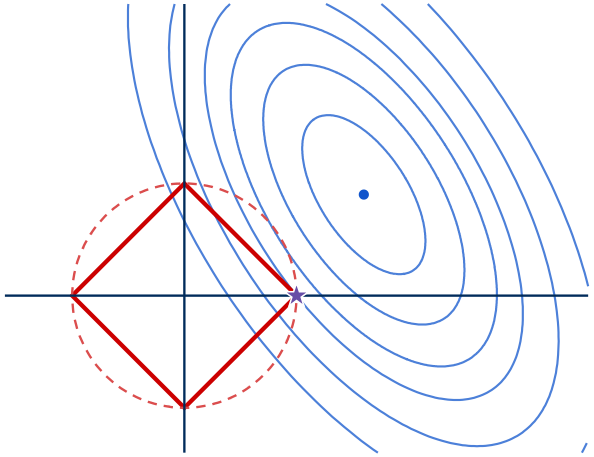
UNIT 02 · CHAPTER 6

Optimization and Regularization



CSE 30124 · Introduction to Artificial Intelligence
William Theisen · University of Notre Dame

An iterative method

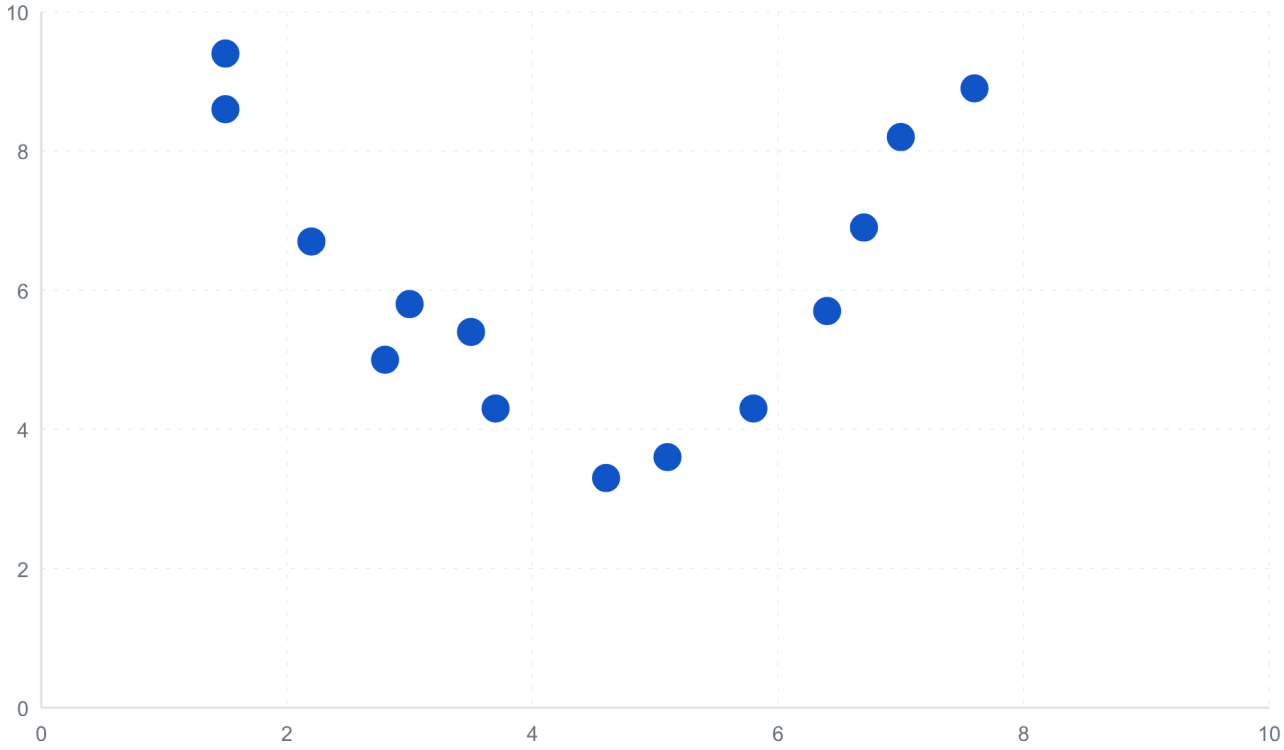
Instead of jumping straight to the bottom of the bowl, we can _____ a little at a time.



