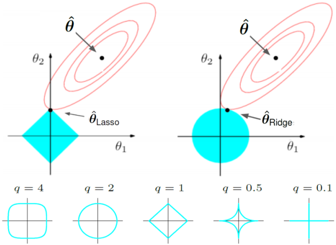
Introduction to Machine Learning

L0 Regularization



Learning goals

- XXX
- XXX

LQ NORM REGULARIZATION

Besides L_1 and L_2 norm we could use any L_q norm for regularization.

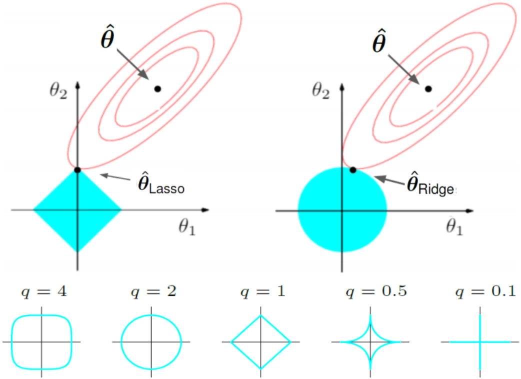


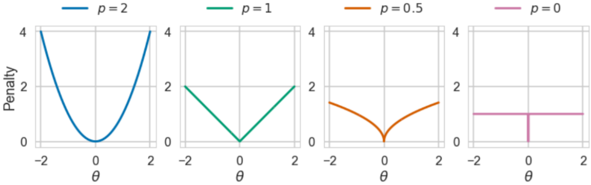
Figure: *Top:* Ridge and Lasso loss contours and feasible regions. *Bottom:* Different feasible region shapes for L_q norms $\sum_j |\theta_j|^q$.

L0 REGULARIZATION

- Consider the L_0-regularized risk of a model $f(\mathbf{x} \mid \boldsymbol{\theta})$

$$\mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_0 := \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda \sum_j |\theta_j|^0.$$

- Unlike the L_1 and L_2 norms, the L_0 "norm" simply counts the number of non-zero parameters in the model.



Credit: Christos Louizos

Figure: L_p norm penalties for a parameter $\boldsymbol{\theta}$ according to different values of p.

L0 REGULARIZATION

- For any parameter $\boldsymbol{\theta}$, the L_0 penalty is zero for $\boldsymbol{\theta} = 0$ (defining $0^0 := 0$) and is constant for any $\boldsymbol{\theta} \neq 0$, no matter how large or small it is.
- L_0 regularization induces sparsity in the parameter vector more aggressively than L_1 regularization, but does not shrink concrete parameter values as L1 and L2 does.
- Model selection criteria such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are special cases of L_0 regularization (corresponding to specific values of λ).
- The L_0-regularized risk is neither continuous, differentiable or convex.
- It is computationally hard to optimize (NP-hard) and likely intractable. For smaller n and p we might be able to solve this nowadays directly, for larger scenarios efficient approximations of the L_0 are still topic of current research.