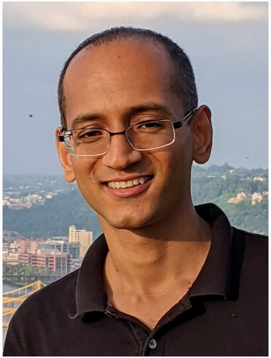
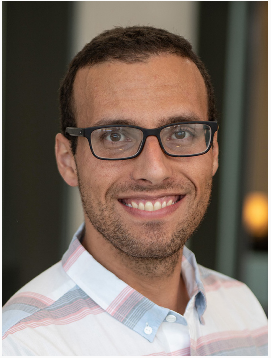
Failures and Successes of Cross-Validation for Early-Stopped Gradient Descent

Yuchen Wu

3rd May 2024

Department of Statistics and Data Science
The Wharton School, University of Pennsylvania





Pratik Patil
Berkeley Stats

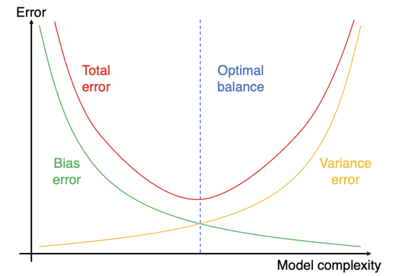
Ryan Tibshirani
Berkeley Stats

Bias-variance tradeoff

Implicit regularization: regularization effect induced by the optimization algorithm

Close connection between ℓ^2 regularization and gradient descent

[Suggala et al., 2018], [Neu and Rosasco, 2018], [Ali et al., 2019], [Ali et al., 2020]



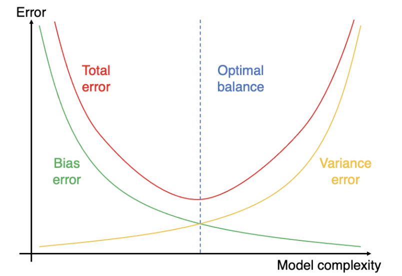
- Ridge regularization: selecting the regularization parameter λ
- Gradient descent: determining whether and when to early stop the GD iteration

Bias-variance tradeoff

Implicit regularization: regularization effect induced by the optimization algorithm

Close connection between ℓ^2 regularization and gradient descent

[Suggala et al., 2018], [Neu and Rosasco, 2018], [Ali et al., 2019], [Ali et al., 2020]



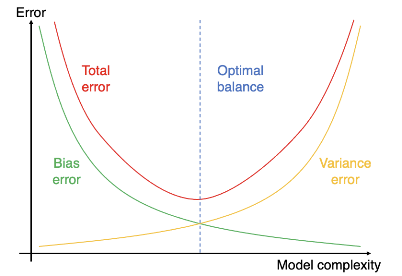
- Ridge regularization: selecting the regularization parameter λ
- Gradient descent: determining whether and when to early stop the GD iteration

Bias-variance tradeoff

Implicit regularization: regularization effect induced by the optimization algorithm

Close connection between ℓ^2 regularization and gradient descent

[Suggala et al., 2018], [Neu and Rosasco, 2018], [Ali et al., 2019], [Ali et al., 2020]



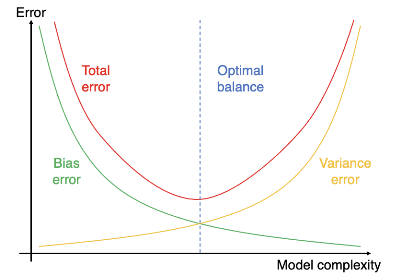
- Ridge regularization: selecting the regularization parameter λ
- Gradient descent: determining whether and when to early stop the GD iteration

Bias-variance tradeoff

Implicit regularization: regularization effect induced by the optimization algorithm

Close connection between ℓ^2 regularization and gradient descent

[Suggala et al., 2018], [Neu and Rosasco, 2018], [Ali et al., 2019], [Ali et al., 2020]



