
Linear Regression, Logistic Classification, and the Definition of Convexity

ECE 57000 — October 2, 2026

David I. Inouye

| Wednesday solved **the linear regression fitting problem**

- Ames sale prices gave us interpretable inputs and a numerical outcome.
- Squared residuals led to closed-form OLS through the normal equations.
- Residual orthogonality showed that no coefficient change can reduce training error.

Today: what does that solution mean, and what changes for classification?

| Orthogonality proves that **no other fit has lower error**

- $\mathbf{r} = [r_1, \dots, r_n]^\top \in \mathbb{R}^n$: **one residual per observation**, $r_i = y_i - \hat{y}_i$.
- $\mathbf{r}$ has n **coordinates for observations**, not d coordinates for features like $\mathbf{x}_i$.
- **Recall OLS orthogonality**: $\widetilde{X}^\top \mathbf{r} = \mathbf{0}$; $\mathbf{r}$ is perpendicular to each n -**dimensional feature column**.

Any other coefficient vector is $\hat{\boldsymbol{\theta}} + \boldsymbol{\delta}$; its prediction change $\widetilde{X}\boldsymbol{\delta} \in \mathbb{R}^n$.

$$\begin{aligned}\|\mathbf{y} - \widetilde{X}(\hat{\boldsymbol{\theta}} + \boldsymbol{\delta})\|^2 &= \|\mathbf{r} - \widetilde{X}\boldsymbol{\delta}\|^2 \\ &= \|\mathbf{r}\|^2 + \|\widetilde{X}\boldsymbol{\delta}\|^2 \\ &\geq \|\mathbf{r}\|^2\end{aligned}$$

(Write the new residual)

$$(\mathbf{r}^\top (\widetilde{X}\boldsymbol{\delta}) = 0)$$

(A squared norm is nonnegative)

The bound is attained: choose $\boldsymbol{\delta} = \mathbf{0}$, so $\|\widetilde{X}\boldsymbol{\delta}\|^2 = 0$.

Thus $\hat{\boldsymbol{\theta}}$ achieves the smallest possible squared error, $\|\mathbf{r}\|^2$: it is a **global minimum**.

| Does zero training residual alignment imply perfect prediction?

The OLS solution satisfies $\widetilde{X}^T r = 0$.

Which conclusion follows?

- A. Every residual is zero.
- B. No linear coefficient adjustment reduces the training squared error.
- C. The model will have zero error on future houses.

B. A nonzero residual can be perpendicular to the entire prediction subspace.

Redundant features can give the same fit with different coefficients

Example: predict a venue's booking price from its room capacities (zero offset).

Feature	Seats	Model A coefficient	Model B coefficient
Room 1: x_1	20	1	0
Room 2: x_2	30	2	1
Total: $x_3 = x_1 + x_2$	50	0	1

$$\hat{y}_A = 1(20) + 2(30) + 0(50) = 80, \quad \hat{y}_B = 0(20) + 1(30) + 1(50) = 80.$$

Same for every venue: $x_2 + x_3 = x_1 + 2x_2$ whenever $x_3 = x_1 + x_2$.

The data cannot distinguish these coefficient choices.

The inverse formula requires **independent columns**; redundant columns make $\widetilde{X}^T \widetilde{X}$ singular.

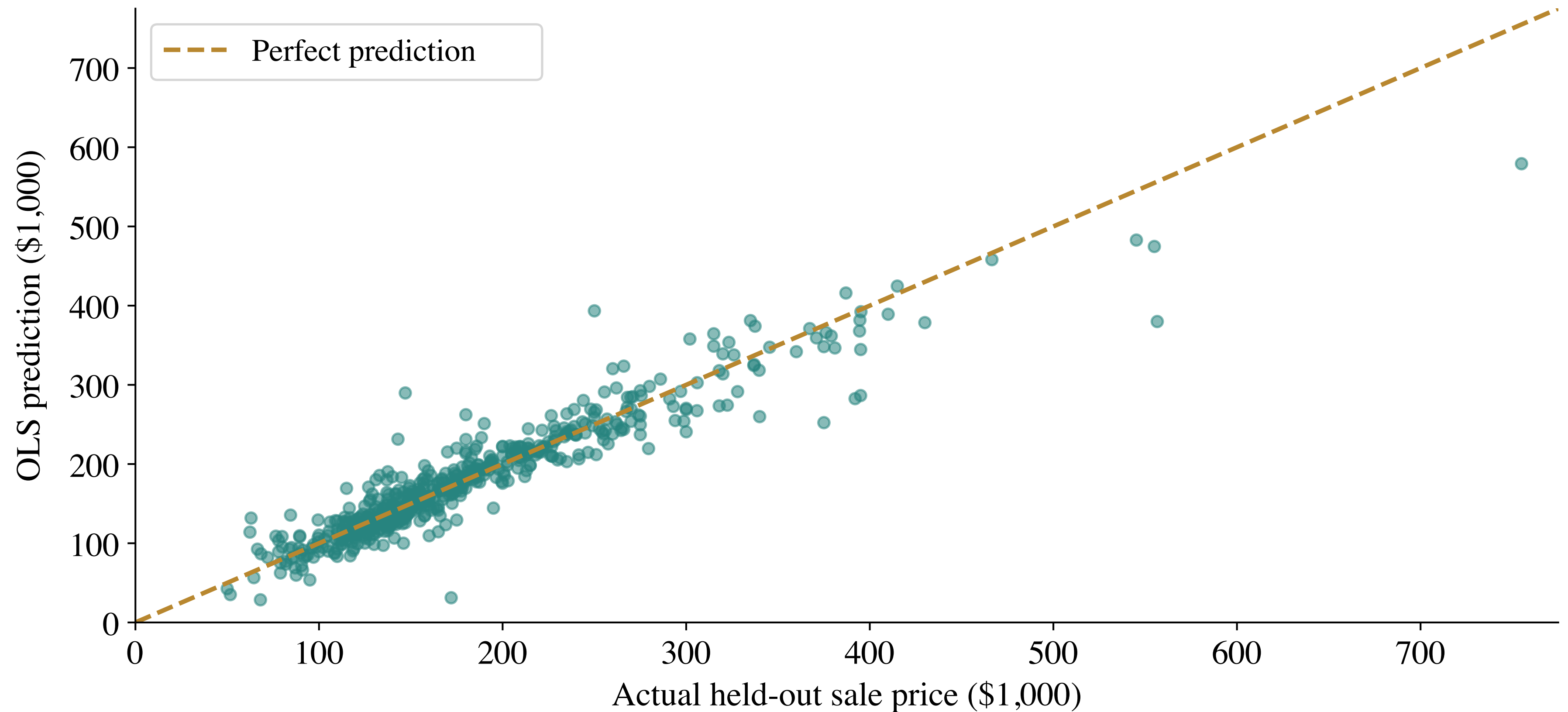
| On real Ames sales, **ordinary least squares is competitive**

