Optimization in Machine Learning (CS 769) **Endsem**
Computer Science and Engineering 30-04-2024 Tuesday, Total time 3 hours
Indian Institute of Technology Bombay

NAME _____ ROLL NUMBER _____

Instructions

1. This paper has 7 questions. The maximum marks are 36.
2. There are two options for Question 6. The students can attempt any one of those.
3. Total time for the examination is 3 hours.
4. This is a closed-book exam. However, students can carry 5 pages of hand-written pr printed notes as announced earlier. No other books, notebooks and printed material are allowed.

CS 769 Spring 2023-24 | Endsem

Date: 30 April /Time: 9:30 AM - 12:30 PM

Total Marks: **36**, Weightage: **30**

Question No: 1: Objective questions including those on Stochastic algorithms

I Which of the following statements is/are FALSE? You need to get all your option(s) correct to get any marks at all. Also, you need to provide a brief explanation for your selected option.

/2

a Adam is an Adaptive variant of Heavy Ball Momentum

b AdaMax is an extension of ADAM that uses the l_∞ norm (*i.e.*, max) instead of square.

c A benefit of AdaGrad is that it sets a fixed learning rate. AdaGrad can significantly improve upon SGD in dense feature sets.

d Though adaptive gradient methods tend to minimize training loss better, they often do so by obtaining more complex and less generalizable solutions!

Error is in (c): A benefit of AdaGrad is that AdaGrad can significantly improve upon SGD in sparse feature sets! It automatically sets the learning rate, and automatically updates the learning rates with a decay schedule.

II Which of the following statements is/are FALSE? You need to get all your option(s) correct to get any marks at all. Also, you need to provide a brief explanation for your selected option.

/2

a Adaptive Gradient Descent Algorithms try to automatically adapt the learning rate.

b Several adaptive and non-adaptive variants can be unified into the following single update equation:

$$w_{t+1} = w_t - \alpha_t H_t^{-1} \nabla f_{i_t}(w_t + \gamma_t(w_t - w_{t-1})) + \beta_t H_t^{-1} H_{t-1}(w_t - w_{t-1})$$

for different choices of G_t, α_t and β_t and with $H_t = \sqrt{G_t}$ (elementwise square root)

c A simplest Adaptive Algorithm called AdaGrad is:

$$w_{t+1} = w_t + \alpha_t H_t^{-1} \nabla f(w_t)$$

where $\alpha_t > 0$, and $g_t = \nabla f(w_t)$ is gradient at the t^{th} iteration and

$$H_t = \text{diag}\left(\left[\sum_{l=1}^{t} \eta_l g_l g_l^T\right]^{1/2}\right) \implies \left[H_t^{-1}\right]_{jj} = \frac{1}{\sqrt{\sum_{l=1}^{t} \eta_l g_{lj}^2}}$$

d Stochastic Momentum improves upon SGD via Momentum update (similar to the Gradient Descent case):

$$w_{t+1} = w_t - \alpha_t \nabla f_{i_t}(w_t + \gamma_t(w_t - w_{t-1})) + \beta_t(w_t - w_{t-1})$$

for different choices of γ_t and β_t

Option (c) is wrong. $w_{t+1} = w_t - \alpha_t H_t^{-1} \nabla f(w_t)$

III Which of the following is NOT a characteristic of stochastic gradient descent (SGD) algorithms? Also, you need to provide a brief explanation for your selected option.

/2

a The convergence analysis of SGD leverages inequalities and equalities in 'expectation' sense. For example, that the vector $g_t^i := \nabla f_i(w_t)$ - a stochastic gradient (a vector of d random variables) with respect to some random data index i is an unbiased estimate of the gradient:

$$\mathbb{E}\left[g_t^i | w_t = w\right] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w) = \nabla f(w), \ w \in \Re^d$$

b SGD is specifically relevant to minimization of Machine Learning loss objectives with respect to parameters w which can be decomposed as a sum over a large number of losses: $L(w) = \frac{1}{n}\sum_{i=1}^{n} L_i(w)$ where L_i is the loss for the i^{th} training example, and n is the total number of training examples.

c Convergence analysis of SGD can never benefit from convexity or strong convexity of the loss function being minimized.

d Stochastic Average Gradient (SAG) is the Hybrid of stochastic gradient with full gradient and SAG incurs more storage overhead (of storing more gradients) than SGD.

