Understanding Deep Learning

Appendices

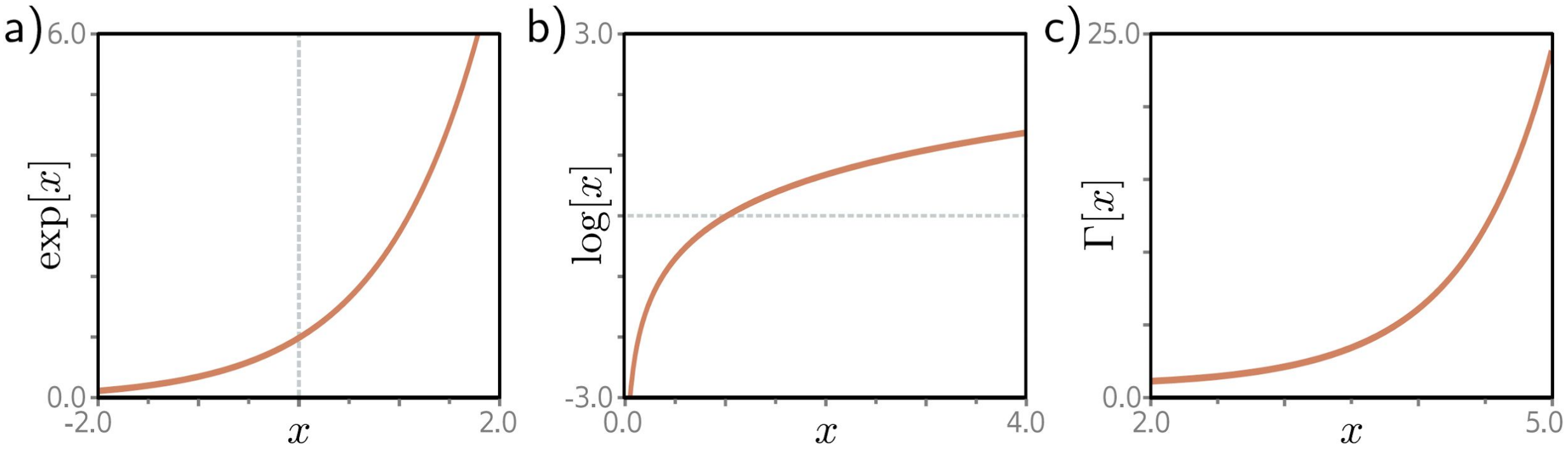


Figure B.1 Exponential, logarithm, and gamma functions. a) The exponential function maps a real number to a positive number. It is a concave function. b) The logarithm is the inverse of the exponential and maps a positive number to a real number. It is a convex function. c) The Gamma function is a continuous extension of the factorial function so that $\Gamma[x] = (x-1)!$ for $x \in \{1, 2, \ldots\}$.

