CSCI 567: Problem Solving

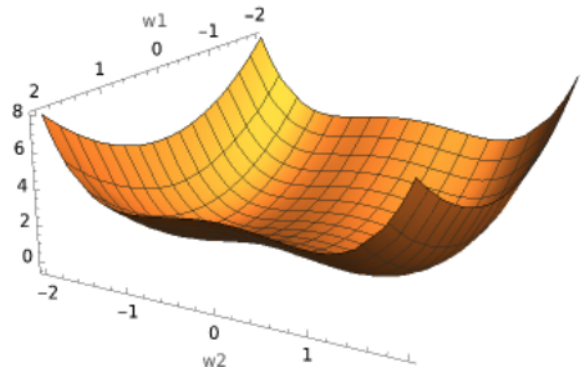
Feb 6, 2026

I. Optimization

II. Calculus & Linear Algebra

III. Probability

Consider a two-dimensional function $F(\boldsymbol{w}) = w_1^2 - w_2^2 + \frac{1}{2}w_2^4$ (a plot is provided below). Which of the following statements are correct?

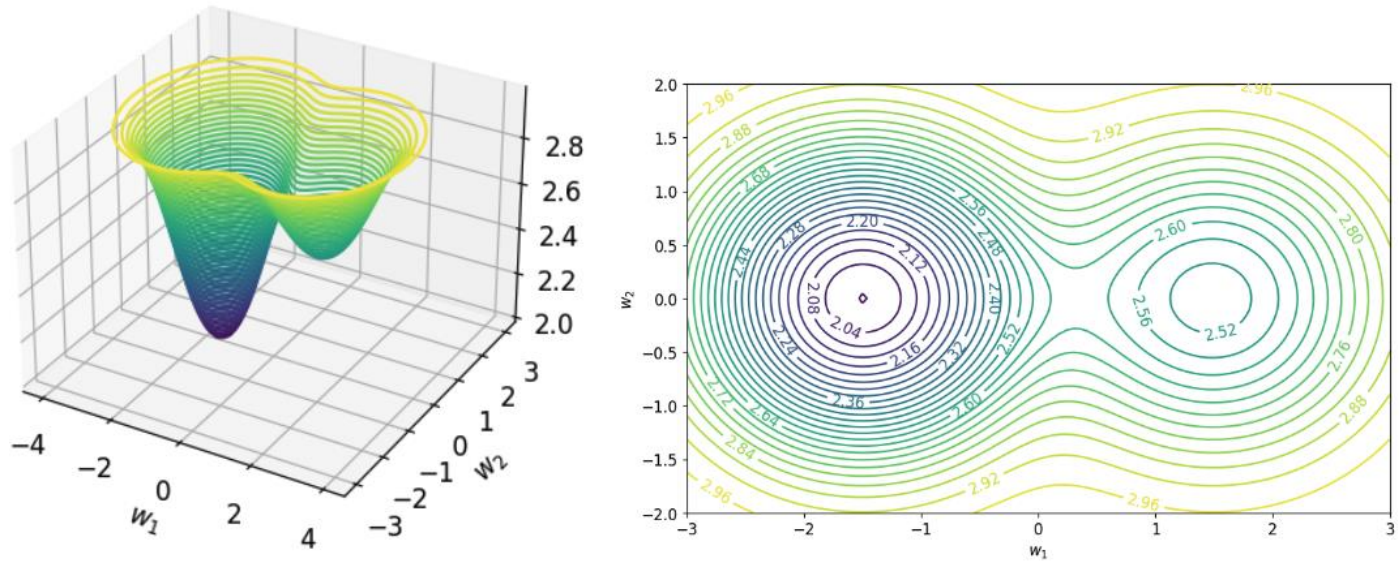


(A) F is convex.
(B) F has three stationary points: $(0, 0)$, $(0, -1)$, and $(0, 1)$.
(C) $(0, 0)$ is a saddle point of F.
(D) Gradient Descent always converges to $(0, -1)$, regardless of initialization.

