Final Exam of MML
13:10 – 16:00, 29 December 2025; INS105
Note: Cell phones and any calculator are forbidden.

Part I: True (T) or False (F) (60%; each for 5%)

1. For any $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{A}^\top \mathbf{A}$ is invertible.
2. The shape of $\frac{d}{d\mathbf{x}}(\mathbf{x}^\top \mathbf{x})\mathbf{x}$ is $\mathbb{R}^{1 \times n}$ for $\mathbf{x} \in \mathbb{R}^n$.
3. Every positive definite matrix $M \in \mathbb{R}^{n \times n}$ is invertible.
4. For any objective convex function $f : \mathbb{R}^d \to \mathbb{R}$, gradient descent with a fixed learning rate $\gamma \geq 0$ always converges to the global minimum of f.
5. $f : \mathbb{R}^2 \to \mathbb{R}$; $f(\mathbf{x}) = \|\mathbf{x}\|_2^2$ is a convex function.
6. For any function ψ with arguments $\mathbf{x}$, $\mathbf{y}$, we have $\min_{\mathbf{y}} \max_{\mathbf{x}} \psi(\mathbf{x}, \mathbf{y}) \leq \max_{\mathbf{x}} \min_{\mathbf{y}} \psi(\mathbf{x}, \mathbf{y})$.
7. Given a set of *N* samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ and probability densities $p(\mathbf{x} \mid \theta)$ parameterized by θ, the negative log-likelihood of the data is $-\prod_{i=1}^N \log p(y_i \mid \mathbf{x}_i, \theta)$.
8. Compared with the maximum a posteriori estimation (MAP), maximum likelihood estimation (MLE) suffers less overfitting issues.
9. Given a sampler of $\mathcal{N}(\mathbf{0}, \mathbf{I})$, we can generate a sample $\mathbf{y} \sim \mathcal{N}(\mu, \mathbf{I})$ by transforming $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to $\mathbf{y} = \mathbf{x} + \mu$.
10. For $\mathbf{x} \sim \mathcal{N}(\mu_x, \Sigma_x)$ and $\mathbf{y} \sim \mathcal{N}(\mu_y, \Sigma_y)$, the distribution of $a\mathbf{x} + b\mathbf{y}$ for $a, b \in \mathbb{R}$ is given as $\mathcal{N}(a\mu_x + b\mu_y, a\Sigma_x + b\Sigma_y)$.
11. Given the Gaussian Mixture Model shown as below, the smaller K is, the more expressive the model will be:

$$p(\mathbf{x} \mid \theta) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k)$$
$$0 \leq \pi_k \leq 1, \ \sum_{k=1}^K \pi_k = 1,$$

12. When we consider maximizing the Lagrangian of SVM, we obtain $\mathbf{w} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i$, which is a linear combination of the data points $\mathbf{x}_i$'s. Then, $\mathbf{x}_i$'s with $\alpha_i = 0$ is called the *support vectors*.

Part II: Calculations. (80%; ONLY THE ANSWERS ARE REQUIRED)

1. (10%) Compute $\frac{d}{d\mathbf{x}}(\mathbf{x}^\top \mathbf{x})\mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^n$.

2. (10%) For $\mu, \sigma \in \mathbb{R}$, compute the derivative $f'(x)$ of the function $f : \mathbb{R} \to \mathbb{R}$;

$$f(x) = \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right).$$

3. (10%) Let X be a continuous random variable with pdf $f_X : [0, 1] \to [0, 1]$: $f_X(x) = 4x^3$. Compute the pdf of $Y = X^2$.