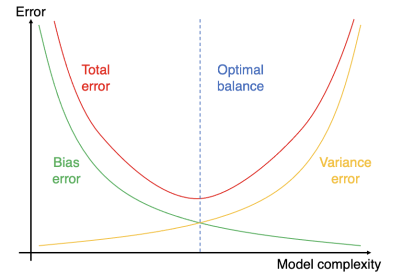
- Ridge regularization: selecting the regularization parameter λ
- Gradient descent: determining whether and when to early stop the GD iteration

Bias-variance tradeoff

Implicit regularization: regularization effect induced by the optimization algorithm

Close connection between ℓ^2 regularization and gradient descent

[Suggala et al., 2018], [Neu and Rosasco, 2018], [Ali et al., 2019], [Ali et al., 2020]



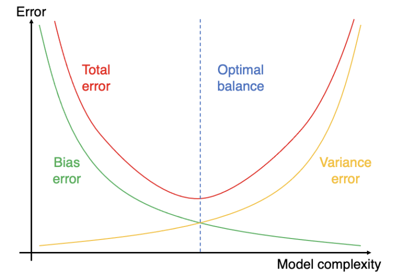
- Ridge regularization: selecting the regularization parameter λ
- Gradient descent: determining whether and when to early stop the GD iteration

Bias-variance tradeoff

Implicit regularization: regularization effect induced by the optimization algorithm

Close connection between ℓ^2 regularization and gradient descent

[Suggala et al., 2018], [Neu and Rosasco, 2018], [Ali et al., 2019], [Ali et al., 2020]



- Ridge regularization: selecting the regularization parameter λ
- Gradient descent: determining whether and when to early stop the GD iteration

Selecting the optimal amount of regularization – cross validation

- Split cross-validation, K-fold cross-validation with a small K (such as 5 or 10)
  Might suffer from significant bias [Rad and Maleki, 2020], [Rad et al., 2020]
- Leave-one-out cross-validation (LOOCV, $K = n$)
  Mitigates bias issues, computationally expensive to implement
- Generalized cross-validation (GCV)
  Approximation to LOOCV for estimators that are linear smoothers
- LOOCV and GCV are consistent for the prediction risk of ridge regression in high-dimensional settings ($p \asymp n$) [Patil et al., 2021]
- Is LOOCV and GCV consistent for GD, in the context of high-dimensional regression?

Selecting the optimal amount of regularization – cross validation

- Split cross-validation, K-fold cross-validation with a small K (such as 5 or 10)
  Might suffer from significant bias [Rad and Maleki, 2020], [Rad et al., 2020]
- Leave-one-out cross-validation (LOOCV, $K = n$)
  Mitigates bias issues, computationally expensive to implement
- Generalized cross-validation (GCV)
  Approximation to LOOCV for estimators that are linear smoothers
- LOOCV and GCV are consistent for the prediction risk of ridge regression in high-dimensional settings ($p \asymp n$) [Patil et al., 2021]
- Is LOOCV and GCV consistent for GD, in the context of high-dimensional regression?

Selecting the optimal amount of regularization – cross validation

- Split cross-validation, K-fold cross-validation with a small K (such as 5 or 10)
  Might suffer from significant bias [Rad and Maleki, 2020], [Rad et al., 2020]
- Leave-one-out cross-validation (LOOCV, $K = n$)
  Mitigates bias issues, computationally expensive to implement
- Generalized cross-validation (GCV)
  Approximation to LOOCV for estimators that are linear smoothers
- LOOCV and GCV are consistent for the prediction risk of ridge regression in high-dimensional settings ($p \asymp n$) [Patil et al., 2021]
- Is LOOCV and GCV consistent for GD, in the context of high-dimensional regression?

Selecting the optimal amount of regularization – cross validation

- Split cross-validation, K-fold cross-validation with a small K (such as 5 or 10)
  Might suffer from significant bias [Rad and Maleki, 2020], [Rad et al., 2020]
- Leave-one-out cross-validation (LOOCV, $K = n$)
  Mitigates bias issues, computationally expensive to implement
- Generalized cross-validation (GCV)
  Approximation to LOOCV for estimators that are linear smoothers
- LOOCV and GCV are consistent for the prediction risk of ridge regression in high-dimensional settings ($p \asymp n$) [Patil et al., 2021]
- Is LOOCV and GCV consistent for GD, in the context of high-dimensional regression?