Statement (c) is False: Let $f : \Re^d \to \Re$ be convex and differentiable, w_* a global minimum; furthermore, suppose that $||w_0 - w_*|| \leq R$, and that $\mathbb{E}\left[||g_t||^2\right] \leq B^2$ for all t. Choosing the constant stepsize

$$\gamma = \frac{R}{B\sqrt{T}}$$

stochastic gradient descent yields

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[f(w_t)\right] - f(w_*) \leq \frac{RB}{\sqrt{T}}$$

Let $f : \Re^d \to \Re$ be differentiable and strongly convex with parameter $\mu > 0$; let w_* be the unique global minimum of f. With decreasing step size specified as

$$\gamma_t := \frac{2}{\mu(t+1)}$$

stochastic gradient descent yields

$$\mathbf{E}\left[f\left(\frac{2}{T(T+1)}\sum_{t=1}^{T} t.w_t\right) - f(w_*)\right] \leq \frac{2B^2}{\mu(T+1)}$$

where $B^2 := max_{t=1}^{T}\ \mathbb{E}\left[||g_t||^2\right]$

IV Which of the following statements is TRUE concerning the theoretical analysis concerning the number of iterations required for attaining an ϵ-approximate solution? Also, you need to provide a brief explanation for your selected option.

/2

a Using gradient descent, smoothness along with strong convexity of the minimization objective, always results in a larger number of iterations than Lipschitz continuity along with strong convexity of the minimization objective

b Using gradient descent, smoothness of the minimization objective, always results in a larger number of iterations than Lipschitz continuity of the minimization objective

c Let $f(x)$ be convex, differentiable, and ∇f be Lipschitz continuous with constant $L > 0$ AND $c(x)$ be convex. Let $F(x) = f(x) + c(x)$. Then, using generalized gradient descent on the function F (assuming $prox_\gamma(\mathbf{z})$ can be solved exactly) generally results in fewer number of iterations than use of subgradient descent on the same function F.

d There is a gap in the convergence analysis between the Generalized gradient descent algorithm and the subgradient descent algorithms, and the Nestorov's acceleration helps fix that gap.

(a) is false. Using gradient descent, smoothness along with strong convexity of the minimization objective results in fewer iterations than Lipschitz continuity along with strong convexity of the minimization objective. Lipschitz continuous + Strongly Convex Functions (CS) with $\gamma_t = \mu(1+t)/2$ attain an an ϵ-approximate solution in $2B^2/\epsilon + 1$ iterations. Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2 L}{2\epsilon})$ iterations.
(b) is also false. Using gradient descent, smoothness of the minimization objective results in fewer number of iterations than Lipschitz continuity of the minimization objective. Lipschitz continuous Functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, attain an ϵ-approximate solution in R^2B^2/ϵ^2 iterations. Smooth Functions (S): With $\gamma = 1/L$, attain an ϵ-approximate solution in $\frac{R^2 L}{\epsilon}$ iterations.
(c) is true. For Subgradient Descent: The subgradient method has convergence rate $\Omega(1/\sqrt{T})$; to get $f(\mathbf{x}_T^{best}) - f(\mathbf{x}_*) \leq \epsilon$, we need $\Omega(1/\epsilon^2)$ iterations. For generalized Gradient Descent: If $f(x)$ is convex, differentiable, and ∇f is Lipschitz continuous with constant $L > 0$ AND $c(x)$ is convex and $prox_\gamma(x)$ can be solved exactly then convergence result (and proof) is similar to that for gradient descent!!

$$f(x_T) - f(x_*) \leq \frac{1}{T}\sum_{t=1}^{T}(f(x_t) - f(x_*)) \leq \frac{\|x_0 - x_*\|^2}{2T\gamma}$$

(d) is completely wrong. Nestorov's acceleration helps fix the gap between the upper bound and the analysed bound for the Lipschitz smooth case. While Nestorov helps improve the bound on the generalized gradient descent, it does not bridge the gap between Subgradient descent and generalized gradient descent algorithms.

