CS 769 Spring 2025 | Midsem

Date: 23rd February /Time: 11:30 AM - 1:30 PM

Total Marks: **24**, Weightage: **20**

Instructions

- This midsem will also be a closed-book exam. However, students can carry 2 pages of hand-written notes. You can interpret 2 pages as 2 sheets of paper which will mean 4 sides of handwritten notes. No other books, notebooks and printed material are allowed. I repeat that other books, notebooks and printed material are NOT allowed.
- No form of collaboration or discussion is allowed.
- No laptops or cell phones are allowed.
- This exam consists of 6 problems. The maximum possible score is 24. Overall weightage (with respect to the course) is 20%
- Write your answers legibly in the answer sheet. (If necessary, use/ask for extra sheets to work out your solutions. These sheets will not be graded.)
- Work efficiently. The questions are not sorted in any order of difficulty. Try and attempt the easier ones first, so that you are not bogged down by the harder questions.
- Good luck!

Problem 1 (6 Marks)

Consider the following loss functions involving regularization. Assume $y_i \in \{-1, +1\}$, $\boldsymbol{\theta} \in \mathbb{R}^n$, $x_i \in \mathbb{R}^n$ for $i = 1 \ldots n$ and $\lambda > 0$, $\mu > 0$, $\nu > 0$, $\lambda_1 \geq 0$, $\lambda_2 \geq 0$. It should be evident that the only variable is $\boldsymbol{\theta} \in \mathbb{R}^n$.

1. $L_1(\boldsymbol{\theta}) = \lambda_1 \|\boldsymbol{\theta}\|_1 + \lambda_2 \|\boldsymbol{\theta}\|_2^2 + \sum_{i=1}^n \log(1 + \exp(-y_i \boldsymbol{\theta}^T x_i))$.
2. $L_2(\boldsymbol{\theta}) = \lambda \|\boldsymbol{\theta}\|_1 + \frac{1}{2} \boldsymbol{\theta}^T H \boldsymbol{\theta} - b^T \boldsymbol{\theta}$, where $H \succ 0$ (positive definite matrix in $\mathbb{R}^{n \times n}$).
3. $L_3(\boldsymbol{\theta}) = \sum_{i=1}^n \log(1 + \exp(-y_i \boldsymbol{\theta}^T x_i)) + \frac{\mu}{2} \|\boldsymbol{\theta}\|_2^2 + \nu \|\boldsymbol{\theta}\|_1$.

Answer the following questions:

(a) Identify which of these functions are Lipschitz smooth and provide a proof or counterexample. (3 marks)

(b) Which of these functions are strongly convex and under what conditions? Justify your answer using necessary and sufficient conditions. (3 marks)

Solution Outline:

- **(a) Lipschitz Smoothness:** For most parts, please refer to Lecture 12 `https://moodle.iitb.ac.in/pluginfile.php/143876/mod_resource/content/13/CS769_2025___Lecture_12-annotated.pdf`. Function 1 and Function 3 are **not** Lipschitz smooth due to the L1 term. However, Function 2 is Lipschitz smooth because the quadratic term dominates and H is positive definite. That is, the hessian will be upper bounded by the maximum of the eigenvalues of H. (3 marks)

- **(b) Strong Convexity:** For most parts, please refer to Lecture 6 at `https://moodle.iitb.ac.in/pluginfile.php/146799/mod_resource/content/4/CS769_2025___Lecture_6-annotated.pdf`. All the three functions are convex for reasons discussed in the class: Log-sum-exp, L_1 and L_2 norm are all convex in θ and so is the quadratic form for positive definite H. Now coming to strong convexity, Function 3 is strongly convex due to the presence of $\frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2$. Function 2 is strongly convex since $H \succ 0$. Function 1 is strongly convex when $\lambda_2 > 0$ but it is **not** strongly convex when $\lambda_2 = 0$, because the L1 norm does not induce strong convexity. (3 marks)

Problem 2 (4 Marks)

Consider the function $f(\mathbf{w})$ defined as follows:

$$f(\mathbf{w}) = \max\left(\|\mathbf{w}\|_2 - \lambda, 0\right) + \mu \max\left(\mathbf{w}^T\mathbf{x}, 0\right), \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^n$ is the only variable and $\mathbf{x} \in \mathbb{R}^n$ is a fixed vector, and $\lambda, \mu > 0$ are also specified to you.

