CS 2243 Fall 2024 Scribes: <names>
09/23 Based on notes by: Kevin Du

Lecture 6: SoS and robust regression

1 Sum-Of-Squares Introduction

The sum-of-squares algorithm (SoS) is a generic framework that can be applied to a wide variety of nonconvex optimization problems, similar to linear programming for linear applications. This SoS framework can convert an inefficient algorithm with a "simple" proof of correctness, i.e. a proof involving a restricted set of convex axioms, into an efficient algorithm with the same guarantees.

SoS is used in a wide variety of statistical algorithms and gives an elegant framework for robust estimation in the presence of corrupted data.

2 Robust Regression Introduction

As an example use case of the SoS algorithm, consider a robust regression setting in which we aim to perform linear regression on arbitrarily corrupted data. Intuitively, we think of an adversary who has access to our dataset and alters a fraction η of the data to make our regression algorithm fail.

Specifically, we are given a corrupted dataset $(x_i, y_i)_{i=1}^N$ with corruption fraction η. The structure of this data is given as follows. The explanatory data is given by

$$\begin{aligned} &(x_i^*, y_i^*, a_i^*)_{i=1}^N \\ \text{with } &y_i^* = \langle w, x_i^* \rangle + \zeta_i \\ &||x_i|| \leq 1 \\ &||w|| \leq R \\ &\zeta_i \sim \mathcal{N}(0, \sigma^2) \\ &a_i^* \in \{0, 1\} \\ &\sum a_i^* \geq N(1-\eta) \\ &a_i^* = 1 \implies (x_i, y_i) = (x_i^*, y_i^*). \end{aligned}$$

We only observe the dataset $(x_i, y_i)_{i=1}^N$ We aim to find

$$\arg\min_w \frac{1}{N} \sum a_i^* (y_i - \langle w, x_i \rangle)^2.$$

Here a_i is an indicator variable with $a_i^* = 1$ if the data point is clean and $a_i^* = 0$ if it is corrupted. The expression we want to minimize then is the clean MSE, i.e. the MSE over the uncorrupted data.

2.1 Initial Approaches

When analyzing different robust regression approaches, we will compare their performance against the optimal baseline

$$OPT = \frac{1}{N} \sum a_i^* (y_i - \langle w^*, x_i \rangle)^2$$

which is the clean MSE obtained if we know which points are corrupted, i.e. if the a_i variables were all given to us. The aim is to find an algorithm which has a clean MSE that is close to "OPT" for small values of η.

If we simply run oridinary least squares regression on the given dataset, the clean MSE we achieve is "OPT" + $O(\eta R^2)$. This gives a naive upper bound on the achievable clean MSE. A lower bound on the clean MSE is given by the information-theoretic bound "OPT" + $O(\eta^2\sigma^2)$.

One may think to use regularization or choose a less sensitive loss function in order to minimize the effect of outliers caused by corrupted data. For instance, statistician Peter J. Huber proposed using the Huber loss function, which grows linearly with respect to the magnitude of the error at the tails and quadratically near the origin.

However, Chen, Koehler, Moitra, and Yau recently showed that

Theorem 1. *Any algorithm based on minimizing a convex loss gets clean MSE $\eta^3 R$ [**?**].*

Thus, these approaches can not be made to achieve the information-theoretic lower bound.

Note that any algorithm can be made to perform poorly when $\eta \geq 1/2$. This is because if there are more corrupted data points than clean data points, an adversary could change many of the data points to match some other parameter w'. Then, no algorithm could tell the clean dataset apart from the fictitious ones generated by w'. However, the focus moving forward will be on cases with small values of η.

3 SoS for Robust Regression

The SoS framework gives an algorithm for robust regression which achieves a clean MSE of $(1+O(C\eta^{1/2}))(OPT + O(C\eta^{1/2}\sigma^2)$ and performs well in practice for distributions with low hypercontractivity constant. To achieve this performance, we first give an inefficient algorithm that finds a low-MSE solution.

We model the problem as an optimization problem with a system of polynomial constraints.

