CS 769 Spring 2022-23 | Endsem

Date: 24 April /Time: 8:30 AM - 11:30 AM

Total Marks: **35**, Weightage: **35**

Question No 1: (3 Marks)

If $f(\mathbf{x})$ is a quasiconvex function, then $\mathbf{g}$ is a quasigradient at $\mathbf{x}_0$ if

$$\mathbf{g}^T(\mathbf{x}-\mathbf{x}_0) \geq 0 \Rightarrow f(\mathbf{x}) \geq f(\mathbf{x}_0)$$

Geometrically, $\mathbf{g}$ defines a supporting hyperplane[1] to the sublevel set

$$\{\mathbf{x} \,|\, f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$$

Consider the linear fractional function $f(\mathbf{x}) = \frac{\mathbf{a}^T\mathbf{x}+b}{\mathbf{c}^T\mathbf{x}+d}$. Let $\mathbf{c}^T\mathbf{x}_0 + d > 0$. Then compute the quasi-gradient $\mathbf{g}$ of $f(\mathbf{x})$ at $\mathbf{x}_0$. Justify/prove your answer.

Solution. $\mathbf{g} = \mathbf{a} - f(\mathbf{x}_0)\mathbf{c}$ is a quasigradient at $\mathbf{x}_0$. This is because, if $\mathbf{c}^T\mathbf{x} + d > 0$, we have $\mathbf{a}^T(\mathbf{x}-\mathbf{x}_0) \geq f(\mathbf{x}_0)\mathbf{c}^T(\mathbf{x}-\mathbf{x}_0) \Rightarrow \mathbf{a}^T\mathbf{x}+b \geq \mathbf{a}^T\mathbf{x}_0+b+\frac{\mathbf{a}^T\mathbf{x}_0+b}{\mathbf{c}^T\mathbf{x}_0+d}\mathbf{c}^T(\mathbf{x}-\mathbf{x}_0)+\frac{\mathbf{a}^T\mathbf{x}_0+b}{\mathbf{c}^T\mathbf{x}_0+d}(d-d) \Rightarrow$ $\mathbf{a}^T\mathbf{x} + b \geq \mathbf{a}^T\mathbf{x}_0 + b + \frac{\mathbf{a}^T\mathbf{x}_0+b}{\mathbf{c}^T\mathbf{x}_0+d}(\mathbf{c}^T\mathbf{x} - d) - \frac{\mathbf{a}^T\mathbf{x}_0+b}{\mathbf{c}^T\mathbf{x}_0+d}(\mathbf{c}^T\mathbf{x}_0 - d) \Rightarrow \mathbf{a}^T\mathbf{x} + b \geq f(\mathbf{x}_0)(\mathbf{c}^T\mathbf{x} - d) \Rightarrow$ $f(\mathbf{x}) \geq f(\mathbf{x}_0)$.

Question No 2 (5 Marks)

Given the general optimization program

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

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.

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

[1] More specifically, for your optional understanding, the set of quasigradients at $\mathbf{x}_0$ form a cone.

2. L is convex in x

3. L is concave in λ and μ

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

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

SOLUTION. Ans: 3. See page 14 of https://moodle.iitb.ac.in/pluginfile.php/377354/mod_resource/content/67/CS769_2023___Lecture20-annotated.pdf

Problem 3 (3 Marks)

Compute the projection step

$$P_{\mathcal{C}}(\mathbf{z}) = \underset{\mathbf{x}\in\mathcal{C}}{\text{argmin}} \frac{1}{2\gamma}||\mathbf{x} - \mathbf{z}||^2$$

where $\mathcal{C} = \{\mathbf{x} \in \Re^n \mid \mathbf{a}^T\mathbf{x} \leq c\}$ for some fixed $\mathbf{a} \in \Re^n$ and $c \in \Re$.