(a) Prove or disprove that $f(\mathbf{w})$ is convex for all $\mathbf{x}$ and for all $\lambda, \mu > 0$. (2 marks)

(b) Compute the subdifferential $\partial f(\mathbf{w})$ at an arbitrary point $\mathbf{w}$. Clearly state and justify different cases based on $\mathbf{w}$. (2 marks)

Solution Outline:

- **(a) Convexity Proof:** For most parts, please refer to Lecture 6 at `https://moodle.iitb.ac.in/pluginfile.php/146799/mod_resource/content/4/CS769_2025___Lecture_6-annotated.pdf`. $f(\mathbf{w})$ is the sum of two convex functions: $\max(\|\mathbf{w}\|_2 - \lambda, 0)$ and $\max(\mathbf{w}^T\mathbf{x}, 0)$. The proof follows from properties of convex functions.

- **(b) Subdifferential:** For most parts, please refer to Lecture 9 at `https://moodle.iitb.ac.in/pluginfile.php/148061/mod_resource/content/9/CS769_2025___Lecture_9-annotated.pdf`. The subdifferential is piecewise-defined based on whether $\mathbf{w}$ is inside or outside the threshold region defined by λ and $\mathbf{x}$. The result follows from subgradients of max functions.

Problem 3 (2 Marks)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable, strictly convex function. The Bregman divergence $D_f(x, y)$ associated with f is defined as:

$$D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle. \tag{2}$$

Suppose that the Bregman divergence $D_f(x, y)$ of the function f satisfies the following inequality for all x, y:

$$D_f(x, y) \geq \frac{\mu}{2} \|x - y\|^2. \tag{3}$$

for a constant $\mu > 0$

Furthermore, let us assume that the function f is Lipschitz smooth with constant L.

What can you say as specifically as possible regarding the convergence rate of the gradient descent to minimize f? You can build upon any of the convergence results proved in the class and **need not** prove the convergence result again here. (2 marks)

Solution Outline: First we show that the quadratic function as the lower bound of the bregman divergence $D_f(x, y$ is equivalent to the strong convexity of f. Even if the student proves it one way - which is shows that the the quadratic function as the lower bound of the bregman divergence $D_f(x, y$ implies the strong convexity of f (part (a)), that should be sufficient.

- **(a) Bregman Lower Bound Implies Strong Convexity:** We assume that for all x, y, the inequality

$$D_f(x, y) \geq \frac{\mu}{2} \|x - y\|^2 \tag{4}$$

holds. Expanding the definition of $D_f(x, y)$:

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq \frac{\mu}{2} \|x - y\|^2. \tag{5}$$

Rearranging this, we obtain:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|^2. \tag{6}$$

This is exactly the definition of strong convexity with parameter μ. (1 marks)

- **(b) Strong Convexity Implies Bregman Lower Bound:** Strong convexity of f means that for some $\mu > 0$,

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|^2. \tag{7}$$

Substituting the definition of Bregman divergence, we obtain:

$$D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq \frac{\mu}{2} \|x - y\|^2. \tag{8}$$

This confirms the required inequality. (0 marks)

- **(c) Invoking result from class for Convergence Rate in Gradient Descent:** We have already seen in Lecture 14 `https://moodle.iitb.ac.in/pluginfile.php/143880/mod_resource/content/15/CS769_2025___Lecture_14-annotated.pdf` (and hints from Lecture 13 (`https://moodle.iitb.ac.in/pluginfile.php/151605/mod_resource/content/12/CS769_2025___Lecture_13-annotated.pdf`) that Strong convexity, in conjunction with Lipschitz smoothness implies that gradient descent with step size η satisfies the inequality:

  $$\|x_{k+1} - x^*\|^2 \leq (1 - \eta\mu) \|x_k - x^*\|^2. \tag{9}$$

  As seen in the class, This means that the distance to the optimal solution decreases geometrically, leading to an exponential convergence rate:

  $$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right) (f(x_k) - f(x^*)). \tag{10}$$

  This confirms quadratic convergence in terms of function values. (1 marks)

Problem 4 (4 Marks)

A differentiable function $g : \mathbb{R}^d \to \mathbb{R}$ is said to satisfy the Restricted Strong Smoothness Property (RSSP) with parameter $\beta_k > 0$, if for all vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in $\mathbb{R}^d$ with $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_0 \leq k$, it holds that

$$g(\boldsymbol{\theta}_2) \leq g(\boldsymbol{\theta}_1) + (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)^T \nabla g(\boldsymbol{\theta}_1) + \frac{\beta_k}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2.$$

Here, $||\mathbf{u}||_0$ is the number of nonzero elements in $\mathbf{u}$. Assume all vectors are column vectors unless stated otherwise.

