{S}_k| = d$.

Problem (7) is a convex quadratic program which can be further cast into a second-order cone program (SOCP) as

$$\boldsymbol{\theta}[k+1] = \arg\min_{\boldsymbol{\theta},\mathbf{t},u\geq 0} \; u \tag{8a}$$

$$\text{s.t.} \quad \left\| \begin{bmatrix} 2\mathbf{t} \\ u-1 \end{bmatrix} \right\|_2 \leq u+1 \tag{8b}$$

$$|x_m - \mathbf{a}_m^T\boldsymbol{\theta}| + \eta\|\boldsymbol{\theta}\|_2 \leq t_m, \forall m \in \mathcal{S}_k \tag{8c}$$

where $\mathbf{t} \in \mathbb{R}^d$ is a vector with elements $\{t_m\}_{m\in\mathcal{S}_k}$.

The iterations can be initialized at $k = 0$ by setting uniformly at random d entries of $\mathbf{w}[0]$ to 1 such that $\|\mathbf{w}[0]\|_0 = d$. It is worth stressing here that in order to estimate $\boldsymbol{\theta}[k+1]$ by solving (7) or (8), we use only reduced-dimensional data of length $d \ll D$ corresponding to the non-zero entries of $\mathbf{w}[k]$, i.e., $\mathbf{y}[k] = \mathrm{diag_r}(\mathbf{w}[k])\mathbf{x}$. Note that the aforementioned two convex programs can be solved with interior-point methods or first-order algorithms (e.g., projected gradient descent and trust-region methods [14]) to further reduce the computational complexity.

IV. Simulations

In this section, simulated tests are presented to verify the merits of the robustified formulation and the proposed alternating minimization approach. The Matlab function `fminunc` along with the `trust region` solver [15] is used to solve the optimization problem (5) when $\mathbf{w}[k]$ is fixed. All numerical tests are implemented on a high-performance computing cluster with 2 Intel Xeon E5-2690 2.9 GHz CPUs and a total of 204 GB RAM.

The performance of the proposed approach is tested on three different datasets: S1) a small synthetic dataset with $D = 10$ and $p = 2$; S2) a big synthetic dataset with $D = 5,000$ and $p = 10$; and S3) the protein structure dataset from the UCI machine learning repository, where $D = 45,730$ observations and $p = 9$ attributes are used to predict the physicochemical properties of the protein tertiary structure. The entries of the regression vectors $\{\bar{\mathbf{a}}_m\}_{m=1}^D$ contain protein structure revealing parameters obtained via clinical experiments. Hence, the regressors are perturbed and noisy [16]. The true $\boldsymbol{\theta}$ is obtained by solving least squares on the entire dataset. For datasets S1 and S2, 100 Monte-Carlo experiments have been implemented to yield the average worst-case residuals. Parameters $\{\mathbf{a}_m, \mathbf{p}_m, n_m\}_{m=1}^D$ and $\boldsymbol{\theta}$ are independent and identically distributed Gaussian random variables with zero mean, and $\|\mathbf{p}_m\|_2 = \eta$ for $m = 1, 2, \ldots, D$ [cf. (1) and (2)].

We plot the objective value of (5) with $\mathbf{w}$ and $\boldsymbol{\theta}$ obtained from the algorithm described in Sec. III-B, i.e., the average worst-case residual (averaged over a number of Monte-Carlo experiments) in Figs. 2, 3, and 4. Specifically, the proposed robust sampling approach is compared with two alternative competitors: i) the exhaustive search serving as a benchmark that simply finds the optimal $\boldsymbol{\theta}$ with the smallest cost among all $\binom{D}{d}$ possible regularized least squares estimators. Since this brute-force method incurs combinatorial complexity, it is only computationally tractable for small values of D; i.e., for the dataset S1 in our simulation setup; and ii) the random