Variables
1. w - estimate of w^* 2. $a_1, \ldots, a_N$
Constraints
1. $a_i^2 = a_i$ for all i 2. $\frac{1}{N}\sum a_i \geq 1 - \eta$
Objective
$\min_w \frac{1}{N} a_i (y_i - \langle w, x_i \rangle)^2$

Table 1: A polynomial optimization problem which models robust linear regression

Solving the regression problem then reduces to solving the polynomial optimization problem above. Unfortunately, polynomial optimization problems are nonconvex and in general NP-hard. However, we can define a convex relaxation of the problem which becomes tractable through the SoS framework.

3.1 Pseudodistributions

Instead of optimizing over a_i and w we optimize over distributions of this quantities. More specifically, we optimize over pseudo-distributions, objects which behave like distributions in that we can take pseudo-expectations of these quantities.

Definition 1. *A degree-t pseudo-distribution is given by the pseudo-expectation operator $\widetilde{\mathbb{E}}$ which takes as input a polynomial in the variables $a_1, \ldots, a_N, w_1, \ldots, w_d$ with degree $\leq d$ and outputs a number. This operator must satisfy:*

Prop. 1. $\widetilde{\mathbb{E}}[1] = 1$

Prop. 2. $\widetilde{\mathbb{E}}[\alpha p + \beta q] = \alpha\,\widetilde{\mathbb{E}}[p] + \beta\,\widetilde{\mathbb{E}}[q]$

Prop. 3. $\widetilde{\mathbb{E}}[p^2] \geq 0$ *for all polynomials* p *with degree* $\leq t/2$

Note that the space of pseudo-distributions is a finite-dimensional object spanned by pseudo-expectations of monomials such as $\widetilde{\mathbb{E}}[a_3^5 a_2^6 w_5]$. Furthermore, the set of pseudo-expectations is convex due to property 3 in Definition 3.1. This concept of the pseudo-distribution will allow us to convert the robust regression problem into a convex optimization problem.

3.2 Including Program Constraints

In order to include the constraints of our polynomial optimization problem, we add another property that $\widetilde{\mathbb{E}}$ must satisfy:

Prop. 4. If there is a "simple", i.e. degree-t, proof that $p \geq 0$ using the problem constraints, then $\widetilde{\mathbb{E}}[p] \geq 0$.

Here, a degree-t proof of a statement is one which uses only polynomials of degree $\leq t$ and the fact that sums of squares are nonnegative, i.e. $\sum x_i^2 \geq 0$. More precisely, a proof is a chain of inequalities with each one derived from the preceding ones by the following derivation:

$$p(x) \geq 0, q(x) = 0 \implies SOS_1(x) + p(x) \cdot SOS_2(x) + q(x) \cdot r(x) \geq 0$$

for any SOS_1, SOS_2 which are sums of squares of polynomials. Note that this property stipulates a convex constraint on $\widetilde{\mathbb{E}}$. Thus, we aim to find an instance of $\mathbb{E}$ that satisfies these convex constraints. Note that a solution always exist as the true objective is feasible under the SoS program. Thus, we can use various convex optimization algorithms such as the ellipsoid method to find a solution instance.

3.3 Examples

We examine a couple examples of conditions we can derive under the SoS framework.

Example 1. *From the constraint* $a_i^2 = a_i$*, there is a simple proof that* $0 \leq a_i \leq 1$.

Proof. Note that $a_i = a_i^2 \geq 0$. Also, we have $(1 - a_i)^2 = 1 - a_i - (a_i^2 - a_i) \geq 0 \implies 1 - a_i \geq 0 \implies a_i \leq 1$. □

Note that we can not prove that $a_i \in \{0, 1\}$ in the SoS system. However, since we've shown that $0 \leq a_i \leq 1$, we know that the SoS program must have $0 \leq \widetilde{\mathbb{E}}[a_i] \leq 1$ for any degree-2 pseudo-expectation.

Example 2 (Cauchy-Schwartz)**.** *If* u_i *and* v_i *are variables, then there is a simple proof that*

$$(\sum_i u_i v_i)^2 \leq (\sum_i u_i^2)(\sum_i v_i^2).$$

Proof. We have

$$(\sum_i u_i^2)(\sum_i v_i^2) - (\sum_i u_i v_i)^2 = \sum_{i,j} (u_i v_j - u_j v_i)^2 \geq 0$$

□

3.4 Reading Off the Objective

Note that the objective is to find parameters w which minimize the clean MSE. Note that by the convexity of $\widetilde{\mathbb{E}}$, we have

Lemma 1. $\sum_i a_i^* (y_i - \langle \widetilde{\mathbb{E}}[w], x_i \rangle)^2 \leq \sum_i a_i^* \widetilde{\mathbb{E}}[(y_i - \langle w, x_i \rangle)^2]$

If we can bound the right hand side of the above equation, we know that $\widetilde{\mathbb{E}}[w]$ achieves a low clean MSE.

