L_0 and $L_{1/2}$ Regularization

SALHI Aymane , AIT HSSAIN Zakaria
ABAYAD Mehdi , SOSSEY Slamane

– 2IA –

December 19, 2025

Outline

From ERM to Regularization

Empirical Risk Minimization (ERM) chooses parameters that fit best the training dataset S.

$$\hat{\beta}_{\text{ERM}} \in \arg\min_{\beta} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i\,,\, f_{\beta}(x_i)\right)}_{L_S(f_\beta)}$$

- If the model is flexible, ERM can fit noise $\Rightarrow$ **overfitting**.
- We want good performance on **unseen data**: minimize **test/true risk** $L_D(f_\beta)$.
- Key idea: control model complexity to reduce the **generalization gap**.

Regularization as a complexity control knob

Regularized objective:

$$\hat{\beta} \in \arg\min_{\beta} \left[\underbrace{L_S(f_\beta)}_{\text{Loss}} + \lambda \times \underbrace{\Omega(f_\beta)}_{\text{Penalty}} \right]$$

- $L_S(f_\beta)$: Empirical error over S (MSE, 0-1 loss, ...)
- $\Omega(f_\beta)$: complexity penalty (norm/quasi-norm)
- $\lambda > 0$: trade-off parameter

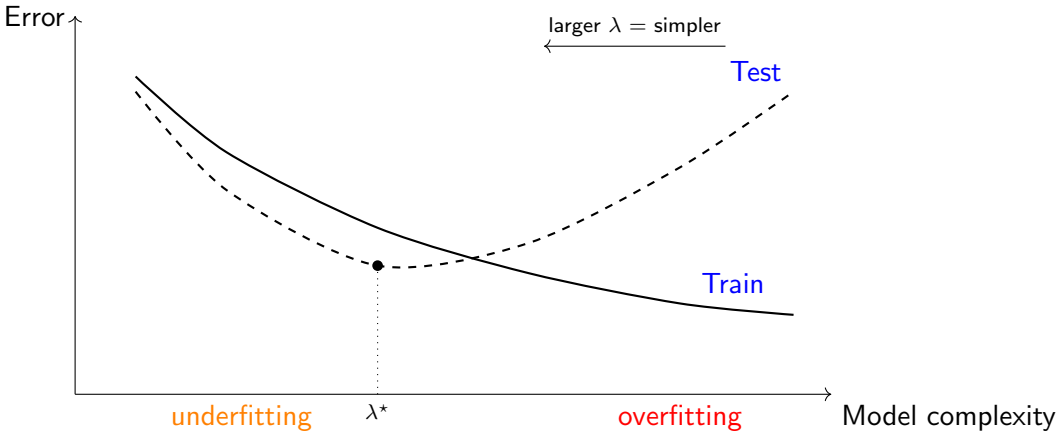
Why overfitting happens (bias variance trade-off)

- Increasing model complexity typically:
  - **decreases bias** (model can fit more patterns),
  - **increases variance** (model becomes sensitive to noise/data changes).

A classic view of expected test error (conceptually):

$$\mathbb{E}\big[L_D(h \in \mathcal{H})\big] \approx \text{Bias}^2 + \text{Variance} + \text{Noise}$$

- Regularization reduces variance by shrinking / simplifying the solution.



Regularization controls complexity $\Rightarrow$ better generalization.

What is sparsity?

Definition : A model is **sparse** if many parameters are **exactly zero**.

- Linear models: $y \approx X\boldsymbol{\beta}$
- Define the **support** (selected features):

$$\mathrm{supp}(\boldsymbol{\beta}) = \{\, j \,:\, \beta_j \neq 0 \,\} \quad , \quad |\mathrm{supp}(\boldsymbol{\beta})| = \|\boldsymbol{\beta}\|_0$$

- Sparsity means: only a small subset of features truly "matters".

Why sparse models are preferred ?

- **Interpretability:** easier to explain which variables drive predictions.
- **Measurement / data cost:** fewer features to collect (medical, finance, IoT, . . .).
- **Robustness:** reduces reliance on noisy/irrelevant variables.

Sparsity illustration

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

Sparse: only 3 non-zeros

Selected features: $\text{supp}(\boldsymbol{\beta}) = \{1, 4, 7\}$

L_0 regularization: definition

L_0 **regularization** adds a penalty equal to the count of non-zero weights (vector's sparsity) in the model:

$$\|\omega\|_0 = card(\{i \ / \ \omega_i \neq 0\}).$$

Penalized objective:

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

- p: the number of weights
- $\mathcal{L}(\omega)$: loss function
- $\lambda > 0$ regularization strength

Interpretation: **"minimize the loss function while using as few features as possible."**

Intuition: best weights subset selection

- L_0 is the most **direct and ideal** sparsity control.
- Often yields **very sparse** and **highly interpretable** models.

