COMS 4721: Review of prerequisites

Daniel Hsu

1 Euclidean space

The $d$-dimensional Euclidean space $\mathbb{R}^d$ is the $d$-dimensional vector space over the real numbers $\mathbb{R}$ where we have the familiar notions of distances and angles from classical plane geometry. The space $\mathbb{R}^d$ is comprised of vectors (or points), which can be added ($z := x + y$ for $x, y \in \mathbb{R}^d$) and scaled by real numbers ($z := cx$ for $c \in \mathbb{R}$ and $x \in \mathbb{R}^d$) to obtain other vectors in $\mathbb{R}^d$.

Every vector $x \in \mathbb{R}^d$ has a (Euclidean) length (or norm; also called the $l_2$ norm), which is denoted by $\|x\|_2$. The length of the scaled vector $cx$ for $c \in \mathbb{R}$ and $x \in \mathbb{R}^d$ is $\|cx\|_2 = |c|\|x\|_2$. Vectors with length one are called unit vectors, and there is a unique vector of length zero which is the zero vector (or origin) $0$.

The Euclidean norm comes from the inner product (or dot product) $\langle x, y \rangle$ between vectors $x, y \in \mathbb{R}^d$, which is defined to be the product of (i) the length of $x$, (ii) the length of $y$, and (iii) the cosine of the angle between $x$ and $y$. (Sometimes we will also write the inner product as $x^\top y$.) Since the cosine of the angle between $x$ and itself is 1, we have $\langle x, x \rangle = \|x\|_2^2$. We say $x$ and $y$ are orthogonal if $\langle x, y \rangle = 0$. Using arguments from plane geometry, it can be shown that the inner product is symmetric ($\langle x, y \rangle = \langle y, x \rangle$) and linear in its first argument (so $\langle cx + y, z \rangle = c\langle x, z \rangle + \langle y, z \rangle$).

The distance dist$(x, y)$ between vectors $x$ and $y$ is measured by the Euclidean norm of their difference, $\|x - y\|_2$. This is a metric (called the Euclidean metric or $l_2$ metric): for all $x, y, z \in \mathbb{R}^d$:

- $\|x - y\|_2 \geq 0$, and $\|x - y\|_2 = 0$ if and only if $x = y$;
- $\|x - y\|_2 = \|y - x\|_2$ (symmetry);
- $\|x - y\|_2 \leq \|x - z\|_2 + \|y - z\|_2$ (triangle inequality).

The triangle inequality is equivalent to $\|x + y\|_2 \leq \|x\|_2 + \|y\|_2$. To see why it holds, observe that $\|x+y\|_2^2 = \langle x+y, x+y \rangle = \|x\|_2^2 + 2\langle x, y \rangle + \|y\|_2^2 \leq \|x\|_2^2 + 2\|x\|_2\|y\|_2 + \|y\|_2^2 = (\|x\|_2 + \|y\|_2)^2$.

The inequality follows from the Cauchy-Schwarz inequality (i.e., $\langle x, y \rangle \leq \|x\|_2\|y\|_2$), which just says that the cosine of an angle between vectors is always at most 1.

A collection of vectors $v_1, v_2, \ldots, v_k$ is linear dependent if there are scalars $c_1, c_2, \ldots, c_k \in \mathbb{R}$, not all zero, such that $c_1 v_1 + c_2 v_2 + \cdots + c_k v_k = 0$. If there no such scalars, then the vectors are linearly independent. An ordered collection of $d$ linearly independent vectors in $\mathbb{R}^d$ is called a basis. A collection of vectors is orthogonal if every pair is orthogonal; and a collection of vectors is orthonormal if they are all of unit length and are orthogonal. Finally, an orthonormal basis for $\mathbb{R}^d$ is a basis comprised of orthonormal vectors.

It will be convenient to refer to a canonical orthonormal basis which we call the coordinate basis $e_1, e_2, \ldots, e_d$. This lets us specify Cartesian coordinates for every vector $x \in \mathbb{R}^d$: that is, we uniquely identify $x$ with the $d$-tuple $x = (x_1, x_2, \ldots, x_d)$ where

$$x_i := \langle e_i, x \rangle \text{ for each } i \in [d] := \{1, 2, \ldots, d\}.$$
In this notation, $e_i$ is the vector whose $i$-th coordinate is 1 and all other coordinates are 0. We can also write the (squared) Euclidean norm of a vector $x \in \mathbb{R}^d$ as $\|x\|^2 = x_1^2 + x_2^2 + \cdots + x_d^2$ (a generalization of the Pythagorean formula), and the inner product between vectors $x, y \in \mathbb{R}^d$ as $\langle x, y \rangle = x_1y_1 + x_2y_2 + \cdots + x_dy_d$.

