Lecture 2: Learning a High-Dimensional Gaussian

- Last time: testing high dimensional distributions (= large domain size) is hard (requires many samples).
- Today: a strong assumption under which efficient learning is possible.

Gaussian Distributions

- Reminder: Gaussian in one dimension:  $\Pr(x) \propto e^{-(x-\mu)^2/2\sigma^2}$
- CLT: add together many indep. r.v.'s → Gaussian
- Also holds in high dimensions!
- So if getting samples from a high-dimensional population where high-dimensional features act like sums of indep rv's, Gaussian assumption can be reasonable.
- Multivariate Gaussian: any affine transformation of

$$\Pr(x) \propto e^{-\|x\|^2/2}$$

- Notation: transform by $x \to Ax+\mu$, distn is called

$N(\mu, A^2)$ (traditionally: $N(\mu, \Sigma)$.)

- **Fact**: $\mathbb{E}_{x \sim N(\mu,\Sigma)} x = \mu$, $\mathbb{E}_{x \sim N(\mu,\Sigma)} (x-\mu)(x-\mu)^T = \Sigma$

"Gaussian is determined by its 1st and 2nd moments"

Task: given $x_1 \ldots x_n \sim N(\mu, \Sigma)$ for some unknown $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$, find some dist'n D s.t. $TV(D, N(\mu,\Sigma)) \leq \varepsilon$.

$\boxed{\text{How big does } n = n(d,\varepsilon) \text{ need to be?}}$

Theorem: $n = O(d^2/\varepsilon^2)$ suffices.

(Compare with $n = \Omega(\sqrt{2^d})$ for uniformity testing on $\{0,1\}^d$, only easier than learning general dist'n on $\{0,1\}^d$.)

(And, can do in polynomial time.)

- Let $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i$, $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (x_i - \hat{\mu})(x_i - \hat{\mu})^T$.

- **Lemma** (directly implies Theorem above):

$$\mathbb{E}_{x_1 \ldots x_n \sim N(\mu,\Sigma)} TV(N(\hat{\mu},\hat{\Sigma}), N(\mu,\Sigma)) \leq O\left(\frac{d}{\sqrt{n}}\right).$$

Proof outline:

① $$\mathrm{TV}(N(\hat\mu,\hat\Sigma), N(\mu,\Sigma)) \le \mathrm{TV}(N(\hat\mu,\Sigma), N(\mu,\Sigma)) + \mathrm{TV}(N(\hat\mu,\hat\Sigma), N(\hat\mu,\Sigma))$$

$$= \underbrace{\mathrm{TV}(N(\Sigma^{-1/2}\hat\mu, I), N(\Sigma^{-1/2}\mu, I))}_{Ⓐ} + \underbrace{\mathrm{TV}(N(0,\hat\Sigma), N(0,\Sigma))}_{Ⓑ}$$

$$\le O(\|\Sigma^{-1/2}(\hat\mu-\mu)\|) \qquad \le O(\|I-\Sigma^{-1/2}\hat\Sigma\Sigma^{-1/2}\|_F)$$

② $$\mathbb{E}\|\Sigma^{-1/2}(\hat\mu-\mu)\|^2 \le O\left(\frac{d}{n}\right)$$

③ $$\mathbb{E}\|I-\Sigma^{-1/2}\hat\Sigma\Sigma^{-1/2}\|_F \le O\left(\frac{d}{\sqrt{n}}\right)$$

<u>Reminder</u>: $$\|M\|_F^2 = \mathrm{Tr}\, MM^T = \mathrm{Tr}\, M^TM$$

$$= \sum_{i,j\le d} M_{ij}^2 = \sum \lambda_i(M)^2$$

Singular values of M

"Frobenius norm"