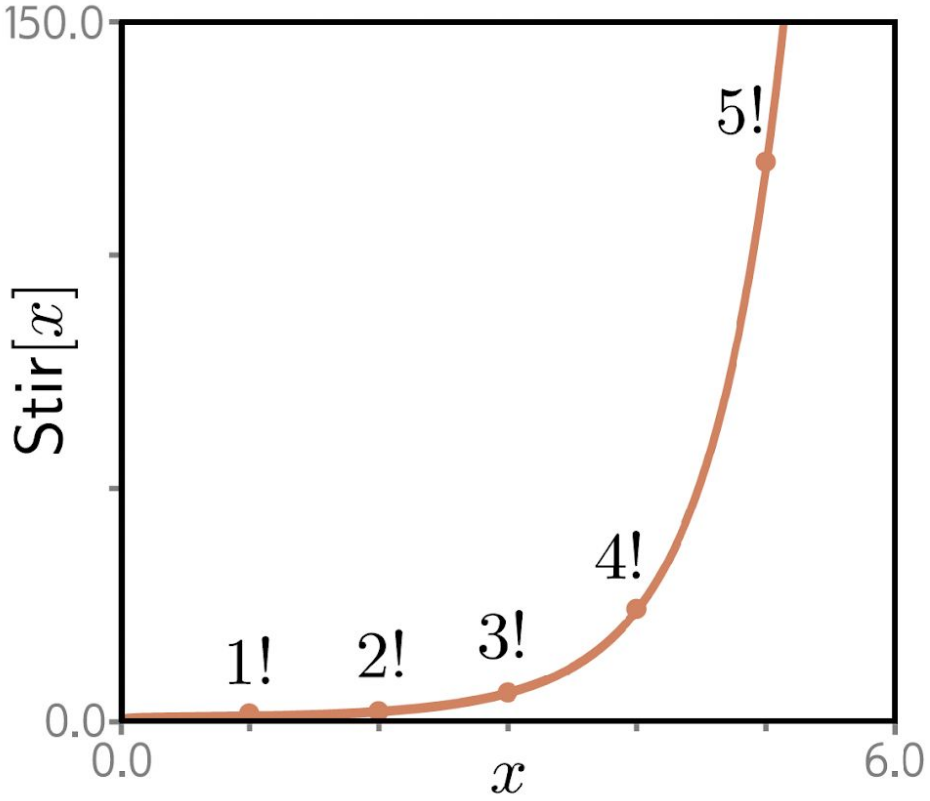