3.5 Bounding Clean MSE of the SoS Output

Suppose $\widetilde{\mathbb{E}}$ is a solution to the convex constraints described above. Then, we aim to bound the clean MSE of the solution vector $\widetilde{\mathbb{E}}[w]$. As seen above, we have

$$\frac{1}{N}\sum_i a_i^*(y_i - \langle\widetilde{\mathbb{E}}[w], x_i\rangle)^2 \le \frac{1}{N}\sum_i a_i^*\,\widetilde{\mathbb{E}}[(y_i - \langle w, x_i\rangle)^2] \le \frac{1}{N}\sum_i \widetilde{\mathbb{E}}[(y_i - \langle w, x_i\rangle)^2] \qquad (*)$$

To bound this quantity, we decompose the cases under the following substitution

$$1 = a_i a_i^* + a_i(1 - a_i^*) + (1 - a_i).$$

Here, $a_i a_i^*$ represents cases where we correctly identified the clean data as clean. $a_i(1-a_i^*)$ are cases where we incorrectly identified corrupted data as clean. $1 - a_i$ are cases where we identified data as corrupted. Then, we can write (*) as

$$\begin{aligned}
(*) &= \frac{1}{N}\sum_i a_i a_i^*(y_i^* - \langle w, x_i^*\rangle)^2 && (1)\\
&+ \frac{1}{N}\sum_i a_i(1-a_i^*)(y_i^* - \langle w, x_i^*\rangle)^2 && (2)\\
&+ \frac{1}{N}\sum_i (1-a_i)(y_i^* - \langle w, x_i^*\rangle)^2 && (3)
\end{aligned}$$

We can bound each of these terms as follows. First,

$$\begin{aligned}
(1) &= \frac{1}{N}\sum_i a_i a_i^*(y_i - \langle w, x_i\rangle)^2\\
&\le \frac{1}{N}\sum_i a_i(y_i - \langle w, x_i\rangle)^2\\
&\le \frac{1}{N}\sum_i a_i^*(y_i^* - \langle w^*, x_i^*\rangle)^2 \le OPT
\end{aligned}$$

To bound the second term, we use the Cauchy Schwarz bound to get

$$\begin{aligned}
(2) &= \frac{1}{N}\sum_i a_i(1-a_i^*)(y_i^* - \langle w, x_i^*\rangle)^2\\
&\le \left(\frac{1}{N}\sum_i (1-a_i^*)^2\right)^{1/2} \cdot \left(\frac{1}{N}\sum_i a_i^2(y_i^* - \langle w, x_i^*\rangle)^4\right)^{1/2}\\
&\le \eta^{1/2}\left(\frac{1}{N}\sum_i (y_i^* - \langle w, x_i^*\rangle)^4\right)^{1/2}
\end{aligned}$$

Similarly, we get

$$
\begin{aligned}
(3) &= \frac{1}{N}\sum_i (1-a_i)(y_i^* - \langle w, x_i^*\rangle)^2 \\
&\le \left(\frac{1}{N}\sum_i (1-a_i)^2\right)^{1/2} \cdot \left(\frac{1}{N}\sum_i (y_i^* - \langle w, x_i^*\rangle)^4\right)^{1/2} \\
&\le \eta^{1/2}\left(\frac{1}{N}\sum_i (y_i^* - \langle w, x_i^*\rangle)^4\right)^{1/2}
\end{aligned}
$$

To bound the quantity $\frac{1}{N}\sum_i (y_i^* - \langle w, x_i^*\rangle)^4$, recall that $y_i^* = \langle w^*, x_i^*\rangle + \zeta_i$ where $\zeta_i \sim \mathcal{N}(0, \sigma^2)$. Then, we have

$$
\frac{1}{N}\sum_i (y_i^* - \langle w, x_i^*\rangle)^4 = \frac{1}{N}\sum_i (\langle w^* - w, x_i^*\rangle + \zeta_i)^4.
$$

