Spring School Lecture-Part 2: Robust Linear Regression with Recovery Guarantees

CHANDRAJIT BAJAJ (bajaj@cs.utexas.edu) http://www.cs.utexas.edu/~bajaj

Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust Regression via Hard Thresholding. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS)*, 2015.

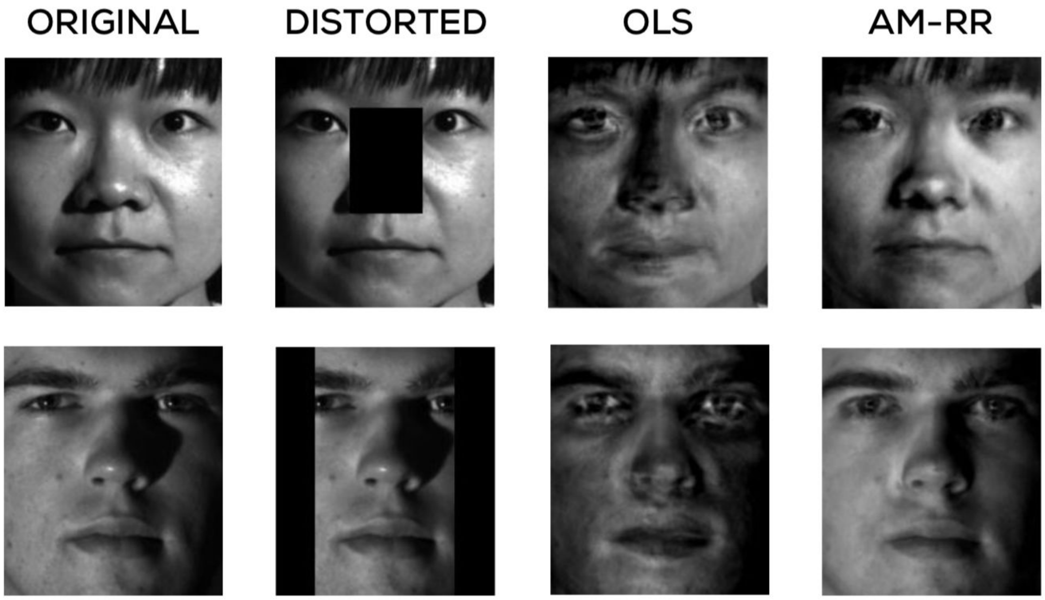


Figure 9.3: An experiment on face reconstruction using robust regression techniques. Two face images were taken and different occlusions were applied to them. Using the model described in § 9.1, reconstruction was attempted using both, ordinary least squares (OLS) and robust regression (AM-RR). It is clear that AM-RR achieves far superior reconstruction of the images and is able to correctly figure out the locations of the occlusions. Images courtesy the Yale Face Database B.

Face Recognition : In biometrics, a fundamental problem is to identify if a new face image belongs to that of a registered individual or not.

Cast as a regression problem by trying to fit various features of the new image to corresponding features of existing images of the individual in the registered database

Assume that images are represented as *n*-dimensional feature vectors say, using simple pixel-based features. Also assume that there already exist *p* images of the person in the database.

Represent the new image $\mathbf{x}^t \in R^n$ in terms of the database images $X = [\mathbf{x}_1, . . . , \mathbf{x}_p] \in R^{n\times p}$ of that person. One solution is to perform *linear* interpolation:

$$\min_{\mathbf{w}\in\mathbb{R}^p} \left\| \mathbf{x}^t - X\mathbf{w} \right\|_2^2 = \sum_{i=1}^{n} (\mathbf{x}_i^t - X^i\mathbf{w})^2.$$

If the person is genuine, then there will exist a combination $\mathbf{w}^*$ such that for all i, we have $\mathbf{x}_i^t \approx X^i\mathbf{w}^*$ i.e., all features can be faithfully reconstructed. Problematic, if however the new image $\mathbf{x}^t$ has occlusions or is otherwise corrupted $\mathbf{x}_i^t = X^i\mathbf{w}^* + \mathbf{b}_i^*$ where $\mathbf{b}_i^*$ = 0 on uncorrupted pixels but can take large and unpredictable values for corrupted pixels .

Nevertheless, one can compute the *least squares fit* in the presence of such corruptions
The *challenge* is to do this without effort to identify the locations of the corrupted pixels.