## 2 Probability spaces and random variables

Let $\Omega$ be a sample space, and $\mathbb{P}$ be a probability distribution over $\Omega$—together, they make up a probability space $(\Omega, \mathbb{P})$. An event $A$ is a subset of $\Omega$, and the probability of $A$ is $\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\omega)$.

A random variable defined on $(\Omega, \mathbb{P})$ is a real-valued function $X: \Omega \to \mathbb{R}$. We’ll also use the notation $X \sim \mathbb{P}$ to declare the random variable; often, we’ll even leave the probability space implicit. The expected value (a.k.a. expectation, mean) of $X$ is the average value that $X$ takes on:

$$\mathbb{E}(X) := \sum_{\omega \in \Omega} X(\omega) \cdot \mathbb{P}(\omega).$$

We’ll often use the more intuitive notation

$$\mathbb{E}(X) = \sum_x x \cdot \mathbb{P}(X = x),$$

where the summation is over the range of $X$. Here, $\mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\})$.

The notation above implicitly assumes that range of $X$ is a discrete set that we can enumerate. But we’ll often encounter a continuous random variable $X$ whose range is a continuous space, like the real line $\mathbb{R}$ or some subset thereof. In the cases we’ll consider, the random variable $X$ will have a probability density function $p: \mathbb{R} \to \mathbb{R}_+$, a non-negative real-valued function on $\mathbb{R}$ satisfying

$$\int_{\mathbb{R}} p(x) \, dx = 1,$$

where for any interval $I \subseteq \mathbb{R}$,

$$\mathbb{P}(X \in I) = \int_I p(x) \, dx.$$

The expected value of $X$ is given by

$$\mathbb{E}(X) = \int_{\mathbb{R}} x \cdot p(x) \, dx.$$

When we have several continuous random variables $X_1, X_2, \ldots, X_d$, they will have a joint probability density function $p: \mathbb{R}^d \to \mathbb{R}_+$ that satisfies

$$\int_{\mathbb{R}^d} p(x) \, dx = \int_{\mathbb{R}^d} p(x_1, x_2, \ldots, x_d) \, dx_1 \, dx_2 \cdots dx_d = 1.$$

Each $X_i$ has a marginal probability density function $p_i: \mathbb{R} \to \mathbb{R}$:

$$p_i(x_i) := \int_{\mathbb{R}^{d-1}} p(x_1, x_2, \ldots, x_d) \, dx_1 \cdots dx_{i-1} \, dx_{i+1} \cdots dx_d,$$

a probability density function in its own right. We’ll collect these random variables into a vector $X = (X_1, X_2, \ldots, X_d)$. The expected value of $X$ is a vector $\mathbb{E}(X) \in \mathbb{R}^d$ whose $i$-th entry is

$$\mathbb{E}(X_i) = \int_{\mathbb{R}} x_i \cdot p_i(x_i) \, dx_i.$$

The random variables $X_1, X_2, \ldots, X_d$ are independent if $p(x_1, x_2, \ldots, x_d) = p_1(x_1)p_2(x_2)\cdots p_d(x_d)$ for all $(x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$. 

2
3 Linearity of expectation

Suppose $X$ and $Y$ are random variables. Then for any scalar $a \geq 0$,

$$\mathbb{E}(aX + Y) = a \cdot \mathbb{E}(X) + \mathbb{E}(Y),$$

regardless of whether $X$ and $Y$ are independent or not. This fact, called linearity of expectation, is very powerful. We are likely to encounter collections of random variables $X_1, X_2, \ldots, X_d$ with all sorts of dependencies between them. In any of these cases, the expected value of any linear combination of random variables is the corresponding linear combination of their expected values:

$$a_1 X_1 + a_2 X_2 + \cdots + a_d X_d = a_1 \mathbb{E}(X_1) + a_2 \mathbb{E}(X_2) + \cdots + a_d \mathbb{E}(X_d).$$

4 Variance

The mean of $X$ tells us what value $X$ takes on average. But we are also interested in how far $X$ is from its mean, on average. This is captured by the variance of $X$,

$$\text{var}(X) := \mathbb{E}(X - \mathbb{E}(X))^2.$$

We look at $\mathbb{E}(X - \mathbb{E}(X))^2$ rather than, say, $\mathbb{E}|X - \mathbb{E}(X)|$ simply for mathematical convenience. Unfortunately, the “units” of $X$ and $\text{var}(X)$ are not the same: if $X$ is measured in “meters”, then $\text{var}(X)$ is measured in “square meters”. Therefore, we will often also look at the square-root of variance $\sqrt{\text{var}(X)}$, which is called the standard deviation.