Same split: **1,758 train / 586 validation / 586 test**; 226 encoded coordinates.

Model	Test RMSE ↓
Predict the training mean	\$78,500
Validation-selected KNN	\$32,800
Ordinary least squares	\$26,300
Random forest	\$25,700
Boosted trees	\$25,200

A compact linear rule is useful here even alongside flexible nonlinear methods.

| Held-out predictions reveal both useful structure and remaining errors



Each dot is one of **586 test sales**. The diagonal means prediction equals outcome.

| Solving the training problem leaves **the prediction claim to evaluate**

We can now explain: the function, the squared-error objective, the exact solution, and its geometry.

An exact fit to the objective does not make the model exact about the world.

Next: keep the linear score, but change from house prices to class probabilities.

| From a solved fit to a fitting algorithm

Stage	Question
1. Interpret the OLS solution	What does the exact fit establish?
2. Choose a classification model and loss	How should a score predict and fit labels?
3. Understand the objective's shape	When is a local optimum globally best?
4. Compute the coefficients	How can local information guide updates?

| Classification changes **what the score must predict**

Regression: a numerical price, $\hat{y} = w^T x + b$.

Classification: a positive/negative review, with an estimated class probability.

Return to our real Cornell case: **18,746 word features**, 1,200 training reviews.

| Logistic regression follows **model** → **objective** → **solution**

1. **Model — now:** turn the score into a probability.
2. **Objective — next:** measure how well those probabilities predict the labels.
3. **Solution — afterward:** learn the coefficients by numerical optimization.

| A learned score combines **signed contributions**

For one review, let $\boldsymbol{x} \in \mathbb{R}^d$ contain its word features.

$$z = \boldsymbol{w}^\top \boldsymbol{x} + b = \sum_{j=1}^d w_j x_j + b.$$

- x_j : the review's measured feature; **given at prediction time**.
- w_j : how that feature contributes; **learned from labeled reviews**.
- b : a learned offset; z : the resulting scalar score.

With $d = 18,746$, the model learns **18,747 parameters**, not one parameter per review.

| Each contribution is a **weight times a measurement**

Illustrative two-feature review, with **invented weights**:

Feature	Measurement x_j	Weight w_j	Contribution $w_j x_j$
“engaging”	2	+0.8	+1.6
“dull”	1	−1.1	−1.1
Offset	—	$b = -0.2$	−0.2

$$z = 1.6 - 1.1 - 0.2 = 0.3 \quad \implies \quad \hat{y} = 1.$$

The model combines evidence; a word’s contribution depends on **both its value and its weight**.

| A zero threshold divides space with a **hyperplane**

$$\hat{y} = \mathbf{1}\{w^\top x + b \geq 0\}, \quad \text{boundary: } w^\top x + b = 0.$$

- In two dimensions, the boundary is a **line**; in three, a **plane**.
- w determines its orientation; b shifts its position.
- Moving in a direction orthogonal to w leaves the score unchanged.