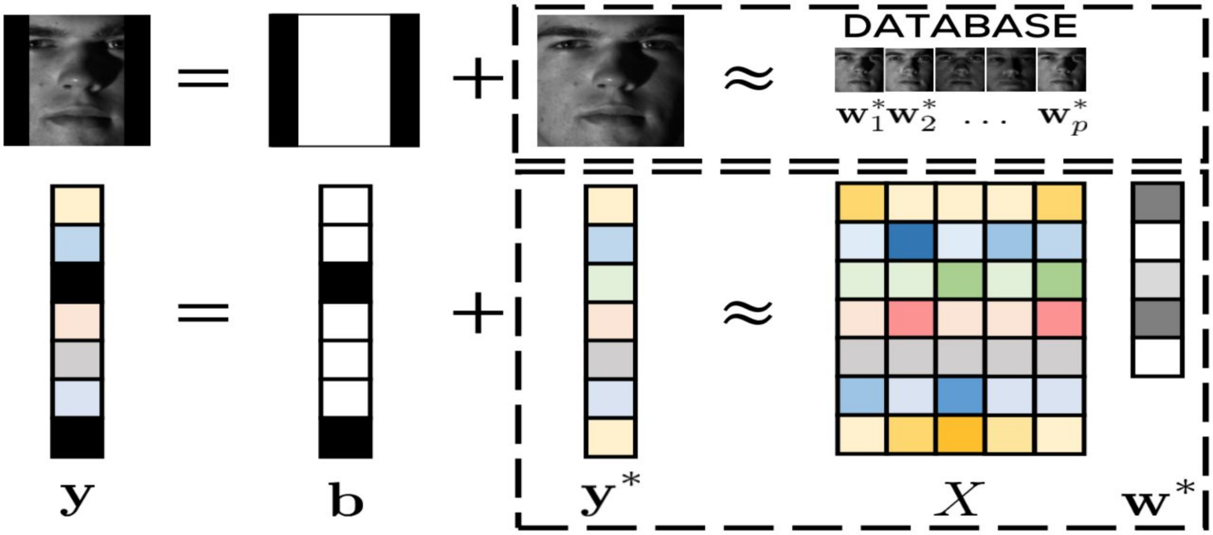


Figure 9.1: A corrupted image $\mathbf{y}$ can be interpreted as a combination of a clean image $\mathbf{y}^*$ and a corruption mask $\mathbf{b}^*$ i.e., $\mathbf{y} = \mathbf{y}^* + \mathbf{b}^*$. The mask encodes the locations of the corrupted pixels as well as the values of the corruptions. The clean image can be (approximately) recovered as an affine combination of existing images in a database as $\mathbf{y}^* \approx X\mathbf{w}^*$. Face reconstruction and recognition in such a scenario constitutes a robust regression problem. Note that the corruption mask $\mathbf{b}^*$ is sparse since only a few pixels are corrupted. Images courtesy the Yale Face Database B.

ROBUST REGRESSION

Goal is to take a set of n (possibly) corrupted data points $(\mathbf{x}_i,y_i)^n_{i=1}$ and recover the underlying parameter vector $\mathbf{w}^*$,

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^p,\mathbf{b}\in\mathbb{R}^n\\ \|\mathbf{b}\|_0\leq k}} \|\mathbf{y}-X\mathbf{w}-\mathbf{b}\|_2^2\,,$$

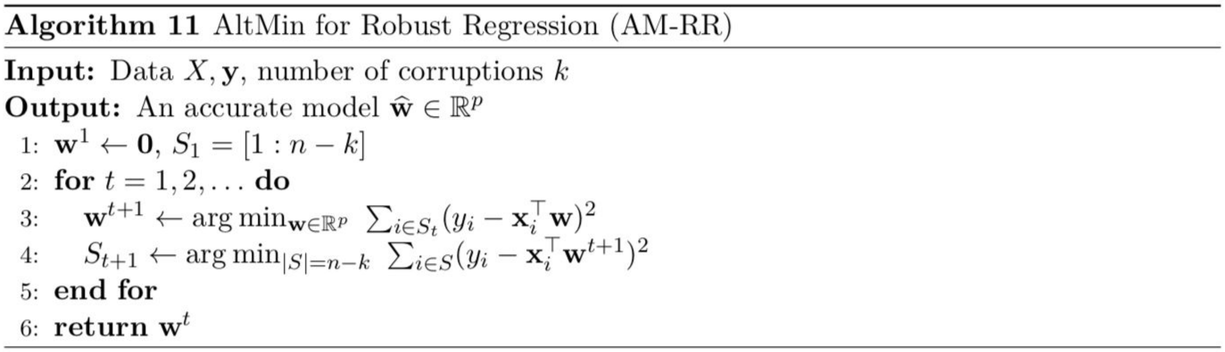
The variables $\mathbf{b}^*_i$ can be unbounded in magnitude and of arbitrary sign. However, we assume that only a few data points are corrupted i.e., the vector $\mathbf{b}^* = [\mathbf{b}^*_1, \mathbf{b}^*_2, \ldots, \mathbf{b}^*_n]$ is sparse $// \mathbf{b}^* //_0 \leq k$ (for as large a k as possible).

