Failures and Successes of Cross-Validation for Early-Stopped Gradient Descent

Pratik Patil [1]   Yuchen Wu [2]   Ryan J. Tibshirani[1]

[1]Department of Statistics, University of California, Berkeley

[2]Department of Statistics and Data Science, Wharton School, University of Pennsylvania

Summary

- We study **LOOCV** and **GCV** for iterative algorithms in linear models.
- GCV is generically **inconsistent** for the prediction risk
- LOOCV is **uniformly consistent** along the algorithm trajectory
- As application, we construct **pathwise prediction** intervals that have asymptotically correct coverage conditional on the training data

Regularization techniques

Explicit regularization

- L_2 Regularization (Ridge)

$$\min_{\beta\in\mathbb{R}^d} \frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

- L_1 Regularization (Lasso)

$$\min_{\beta\in\mathbb{R}^d} \frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

- Elastic Net Regularization

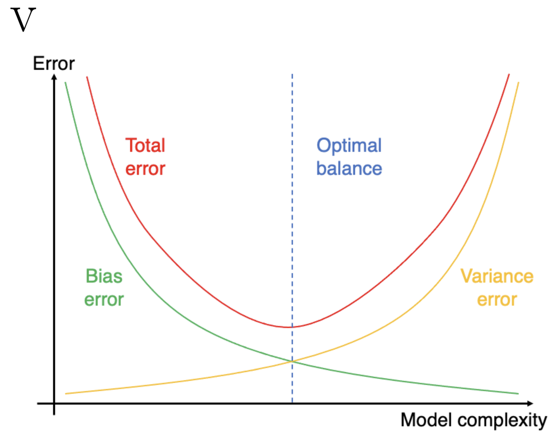
$$\min_{\beta\in\mathbb{R}^d} \frac{1}{n}\|y - X\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2$$

Implicit regularization

- Early stopping
- Gradient descent & stochastic gradient descent

Bias-variance tradeoff

V



Question: How to select the optimal amount of regularization?

- Ridge regularization: selecting the regularization parameter λ
- Gradient descent: determining whether and when to early stop the process
- Close connection between ℓ^2 regularization and gradient descent

Cross validation (CV)

- Split-sample CV, K-fold CV with a small K (such as 5 or 10)
  Might suffer from significant bias
- Leave-one-out CV (LOOCV)
  Mitigates bias issues, computationally expensive
- Generalized CV (GCV)
  Approximation to LOOCV for estimators that are linear smoothers
- LOOCV and GCV are consistent for the high-dimensional ridge regression ($p \asymp n$)
- Are LOOCV and GCV consistent for GD?

LOOCV consistency

- $\hat{\beta}_{k,i}$: GD with k iterations trained on (X_{-i}, y_{-i})

$$\hat{R}^{\mathrm{loo}}(\hat{\beta}_k) = n^{-1}\Sigma_{i=1}^n (y_i - x_i^\intercal \hat{\beta}_{k,-i})^2$$

- **(Main theorem)** Under our assumptions, LOOCV is uniformly consistent:

$$\max_{k\in[K]} \left|\hat{R}^{\mathrm{loo}}(\hat{\beta}_k) - R(\hat{\beta}_k)\right| \xrightarrow{\text{a.s.}} 0$$

- Application: use LOOCV to tune early stopping:

$$k_* = \arg\min_{k\in[K]} \hat{R}^{\mathrm{loo}}(\hat{\beta}_k),$$

$$|R(\hat{\beta}_{k_*}) - \min_{k\in[K]} R(\hat{\beta}_k)| \xrightarrow{\text{a.s.}} 0$$

Definition (T_2-inequality)

We say a distribution μ satisfies the T_2-inequality if there exists a constant $\sigma(\mu) \geq 0$, such that for every distribution ν,

$$W_2(\mu, \nu) \leq \sqrt{2\sigma^2(\mu) D_{\mathrm{KL}}(\nu \parallel \mu)}$$

High-dim least squares regression

- Data $\{(x_i, y_i)\}_{i\leq n} \subseteq \mathbb{R}^p \times \mathbb{R}$, $\mathbf{p} \asymp \mathbf{n}$
- OLS problem: $\text{minimize}_{\beta\in\mathbb{R}^p} \ \frac{1}{2n}\|y - X\beta\|_2^2$
- Solve with GD: $\hat{\beta}_k = \hat{\beta}_{k-1} + \frac{\delta_{k-1}}{n} X^\intercal (y - X\hat{\beta}_{k-1})$
- Out-of-sample prediction risk:

$$R(\hat{\beta}_k) = \mathbb{E}_{x_0,y_0}[(y_0 - x_0^\intercal \hat{\beta}_k)^2 \mid X, y]$$