Consider a statistical setting where the function g corresponds to the empirical risk function with a logistic loss. More precisely, suppose we observe n input-output pairs $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$ sampled *i.i.d.* from some distribution P, following the model:

1. The conditional probability is given by $P(Y = 1|X = \mathbf{x}) = \sigma(\mathbf{x}^T\boldsymbol{\theta})$, where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function, and $\boldsymbol{\theta} \in \mathbb{R}^d$ is an unknown parameter.
2. The empirical risk function is defined as:

   $$g(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-y_i \mathbf{x}_i^T \boldsymbol{\theta}}).$$

1. Show that $g(\boldsymbol{\theta})$ satisfies the Restricted Strong Smoothness Property (RSSP) with an appropriate parameter β_k. Express β_k in terms of properties of the data $(\mathbf{x}_i)_{i=1}^n$. (3 marks)
2. Qualitatively (verbally) discuss the significance of RSSP in high-dimensional learning problems with respect to convergence of an algorithm such as gradient descent. (1 marks)

Solution and Marking Scheme:

This is the final sol for Q1: $\lambda_{\max,k}$ be the largest eigenvalue of the submatrix corresponding to the subset of coordinates where $\boldsymbol{\theta}_1$ and ($\boldsymbol{\theta}_2$ where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ differ by at most k nonzero entries).

(a) Showing Restricted Strong Smoothness Property (RSSP) (3 Marks) The splitting of marks given below is only indicative. The students could have solved the problem differently in which case you can reweight appropriately.

1. Compute the Hessian $\nabla^2 g(\boldsymbol{\theta})$ explicitly. Recall that for the logistic loss, the Hessian is given by:

   $$\nabla^2 g(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \sigma(x_i^T\boldsymbol{\theta})(1-\sigma(x_i^T\boldsymbol{\theta}))x_i x_i^T.$$

   This term provides an upper bound on the curvature of $g(\theta)$. (1 Mark)

2. Derive the maximum eigenvalue of the Hessian and establish that it is upper-bounded by a quantity dependent on the data. In particular, we can bound the Hessian as:

   $$\nabla^2 g(\boldsymbol{\theta}) \preceq \frac{1}{4n}\sum_{i=1}^{n} x_i x_i^T$$

   This follows from the fact that $\sigma(z)(1-\sigma(z)) \leq 1/4$ for all z. (1 Mark)

3. Using the sparsity constraint $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_0 \leq k$, we derive that the restricted smoothness parameter β_k given by the largest eigenvalue of the restricted data covariance matrix:

   $$\beta_k = \frac{1}{4n}\lambda_{\max}\left(\sum_{i=1}^{n} x_i x_i^T\right),$$

   where X_k denotes the subset of X with at most k nonzero rows. (1 Mark)

(b) Discussion on restricted strong smoothness and its qualitative effect on convergence (1 Marks) - Explain why RSSP is crucial in optimization algorithms such as gradient descent. The restricted smoothness property ensures that the function does not change too rapidly, which guarantees stable updates when using gradient-based optimization methods. Without RSS, gradient descent steps might be too aggressive, leading to instability. (1 Mark)

Problem 5 (4 Marks)

In many real-world applications, classification problems involve ambiguous or uncertain labels. One way to handle such ambiguity is by incorporating probabilistic label representations into an optimization framework. Your task is to analyze and formulate an optimization problem for classification using probabilistic labels.

Consider a classification problem with C classes, i.e., $\mathcal{C} = \{c_1, c_2, \ldots, c_C\}$, where each instance $\mathbf{x}_i$ is associated with a probability distribution over the classes instead of a single deterministic label. Specifically, the ground truth dataset consists of n examples:

$$\mathcal{D} = \{(\mathbf{x}_1, \mathbf{p}_1), (\mathbf{x}_2, \mathbf{p}_2), \ldots, (\mathbf{x}_n, \mathbf{p}_n)\},$$

where $\mathbf{p}_i = [p_{i1}, p_{i2}, \ldots, p_{iC}]$ is a probability vector satisfying $\sum_{j=1}^{C} p_{ij} = 1$ and $p_{ij} \geq 0$ for all j.