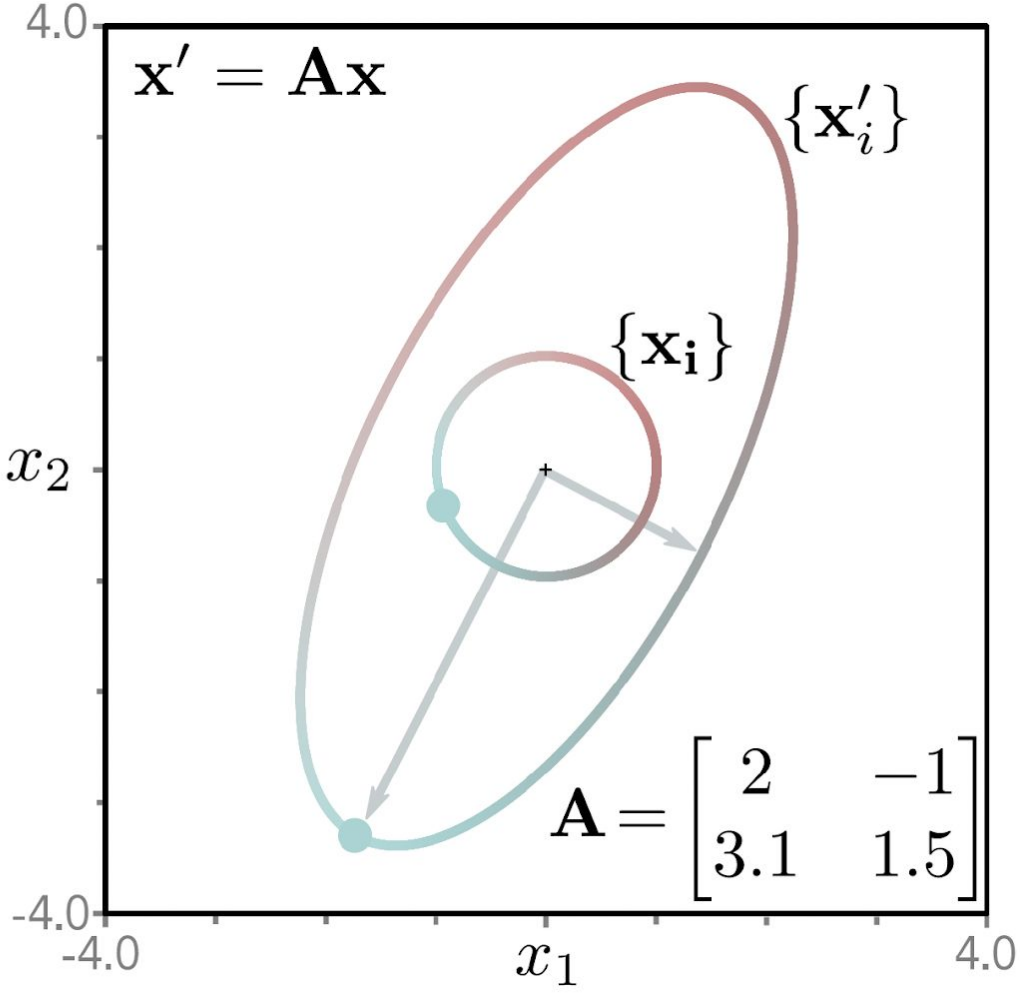