Note, it is impossible to recover the model $\mathbf{w}^*$ *if more than half the points are corrupted i.e.,* $k \geq n/2$.

It can be seen that $\mathbf{w}^*$ and supp($\mathbf{b}^*$) =: S_* i.e., *the true model* and *the locations of the uncorrupted points*, are the *two* most crucial elements since given one, finding the other is very simple.

Indeed, if someone were to magically hand us $\mathbf{w}^*$, it is trivial to identify S_* by simply identifying data points where $y_i = \mathbf{x}_i^\top \mathbf{w}_*$. On the other hand, given S_*, it is simple to obtain $\mathbf{w}^*$ by simply solving a least squares regression problem on the set of data points in the set S∗ namely,

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^p\\ |S|=n-k}} \left\|\mathbf{y}_S - X^S\mathbf{w}\right\|_2^2$$



Algorithm 11 AltMin for Robust Regression (AM-RR)

Input: Data $X, \mathbf{y}$, number of corruptions k

Output: An accurate model $\widehat{\mathbf{w}} \in \mathbb{R}^p$

```
$\mathbf{w}^1 \leftarrow \mathbf{0}$, $S_1 = [1 : n-k]$
for $t = 1, 2, \ldots$ do
    $\mathbf{w}^{t+1} \leftarrow \arg\min_{\mathbf{w} \in \mathbb{R}^p} \sum_{i \in S_t} (y_i - \mathbf{x}_i^\top \mathbf{w})^2$
    $S_{t+1} \leftarrow \arg\min_{|S|=n-k} \sum_{i \in S} (y_i - \mathbf{x}_i^\top \mathbf{w}^{t+1})^2$
end for
return $\mathbf{w}^t$
```

(0) AM-RR maintains a *model estimate* **w**^{t} and an *active set* $S_t \subset [n]$ of points that are deemed clean at the moment.

(1) Initially the *active set* S_1 is taken to be the first n-k points

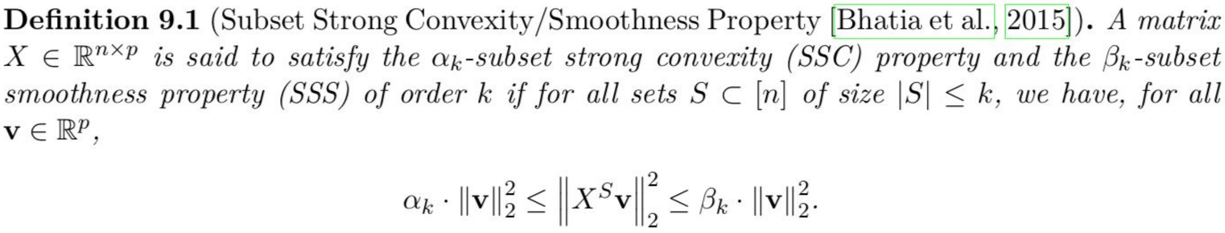
(2) At every time step t,

(3) AM-RR first fixes the active set S_t and updates the model **w**^{t} *(via least squares over active set)*,

and then

(4) fixes the model **w**$^{t+1}$ and updates the active set S_{t+1} (*by taking the n − k data points of S with the smallest residuals (by magnitude) with respect to the updated model and designating them to be the active set*)

Robust Recovery Guarantee for AM-RR

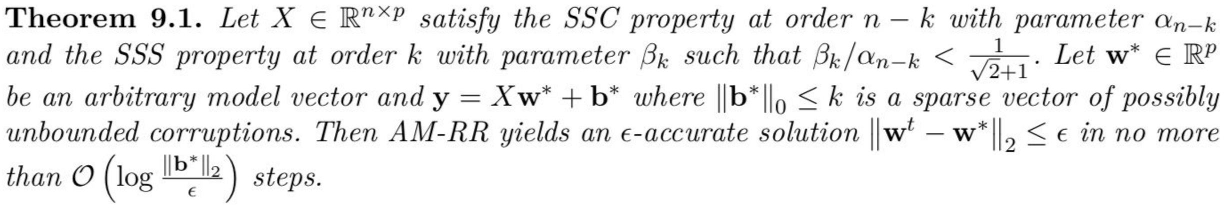


Definition 9.1 (Subset Strong Convexity/Smoothness Property [Bhatia et al., 2015]). *A matrix $X \in \mathbb{R}^{n \times p}$ is said to satisfy the α_k-subset strong convexity (SSC) property and the β_k-subset smoothness property (SSS) of order k if for all sets $S \subset [n]$ of size $|S| \leq k$, we have, for all $\mathbf{v} \in \mathbb{R}^p$,*

$$\alpha_k \cdot \|\mathbf{v}\|_2^2 \leq \left\| X^S \mathbf{v} \right\|_2^2 \leq \beta_k \cdot \|\mathbf{v}\|_2^2.$$

The SSC/SSS properties require that the *design matrix* X formed by taking any subset of *k* pixels from the data set of *n* pixels act as an approximate isometry on all *p* dimensional points.