<u>Reminder 2</u>: $$\|v\|^2 = \mathrm{Tr}\, vv^T = \sum_i v_i^2$$

(A) Goal: $TV(N(\mu, I), N(\tau, I)) \le O(\|\mu - \tau\|)$

Enough: $KL(N(\mu, I) \| N(\tau, I)) \le O(\|\mu - \tau\|^2)$

$$= \mathop{\mathbb{E}}_{x \sim N(\mu, I)} \log \frac{e^{-\|x-\mu\|^2/2}}{e^{-\|x-\tau\|^2/2}}$$

$$\frac{1}{2}\left[\mathop{\mathbb{E}}_{x \sim N(\mu, I)} -\|x-\mu\|^2 + \|x-\tau\|^2\right]$$

$$= \frac{1}{2}\mathbb{E}\left[-\|x-\mu\|^2 + \|x-\mu+\mu-\tau\|^2\right]$$

$$= \frac{1}{2}\mathbb{E}\left[-\|x-\mu\|^2 + \|x-\mu\|^2 + 2\langle x-\mu, \mu-\tau\rangle + \|\mu-\tau\|^2\right]$$

$$= \frac{1}{2}\|\mu - \tau\|^2$$

$\square$

(B) Enough: $KL(N(0,\Sigma) \| N(0,\Gamma)) \leq O(\|I - \Gamma^{-1/2}\Sigma\Gamma^{-1/2}\|_F^2)$

Recall that $P_{N(0,\Sigma)}(x) = \left(\frac{1}{2\pi}\right)^{d/2} \frac{1}{\sqrt{\det \Sigma}} e^{-\|\Sigma^{-1/2}x\|^2/2}$

$$KL(\cdots) = \mathbb{E}_{x\sim N(0,\Sigma)} \log \frac{\frac{1}{\sqrt{\det\Sigma}} e^{-\|\Sigma^{-1/2}x\|^2/2}}{\frac{1}{\sqrt{\det\Gamma}} e^{-\|\Gamma^{-1/2}x\|^2/2}}$$

$$= \frac{1}{2}\left[-\log \frac{\det\Sigma}{\det\Gamma} + \mathbb{E}_{x\sim N(0,\Sigma)} \|\Gamma^{-1/2}x\|^2 - \|\Sigma^{-1/2}x\|^2\right]$$

$$= \frac{1}{2}\left[-\log \det(\Gamma^{-1/2}\Sigma\Gamma^{-1/2}) + \mathbb{E}\operatorname{Tr}(\Gamma^{-1/2}xx^T\Gamma^{-1/2} - \Sigma^{-1/2}xx^T\Sigma^{-1/2})\right]$$

$$= -\frac{1}{2}\log\det(\Gamma^{-1/2}\Sigma\Gamma^{-1/2}) + \frac{1}{2}\operatorname{Tr}(\Gamma^{-1/2}\Sigma\Gamma^{-1/2} - I)$$

let $\lambda_1 \ldots \lambda_d$ be the eigs of $\Gamma^{-1/2}\Sigma\Gamma^{-1/2} - I$ Then,

$$= -\frac{1}{2}\sum \log(1+\lambda_i) + \frac{1}{2}\sum \lambda_i$$

$$= O\left(\sum \lambda_i^2\right) \quad \text{if } |\lambda_i| < 1 \ \forall i \text{ (Convergence of Taylor).}$$

ok to assume, since o.w. $\|I - \Gamma^{-1/2}\Sigma\Gamma^{-1/2}\|_F \gtrsim 1 \gtrsim TV(\cdots)$.

$\square$

(2) $\mathbb{E}\| \Sigma^{-1/2}(\hat{\mu} - \mu) \|^2 = \mathbb{E}\| \frac{1}{n}\sum_{i=1}^{n} \Sigma^{-1/2} X_i - \mu \|^2$

