CSE 840: Computational Foundations of Artificial Intelligence April 02, 2025

Expectation and Variance in the General Setting

Instructor: Vishnu Boddeti Scribes: Jiaming Zhang, Yuyuan Tian

1 Expectation and Variance

Definition 1.1 (L^p-space) *Let $(\Omega, \mathcal{A}, P)$ be a probability space. For $1 \le p < \infty$ we define*

$$L^p(\Omega, \mathcal{A}, P) \; := \; \big\{X : \Omega \to \mathbb{R} \;\big|\; X \text{ measurable and } \int_\Omega |X|^p \, dP < \infty\big\}.$$

For $p = \infty$ we write L^∞ for the space of essentially bounded random variables.

Definition 1.2 (Expectation & Moments) *If $X \in L^1(\Omega, \mathcal{A}, P)$, its* expectation *(or first moment) is*

$$\mathbb{E}[X] \; := \; \int_\Omega X \, dP = \int_\mathbb{R} x \, dP_X(x).$$

More generally, if $k \in \mathbb{N}$ and $X^k \in L^1$, the k-th moment is

$$\mathbb{E}\big[X^k\big] \; = \; \int_\Omega X^k \, dP.$$

Definition 1.3 (Variance & Covariance) *For $X, Y \in L^2(\Omega, \mathcal{A}, P)$ the* variance *and* covariance *are defined by*

$$\operatorname{Var}(X) \; := \; \mathbb{E}\big[(X - \mathbb{E}[X])^2\big], \qquad \operatorname{Cov}(X, Y) \; := \; \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big].$$

2 Markov and Chebyshev Inequalities

2.1 Cauchy–Schwarz Inequality

Theorem 2.1 (Cauchy–Schwarz Inequality) *Let $x, y \in L^2(\Omega, \mathcal{A}, P)$. Then,*

$$\big|\mathbb{E}[x\,y]\big|^2 \; \le \; \mathbb{E}\big[x^2\big]\,\mathbb{E}\big[y^2\big].$$

2.2 Markov Inequality

Theorem 2.2 (Markov Inequality) *Let $g : [0, \infty) \to [0, \infty)$ be a non-decreasing measurable function and let X be a non-negative r.v. Then for every $a > 0$*

$$P\big\{X \ge a\big\} \; \le \; \frac{\mathbb{E}\big[g(X)\big]}{g(a)}.$$

In particular, with $g(x) = x$ we obtain $P\{X \geq a\} \leq \frac{\mathbb{E}[X]}{a}$.

2.3 Chebyshev Inequality

Theorem 2.3 (Chebyshev) *For any $X \in L^2$ and $\varepsilon > 0$*

$$P\big\{|X - \mathbb{E}[X]| \geq \varepsilon\big\} \;\leq\; \frac{\mathrm{Var}(X)}{\varepsilon^2}.$$

Chebyshev's inequality provides a distribution-free upper bound on the probability of large deviations. It is a key tool in proving the Weak Law of Large Numbers.

3 Probability Distributions

3.1 Discrete Distributions

Definition 3.1 (Uniform Distributions on $\{1, \dots, n\}$) *A discrete r.v. X is* uniform *on $\{1, \dots, n\}$ if $P\{X = i\} = \frac{1}{n}$ for each i.*

Definition 3.2 (Binomial Distributions $\mathrm{Bin}(n, p)$) *Let $n \in \mathbb{N}$ and $p \in (0, 1)$. If X counts the number of heads in n independent Bernoulli(p) trials then*

$$P\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, \dots, n.$$

Definition 3.3 (Poisson Distributions $\mathrm{Pois}(\lambda)$) *For $\lambda > 0$, a r.v. X is* Poisson *with rate λ if*

Parameter $\lambda > 0$

$$P\{X = k\} = e^{-\lambda} \frac{\lambda^k}{k!}, \qquad k \in \mathbb{N}_0.$$

It often models the number of arrivals in a fixed time interval.

3.2 Continuous Distributions

Definition 3.4 (Uniform on $[a, b)$]

A continuous r.v. X is uniform *on $[a, b]$ if its density is*

$$f_X(x) = \begin{cases} (b-a)^{-1}, & x \in [a, b], \\ 0, & \textit{otherwise.} \end{cases}$$

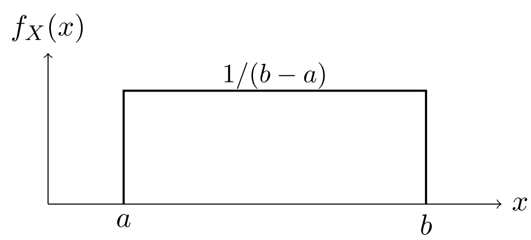

Figure 1: Density of a continuous uniform distribution on $[a, b]$.