Selecting the optimal amount of regularization – cross validation

- Split cross-validation, K-fold cross-validation with a small K (such as 5 or 10)
  Might suffer from significant bias [Rad and Maleki, 2020], [Rad et al., 2020]
- Leave-one-out cross-validation (LOOCV, $K = n$)
  Mitigates bias issues, computationally expensive to implement
- Generalized cross-validation (GCV)
  Approximation to LOOCV for estimators that are linear smoothers
- LOOCV and GCV are consistent for the prediction risk of ridge regression in high-dimensional settings ($p \asymp n$) [Patil et al., 2021]
- Is LOOCV and GCV consistent for GD, in the context of high-dimensional regression?

Selecting the optimal amount of regularization – cross validation

- Split cross-validation, K-fold cross-validation with a small K (such as 5 or 10)
  Might suffer from significant bias [Rad and Maleki, 2020], [Rad et al., 2020]
- Leave-one-out cross-validation (LOOCV, $K = n$)
  Mitigates bias issues, computationally expensive to implement
- Generalized cross-validation (GCV)
  Approximation to LOOCV for estimators that are linear smoothers
- LOOCV and GCV are consistent for the prediction risk of ridge regression in high-dimensional settings ($p \asymp n$) [Patil et al., 2021]
- Is LOOCV and GCV consistent for GD, in the context of high-dimensional regression?

Selecting the optimal amount of regularization – cross validation

- Split cross-validation, K-fold cross-validation with a small K (such as 5 or 10)
  Might suffer from significant bias [Rad and Maleki, 2020], [Rad et al., 2020]
- Leave-one-out cross-validation (LOOCV, $K = n$)
  Mitigates bias issues, computationally expensive to implement
- Generalized cross-validation (GCV)
  Approximation to LOOCV for estimators that are linear smoothers
- LOOCV and GCV are consistent for the prediction risk of ridge regression in high-dimensional settings ($p \asymp n$) [Patil et al., 2021]
- Is LOOCV and GCV consistent for GD, in the context of high-dimensional regression?

Selecting the optimal amount of regularization – cross validation

- Split cross-validation, K-fold cross-validation with a small K (such as 5 or 10)
  Might suffer from significant bias [Rad and Maleki, 2020], [Rad et al., 2020]
- Leave-one-out cross-validation (LOOCV, $K = n$)
  Mitigates bias issues, computationally expensive to implement
- Generalized cross-validation (GCV)
  Approximation to LOOCV for estimators that are linear smoothers
- LOOCV and GCV are consistent for the prediction risk of ridge regression in high-dimensional settings ($p \asymp n$) [Patil et al., 2021]
- Is LOOCV and GCV consistent for GD, in the context of high-dimensional regression?

High-dimensional least squares regression

- Consider i.i.d. data $\{(x_i, y_i)\}_{i \le n} \subseteq \mathbb{R}^p \times \mathbb{R}$, $p \asymp n$
- The ordinary least squares problem:

$$\text{minimize}_{\beta \in \mathbb{R}^p} \ \frac{1}{2n} \|y - X\beta\|_2^2$$

- Solve with gradient descent:

$$\hat{\beta}_k = \hat{\beta}_{k-1} + \frac{\delta_{k-1}}{n} X^\top (y - X\hat{\beta}_{k-1}), \qquad k = 1, 2, \cdots, K$$

  K steps, step size δ_k
- Wish to estimate the out-of-sample prediction risk:

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

  (x_0, y_0) is a test data point. Expectation is taken over (x_0, y_0)
- Can we use LOOCV and GCV to tune early-stopped gradient descent?

