Lecture 22: Learning Gaussian Covariance Matrices, and Learning and Testing Product Distributions

Daniel Kane
Scribe: Ken Hoover

May 22, 2017

1 Last Lecture

In the last lecture, we covered upper bounds on the sample complexity of A_k-closeness testing. We also gave tight bounds of $\Theta(n/\epsilon^2)$ for learning the mean of a Gaussian distribution over $\mathbb{R}^n$ with the identity covariance matrix to error ϵ, and of $\Theta(\sqrt{n}/\epsilon^2)$ for testing whether the mean of a given Gaussian, also with the identity covariance matrix, is equal to 0 or has L_2-norm greater-than ϵ.

2 Learning the Covariance Matrix of a Gaussian

Upper Bound

Our upper bound is obtained by examining the sample covariance matrix. Namely, given a collection of N samples over $\mathbb{R}^n$, $\{\mathbf{x}^{(i)} \sim \mathcal{N}(0, \Sigma)\}$, the algorithm returns

$$\widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^{(i)}(\mathbf{x}^{(i)})^T \ .$$

For analyzing how large N must be, we can assume that $\Sigma = \mathsf{I}$, the identity matrix; this is because there is an affine transformation between $\mathcal{N}(0, \Sigma)$ and $\mathcal{N}(0, \mathsf{I})$. We know, from the previous class, that

$$d_{TV}(\mathcal{N}(0, \mathsf{I}), \mathcal{N}(0, \Sigma)) = \Theta(\min(1, ||\mathsf{I} - \Sigma||_F)) \ ,$$

where $||\cdot||_F$ is the Frobenius norm. We also note that $d_{TV}(\mathcal{N}(0, \mathsf{I}), \mathcal{N}(0, \widehat{\Sigma}))$, treated as a distribution over $\widehat{\Sigma}$, has an expected value of 0. So, we then estimate the variance of $d_{TV}(\mathcal{N}(0, \mathsf{I}), \mathcal{N}(0, \widehat{\Sigma}))$ as

$$||\mathsf{I} - \widehat{\Sigma}||_F^2 = \sum_{i,j}\left[\frac{1}{N}\sum_{k=1}^{N}(x_i^{(k)}x_j^{(k)} - \delta_{ij})\right]^2 \ .$$

Each (i, j) term contributes $1/N$ to the sum, thus the variance in our estimator is approximately n^2/N. Thus, $N = C(n^2/\epsilon^2)$ samples are required, for some constant C.

Lower Bound

For our lower bound, an outline of the proof is given. Consider the family of symmetric matrices defined by $\Sigma_{ij} = \Sigma_{ji} = \delta_{ij} \pm \frac{c\epsilon}{n}$, where c is some constant and plus or minus is chosen uniformly and independently for each coordinate. Any algorithm which then wants to learn $\mathcal{D} = \mathcal{N}(0, \Sigma)$ to within ϵ must then be able to guess 2/3rds of the signs correctly.

For a given index (i, j) of Σ, and an arbitrary vector $\mathbf{x}$, we have that changing the sign of $\pm c\epsilon/n$ at Σ_{ij} results in a multiplicative difference of approximately $\exp(\pm x_i x_j \epsilon/n)$ in the pdf at $\mathbf{x}$. This follows from the fact that, for $\Sigma \approx \mathsf{I}$,

$$\mathbf{x}^T \Sigma^{-1} \mathbf{x} \approx \mathbf{x}^T (2\mathsf{I} - \Sigma)\mathbf{x} \ .$$

Since $\epsilon/n << 1$, we have that $\pm x_i x_j \epsilon/n$ is very close to 0. Thus, using the linear approximation of $e^x \approx 1 + x$ near 0, we have that the information gain between Σ and $\mathbf{x}$ is approximately $x_i^2 x_j^2 \epsilon^2/n^2$. Also note that x_i^2 and x_j^2 are approximately 1, on average, so we can expect each sample $\mathbf{x}$ to give ϵ^2/n^2 bits of information about Σ. We then need that $N\epsilon^2/n^2 >> 1$, which gives the lower bound.[1]

3 Binary Product Distributions

A *binary product distribution* $\mathcal{P}$ over $\{0,1\}^n$ is a distribution such that, given $\mathbf{x} \sim \mathcal{P}$, x_i is independent of every x_j for $j \neq i$. We can define $\mathcal{P}$ by $\mathbf{p} \in [0,1]^n$, where $p_i = \mathbb{E}[x_i]$.