Question 2: Submodular Subset Selection

1. Write a summary listing 4 different kinds of set functions (several of them could be submodular but all need not be submodular) categorizing them suitably as discussed in the class.

   /2

2. Also point out which of them are interesting for maximization and which of them for minimization.

   /2

3. Finally, write down the mathematical forms of those four objective functions

   /2

As described in https://moodle.iitb.ac.in/pluginfile.php/103265/mod_resource/content/23/Lecture23-CS769-2023-Final-SubmodularOpt1-latest-annotated-pages-1-39.pdf, submodular and related set functions can be classified into

1. The following functions are natural choices for minimization: Cooperative costs, Attractive Potentials, Complexity,
2. The following functions are natural choices for maximization: Representation, Diversity, Coverage

Question 3: Projection

Often, in optimization problems in machine learning, we have a simple constraint[1] requiring that parameters lie in a particular interval. Such optimization problems can be effectively solved using the **projected gradient descent algorithm** discussed in the class.

Derive the exact **projection operation** of the **projected gradient descent algorithm**, for the following optimization problem which has the simplest form of such an interval constraint:

$$\begin{aligned} &\min_{\mathbf{x}\in\Re} && f(\mathbf{x}) \\ &\text{subject to } \mathbf{x} \leq r \text{ and } \mathbf{x} \geq l \end{aligned} \tag{1}$$

for some fixed given scalars $l < r \in \Re$ and for some machine learning convex loss function $f(\mathbf{x})$. You can use the Karush-Kuhn-Tucker (KKT) conditions for deriving the projection step.

/5

[1] For example, in Support Vector Classification, the constraint is that each parameter $\xi_i \in [0, C]$ for some fixed value of C.

$$P_C(\mathbf{z}) = prox_{I_C}(\mathbf{z}) = \underset{\mathbf{x}\in C}{\operatorname{argmin}} \frac{1}{2\gamma}||\mathbf{x}-\mathbf{z}||^2 + I_C(\mathbf{x}) = \underset{\mathbf{x}\in C}{\operatorname{argmin}} \frac{1}{2\gamma}||\mathbf{x}-\mathbf{z}||^2$$

The projection step amounts to solving the following optimization problem

$$\begin{array}{ll} \min\limits_{\mathbf{x}\in\Re} & \frac{1}{2\gamma}||\mathbf{x}-\mathbf{z}||^2 \\ \text{subject to } \mathbf{x} \leq r \text{ and} & -\mathbf{x} \leq -l \end{array} \tag{2}$$

Note that $||\mathbf{x}-\mathbf{z}||^2$ is a convex function of the scalar $\mathbf{x}$ and the inequality constraints are both linear.

We can now write down the KKT necessary and sufficient conditions for optimality as

$$\begin{array}{ll} (1) & \nabla\frac{1}{2\gamma}||\mathbf{x}-\mathbf{z}||^2 + \lambda_1\nabla(\mathbf{x}-r) + \lambda_2\nabla(-\mathbf{x}+l) = \mathbf{0} \\ (2a) & \mathbf{x}-r \leq 0 \\ (2b) & -\mathbf{x}+l \leq 0 \\ (4a) & \widehat{\lambda}_1(\mathbf{x}-r) = 0 \\ (4b) & \widehat{\lambda}_2(-\mathbf{x}+l) = 0 \end{array} \tag{3}$$

