Learning, Algorithm Design and Beyond Worst-Case Analysis **Speaker: Ankur Moitra**

Robust Statistics, Revisited

Simons Insitute (online) **Scribe: Chi-Ning Chou**

Robust statistics considers the settings where the observed data is contaminated with some noise so that the underlying distribution might not lie in the model we concern. In literature, some approaches can guarantee to be robust in the sense that the error won't scale with the dimension. However, the running time is not polynomial. In this work, the the authors were trying to answer the following question:

Is robust estimation algorithmically possible in high-dimensions?

1 Classic and robust estimation

Let's first have some quick introduction to classic statistics and robust statistics to see the different rationales. For convenience, here we consider the parameter estimation problem.

1.1 Classic statistics

The following is a common parameter estimation problem in a classic statistics setting.

Problem 1 (1-dimensional Gaussian model) *Let $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, suppose the data is drawn from $\mathcal{N}(\mu, \sigma^2)$, how many samples required to estimated μ and σ^2 within ϵ error with high probability?*

Given data $X_1, \dots, X_n$, using empirical mean $\hat{\mu} = \frac{1}{n}\sum_{i\in[n]} X_i$ and empirical variance $\frac{1}{n}\sum_{i\in[n]})(X_i - \hat{\mu})^2$, one can have quadratic convergence result by central limit theorem.

In general, when dealing with such classic estimation problem, a rule of thumb is using the *maximum likelihood estimator (MLE)*. Fisher proved that under certain regularity condition, MLE is asymptotically efficient.

1.2 Robust statistics

In 1960, famous statistician Tukey asked the following question:

What if there's error in the model itself?

In other words, the goal would be finding estimators that behave well in a neighborhood around the model. Consider the following variants of 1-dimensional Gaussian model.

Problem 2 (robust 1-dimensional Gaussian model) *Let $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Given the corrupted sample from $\mathcal{N}(\mu, \sigma^2) + Z$, where Z is some noise distribution. How many samples required to estimated μ and σ^2 within ϵ error with high probability?*

Obviously, if we do not address any constraint on the noise, it's impossible to do anything. Thus, we need to come up with some useful notion to quantify the noise. Here, we consider the ℓ_1 norm of the noise. Specifically, when we constraint the model to have noise with $O(\epsilon)$ ℓ_1 norm, it's equivalent to consider arbitrary corruption of $O(\epsilon)$ fraction of the samples, which is a generalization of the *Hubers Contamination Model*: An adversary can add an -fraction of samples.

Apart from measuring the noise, we also need a notion to measure the performance of our estimator. It turns out that the total variation of distributions is a good choice.

Definition 3 (total variation) *Let f, g be probability distributions over $\mathbb{R}$, the total variation of them is*

$$d_{TV}(f,g) := \frac{1}{2}\int_{-\infty}^{\infty} |f(x) - g(x)| dx. \tag{1}$$

Remark 4 There are other equivalent definition of total variation. Here we adopt the one that corresponds to ℓ_1 distance,

Suppose f^* is the underlying distribution and f is the distribution we observed. Our goal would be finding an estimator with distribution g such that

$$d_T V(g, f^*) = O(\epsilon).$$

However, we might not have the underlying distribution f^*. Nevertheless, we have the observation$f$, and by the triangle inequality of total variation, it suffices to have

$$d_{TV}(g, f) = O(\epsilon).$$

Now, it's good time for us to play with the setting we set up above on the 1-dimensional Gaussian model. From the adversary view of noise, one can immediately see that the empirical mean and empirical variance estimators in the classic setting won't work anymore since we can add arbitrarily far mass from the true mean.

Luckily, using the *median* and the *median absolute deviation* resolve the problem. Recall that the median of an dataset $X = (X_1, \dots, X_n)$ is the data point that is no less than half of the data and no greater than half of the data point. The median absolute deviation is defined as

$$\mathrm{MAD}(X) := \mathrm{median}(X_i - \mathrm{median}(X)).$$

It turns out when given samples from distribution that is ϵ-close to $\mathcal{N}(\mu, \sigma^2)$, let $\hat{\mu} = \mathrm{median}(X)$ and $\hat{\sigma} = \frac{\mathrm{MAD}}{\Phi^{-1}(3/4)}$, we have

$$d_{TV}(\mathcal{N}(\mu, \sigma^2), \mathcal{N}(\hat{\mu}, \hat{\sigma}^2) = O(\epsilon).$$

Technically, we said the 1-dimensional Gaussian model is agnostically learnable in polynomial time.

2 Robustness vs. Hardness in High-dimensions

2.1 High-dimensional noisy Gaussian model

It's natural to generalize the previous 1-dimensional Gaussian model to d-dimensional Gaussian $\mathcal{N}(\mu, \Sigma)$ where μ is a d-dimensional vector and Σ is a $d \times d$ positive semidefinite matrix. The goal is to efficient estimator $\hat{\mu}$ and $\hat{\Sigma}$ such that given samples with distribution ϵ-close to $\mathcal{N}(\mu, \Sigma)$,

$$d_{TV}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\hat{\mu}, \hat{\Sigma})) = \tilde{O}(\epsilon).$$

Remark 5 It suffices to consider the following two special cases.

1. Unknown mean: $\mathcal{N}(\mu, I)$.

2. Unknown covariance matrix: $\mathcal{N}(0, \Sigma)$.

A sad news is that before this work, all known estimators are either hard to compute or lose polynomial factors in the dimension. For instance, consider the Tuckey median we saw in the 1-dimensional case, when generalizing to d-dimensional case, the running time is NP-hard. Another geometric median method is efficiently computable, however, only has $O(\epsilon\sqrt{d})$ error guarantee. Namely, it can only handle the situation when $\epsilon \leq \frac{1}{\sqrt{d}}$.

2.2 Robust estimation for high-dimensional Gaussian model [DKK+16]

Diakonikolas, Li, Kamath, Kane, Moitra, Stewart provided a positive answer to the robust estimation for high-dimensional Gaussian model with the following theorem.

Theorem 6 *There's an algorithm given $n = \tilde{O}(d^3/\epsilon^2)$ samples from a distribution that is ϵ-close in total variation distance to a d-dimensional Gaussian finds parameters that satisfy*

$$d_{TV}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\hat{\mu}, \hat{\Sigma})) = O(\epsilon \log^{3/2} 1\epsilon). \qquad (2)$$

Moreover, the running time of the algorithm is polynomial in n and d.

References

[DKK+16] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. *arXiv preprint arXiv:1604.06443*, 2016.