

# Algorithmic Statistics

## Lecture 4: Learning a Gaussian

Samuel B. Hopkins

In lecture 2, we saw that learning high-dimensional probability distributions without some assumptions requires a number of samples growing exponentially with dimension. Today we will see how adding assumptions can bring the sample complexity down to a polynomial in dimension.

We start with a reminder about multivariate Gaussians.

**Definition 0.1** (Multivariate Gaussian). A distribution on  $\mathbb{R}^d$  is  $\mathcal{N}(\mu, \Sigma)$  if it has density

$$p_{\mu, \Sigma}(x) = \frac{1}{(2\pi)^{d/2}(\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right),$$

where  $\mu \in \mathbb{R}^d$  and  $\Sigma \in \mathbb{R}^{d \times d}$  is PSD.

Basic facts:

- $\mathbb{E}[X] = \mu$  and  $\text{Cov}(X) = \Sigma$  when  $X \sim \mathcal{N}(\mu, \Sigma)$ .
- Every Gaussian distribution is an affine transformation of the standard Gaussian  $\mathcal{N}(0, I)$ . That is, if  $Z \sim \mathcal{N}(0, I)$  and  $X = \Sigma^{1/2}Z + \mu$ , then  $X \sim \mathcal{N}(\mu, \Sigma)$ .
- The central limit theorem says that if you add a bunch of independent random variables, the resulting random variable has an approximately Gaussian distribution. (We won't need a formal statement.)

*Main question today:* Suppose you are willing to assume that the population from which you're drawing samples is Gaussian. How many samples do you need to learn the distribution?

Such an assumption might be reasonable if you believe that the samples you're observing are well described by adding a bunch of independent contributions.

**Theorem 0.2.** Suppose  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \Sigma)$  are iid from a  $d$ -dimensional Gaussian. Then

$$\mathbb{E} \text{TV}(\mathcal{N}(\hat{\mu}, \hat{\Sigma}), \mathcal{N}(\mu, \Sigma)) \leq O\left(\frac{d}{\sqrt{n}}\right),$$

where  $\hat{\mu} = \frac{1}{n} \sum_{i \leq n} X_i$  and  $\hat{\Sigma} = \frac{1}{n} \sum_{i \leq n} (X_i - \hat{\mu})(X_i - \hat{\mu})^\top$ . Consequently,  $O(d/\varepsilon^2)$  samples are sufficient to learn a  $d$ -dimensional Gaussian to total variation distance  $\varepsilon$ .

How to prove the theorem? We will first prove that it would be enough to learning "parameter distance". The "parameters" of a Gaussian are the mean and covariance  $(\mu, \Sigma)$ , and learning in parameter distance means that we find  $\hat{\mu} \approx \mu$  and  $\hat{\Sigma} \approx \Sigma$ . Here the choice of norm we use to capture " $\approx$ " will be quite important! Then we will show that learning in parameter distance is possible with  $O(d^2/\varepsilon)$  samples. More formally, we will prove the following three lemmas, which immediately imply the theorem.

**Lemma 0.3.** For any  $\mu, \nu, \Sigma, \Gamma$ , we have  $\text{TV}(\mathcal{N}(\nu, \Gamma), \mathcal{N}(\mu, \Sigma)) \leq O(\|\Sigma^{-1/2}(\nu - \mu)\|) + O(\|I - \Sigma^{-1/2}\Gamma\Sigma^{-1/2}\|_F)$ .

**Lemma 0.4.**  $\mathbb{E} \|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \leq O(\sqrt{d/n})$ .

**Lemma 0.5.**  $\mathbb{E} \|I - \Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}\|_F \leq O(d/\sqrt{n})$ .

As a reminder, the Frobenius norm of a matrix  $M$ , denoted  $\|M\|_F$ , is the entry-wise Euclidean norm. And,  $\|M\|_F = \sqrt{\sum \sigma_i^2}$  where the  $\sigma$ s are the singular values of  $M$ .