4.1 Jensen’s inequality

Another (easily checked) formula for variance is $\text{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$. And since the variance of a random variable is always non-negative, we have the inequality

$$(\mathbb{E}(X))^2 \leq \mathbb{E}(X^2).$$

This inequality is actually a special case of a more general inequality, called Jensen’s inequality, which states that for any convex function $f : \mathbb{R} \to \mathbb{R}$, we have $f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$. A function $f$ is convex if

$$f(px + (1-p)y) \leq pf(x) + (1-p)f(y)$$

for any $p \in [0,1]$ and $x, y \in \mathbb{R}$. The above inequality is the special case where $f(x) = x^2$, which is convex.

Applying the inequality to the random variable $|X - \mathbb{E}(X)|$, we obtain the convenient fact

$$\mathbb{E}|X - \mathbb{E}(X)| \leq \sqrt{\mathbb{E}(X - \mathbb{E}(X))^2} = \sqrt{\text{var}(X)}.$$

4.2 Variance of linear combinations

If $X$ and $Y$ are random variables, then for any scalar $a$,

$$\text{var}(aX + Y) = a^2 \text{var}(X) + \text{var}(Y) + 2a \text{cov}(X,Y)$$

where

$$\text{cov}(X,Y) := \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$$

is the covariance between $X$ and $Y$, which need not be zero. However, if $X$ and $Y$ are independent, then

$$\text{cov}(X,Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(X - \mathbb{E}(X))\mathbb{E}(Y - \mathbb{E}(Y)) = 0,$$
upon which we have \( \text{var}(aX + Y) = a^2 \text{var}(X) + \text{var}(Y) \). More generally, if \( X_1, X_2, \ldots, X_d \) are independent random variables, then
\[
\text{var}(a_1X_1 + a_2X_2 + \cdots + a_dX_d) = a_1^2 \text{var}(X_1) + a_2^2 \text{var}(X_2) + \cdots + a_d^2 \text{var}(X_d).
\]
(Actually, this holds even if we just have pairwise independence—i.e., any two \( X_i \) and \( X_j \) are independent.)

4.3 Example: symmetric random walk

A symmetric random walk on \( \mathbb{Z} \) is a stochastic process \((S_t)_{t \in \mathbb{Z}_+}\). Here, \( S_0 := 0 \), and for each time \( t \geq 1 \), \( S_t := S_{t-1} + X_t \), where \( \mathbb{P}(X_t = -1) = \mathbb{P}(X_t = 1) = 1/2 \). Clearly \( S_n = \sum_{t=1}^{n} X_t \). Each \( X_t \) has mean \( \mathbb{E}(X_t) = 0 \), and variance \( \text{var}(X_t) = 1 \). By linearity of expectation, \( \mathbb{E}(S_n) = \sum_{t=1}^{n} \mathbb{E}(X_t) = 0 \).

We assume \( \{X_t : t \in \mathbb{N}\} \) are independent, so we have that \( \text{var}(S_n) = \sum_{t=1}^{n} \text{var}(X_t) = n \), which in turn implies
\[
\mathbb{E}|S_n| \leq \sqrt{\text{var}(S_n)} = \sqrt{n}.
\]

5 Tail bounds from moments

In a symmetric random walk, the position after \( n \) steps is, in expectation, within \( \sqrt{n} \) of the origin. Can something similar be said about any particular realization of the random walk? To answer this question, we appeal to Markov’s inequality.

**Theorem 1** (Markov’s inequality). For any \( t \geq 0 \), \( \mathbb{P}(|X| \geq t) \leq \mathbb{E}|X|/t \).

**Proof.** For any \( t \geq 0 \),
\[
|X| \geq \mathbb{1}\{|X| \geq t\} \cdot t.
\]
(Here, \( \mathbb{1}\{\phi\} := 1 \) if \( \phi \) is true, and \( \mathbb{1}\{\phi\} := 0 \) otherwise.) Take expectations of both sides to obtain
\[
\mathbb{E}|X| \geq \mathbb{E}(\mathbb{1}\{|X| \geq t\}) \cdot t.
\]
Now interpret the right-hand side as \( \mathbb{P}(|X| \geq t) \cdot t \) and rearrange. \( \square \)

Suppose, for \( t > 0 \) and \( \delta \in (0, 1) \), that \( \mathbb{P}(X = 0) = 1 - \delta \) and \( \mathbb{P}(X = t) = \delta \). Then \( \mathbb{P}(X \geq t) = \delta = \mathbb{E}(X)/t \). So Markov’s inequality is tight at least in some cases.