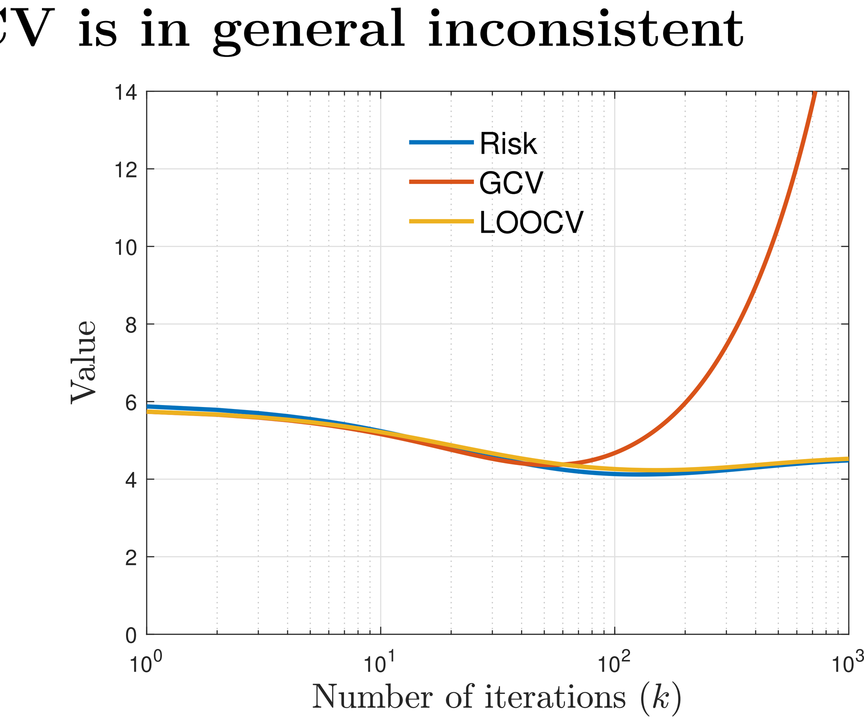
- How well do LOOCV and GCV estimate $R(\hat{\beta}_k)$?

Assumptions

- $x_i = \Sigma^{1/2} z_i$, $z_{ij} \sim_{i.i.d.} \mu_z$, μ_z has mean 0, variance 1, and satisfies the T_2-inequality
- $0 < \zeta_L \leq p/n \leq \zeta_U < \infty$, $\|\Sigma\|_{\mathsf{op}} \leq \sigma_\Sigma$
- $y_i = f(x_i, \varepsilon_i)$, f is L_f-Lipschitz, $\mathbb{E}[y_1^8] \leq m_8$
- $\varepsilon_i \sim_{i.i.d.} \mu_\varepsilon$, μ_ε has mean zero and satisfies the T_2-inequality
- $\Sigma_{k=1}^K \delta_{k-1} \leq \Delta$, $K = o(n(\log n)^{-3/2})$
- $\|\hat{\beta}_0\|_2 \leq B_0$

GCV inconsistency

- LOOCV is consistent, while in most cases computationally expensive
- For predictors that are linear smoothers, can use GCV to approximate LOOCV [Golub et al., 1979]
- GCV is consistent for high-dim ridge regression with mild data assumptions [Patil et al., 2021, 2022]
- *Question: Is GCV also consistent for gradient descent?*
- **GCV is in general inconsistent**



LOOCV shortcut

- Computation is an issue for LOOCV
- We propose a shortcut implementation of LOOCV that has complexity $O(n^3 + nK^2)$ (recall $p \asymp n$)
- When $K \lesssim n$, complexity of the shortcut implementation is at most the same as that for GCV ($O(n^3)$)

Discussion and future directions

Summary

- LOOCV is uniformly consistent along the GD path under mild assumptions
- GCV is inconsistent in even standard examples
- Propose shortcut formula to reduce computational cost

Future directions

- Extension to general iterative algorithms
- Universality result without the T_2 assumption?
- Develop approximate LOOCV approach

References

[1] Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

[2] Pratik Patil, Yuting Wei, Alessandro Rinaldo, and Ryan Tibshirani. Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pages 3178–3186. PMLR, 2021.

[3] Pratik Patil, Alessandro Rinaldo, and Ryan J. Tibshirani. Estimating functionals of the out-of-sample error distribution in high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, 2022.