The first question we ask of such distributions is what metric do we use in comparing them? The natural answer is the total variation distance. Since binary product distributions are discrete, we have an explicit formula for this distance. Namely, given binary product distributions $\mathcal{P}$ and $\mathcal{Q}$ with mean vectors $\mathbf{p}$ and $\mathbf{q}$, we have that

$$d_{TV}(\mathcal{P}, \mathcal{Q}) = \frac{1}{2} \sum_{\mathbf{x} \in \{0,1\}^n} \left| \prod_{i=1}^n [p_i^{x_i}(1-p_i)^{1-x_i}] - \prod_{i=1}^n [q_i^{x_i}(1-q_i)^{1-x_i}] \right| \ .$$

However, this usually isn't that useful for our proofs. Instead, we want other bounds that are more amenable to analysis. One such bound is

$$d_{TV}(\mathcal{P}, \mathcal{Q}) \leq ||\mathbf{p} - \mathbf{q}||_1 \ .$$

However, this is not a great upper bound. A better one can be obtained as follows. Suppose X is a random bit, and that $Y = \mathcal{P}$ if $X = 0$, otherwise $Y = \mathcal{Q}$, where $\mathcal{P}$ and $\mathcal{Q}$ are both binary product distributions. We then have that

$$d_{TV}(\mathcal{P}, \mathcal{Q})^2 << I(X;Y) \ .$$

Since each Y_i is independent conditioned on X, we have that

$$\begin{aligned} I(X;Y) &\leq \sum_i I(X;Y_i) \\ &= \Theta\left(\sum_i \frac{(p_i - q_i)^2}{(p_i + q_i)(2 - p_i - q_i)} \right) \ . \end{aligned}$$

[1] As mentioned in class, a lot of details about this proof have been shoved under the rug.

When we use one versus the other depends on how "unbalanced" the probability of each coordinate is. If for every i, we have that p_i and q_i both lie in some constant range, say $[1/10, 9/10]$, then the denominators of the second bound are constants, and so $d_{TV}(\mathcal{P}, \mathcal{Q}) \approx ||\mathbf{p} - \mathbf{q}||_2$. Otherwise, $d_{TV}(\mathcal{P}, \mathcal{Q}) \approx ||\mathbf{p} - \mathbf{q}||_1$.

Learning Binary Product Distributions

Upper Bound: For the upper bound, we use the natural algorithm of taking the vector of sample means $\widehat{\mathbf{p}}$. Note that each $\widehat{p_i}$ is a random variable with mean p_i, and variance $p_i(1-p_i)/N$, where N is the number of samples. We can upper bound the squared total variation distance between $\mathcal{P}$ and $\widehat{\mathcal{P}}$, the distribution defined by $\widehat{\mathbf{p}}$, by

$$d_{KL}(\mathcal{P}||\widehat{\mathcal{P}}) = \sum_i \frac{(\widehat{p_i} - p_i)^2}{p_i(1-p_i)} \ ,$$

the Kullback-Leibler divergence. Note that in expectation, this is just $\sum_i \mathrm{Var}(\widehat{p_i})/(N \cdot \mathrm{Var}(\widehat{p_i}))$, and so the expected squared error is n/N. Thus we just need $N \geq n/\epsilon^2$.

Lower Bound: Consider the family of product distributions defined by $p_i = 1/2 \pm c\epsilon/\sqrt{n}$, where plus or minus is chosen uniformly and independently at random for each coordinate. Then

$$\begin{aligned} I(X; \text{Samples}) &\leq N \cdot I(X; \text{A Sample}) \\ &\leq Nn \cdot I(X; \text{One coordinate of one sample}) \\ &= Nn \cdot \frac{c^2\epsilon^2}{n} \\ &= Nc^2\epsilon^2 \ , \end{aligned}$$

which we require to be much larger than n. Thus, $N \geq n/\epsilon^2$ is a matching lower bound.

Identity Testing for Product Distributions[2]

For this, we are given a binary product distribution $\mathcal{Q}$, along with its mean vector $\mathbf{q}$, and want to determine if some unknown product distribution $\mathcal{P}$ is equal to $\mathcal{Q}$, or has total variation distance greater-than ϵ. First, we can suppose that $q_i \leq 1/2$ for each coordinate i; if this isn't the case, then replace q_i with $1 - q_i$, and flip the ith bit of each sample from $\mathcal{P}$. This leaves the total variation distance unchanged.