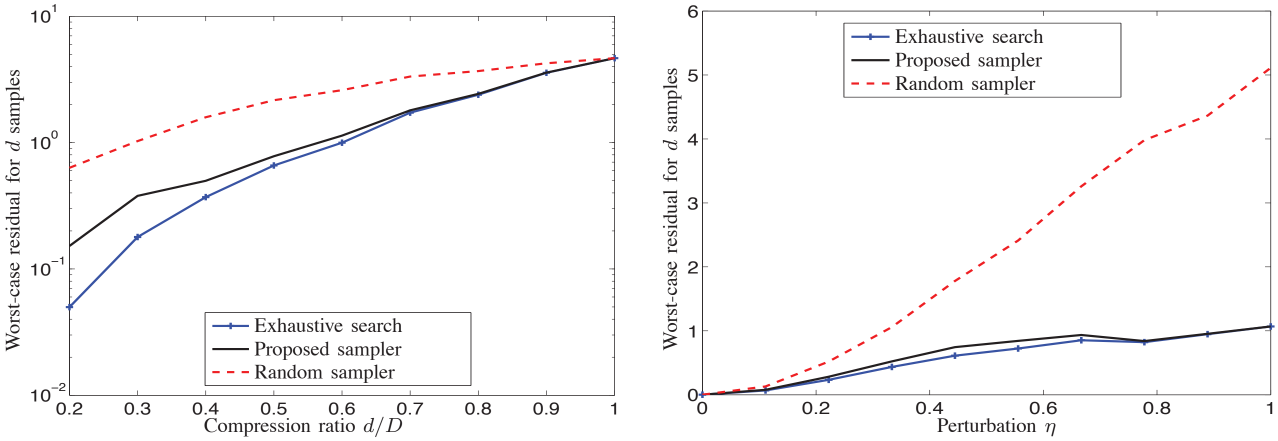

Fig. 2: Small synthetic dataset S1. The average objective value of (5) for: (a) different values of compression ratio fixing $\eta = 0.5$, and (b) different values of η fixing $d/D = 0.5$.

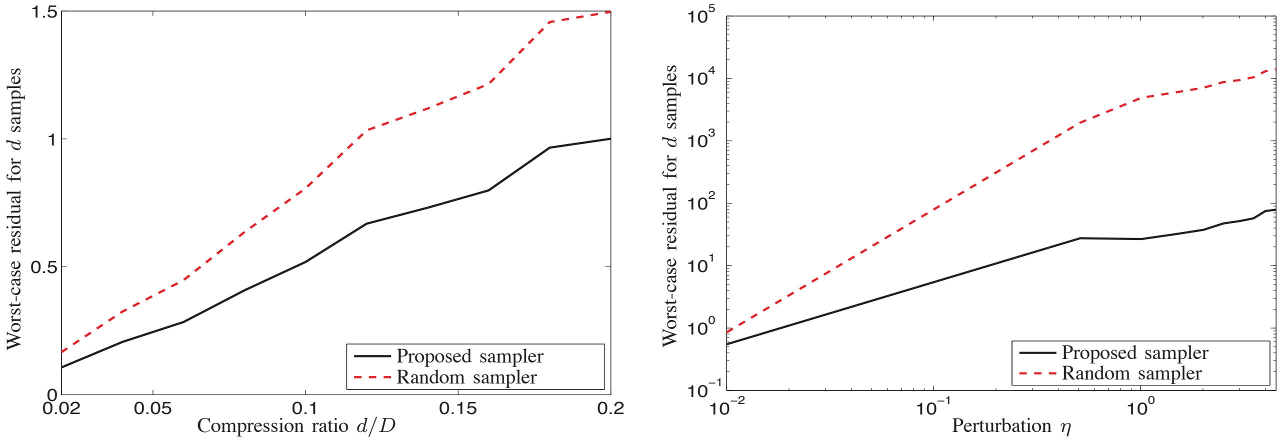

Fig. 3: Big synthetic dataset S2. The average objective value of (5) for: (a) different values of compression ratio fixing $\eta = 0.01$, and (b) different values of η fixing $d/D = 0.1$.

sampler, which randomly chooses d observations, and solves the reduced-dimension robust least squares problem [cf. (5)] to obtain the robust estimator $\hat{\boldsymbol{\theta}}$. Note that the random sampler can be regarded as an initial step of our proposed approach, which updates the sampler $\mathbf{w}[k]$ and the regression coefficient $\boldsymbol{\theta}[k]$ in an alternating descent manner until convergence.