We define a classifier parameterized by $\boldsymbol{\theta}$ that outputs class probabilities $q_{ij}(\boldsymbol{\theta})$ for each instance $\mathbf{x}_i$ and class c_j. The objective is to learn $\boldsymbol{\theta}$ by minimizing a loss function based on the cross-entropy between the ground truth probabilities and the predicted probabilities:

$$L(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{C} p_{ij} \log q_{ij}(\theta).$$

1. Present the problem as a constrained optimization problem, incorporating additional constraints to ensure valid probability distributions for $q_{ij}(\boldsymbol{\theta})$. Try and reformulate the problem so that the constraints can be removed and incorporated directly into the objective. (2 marks)

2. Derive the gradient of the loss function with respect to $\boldsymbol{\theta}$ and discuss its implications in terms of optimization. (2 marks)

Solution and Marking Scheme:
(a) Reformulating as a Constrained Optimization Problem (2 Marks)

- Introduce constraints ensuring that the probabilities sum to one for each instance:

$$\sum_{j=1}^{C} q_{ij}(\boldsymbol{\theta}) = 1, \quad \forall i.$$

(1 Mark)

- Recap the softmax relaxation in contextual bandits from Lecture 3 `https://moodle.iitb.ac.in/pluginfile.php/143858/mod_resource/content/4/CS769_2025__Lecture_3-annotated.pdf` as well as use of softmax in Lecture 1 for Logistic Regression `https://moodle.iitb.ac.in/pluginfile.php/143894/mod_resource/content/7/CS769_2025__Intro_Lecture_1-annotated.pdf` which stayed through most subsequent lectures. Taking inspiration from all these uses of softmax, express $q_{ij}(\boldsymbol{\theta})$ using a softmax function:

$$q_{ij}(\boldsymbol{\theta}) = \frac{e^{z_{ij}(\boldsymbol{\theta})}}{\sum_{k=1}^{C} e^{z_{ik}(\boldsymbol{\theta})}}$$

where $z_{ij}(\boldsymbol{\theta})$ are the raw class scores produced by the classifier. (1 Mark)

Now identify the role of softmax: Since softmax inherently ensures that the probabilities sum to one, it naturally satisfies the constraints imposed in part (a). This property simplifies constrained optimization into an unconstrained problem in practical implementation. (1 Mark)

- Reformulate the constrained optimization problem using the Lagrangian method:

$$\mathcal{L}(\boldsymbol{\theta}, \lambda) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{C} p_{ij} \log q_{ij}(\boldsymbol{\theta}) + \sum_{i=1}^{n} \lambda_i \left(\sum_{j=1}^{C} q_{ij}(\boldsymbol{\theta}) - 1 \right).$$

  (Deferred this to beyond midsem so this wont carry weightage)

(b) Deriving the Gradient and Its Implications (2 Marks)

- Compute the gradient of the loss function with respect to $z_{ij}(\theta)$:

$$\frac{\partial L}{\partial z_{ij}} = q_{ij}(\boldsymbol{\theta}) - p_{ij}.$$

  This result follows from differentiating the cross-entropy loss. (1 Mark)

- Discuss the interpretation: The gradient shows that updating the parameters in gradient descent pushes $q_{ij}(\boldsymbol{\theta})$ towards p_{ij}, aligning predicted probabilities with the given label distributions. (1 Mark)

Problem 6 (4 Marks)

Find and correct the flawed steps in our attempt to prove that for a Lipschitz smooth convex function f with smoothness constant L, gradient descent with step size $\gamma = \frac{1}{L}$ achieves a solution x_T *s.t.* $|f(x_T) - f(x^*)| \leq \epsilon$ in $\frac{R^2 L}{\epsilon}$ iterations. That is, for smooth convex functions $f(x_T) - f(x^*) \leq \frac{L}{2T}||x_0 - x^*||^2 = \frac{LR^2}{2T}$ and therefore to ensure that $f(x_T) - f(x^*) \leq \epsilon$, we require $\frac{LR^2}{2T} \leq \epsilon$ which implies that $T \geq \frac{R^2 L}{2\epsilon}$. Following are the steps of our attempted proof. You need to identify all the flawed steps

