Q1 [10 marks]. (3 marks) Give an example of machine learning models, where the training loss is different from the error measure (i.e., measure of success). **(3 marks)** Why is such a difference desired? **(4 marks)** Should the validation criterion be the loss or error measure? And why?

Training loss: cross-entropy
error measure: accuracy
Advantage: differentiable training.
Val criterion: error measure

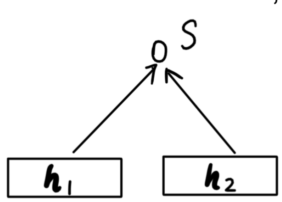
Other solutions are also acceptable:
Training loss: MSE + ℓ_2-penalty
Error: MSE
Advantage: regularization

Q2 [10 marks]. a) Show an example of two models, where the one with more parameters is more overfitting. b) Show another example, where the one with more parameters is **not** more overfitting.

a) NN with 10 hidden units VS NN with 100 hidden units
(more parameters, more overfitting)

b) logistic regression VS 2-way softmax
(more parameters, equivalent model)

Q3 [10 marks]. Consider a local structure of some neural networks: $s = \mathbf{h}_1^\top \mathbf{W} \mathbf{h}_2$, where $\mathbf{h}_1, \mathbf{h}_2 \in \mathbb{R}^d, s \in \mathbb{R}$, and $\mathbf{W}$ is the parameter matrix.



a) **[2 marks]** Give the dimension of $\mathbf{W}$.

b) **[5 marks]** Assume $\partial J / \partial s$ is known. Give the recursion formulas for backpropagation through the local structure.

c) **[3 marks]** Derive the partial derivative of J with respect to $\mathbf{W}$.

Hints: You may use either 1) scalar calculus and organize partial derivatives in the vector form, or 2) the matrix calculus identities

- If $\mathbf{A}$ is not a function of $\mathbf{x}$, then $\nabla_{\mathbf{x}} \mathbf{A}\mathbf{x} = \mathbf{A}^\top$
- If $\mathbf{A}$ is not a function of $\mathbf{x}$ and $\mathbf{A}$ is symmetric, then $\nabla_{\mathbf{x}} \mathbf{x}^\top \mathbf{A} \mathbf{x} = 2\mathbf{A}\mathbf{x}$

a) $d \times d$

b) Notice that $s \in \mathbb{R}$, $\frac{\partial J}{\partial h_1} = \frac{\partial J}{\partial s} \cdot \frac{\partial s}{\partial h_1} = \frac{\partial J}{\partial s} \cdot W h_2$

$\frac{\partial J}{\partial h_2} = \frac{\partial J}{\partial s} \cdot \frac{\partial s}{\partial h_2} = \frac{\partial J}{\partial s} \cdot W^T h_1$

c) $\frac{\partial s}{\partial w_{ij}} = h_{1,i} h_{2j}$

Thus $\frac{\partial s}{\partial W} = h_1 h_2^T$

$\frac{\partial J}{\partial W} = \frac{\partial J}{\partial s} \cdot \frac{\partial s}{\partial W}$
$= \frac{\partial J}{\partial s} h_1 h_2^T$

$[\ h_1^T\ ]$ $\begin{array}{cc} & j \\ i & \begin{bmatrix} - - W & w_{ij} \end{bmatrix} \end{array}$ $\begin{bmatrix} h_2 \end{bmatrix}$

Q4 [25 marks]. Consider a K-way softmax classification based on input $\mathbf{x} \in \mathbb{R}^d$.

a) [10 marks] Write the formula of the softmax classifier, and give the cross-entropy loss function.

b) [15 marks] Show that the optimization is convex.