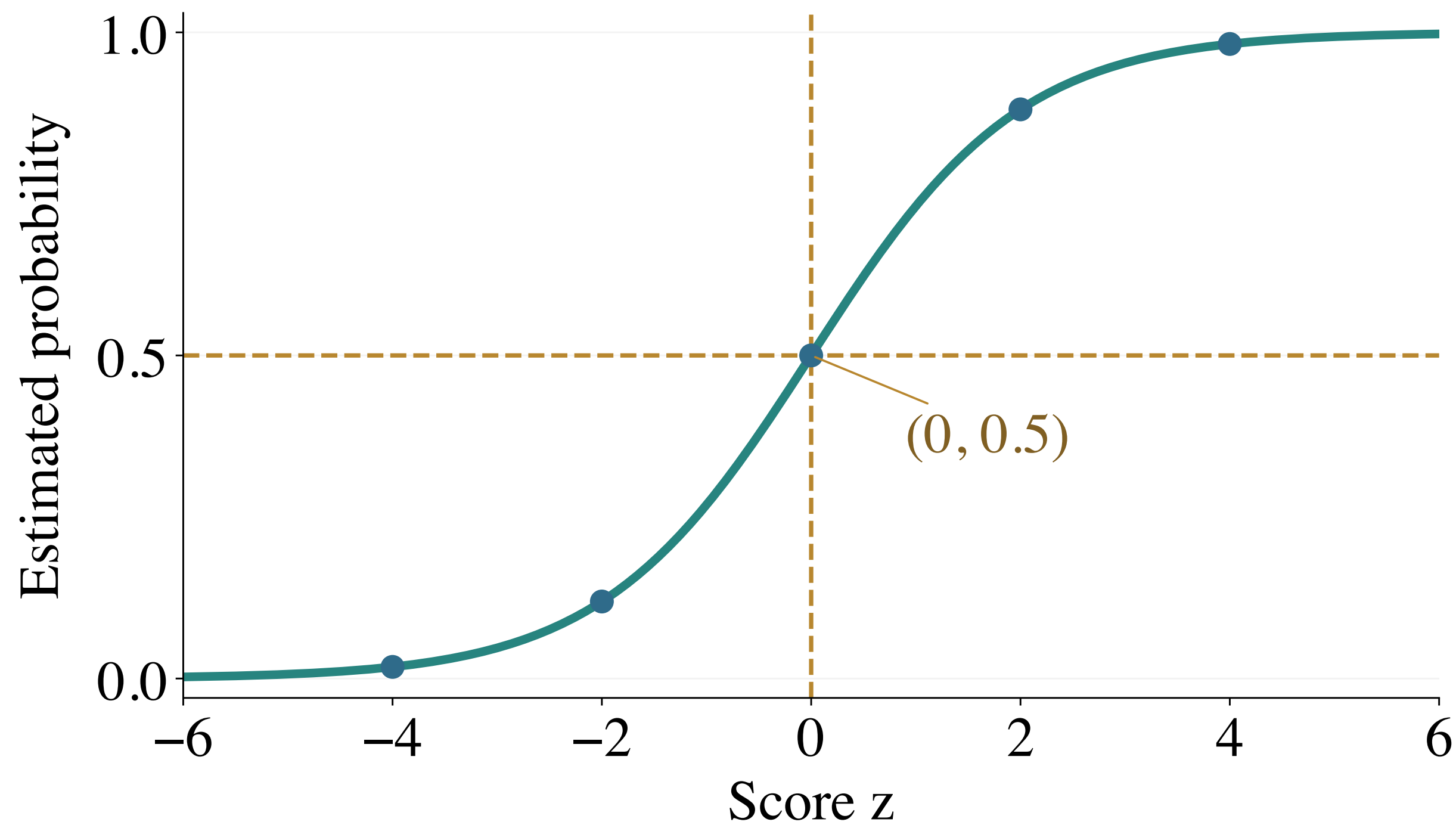
Learning chooses **which direction matters for the labels**, rather than which examples are nearby in every coordinate.

| A sigmoid converts the score into a **probability estimate**

$$\hat{\eta}(x) = \sigma(z) = \frac{1}{1 + e^{-z}}, \quad z = w^T x + b.$$

Score z	Probability of label 1
-4	1.8%
-2	11.9%
0	50%
2	88.1%
4	98.2%

Logistic regression uses this probability model.



| The probability changes smoothly; the **class boundary stays linear**

$$\hat{\eta}(x) \geq \frac{1}{2} \iff w^T x + b \geq 0.$$

- A larger positive score means a larger estimated probability of label 1.
- Multiplying all coefficients by 10 keeps the decisions but makes probabilities more extreme.
- An S-shaped probability function does **not** make this decision boundary curved.

| The probability model is fixed; now choose the fitting objective

Model: a linear score followed by a sigmoid.

Now: what loss should tell us whether its probabilities fit the labels?

Next: solve for the coefficients that minimize that loss.

| Accuracy gives little guidance **between boundary crossings**

For one positive review, compare scores $z = -3$ and $z = -0.1$.

Score	Predicted probability	Predicted class	Classification error
-3	4.7%	0	1
-0.1	47.5%	0	1

The probability moved toward the observed label, but accuracy sees **no improvement yet**.

A smooth loss can provide information about such changes before the class flips.

| Log loss rewards probability assigned to the **observed outcome**

Let $a = \hat{\eta}(x)$ and $y \in \{0, 1\}$.

$$\ell(y, a) = -y \log a - (1 - y) \log(1 - a) = \begin{cases} -\log a, & y = 1, \\ -\log(1 - a), & y = 0. \end{cases}$$

Give the observed outcome high probability $\rightarrow$ **small loss**.

Give the observed outcome nearly zero probability $\rightarrow$ **large loss**.

| Confident mistakes receive a larger penalty

For a positive outcome, $y = 1$:

Predicted probability a	Class at threshold 0.5	Log loss — $\log a$
0.99	Correct	0.010
0.60	Correct	0.511
0.40	Wrong	0.916
0.01	Wrong	4.605

Unlike classification error, log loss distinguishes **how confidently** a prediction was made.

| Would accuracy and log loss choose the same model?

Two observed outcomes are both positive: $y_1 = y_2 = 1$.

Model	Predicted probabilities	Accuracy	Average log loss
A	(0.60, 0.60)	100%	0.511
B	(0.99, 0.49)	50%	0.362

What does this comparison establish?

- A. Accuracy and log loss always rank models identically.
- B. Better log loss can coexist with worse thresholded accuracy.
- C. Model B must generalize better to every new population.

B. The criteria assess different outputs. Neither two-example score establishes population performance.

| MSE and log loss express **different prediction criteria**

Numerical regression	Logistic probability prediction
Output: $\hat{y} = z$	Output: $a = \sigma(z)$
Loss: $\frac{1}{2}(z - y)^2$	Loss: $-y \log a - (1 - y) \log(1 - a)$
Penalize numerical residual size.	Penalize low probability of the observed label.