[These properties are related to the traditional RSC/RSS properties and it can be shown that RIP-inducing distributions over matrices also produce matrices that satisfy the SSC/SSS properties, with high probability.]

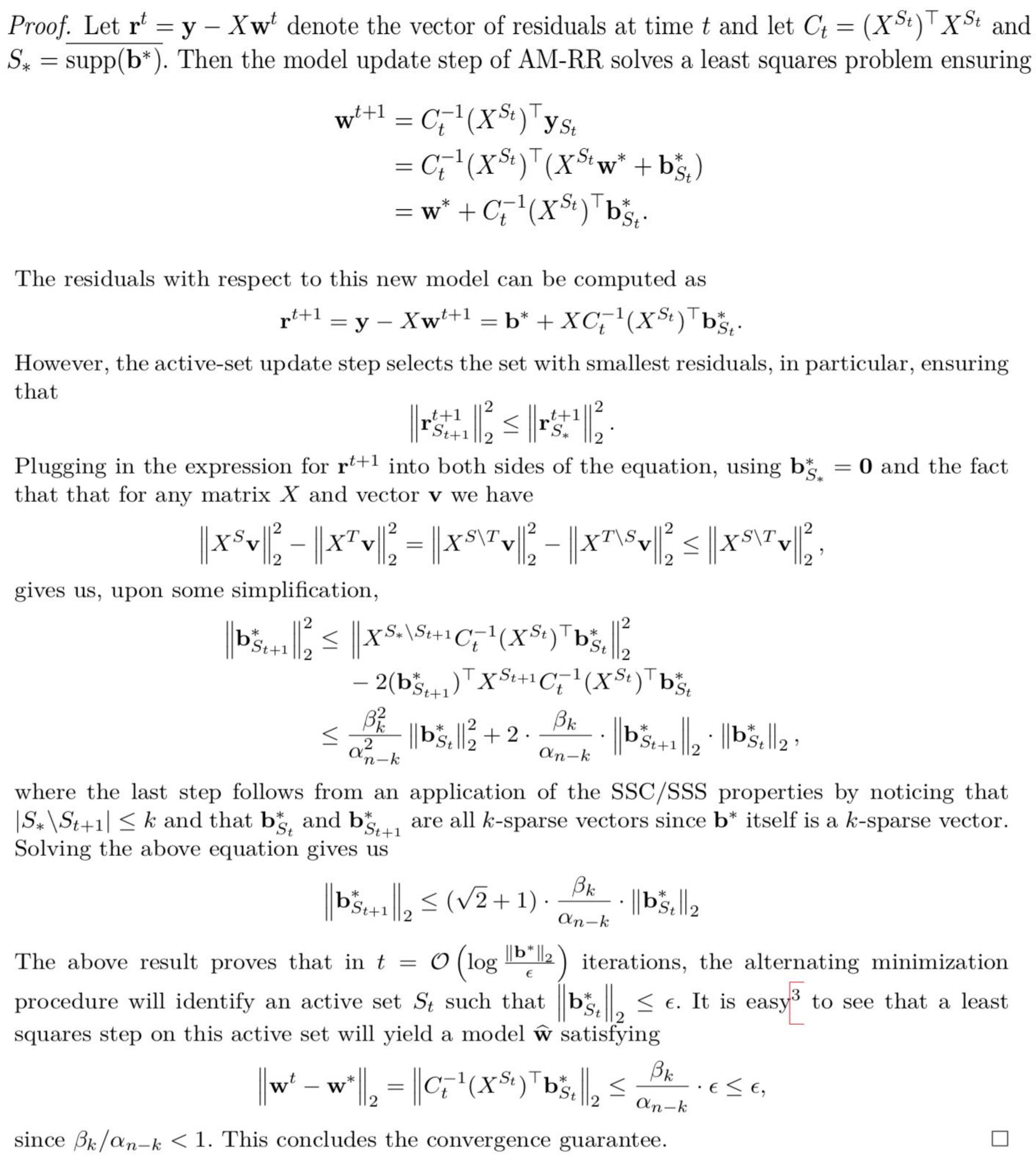


Theorem 9.1. *Let $X \in \mathbb{R}^{n \times p}$ satisfy the SSC property at order $n-k$ with parameter α_{n-k} and the SSS property at order k with parameter β_k such that $\beta_k / \alpha_{n-k} < \frac{1}{\sqrt{2}+1}$. Let $\mathbf{w}^* \in \mathbb{R}^p$ be an arbitrary model vector and $\mathbf{y} = X\mathbf{w}^* + \mathbf{b}^*$ where $\|\mathbf{b}^*\|_0 \leq k$ is a sparse vector of possibly unbounded corruptions. Then AM-RR yields an ϵ-accurate solution $\left\|\mathbf{w}^t - \mathbf{w}^*\right\|_2 \leq \epsilon$ in no more than $\mathcal{O}\left(\log \frac{\|\mathbf{b}^*\|_2}{\epsilon}\right)$ steps.*

Intuitive Proof **:**

Since the algorithm uses only a subset of data points to estimate the model vector, it is essential that smaller subsets of data points of size n − k (in particular the true subset of clean points S_*) also allow the model to be recovered.

*This is equivalent to requiring that the **design matrices** formed by smaller subsets of data points not identify distinct model vectors. This is exactly what the SSC property demands.*



Proof. Let $\mathbf{r}^t = \mathbf{y} - X\mathbf{w}^t$ denote the vector of residuals at time t and let $C_t = (X^{S_t})^\top X^{S_t}$ and $S_* = \text{supp}(\mathbf{b}^*)$. Then the model update step of AM-RR solves a least squares problem ensuring

$$\begin{aligned}\mathbf{w}^{t+1} &= C_t^{-1}(X^{S_t})^\top \mathbf{y}_{S_t} \\ &= C_t^{-1}(X^{S_t})^\top (X^{S_t}\mathbf{w}^* + \mathbf{b}^*_{S_t}) \\ &= \mathbf{w}^* + C_t^{-1}(X^{S_t})^\top \mathbf{b}^*_{S_t}.\end{aligned}$$