Figure B.2 Stirling's formula. The factorial function $x!$ can be approximated by Stirling's formula Stir$[x]$ which is defined for every real value.

Figure B.3 Eigenvalues. When the points $\{\mathbf{x}_i\}$ on the unit circle are transformed to points $\{\mathbf{x}'_i\}$ by a linear transformation $\mathbf{x}'_i = \mathbf{A}\mathbf{x}_i$, they are mapped to an ellipse. For example, the light blue point on the unit circle is mapped to the light blue point on the ellipse. The length of the major (longest) axis of the ellipse (long gray arrow) is the magnitude of the first eigenvalue of the matrix, and the length of the minor (shortest) axis of the ellipse (short gray arrow) is the magnitude of the second eigenvalue.



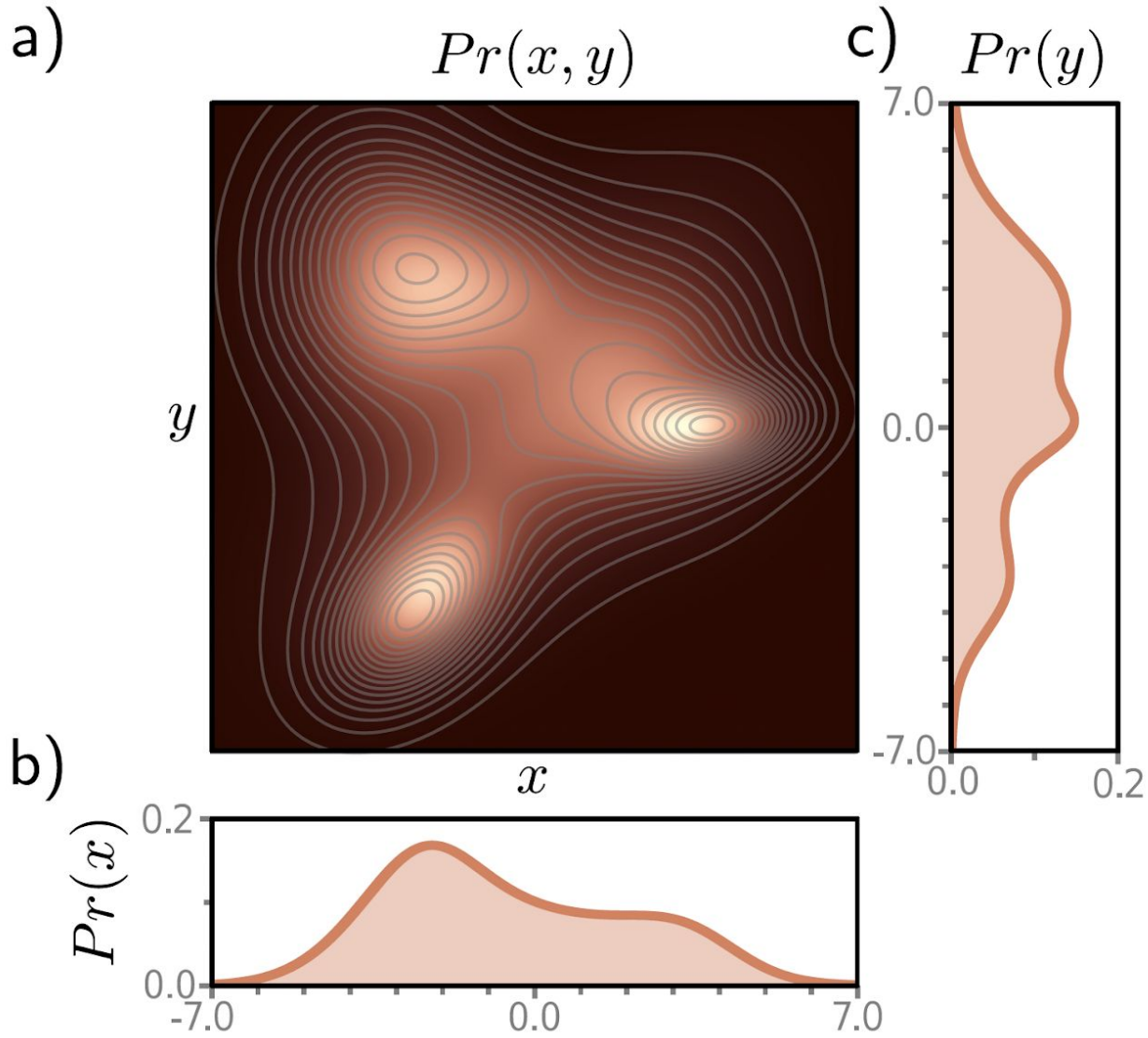


Figure C.1 Joint and marginal distributions. a) The joint distribution $Pr(x, y)$ captures the propensity of variables x and y to take different combinations of values. Here, the probability density is represented by the color map, so brighter positions are more probable. For example, the combination $x = 6, y = 6$ is much less likely to be observed than the combination $x = 5, y = 0$. b) The marginal distribution $Pr(x)$ of variable x can be recovered by integrating over y. c) The marginal distribution $Pr(y)$ of variable y can be recovered by integrating over x.