A machine learning objective function is usually of the form $F(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{w})$ for some loss function $f_i : \mathbb{R}^d \to \mathbb{R}$ corresponding to the i-th training point. We know that $\nabla f_i(\boldsymbol{w})$ for i chosen uniformly at random is a stochastic gradient of this objective since $\mathbb{E}[\nabla f_i(\boldsymbol{w})] = \nabla F(\boldsymbol{w})$‘. Which of the following is also a stochastic gradient?

(A) $\sum_{i\in S} \nabla f_i(\boldsymbol{w})$ for a subset $S \subset \{1, \ldots, n\}$ (of some fixed size) chosen uniformly at random.
(B) $\frac{1}{|S|}\sum_{i\in S} \nabla f_i(\boldsymbol{w})$ for a subset $S \subset \{1, \ldots, n\}$ (of some fixed size) chosen uniformly at random.
(C) $\nabla f_i(\boldsymbol{w}) - \nabla f_i(\boldsymbol{w}_0) + \nabla F(\boldsymbol{w}_0)$ for i chosen uniformly at random and some fixed point $\boldsymbol{w}_0$.
(D) $\frac{\partial F(\boldsymbol{w})}{\partial w_i} \cdot d\boldsymbol{e}_i$ for a coordinate i chosen uniformly at random ($\boldsymbol{e}_i \in \mathbb{R}^d$ is the standard basis vector with 1 in the i-th coordinate and 0 in all other coordinates).

Consider the function depicted in Figs. 2a and 2b. On the left we have a 3-d plot of the function, on the right we have a contour plot of the same function

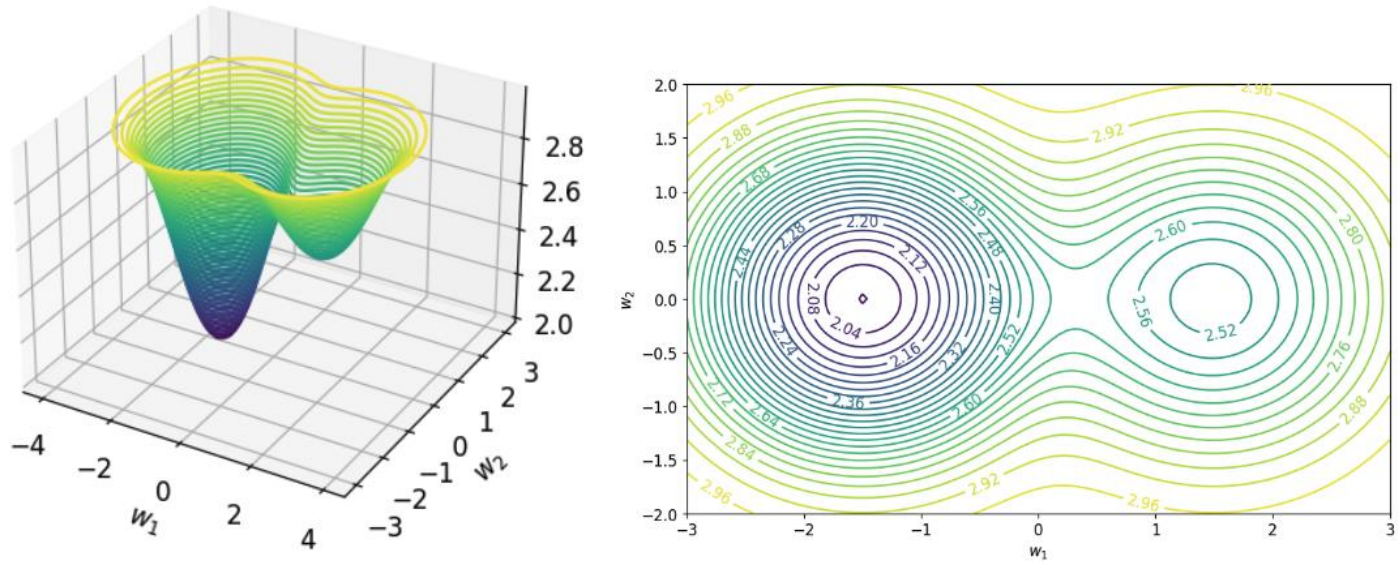


(a) (b)

Figure 2: A 3-d plot and contour plot of the same function.