The residuals with respect to this new model can be computed as

$$\mathbf{r}^{t+1} = \mathbf{y} - X\mathbf{w}^{t+1} = \mathbf{b}^* + XC_t^{-1}(X^{S_t})^\top \mathbf{b}^*_{S_t}.$$

However, the active-set update step selects the set with smallest residuals, in particular, ensuring that

$$\left\|\mathbf{r}^{t+1}_{S_{t+1}}\right\|_2^2 \leq \left\|\mathbf{r}^{t+1}_{S_*}\right\|_2^2.$$

Plugging in the expression for $\mathbf{r}^{t+1}$ into both sides of the equation, using $\mathbf{b}^*_{\overline{S_*}} = \mathbf{0}$ and the fact that that for any matrix X and vector $\mathbf{v}$ we have

$$\left\|X^S\mathbf{v}\right\|_2^2 - \left\|X^T\mathbf{v}\right\|_2^2 = \left\|X^{S\backslash T}\mathbf{v}\right\|_2^2 - \left\|X^{T\backslash S}\mathbf{v}\right\|_2^2 \leq \left\|X^{S\backslash T}\mathbf{v}\right\|_2^2,$$

gives us, upon some simplification,

$$\begin{aligned}\left\|\mathbf{b}^*_{S_{t+1}}\right\|_2^2 \leq{} & \left\|X^{S_*\backslash S_{t+1}} C_t^{-1}(X^{S_t})^\top \mathbf{b}^*_{S_t}\right\|_2^2 \\ & - 2(\mathbf{b}^*_{S_{t+1}})^\top X^{S_{t+1}} C_t^{-1}(X^{S_t})^\top \mathbf{b}^*_{S_t} \\ \leq{} & \frac{\beta_k^2}{\alpha_{n-k}^2}\left\|\mathbf{b}^*_{S_t}\right\|_2^2 + 2 \cdot \frac{\beta_k}{\alpha_{n-k}} \cdot \left\|\mathbf{b}^*_{S_{t+1}}\right\|_2 \cdot \left\|\mathbf{b}^*_{S_t}\right\|_2,\end{aligned}$$

where the last step follows from an application of the SSC/SSS properties by noticing that $|S_*\backslash S_{t+1}| \leq k$ and that $\mathbf{b}^*_{S_t}$ and $\mathbf{b}^*_{S_{t+1}}$ are all k-sparse vectors since $\mathbf{b}^*$ itself is a k-sparse vector. Solving the above equation gives us

$$\left\|\mathbf{b}^*_{S_{t+1}}\right\|_2 \leq (\sqrt{2}+1) \cdot \frac{\beta_k}{\alpha_{n-k}} \cdot \left\|\mathbf{b}^*_{S_t}\right\|_2$$