Within that red highlight, I have presented the correct solution)

- **Gradient Descent for Smooth Functions**

  1. Based on Lipschitz smoothness of f we have $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{L}{2}||\mathbf{y} - \mathbf{x}||^2$

$$f(x_{t+1}) \leq f(x_t) - g_t^T(x_{t+1} - x_t) + \frac{L}{2}||x_{t+1} - x_t||^2 \tag{11}$$

$$\leq f(x_t) + \gamma||g_t||^2 + \frac{L}{2}\gamma^2||g_t||^2 \tag{12}$$

  2. For step size $\gamma = 1/L$, $f(x_t) - \gamma||g_t||^2 + \frac{L}{2}\gamma^2||g_t||^2$ gets minimized. With this γ, the above result becomes:

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}||g_t||^2$$

  3. $f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}||g_t||^2 \Rightarrow \frac{1}{2L}||g_t||^2 \leq f(x_t) - f(x_{t+1})$

4. Summing the inequality in step 4 for $t = 0$ to $T - 1$:

$$\frac{1}{2L}\sum_{t=0}^{T-1} ||g_t||^2 \leq \sum_{t=0}^{T-1} [f(x_t) - f(x_{t+1})] = [f(x_0) - f(x_T)] \tag{13}$$

- **Invoking Convexity and Simple Expansion:** We now continue the analysis by invoking convexity.

6. From the definition of Gradient descent:

$$g_t^T(x_t - x^*) = \frac{1}{\gamma}(x_t - x_{t+1})^T(x_t - x^*)$$

7. Note that $2v^T w = ||v||^2 + ||w||^2 - ||v - w||^2$
8. We can then rewrite the RHS in step 6 as:

$$\begin{aligned} g_t^T(x_t - x^*) &= \frac{1}{2\gamma}(||x_t - x_{t+1}||^2 + ||x_t - x^*||^2 - ||x_{t+1} - x^*||^2) \\ &= \frac{\gamma}{2}||g_t||^2 + \frac{1}{2\gamma}(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2) \end{aligned} \tag{14}$$

9. Summing (14) in step 8 over $t = 0...T - 1$ iterations:

$$\sum_{t=0}^{T-1} g_t^T(x_t - x^*) = \frac{1}{2\gamma}(||x_0 - x^*||^2 - ||x_T - x^*||^2) + \frac{\gamma}{2}\sum_{t=0}^{T-1} ||g_t||^2$$

10. Invoking convexity with $x = x_t, y = x^*$,

$$f(x_t) - f(x^*) \leq g_t^T(x_t - x^*) \tag{15}$$

11. Recalling from Step 9

$$\sum_{t=0}^{T-1} g_t^T(x_t - x^*) = \frac{1}{2\gamma}(||x_0 - x^*||^2 - ||x_T - x^*||^2) + \frac{\gamma}{2}\sum_{t=0}^{T-1} ||g_t||^2$$

which, based on $||x_T - x^*||^2 > 0$, implies:

$$\sum_{t=0}^{T-1} g_t^T(x_t - x^*) \leq \frac{\gamma}{2}\sum_{t=0}^{T-1} ||g_t||^2 - \frac{1}{2\gamma}(||x_0 - x^*||^2) \tag{16}$$

12. Combining (15) in step 10 with (16) in step 11, we have:

$$\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq \frac{\gamma}{2}\sum_{t=0}^{T-1}||g_t||^2 - \frac{1}{2\gamma}(||x_1 - x^*||^2)$$

13. The RHS of step 12, on setting $\gamma = 1/L$, yields

$$\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq \frac{1}{2L}\sum_{t=0}^{T-1}||g_t||^2 + \frac{L}{2}(||x_0 - x^*||^2)$$

14. Further, on invoking (13) in step 5 on part of the RHS above

$$\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq f(x_0) - f(x_T) + \frac{L}{2}||x_0 - x^*||^2$$

- **Reinvoking Gradient Descent for Smooth Functions: III**

15. Re-writing the expression in step 14:

$$\sum_{t=0}^{T}(f(x_t) - f(x^*)) \leq \frac{L}{2}||x_0 - x^*||^2$$

16. The inequality in step 15 implies

$$f(x_T) - f(x^*) \leq \sum_{t=0}^{T}\frac{(f(x_t) - f(x^*))}{T} \leq \frac{L}{2T}||x_0 - x^*||^2$$