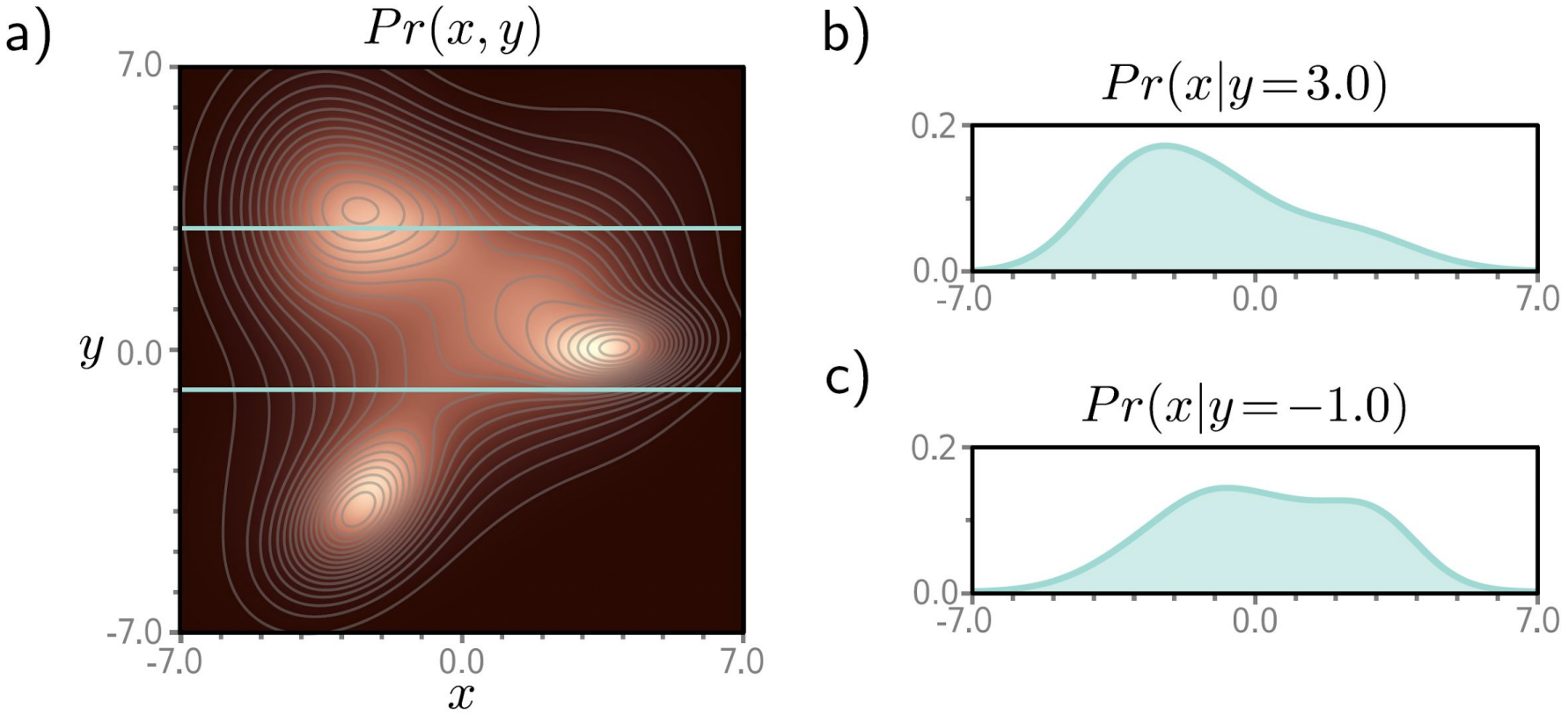


Figure C.2 Conditional distributions. a) Joint distribution $Pr(x, y)$ of variables x and y. b) The conditional probability $Pr(x|y = 3.0)$ of variable x, given that y takes the value 3.0, is found by taking the horizontal "slice" $Pr(x, y = 3.0)$ of the joint probability (top cyan line in panel a), and dividing this by the total area $Pr(y = 3.0)$ in that slice so that it forms a valid probability distribution that integrates to one. c) The joint probability $Pr(x, y=-1.0)$ is found similarly using the slice at $y=-1.0$.