Definition 3.5 (Normal $\mathcal{N}(\mu, \sigma^2)$) *A r.v. X is* normal *with mean μ and variance $\sigma^2 > 0$ if its density is*

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big).$$

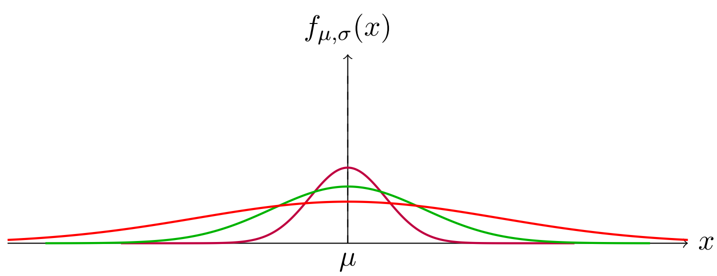

Figure 2: Normal densities with identical mean μ and different variances ($\sigma_{\text{orange}} < \sigma_{\text{green}} < \sigma_{\text{red}}$).

4 Multivariate Normal Distribution

Let $X = (X_1, \ldots, X_n)^\top \in \mathbb{R}^n$ with mean vector μ and covariance matrix Σ. We write $X \sim \mathcal{N}(\mu, \Sigma)$ if

$$f_{\mu,\Sigma}(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\Big(-\tfrac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\Big).$$

Key facts.

- Σ is symmetric positive semi-definite and thus possesses an eigen-decomposition $\Sigma = Q\Lambda Q^\top$.
- The contour ellipsoids of $f_{\mu,\Sigma}$ are aligned with the eigenvectors of Σ.
- Independence of components X_i is equivalent to Σ being diagonal.
- If $X \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $Y \sim \mathcal{N}(\mu_2, \Sigma_2)$ are independent, then $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$.

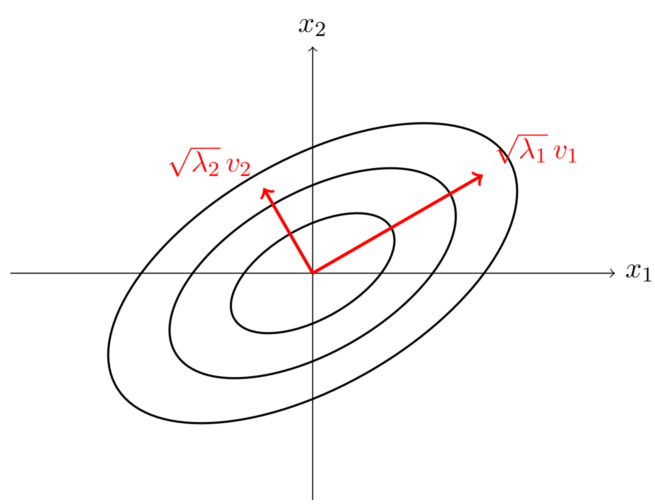

Figure 3: Contours of a bivariate normal distribution with mean $(0, 0)$. Ellipses represent constant Mahalanobis distance. Red arrows indicate eigenvectors v_1, v_2 of Σ scaled by $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$.

5 Mixture of Gaussians

Definition 5.1 (Gaussian Mixture Model) *Let* $\{\pi_i\}_{i=1}^k$ *be non-negative weights satisfying* $\sum_{i=1}^k \pi_i = 1$*, and let* f_{μ_i,Σ_i} *denote Gaussian densities. The* Gaussian mixture *density is*

$$f(x) = \sum_{i=1}^{k} \pi_i \, f_{\mu_i,\Sigma_i}(x).$$

GMMs combine multiple Gaussian "clusters" and can approximate arbitrary continuous densities. The Expectation–Maximisation (EM) algorithm is the canonical method for parameter estimation.

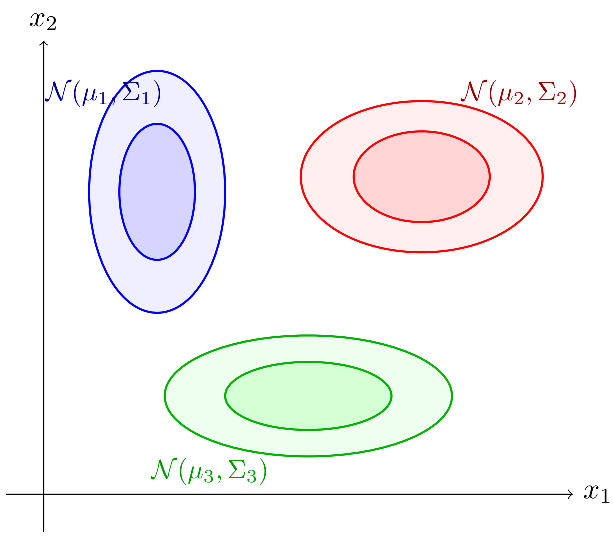

Figure 4: Contour plot of a 3-component Gaussian mixture in $\mathbb{R}^2$. Shaded ellipses depict $1\,\sigma$ and $2\,\sigma$ level sets for each component.

6 Additional Results

6.1 Weak Law of Large Numbers

Theorem 6.1 (WLLN) *Let* $X_1, X_2, \ldots$ *be i.i.d. with* $\mathbb{E}[X_i] = \mu$ *and* $\mathrm{Var}(X_i) = \sigma^2 < \infty$. *Then for any* $\varepsilon > 0$

$$\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \varepsilon\right\} \longrightarrow 0 \quad (n \to \infty).$$

6.2 Central Limit Theorem

Theorem 6.2 (CLT) *Under the same assumptions as the WLLN,*

$$\frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow{d} \mathcal{N}(0,1).$$