An iterative method

Overfitting

Unfortunately, **Linear Regressions** (a form of MLE), are **prone to overfitting**.



Regularization: L_1 and L_2

To help **prevent overfitting** when using **linear regressions**, we can perform a technique called **regularization**.

The goal of **regularization** is to **encourage models to learn the broader patterns** within the data rather than memorizing it.

L_2 regularization: Ridge

L2 regularization (called **ridge regularization**) **encourages** _____ _____ **weights**.

L_1 regularization: LASSO

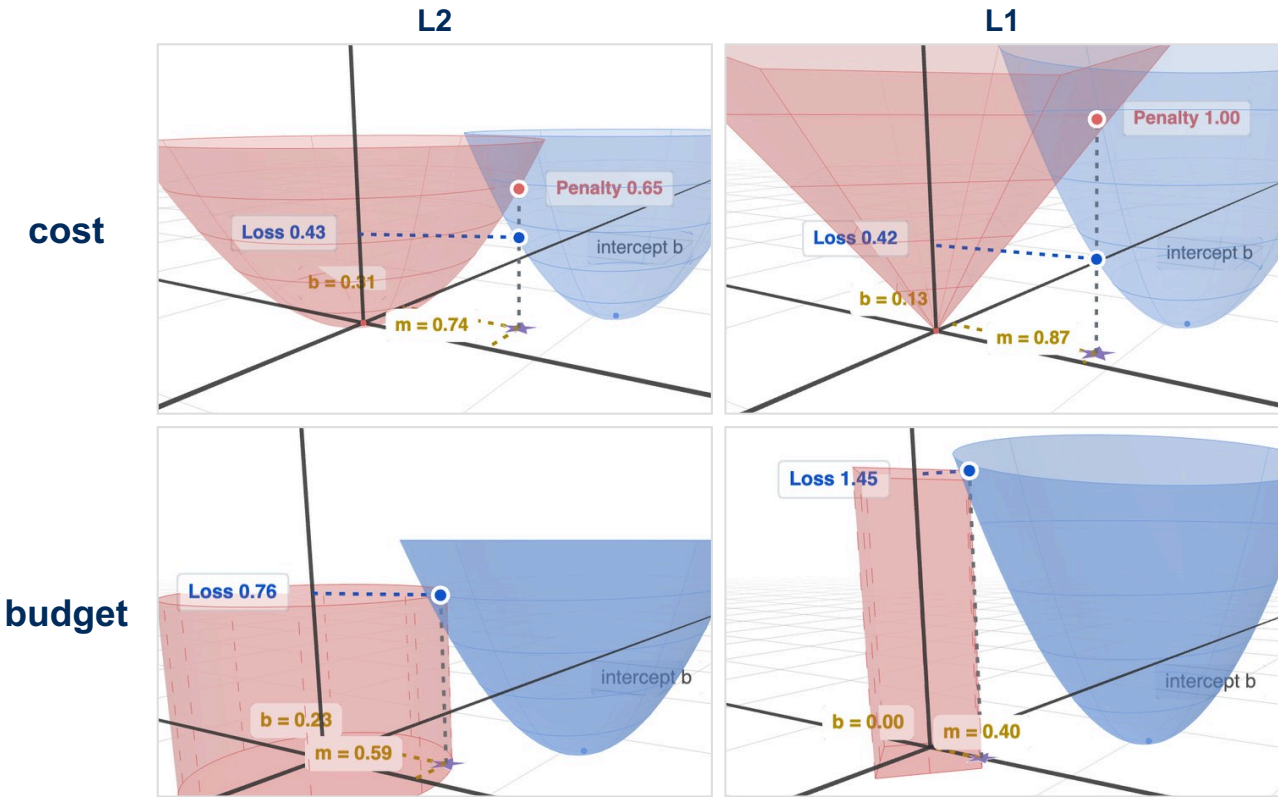
L1 regularization (also called **LASSO**) leads to _____ and **sparse models**.