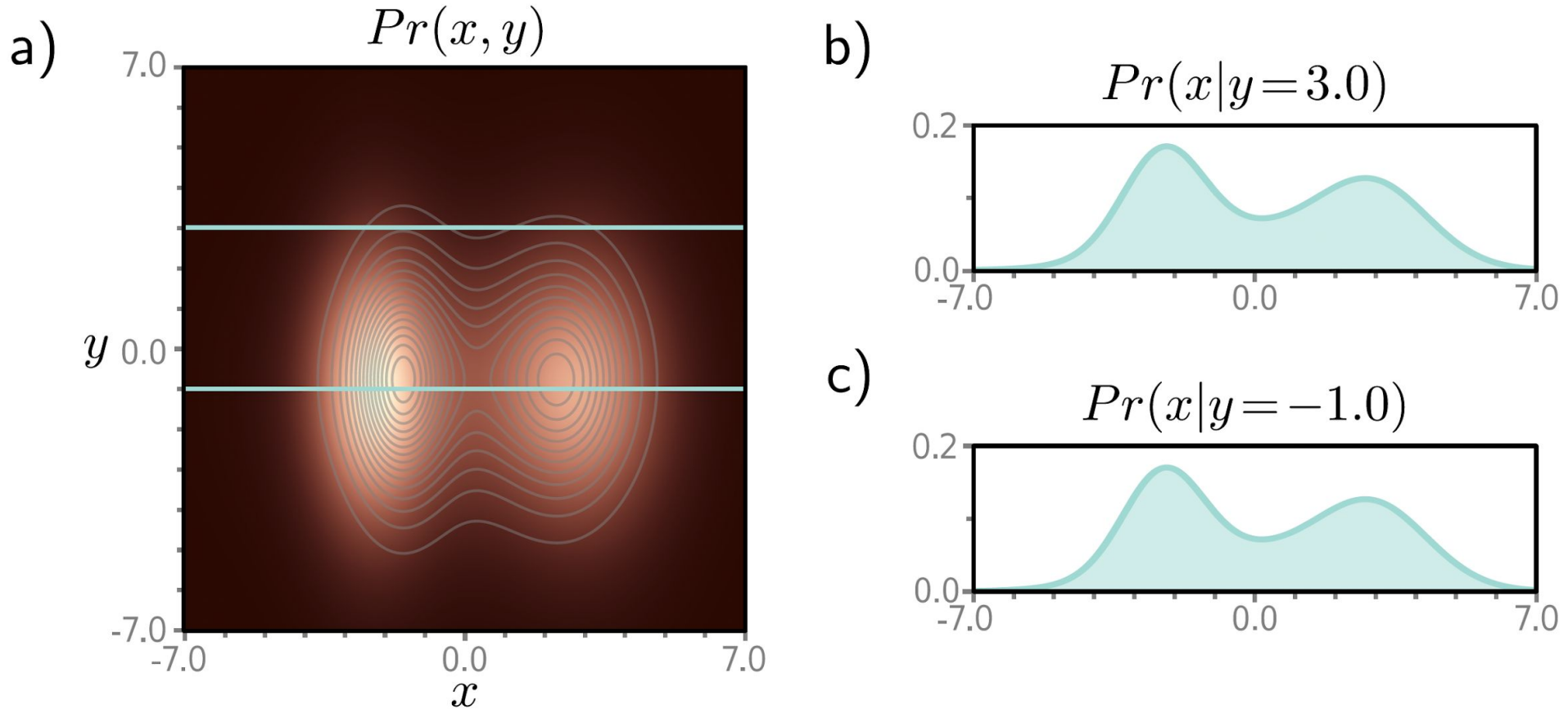


Figure C.3 Independence. a) When two variables x and y are independent, the joint distribution factors into the product of marginal distributions, so $Pr(x, y) = Pr(x)Pr(y)$. Independence implies that knowing the value of one variable tells us nothing about the other. b–c) Accordingly, all of the conditional distributions $Pr(x|y = \bullet)$ are the same and are equal to the marginal distribution $Pr(x)$.