Hint: The bias term may be omitted for simplicity, as it may be absorbed into weights by introducing a constant feature.

a) $y_i = \text{softmax}(Wx) = \dfrac{\exp(w_i^T x)}{\sum_{i'} \exp(w_{i'}^T x)}$

$$\frac{\partial J}{\partial W_{ij}} = -t_i x_j + \frac{1}{\sum_{k'} \exp(w_{k'}^T x)} \cdot \exp(w_i^T x) \cdot x_j$$

$$= (y_i - t_i)\, x_j$$

Loss: $J = -\sum_{k=1}^{K} t_k \log y_k$

for $t_1, \cdots t_K$ being one-hot target

b) $$J = -\sum_{k=1}^{K} t_k \log y_k = -\sum_{k=1}^{K} t_k \log \frac{\exp(w_k^T x)}{\sum_{k'} \exp(w_{k'}^T x)}$$

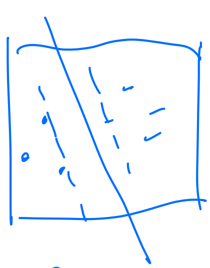
$$= \sum_{k=1}^{K} t_k \left[-w_k^T x + \log \sum_{k'} \exp(w_{k'}^T x) \right] = \left(-\sum_{k=1}^{K} t_k w_k^T x \right) + \left(\log \sum_{k'} \exp(w_{k'}^T x) \right)$$

Q5 [20 marks]. Consider a linearly separable binary classification, where the input is $\mathbf{x} \in \mathbb{R}^d$ and the target is $t \in \{+1, -1\}$.

a) [10 marks] **Explain** in text the intuition of a support vector machine (SVM) and **formulate** the intuition in math. *Hint*: the distance between a point $\mathbf{x}_*$ and a hyperplane $\mathbf{w}^\top \mathbf{x} + b = 0$ is $|\mathbf{w}^\top \mathbf{x}_* + b| / \|\boldsymbol{w}\|$.

b) [10 marks] Transform the problem into the canonical form of convex optimization. Give brief derivation steps and explanations. *Hint*: Cheatsheet in page 3. Solving SVM is not needed.

a) Maximize the margin (minimal distance between a sample and the decision boundary)



$$\underset{w,b}{\text{maximize}} \ \min_{m=1}^{M} \frac{t^{(m)}(w^T x^{(m)} + b)}{\|w\|}$$

b) Since w and b may scale freely, we may set $\min_{m=1}^{M} t^{(m)}(w^T x^{(m)} + b) = 1$

Then the problem becomes

$$\underset{w,b}{\text{maximize}} \ \frac{1}{\|w\|}$$
$$\text{s.t.} \ \min_{m=1}^{M} t^{(m)}(w^T x^{(m)} + b) = 1$$

which is equivalent to

$$\underset{w,b}{\text{minimize}} \ \frac{1}{2}\|w\|^2$$
$$\text{s.t.} \ 1 - t^{(m)}(w^T x^{(m)} + b) \leq 0 \quad \text{for } m = 1, \cdots M.$$

Q6 [25 marks]. Consider the parameters $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$ for a K-way categorical distribution.

a) [5 marks] Given a dataset $\mathcal{D} = \{x^{(m)}\}_{m=1}^M$, where $x^{(m)} \in \{1, \cdots, K\}$ is iid sampled from the categorical distribution $\text{Cat}(\pi_1, \cdots, \pi_K)$. Write out the likelihood of $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$.

$$\mathcal{L}(\pi; \mathcal{D}) = p(\mathcal{D}; \pi) = \prod_{m=1}^{M} \prod_{k=1}^{K} \pi_k^{\mathbb{1}\{x^{(m)}=k\}} = \prod_{m=1}^{M} \prod_{k=1}^{K} \pi_k^{x_k^{(m)}}$$

where we define $x_k^{(m)} = \mathbb{1}\{x^{(m)}=k\}$ for simplicity.

b) [10 marks] A Dirichlet distribution is parameterized by $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)$, where $\alpha_k > 0$ for $k = 1, \cdots, K$. The density of a Dirichlet distribution is $p(\boldsymbol{\pi}; \boldsymbol{\alpha}) = C(\boldsymbol{\alpha}) \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$, where $C(\boldsymbol{\alpha})$ is a normalizing constant depending on $\boldsymbol{\alpha}$.