Formalizations of regularization

There are two ways to formalize regularization:

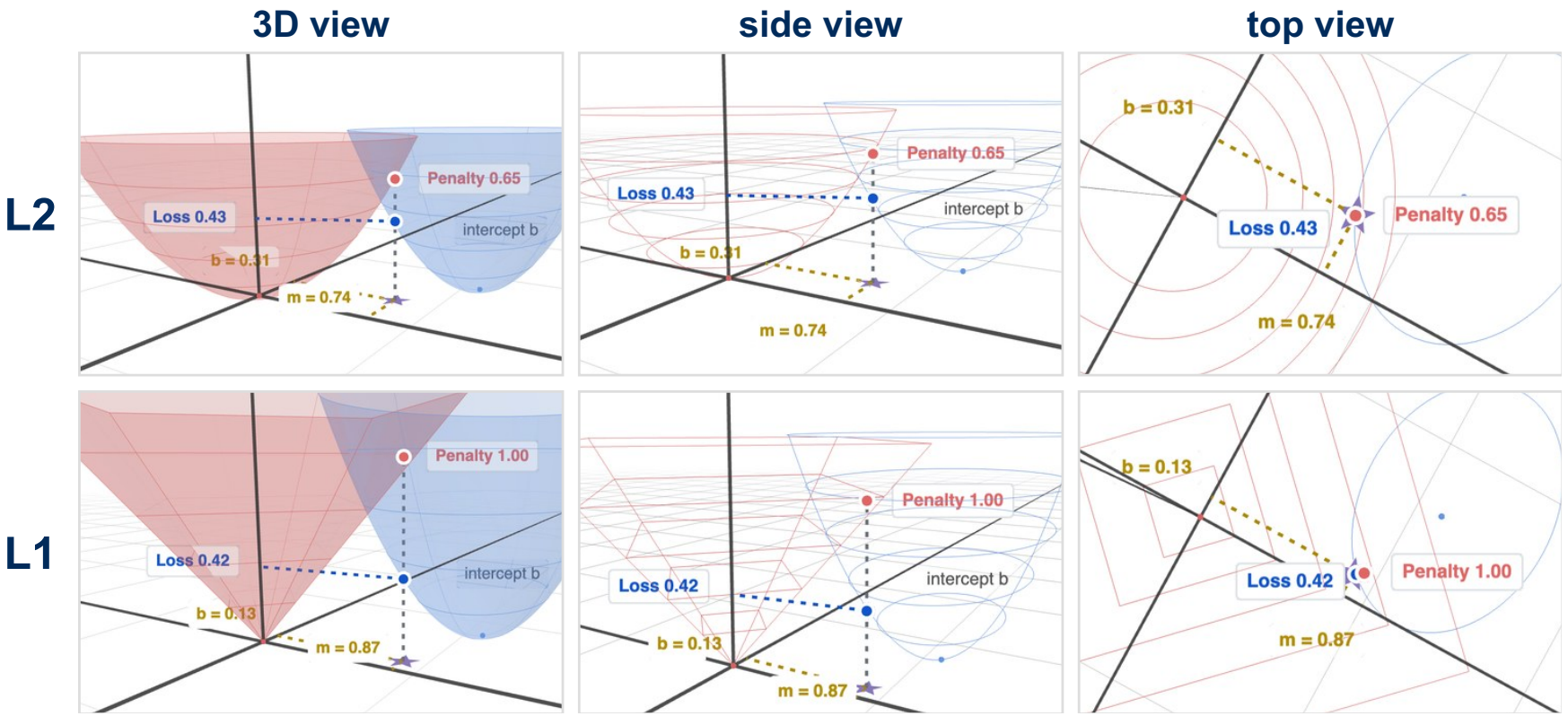
- As a _____ placed on the **weights**
- As a _____ for the **weights**

Formalizations of regularization



Regularization as a cost

When we formalize **regularization** as a **cost**, we modify the **loss function** such that increasing the weights incurs a "cost".