High-dimensional least squares regression

- Consider i.i.d. data $\{(x_i, y_i)\}_{i \le n} \subseteq \mathbb{R}^p \times \mathbb{R}$, $p \asymp n$
- The ordinary least squares problem:

$$\text{minimize}_{\beta \in \mathbb{R}^p} \;\; \frac{1}{2n}\|y - X\beta\|_2^2$$

- Solve with gradient descent:

$$\hat{\beta}_k = \hat{\beta}_{k-1} + \frac{\delta_{k-1}}{n} X^\top (y - X\hat{\beta}_{k-1}), \qquad k = 1, 2, \cdots, K$$

  K steps, step size δ_k
- Wish to estimate the out-of-sample prediction risk:

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

  (x_0, y_0) is a test data point. Expectation is taken over (x_0, y_0)
- Can we use LOOCV and GCV to tune early-stopped gradient descent?

High-dimensional least squares regression

- Consider i.i.d. data $\{(x_i, y_i)\}_{i \le n} \subseteq \mathbb{R}^p \times \mathbb{R}$, $p \asymp n$
- The ordinary least squares problem:

$$\text{minimize}_{\beta \in \mathbb{R}^p} \ \frac{1}{2n} \|y - X\beta\|_2^2$$

- Solve with gradient descent:

$$\hat{\beta}_k = \hat{\beta}_{k-1} + \frac{\delta_{k-1}}{n} X^\top (y - X\hat{\beta}_{k-1}), \qquad k = 1, 2, \cdots, K$$

  K steps, step size δ_k

- Wish to estimate the out-of-sample prediction risk:

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

  (x_0, y_0) is a test data point. Expectation is taken over (x_0, y_0)

- Can we use LOOCV and GCV to tune early-stopped gradient descent?

High-dimensional least squares regression

- Consider i.i.d. data $\{(x_i, y_i)\}_{i \le n} \subseteq \mathbb{R}^p \times \mathbb{R}$, $p \asymp n$
- The ordinary least squares problem:

$$\text{minimize}_{\beta \in \mathbb{R}^p} \ \frac{1}{2n} \|y - X\beta\|_2^2$$

- Solve with gradient descent:

$$\hat{\beta}_k = \hat{\beta}_{k-1} + \frac{\delta_{k-1}}{n} X^\mathsf{T}(y - X\hat{\beta}_{k-1}), \qquad k = 1, 2, \cdots, K$$

  K steps, step size δ_k
- Wish to estimate the out-of-sample prediction risk:

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

  (x_0, y_0) is a test data point. Expectation is taken over (x_0, y_0)
- Can we use LOOCV and GCV to tune early-stopped gradient descent?

High-dimensional least squares regression

- Consider i.i.d. data $\{(x_i, y_i)\}_{i \le n} \subseteq \mathbb{R}^p \times \mathbb{R}$, $p \asymp n$
- The ordinary least squares problem:

$$\text{minimize}_{\beta \in \mathbb{R}^p} \quad \frac{1}{2n} \|y - X\beta\|_2^2$$

- Solve with gradient descent:

$$\hat{\beta}_k = \hat{\beta}_{k-1} + \frac{\delta_{k-1}}{n} X^\intercal (y - X\hat{\beta}_{k-1}), \qquad k = 1, 2, \cdots, K$$

  K steps, step size δ_k
- Wish to estimate the out-of-sample prediction risk:

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

  (x_0, y_0) is a test data point. Expectation is taken over (x_0, y_0)
- Can we use LOOCV and GCV to tune early-stopped gradient descent?

High-dimensional least squares regression

- Consider i.i.d. data $\{(x_i, y_i)\}_{i \le n} \subseteq \mathbb{R}^p \times \mathbb{R}$, $p \asymp n$
- The ordinary least squares problem:

$$\text{minimize}_{\beta \in \mathbb{R}^p} \;\; \frac{1}{2n} \|y - X\beta\|_2^2$$

- Solve with gradient descent:

$$\hat{\beta}_k = \hat{\beta}_{k-1} + \frac{\delta_{k-1}}{n} X^\intercal (y - X\hat{\beta}_{k-1}), \qquad k = 1, 2, \cdots, K$$

  K steps, step size δ_k
- Wish to estimate the out-of-sample prediction risk:

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

  (x_0, y_0) is a test data point. Expectation is taken over (x_0, y_0)
- Can we use LOOCV and GCV to tune early-stopped gradient descent?