In Fig. 2(a), the worst-case residuals of the three sampling approaches are compared for different compression ratios $d/D = 0.2, 0.3, \ldots, 1$, while the perturbation size is fixed to $\eta = 0.5$. Clearly, for all the samplers, the worst-case residual of d samples increases with the increase of d (also the compression ratio). When $d = D$, all schemes solve the same robust least squares problem (5) with $\mathbf{w} = \mathbf{1}$, and hence yield the same worst-case residual. Since the proposed approach converges to find the best d measurements while the random sampler uniformly chooses d ones, the former edges over the latter across the entire range of the compression ratio, and the performance gain increases with decreasing d/D values. Note that the performance gap between the proposed sampler and the exhaustive search is relatively small, and diminishes with the increase of the compression ratio. For $d/D \geq 0.7$, the proposed one has the same performance as the one of the exhaustive search.

In Fig. 2(b), the worst-case residuals are shown with various perturbation size $\eta \in [0, 1]$ while fixing $d/D = 0.5$. As expected, the worst-case residual increases with an increasing perturbation size. Note that all three curves coincide at the point when $\eta = 0$. In this case, the objective value (5) boils down to the estimate of the noise variance. Hence, there is no difference among the three sampling approaches without model uncertainty. The worst-case residual of our proposed approach is very close to the exhaustive search for different values of η, which also offers a superior performance over the random sampler, especially in the range of large uncertainties.

The worst-case residuals of the proposed approach vis-à-vis the random one for different values of d/D and η are also tested on the big synthetic dataset S2 and the real dataset S3 related to protein structure modeling, as shown in Figs. 3 and 4. Our novel solver achieves good scalability for these large datasets. In all cases, we consistently observe performance gains of the proposed sampler, with a similar trend as the one exhibited in Fig. 2. Note that for the curves

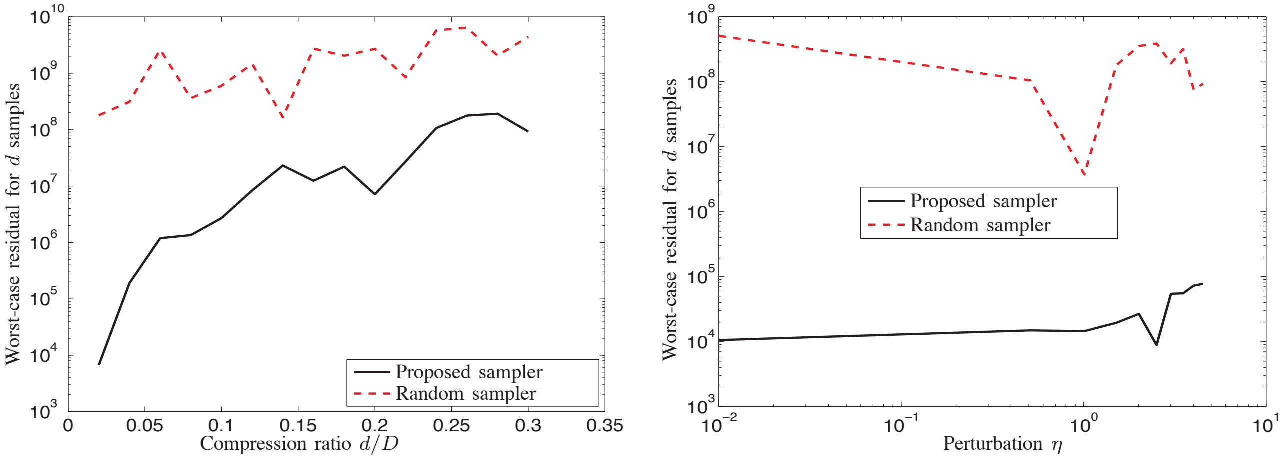

Fig. 4: Real dataset S3. The average objective value of (5) for: (a) different values of compression ratio fixing $\eta = 0.01$, and (b) different values of η fixing $d/D = 0.01$.

in Fig. 4 there are no Monte-Carlo experiments, and hence no smoothing effect for the real dataset S3. Moreover, the mean squared error (MSE) of the estimator $\hat{\boldsymbol{\theta}}$, namely $\mathbb{E}\left[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2\right]$ yielded by the proposed approach is not necessarily better than the one given by the random sampler. The reason is that the goal of problem (5) is to minimize the worst case residual rather than the MSE of the estimator.

