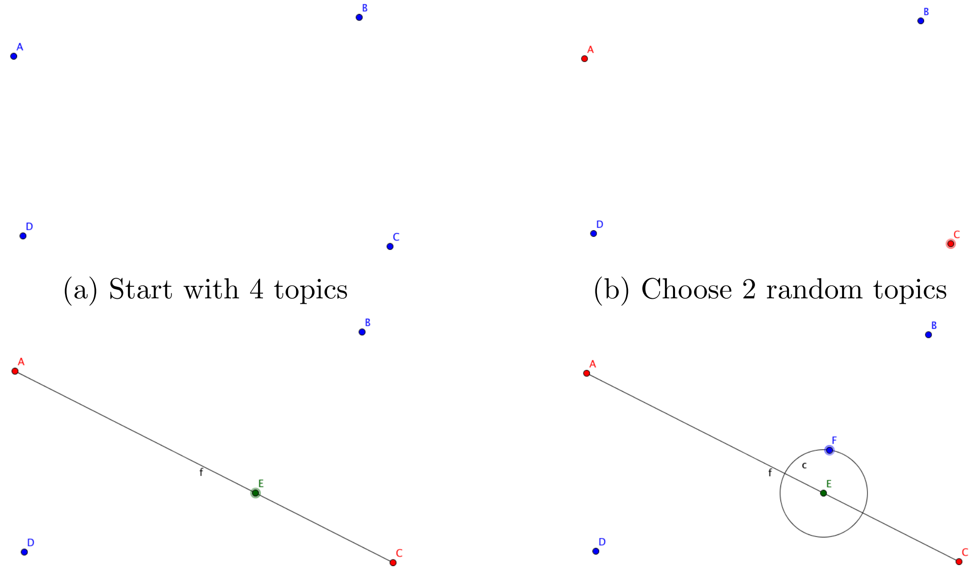
Generative Models

Ananya Kumar

January 15, 2017

Suppose we have T topics in d dimensional space. We assume that no 3 topics are collinear. We repeatedly pick a random pair of topics T_1 and T_2. Each pair is picked with equal probability. Then we pick a uniformly random data point in the line segment connecting T_1 and T_2. We repeat this process m times to generate m data points.



(a) Start with 4 topics

(b) Choose 2 random topics

(c) Pick point in the middle

(d) Pick data point in circle

Figure 1: Illustration of uniform noise case of generating points

In the noisy case, we perturb the generated data points. In the uniform noise case, the data points are perturbed so that they are uniformly random in a sphere of radius δ centered around the original location of the data point. We can also relax the distributional assumptions on how the data is generated, consider different noise models, or consider k-sparsity models where data points are generated from k topics.

1 Goals

There are a few possible goals:

1. In the noiseless case, retrieve all T topics, exactly. Assuming we generated $\Omega(T^2 \log T)$ data points, we have a polynomial time algorithm. The probability of failure decreases exponentially with the constant factor. As an example, if we have around 50 topics, about 100,000 points are sufficient for a 99% probability of recovering all topics. Can we recover the topics with fewer points?

2. In the uniform noise case, assuming we know the number of topics, can we recover all topics with error ϵ? We probably need additional conditions to make this problem tractable.

3. Instead of recovering the topics, it could be sufficient to find $O(T)$ topic proposals. We want every data point to be within ϵ from the line segment connecting some pair of topic proposals.

2 Noiseless Case

Reading this section is not too important for the rest of the document. I wrote up the details in another document, but I'll recap the key ideas for completeness.

If you generate $\Omega(T^2 \log T)$ data points, with high probability every pair of topics has at least 3 data points generated between them. Pick all triplets of collinear data points, and draw an (infinite) line through them. Consider all intersections between pairs of lines. Only select pairs of intersections with data points on the (finite) line segment connecting them. Remove any intersection that is not in a selected pair. Let S be the set of remaining intersections.

The convex hull of S is the convex hull of the topics, so we have recovered the topic convex hull. To recover the remaining topics, remove all the points on the topic convex hull from S. Also remove any points in S that lie on a segment connecting pairs of topics in the topic convex hull. Repeat the process of selecting the convex hull and removing points until we have identified all topics (removed all the data points).

3 Densest Point Algorithm

I'll motivate the algorithm with a simple 2D example. Suppose we have 3 topics, T_1, T_2, T_3, that form an equilateral triangle in 2D space with side length 1. Suppose we generate m data points through our uniform noise generation process, with noise parameter δ. We choose radius $r = \epsilon/2$, assuming ϵ is much smaller than the side length of the equilateral triangle.

Definition 3.1. Given a point p in 2D space, the *density* around p is the proportion of generated data points that are within distance r of p.

In this case, the expected density is maximized close to the 3 topics. If p is a point on the segment connecting T_1 and T_2, that is distance r away from T_1, the density is $3r/3 = r$. If p is a point on the segment connecting T_1 and T_2, that is distance $> 2r = \epsilon$ away from T_1 and T_2, the density is $\leq 2r/3$.

We want a theorem that says that if m is large enough, with high probability the actual density is highest near the topics. The VC dimension of a circle is 3. So we can get a uniform bound that bounds the probability that the density deviates from the expected density at any point by more than $r/3$. We can choose m such that this probability is very low. Then, if we have m points, with high probability we can locate the topics with error ϵ by finding the highest density point. This can be done in time polynomial in m.

Inspired by this success, we propose the following, not yet fully precise algorithm, to recover the topics. Assume that we have generated m data points, for some large m. We pick the data point p with the highest density (as defined above, for some pre-defined r). p is our estimate for one of the topics, in that it should be within ϵ from some topic. We then remove all the data points within distance r' (for some pre-defined r') of p. We continue the process until we recover all T topics.

For this particular approach, a few important tasks remain:

1. This algorithm does not always work efficiently. We need to impose some additional constraints. For example, if we have 3 topics that are almost collinear, then we need to generate many data points to recover the topics. What constraints does the algorithm depend on?

2. Consider the case where we have many topics, and draw the segments connecting all pairs of topics. Some of the segments intersect, and the expected density at the intersection points may be high. Does the algorithm still work in these cases, or do we need additional constraints?

3. r and r' might depend on constraints specific to the problem. We need to choose suitable constants for r and r' and complete the analysis.

4. Intuitively this algorithm could work for k-sparsity versions of the problem, where we generate data points from combinations of k topics. It could work for other noise distributions (e.g. Gaussian) as well. This would be significant, so we could examine this case.