$X_i = \Sigma(z_i + \mu)$ for $z_i \sim N(0, I)$, so

$$= \mathbb{E}_{z_1 \ldots z_n} \| \frac{1}{n}\sum_{i=1}^{n} z_i \|^2 = \frac{1}{n^2}\mathbb{E}\sum_{i,j} z_i^T z_j = \frac{1}{n^2}\sum_i \mathbb{E}\|z_i\|^2$$

$$= \frac{d}{n}$$

$\square$

③ $$I - \Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} = I - \frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1/2}(x_i-\hat{\mu})(x_i-\hat{\mu})^T\Sigma^{-1/2}$$

$$= I - \frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1/2}(x_i-\mu+\mu-\hat{\mu})(x_i-\mu+\mu-\hat{\mu})^T\Sigma^{-1/2}$$

$$= I - \frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1/2}(x_i-\mu)(x_i-\mu)^T - \Sigma^{-1/2}(\mu-\hat{\mu})(\mu-\hat{\mu})^T\Sigma^{-1/2}$$

$$\underbrace{- \frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1/2}(x_i-\mu)(\mu-\hat{\mu})^T\Sigma^{-1/2}}_{= \Sigma^{-1/2}(\hat{\mu}-\mu)(\mu-\hat{\mu})^T\Sigma^{-1/2}} - \frac{1}{n}\sum_{i=1}\Sigma^{-1/2}(\mu-\hat{\mu})(x_i-\mu)\Sigma^{-1/2}$$

By triangle inequality,

$$\mathbb{E}\|I-\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}\|_F \le \mathbb{E}\left\|I-\Sigma^{-1/2}\left(\frac{1}{n}\sum_{i=1}(x_i-\mu)(x_i-\mu)^T\right)\Sigma^{-1/2}\right\|_F$$
$$+\ \mathbb{E}\|\Sigma^{-1/2}(\hat{\mu}-\mu)\|_2^2$$

$$\le \mathbb{E}\left\|I-\Sigma^{-1/2}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right)\Sigma^{-1/2}\right\|_F + O\left(\frac{d}{n}\right) \text{ by } ②.$$

$\Sigma^{-1/2}(x_i-\mu)$ is distributed as $z \sim N(0,I)$, so get

$$\le \mathbb{E}_{z_1\ldots z_n}\left\|I-\frac{1}{n}\sum z_iz_i^T\right\|_F + O\left(\frac{d}{n}\right).$$

$$\le \left(\mathbb{E}_{z_1\ldots z_n}\left\|I-\frac{1}{n}\sum z_iz_i^T\right\|_F^2\right)^{1/2} + O\left(\frac{d}{n}\right)$$

$$= O\left(\frac{d}{n}\right) + \left(\frac{1}{n^2}\sum_{i,j\le n}\mathbb{E}\left[\mathrm{Tr}\left(z_iz_i^T-\mathbb{E}z_iz_i^T\right)\cdot\left(z_jz_j^T-\mathbb{E}z_jz_j^T\right)\right]\right)^{1/2}$$

$$= O\left(\frac{d}{n}\right) + \left(\frac{1}{n}\, \mathbb{E}\, \mathrm{Tr}\left(ZZ^T - \mathbb{E}ZZ^T\right)^2\right)^{1/2}$$

$$= O\left(\frac{d}{n}\right) + \left(\frac{1}{n}\, \mathbb{E} \sum_{i,j \le d} \left(Z(i)Z(j) - \mathbb{E}\, Z(i)Z(j)\right)^2\right)^{1/2}$$

$$= O\left(\frac{d}{\sqrt{n}}\right) \qquad \square$$

What if we only cared about learning the "shape", not learning in TV dist ?

<u>Task</u>: Given $X_1 \dots X_n \sim N(0,\Sigma)$, find $\hat{\Sigma}$ s.t.

$$\|\Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2} - I\|_{op} \leq \varepsilon$$

<u>Reminder</u>: $\|M\|_{op} = \max_{v} \frac{v^T M v}{\|v\|^2}$. So,

$$\forall v, \quad v^T \hat{\Sigma} v = (1 \pm \varepsilon) v^T \Sigma v$$

- simultaneously estimate the variance in every direction.
- good for e.g. PCA.

<u>Theorem</u>: $n = O\left(\frac{d}{\varepsilon^2}\right)$ samples suffice — compare to $\frac{d^2}{\varepsilon^2}$ for T.V.