For the symmetric random walk, let \( t = c\sqrt{n} \) for some \( c > 0 \). Applying Markov’s inequality, we obtain the inequality
\[
\mathbb{P}(|S_n| \geq c\sqrt{n}) \leq \frac{\mathbb{E}|S_n|}{c\sqrt{n}} \leq \frac{1}{c}.
\]
This means that for at least a \( 1-1/c \) fraction of the realizations, the position after \( n \) steps is within \( c\sqrt{n} \) of the origin.

This tail bound can sometimes be improved by a more effective use of Markov’s inequality. The essence of this idea is captured in Chebyshev’s inequality.

**Theorem 2** (Chebyshev’s inequality). For any \( t \geq 0 \), \( \mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \text{var}(X)/t^2 \).

**Proof.** Apply Markov’s inequality to the random variable \((X - \mathbb{E}(X))^2\). \( \square \)
Again with the symmetric random walk, applying Chebyshev’s inequality gives
\[
P(|S_n| \geq c \sqrt{n}) \leq \frac{\text{var}(S_n)}{c^2 n} = \frac{1}{c^2}.
\]
The improvement comes from exploiting the variance of $|S_n|$, whereas previously we just used its expected value. By exploiting higher-order moments ($\mathbb{E}(X^3)$, $\mathbb{E}(X^4)$, etc.), we may obtain further improvements.

5.1 Weak Law of Large Numbers

Using Chebyshev’s inequality, we can prove a simple version of the Law of Large Numbers (LLN). We prove the non-asymptotic form of the “Weak” LLN by applying Chebyshev’s inequality as applied to the average of $n$ independent and identically distributed (iid) random variables.

**Theorem 3** (Weak Law of Large Numbers). Let $X_1, X_2, \ldots, X_n$ be iid random variables. Let $\mu_n := (1/n) \sum_{i=1}^{n} X_i$, $\mu := \mathbb{E}(X_1)$, and $\sigma^2 := \text{var}(X_1)$. For any $\varepsilon > 0$, $P(|\mu_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon}$.

**Proof.** By linearity of expectation, $\mathbb{E}(\mu_n) = \mu$. Therefore, applying Chebyshev’s inequality and using the fact that $X_1, X_2, \ldots, X_n$ are iid, $P(|\mu_n - \mu| > \varepsilon) \leq \frac{\text{var}(\mu_n)}{\varepsilon} = \frac{\sigma^2}{n\varepsilon}$.

6 Eigenvalues and eigenvectors

A scalar $\lambda$ is an eigenvalue of a $d \times d$ matrix $M$ if there is a vector $v \neq 0$ such that $Mv = \lambda v$. This vector $v$ is the eigenvector corresponding to the eigenvalue $\lambda$. (Note that corresponding eigenvectors are not unique: if $v$ is an eigenvector corresponding to $\lambda$, then so is $cv$ for $c \neq 0$.)

If $M$ is regarded as a linear transformation, then it maps its eigenvectors $v$ into a scaling of $v$ itself (by $\lambda$). Linear transformations always have eigenvalues, though they may be complex-valued:

\[
\lambda \text{ is an eigenvalue of } M \\
\iff \text{ there exists } v \neq 0 \text{ such that } Mv = \lambda v \\
\iff \text{ there exists } v \neq 0 \text{ such that } (M - \lambda I)v = 0 \\
\iff M - \lambda I \text{ is singular (i.e., not invertible)} \\
\iff \det(M - \lambda I) = 0. 
\]

The function $\lambda \mapsto \det(M - \lambda I)$ is a degree-$d$ polynomial in $\lambda$, and hence it has $d$ roots (some roots may be repeated, and some may be complex).

An important special case is when $M \in \mathbb{R}^{d \times d}$ is symmetric (i.e., $M = M^\top$), because then all of its eigenvalues (and eigenvectors) are real.