By the elementary inequality $(a+b)^4 \le 8(a^4 + b^4)$, we can bound the above expression by

$$
\frac{8}{N}\sum_i \langle w^* - w, x_i^*\rangle^4 + \frac{8}{N}\sum_i \zeta_i^4.
$$

Note that in expectation $8\,\mathbb{E}[\zeta_i^4] = 24\sigma^4$. To summarize, we have shown that

$$
\begin{aligned}
(*) &= (1) + (2) + (3) \\
&\le OPT + 2\eta^{1/2}\left(\frac{1}{N}\sum_i (y_i^* - \langle w, x_i^*\rangle)^4\right)^{1/2} \\
&\le OPT + 2\eta^{1/2}\left(\frac{8}{N}\sum_i \langle w^* - w, x_i^*\rangle^4 + O(\sigma^4)\right)^{1/2} \\
&\le OPT + O(\eta^{1/2})\left[\left(\frac{1}{N}\sum_i \langle w^* - w, x_i^*\rangle^4\right)^{1/2} + \sigma^2\right]
\end{aligned}
$$

Note that that last inequality is true when σ^4 and $\sum_i \langle w^* - w, x_i^*\rangle^4$ are roughly proportional. This is usually the case practically so we will assume this is true going forward. However, we note that the rigorous inequality is in fact

$$
((*) - OPT)^2 \le O(\eta)\left[\frac{1}{N}\sum_i \langle w^* - w, x_i^*\rangle^4 + \sigma^4\right]
$$

Finally, to bound the value $\frac{1}{N}\sum_i \langle w^* - w, x_i^*\rangle^4$, we will need an assumption on the distribution.

Definition 2. *A distribution q is 4-hypercontractive if*

$$
\mathbb{E}_{x\sim q}[\langle v, x\rangle^4] \le (C \cdot \mathbb{E}_{x\sim q}[\langle v, x\rangle^2])^2 \tag{**}
$$

for all $v \in \mathbb{R}^d$ for some $C = O(1)$. q is certifiably 4-hypercontractive if (**) has an SoS proof.

For instance, any rotation of a product distribution (e.g. $\mathcal{N}(\mu, \Sigma)$) is certifiably 4-hypercontractive.
In our case, we can assume that the distribution of x is 4-hypercontractive meaning

$$\begin{aligned}\left(\frac{1}{N}\sum_i \langle w^* - w, x_i\rangle^4\right)^{1/2} &\le \frac{C}{N}\sum_i \langle w^* - w, x_i\rangle^2 \\ &= \frac{C}{N}\sum_i (y_i^* - \langle w, x_i\rangle - \zeta_i)^2 \\ &\le \frac{2C}{N}\sum_i (y_i^* - \langle w, x_i^*\rangle)^2 + \frac{2C}{N}\sum_i \zeta_i^2 \\ &\le \frac{2C}{N}\sum_i (y_i^* - \langle w, x_i^*\rangle)^2 + O(C\sigma^2)\end{aligned}$$

Thus, we have shown that

$$\frac{1}{N}\sum_i (y_i^* - \langle w, x_i^*\rangle)^2 = (*) \le OPT + O(\eta^{1/2})\left[\frac{2C}{N}\sum_i (y_i^* - \langle w, x_i^*\rangle)^2 + O(C\sigma^2)\right]$$

Note that the term on the left hand side appears on the right hand side as well. Thus, rearranging, we get

$$(1 - O(C\eta^{1/2}))\frac{1}{N}\sum_i (y_i^* - \langle w, x_i^*\rangle)^2 \le OPT + O(C\eta^{1/2}\sigma^2)$$

Putting this all together, if $C\eta^{1/2}$ is sufficiently small, we have

$$\text{Clean MSE } \le (*) \le (1 + O(C\eta^{1/2}))(OPT + O(C\eta^{1/2}\sigma^2))$$

We have thus shown that under some mild assumptions about the problem distribution such as the certifiable 4-hypercontractivity of x, the clean MSE obtained from a solution found using the SoS program achieves an MSE bounded by the expression above.

References