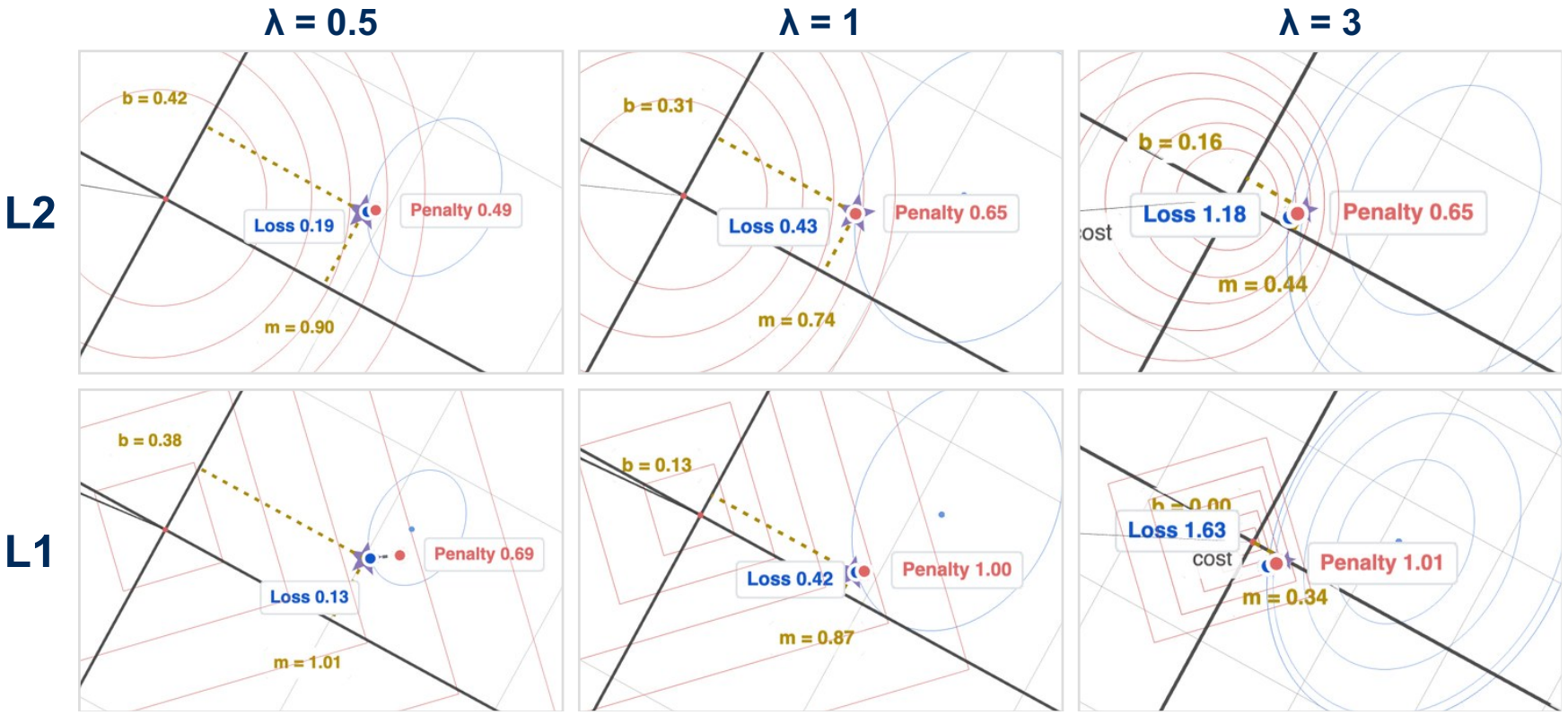


Regularization as a cost

We use an argument called λ to determine how heavily the **cost** is weighed. A larger λ increases the **cost**.