Simplifying

$$\begin{array}{ll} (1) & \frac{1}{\gamma}(\mathbf{x}-\mathbf{z}) + \lambda_1 - \lambda_2 = \mathbf{0} \\ (2a) & \mathbf{x}-r \leq 0 \\ (2b) & -\mathbf{x}+l \leq 0 \\ (4a) & \widehat{\lambda}_1(\mathbf{x}-r) = 0 \\ (4b) & \widehat{\lambda}_2(-\mathbf{x}+l) = 0 \end{array} \tag{4}$$

This implies that, at point of optimality $l \leq x \leq r$, and

- if $l < x < r$ then $\widehat{\lambda}_1 = 0$ and $\widehat{\lambda}_2 = 0$ implying that $x = z$
- else $x = l$ of $x = r$

Which summarily means, the projection operation is $P_C(\mathbf{z}) = min\{max\{z, l\}, r\}$

Recall that the Prox operator of a function h for some argument z is $\text{prox}_h(z) = \text{argmin}_x \frac{1}{2\gamma}||x-z||^2 + h(x)$.

Here the function $h(x)$ is the $I_C(x)$ indicator function.

Question 4 [Option 1] (5 Marks)

Consider the following optimization problem in (5), in the backdrop of the notation in Table 1. The problem concerns the learning the adjacency matrix W of a graph of N nodes through its Laplacian L. The learning is based on measurements taken over the N entities over T time steps. All the definitions are in the Table 1.

Notation	Dimension	Meaning
N	scalar	**Total** Number of **Entities** among which we wish to learn the graphical relation.
T	scalar	Total **Time Steps** recorded over Entities.
X	N $\times$ T	**Signals**: T measurements taken over each of the N Entities.
V	N $\times$ N	**Eigen Vectors** of the Graph Laplacian **L**.
Δ	N $\times$ N	A Diagonal Matrix storing **Eigen Values** of the Graph Laplacian **L**.
α	N $\times$ T	Representation of Signals in Fourier Space with Basis V also called as **GFT** coefficients. (**G**raph **F**ourier **T**ransform)
W	N $\times$ N	**Adjacency Matrix** of the underlying Graphical relation between the entities that we wish to learn.
D	N $\times$ N	Diagonal Matrix storing **Degrees** of each Node (or Entity) in the present in the Graph. Mathematically expressed as $D_{i,i} = \sum_j W_{i,j}$.
L	N $\times$ N	Combinatorial Graph Laplacian Matrix L (or **Unnormalized Graph Laplacian L**) corresponding to the Adjacency Matrix W. Mathematically expressed as $L = D - W$. The Eigen Decomposition of Graph Laplacian L is expressed as $L = V \cdot \Delta \cdot V^\top$
$\hat{\Theta}$	N $\times$ N	**Inverse Covariance Matrix** Corresponding to the Signals stored in X. Mathematically expressed as $\hat{\Theta} = \left(\frac{X \cdot X^\top}{T-1}\right)^{-1}$.
σ	scalar	The Relation between the Inverse Covariance Matrix and the Combinatorial Graph Laplacian Matrix L is expressed as $\hat{\Theta} = L + \frac{1}{\sigma^2} \cdot I$, here the parameter σ, controls the amount of **regularization**.

Table 1: Table of notations

$$\min_{\alpha,\Delta,V} \|X - V\cdot\alpha\|_F^2 + \mu_1 \cdot \left\|\Delta^{1/2}\cdot\alpha\right\|_F^2 + \mu_2 \cdot \|\alpha\|_S$$

$$\text{s.t.} \quad \begin{cases} V^\top V = I_N, v_1 = \frac{1}{\sqrt{N}}\mathbf{1}_N, & \text{(a)} \\ \left(V\Delta V^\top\right)_{k,\ell} \leq 0 \quad k \neq \ell, & \text{(b)} \\ \Delta = \operatorname{diag}(0,\lambda_2,\ldots,\lambda_N) \succeq 0, & \text{(c)} \\ \operatorname{tr}(\Delta) = N \in \mathbb{R}_*^+, & \text{(d)} \end{cases} \tag{5}$$