SOLUTION. See pages 7-9 of https://moodle.iitb.ac.in/pluginfile.php/377354/mod_resource/content/67/CS769_2023___Lecture20-annotated.pdf

Problem 4 (3 Marks)

Now instead consider the *prox* operator discussed in the class:

$$prox_c(\mathbf{z}) = \text{argmin}_{\mathbf{x}} \frac{1}{2\gamma}||\mathbf{x} - \mathbf{z}||^2 + c(\mathbf{x})$$

Let $\mathbf{x} \in \Re$ and $\mathbf{z} \in \Re$ be a fixed value. For the following choice of $c(x)$, what is $prox_c(\mathbf{z})$?

$$c(x) = \begin{matrix} -\lambda \log x & \text{if } x > 0 \\ \infty & \text{if } x \leq 0 \end{matrix}$$

It is ok even if you derive the answer for the special case of $\gamma = 1$.

SOLUTION. $prox_c(\mathbf{z}) = \dfrac{z + \sqrt{z^2 + 4\lambda}}{2}$. See page 11 of https://moodle.iitb.ac.in/pluginfile.php/363702/mod_resource/content/68/CS769_2023___Lecture19-annotated.pdf.

The answer can be derived by substituting for $c(x) = -\lambda \log x$ (since minimizer cannot be for $x \leq 0$ since that will lead to an unacceptable ∞ value for an objective value that needs to be minimized. Setting the gradient/derivative wrt x to 0 yields
$(x - z) - \frac{\lambda}{x} = 0$ which implies that at the solution x, we must have $x^2 - zx - \lambda = 0$ under the condition that $x > 0$. That means even if $x = \frac{-b\pm\sqrt{b^2-4ac}}{2a} = \frac{z\pm\sqrt{z^2+4\lambda}}{2a}$, we must restrict x to the positive possibility $x = \frac{z+\sqrt{z^2+4\lambda}}{2a}$

Problem 5 (6 Marks)

Let V be the ground set of discrete instances $1, 2, \ldots, i, \ldots$. Let s_{ij} be the similarity kernel between any two points $i, j \in V$. Is the following function $f(X)$ always submodular for any $X \subseteq V$? (4 marks)

$$\sum_{i \in V} \max_{k \in X} s_{ik}$$

SOLUTION. Ans: Yes. This is the facility location function and it is indeed submodular as well as monotone. See pages 12-13 of https://moodle.iitb.ac.in/pluginfile.php/363715/mod_resource/content/6/CS769-2023-Final-SubmodularOpt22-annotated.pdf

Is the following function $g(X)$ always submodular for any $X \subseteq V$? (2 marks)

$$\sum_{i \in V} \min_{k \in X} s_{ik}$$

SOLUTION. Ans: No. This function is not always submodular. See pages 21-23 of https://moodle.iitb.ac.in/pluginfile.php/363714/mod_resource/content/4/CS769-2023-Final-Submod pdf for a counter example on how the min of two look up functions is not submodular. You can extend that same proof to min of two elements from a similarity kernel s_{ik}

Problem 6 (2 Marks)

Match the most appropriate proof (sketch) technique from the RHS that was adopted to prove the submodularity of the function of the LHS. You just need to enumberate the answers in the format

LHS index → RHS index. Example: (e) → (v).

(a) Attractive potentials	(i) Concave over modular
(b) Discrete entropy	(ii) Strict concavity of log
(c) Complexity functions	(iii) Diminishing returns on volume
(d) Determinantal Point Processes	(iv) Second equivalent definition of submodularity in terms of intersection and union

SOLUTION. (a) - (iv) Page 7 of https://moodle.iitb.ac.in/pluginfile.php/363715/mod_resource/content/6/CS769-2023-Final-SubmodularOpt22-annotated.pdf
(b) - (ii) Page 6 of Page23ofhttps://moodle.iitb.ac.in/pluginfile.php/379216/mod_resource/content/6/CS769-2023-Final-SubmodularOpt23-annotated.pdf
(c) - (i) Page 7 of https://moodle.iitb.ac.in/pluginfile.php/363715/mod_resource/content/6/CS769-2023-Final-SubmodularOpt22-annotated.pdf
(d) - (iii) Page 10 of https://moodle.iitb.ac.in/pluginfile.php/379216/mod_resource/content/6/CS769-2023-Final-SubmodularOpt23-annotated.pdf

