Sample Complexity and Good Sets

Daniel M. Kane

Departments of CS/Math
University of California, San Diego
dakane
dakane@ucsd.edu

August 13th, 2018

Sample Complexity

Ilias described how the filter and convex programming methods work with infinitely many samples. Here we will discuss how to get the details of the analysis correct for finitely many samples, and how to optimize the sample complexity in the analysis.

Outline

- Convex Program Approach and Covers
- Filtering Easy Analysis
- Good Sets and Full Analysis

Setup

We assume that we are trying to robustly learn a Gaussian $G = N(\mu, I) \subset \mathbb{R}^n$.

Setup

We assume that we are trying to robustly learn a Gaussian $G = N(\mu, I) \subset \mathbb{R}^n$.

For simplicity, we assume $\mu = 0$ (as everything we do is translation invariant).

Convex Program, Basic Conditions

Minimally, for the convex program to work, we need our set S of N unbiased samples to satisfy:

$$\left| \frac{1}{N} \sum_{X \in S} X \right| \leq \epsilon$$

and

$$\left| \frac{1}{N} \sum_{X \in S} XX^T - I \right|_2 \leq \epsilon$$

Mean

Note that

$$\frac{1}{N}\sum_{X\in S} X \sim N(0, I/\sqrt{N}).$$

So n/ϵ^2 samples suffice to ensure that it has norm less than ϵ.

Covariance

Note that

$$\left| \frac{1}{N} \sum_{X \in S} XX^T - I \right|_2 = \sup_{|v|_2 = 1} \left| \frac{1}{N} \sum_{X \in S} [(v \cdot X)^2 - 1] \right|.$$

Covariance

Note that

$$\left|\frac{1}{N}\sum_{X\in S} XX^T - I\right|_2 = \sup_{|v|_2=1}\left|\frac{1}{N}\sum_{X\in S}[(v\cdot X)^2 - 1]\right|.$$

- Need to show that this is small for *all* v.
- Could do if finitely many.

Covers

There exists a cover C of the unit sphere with:

- $|C| = 2^{O(n)}$.
- For all $|v|_2 = 1$, there is a $w \in C$ so that $v \cdot w \geq 9/10$.
- For any symmetric A, there is a $w \in C$ with $|w^T A w| \geq |A|_2/2$.

Covers

There exists a cover C of the unit sphere with:

- $|C| = 2^{O(n)}$.
- For all $|v|_2 = 1$, there is a $w \in C$ so that $v \cdot w \geq 9/10$.
- For any symmetric A, there is a $w \in C$ with $|w^T A w| \geq |A|_2/2$.

Enough to apply union bound over C.

Concentration

For each $w \in C$ need with high probability that

$$\left|\sum_{X \in S} (w \cdot X)^2 - 1\right| \leq N\epsilon/2.$$

Concentration

For each $w \in C$ need with high probability that

$$\left|\sum_{X \in S} (w \cdot X)^2 - 1\right| \leq N\epsilon/2.$$

Probability of failure

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\pm t \sum_{X \in S} (w \cdot X)^2 - 1\right)\right] e^{-Nt\epsilon/2} \\
=&\mathbb{E}\left[\exp(\pm t(G^2 - 1))\right]^N e^{-Nt\epsilon/2} \\
=&\exp(N(O(t^2) - t\epsilon/2)).
\end{aligned}$$

Concentration

For each $w \in C$ need with high probability that

$$\left| \sum_{X \in S} (w \cdot X)^2 - 1 \right| \leq N\epsilon/2.$$

Probability of failure

$$\begin{aligned} & \mathbb{E}\left[\exp\left(\pm t \sum_{X \in S} (w \cdot X)^2 - 1\right)\right] e^{-Nt\epsilon/2} \\ = & \mathbb{E}\left[\exp(\pm t(G^2 - 1))\right]^N e^{-Nt\epsilon/2} \\ = & \exp(N(O(t^2) - t\epsilon/2)). \end{aligned}$$

Setting t to be a sufficiently small multiple of ϵ, this is $\exp(-\Omega(N\epsilon^2))$, so $N = O(n/\epsilon^2)$ suffices.

