In-Class Review

True or false:

FALSE! (A) A linear combination is a special case of affine combination

TRUE! (B) A convex combination is a special case of linear combination

TRUE! (C) A linear combination of vectors has the same number of dimensions as the vectors

TRUE! (D) A vector in a linearly independent set of vectors cannot be written as a linear combination of other vectors

In-Class Review

True or false:

TRUE! (A) Any vector in a subspace can be written as a linear combination of the basis vectors for that subspace

TRUE! (B) All basis vectors must be linearly independent of one another

TRUE! (C) Any linearly independent set of vectors must be a basis

TRUE! (D) It is possible for a basis to have non-orthogonal vectors

TRUE! (E) Any orthonormal basis must be an orthogonal basis

In-Class Review

True or false:

FALSE! (A) The outer product of two vectors is the transpose of their inner product

TRUE! (B) The transpose of the transpose of a matrix is itself

TRUE! (C) $(AB)^\top$ is not necessarily the same as $A^\top B^\top$

TRUE! (D) The outer product of a vector with itself must be a square matrix

TRUE! (E) The inner product of two vectors of different dimensions cannot be computed

In-Class Review

True or false:

TRUE! (A) Left-singular vectors are not necessarily the same as right-singular vectors

TRUE! (B) Right-singular vectors could be the same as eigenvectors

TRUE! (C) Singular values of a matrix always exist

TRUE! (D) Singular values are always non-negative

TRUE! (E) Left-singular vectors could be the same as right-singular vectors

In-Class Review

Q1: Which of the following is **always** a valid mathematical operation? (An operation that results in complex numbers is not considered valid.)

No (A) Computing A^{-1}, where A is full-rank but non-square

Yes (B) Computing $\sqrt{\vec{x}^\top A^\top A\vec{x}}$, where A is non-square

No (C) Computing $\sqrt{\vec{x}^\top A\vec{x}}$, where A is symmetric

No (D) Computing the eigendecomposition of A, where A is non-square

No (E) Computing $\frac{A}{B}$,where A and B are positive definite

In-Class Review

Do each of the following characterize a convex function?

No a) $\vec{x}^{\top}\left(\frac{\partial^2 f}{\partial \vec{x}\partial \vec{x}^{\top}}(\vec{x})\right)\vec{x} \geq 0 \; \forall \vec{x}$

Yes b) $\vec{x}^{\top}\left(\frac{\partial^2 f}{\partial \vec{x}\partial \vec{x}^{\top}}(\vec{y})\right)\vec{x} \geq 0 \; \forall \vec{x}, \vec{y}$

Yes c) $\frac{\partial^2 f}{\partial \vec{x}\partial \vec{x}^{\top}}(\vec{x}) \succcurlyeq 0 \; \forall \vec{x}$

Yes d) All eigenvalues of $\frac{\partial^2 f}{\partial \vec{x}\partial \vec{x}^{\top}}$ are non-negative $\forall \vec{x}$

Yes e) Let $U\Lambda U^{\top} = \frac{\partial^2 f}{\partial \vec{x}\partial \vec{x}^{\top}}$ be the eigendecomposition. The diagonal entries of Λ are all non-negative $\forall \vec{x}$

In-Class Review

Q1: Which of the following facts about ridge regression is true?

True (A) Ridge regression is less prone to overfitting compared to ordinary least squares

True (B) Ridge regression always has a unique optimal parameter vector

True (C) Compared to ordinary least squares, ridge regression adds a regularizer

True (D) Ridge regression uses more hyperparameters than ordinary least squares

False (E) Ridge regression uses more parameters than ordinary least squares

True (F) Ridge regression uses a strictly convex loss function

In-Class Review

Which one of the following **must be** a Gaussian random variable?