## 1 KL divergence between Gaussians

**Proposition 1.1** (KL formula for Gaussians). For  $P = \mathcal{N}(\mu, \Sigma)$  and  $Q = \mathcal{N}(\nu, \Gamma)$ ,

$$\text{KL}(P\|Q) = \frac{1}{2} \left( (\mu - \nu)^\top \Gamma^{-1} (\mu - \nu) + \text{tr}(\Gamma^{-1} \Sigma) - \log \det(\Gamma^{-1} \Sigma) - d \right).$$

*Derivation.* By definition,  $\text{KL}(P\|Q) = \mathbb{E}_{X \sim P} \left[ \log \frac{p_{\mu, \Sigma}(X)}{p_{\nu, \Gamma}(X)} \right]$ . Taking logs of the densities and simplifying the quadratic forms yields

$$\text{KL}(P\|Q) = \frac{1}{2} \mathbb{E}_P \left[ (X - \nu)^\top \Gamma^{-1} (X - \nu) - (X - \mu)^\top \Sigma^{-1} (X - \mu) \right] + \frac{1}{2} \log \frac{\det \Gamma}{\det \Sigma}.$$

We have  $\mathbb{E}_P(X - \mu)^\top \Sigma^{-1} (X - \mu) = \text{Tr} \Sigma \Sigma^{-1} = d$ . For the other quadratic term, we can use

$$\begin{aligned} \mathbb{E}_P(X - \nu)(X - \nu)^\top &= \mathbb{E}_P(X - \mu + (\mu - \nu))(X - \mu + (\mu - \nu))^\top \\ &= \Sigma + (\mu - \nu)(\mu - \nu)^\top \end{aligned}$$

and hence that

$$\mathbb{E}_P(X - \nu)^\top \Gamma^{-1} (X - \nu) = \text{Tr}(\Sigma \Gamma^{-1} + (\mu - \nu) \Gamma^{-1} (\mu - \nu)).$$

Finally, for the  $\log(\det \Gamma / \det \Sigma)$  term, we can simplify to  $-\log \det \Gamma^{-1} \Sigma$  by determinant rules.  $\square$

Two immediate corollaries of Proposition 1.1 that we will use:

$$\text{(Same covariance)} \quad \text{KL}(\mathcal{N}(\mu, \Sigma) \parallel \mathcal{N}(\nu, \Sigma)) = \frac{1}{2} \|\mu - \nu\|_{\Sigma^{-1}}^2. \quad (1.1)$$

$$\text{(Same mean)} \quad \text{KL}(\mathcal{N}(\mu, \Sigma) \parallel \mathcal{N}(\mu, \Gamma)) = \frac{1}{2} (\text{tr}(A) - \log \det A - d), \quad A := \Gamma^{-1/2} \Sigma \Gamma^{-1/2}. \quad (1.2)$$

## 2 Proof of Lemma 0.3

Let  $A = \Sigma^{-1/2} \Gamma \Sigma^{-1/2}$ . If any eigenvalue of  $I - A$  is  $\Omega(1)$ , then we are done, since in this case  $\|I - \Sigma^{-1/2} \Gamma \Sigma^{-1/2}\|_F = \Omega(1)$  while  $\text{TV}(\cdot, \cdot) \leq 1$ .

Recall that for a symmetric matrix  $A$  with eigenvalues  $\lambda_1, \dots, \lambda_d$ ,  $\text{Tr} A = \sum_{i \leq d} \lambda_i$ , and  $\det A = \prod_{i \leq d} \lambda_i$ , so  $\log \det A = \sum_{i \leq d} \log \lambda_i$ . Hence,

$$\text{KL}(\mathcal{N}(\mu, \Sigma) \parallel \mathcal{N}(\mu, \Gamma)) = \frac{1}{2} (\text{tr}(A) - \log \det A - d) = \frac{1}{2} \left( \sum_{i \leq d} \lambda_i - 1 - \log \lambda_i \right)$$

where  $\lambda_1, \dots, \lambda_d$  are the eigenvalues of  $\Gamma^{-1/2} \Sigma \Gamma^{-1/2}$ . Now,

$$(x - 1) - \log x = (x - 1) - \log(1 + (x - 1)).$$

Since the first term in the Taylor expansion of  $\log(1 + (x - 1))$  is  $(x - 1)$ , we should expect this difference to act like the quadratic  $(x - 1)^2$ . Since we get to assume that the eigenvalues of  $A$  are smaller than some universal constant, we can therefore assume that for each  $\lambda_i$ , we have  $(x - 1) - \log(x) \leq O(x - 1)^2$ , meaning that

$$\sum_{i \leq d} \lambda_i - 1 - \log \lambda_i \leq O(\|I - A\|_F^2).$$

Now the lemma follows from Pinsker's inequality.