Subset Bounds

For the analysis to work also need for any weight function $0 \leq w_x \leq 1/(1-2\epsilon)$, $\sum_{x \in S} w_x = 1$ that

$$\left| \sum_{X \in S} w_X X \right| \ll \epsilon \sqrt{\log(1/\epsilon)}$$

and

$$\left| \sum_{X \in S} w_X X X^T - I \right|_2 \ll \epsilon \log(1/\epsilon).$$

Simplifications

Enough to check for w the indicator of any subset of S of size $N(1-2\epsilon)$.

Simplifications

Enough to check for w the indicator of any subset of S of size $N(1-2\epsilon)$.

Enough to show that whp over each $v \in C$ and $S' \subset S$ not too small,

$$\left|\sum_{X \in S'} v \cdot X\right| \ll \epsilon\sqrt{\log(1/\epsilon)}N$$

and

$$\left|\sum_{X \in S'} (v \cdot X)^2 - 1\right| \ll \epsilon\log(1/\epsilon)N.$$

Simplifications

Enough to check for w the indicator of any subset of S of size $N(1-2\epsilon)$.

Enough to show that whp over each $v \in C$ and $S' \subset S$ not too small,

$$\left|\sum_{X \in S'} v \cdot X\right| \ll \epsilon\sqrt{\log(1/\epsilon)}N$$

and

$$\left|\sum_{X \in S'} (v \cdot X)^2 - 1\right| \ll \epsilon \log(1/\epsilon)N.$$

Assuming that the sum over all of S is good, suffices to show that the sums over $S \backslash S'$ are bounded.

Mean Bound

Need to bound

$$\left|\sum_{X\in S\setminus S'} v\cdot X\right|.$$

Mean Bound

Need to bound

$$\left| \sum_{X \in S \setminus S'} v \cdot X \right| .$$

At most

$$O(\epsilon\sqrt{\log(1/\epsilon)}N) + \sum_{X \in S, |v \cdot X| \geq 10\sqrt{\log(1/\epsilon)}} |v \cdot X| .$$

Mean Bound

Need to bound

$$\left|\sum_{X \in S \setminus S'} v \cdot X\right|.$$

At most

$$O(\epsilon\sqrt{\log(1/\epsilon)}N) + \sum_{X \in S, |v\cdot X| \geq 10\sqrt{\log(1/\epsilon)}} |v \cdot X|.$$

$$\begin{aligned} &\mathbb{E}\left[\exp\left(\sum_{X \in S, |v\cdot X| \geq 10\sqrt{\log(1/\epsilon)}} |v \cdot X|\right)\right] \\ =&\mathbb{E}_{Y\sim N(0,1)}[\exp(\mathbf{1}_{|Y|\geq 10\sqrt{\log(1/\epsilon)}}|Y|)]^N \\ \leq&(1+\epsilon^2)^N. \end{aligned}$$

Mean Bound

Need to bound

$$\left|\sum_{X \in S \setminus S'} v \cdot X\right|.$$

At most

$$O(\epsilon\sqrt{\log(1/\epsilon)}N) + \sum_{X \in S, |v \cdot X| \geq 10\sqrt{\log(1/\epsilon)}} |v \cdot X|.$$

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\sum_{X \in S, |v \cdot X| \geq 10\sqrt{\log(1/\epsilon)}} |v \cdot X|\right)\right] \\
=&\mathbb{E}_{Y \sim N(0,1)}[\exp(\mathbf{1}_{|Y| \geq 10\sqrt{\log(1/\epsilon)}}|Y|)]^N \\
\leq&(1+\epsilon^2)^N.
\end{aligned}$$

The probability of being too big is $\exp(-\Omega(\epsilon\sqrt{\log(1/\epsilon)}N))$, so $N = n/\epsilon^2$ suffices.

Covariance Bound

Need to bound

$$\left| \sum_{X \in S \setminus S'} (v \cdot X)^2 - 1 \right|.$$

Covariance Bound

Need to bound

$$\left| \sum_{X \in S \setminus S'} (v \cdot X)^2 - 1 \right|.$$

At most

$$O(\epsilon \log(1/\epsilon) N) + \sum_{X \in S, |v \cdot X| \geq 10\sqrt{\log(1/\epsilon)}} |v \cdot X|^2.$$

Covariance Bound

Need to bound

$$\left|\sum_{X\in S\setminus S'}(v\cdot X)^2-1\right|.$$

At most

$$O(\epsilon\log(1/\epsilon)N)+\sum_{X\in S,|v\cdot X|\geq 10\sqrt{\log(1/\epsilon)}}|v\cdot X|^2.$$