Again, we use the KL distance bound, and want to see if

$$\sum_i \frac{(p_i - q_i)^2}{q_i(1-q_i)} > \epsilon^2 \ .$$

Our approach for doing this is similar to how we derived χ^2 testers in the discrete univariate case. Namely, for each $i \in [n]$, let M_i be sampled i.i.d. from $\mathrm{Poi}(m)$. Then, take $M = max_{i \in [n]}\{M_i\}$

[2]This and the closeness testing portion of the notes can all be found with further detail in section 4 of the Canonne et al. paper on the course website, if further clarification is needed.

samples $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}$ from $\mathcal{P}$, and for each i record the number of times $x_i = 1$ in the first M_i samples as X_i. Note that $M \in O(m)$ for $m >> \log n$. Now define

$$Z = \sum_i \frac{(X_i - mq_i)^2 - X_i}{q_i(1-q_i)} \ .$$

We then have that

$$\mathbb{E}[Z] = m^2 \sum_i \frac{(p_i - q_i)^2}{q_i(1-q_i)} \ .$$

Furthermore, we can prove that

$$\mathrm{Var}(Z) = O\left(\sum_i \frac{m^2 p_i^2 + m^3 p_i (p_i - q_i)^2}{q_i^2} \right) \ .$$

Together, these two can be used to derive a $\sqrt{n}/\epsilon^2$ sample upper bound. Note that a matching lower bound can also be found.

Lower Bounds for Closeness Testing for Product Distributions

For the balanced case (i.e. $p_i, q_i \in [c, 1-c]$ for some constant c and every i), we note that closeness testing here is similar to closeness testing for Gaussians with diagonal covariance matrices. This requires $\Omega(\sqrt{n}/\epsilon^2)$ samples.

In the unbalanced case, suppose an algorithm decides closeness with m samples, possibly depending on ϵ and n. Let X be a uniform random bit. Define two product distributions $\mathcal{P}$ and $\mathcal{Q}$ as follows. For each $i \in [n]$, with $1/2$ probability set $p_i = q_i = 1/m$. Otherwise, if $X = 0$, set $p_i = q_i = 1/n$, and if $X = 1$, set $p_i = (1 \pm \epsilon)/n$, and $q_i = (1 \mp \epsilon)/n$. The intuition here is the same as when proving the lower bound for closeness testing on discrete distributions. However, here we can have many more noise bins (the $\approx n/2$ coordinates where $p_i = q_i = 1/m$), since we are no longer constrained to $||\mathbf{p}||_1 = 1$.

Note that if $X = 0$, then $\mathcal{P} = \mathcal{Q}$. If $X = 1$, then with high probability $d_{TV}(\mathcal{P}, \mathcal{Q}) >> \epsilon$. Consider the m samples from each distribution, $\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(m)}$ and $\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(m)}$. Note that

$$\begin{aligned} I(X; \{\mathbf{p}^{(j)}\}, \{\mathbf{q}^{(j)}\}) &\leq n \cdot I\left(X; \sum_{j=1}^m p_1^{(j)}, \sum_{j=1}^m q_1^{(j)} \right) \\ &= n \cdot \sum_{k_p, k_q \geq 0} \frac{(\Pr(A_1 = k_p, B_1 = k_q | X = 0) - \Pr(A_1 = k_p, B_1 = k_q | X = 1))^2}{\Pr(A_1 = k_p, B_1 = k_q | X = 0) + \Pr(A_1 = k_p, B_1 = k_q | X = 1)} \end{aligned}$$

where $A_i = \sum_{j=1}^m p_i^{(j)}$ and $B_i = \sum_{j=1}^m q_i^{(j)}$. This sum can be shown to be equal to $(m\epsilon/n)^4$; since we require $I(X; \{\mathbf{p}^{(j)}\}, \{\mathbf{q}^{(j)}\}) >> 1$, we need $n(m\epsilon/n)^4 >> 1$, which implies $m >> n^{3/4}/\epsilon$. Thus, a lower bound on the sample complexity of $\max(\sqrt{n}/\epsilon^2, n^{3/4}/\epsilon)$ is obtained. Next class, we show that this is sufficient for closeness testing.