L0 and L1/2 Regularization
Sparsity Beyond L1

Team Name (A, B, C, D)

Machine Learning

December 17, 2025

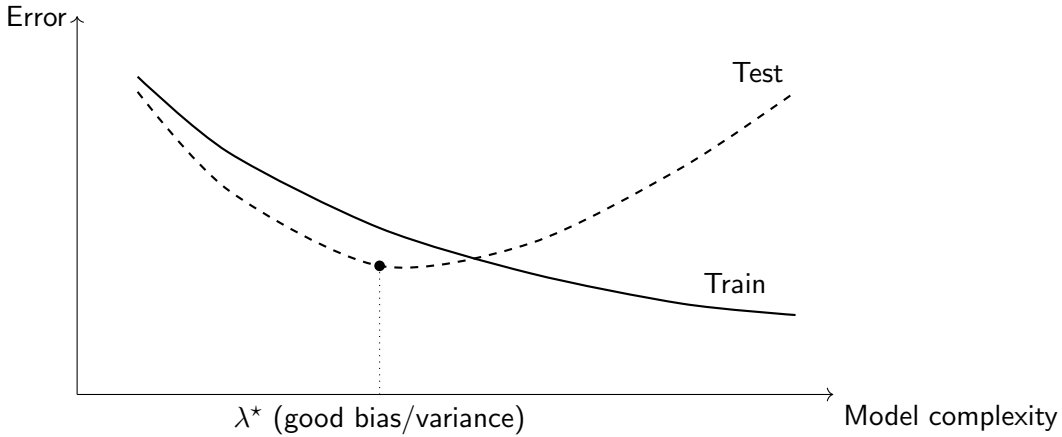
Outline

Why regularize? (Motivation)

- **Overfitting vs generalization:** training error decreases, test error may increase.
- **High-dimensional ML:** many features, noisy / correlated inputs.
- **Why we like sparsity:** interpretability, lower measurement cost, faster inference, robustness.

Goal today: encourage models with **few non-zero coefficients**.

Overfitting picture (intuition)



Regularization controls complexity $\Rightarrow$ better generalization.

What is sparsity?

- A model is **sparse** if many parameters are **exactly zero**.
- In linear models: $y \approx X\beta$, sparsity means only a few features are used.
- Benefits:
  - interpretability (feature selection),
  - lower cost (measure fewer features),
  - deployment (faster + simpler models).

Sparsity illustration

$$\boldsymbol{\beta} = \begin{bmatrix} 2.3 \\ 0 \\ 0 \\ -1.1 \\ 0 \\ 0 \\ 0.7 \\ 0 \end{bmatrix}$$

Sparse: only 3 non-zeros
(feature selection)

L0 regularization: definition

L0 "norm" counts non-zero coefficients:

$$\|\boldsymbol{\beta}\|_0 = \#\{j : \beta_j \neq 0\}.$$

Penalized objective:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \ \mathcal{L}(\boldsymbol{\beta}) + \lambda \ \|\boldsymbol{\beta}\|_0$$

where $\mathcal{L}(\boldsymbol{\beta})$ is the loss (e.g., MSE, logistic loss) and $\lambda > 0$ controls sparsity.

Interpretation: "fit the data while using as few features as possible."

Intuition: best subset selection

- L0 regularization is closely related to **best subset selection**:
  - choose a subset $S \subset \{1, \ldots, p\}$,
  - fit the model using only features in S,
  - pick the best subset size (via λ or k).
- Often yields **very sparse** and **highly interpretable** models.

Why L0 is hard

- The objective is **non-convex** and **discontinuous**.
- It implies a **combinatorial search** over feature subsets:

$$\text{number of subsets} \approx 2^p.$$

- Practical consequence: exact optimization becomes infeasible when p is large.

How people handle L0 in practice

Common strategies (choose depending on p and compute budget):

- **Greedy methods:** forward selection, Orthogonal Matching Pursuit (OMP).
- **Relaxations:** replace L0 by L1 (convex surrogate).
- **Mixed-Integer Optimization:** can solve exact/near-exact for smaller p.
- **Smooth surrogates:** continuous approximations to the counting function.

L1/2 regularization: definition

L1/2 is an L_p penalty with $p = \frac{1}{2}$ (non-convex):

$$\sum_{j=1}^{p} |\beta_j|^{1/2}.$$

Penalized objective:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \; \mathcal{L}(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|^{1/2}.$$

Note: this is **not** Elastic Net (which is a mix of L1 and L2).

Why L1/2 can be more sparse than L1

- For $p < 1$, the penalty is **more aggressive near zero** than L1:
  - small coefficients are pushed to exactly 0 more strongly,
  - often yields **stronger sparsity** than L1.
- Intuition often cited: **less shrinkage bias** on large coefficients than L1 (depends on solver/settings).
- Trade-off: **non-convex** objective $\Rightarrow$ optimization is harder.

Optimization intuition (high-level)

Typical solver families (no heavy math needed):

- **Iterative reweighting:** solve a sequence of easier weighted problems that approximate L1/2.
- **Thresholding / proximal-style updates:** non-linear shrinkage rules.

Practical notes:

- initialization matters,
- possible local minima (non-convex).

Comparison: L0 vs L1/2 vs L1

	L0	**L1/2**	**L1**
Sparsity strength	Very high (ideal)	High	Medium/High
Convex?	No (discrete)	No ($p < 1$)	Yes
Optimization	Hard (combinatorial)	Hard/medium (non-convex)	Easier (convex)
Stability	Can be unstable	Depends on solver/init	Usually stable
Typical use-cases	Strict subset selection	Need more sparsity than L1	Strong baseline, interpretable

When to use what?

Use L1

- Convex + reliable
- Good baseline
- Fast training

Use L1/2

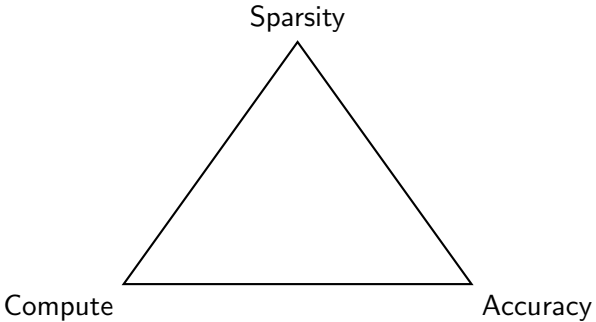
- Need stronger sparsity
- Accept non-convexity
- Careful optimization

Use L0

- Strict feature subset
- Manageable p
- Specialized solvers

Conclusion

- **L0:** direct sparsity objective, very interpretable, but computationally hard.
- **L1/2:** a bridge between L0 and L1; often yields stronger sparsity than L1, but non-convex.
- Always balance: **accuracy** vs **sparsity** vs **compute**.



References (minimal)

- Course notes / lecture slides on regularization and sparse modeling.
- Best subset selection / L0 regularization: classic regression model selection literature.
- Non-convex sparse penalties (L_p with $p < 1$): overview papers on non-convex regularization (optional).