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\sum_{X\in S,|v\cdot X|\geq 10\sqrt{\log(1/\epsilon)}}|v\cdot X|^2/10\right)\right]\\
=&\mathbb{E}_{Y\sim N(0,1)}[\exp(\mathbf{1}_{|Y|\geq 10\sqrt{\log(1/\epsilon)}}|Y|^2/10)]^N\\
\leq&(1+\epsilon^2)^N.
\end{aligned}$$

Covariance Bound

Need to bound

$$\left|\sum_{X\in S\setminus S'} (v\cdot X)^2 - 1\right|.$$

At most

$$O(\epsilon\log(1/\epsilon)N) + \sum_{X\in S, |v\cdot X|\geq 10\sqrt{\log(1/\epsilon)}} |v\cdot X|^2.$$

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\sum_{X\in S, |v\cdot X|\geq 10\sqrt{\log(1/\epsilon)}} |v\cdot X|^2/10\right)\right]\\
=&\mathbb{E}_{Y\sim N(0,1)}[\exp(\mathbf{1}_{|Y|\geq 10\sqrt{\log(1/\epsilon)}}|Y|^2/10)]^N\\
\leq&(1+\epsilon^2)^N.
\end{aligned}$$

The probability of being too big is $\exp(-\Omega(\epsilon\log(1/\epsilon)N))$, so $N = n/\epsilon^2$ suffices.

Summary

So the conditions needed for the convex program hold with high probability so long as $N \gg n/\epsilon^2$.

Filter: Naive Analysis

If the bad points are coming i.i.d. from some distribution X, there is an easy analysis that uses separate samples for each filter step.

Filter: Naive Analysis

If the bad points are coming i.i.d. from some distribution X, there is an easy analysis that uses separate samples for each filter step.

1. Find a rough approximation to μ.
2. Throw out samples $10\sqrt{n}$ far from μ (an exponentially small fraction of good samples).
3. Take samples, and compute empirical covariance.
4. If no large eigenvalues, return sample mean.
5. If large eigenvalue v, use samples to approximate the distribution $v \cdot X$ and find a threshold for a filter.
6. Return to step 3, applying the filter to all future samples.

Sample Complexity

- By Chernoff Bounds, $O(n^2/\epsilon^2)$ samples suffice to ensure good approximations to mean and covariance.
- $O(n/\epsilon^2)$ samples suffices to approximate cumulative density distribution of $v \cdot X$ to error $\epsilon/\sqrt{n}$.
- Need to take this many samples *every round*.

Good Sets

If you want to reuse the same samples between rounds or work against stronger error models, need to have a condition on the set of uncorrupted samples that implies the algorithm will work.

Good Sets

If you want to reuse the same samples between rounds or work against stronger error models, need to have a condition on the set of uncorrupted samples that implies the algorithm will work.

A *good set* of samples S should:

- Have appropriate mean and covariance even when restricting to $(1-\epsilon)$-dense subsets.
- Not loose too many points to filters.
- Have a set of N i.i.d. points of G be good with high probability.

Analysis Template

Assume S is a good set. Let $D(S, S') = |S \Delta S'|/|S|$.

Analysis Template

Assume S is a good set. Let $D(S, S') = |S \Delta S'|/|S|$.

Want a procedure that given S' with $D(S, S') \leq 2\epsilon$ for a good set S either:

- Returns a $\tilde{\mu}$ with $|\mu - \tilde{\mu}| = O(\epsilon\sqrt{\log(1/\epsilon)})$.
- OR returns an S'' with $D(S, S'') < D(S, S')$.

Analysis Template

Assume S is a good set. Let $D(S, S') = |S \Delta S'|/|S|$.

Want a procedure that given S' with $D(S, S') \leq 2\epsilon$ for a good set S either:

- Returns a $\tilde{\mu}$ with $|\mu - \tilde{\mu}| = O(\epsilon\sqrt{\log(1/\epsilon)})$.
- OR returns an S'' with $D(S, S'') < D(S, S')$.
  - Usually $S'' \subset S'$ and less than half of the elements of $S' \backslash S''$ are in S.

Analysis Template

Assume S is a good set. Let $D(S, S') = |S \Delta S'|/|S|$.

