Deep Learning Quiz

Date: 9/22/2022. Duration: 50 minutes

1. Answer True or False in the following statements about the Gaussian distribution $\mathcal{N}$:
   (a) If $\mathrm{x} \sim \mathcal{N}$, then $a\mathrm{x} + b \sim \mathcal{N}$ for any constants $a, b \in \mathbb{R}$ (5%)
   (b) If $\mathrm{z} = \mathrm{x} + \mathrm{y}$, where $\mathrm{x}, \mathrm{y} \sim \mathcal{N}$, then $\mathrm{z} \sim \mathcal{N}$. (5%)
2. Given N i.i.d samples $\boldsymbol{X}^{\in N \times D} = [x^{(1)}, \dots, x^{(N)}]^{\mathrm{T}}$ of a random variable $\mathbf{x}$, the Principal Components Analysis (PCA) finds K orthonormal vector $\boldsymbol{W} = [w^{(1)}, \dots, w^{(K)}]$ such that the transformed variable $\mathbf{z} = \boldsymbol{W}^{\mathrm{T}}\mathbf{x}$ has the most "spread out" attributes, i.e., each attribute $\mathrm{z_i} = w^{(i)\mathrm{T}}\mathbf{x}$ has the maximum variance $\mathrm{Var}(\mathrm{z_i})$. Now consider the problem of finding $w^{(1)}$:
   (a) Assuming that x has zero mean, show that $\sigma_{\mathrm{z}1}^2 = \frac{1}{N} w^{(1)\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} w^{(1)}$. (10%)
   (b) Use the Rayleigh's Quotient to explain that the optimal $w^{(1)}$ is given by the eigenvector of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ corresponding to the largest eigenvalue. (10%)
3. Consider a situation where a doctor wants to inference if a patient is having either the disease $\mathrm{y}^{(1)}$ or $\mathrm{y}^{(2)}$ by examining the patient's symptoms x. Explain why the Bayes' rule,
   $$\mathrm{P}(\mathrm{y}|\mathrm{x}) = \frac{\mathrm{P}(\mathrm{x}|\mathrm{y})\mathrm{P}(\mathrm{y})}{\mathrm{P}(\mathrm{x})},$$
   can make the inference easier. (10%)
4. Give an example of two distributions P and Q to show that the Kullback-Leibler (KL) Divergence $\mathrm{D_{KL}}(\mathrm{P}||\mathrm{Q}) = \mathrm{E}_{\mathrm{x}\sim\mathrm{P}}[\log\frac{\mathrm{P}(\mathrm{x})}{\mathrm{Q}(\mathrm{x})}]$ is asymmetric, i.e., $\mathrm{D_{KL}}(\mathrm{P}||\mathrm{Q}) \neq \mathrm{D_{KL}}(\mathrm{Q}||\mathrm{P})$. (10%)
5. Consider a continuous, differentiable function $f: \mathbb{R}^d \to \mathbb{R}$ and an input point $\boldsymbol{a} \in \mathbb{R}^d$.
   (a) For any direction $\boldsymbol{u}$ in the input space, show that the directional derivative of f at $\boldsymbol{a}$ along $\boldsymbol{u}$ equals to $\nabla f(\boldsymbol{a})^T \boldsymbol{u}$. (10% Hint: the directional derivative of f at $\boldsymbol{a}$ along $\boldsymbol{u}$ is the derivative of function $f(\boldsymbol{a} + \varepsilon \boldsymbol{u})$ with respect to ε, evaluated at $\varepsilon = 0$.)
   (b) What is the direction in the input space that leads to the steepest decent of f starting from $\boldsymbol{a}$, i.e., what is the solution of $\mathrm{argmin}_{\boldsymbol{u},\|\boldsymbol{u}\|=1} \nabla f(\boldsymbol{a})^T \boldsymbol{u}$? (10%)
6. Given a quadratic function $f: \mathbb{R}^2 \to \mathbb{R}$, $f(x) = \frac{1}{2} x^{\mathrm{T}} \boldsymbol{A} x - b^{\mathrm{T}} x + c$, where $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$ is symmetric. Explain why the problem
   $$\mathrm{argmin}_x f(x)$$
   is hard to solve by Gradient Descent algorithm when $\boldsymbol{A}$ is ill-conditioned (i.e., when the condition number $\varkappa(\boldsymbol{A}) = \max_{i,j} |\frac{\lambda_i}{\lambda_j}|$ is large). (10%)
7. Given a vector x, let $z = x - \max_i x_i 1$. When you implement the line $\mathrm{c} = \log(softmax(z)_i)$ for some $i > 0$ in a computer program which stores z_i as a float, what numerical issues may occur? (10%) How to walk around these issues in your implementations? (10%)

8. Consider a constrained optimization problem:

$$\min_x f(x)$$

$$\text{subject to } x \in \left\{x: g^{(i)}(x) \leq 0, h^{(j)}(x) = 0\right\}_{i,j},$$

for some positive integers i and j. Explain why the following unconstrained problem:

$$\min_x \max_{\alpha,\beta,\alpha\geq 0} f(x) + \sum_i \alpha_i g^{(i)}(x) + \sum_j \beta_j h^{(j)}(x)$$

gives the same optimal solution. (10%)