Squared error on probabilities is also possible, but composing it with a sigmoid generally loses convexity **in the coefficients**.

| The logistic objective has **no general closed-form coefficient solution**

With $z_i = \mathbf{w}^\top \mathbf{x}_i + b$ and $a_i = \sigma(z_i) = e^{z_i} / (1 + e^{z_i})$:

$$\log a_i = z_i - \log(1 + e^{z_i}), \quad \log(1 - a_i) = -\log(1 + e^{z_i}).$$

$$\ell(y_i, a_i) = -y_i \log a_i - (1 - y_i) \log(1 - a_i)$$

(Usual binary log loss)

$$= -y_i z_i + [y_i + (1 - y_i)] \log(1 + e^{z_i})$$

(Substitute the two log identities)

$$= \log(1 + e^{z_i}) - y_i z_i$$

(Combine the coefficients)

Average the same loss over observations:

$$J(\mathbf{w}, b) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, a_i) = \frac{1}{n} \sum_{i=1}^n [\log(1 + e^{z_i}) - y_i z_i].$$

OLS gave linear stationary equations; logistic loss gives **nonlinear equations** in the coefficients.

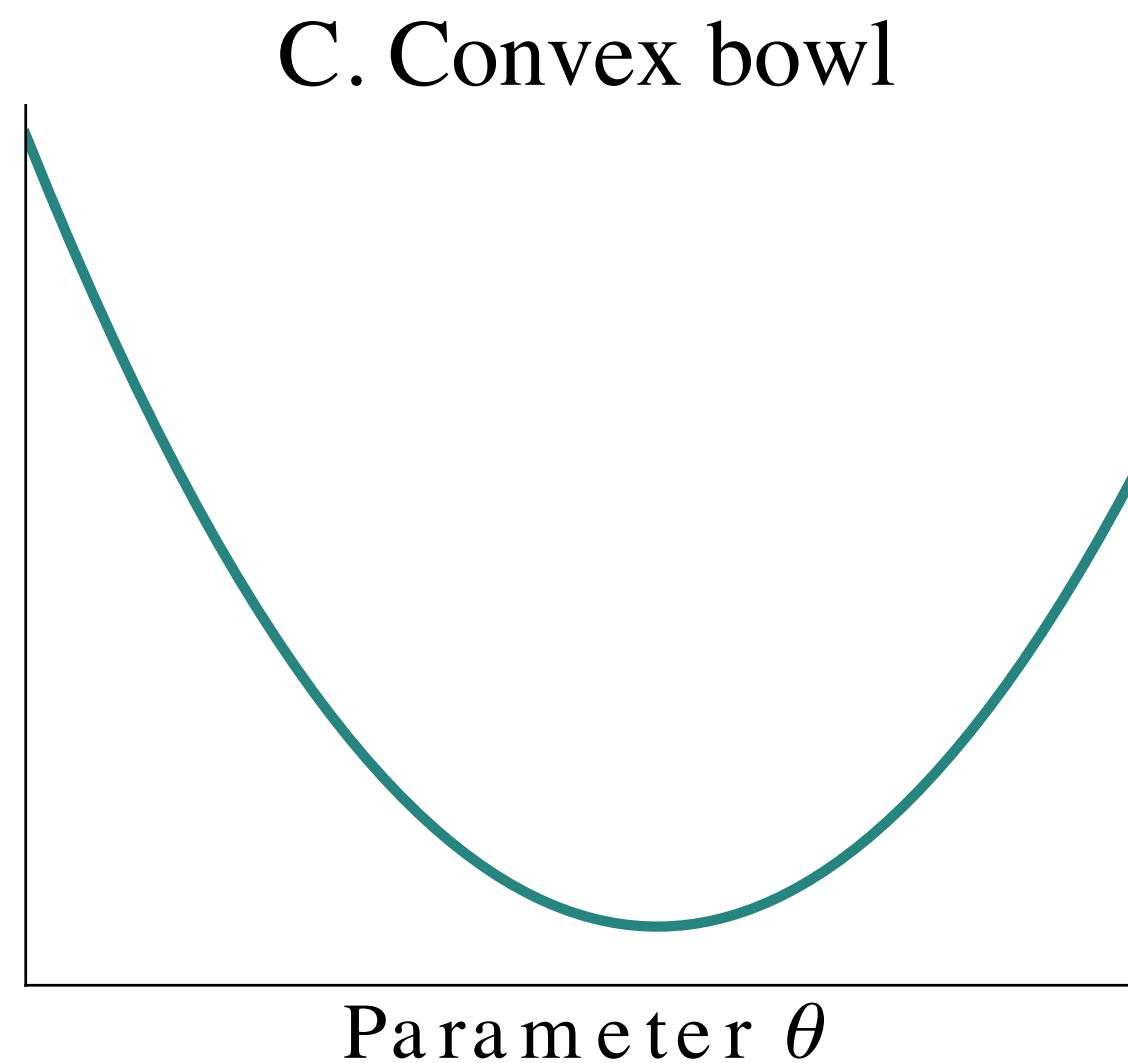
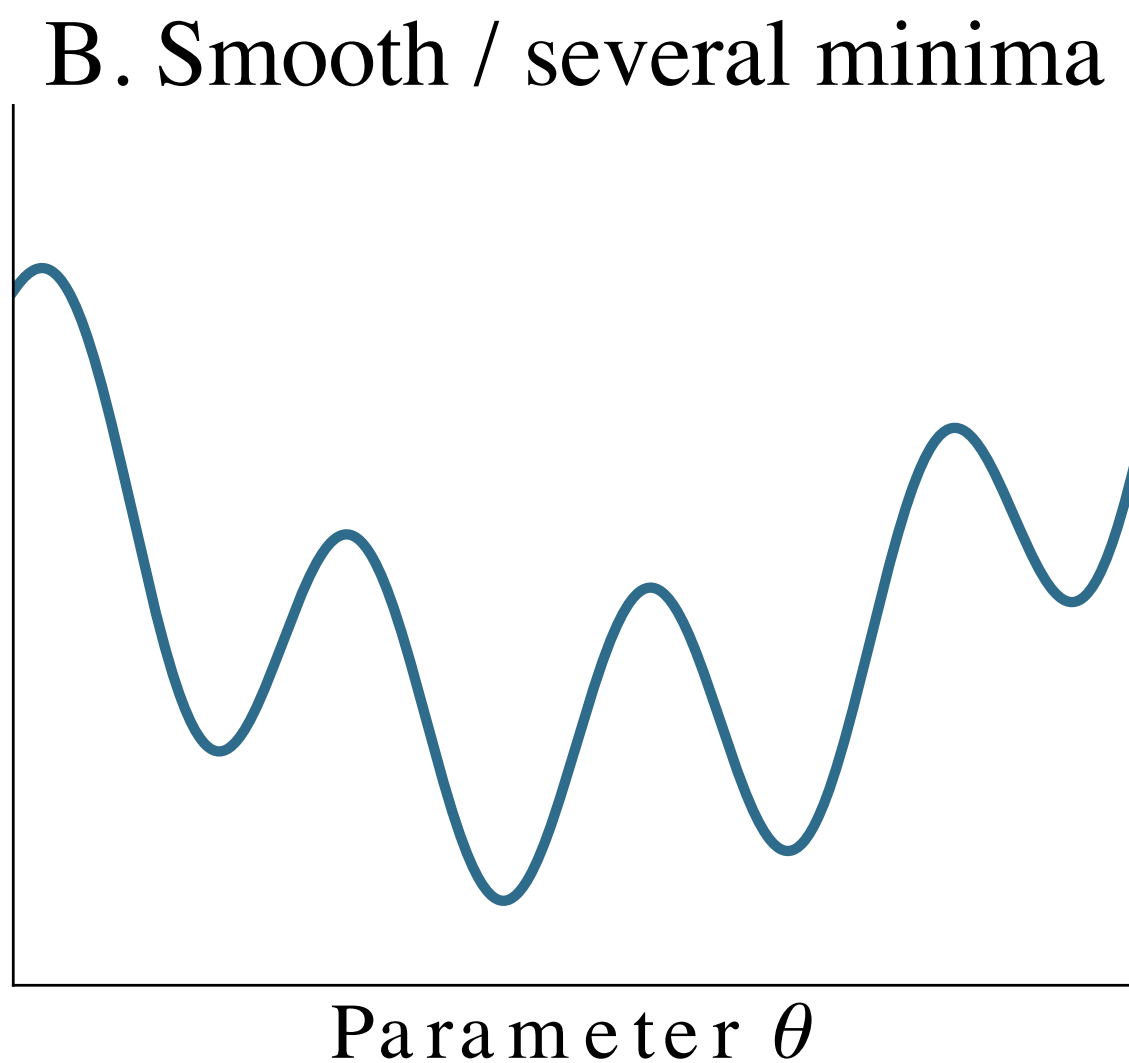
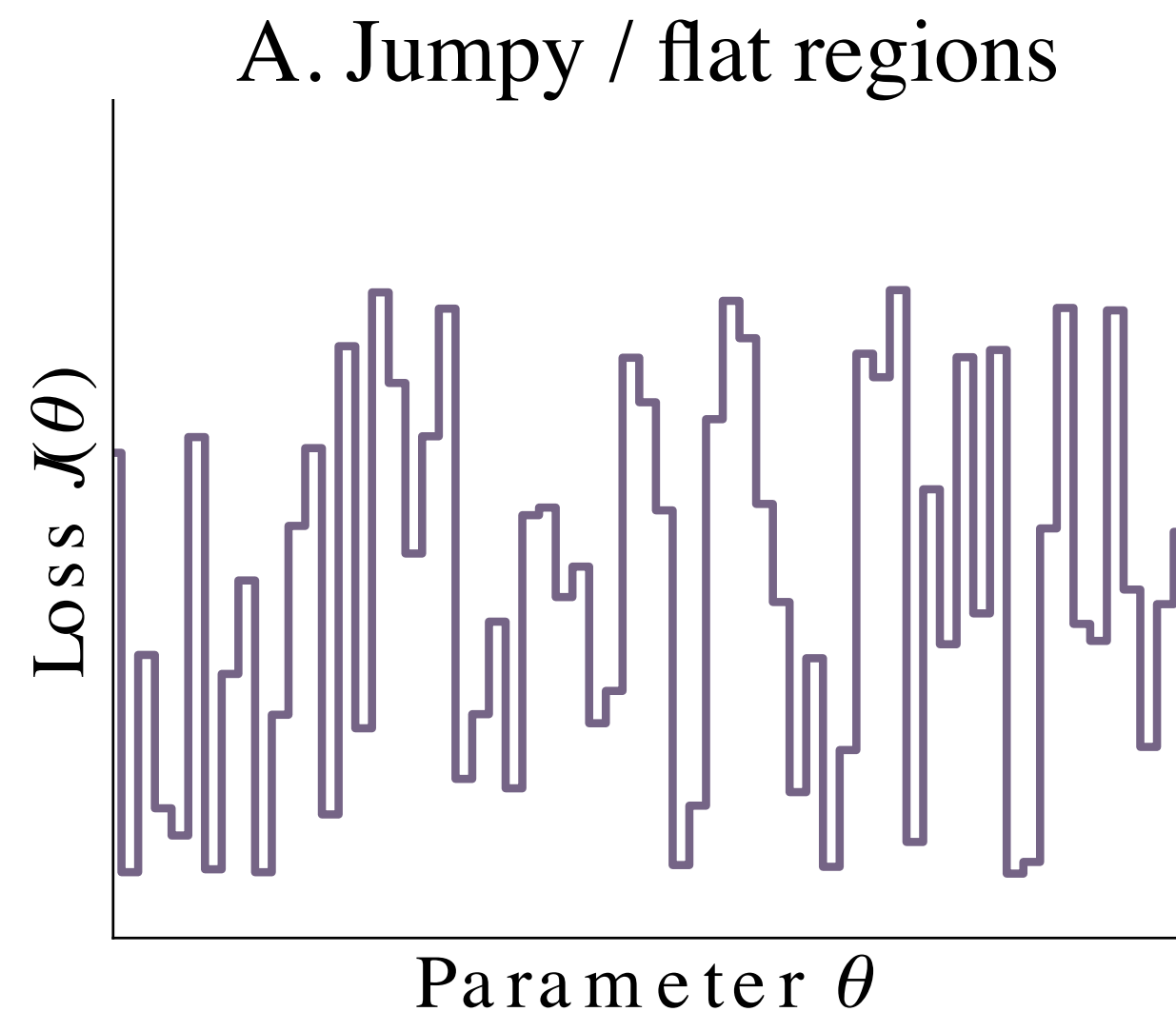
Next: what does the objective's shape guarantee?