### 3 Estimators and parameter concentration

Given i.i.d.  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \Sigma)$ , define the empirical mean  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$  and empirical covariance

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu})(X_i - \hat{\mu})^\top.$$

It is convenient to “whiten”: let  $Z_i = \Sigma^{-1/2}(X_i - \mu) \sim \mathcal{N}(0, I)$ . Then

$$\Sigma^{-1/2}(\hat{\mu} - \mu) = \frac{1}{n} \sum_{i=1}^n Z_i, \quad \Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2} - I = \frac{1}{n} \sum_{i=1}^n (Z_i Z_i^\top - I) + \Sigma^{-1/2}(\mu - \hat{\mu})(\mu - \hat{\mu})^\top \Sigma^{-1/2}.$$

*Proof of Lemma 0.4.*  $\frac{1}{n} \sum_{i=1}^n Z_i$  has coordinates distributed as  $\mathcal{N}(0, \frac{1}{n})$ . So

$$\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n Z_i \right\| \leq (\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n Z_i \right\|^2)^{1/2} \leq \sqrt{d/n}. \quad \square$$

*Proof of Lemma 0.5.* For any rank-one matrix  $vv^\top$ , we have  $\|vv^\top\|_F = \|v\|^2$ , so

$$\mathbb{E} \|\Sigma^{-1/2}(\mu - \hat{\mu})(\mu - \hat{\mu})^\top \Sigma^{-1/2}\|_F = \mathbb{E} \|\Sigma^{-1/2}(\hat{\mu} - \mu)\|^2 \leq d/n.$$

We can do an explicit calculation for the other term,  $\frac{1}{n} \sum_{i=1}^n (Z_i Z_i^\top - I)$ . First consider an off-diagonal entry of the matrix. Let  $\alpha_1, \dots, \alpha_n, \beta_1, \dots, \beta_n$  all be iid from  $\mathcal{N}(0, 1)$ . Then an off diagonal entry is distributed exactly like  $\frac{1}{n} \sum_{i=1}^n \alpha_i \beta_i$ . So

$$\mathbb{E} \left( \frac{1}{n} \sum_{i=1}^n \alpha_i \beta_i \right)^2 = \frac{1}{n^2} \sum_{i,j} \mathbb{E} \alpha_i \alpha_j \beta_i \beta_j = \frac{1}{n},$$

since only the  $i = j$  terms in the sum contribute. An on-diagonal entry is distributed as  $\frac{1}{n} \sum_{i=1}^n (\alpha_i^2 - 1)$ . So

$$\mathbb{E} \left( \frac{1}{n} \sum_{i=1}^n (\alpha_i^2 - 1) \right)^2 = \frac{1}{n^2} \sum_{i,j} \mathbb{E} (\alpha_i^2 - 1)(\alpha_j^2 - 1) \leq O(1/n).$$

Putting together the diagonal and off-diagonal cases, we have  $\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n (Z_i Z_i^\top - I) \right\|_F^2 = O(d^2/n)$ , which completes the proof.  $\square$

As a remark, the essentially the same rates of estimation error for the parameters would hold if we weakened the assumptions on the underlying distribution significantly – it would be enough, for instance, for the distribution to be subgaussian. (Even much weaker assumptions are enough.) But for such a broad class of distributions, we would not be able to relate parameter distance to total variation.

## 4 Matrix Concentration: Learning the Covariance in Spectral Order

In the foregoing, we saw how to learn the covariance  $\Sigma$  of a Gaussian in Frobenius norm using  $\approx d^2$  samples. What if we found a weaker error norm acceptable? Here I want to illustrate that the sample complexity of parameter learning is heavily dependent on the way that errors are measured.