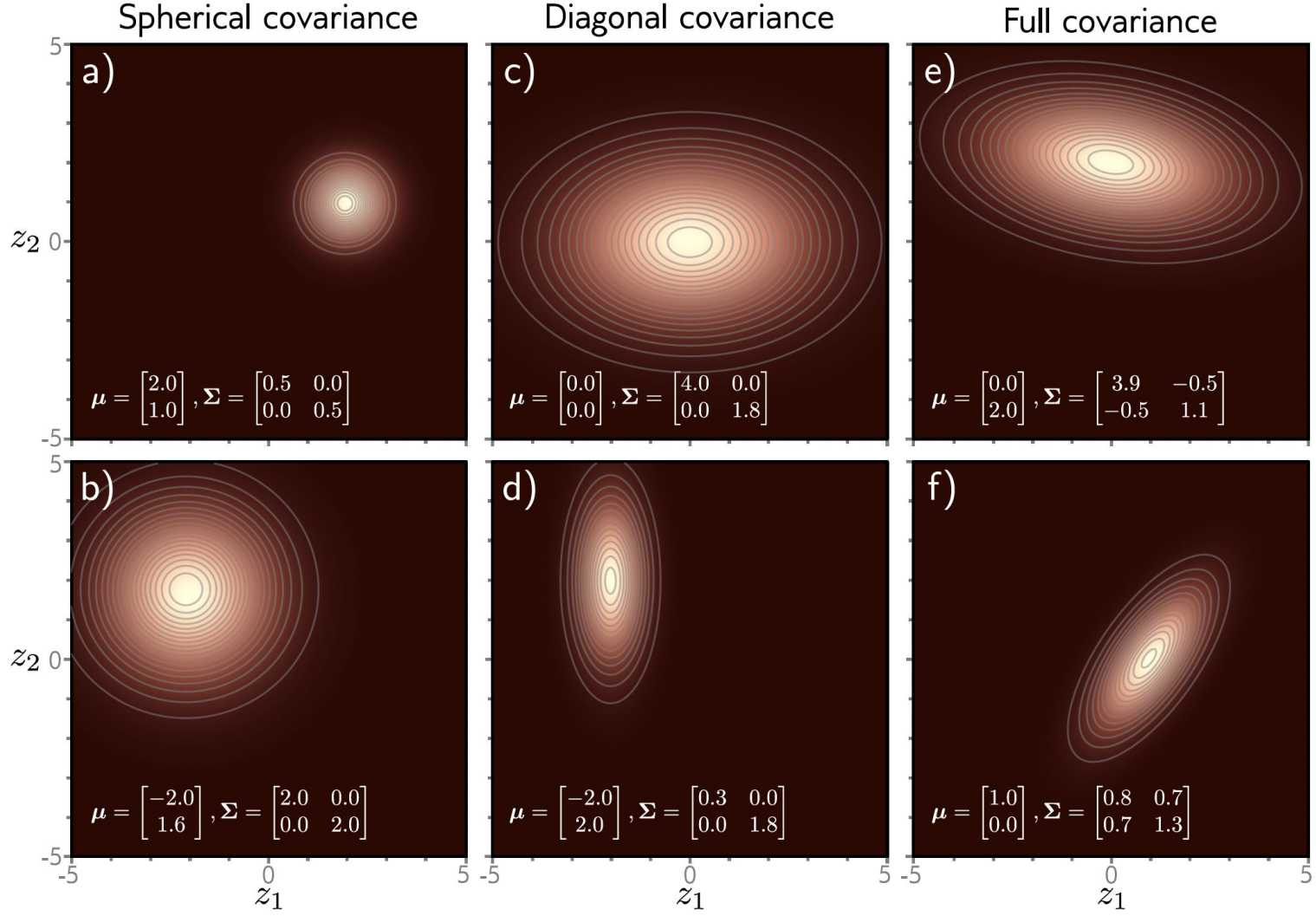


Figure C.4 Bivariate normal distribution. a–b) When the covariance matrix is a multiple of the diagonal matrix, the isocontours are circles, and we refer to this as spherical covariance. c–d) When the covariance is an arbitrary diagonal matrix, the isocontours are axis-aligned ellipses, and we refer to this as diagonal covariance e–f) When the covariance is an arbitrary symmetric positive definite matrix, the iso-contours are general ellipses, and we refer to this as full covariance.