Why the shape of the objective matters

| From a solved fit to a fitting algorithm

Stage	Question
1. Interpret the OLS solution	What does the exact fit establish?
2. Choose a classification model and loss	How should a score predict and fit labels?
3. Understand the objective's shape	When is a local optimum globally best?
4. Compute the coefficients	How can local information guide updates?

| Which loss would be easiest to optimize?



Which shape gives us the clearest route to a global minimum?

A. The jumpy loss.

B. The smooth loss with several minima.

C. The convex bowl.

C. A local minimum is already global; there are no worse local minima to escape.

| A useful objective lets us **reason about the solution**

- **Jumpy:** a small change may reveal little—or abruptly change the loss.
- **Smooth, several minima:** downhill progress can end at a worse local minimum.
- **Convex:** a local minimum is global, making optimality easier to analyze.

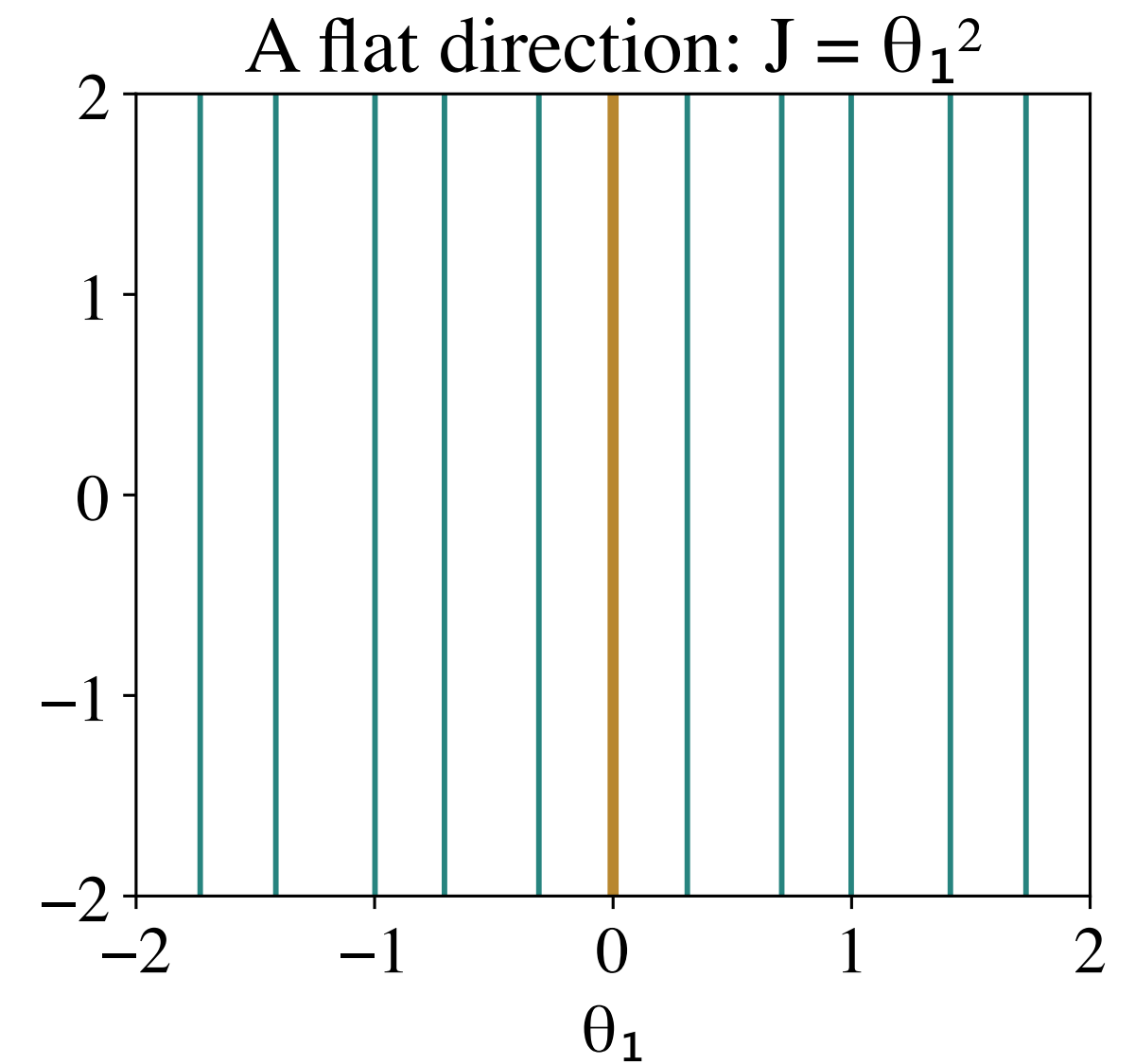
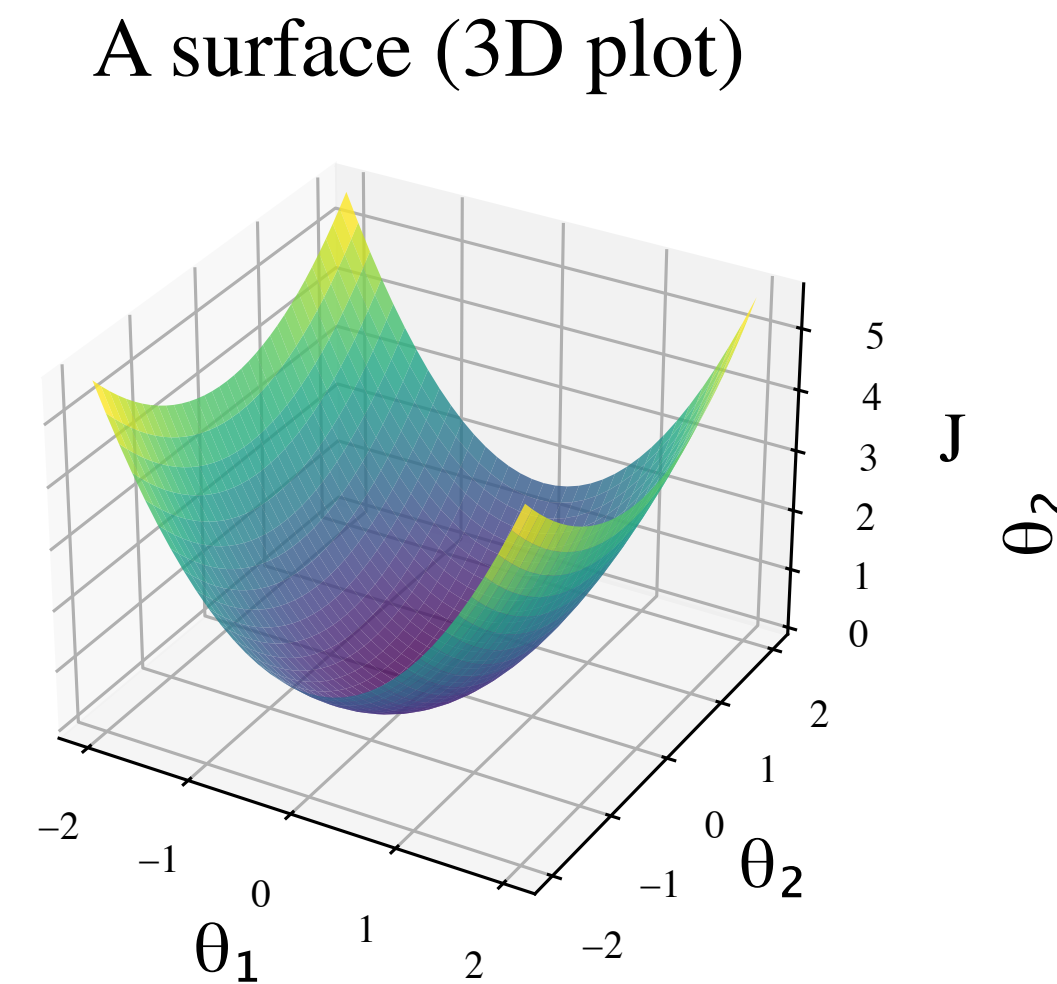
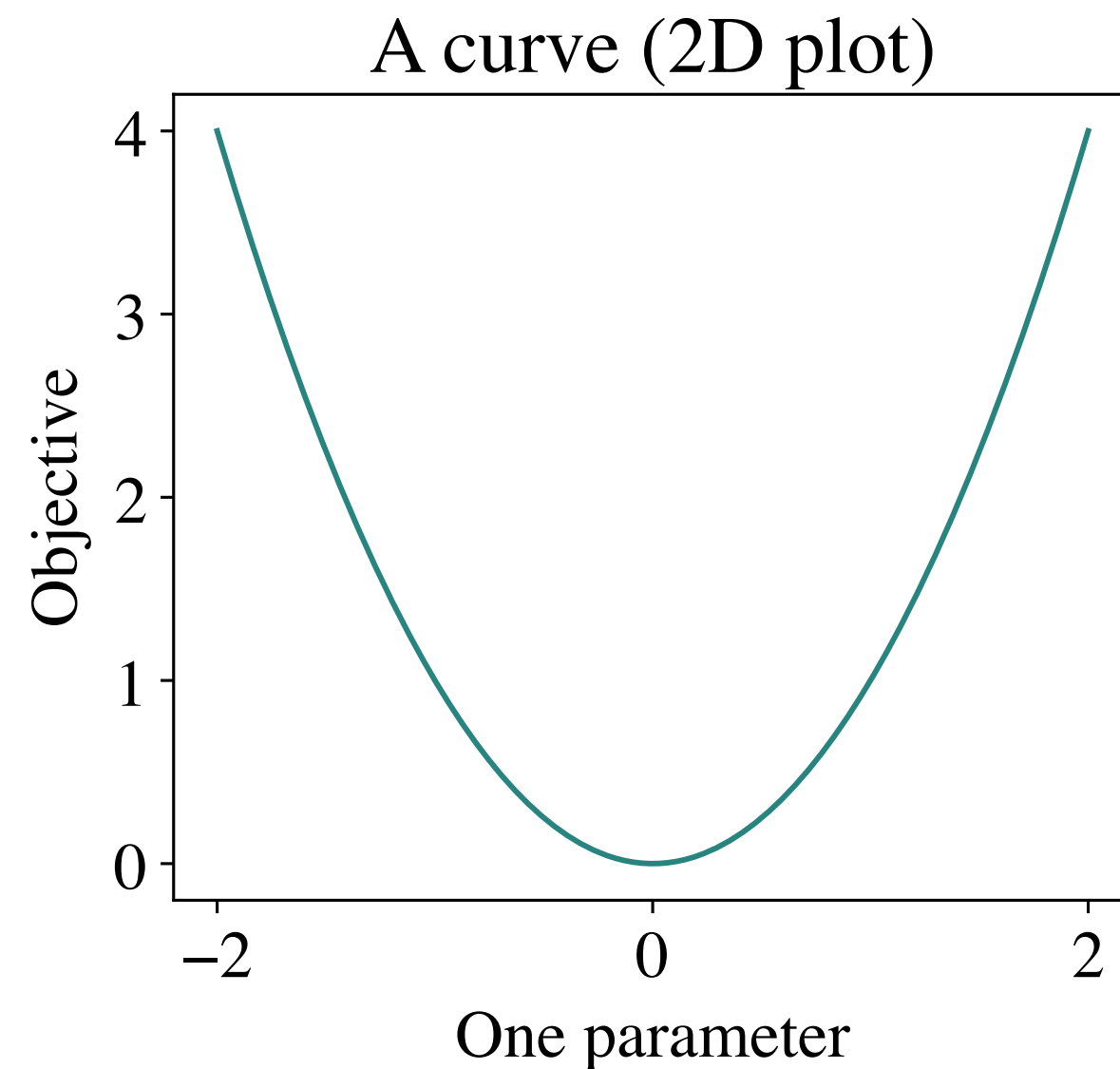
Convexity is a useful class of well-behaved objectives.