Problem 7 (4 Marks)

Recall the unification of several adaptive and non-adaptive first order update variants into a single update equation:

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

where $H_T = \text{diag}\left(\left[\sum_{t=1}^{T} \eta_t g_t g_t^T\right]^{1/2}\right) \implies \left[H_t^{-1}\right]_{jj} = \frac{1}{\sqrt{\sum_{t=1}^{T} \eta_t g_{tj}^2}}$. and G_T is such that $H_t = \sqrt{G_t}$ (elementwise)

Consider the table below that was discussed in the class.

	SGD	HB	NAG	AdaGrad	RMSProp	Adam
G_t	I	I	I	$G_{t-1} + D_t$	$\beta_2 G_{t-1} + (1 - \beta_2) D_t$	$\frac{\beta_2}{1-\beta_2^t} G_{t-1} + \frac{1-\beta_2}{1-\beta_2^t} D_t$
α_t	α	α	α	α	α	$\alpha \frac{1-\beta_1}{1-\beta_1^t}$
β_t	0	β	β	0	0	$\frac{\beta_1(1-\beta_1^{t-1})}{1-\beta_1^t}$
γ	0	0	β	0	0	0

1. Substitute the values of G_t, α_t, β_t and γ and write each update type in the most specific form that is possible. (2 marks)

2. Present an important characteristic of each update type in simple English. For example, you could comment on the algorithms in the table along the lines of diminishing learning rate, adaptive, momentum type, *etc.* (2 marks)

SOLUTION. Pages 32-33 of https://moodle.iitb.ac.in/pluginfile.php/375856/mod_resource/content/70/CS769_2023___Lecture16-annotated.pdf and 3-12 of https://moodle.iitb.ac.in/pluginfile.php/376393/mod_resource/content/68/CS769_2023___Lecture17-annotated.pdf. Eg: Benefits of AdaGrad: AdaGrad can significantly improve upon SGD in sparse feature sets!

It automatically sets the learning rate, and secondly, automatically updates the learning rates with a decay schedule! Also, it has a per coordinate learning rate! Adagrad is starting point of numerous new techniques for adaptive methods. RMSProp fixes the significantly diminishing learning rate created by Adagrad. Adam is basically HB Momentum + Adaptive (RMSProp).

Problem 8 (4 Marks)

Which of the following statements is/are TRUE concerning the theoretical analysis concerning the number of iterations required for attaining an ϵ-approximate solution? Justify each choice, briefly.

1. 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

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

3. 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.

4. 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.

Solution. Note that this problem is adapted from last year. So anyone who took the effort to go over the last year problems and solutions should have got this.
(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 $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ 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 $\frac{R^2L}{\epsilon}$ 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.

Problem 9: (5 Marks)

Consider a quasi-Netwon algorithm for minimizing a function $f(\mathbf{x})$ without any constraints. In one such quasi-Newton algorithm, the $B^{(k+1)}$ (approximation to the Hessian) is obtained from a positive-definite matrix $B^{(k)}$ from the previous iteration on k, by using the Davidon-Fletcher-Powell (DFP) updating formula, which is specified below:

$$B^{(k+1)} = B^{(k)} + \frac{\Delta\mathbf{x}^{(k)}\left(\Delta\mathbf{x}^{(k)}\right)^T}{\left(\Delta\mathbf{x}^{(k)}\right)^T\Delta\mathbf{g}^{(k)}} - \frac{B^{(k)}\Delta\mathbf{g}^{(k)}\left(\Delta\mathbf{g}^{(k)}\right)^T B^{(k)}}{\left(\Delta\mathbf{g}^{(k)}\right)^T B^{(k)}\Delta\mathbf{g}^{(k)}}$$