Interlude: Matrix Concentration

- n copies of a $\mathbb{R}$-valued random variable X —
  how close is $\frac{1}{n}\sum X_i$ to $\mathbb{E}X$?

$$\left[\mathbb{E}\left(\frac{1}{n}\sum X_i - \mathbb{E}X\right)^2\right]^{1/2} = \sqrt{\frac{Var(X)}{n}}$$

If X is a little bit "nice" (bdd, subGaussian, etc.), get concentration — $\frac{1}{n}\sum X_i$ acts like Gaussian

$$\Pr\left(\left|\frac{1}{n}\sum X_i - \mathbb{E}X\right| > t\right) \lesssim e^{-t^2 n / Var(X)}$$

- n copies of a $\mathbb{R}^d$-valued random vector X:

$$\left[\mathbb{E}\left\|\frac{1}{n}\sum X_i - \mathbb{E}X\right\|^2\right]^{1/2} = \sqrt{\frac{\mathbb{E}\|X-\mathbb{E}X\|^2}{n}} = \sqrt{\frac{\operatorname{Tr}\operatorname{Cov}(X)}{n}}$$

- n copies of a $\mathbb{R}^{d\times d}$ random matrix X (symmetric)?

$$\mathbb{E}\left\|\frac{1}{n}\sum X_i - \mathbb{E}X\right\|_{op} \leq ??$$

- What should we expect? Suppose $X = \begin{pmatrix} X^{(1)} & & 0 \\ & \ddots & \\ 0 & & X^{(d)} \end{pmatrix}$ is diagonal.

$$\frac{1}{n}\sum X_i = \begin{pmatrix} \frac{1}{n}\sum X_i^{(1)} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{n}\sum X_i^{(d)} \end{pmatrix} \text{ has}$$

Singular vals $\frac{1}{n}\sum X_i^{(j)}$.

For each $j \leq d$, expect

$$\Pr\left(\left|\frac{1}{n}\sum X_i^{(j)} - \mathbb{E}X^{(j)}\right| > t\right) \lesssim e^{-t^2 n/\operatorname{Var}(X^{(j)})} \lesssim \frac{1}{100d}$$

if $t = \Theta\left(\sqrt{\frac{\operatorname{Var}(X^{(j)})\cdot \log d}{n}}\right)$

By union bound, wp $\gtrsim 0.99$,

$$\left\|\frac{1}{n}\sum X_i - \mathbb{E}X\right\| \le O\left(\max_j \sqrt{\mathrm{Var}(X^{(j)})} \cdot \sqrt{\frac{\log d}{n}}\right)$$

Is something like this true in general, ie even for non-diagonal X?

Matrix Bernstein Inequality: Let X be d×d random matrix, $\mathbb{E}X = 0$, and $\|X\| \le R$ w.p. 1, $X_1 \ldots X_n$ indep. copies.

$$\mathbb{E}\left\|\frac{1}{n}\sum X_i\right\|_{op} \le O\left(\frac{\|\mathbb{E}XX^T\| + \|\mathbb{E}X^TX\|}{\sqrt{n}} \cdot \sqrt{\log d} + \frac{R\log d}{n}\right)$$

Application to prove thm, up to $\log d$ factors:

WLOG, $\Sigma = I$, and goal is to show

$$\mathbb{E} \left\| \frac{1}{n} \sum x_i x_i^T - I \right\|_{op} \leq O\left(\sqrt{\frac{d \log d}{n}}\right).$$

$\mathbb{E}[X_i X_i^T - I] = 0$, but don't have $\| X_i X_i^T - I \|_{op} \leq R$ wp 1.

But let's pretend... $\| X_i X_i^T - I \|_{op} \leq O\left(\frac{d}{n}\right)$ wp 1 (almost true).

Then, just need to calculate $\| \mathbb{E}(XX^T - I)(XX^T - I) \|$

$$\leq \| \mathbb{E} \|X\|^2 XX^T \| + O(1) \leq O(d)$$

which proves the theorem $\square$