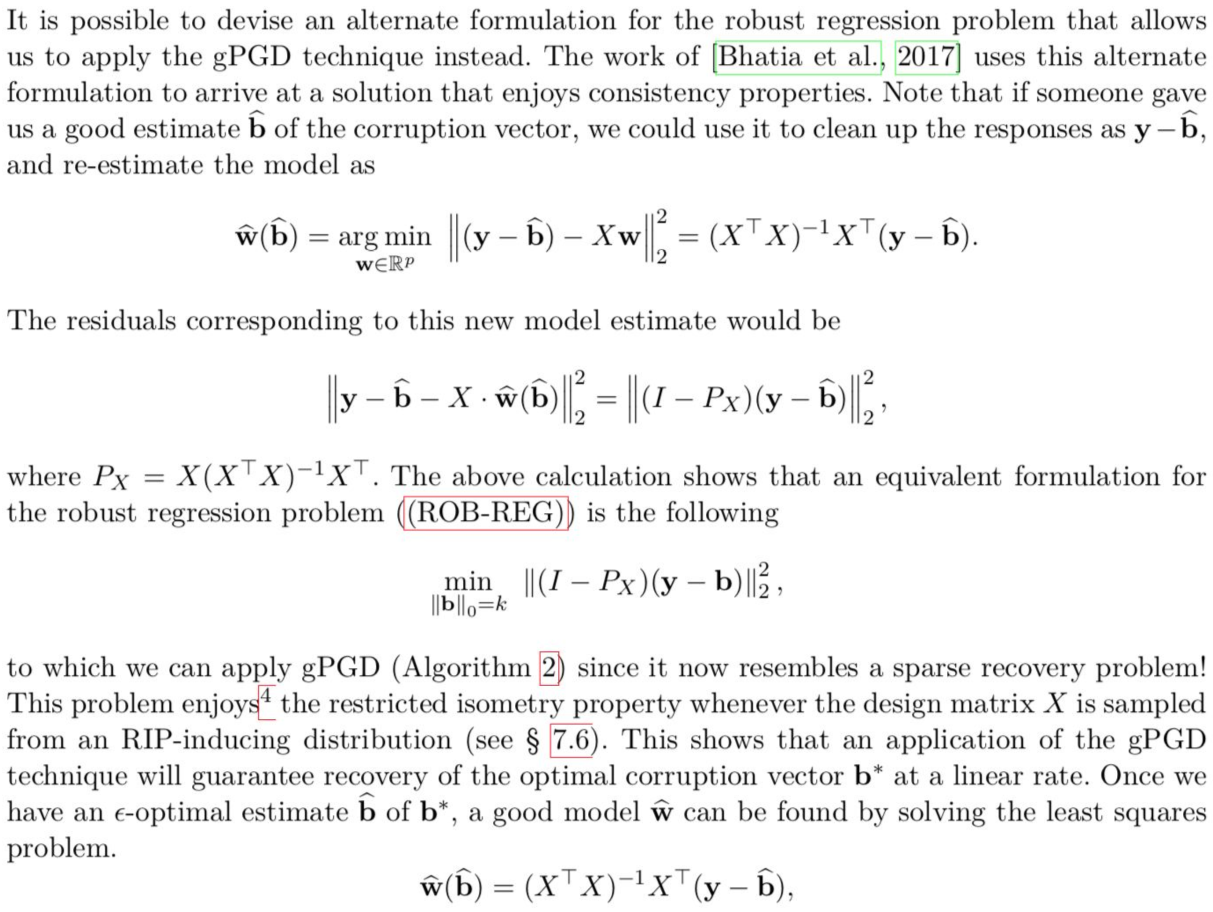
The above result proves that in $t = \mathcal{O}\left(\log \frac{\|\mathbf{b}^*\|_2}{\epsilon}\right)$ iterations, the alternating minimization procedure will identify an active set S_t such that $\left\|\mathbf{b}^*_{S_t}\right\|_2 \leq \epsilon$. It is easy[3] to see that a least squares step on this active set will yield a model $\widehat{\mathbf{w}}$ satisfying

$$\left\|\mathbf{w}^t - \mathbf{w}^*\right\|_2 = \left\|C_t^{-1}(X^{S_t})^\top \mathbf{b}^*_{S_t}\right\|_2 \leq \frac{\beta_k}{\alpha_{n-k}} \cdot \epsilon \leq \epsilon,$$

since $\beta_k/\alpha_{n-k} < 1$. This concludes the convergence guarantee. □

The crucial assumption in the previous result is the requirement $\beta_k/\alpha_{n-k} < \frac{1}{\sqrt{2}+1}$. Clearly, as $k \to 0$, we have $\beta_k \to 0$ but if the matrix X is well conditioned we still have $\alpha_{n-k} > 0$. Thus, for small enough k, it is assured that we will have $\beta_k/\alpha_{n-k} < \frac{1}{\sqrt{2}+1}$. The point at which this occurs is the so-called *breakdown point* of the algorithm – it is the largest number k such that the algorithm can tolerate k possibly adversarial corruptions and yet guarantee recovery.