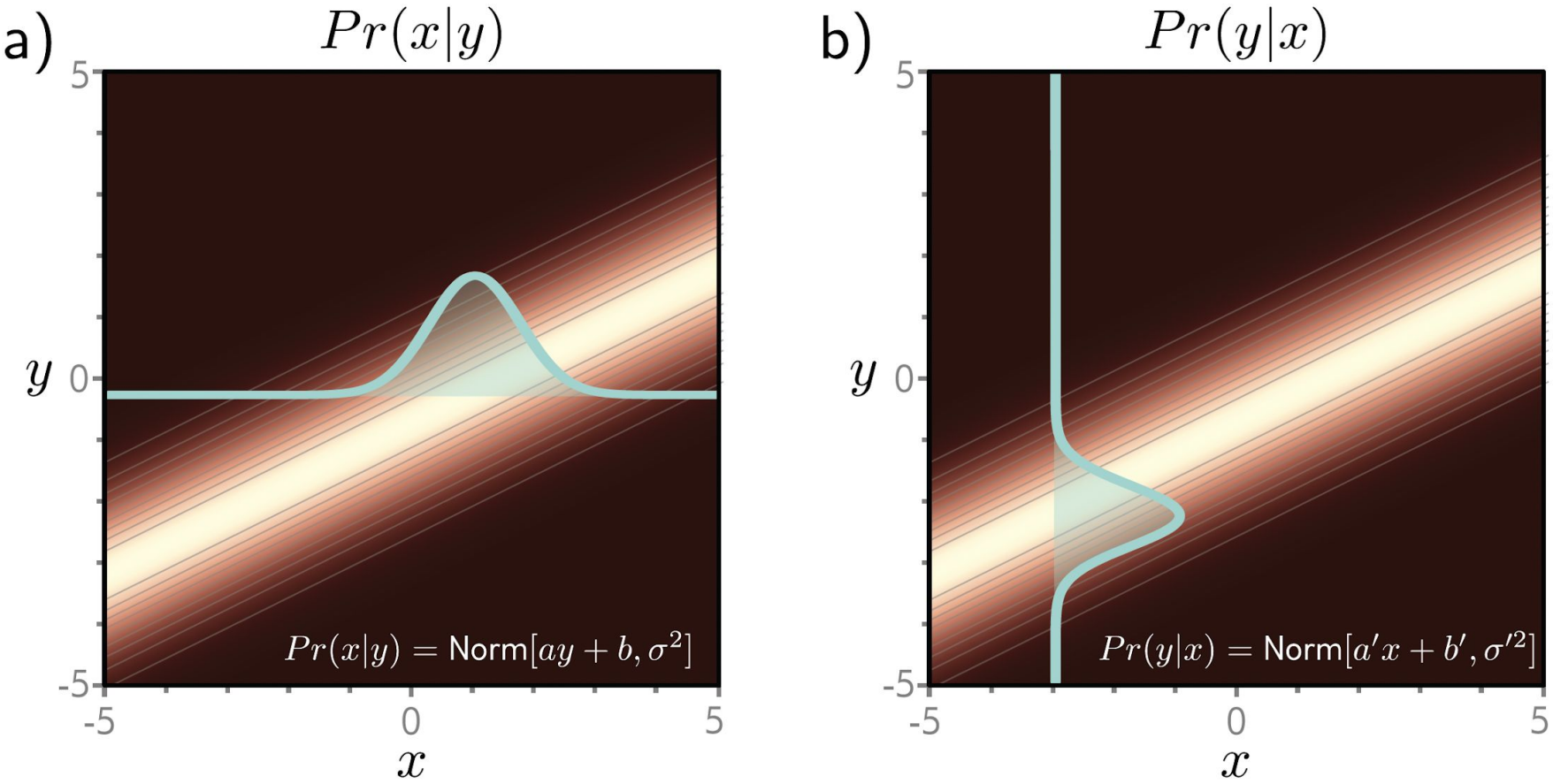


Figure C.5 Change of variables. a) The conditional distribution $Pr(x|y)$ is a normal distribution with constant variance and a mean that depends linearly on y. Cyan distribution shows one example for $y = -0.2$. b) This is proportional to the conditional probability $Pr(y|x)$, which is a normal distribution with constant variance and a mean that depends linearly on x. Cyan distribution shows one example for $x = -3$.

Figure C.6 Lower bound on negative logarithm. The function $1 - y$ is always less than the function $-\log[y]$. This relation is used to show that the Kullback-Leibler divergence is always greater than or equal to zero.