(a) Is the function convex? Explain. How many local minima does the function have?(2 points)

Consider the function depicted in Figs. 2a and 2b. On the left we have a 3-d plot of the function, on the right we have a contour plot of the same function



(a) (b)

Figure 2: A 3-d plot and contour plot of the same function.

(b) Suppose we use gradient descent to minimize the function. Will gradient descent always find the global minimizer of the function? Explain. (2 points)

(c) Fig 3 shows the iterates of some run of gradient descent. Comment on the behavior of gradient descent seen in this plot, and suggest how you can improve the convergence. (2 points)

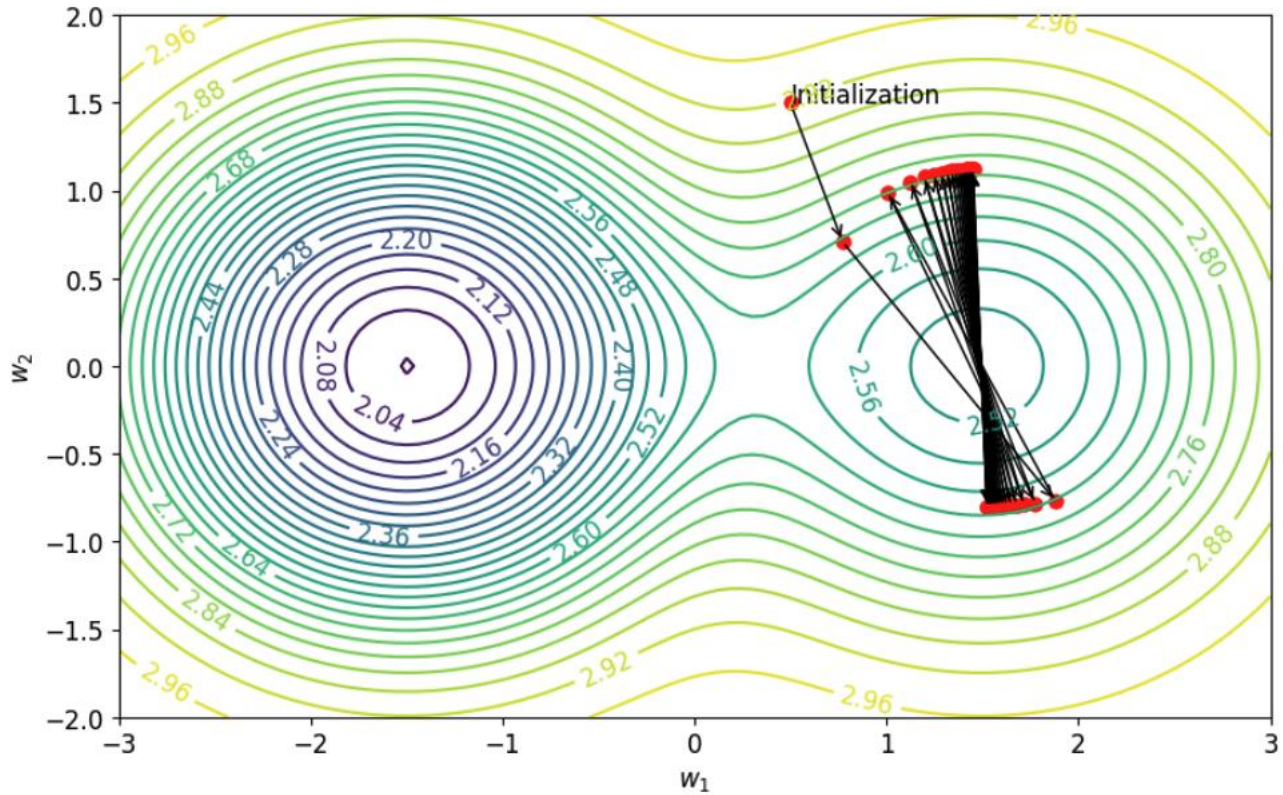


Figure 3: Gradient descent iterates denoted from some initialization, using red dots.