Robust Regression via Projected Gradient Descent



It is possible to devise an alternate formulation for the robust regression problem that allows us to apply the gPGD technique instead. The work of [Bhatia et al., 2017] uses this alternate formulation to arrive at a solution that enjoys consistency properties. Note that if someone gave us a good estimate $\widehat{\mathbf{b}}$ of the corruption vector, we could use it to clean up the responses as $\mathbf{y}-\widehat{\mathbf{b}}$, and re-estimate the model as

$$\widehat{\mathbf{w}}(\widehat{\mathbf{b}}) = \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min}\ \left\|(\mathbf{y}-\widehat{\mathbf{b}}) - X\mathbf{w}\right\|_2^2 = (X^\top X)^{-1}X^\top(\mathbf{y}-\widehat{\mathbf{b}}).$$

The residuals corresponding to this new model estimate would be

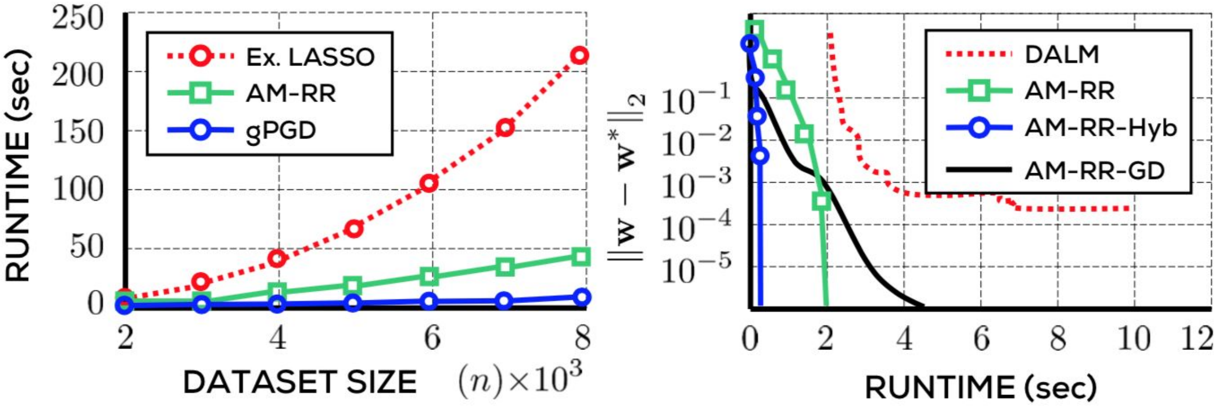
$$\left\|\mathbf{y}-\widehat{\mathbf{b}} - X\cdot\widehat{\mathbf{w}}(\widehat{\mathbf{b}})\right\|_2^2 = \left\|(I-P_X)(\mathbf{y}-\widehat{\mathbf{b}})\right\|_2^2,$$

where $P_X = X(X^\top X)^{-1}X^\top$. The above calculation shows that an equivalent formulation for the robust regression problem ((ROB-REG)) is the following

$$\min_{\|\mathbf{b}\|_0=k}\ \|(I-P_X)(\mathbf{y}-\mathbf{b})\|_2^2,$$

to which we can apply gPGD (Algorithm 2) since it now resembles a sparse recovery problem! This problem enjoys[4] the restricted isometry property whenever the design matrix X is sampled from an RIP-inducing distribution (see § 7.6). This shows that an application of the gPGD technique will guarantee recovery of the optimal corruption vector $\mathbf{b}^*$ at a linear rate. Once we have an ϵ-optimal estimate $\widehat{\mathbf{b}}$ of $\mathbf{b}^*$, a good model $\widehat{\mathbf{w}}$ can be found by solving the least squares problem.

$$\widehat{\mathbf{w}}(\widehat{\mathbf{b}}) = (X^\top X)^{-1}X^\top(\mathbf{y}-\widehat{\mathbf{b}}),$$



(a) Run-time Comparison

(b) Convergence Comparison

compares various solvers on a robust regression problem in $p = 300$ dimensions with 1800 data points of which 40% are corrupted. The solvers include the gAM-style solver AM-RR, a variant using gradient-based updates, a hybrid method (see § 9.5), and the DALM method [Yang et al., 2013], a state-of-the-art solver for relaxed LASSO-style formulations. The hybrid method is the fastest of all the techniques. In general, all AM-RR variants are much faster than the relaxation-based method.