Recall that for matrices  $A, B$  we write  $A \leq B$  if  $B - A$  is PSD. What would it mean if given samples  $X_1, \dots, X_n \sim \mathcal{N}(0, \Sigma)$ , we learned a matrix  $\hat{\Sigma}$  such that  $(1 - \varepsilon)\Sigma \leq \hat{\Sigma} \leq (1 + \varepsilon)\Sigma$ ? One nice interpretation is that  $\hat{\Sigma}$  would tell us an accurate estimate of every one-directional variance – for every unit  $v$ , we would have  $v^\top \hat{\Sigma} v = (1 \pm \varepsilon) \mathbb{E}\langle X, v \rangle^2$ . (Learning in Frobenius norm corresponds to a stronger guarantee about all degree-2 polynomials of  $X$ , rather than just those of the form  $\langle X, v \rangle^2$ , although this is a little more complicated to state – it’s not as simple as learning  $(1 \pm \varepsilon) \mathbb{E} p(X)$ .)

Note that the minimum  $\varepsilon$  such that  $(1 - \varepsilon)\Sigma \leq \hat{\Sigma} \leq (1 + \varepsilon)\Sigma$  is equivalently  $\|I - \Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2}\|$ , where  $\|\cdot\|$  denotes operator norm, or maximum-magnitude eigenvalue, of a matrix.

**Theorem 4.1.** *Let  $X_1, \dots, X_n \sim \mathcal{N}(0, \Sigma)$  be iid  $d$ -dimensional Gaussians. Then  $\mathbb{E} \|I - \Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2}\| \leq O(\sqrt{d/n} + d/n)$ , where  $\hat{\Sigma}$  is the empirical covariance.*

There are many ways to prove this theorem. We will take the opportunity to introduce a very useful general-purpose tool, the *Matrix Bernstein inequality*, which will allow us to prove a slightly weaker statement (losing a  $\log d$  factor).

### 4.1 Reminder about scalar concentration of measure

Let  $\alpha_1, \dots, \alpha_n$  be independent random variables taking values in  $[-1, 1]$  with  $\mathbb{E} \alpha_i = 0$ . Then standard measure concentration tells us that  $\sum \alpha_i$  acts like a Gaussian with variance  $\sigma^2 = \sum \alpha_i^2$ , in the sense that for every  $t > 0$ ,

$$\Pr(|\sum \alpha_i| > t) \leq \exp(-\Omega(t^2/(\sigma^2 + t))).$$

The form of this inequality is interesting – a Gaussian would be missing the additive  $+t$  in the denominator of the fraction  $t^2/(\sigma^2 + t)$ . The usual interpretation here is that  $\sum \alpha_i$  acts Gaussian for  $t = O(\sigma^2)$ , but for larger  $t$  the sum acts sub-exponential instead. If instead of  $\alpha_i \in [-1, 1]$  w.p. 1 we know  $\alpha_i \in [-R, R]$  w.p. 1 for some other number  $R$ , the same inequality holds with  $\sigma^2 + Rt$  in place of  $\sigma^2 + t$ ; this is known as Bernstein’s inequality.

### 4.2 Diagonal matrices

We want to work up to the setting that  $\alpha_1, \dots, \alpha_n$  are each  $d \times d$  random matrices, so that we eventually take  $\alpha_i = Z_i Z_i^\top - I$ . On the way there, let’s imagine that  $\alpha_1, \dots, \alpha_n$  are  $d \times d$  diagonal random matrices with  $\mathbb{E} \alpha_i = 0$  and  $\|\alpha_i\| \leq 1$  w.p. 1. Then  $\sum \alpha_i$  is also a diagonal matrix, and  $\|\sum \alpha_i\|$  is the maximum of  $d$  random variables, each obeying Bernstein’s inequality. Then using a union bound across all  $d$  of these random variables, we obtain

$$\Pr(\|\sum \alpha_i\| > t) \leq d \exp(-\Omega(t^2/(\sigma^2 + 1))).$$

### 4.3 Matrix Bernstein

Remarkably, it turns out that the same inequality holds even without the diagonal assumption on the summands.