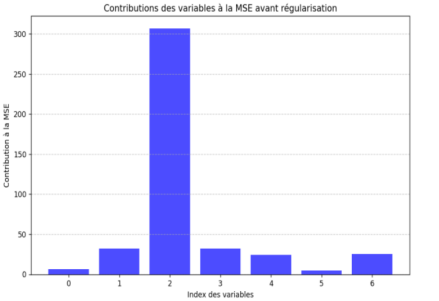
L_0 regularization algorithm

- choose an initial subset of weights : $best_S = S_0$
- for each subset S of the p weights:
  1. Fix weights outside S to zero
  2. Optimize the weights in S to minimize the loss function
  3. Compute the objective function $Obj(\omega_S)$
  4. if $Obj(\omega_S) \leq Obj(\omega_{best_S})$: set $best_S$ to S

Why L0 is not Practical

- **Combinatorial problem in discrete space**:
  - for each weight, whether use it or prune it
  - NP-Hard
  - 2^p possiblities (far more than atoms in the universe for only 1million params)
- **Non-differentiability**: Can't use Gradient based optimization (required by DL).

Regularization strength




```
lambda_reg = 11
Indices des variables supprimées : [0, 1, 3, 5]
MSE avant régularisation: 7.898610469088357
MSE après régularisation: 18.766727123292448
Nombre de variables avant régularisation: 7
Nombre de variables après régularisation: 3
```

```
lambda_reg = 20
Indices des variables supprimées : [0, 1, 3, 4, 5, 6
MSE avant régularisation: 7.898610469088357
MSE après régularisation: 22.202762820115172
Nombre de variables avant régularisation: 7
Nombre de variables après régularisation: 1
```

```
lambda_reg = 0.1
Aucune variable n'a été supprimée par la régularisation L0.
MSE avant régularisation: 7.898610469088357
MSE après régularisation: 7.898610469088357
Nombre de variables avant régularisation: 7
Nombre de variables après régularisation: 7
```

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:

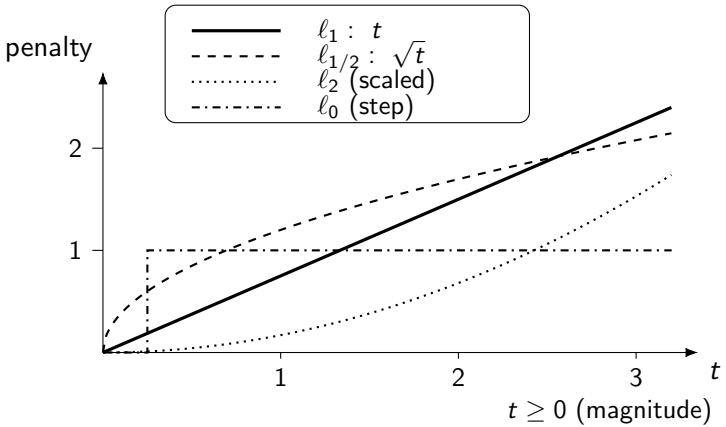
$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \; \mathcal{L}(\boldsymbol{\beta}) \; + \; \lambda \sum_{j=1}^{p} |\beta_j|^{1/2}.$$

Note: this is **not** Elastic Net (which mixes 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.
- Often cited: **less shrinkage bias** on large coefficients than L1 (problem and solver dependent).
- Trade-off: **non-convex** objective $\Rightarrow$ harder optimization.

Visual intuition: penalty shape (TikZ)



Message: $\sqrt{t}$ is **steep near** 0 (kills small coefficients) and **concave** (closer to L0 than L1).

Why "steeper near zero" matters

For $t > 0$:

$$\frac{d}{dt}\sqrt{t} = \frac{1}{2\sqrt{t}} \quad \Rightarrow \quad \text{very large when } t \to 0^+.$$

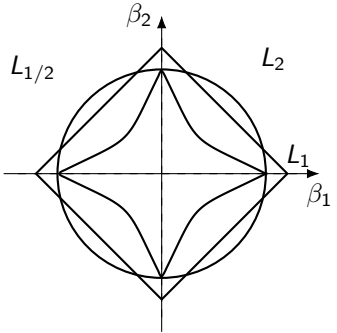
- Near zero, the penalty gradient is very large.
- Small coefficients experience a strong push toward 0.
- This explains stronger sparsity compared to L1.