High-dimensional least squares regression

- Consider i.i.d. data $\{(x_i, y_i)\}_{i\le n} \subseteq \mathbb{R}^p \times \mathbb{R}$, $p \asymp n$
- The ordinary least squares problem:

$$\text{minimize}_{\beta\in\mathbb{R}^p} \;\; \frac{1}{2n}\|y - X\beta\|_2^2$$

- Solve with gradient descent:

$$\hat\beta_k = \hat\beta_{k-1} + \frac{\delta_{k-1}}{n} X^\intercal (y - X\hat\beta_{k-1}), \qquad k = 1, 2, \cdots, K$$

  K steps, step size δ_k
- Want to estimate the out-of-sample prediction risk:

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

  (x_0, y_0) is a test data point. Expectation is taken over (x_0, y_0)
- How well do LOOCV and GCV estimate $R(\hat\beta_k)$?

LOOCV

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

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

- Under certain conditions,

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

- Application: use LOOCV to tune early-stopped gradient descent:

$$k_* = \arg\min_{k\in[K]} \hat{R}^{\text{loo}}(\hat{\beta}_k), \qquad |R(\hat{\beta}_{k_*}) - \min_{k\in[K]} R(\hat{\beta}_k)| \overset{\text{a.s.}}{\to} 0$$

LOOCV

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

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

- Under certain conditions,

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

- Application: use LOOCV to tune early-stopped gradient descent:

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

LOOCV

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

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

- Under certain conditions,

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

- Application: use LOOCV to tune early-stopped gradient descent:

$$k_* = \arg\min_{k\in[K]} \hat{R}^{\mathrm{loo}}(\hat{\beta}_k), \qquad |R(\hat{\beta}_{k_*}) - \min_{k\in[K]} R(\hat{\beta}_k)| \xrightarrow{\text{a.s.}} 0$$

LOOCV

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

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

- Under certain conditions,

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

- Application: use LOOCV to tune early-stopped gradient descent:

$$k_* = \arg\min_{k\in[K]} \hat{R}^{\mathrm{loo}}(\hat{\beta}_k), \qquad |R(\hat{\beta}_{k_*}) - \min_{k\in[K]} R(\hat{\beta}_k)| \xrightarrow{\text{a.s.}} 0$$

Assumptions

- Feature vector decomposition: $x_i = \Sigma^{1/2} z_i$, $z_{ij} \sim_{i.i.d.} \mu_z$, $\|\Sigma\|_{\mathsf{op}} \leq \sigma_\Sigma$
- $y_i = f(x_i, \varepsilon_i)$, f is L_f-Lipschitz, $\mathbb{E}[y_1^8] \leq m_8$, $\varepsilon_i \sim_{i.i.d.} \mu_\varepsilon$
- μ_z, μ_ε satisfy the T_2-inequality
- $0 < \zeta_L \leq p/n \leq \zeta_U < \infty$
- $K = o(n(\log n)^{-3/2})$
- Initialization is bounded: $\|\hat{\beta}_0\|_2 \leq B_0$

Assumptions

- Feature vector decomposition: $x_i = \Sigma^{1/2} z_i$, $z_{ij} \sim_{i.i.d.} \mu_z$, $\|\Sigma\|_{\mathsf{op}} \le \sigma_\Sigma$
- $y_i = f(x_i, \varepsilon_i)$, f is L_f-Lipschitz, $\mathbb{E}[y_1^8] \le m_8$, $\varepsilon_i \sim_{i.i.d.} \mu_\varepsilon$
- μ_z, μ_ε satisfy the T_2-inequality
- $0 < \zeta_L \le p/n \le \zeta_U < \infty$
- $K = o(n(\log n)^{-3/2})$
- Initialization is bounded: $\|\hat{\beta}_0\|_2 \le B_0$

Assumptions

- Feature vector decomposition: $x_i = \Sigma^{1/2} z_i$, $z_{ij} \sim_{i.i.d.} \mu_z$, $\|\Sigma\|_{\mathsf{op}} \leq \sigma_\Sigma$
- $y_i = f(x_i, \varepsilon_i)$, f is L_f-Lipschitz, $\mathbb{E}[y_1^8] \leq m_8$, $\varepsilon_i \sim_{i.i.d.} \mu_\varepsilon$
- μ_z, μ_ε satisfy the T_2-inequality
- $0 < \zeta_L \leq p/n \leq \zeta_U < \infty$
- $K = o(n(\log n)^{-3/2})$
- Initialization is bounded: $\|\hat{\beta}_0\|_2 \leq B_0$