**Theorem 4.2** (Matrix Bernstein inequality). *Let  $\alpha_1, \dots, \alpha_n$  be independent  $d \times d$  random matrices with  $\|\alpha_i\| \leq R$  a.s.. And let  $\sigma^2 = \max(\|\mathbb{E} \sum \alpha_i \alpha_i^\top\|, \|\mathbb{E} \sum \alpha_i^\top \alpha_i\|)$ . Then for every  $t > 0$ ,*

$$\Pr(\|\sum \alpha_i\| > t) \leq d \exp(-\Omega(t^2/(\sigma^2 + Rt))).$$

Furthermore,  $\mathbb{E} \|\sum \alpha_i\| \leq O(\sigma \sqrt{\log d} + R \log d)$ .

E

### 4.4 Analysis of the empirical covariance

Now we are equipped to analyze the empirical covariance  $\hat{\Sigma}$ . We want to analyze  $\|\Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1} - I\|$ . It will be enough to analyze  $\|\frac{1}{n} \sum (Z_i Z_i^\top - I)\|$ . We would like to take  $\alpha_i = Z_i Z_i^\top - I$  and apply Matrix Bernstein. There's one issue – there is no almost-sure upper bound on  $\|Z_i Z_i^\top - I\|$ .

There's a standard trick here – we impose the almost-sure bound by fiat. Let

$$\alpha_i = (Z_i Z_i^\top - I) \cdot 1(\|Z_i\| \leq 10\sqrt{d}) - \mathbb{E}(Z_i Z_i^\top - I) \cdot 1(\|Z_i\| \leq 10\sqrt{d}).$$

Then  $\mathbb{E} \alpha_i = 0$  and  $\|\alpha_i\| \leq O(d)$  a.s.. To apply Matrix Bernstein, we need to calculate (an upper bound on)  $\sigma^2 = \sum \mathbb{E} \alpha_i^2$ . (Each  $\alpha_i$  is symmetric so the two terms in the max in the definition of  $\sigma^2$  are identical.)

Since the  $\alpha_i$ s are iid, it's enough to look at just one of them; let  $\alpha = (ZZ^\top - I) \cdot 1(\|Z\| \leq 10\sqrt{d}) - \mathbb{E}(ZZ^\top - I) \cdot 1(\|Z\| \leq 10\sqrt{d})$  for a standard Gaussian  $Z$ . Then  $\mathbb{E} \alpha^2 \leq O(1) \mathbb{E}(ZZ^\top - I)^2 1(\|Z\| \leq 10\sqrt{d}) \leq O(d)I$ . So  $\sigma^2 \leq O(nd)$ , and we get  $\mathbb{E} \|\sum \alpha_i\| \leq O(d \log d + \sqrt{nd \log d})$ ; when we divide by  $n$  we get  $O((d \log d)/n + \sqrt{d \log d/n})$ .

The only missing piece now is that  $\sum \alpha_i$  is not exactly  $\frac{1}{n} \sum Z_i Z_i^\top - I$ , because we introduced the indicators  $1(\|Z_i\| \leq 10\sqrt{d})$ . So we need to analyze

$$\mathbb{E} \|\sum \alpha_i - (Z_i Z_i^\top - I)\|.$$

It's enough to analyze  $\mathbb{E} \|\sum (Z_i Z_i^\top - I) 1(\|Z_i\| > 10\sqrt{d})\|$ . Sloppy bounds are enough here. By triangle inequality, this is at most  $n \mathbb{E}(1 + \|Z\|^2) 1(\|Z\| > 10\sqrt{d})$ . By Cauchy-Schwarz, this is at most  $O(nd) \sqrt{\Pr(\|Z\| > 10\sqrt{d})}$ . Standard Gaussian concentration shows that this probability is  $\exp(-\Omega(d))$ .

So, we got  $\mathbb{E} \|\Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2} - I\| \leq \tilde{O}(d/n + \sqrt{d/n} + nd \exp(-\Omega(d)))$ . (The  $\tilde{O}$  notation hides logs.) We can get slightly different balance between the three terms if we change the threshold  $10\sqrt{d}$  to something else, e.g.  $\sqrt{d} \log n$ .