The values $\Delta\mathbf{x}^{(k)}$, $B^{(k)}$ and $\Delta\mathbf{g}^{(k)}$ are iteratively obtained in the following manner:

1. $\Delta\mathbf{x}^{(k)} = -B^{(k)}\nabla f\left(\mathbf{x}^{(k)}\right)$

2. $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t^{(k)}\Delta\mathbf{x}^{(k)}$, where $t^{(k)}$ can be obtained using any method such as line search, *etc.*

3. $\Delta\mathbf{g}^{(k)} = \nabla f\left(\mathbf{x}^{(k+1)}\right) - \nabla f\left(\mathbf{x}^{(k)}\right)$

Show that the condition

$$\left(\Delta\mathbf{x}^{(k)}\right)^T\Delta\mathbf{g}^{(k)} > 0$$

will ensure that $B^{(k+1)}$ remains positive definite.

Solution. See next page

Ans:

$$x^T B^{(k+1)} x = x^T B^{(k)} x - \frac{x^T B^{(k)} \Delta g^{(k)} (\Delta g^{(k)})^T B^{(k)} x}{(\Delta g^{(k)})^T B^{(k)} \Delta g^{(k)}} + \frac{x^T \Delta x^{(k)} (\Delta x^{(k)})^T x}{(\Delta x^{(k)})^T \Delta g^{(k)}}$$

$$= \frac{\|u\|^2 \|v\|^2 - (u^T v)^2}{\|v\|^2} + \frac{(x^T \Delta x^{(k)})^2}{(\Delta g^{(k)})^T (\Delta x^{(k)})} \quad ①$$

where $u = (B^k)^{1/2} x$ & $v = (B^k)^{1/2} \Delta g^{(k)}$

Using the Cauchy Schwartz inequality . . . .

$$(u^T v)^2 \le (\|u\| \|v\|)^2$$

Equality holds iff $u = \theta v$ for some $\theta \neq 0$

Case ⓐ $u = \theta v \Rightarrow [u^T v]^2 = \|u\|^2 \|v\|^2$

$\Rightarrow$ ① becomes

$$x^T B^{(k+1)} x = \frac{(x^T \Delta x^{(k)})^2}{(\Delta x^{k})^T \Delta g^{(k)}}$$

Since $u = \theta v$, $(B^{(k)})^{1/2} x = \theta (B^k)^{1/2} \Delta g^{(k)}$

$\Rightarrow x = \theta \Delta g^{(k)}$ [$\det(B) = \det(B^{1/2} B^{1/2}) \neq 0$ $\Rightarrow \det(B^{1/2}) \neq 0 \Rightarrow B^{1/2}$ has independent rows/columns

$$\therefore\ x^T B^{(k+1)} x = \theta^2 \frac{\left[(\Delta g^{(k)})^T \Delta x^{(k)}\right]^2}{\left[(\Delta x^{(k)})^T \Delta g^{(k)}\right]}$$

$$= \theta^2 \underbrace{(\Delta g^{(k)})^T \Delta x^{(k)}}_{\text{given to be } > 0} > 0$$

Case (b) If $u \neq \theta v$

$$\underbrace{\frac{\|u\|^2 \|v\|^2 - [u^T v]^2}{\|v\|^2}}_{\text{First term in (1)}} > 0$$

Since $(\Delta g^{(k)})^T \Delta x^{(k)} > 0$

$$\underbrace{\frac{(x^T \Delta x^{(k)})^2}{(\Delta g^{(k)})^T \Delta x^{(k)}}}_{\text{Second term in (1)}} \geq 0$$

Together imply

$x^T B^{k+1} x > 0$

$\Downarrow$

$B^{(k+1)}$ is positive definite