Now based on the notation in Table 1,

1. explain/interpret each of the three terms in the the objective function in (5). Interpretation of each of the three terms carries 1 mark. /3

2. and similarly interpret any of the two constraints in the objective function in (5) /2

Interpretation of the terms

In the objective function (5), the first term corresponds to the quadratic approximation error of X by $V\cdot\alpha$. The second term is a smoothness regularization equally imposed on each column of $V\cdot\alpha$ and could also be expressed in an alternate way as:

$$\left\|\Delta^{1/2}\alpha\right\|_F^2 = \sum_{i=1}^{N} \lambda_i \left\|\alpha_{i,:}\right\|_2^2,$$

where $\alpha_{i,:}$ is the i-th row of the matrix α. From its definition, we can see that it tends to be low when high values of $\{\lambda_i\}_{i=1}^N$ are associated with rows of α with low ℓ_2-norm. This corroborates the idea that the $\{\lambda_i\}_{i=1}^N$ can be interpreted as frequencies and the elements of α as Fourier coefficients.

Finally, $\mu_2 \cdot \|\alpha\|_S$, is a sparsity regularization. Authors propose to either use the $\ell_{2,1}$ (the sum of the ℓ_2-norm of each row of α) or $\ell_{2,0}$ (the number of rows with ℓ_2-norm different than 0) that induces a row-sparse solution $\widehat{\alpha}$. Remark on the choice of $\|\cdot\|_S$. In the context of GSP, it is natural to assume that the graph signals are bandlimited, all at the same dimensions. This property is enforced by $\|\cdot\|_S$ and has two main advantages: it is a key assumption for sampling over a graph and this particular structure is better for inferring graphs with clusters. Therefore, in this article, the use of the classical ℓ_0-norm and the ℓ_1-norm has not been investigated since they would impose sparsity at every dimension of the matrix α 'independently', which would consequently break the bandlimitedness assumption.

The hyperparameters, $\mu_1, \mu_2 > 0$ are controlling respectively the smoothness of

the approximated signals and the sparsity of α. Finally, the first three constraints (1a), (1b), (1c) enforce $V\Delta V^\top$ to be a Laplacian matrix of an undirected graph. More specifically, by definition, $L = D - W$ with $W \in \mathbb{R}_+^{N\times N}$, thus we necessary have $\forall k \neq \ell, L_{k,\ell} = \left(V\Delta V^\top\right)_{k,\ell} \leq 0$ (constraint (1b)). Furthermore, as $V\Delta V^\top$ is the eigendecomposition of the Laplacian matrix of an undirected graph, $V^\top V = I_N, v_1 = \frac{1}{\sqrt{N}}\mathbf{1}_N$ and $\lambda_1 = 0 \leq \lambda_2 \leq \ldots \leq \lambda_N$ (constraints (1a) and (1c)). The last constraint (1d) was proposed in Dong et al. (2016) as to impose structure in the learned graph while avoiding that the trivial solution $\widehat{\Delta} = 0$.
The objective function (5) is not jointly convex but when $\|\cdot\|_S$ is taken to be the $\ell_{2,1}$ norm, it is convex with respect to each of the block-variables α, Δ, or V, taken independently.

Question 4 [Option 2] (5 Marks)

Let $\mathcal{O} = \{x_1, x_2, \ldots x_n\}$ denote n identically distributed observations and suppose that we are interested in estimating some parameter $\theta \in \Re^p$ of the model/function we are trying to fit to the observations $\mathcal{O}$. Let $L : \Re^p \times \mathcal{O}$ be a convex and differentiable loss function that, for a given set of observations $\mathcal{O}$, assigns a cost $L(\theta; \mathcal{O})$ to any parameter $\theta \in \Re^p$.