Want a procedure that given S' with $D(S, S') \leq 2\epsilon$ for a good set S either:

- Returns a $\tilde{\mu}$ with $|\mu - \tilde{\mu}| = O(\epsilon\sqrt{\log(1/\epsilon)})$.
- OR returns an S'' with $D(S, S'') < D(S, S')$.
  - Usually $S'' \subset S'$ and less than half of the elements of $S' \backslash S''$ are in S.

Iterating procedure eventually returns a valid approximation.

Basic Analysis

The existing conditions ensure that if $\mathrm{Cov}(S') \leq I(1 + O(\epsilon \log(1/\epsilon)))$, then we get a good approximation to the mean.

Basic Analysis

The existing conditions ensure that if $\mathrm{Cov}(S') \leq I(1 + O(\epsilon \log(1/\epsilon)))$, then we get a good approximation to the mean.

Otherwise, if v is an eigenvector with eigenvalue $1 + \delta$, can approximate $v \cdot \mu$ to error $O(1)$ (by taking a median), and can find a threshold T beyond which there are more points than there should be. Need S to not have too many points beyond this threshold.

Samples

If we threw out points more than $10\sqrt{n}$ from μ, can find T so that

$$\mathrm{Pr}_{Y \in_u S'}(v \cdot Y \geq T) \geq 2\mathrm{Pr}(v \cdot G \geq T) + (\epsilon/n).$$

It is enough to have

$$\mathrm{Pr}_{X \in_u S}(v \cdot X \geq T) = \mathrm{Pr}(v \cdot G \geq T) + O(\epsilon/n).$$

for every v, T.

Samples

If we threw out points more than $10\sqrt{n}$ from μ, can find T so that

$$\mathrm{Pr}_{Y \in_u S'}(v \cdot Y \geq T) \geq 2\mathrm{Pr}(v \cdot G \geq T) + (\epsilon/n).$$

It is enough to have

$$\mathrm{Pr}_{X \in_u S}(v \cdot X \geq T) = \mathrm{Pr}(v \cdot G \geq T) + O(\epsilon/n).$$

for every v, T.

The set of halfspaces as test has VC-dimension n, and so by the VC-inequality, this happens whp when $N \gg n^3/\epsilon^2$.

Improved Sample Complexity

This analysis requires many samples in order to get such precise control over the tail bounds. If we filter in a more relaxed manner, we can get by with weaker bounds.

Improved Sample Complexity

This analysis requires many samples in order to get such precise control over the tail bounds. If we filter in a more relaxed manner, we can get by with weaker bounds.

If an ϵ-fraction of errors increases the variance in the v-direction by much more than $\epsilon \log(1/\epsilon)$, then

$$\sum_{X \in S' \setminus S, |v \cdot (X-\mu)| > 10\sqrt{\log(1/\epsilon)}} |v \cdot (X - \mu)|^2 \gg \epsilon \log(1/\epsilon)|S|.$$

Improved Sample Complexity

This analysis requires many samples in order to get such precise control over the tail bounds. If we filter in a more relaxed manner, we can get by with weaker bounds.

If an ϵ-fraction of errors increases the variance in the v-direction by much more than $\epsilon \log(1/\epsilon)$, then

$$\sum_{X \in S' \setminus S, |v \cdot (X-\mu)| > 10\sqrt{\log(1/\epsilon)}} |v \cdot (X-\mu)|^2 \gg \epsilon \log(1/\epsilon)|S|.$$

We note that this is much more than this sum should be over good samples. So if

$$\sum_{X \in S, |v \cdot (X-\mu)| > 10\sqrt{\log(1/\epsilon)}} |v \cdot (X-\mu)|^2 \ll \epsilon \log(1/\epsilon)|S|.$$

we can filter by throwing away X with $|v \cdot (X-\tilde{\mu})| > 10\sqrt{\log(1/\epsilon)}$ with probability proportional to $|v \cdot (X-\mu)|^2$.

Improved Sample Complexity

By the bound shown before, if $N \gg n/\epsilon^2$, this happens with high probability for all $w \in C$. It is not hard to modify to work for all v.

Improved Sample Complexity

By the bound shown before, if $N \gg n/\epsilon^2$, this happens with high probability for all $w \in C$. It is not hard to modify to work for all v.

Upshot: With this filtering method $O(n/\epsilon^2)$ samples suffices.