Geometry intuition: why L1/2 promotes sparsity

Many penalized problems can be seen as constrained ones:

$$\min_{\boldsymbol{\beta}} \ \mathcal{L}(\boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta}) \quad \Longleftrightarrow \quad \min_{\boldsymbol{\beta}} \ \mathcal{L}(\boldsymbol{\beta}) \text{ s.t. } P(\boldsymbol{\beta}) \leq c.$$

Sparsity intuition: sharper corners $\Rightarrow$ optimum more likely on an axis ($\beta_j = 0$).



Optimization intuition (high-level)

Typical solver families:

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

Practical notes:

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

Iterative reweighting idea

Approximate the non-convex penalty by a weighted L1 penalty:

$$\sum_{j=1}^{p} |\beta_j|^{1/2} \approx \sum_{j=1}^{p} w_j |\beta_j|, \qquad w_j = \frac{1}{2\sqrt{|\beta_j| + \varepsilon}}.$$

- Small $|\beta_j| \Rightarrow$ large w_j.
- Large w_j penalizes that coefficient more $\Rightarrow$ it tends to vanish.

Result: iteratively solving weighted L1 problems approximates L1/2.

Pros / cons summary

Advantages

- Very strong sparsity (closer to L0 than L1).
- Potentially less bias on large coefficients than L1.

Limitations

- Non-convex $\Rightarrow$ harder optimization; local minima possible.
- Sensitive to initialization and solver choice.

One-line takeaway: L1/2 is a compromise — much sparser than L1, but not as combinatorial as L0.

Comparison: L0 vs L1/2 vs L1

	L0	**L1/2**	**L1 (Lasso)**
Norm	$\|\theta\|_0 = \|\{i : \theta_i \neq 0\}\|$	$\|\theta\|_{1/2} = \sum_i \|\theta_i\|^{1/2}$	$\|\theta\|_1 = \sum_i \|\theta_i\|$
Sparsity	Very high (exact)	High (aggressive)	Medium/High
Convex?	No (discrete)	No ($p < 1$)	Yes
Optimization	NP-hard (combinatorial)	Non-convex (ITA, MM)	Convex (proximal)
Bias	None (if found)	Low on large θ	Linear with $\|\theta\|$
Stability	Very unstable	Sensitive to init.	Stable
Complexity	$O(2^p)$ exact	$O(n \cdot iter)$	$O(n \cdot iter)$
Use-case	Strict subset	Strong sparsity needs	General baseline

When to use what?

Use L1 (Lasso)

- Convex + reliable
- Fast solvers (LARS, proximal)
- Theory well-established
- Good baseline
- × Bias on large coefs
- × Moderate sparsity

Use L1/2

- Stronger sparsity than L1
- Less biased estimates
- Better signal recovery
- × Non-convex (local minima)
- × Needs good init.
- × Careful tuning

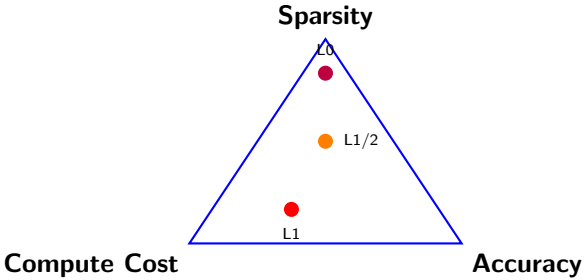
Use L0

- Optimal subset selection
- Maximal interpretability
- Unbiased
- × NP-hard
- × Only small p ($<$ 30-40)
- × Greedy approx needed

Conclusion

- **L0:** Direct sparsity control, interpretable, but computationally intractable for large p. *Best for: feature selection with $p < 40$*
- **L1/2:** Bridges L0 and L1; achieves stronger sparsity with less bias than Lasso. *Best for: compressed sensing, medical imaging, signal reconstruction*
- **L1 (Lasso):** Industry standard—convex, stable, fast. *Best for: most real-world applications, production systems*

- **Key insight:** As p decreases from 1 to 0, sparsity increases but optimization difficulty grows exponentially.
- Always balance the **trade-off triangle**:



References

Foundational papers:

- Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso*. JRSS-B.
- Xu, Z. et al. (2012). *$L_{1/2}$ regularization: A thresholding representation theory*. IEEE Trans. Neural Networks.
- Bertsimas, D. et al. (2016). *Best subset selection via a modern optimization lens*. Annals of Statistics.

Books & surveys:

- Hastie, T. et al. (2015). *Statistical Learning with Sparsity* (Ch. 2-3).
- Mallat, S. (2008). *A Wavelet Tour of Signal Processing* (Sparse representations).

Software:

- `scikit-learn` (Lasso, ElasticNet), `L0Learn` (R package), `CVXPY` (convex optimization)