4. (5%) Given $\mathbf{x}, \mathbf{y}, \mathbf{b} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{n \times n}$, if $\mathbf{x}, \mathbf{y}$ are random vectors such that $\mathbf{y} = 2\mathbf{A}\mathbf{x} - \mathbf{b}$ and $\mathbb{V}[\mathbf{x}] = \sigma$, then compute the variance $\mathbb{V}[\mathbf{y}]$. (Hint: $\mathbb{V}[\mathbf{x}] = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^\top$)

5. (10%) Consider the problem:

$$\min_{\mathbf{x} \in \mathbb{R}^d} \mathbf{c}^\top \mathbf{x}$$
$$\text{subject to } \mathbf{A}\mathbf{x} \preceq \mathbf{b}, \text{ for } \mathbf{A} \in \mathbb{R}^{m \times d}, \mathbf{b} \in \mathbb{R}^m \text{ and } \mathbf{c} \in \mathbb{R}^d.$$

   a. (5%) Please provide its Lagrangian function $\mathcal{L}(\mathbf{x}, \lambda)$.

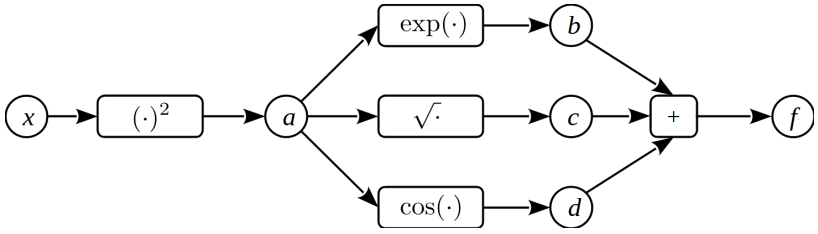
   b. (5%) Please list the dual problem.

6. (5%) Given $f : \mathbb{R}^+ \to \mathbb{R}$; $f(x) = x \ln x$. Compute $f(t) + (\nabla_x f)(t)^\top (z - t)$ for $t = e^2, z = e^3$.

7. (5%) Consider $f : \mathbb{R}^2 \to \mathbb{R}$; $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix} \mathbf{x} - \begin{bmatrix} 2 \\ 3 \end{bmatrix}^\top \mathbf{x} + [1 \;\; 0]^\top$. Please compute $\nabla f(\mathbf{x})$.

8. (5%) Given the function $f(x, y) = x^2 + 2xy$ for $x, y \in \mathbb{R}$, please compute the Hessian matrix of f.

9. (10%) Consider the following workflow



   According to the automatic differentiation rule of the reverse mode (backpropagation), please write down how we can compute $\frac{\partial f}{\partial a}$ and $\frac{\partial f}{\partial x}$.

   (Hint: $\frac{\partial f}{\partial b} = \frac{\partial f}{\partial c} = \frac{\partial f}{\partial d} = 1$)

10. (10%) In a Markov Decision Process, a trajectory $\tau = (\mathbf{s}_0, \mathbf{a}_0, \mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_T, \mathbf{a}_T)$ occurs with probability

$$\begin{aligned} p_\theta(\tau) \;\; = \;\; & p(\mathbf{s}_0)\, \pi_\theta(\mathbf{a}_0 \mid \mathbf{s}_0)\, p(\mathbf{s}_1 \mid \mathbf{s}_0, \mathbf{a}_0)\, \pi_\theta(\mathbf{a}_1 \mid \mathbf{s}_1)\, p(\mathbf{s}_2 \mid \mathbf{s}_1, \mathbf{a}_1) \\ & \cdots p(\mathbf{s}_T \mid \mathbf{s}_{T-1}, \mathbf{a}_{T-1})\, \pi_\theta(\mathbf{a}_T \mid \mathbf{s}_T). \end{aligned}$$

   The gradient of the expected return $r(\tau)$ turns out to be $\frac{\partial}{\partial \theta} \mathbb{E}_{p_\theta(\tau)}[\, r(\tau)\,] = \frac{\partial}{\partial \theta} \sum_\tau r(\tau)\, p_\theta(\tau)$.

   Please prove that $\frac{\partial}{\partial \theta} \mathbb{E}_{p_\theta(\tau)}[\, r(\tau)\,] = \mathbb{E}_{p_\theta(\tau)}\Big[\, r(\tau)\, \frac{\partial}{\partial \theta} \log p_\theta(\tau)\,\Big]$.