(use notations and conventions from the class) Consider the problem of linear regression where we minimize the loss

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N} \alpha_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda_1 g(\mathbf{w})$$

where $g()$ is a regularization term. We also write the loss in matrix form as

$$\mathcal{L}_2 = \frac{1}{N}[Y - \mathbf{X}\mathbf{w}]^T A [Y - \mathbf{X}\mathbf{w}] + \lambda_2 g(\mathbf{w}).$$

If $\mathcal{L}_1 = \mathcal{L}_2$ for all $\mathbf{w}$, then

(A) **[Ans]** A is a diagonal matrix

(B) $A_{ij} = \alpha_i \cdot \alpha_j$

(C) **[Ans]** $A_{ii} = \alpha_i$ else zero

(D) $A_{ii} = \frac{1}{\alpha_i}$ else zero

(E) none of the above

(use notations and conventions from the class) Consider the problem of linear regression where we minimize the loss

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N} \alpha_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda_1 g(\mathbf{w})$$

where $g()$ is a regularization term. We also write the loss in matrix form as

$$\mathcal{L}_2 = \frac{1}{N}[Y - \mathbf{X}\mathbf{w}]^T A [Y - \mathbf{X}\mathbf{w}] + \lambda_2 g(\mathbf{w}).$$

If $\mathbf{A} = I$, $\alpha_i = 1$ for all i, and $\lambda_1 = \lambda_2 = 1$, then

(A) **[Ans]** Both the loss functions are identical i.e., $\mathcal{L}_1 = \mathcal{L}_2$

(B) **[Ans]** The optima of the first objective $\mathbf{w}_1^*$ is same as the optima of $\mathcal{L}_2$, i.e., $\mathbf{w}_2^*$

(C) **[Ans]** At the optima, value of the losses are same. i.e., $\mathcal{L}_1^* = \mathcal{L}_2^*$

(D) $\mathcal{L}_1$ is a scalar and $\mathcal{L}_2$ is a vector

(E) none of the above

(use notations and conventions from the class) Consider the problem of linear regression where we minimize the loss

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N} \alpha_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda_1 g(\mathbf{w})$$

where $g()$ is a regularization term. We also write the loss in matrix form as

$$\mathcal{L}_2 = \frac{1}{N}[Y - \mathbf{X}\mathbf{w}]^T A [Y - \mathbf{X}\mathbf{w}] + \lambda_2 g(\mathbf{w}).$$

If $\mathbf{A} = I$, $\alpha_i = 2$ for all i, and $\lambda_1 = \lambda_2 = 0$, then

(A) Both the loss functions are identical i.e., $\mathcal{L}_1 = \mathcal{L}_2$

(B) **[Ans]** The optima of the first objective $\mathbf{w}_1^*$ is same as the optima of $\mathcal{L}_2$, i.e., $\mathbf{w}_2^*$

(C) At the optima, value of the losses are same. $\mathcal{L}_1^* = \mathcal{L}_2^*$

(D) $\mathcal{L}_1$ is a scalar and $\mathcal{L}_2$ is a vector

(E) none of the above

(use notations and conventions from the class) Consider the problem of linear regression where we minimize the loss

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N} \alpha_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda_1 g(\mathbf{w})$$

where $g()$ is a regularization term. We also write the loss in matrix form as

$$\mathcal{L}_2 = \frac{1}{N}[Y - \mathbf{X}\mathbf{w}]^T A [Y - \mathbf{X}\mathbf{w}] + \lambda_2 g(\mathbf{w}).$$

If $\mathbf{A} = I$, $\alpha_i = 1$ for all i, and $\lambda_1 \neq \lambda_2 \neq 0$, then

(A) The optimal parameters $\mathbf{w}^*$ is independent of λ_i.

(B) The larger the lambda, the better the solution.

(C) The smaller the lambda, the better the solution

(D) When lambda is nonzero (positive), loss will increase (since $g(w)$ is also positive in practice), better to use $\lambda = 0$.

(E) **[Ans]** None of the above.

(use notations and conventions from the class) Consider the problem of linear regression where we minimize the loss

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N} \alpha_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda_1 g(\mathbf{w})$$

where $g()$ is a regularization term. We also write the loss in matrix form as

$$\mathcal{L}_2 = \frac{1}{N}[Y - \mathbf{X}\mathbf{w}]^T A [Y - \mathbf{X}\mathbf{w}] + \lambda_2 g(\mathbf{w}).$$

See $\mathcal{L}_2$ closely,

(A) **[Ans]** When A is a diagonal matrix, this is equivalent to weighing each sample independently.

(B) When A is not a diagonal matrix, this loss does not make any sense. Don't use.

(C) **[Ans]** When A is PD, we can do cholesky decomposition of A as LL^T and an equivalent formulation is possible in $\mathcal{L}_1$ is each sample getting transformed as $\mathbf{L}^T \mathbf{x}_i$ (as in LMNN/Metric Learning)

(D) **[Ans]** When A is a rank deficient matrix, an equivalent formulation is possible in $\mathcal{L}_1$ with a dimensionality reduction (this could be proved with eigen decomposition).

(E) None of the above