**Theorem 4.** Pick any symmetric matrix $M \in \mathbb{R}^{d \times d}$. Then:

- $M$ has $d$ real eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_d \in \mathbb{R}$ (some may be repeated).
- There is a set of $d$ corresponding eigenvectors $v_1, v_2, \ldots, v_d \in \mathbb{R}^d$ that form an orthonormal basis for $\mathbb{R}^d$.
- $M = \sum_{i=1}^{d} \lambda_i v_i v_i^\top = V \Lambda V^\top$, where

\[
V := \begin{bmatrix}
v_1 & v_2 & \cdots & v_d
\end{bmatrix}, \quad \Lambda := \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d) = \begin{bmatrix}
\lambda_1 & & \\
& \lambda_2 & \\
& & \ddots \\
& & & \lambda_d
\end{bmatrix}.
\]
Proof. The proof of first two points can be found in most linear algebra texts. We just prove the last point to illustrate a proof technique. To establish equality of \( M \) and \( \sum_{i=1}^{d} \lambda_i v_i v_i^\top \), we just need to show that they have the same behavior on some set of \( d \) linearly independent vectors—we have the freedom to choose this set of vectors. The set of vectors we’ll use is \( v_1, v_2, \ldots, v_d \). Pick any \( j \in [d] \), and consider \( M v_j \) and \( (\sum_{i=1}^{d} \lambda_i v_i v_i^\top) v_j \). The former is \( \lambda_j v_j \) since \( v_j \) is the eigenvector corresponding to eigenvalue \( \lambda_j \); the latter is \( \sum_{i=1}^{d} \lambda_i (v_j, v_i) v_i = \lambda_j v_j \), where we use the fact that \( v_1, v_2, \ldots, v_d \) are orthonormal. \( \Box \)

We say a symmetric matrix \( M \in \mathbb{R}^{d \times d} \) is positive definite if for all non-zero vectors \( x \in \mathbb{R}^d \), \( x^\top M x > 0 \). If the inequality only holds with \( > \) replaced by \( \geq \), then we say \( M \) is positive semidefinite (psd).

**Lemma 1.** Let \( M \) be a symmetric \( d \times d \) matrix. \( M \) is positive definite (resp., positive semidefinite) if and only if every eigenvalue of \( M \) is positive (resp., non-negative).

**Proof.** Let the eigenvalues of \( M \) be \( \lambda_1, \lambda_2, \ldots, \lambda_d \in \mathbb{R} \), and the corresponding orthonormal eigenvectors be \( v_1, v_2, \ldots, v_d \in \mathbb{R}^d \). Suppose \( M \) is positive definite. Then for each \( i \in [d] \), \( v_i^\top M v_i > 0 \). But

\[
 v_i^\top M v_i = v_i^\top (\lambda_i v_i) = \lambda_i
\]

for each \( i \in [d] \), so every eigenvalue of \( M \) is positive. Now suppose instead that \( \lambda_i > 0 \) for all \( i \in [d] \). Then for any non-zero \( x \in \mathbb{R}^d \),

\[
 x^\top M x = x^\top \left( \sum_{i=1}^{d} \lambda_i v_i v_i^\top \right) x = \sum_{i=1}^{d} \lambda_i (v_i^\top x)^2 > 0.
\]

The proofs of the remaining claims (concerning positive semidefiniteness) are similar. \( \Box \)

## 7 Multivariate Gaussian distributions

If \( Z_1, Z_2, \ldots, Z_d \) are iid \( N(0, 1) \) random variables, then we say that the random vector \( Z = (Z_1, Z_2, \ldots, Z_d) \) has a **standard \( d \)-dimensional Gaussian distribution**, written as \( Z \sim N(0, I) \), which says that the mean of \( Z \) is 0, and the covariance is the identity matrix \( I \). Recall that the covariance matrix of a random vector \( X \) is

\[
 \text{cov}(X) = E[(X - E(X))(X - E(X))^\top]
\]

(so the \((i, j)\)-th entry of \( \text{cov}(X) \) is \( \text{cov}(X_i, X_j) \)). More generally, if \( \mu \in \mathbb{R}^d \) and \( \Sigma \) is a symmetric positive definite matrix, then \( X \sim N(\mu, \Sigma) \) means that \( X \) has a \( d \)-dimensional Gaussian density