Yes (A) $X|Y=1$, where $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}(\vec{0}, I)$

Yes (B) $Y|X=100$, where $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}(\vec{0}, I)$

Yes (C) $\frac{1}{2}X - \frac{1}{3}Y$, where $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}(\vec{0}, I)$

Yes (D) $-10X + 5$, where $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}(\vec{0}, I)$

Yes (E) $Y - 10X$, where $X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0,1)$ and $X \perp Y$

Yes (F) $X - Y$, where $X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0,1)$ and $\mathrm{Cov}(X,Y) = -1$

No (G) $X - Y$, where $X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0,1)$ and $\mathrm{Cov}(X,Y) = -0.5$

Yes (H) $-10X + 5$, where $X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0,1)$ and $\mathrm{Cov}(X,Y) = -0.5$

In-Class Review

Which of the following statements about the difference between MLE and MAP parameter estimates is true?

True (A) MLE does not depend on the prior, whereas MAP does

False (B) MLE depends on the likelihood function, whereas MAP does not

True (C) The derivation of MLE does not use Bayes' rule, whereas the derivation of MAP does

True (D) Ridge regression can be interpreted as MAP, but not as MLE

True (E) MAP corresponds to optimizing a loss function with a regularizer on the parameters, whereas MLE corresponds to optimizing a loss function without a regularizer

In-Class Review

Which of the following statements about gradient descent is true?

True (A) Gradient descent may not find the global minimum on non-convex functions

True (B) Gradient descent may diverge on non-Lipschitz functions

True (C) With a sufficiently small step size, gradient descent always converges at a rate of $\Theta(1/\epsilon^2)$ to the minimal objective value on convex Lipschitz functions

False (D) With a sufficiently small step size, gradient descent always converges at a rate of $\Theta(1/\epsilon^2)$ to the minimal objective value on functions whose every local minimum is a global minimum

False (E) With a sufficiently small step size, gradient descent always converges at a rate of $\Theta(1/\epsilon^2)$ to the optimal parameters when the objective function is both convex and Lipschitz

In-Class Review

Q1: Which of the following optimization algorithms uses a non-diagonal preconditioner?

No (A) Gradient descent

No (B) Gradient descent with momentum

No (C) Nesterov's accelerated gradient

No (D) AdaGrad

No (E) Adam

Yes (F) Newton's method

In-Class Review

Which of the following would NOT be an effective way around optimization difficulties caused by exploding gradients?

Effective (A) Perform gradient clipping

Effective (B) Make the network shallower

Not Effective (C) Use a non-saturating activation function like ReLU

Effective (D) Add weight decay

Effective (E) Add layer normalization

Effective (F) Initialize the weights using the He initializer

In-Class Review

Which of the following statements about support vector machines are true?

True (A) An SVM is a binary classifier

True (B) An SVM is a linear classifier

True (C) The decision boundary of an SVM is a hyperplane

False (D) The decision boundary of an SVM may not pass through the origin

True (E) The width of the margin is not known a priori

True (F) An SVM tries to maximize the margin

In-Class Review

Which of the following statements about SVMs are true?

True (A) Maximizing $\frac{1}{\|\vec{w}\|_2}$ is equivalent to minimizing $\|\vec{w}\|_2$

True (B) Minimizing $\|\vec{w}\|_2$ is equivalent to minimizing $\|\vec{w}\|_2^2$

True (C) $\vec{w}$ is perpendicular to the decision boundary

True (D) The decision boundary must pass through the origin when $b = 0$

False (E) Maximizing the margin is equivalent to maximizing $\|\vec{w}\|_2^2$

In-Class Review

Consider a linearly inseparable dataset. Which of the following statements is true?

True (A) Hard-margin SVMs with linear kernels never have a feasible solution.

True (B) Hard-margin SVMs with quadratic kernels sometimes have a feasible solution.

False (C) Hard-margin SVMs with quadratic kernels always have a feasible solution.

True (D) Soft-margin SVMs with linear kernels always have a feasible solution.

True (E) Soft-margin SVMs with quadratic kernels sometimes have a feasible solution.

True (F) Soft-margin SVMs with quadratic kernels always have a feasible solution.