Allen Y. Yang, Zihan Zhou, Arvind Ganesh Balasubramanian, S Shankar Sastry, and Yi Ma. Fast ℓ_1-Minimization Algorithms for Robust Face Recognition. *IEEE Transactions on Image Processing*, 22(8):3234–3246, 2013.

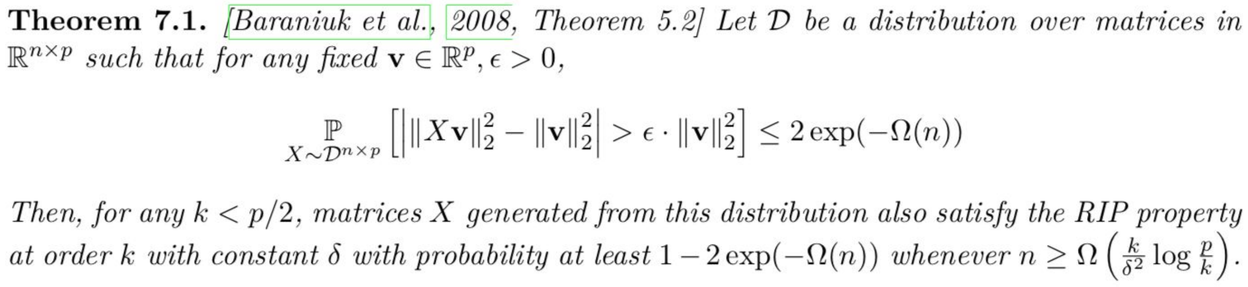
[DALM]

Nam H Nguyen and Trac D Tran. Robust Lasso With Missing and Grossly Corrupted Observations. *IEEE Transaction on Information Theory*, 59(4):2036–2058, 2013b.

[Extended Lasso]

Ensuring RIP and other Properties

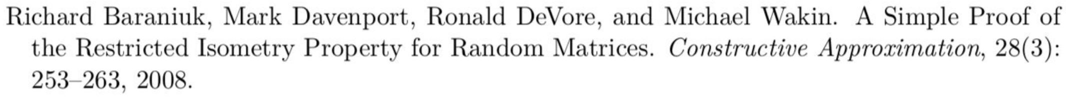
Random Designs: The simplest of these results are the so-called *random* design constructions which guarantee that if the matrix is sampled from certain well behaved distributions, then it will satisfy the RIP property with high probability. For instance, the work of Baraniuk et al. [2008] shows the following result:



Theorem 7.1. *[Baraniuk et al., 2008, Theorem 5.2] Let $\mathcal{D}$ be a distribution over matrices in $\mathbb{R}^{n\times p}$ such that for any fixed $\mathbf{v} \in \mathbb{R}^p, \epsilon > 0$,*

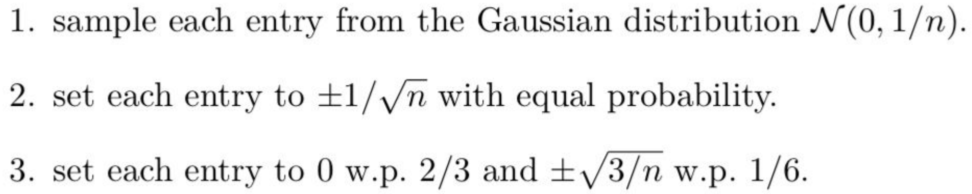
$$\underset{X\sim\mathcal{D}^{n\times p}}{\mathbb{P}}\left[\left|\|X\mathbf{v}\|_2^2 - \|\mathbf{v}\|_2^2\right| > \epsilon \cdot \|\mathbf{v}\|_2^2\right] \leq 2\exp(-\Omega(n))$$

Then, for any $k < p/2$, matrices X generated from this distribution also satisfy the RIP property at order k with constant δ with probability at least $1 - 2\exp(-\Omega(n))$ whenever $n \geq \Omega\left(\frac{k}{\delta^2}\log\frac{p}{k}\right)$.



Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A Simple Proof of the Restricted Isometry Property for Random Matrices. *Constructive Approximation*, 28(3): 253–263, 2008.

Thus, a distribution over matrices that, for every *fixed* vector, acts as an almost isometry with high probability, is also guaranteed to, with very high probability, generate matrices that act as a restricted isometry *simultaneously* over all sparse vectors. Such matrix distributions are easy to construct – one simply needs to sample each entry of the matrix independently according to one of the following distributions:



1. sample each entry from the Gaussian distribution $\mathcal{N}(0, 1/n)$.
2. set each entry to $\pm 1/\sqrt{n}$ with equal probability.
3. set each entry to 0 w.p. 2/3 and $\pm\sqrt{3/n}$ w.p. 1/6.