Below, we investigate some important aspects of the convex optimization problem $\theta^* \in \arg\min_{\theta\in\Re^p} \{L(\theta; \mathcal{O}) + \lambda\Omega(\theta)\}$ where $\lambda > 0$ is a user-defined regularization penalty and $\Omega : \Re^p \to \Re^+$ is a norm which serves to regularize. Specifically, in the subproblems 1 and 2 below, we investigate an important property of the regularizer $\Omega(.)$, *viz.*, **decomposibility**. This property has been used to prove faster convergence rates of the regularized loss.

a **Orthogonal Complements:** We will set the notation for decomposibility of the regulariser Ω is defined in terms of a pair of subspaces $\mathcal{M} \subseteq \overline{\mathcal{M}}$ in $\Re^p$. An example is when $\mathcal{M} = \overline{\mathcal{M}}$. The role of the model subspace $\mathcal{M}$ is to capture the constraints specified by the model; for instance, it might be the subspace of vectors with a particular support (support is the set of non-zero elements of a vector, as we will see soon). The orthogonal complement of the space $\overline{\mathcal{M}}$, namely, the set

$$\overline{\mathcal{M}}^\perp = \left\{v \in \Re^p | \langle u, v\rangle = 0 \ \forall u \in \overline{\mathcal{M}}\right\}$$

represents deviations away from the model subspace $\mathcal{M}$. Recall that $\langle u, v\rangle$ represents the dot product between u and v. The definition of $\mathcal{M}^\perp$ is very

similar

$$\mathcal{M}^{\perp} = \{v \in \Re^p | \langle u, v \rangle = 0 \ \forall u \in \mathcal{M}\}$$

Is it that if $\mathcal{M} \subseteq \overline{\mathcal{M}}$ then we must have that $\overline{\mathcal{M}}^{\perp} \subseteq \mathcal{M}^{\perp}$? State your answer and substantiate it.

/2

Yes. We will prove by contradiction. Let v' be such that $v' \in \Re^p | \langle u, v' \rangle = 0 \ \forall u \in \overline{\mathcal{M}}$ but $\langle u, v' \rangle \neq 0$ *for some* $u \in \mathcal{M}$. But this is not possible since $\mathcal{M} \subseteq \overline{\mathcal{M}}$!

b **Decomposibility of the Regulariser Ω:** Given a pair of subspaces $\mathcal{M} \subseteq \overline{\mathcal{M}}$, a norm-based regularizer Ω is said to be decomposable with respect to $(\mathcal{M}, \mathcal{M}^{\perp})$ if $\Omega(\theta + \gamma) = \Omega(\theta) + \Omega(\gamma)$ for all $\theta \in \mathcal{M}$ and $\gamma \in \overline{\mathcal{M}}^{\perp}$. By the triangle inequality for a norm, we always have $\Omega(\theta + \gamma) \leq \Omega(\theta) + \Omega(\gamma)$. Now if $\mathcal{M} = \overline{\mathcal{M}}$, it is obvious that the decomposability condition above holds if and only if the triangle inequality is tight for all pairs $(\theta, \gamma) \in (\mathcal{M}, \mathcal{M}^{\perp})$.

As an example let us consider the decomposibility of the L_1-norm regularizer. Provide step-wise detailed proof that the L_1-norm is decomposable.

/3

Hint: For any particular subset $S \subseteq \{1, 2, ..., p\}$ with cardinality s, you can define the model subspace $\mathcal{M}(S) := \{\theta \in \Re^p | \theta_j = 0 \ \forall \ j \notin S\}$. Here our notation reflects the fact that $\mathcal{M}$ depends explicitly on the chosen subset S.