Now suppose the prior distribution of $\boldsymbol{\pi}$ is a Dirichlet distribution parameterized by $\boldsymbol{\alpha}$. Given $\mathcal{D} = \{x^{(m)}\}_{m=1}^M$, what is the posterior distribution of $\boldsymbol{\pi}$? *Hint:* The posterior distribution would be another Dirichlet distribution parameterized by $\tilde{\boldsymbol{\alpha}}$. What is $\tilde{\boldsymbol{\alpha}}$?

Posterior $p(\pi \mid \mathcal{D}) \propto p(\pi) \cdot p(\mathcal{D} \mid \pi) \propto \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \cdot \prod_{k=1}^{K} \pi_k^{\sum_{m=1}^{M} x_k^{(m)}}$

$$= \prod_{k=1}^{K} \pi_k^{\alpha_k + \sum_{m=1}^{M} x_k^{(m)} - 1}$$

Thus $\pi \mid \mathcal{D} \sim \text{Dir}(\alpha_1 + M_1, \cdots, \alpha_K + M_K)$ where $M_k = \sum_{m=1}^{M} x_k^{(m)}$

c) [10 marks] Derive the max a posteriori (MAP) estimation of $\boldsymbol{\pi}$. *Hint:* Notice the constraint $\pi_1 + \cdots + \pi_K = 1$. The constraints of $\pi_k > 0$ (for $k = 1, \cdots, K$) will be automatically satisfied and thus can be ignored. (log)

Define $N_k = \alpha_k + \sum_{m=1}^{M} x_k^{(m)} - 1$

Goal: $\underset{\pi}{\text{maximize}} \prod_{k=1}^{K} \pi_k^{N_k}$

$\Leftrightarrow \underset{\pi}{\text{minimize}} - \sum_{k=1}^{K} N_k \log \pi_k$

Constraints: $\pi_1 + \cdots + \pi_K = 1$

Lagrangian:

$$\mathcal{L} = -\sum_{k=1}^{K} N_k \log \pi_k + \lambda(\pi_1 + \cdots + \pi_K - 1) = 0$$

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \frac{N_k}{\pi_k} + \lambda \overset{\text{set}}{=} 0$$

Thus $\pi_k = -\dfrac{N_k}{\lambda}$

Since $\pi_1 + \cdots + \pi_K = 1$

$$-\frac{N_1}{\lambda} - \cdots - \frac{N_K}{\lambda} = 1$$

$$\lambda = -(N_1 + \cdots + N_K)$$

$$\pi_k = -\frac{N_k}{\lambda} = \frac{N_k}{N_1 + \cdots + N_K} = \frac{\alpha_k + \sum_{m=1}^{M} x_k^{(m)} - 1}{\sum_{k'} \alpha_k + M - K}$$

END

Cheatsheet: For a convex optimization

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{maximize}} \quad & f_0(\boldsymbol{x}) \\ \text{subject to} \quad & f_i(\boldsymbol{x}) \leq 0 \quad \text{for } i = 1, \cdots, m \\ & h_i(\boldsymbol{x}) = 0 \quad \text{for } i = 1, \cdots, n \end{aligned}$$

where f_i is a convex function and h_i is an affine function. The Lagrangian is defined to be

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}) + \sum_{i=1}^{n} \nu_i h_i(\boldsymbol{x})$$

For a differentiable convex optimization problem, the sufficient and necessary conditions for the optimality are

1) $f_i(\boldsymbol{x}) \leq 0$ for $i = 1, \cdots, m$
2) $h_i(\boldsymbol{x}) = 0$ for $i = 1, \cdots, n$
3) $\lambda_i \geq 0$ for $i = 1, \cdots, m$
4) $\lambda_i f_i(\boldsymbol{x}) = 0$ for $i = 1, \cdots, m$
5) $\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = 0$

Scrap paper. May be detached. Additional paper is available upon request as appropriate.
May be used as answer sheets if you

- Print your name and ID on every answer sheet (including additional sheets) submitted (1 bonus mark)
- Mark your solution and corresponding problem number clearly
- Submit the sheet by the end of the exam