It gives a global guarantee; speed still depends on the problem and algorithm.

| Objective shape answers **three linked questions**

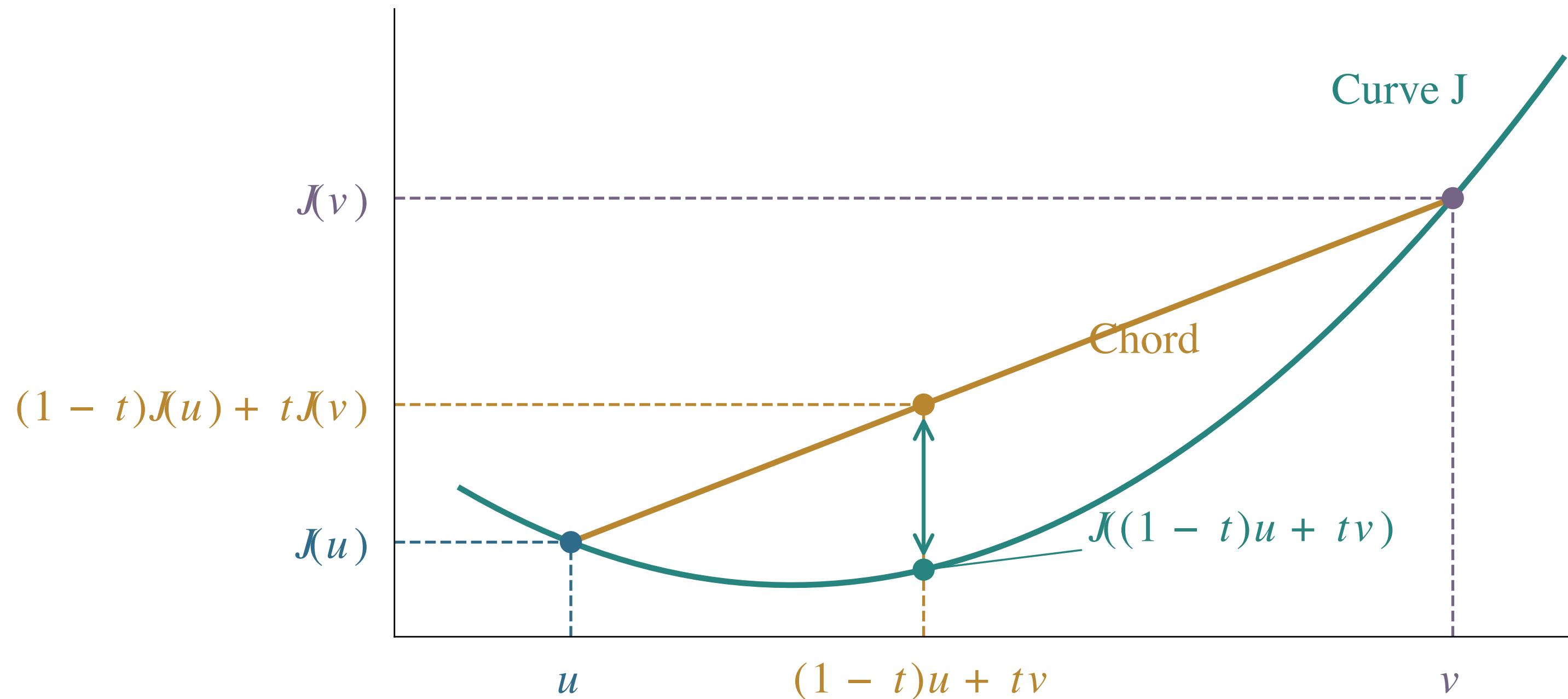
1. **Define:** what does convexity mean geometrically?
2. **Check:** are our OLS and logistic objectives convex?
3. **Interpret:** what does that guarantee—and what else do we need?

| Convexity is a statement about the entire objective shape



A convex objective may have a single bottom—or a flat set of equally good solutions.

| Definition: a convex function lies below every chord



J is **convex** on a convex domain if, for **every** u, v in the domain and $0 \leq t \leq 1$,

$$J((1-t)u + tv) \leq (1-t)J(u) + tJ(v).$$

Loss at the averaged input $\leq$ average of the endpoint losses.