SOLUTION. In this case, we may define $\overline{\mathcal{M}}(S) = \mathcal{M}(S)$ and note that the orthogonal complement with respect to the Euclidean inner product is given by $\overline{\mathcal{M}}^{\perp}(S) = \mathcal{M}^{\perp}(S) = \{\gamma \in \Re^p | \gamma_j = 0 \ \forall \ j \in S\}$.
This set corresponds to the perturbation subspace, capturing deviations away from the set of vectors with support S. We claim that for any subset S, the L_1-norm $\Omega(\theta) = \|\theta\|_1$ is decomposable with respect to the pair $(\mathcal{M}(S), \mathcal{M}^{\perp}(S))$. Indeed, by construction of the subspaces, any $\theta \in \mathcal{M}(S)$ can be written in the partitioned form $\theta = (\theta_S, 0_{S^c})$, where $\theta_S \in \Re^s$ and $0_{S^c} \in \Re^{p-s}$ is a vector of zeros, where S^c is the complement of the set S. Similarly, any vector $\gamma \in \mathcal{M}^{\perp}(S)$ has the partitioned representation $(0_S, \gamma_{S^c})$. Putting together the pieces, we obtain

$$\|\theta + \gamma\| = \|(\theta_S, 0) + (0, \gamma_{S^c})\|_1 = \|\theta\|_1 + \|\gamma\|_1$$

showing that the L_1-norm is decomposable as claimed.
Similarly, one can prove that group structured norms on vectors, nuclear norms on low rank matrices, etc are all decomposable.

Question 5: Matrix Norm

Show that the matrix norm (on matrix $A \in \Re^{m \times n}$) induced by vector norm N: $M_N(A) = \sup_{\mathbf{x} \neq 0} \frac{N(A\mathbf{x})}{N(\mathbf{x})}$ is a convex function by using any property of convex function except that you CANNOT assume that $M_N(A)$ is a valid norm. You can only assume the given definition of $M_N(A)$, the fact that N is itself a valid norm and make use of the properties of convex functions. You can also assume the following inequality: $\sup_{\mathbf{x} \in \mathcal{C}} [f(\mathbf{x}) + g(\mathbf{x})] \leq \sup_{\mathbf{x} \in \mathcal{C}} f(\mathbf{x}) + \sup_{\mathbf{x} \in \mathcal{C}} g(\mathbf{x})$

Recall that $\sup_{s \in S} f(s) = \widehat{f}$ if $\widehat{f}$ is the minimum upper bound for $f(s)$ over $s \in S$.
Aside: We saw several instances of the vector induced matrix norm such as the following:

Eg:

$M_N(I) = 1$ (i.e., A is identity matrix I) irrespective of N

If $N = \|.\|_2$, $M_N(A) = \sqrt{\lambda_{max}(A^T A)} = \sigma_{max}(A)$ is the spectral norm, where $\lambda_{max}(A^T A)$ is the dominant eigenvalue of $A^T A$ and $\sigma_{max}(A)$ is called the largest singular value of A $N = \|.\|_1 \implies M_N(A) = \max_j \sum_{i=1}^{n} |a_{ij}|$ $N = \|.\|_\infty \implies M_N(A) = \max_i \sum_{j=1}^{m} |a_{ij}|$

/3

Proof of Spectral Norm is Convex:

$||\theta \mathbf{X_1} + (1-\theta)\mathbf{X_2}||_{spec} = sup_{\mathbf{v}} \frac{||\theta \mathbf{X_1 v} + (1-\theta)\mathbf{X_2 v}||_2}{||\mathbf{v}||_2}$

$\leq sup_{\mathbf{v}} [\frac{\theta ||\mathbf{X_1 v}||_2}{||\mathbf{v}||_2} + \frac{(1-\theta)||\mathbf{X_2 v}||_2}{||\mathbf{v}||_2}]$ *[using Cauchy Schwarz]*

[since supremum/max of sums $\leq$ sum of supremums/maximums]

$= \theta ||\mathbf{X_1}||_{spec} + (1-\theta)||\mathbf{X_2}||_{spec}$

Question 6: Prox Operator

3. Let that the Prox operator of a function h for some argument z is

$$\text{prox}_h(z) = \text{argmin}_x \frac{1}{2\gamma}\|x - z\|^2 + h(x)$$

Compute the Prox operator for $h(x)$ defined as $h(x) = 0$ if $0 \leq x \leq \theta$ (for some fixed $\theta > 0$) and $h(x) = 5000$ for any other value of x.

