# 5SSD0 Formula Sheet

August 24, 2025

## Distributions

No matter how $x$ is distributed,

$$\mathrm{E}[Ax+b] = A\mathrm{E}[x] + b \tag{1}$$

$$\mathrm{Cov}[Ax+b] = A\mathrm{Cov}[x]A^T \tag{2}$$

### Gaussian distribution

The (*moment* parameterization of the) **Gaussian distribution** is given by

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{M/2}|\Sigma|^{1/2}} \exp\left( -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right). \tag{3}$$

Alternatively, the *canonical* parameterization of the Gaussian distribution is given by

$$\mathcal{N}_c(x|\eta,\Lambda) = \exp\left( a + \eta^T x - \frac{1}{2}x^T\Lambda x\right), \tag{4}$$

where where $\eta = \Sigma^{-1}\mu$ is the canonical mean, $\Lambda = \Sigma^{-1}$ is the precision matrix, and $a$ is a normalization constant.

#### Multiplication of Gaussians

$$\mathcal{N}(x|\mu_a,\Sigma_a)\mathcal{N}(x|\mu_b,\Sigma_b) = \underbrace{\mathcal{N}(\mu_a|\mu_b,\Sigma_a+\Sigma_b)}_{\text{normalization constant}}\mathcal{N}(x|\mu_c,\Sigma_c) \tag{5}$$

where $\Sigma_c^{-1} = \Sigma_a^{-1} + \Sigma_b^{-1}$, and $\Sigma_c^{-1}\mu_c = \Sigma_a^{-1}\mu_a + \Sigma_b^{-1}\mu_b$.

#### Conditioning and marginalization

Let $z = \begin{bmatrix} x \\ y \end{bmatrix}$ be jointly normal distributed as

$$p(z) = \mathcal{N}(z|\mu,\Sigma) = \mathcal{N}\left( \begin{bmatrix} x \\ y \end{bmatrix} \middle| \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_y \end{bmatrix} \right)$$

Then,

$$p(y|x) = \mathcal{N}\left(y \,|\, \mu_y + \Sigma_{xy}^T \Sigma_x^{-1}(x - \mu_x),\, \Sigma_y - \Sigma_{xy}^T \Sigma_x^{-1} \Sigma_{xy}\right) \tag{6}$$

$$p(x) = \mathcal{N}\left(x|\mu_x, \Sigma_x\right) \tag{7}$$

### Beta distribution

For a variable $x \in [0,1]$, the **Beta distribution** is given by

$$\mathrm{Beta}(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\,. \tag{8}$$

The expected value of the Beta distribution is $\mathrm{E}[x] = \alpha/(\alpha+\beta)\,$.

## Matrix Calculus

We define the **gradient** of a scalar function $f(A)$ with respect to an $n \times k$ matrix $A$ as

$$\nabla_A f = \begin{bmatrix} \frac{\partial f}{\partial a_{11}} & \frac{\partial f}{\partial a_{12}} & \cdots & \frac{\partial f}{\partial a_{1k}} \\ \frac{\partial f}{\partial a_{21}} & \frac{\partial f}{\partial a_{22}} & \cdots & \frac{\partial f}{\partial a_{2k}} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial f}{\partial a_{n1}} & \frac{\partial f}{\partial a_{n2}} & \cdots & \frac{\partial f}{\partial a_{nk}} \end{bmatrix} \tag{9}$$

The following maxtrix calculus formulas are useful, see also Bishop PRML, appendix C:

$$(AB)^T = B^T A^T \tag{10}$$

$$(AB)^{-1} = B^{-1} A^{-1} \tag{11}$$

$$|A^{-1}| = |A|^{-1} \tag{12}$$

$$\nabla_A \log|A| = (A^T)^{-1} = (A^{-1})^T \tag{13}$$

$$\mathrm{Tr}[ABC] = \mathrm{Tr}[CAB] = \mathrm{Tr}[BCA] \tag{14}$$

$$\nabla_A \mathrm{Tr}[AB] = \nabla_A \mathrm{Tr}[BA] = B^T \tag{15}$$

$$\nabla_A \mathrm{Tr}[ABA^T] = A(B + B^T) \tag{16}$$

$$\nabla_x x^T A x = (A + A^T)x \tag{17}$$

$$\nabla_X a^T X b = \nabla_X \mathrm{Tr}[ba^T X] = ab^T \tag{18}$$

## Factor Graphs

For a node $f(y, x_1, \ldots, x_n)$ with incoming messages $\overrightarrow{\mu}_{X_1}(x_1),\ \ldots,\ \overrightarrow{\mu}_{X_n}(x_n)$, the outgoing message is given by the **sum-product rule**:

$$\overrightarrow{\mu}_Y(y) = \sum_{x_1,\dots,x_n} \underbrace{\overrightarrow{\mu}_{X_1}(x_1)\overrightarrow{\mu}_{X_2}(x_2)\cdots\overrightarrow{\mu}_{X_n}(x_n)}_{\text{incoming messages}} \cdot \underbrace{f(y,x_1,\dots,x_n)}_{\text{node function}} \tag{19}$$

# (Variational) Bayes

For a model $p(x,z) = p(x|z)p(z)$, where $x$ and $z$ are observed and unobserved variables, respectively, and a variational posterior distribution $q(z)$, the **variational free energy** (VFE) functional is defined by

$$F[q] = \underbrace{\sum_z q(z)\log\frac{q(z)}{p(z)}}_{\text{complexity}} - \underbrace{\sum_z q(z)\log p(x|z)}_{\text{accuracy}} \,. \tag{20}$$