V. Conclusions and Future Directions

In this paper, we have discussed data sketching (i.e., a data reduction tool) for large-scale inverse problems with model mismatch. This is relevant as the model information might not be completely available for massive datasets. Specifically, we have considered cases where the regressors are known only up to a certain bounded perturbation. We pose the problem of sketching that censors less informative samples (i.e., samples with large residual values) as well as provides an estimate that performs well for any allowed perturbation. Mathematically, this design of sparse sketching is a Boolean min-max problem. We develop an efficient and scalable solver leveraging the BCD algorithm, whose performance is corroborated by the numerical results with both synthetic and real datasets.

The assumption throughout this paper was that the observation noise is independent across the data samples and the dataset is available as a batch. As a topic for future research, it is interesting to investigate the same problem, but with dependent observations. Further, sketching for streaming big data with model mismatch is certainly an interesting topic for future work.

References

[1] S. Joshi and S. Boyd, “Sensor selection via convex optimization,” *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 451–462, Feb. 2009.

[2] S. P. Chepuri and G. Leus, “Sparsity-promoting sensor selection for non-linear measurement models,” *IEEE Trans. Signal Process.*, vol. 63, no. 3, pp. 684–698, Feb. 2015.

[3] C. Rago, P. Willett, and Y. Bar-Shalom, “Censoring sensors: A low-communication-rate scheme for distributed detection,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 32, no. 2, pp. 554–568, Apr. 1996.

[4] E. J. Msechu and G. B. Giannakis, “Sensor-centric data reduction for estimation with WSNs via censoring and quantization,” *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 400–414, Jan. 2012.

[5] G. Wang, D. K. Berberidis, V. Kekatos, and G. B. Giannakis, “Online reconstruction from big data via compressive censoring,” in *Proc. of 3rd IEEE Global Conf. on Signal and Information Process.*, Atlanta, GA, USA, Dec. 2014.

[6] D. Berberidis, V. Kekatos, and G. B. Giannakis, “Online censoring for large-scale regressions with application to streaming big data,” Jul. 2015, [Online]. Available: http://arxiv.org/pdf/1507.07536v1.pdf.

[7] P. Drineas, M. Mahoney, and S. Muthukrishnan, “Sampling algorithms for ℓ_2 regression and applications,” in *Proc. of 17th annual ACM-SIAM symp. on Discrete algorithm*, Miami, FL, USA, Jan. 2006, pp. 1127–1136.

[8] T. Strohmer and R. Vershynin, “A randomized Kaczmarz algorithm with exponential convergence,” *J. of Fourier Anal. Appl.*, vol. 15, no. 2, pp. 262–278, Apr. 2009.

[9] G. Kail, S. P. Chepuri, and G. Leus, “Robust censoring for linear inverse problems,” in *Proc. of 16th IEEE Intl. Wrksp. on Signal Process. Advances in Wireless Commun.*, Stockholm, Sweden, Jun. 2015, pp. 495–499.

[10] E. Ghaoui and H. Lebret, “Robust solutions to least-squares problems with uncertain data,” *SIAM J. Matrix Anal. Appl.*, vol. 18, no. 4, pp. 1035–1064, Oct. 1997.

[11] S. Chandrasekaran, G. Golub, M. Gu, and A. Sayed, “Parameter estimation in the presence of bounded data uncertainties,” *SIAM J. Matrix Anal. Appl.*, vol. 19, no. 1, pp. 235–252, Jan. 1998.

[12] M. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, “Applications of second-order cone programming,” *Linear Algebra Appl.*, vol. 284, no. 1–3, pp. 193–228, Nov. 1998.

[13] P. J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*, John Wiley & Sons, New York, NY, 2003.

[14] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 2nd edition, 1999.

[15] The MathWorks, Inc., “Optimization toolbox user’s guide,” Sep. 2015, [Online]. Available: http://www.mathworks.com/help/releases/R2015b/pdf_doc/optim/optim_tb.pdf.

[16] M. Lichman, “UCI machine learning repository,” Mar. 2013, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Physicochemical+Properties+of+Protein+Tertiary+Structure.