/4

We consider different cases based on the value of z. Case 1 : $z < 0$ and $\frac{1}{2\gamma}\|z\|^2 > 5000$ In this case, it's better to choose $h(x) = 5000$ and $x = z$. So, we have:

$$\text{prox}_h(z) = z$$

Case 2 : $z < 0$ and $\frac{1}{2\gamma}\|z\|^2 \leq 5000$ In this case, the minimum within the domain $0 \leq x \leq \theta$ is achieved at $x = 0$. So, we have:

$$\text{prox}_h(z) = 0$$

Case 3 : $0 \leq z \leq \theta$ In this case, the minimum is achieved at $x = z$. So, we have:

$$\text{prox}_h(z) = z$$

Case 4 : $z > \theta$ and $\frac{1}{2\gamma}\|z-\theta\|^2 \leq 5000$ In this case, the minimum within the domain $0 \leq x \leq \theta$ is achieved at $x = \theta$. So, we have:

$$\text{prox}_h(z) = \theta$$

Case 5: $z > \theta$ and $\frac{1}{2\gamma}\|z-\theta\|^2 > 5000$ In this case, it's better to choose $h(x) = 5000$ and $x = z$. So, we have:

$$\text{prox}_h(z) = z$$

Combining all cases, the Prox operator for the given function $h(x)$ can be defined as: 4

$$\text{prox}_h(z) = \begin{cases} z, & \text{if } z < 0 \text{ and } \frac{1}{2\gamma}\|z\|^2 > 5000 \\ 0, & \text{if } z < 0 \text{ and } \frac{1}{2\gamma}\|z\|^2 \leq 5000 \\ z, & \text{if } 0 \leq z \leq \theta \\ \theta, & \text{if } z > \theta \text{ and } \frac{1}{2\gamma}\|z-\theta\|^2 \leq 5000 \\ z, & \text{if } z > \theta \text{ and } \frac{1}{2\gamma}\|z-\theta\|^2 > 5000 \end{cases}$$

Question No 7 (5 Marks)

Given the general optimization program

$$\begin{aligned} &\min_{\mathbf{x}\in\mathcal{D}} \quad f(\mathbf{x}) \\ &\text{subject to } g_i(\mathbf{x}) \leq 0,\ i = 1, 2, \ldots, m \\ &\text{subject to } h_j(\mathbf{x}) = 0,\ j = 1, 2, \ldots, p \end{aligned} \tag{6}$$

consider its Lagrangian $L(x, \lambda_1, \ldots, \lambda_m, \mu_1, \ldots, \mu_p)$ with $\lambda_1, \ldots, \lambda_i, \ldots, \lambda_m$ and $\mu_1, \ldots, \mu_j, \ldots, \mu_p$ introduced for the inequality constraints $g_1, \ldots, g_i, \ldots, g_m$ and equality constraints $h_1, \ldots, h_j, \ldots, h_p$ respectively. Which of the following is/are true? Provide very brief justification for the truth/false assertion you make for each of the 5 answers.

a $L(x^*, \lambda, \mu) = L(x, \lambda^*, \mu^*)$ for primal-optimal x^*, dual-optimal (λ^*, μ^*), and all x, λ, μ

b L is convex in x

c L is concave in λ and μ

d $f(x) \geq L(x, \lambda, \mu)$ for all x, λ, μ

e λ_i^* . $h_i(x) = 0$ at dual-optimal λ^* for all x and $i = 1, \ldots, min(m, p)$

/5

Ans: (c). Anyone who attempted problems from the last year's question paper should get this right. See page 14 of https://moodle.iitb.ac.in/pluginfile.php/377354/mod_resource/content/67/CS769_2023___Lecture20-annotated.pdf

Total: 41