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

so \( E(X) = \mu \) and \( \text{cov}(X) = \Sigma \). This is much like the standard \( d \)-dimensional Gaussian density, except centered at \( \mu \), and the level sets of equal density are ellipsoidal shells of the form

\[
 \{x \in \mathbb{R}^d : (x - \mu)^\top \Sigma^{-1} (x - \mu) = r^2 \}
\]

for some \( r \geq 0 \). Let \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0 \) be the eigenvalues of \( \Sigma \), and let the corresponding eigenvectors be \( v_1, v_2, \ldots, v_d \in \mathbb{R}^d \), which we assume form an orthonormal basis for \( \mathbb{R}^d \). The
eigenvectors define the principal axes of the ellipsoidal contours of equal density, and the eigenvalues give the variance of $X$ in the corresponding eigenvector direction:

$$\text{var}(\langle v_i, X \rangle) = \mathbb{E}(v_i, X - \mu)^2 = \mathbb{E} v_i^\top (X - \mu) (X - \mu)^\top v_i = v_i^\top \Sigma v_i = v_i^\top V \Lambda V^\top v_i = \lambda_i$$

(where $V = [v_1 | v_2 | \cdots | v_d]$ and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$).

Note that if we define $W_i := \langle v_i, X \rangle$ for each $i \in [d]$, then $W = (W_1, W_2, \ldots, W_d) = V^\top X \sim N(V^\top \mu, \Lambda)$; in particular, $W_1, W_2, \ldots, W_d$ are independent. This means that if $X$ has a multivariate Gaussian distribution, then it is a rotation away from being a vector of independent random variables. In fact, if

$$Z := \Lambda^{-1/2} V^\top (X - \mu),$$

then $Z \sim N(0, I)$.

A more general phenomenon at work here is that a \emph{linear transformation} of a Gaussian random vector is again a Gaussian random vector: for $X \sim N(\mu, \Sigma)$ and any $A \in \mathbb{R}^{d \times k}$,

$$A^\top X \sim N(A^\top \mu, A^\top \Sigma A).$$

8 Fitting a Gaussian distribution to data

Suppose you want to model some data $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ (regarded as an iid sample) with a Gaussian distribution. This amounts to choosing a mean vector $\hat{\mu}$ and a covariance matrix $\hat{\Sigma}$. How should these parameters be chosen?

There is a venerable principle of statistical estimation called the \emph{maximum likelihood principle}, which prescribes the following procedure. Regard the data $S := (x_1, x_2, \ldots, x_n)$ as an iid sample, and then set $(\hat{\mu}, \hat{\Sigma})$ to be the parameters $(\mu, \Sigma)$ with the highest likelihood given the data:

$$\mathcal{L}(\mu, \Sigma; S) := \prod_{i=1}^n p(x_i; \mu, \Sigma) = \prod_{i=1}^n \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2} (x_i - \mu)^\top \Sigma^{-1} (x_i - \mu) \right).$$

It turns out to be convenient to instead consider the \emph{log-likelihood} of the parameters, since the logarithm of a product is a sum of logarithms:

$$\ln \mathcal{L}(\mu, \Sigma; S) = \ln \prod_{i=1}^n p(x_i; \mu, \Sigma) = \sum_{i=1}^n \ln \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2} (x_i - \mu)^\top \Sigma^{-1} (x_i - \mu) \right)$$

$$= -\frac{n}{2} \ln \det(\Sigma) - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^\top \Sigma^{-1} (x_i - \mu) + \text{constant}.$$

(The logarithm is a strictly increasing function on the positive reals, so a maximizer of the log-likelihood is a maximizer of the likelihood.) In this case, we can derive closed-form expressions for the maximizing $(\hat{\mu}, \hat{\Sigma})$ of $\ln \mathcal{L}(\cdot; S)$ using (matrix) calculus:

$$\hat{\mu} := \frac{1}{n} \sum_{i=1}^n x_i,$$

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

We arrive at these formulas for $\hat{\mu}$ and $\hat{\Sigma}$ by considering an “optimality condition” that any maximizer of the log-likelihood must satisfy. This optimality condition says that the derivatives of the
log-likelihood with respect to the optimization variables ($\mu$ and $\Sigma$ in this case) must equal to zero when evaluated at the purported maximizers.

We emphasize one important caveat, which is that the maximum likelihood setting of $\hat{\Sigma}$ is not valid whenever $\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^\top$ is singular, as such a matrix is not positive definite. In fact, in this case, the value of the log-likelihood $\ln L(\mu, \Sigma; S)$ function can be made to diverge to $+\infty$ by choosing a sequence of matrices $\Sigma$ (all symmetric and positive definite) appropriately.