I. Optimization

II. Calculus & Linear Algebra

III. Probability

1 Weighted linear regression (10 points)

Consider a modification of the standard linear regression setup where each datapoint is associated with an importance weight. Formally, given a dataset of n datapoints $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, each datapoint is associated with an importance weight $r_i > 0$. The weighted residual sum of squares objective is defined as,

$$\text{WRSS}(\boldsymbol{w}) = \sum_{i=1}^{n} r_i \left(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i - y_i\right)^2 .$$

(a) Let $\boldsymbol{X}$ be the $n \times d$ matrix whose i-th row is $\boldsymbol{x}_i^{\mathrm{T}}$, $\boldsymbol{y}$ be the n-dimensional column vector whose i-th entry is y_i and $\boldsymbol{R}$ be the diagonal matrix where $\boldsymbol{R}_{ii} = r_i$ for all i and 0 for all other entries. Show that the WRSS objective can be written as follows in matrix form, (3 points)

$$\text{WRSS}(\boldsymbol{w}) = (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^{\mathrm{T}} \boldsymbol{R} (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}). \tag{1}$$

$$\text{WRSS}(\boldsymbol{w}) = (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^{\text{T}} \boldsymbol{R} (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}). \tag{1}$$

(b) Solve for the closed-form solution $\boldsymbol{w}^*$ which minimizes Eq 1 (assuming invertibility of any matrices as needed). (7 points)

Hint: You might find it helpful to rewrite WRSS($\boldsymbol{w}$) in the form WRSS($\boldsymbol{w}$) = $\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w} - 2\boldsymbol{b}^T\boldsymbol{w} + \boldsymbol{y}^T\boldsymbol{R}\boldsymbol{y}$, for some matrix $\boldsymbol{A}$ and some vector $\boldsymbol{b}$ (which you would need to find)

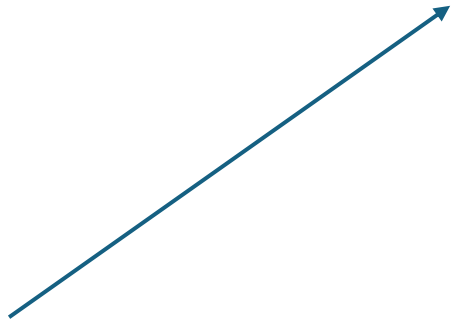
A function $f : \mathbb{R}^{n \times n} \to \mathbb{R}$ is defined as $f(\mathbf{A}) = \mathbf{x}^\top \mathbf{A} \mathbf{x}$ for some $\mathbf{x} \in \mathbb{R}^{n \times 1}$. What is the derivative $\dfrac{\partial f}{\partial \mathbf{A}}$?

A symmetric matrix A is **positive semidefinite** (PSD) if

$$x^\top Ax \geq 0 \quad \forall x$$

Intuition: Ax never "pushes against" the vector x.

Equivalent characterization: Symmetric A is PSD if and only if all eigenvalues are ≥ 0.



Q4 Suppose A is a PSD matrix and M is any (not necessarily square) matrix of compatible dimensions. Prove that $M^\top AM$ is PSD.

Q5 Let A be a symmetric matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$.

(a) What are the eigenvalues of $A + \mu I$ for $\mu \in \mathbb{R}$?

(b) For what values of μ is $A + \mu I$ positive semidefinite?

I. Optimization
II. Calculus & Linear Algebra
III. Probability

Q Suppose your spam classification software gives the guarantee that (1) if an email is spam, it will mark it as spam with probability 90%, (1) if an email is not spam, it will only mark it as spam with probability 10%. Suppose you know that 1% of all your emails are spam. If your spam classification software marks a certain email as spam, what is the probability that it is actually spam?

A **fair coin** is flipped **100 times**.

Let X be the **number of heads** obtained.

1. Compute $\mathbb{E}[X]$.
2. Compute $\mathrm{Var}(X)$.
3. Compute the standard deviation of X.

Take the numbers

$$1, 2, 3, \ldots, n$$

and **shuffle them uniformly at random**, writing them in a row.

We say that position k is a **peak** if:

- $k = 1$ and the first number is larger than the second, or
- $1 < k < n$ and the number in position k is larger than **both** its immediate neighbors, or
- $k = n$ and the last number is larger than the one before it.

Let X be the **total number of peaks** in the row.

Question: What is the *average value* of X?