Regularization as a cost, increasing λ

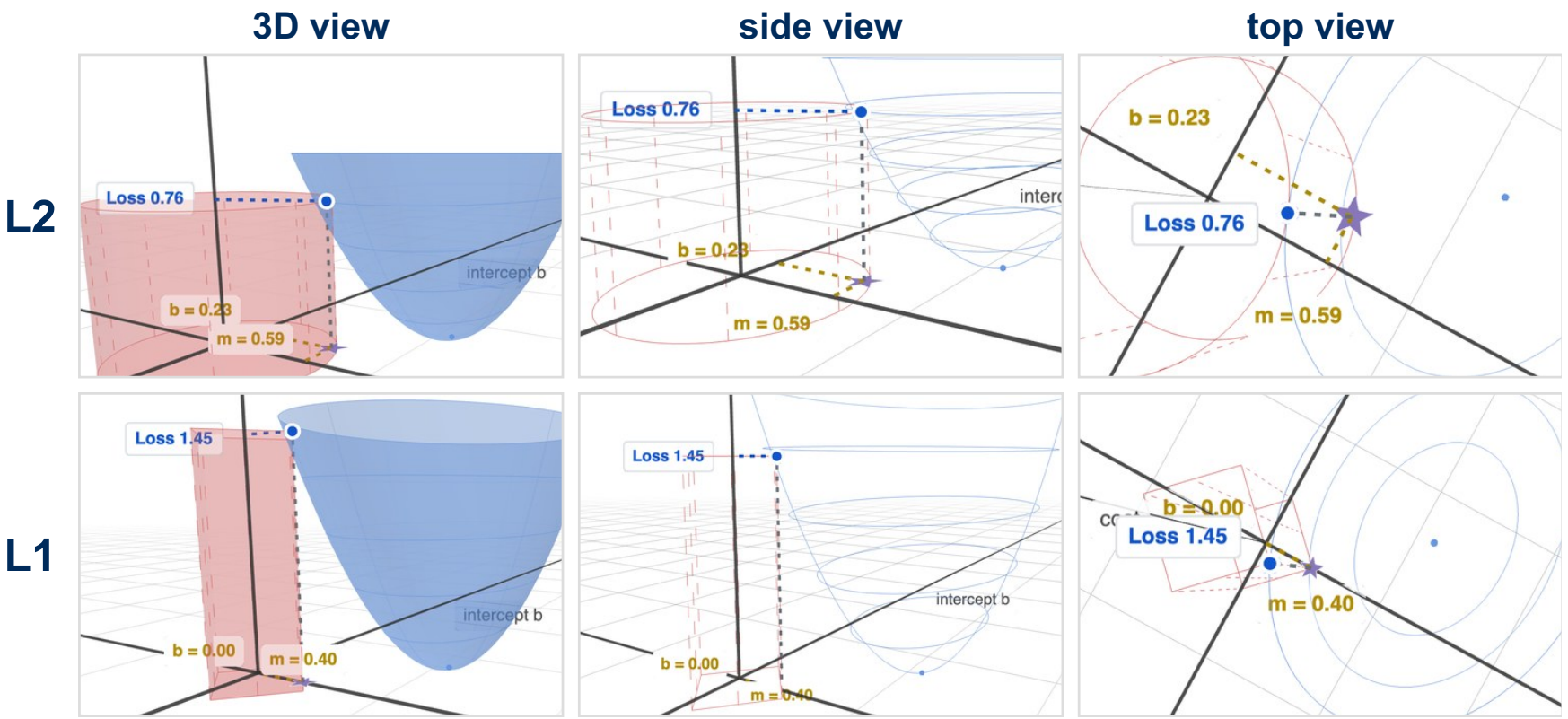
A larger λ increases the **cost** of our weights, typically resulting in the optimum being pulled closer to the origin.



