REGULARIZATION | SPARSITY

Feature crosses can significantly increase ~~feature~~ space, model size, eval time, etc.

Want to zero-out some weights / eliminate coefficients.

L1 regularization = penalize sum of abs. vals. of all the weights
↳ encourages sparsity

Differences { L2 penalizes $weight^2$
L1 penalizes $|weight|$ }

↳ So they have different derivatives!
The L1 derivative drives L1 to zero faster (since L2 deriv. is smaller as L2 shrinks)

Caveats of L1:

- Can eliminate weakly informative features
- or strongly informative w/ diff. scale
  ↳ Features need to be normalized to similar scales

E.g. $\frac{\text{val} - \text{mean}}{\text{std. dev.}}$ (# of std. dev.s from the mean)