Regularization as a budget

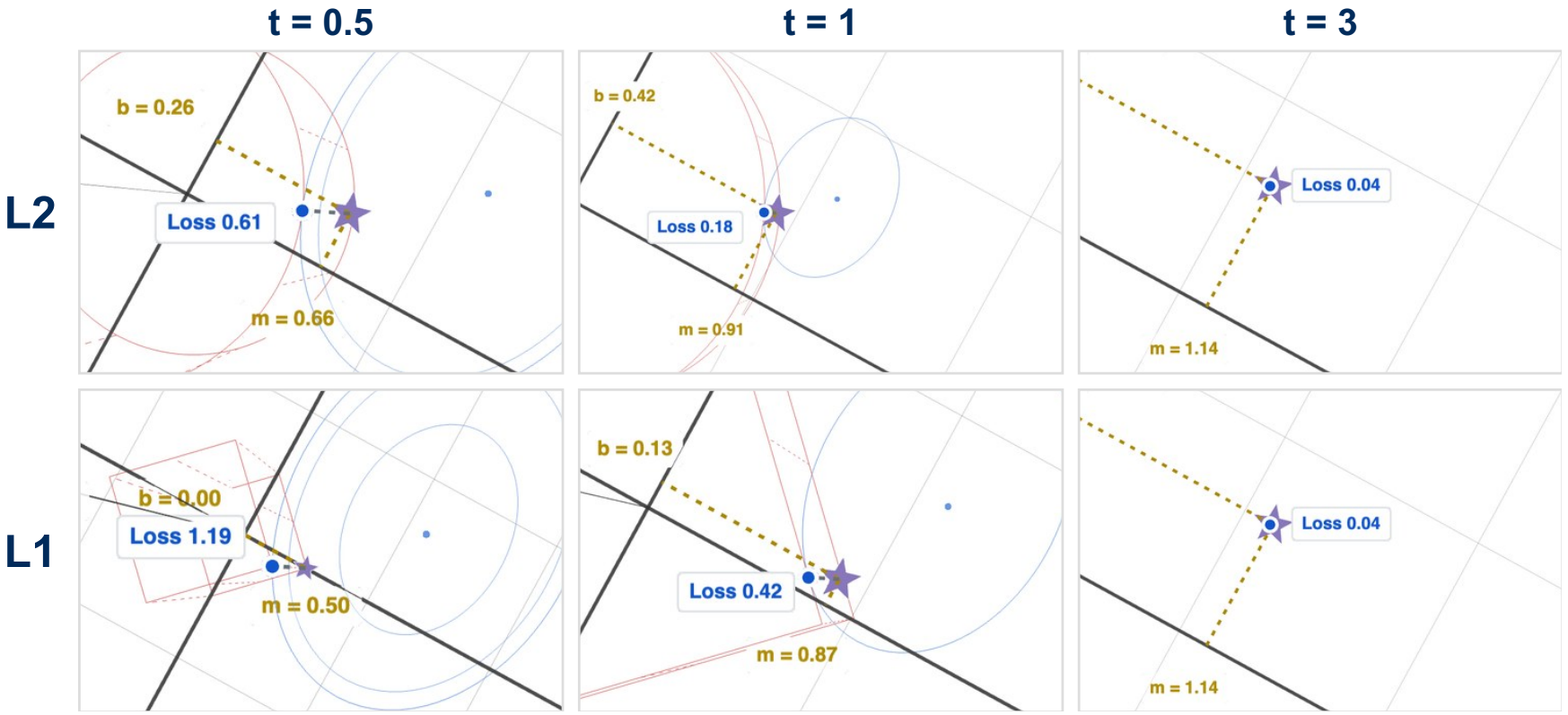
When we formalize **regularization** as a **budget**, we constrain our **loss function** such that the weights chosen must remain within "budget".

Regularization as a budget



Regularization as a budget, increasing t

A larger t increases the **budget** for our weights, typically resulting in the optimum being closer to the unconstrained optimum.



A Budget in the Brain

The brain is about 2% of our body weight but uses about 20% of our energy (around 20 watts). Firing neurons and keeping synapses going is expensive, so it **works on a _____**.

One result is **sparse activity**: at any moment only a small fraction of neurons are firing, a bit like L1 pushing most weights to zero.

Another is **synaptic pruning**. Synapse density peaks in early childhood and drops through adolescence (Huttenlocher, 1979), so the brain **keeps the connections it uses and removes the rest**.

MAP estimation

L1 and **L2 regularization** are a form of **MAP Estimation**.

MAP stands for _____ which is a **Bayesian** approach for **applying a prior to our weights**.

Review

Complete the review: **L1 regularization** penalizes _____ _____. **L2 regularization** penalizes _____. In a penalty formulation, increasing **lambda** makes the penalty _____. In a budget formulation, increasing the budget makes the constraint _____.

Matching: regularization

Write the letter of the matching definition next to each term.

_____ **MAP Estimation**	**A.** Penalizes the squared size of the weights, encouraging small weights
_____ **L2 (Ridge)**	**B.** Controls how heavily the regularization cost is weighed
_____ **L1 (LASSO)**	**C.** The budget the weights must remain within
_____ λ	**D.** Penalizes the absolute size of the weights, driving some to exactly 0
_____ t	**E.** A Bayesian approach for applying a prior to our weights

Practice problems

1. Units and the penalty. Last chapter's line is $\hat{y} = 9.0\ x_1 - 6.7$ with x_1 in hundreds of yards, or $\hat{y} = 0.090\ x_1 - 6.7$ with x_1 in plain yards. The two make exactly the same predictions. Why might ridge treat them differently?

2. Ridge or LASSO? Total yards is pass yards plus rush yards. We put all three in one model. Plain OLS gives weights of -64, $+72$ and $+74$. What would ridge do with these three features, and what would LASSO do? Which would you use to find out which stats matter?

3. Reading λ as a prior. Our yards-and-turnovers line has $\sigma^2 \approx 60$. On the MAP pages, a prior of $\tau = 5$ gave $\lambda \approx 2.4$. What λ do we get if we expect smaller weights, $\tau = 2$? Will the fitted weights get bigger or smaller?