Assumptions

- Feature vector decomposition: $x_i = \Sigma^{1/2} z_i$, $z_{ij} \sim_{i.i.d.} \mu_z$, $\|\Sigma\|_{\mathsf{op}} \leq \sigma_\Sigma$
- $y_i = f(x_i, \varepsilon_i)$, f is L_f-Lipschitz, $\mathbb{E}[y_1^8] \leq m_8$, $\varepsilon_i \sim_{i.i.d.} \mu_\varepsilon$
- μ_z, μ_ε satisfy the T_2-inequality
- $0 < \zeta_L \leq p/n \leq \zeta_U < \infty$
- $K = o(n(\log n)^{-3/2})$
- Initialization is bounded: $\|\hat{\beta}_0\|_2 \leq B_0$

Assumptions

- Feature vector decomposition: $x_i = \Sigma^{1/2} z_i$, $z_{ij} \sim_{i.i.d.} \mu_z$, $\|\Sigma\|_{\mathsf{op}} \leq \sigma_\Sigma$
- $y_i = f(x_i, \varepsilon_i)$, f is L_f-Lipschitz, $\mathbb{E}[y_1^8] \leq m_8$, $\varepsilon_i \sim_{i.i.d.} \mu_\varepsilon$
- μ_z, μ_ε satisfy the T_2-inequality
- $0 < \zeta_L \leq p/n \leq \zeta_U < \infty$
- $K = o(n(\log n)^{-3/2})$
- Initialization is bounded: $\|\hat{\beta}_0\|_2 \leq B_0$

Assumptions

- Feature vector decomposition: $x_i = \Sigma^{1/2} z_i$, $z_{ij} \sim_{i.i.d.} \mu_z$, $\|\Sigma\|_{\mathsf{op}} \leq \sigma_\Sigma$
- $y_i = f(x_i, \varepsilon_i)$, f is L_f-Lipschitz, $\mathbb{E}[y_1^8] \leq m_8$, $\varepsilon_i \sim_{i.i.d.} \mu_\varepsilon$
- μ_z, μ_ε satisfy the T_2-inequality
- $0 < \zeta_L \leq p/n \leq \zeta_U < \infty$
- $K = o(n(\log n)^{-3/2})$
- Initialization is bounded: $\|\hat{\beta}_0\|_2 \leq B_0$

Assumptions

- Feature vector decomposition: $x_i = \Sigma^{1/2} z_i$, $z_{ij} \sim_{i.i.d.} \mu_z$, $\|\Sigma\|_{\mathsf{op}} \leq \sigma_\Sigma$
- $y_i = f(x_i, \varepsilon_i)$, f is L_f-Lipschitz, $\mathbb{E}[y_1^8] \leq m_8$, $\varepsilon_i \sim_{i.i.d.} \mu_\varepsilon$
- μ_z, μ_ε satisfy the T_2-inequality
- $0 < \zeta_L \leq p/n \leq \zeta_U < \infty$
- $K = o(n(\log n)^{-3/2})$
- Initialization is bounded: $\|\hat{\beta}_0\|_2 \leq B_0$

T_2-inequality

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,

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

Examples of distributions that satisfy T_2:

1. Distributions that satisfy log Sobolev inequality
2. Log-concave distributions
3. Gaussian convolutions of distributions with bounded support

T_2-inequality

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,

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

Examples of distributions that satisfy T_2:

1. Distributions that satisfy log Sobolev inequality
2. Log-concave distributions
3. Gaussian convolutions of distributions with bounded support

T_2-inequality

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,

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

Examples of distributions that satisfy T_2:

1. Distributions that satisfy log Sobolev inequality
2. Log-concave distributions
3. Gaussian convolutions of distributions with bounded support

T_2-inequality

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,

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

Examples of distributions that satisfy T_2:

1. Distributions that satisfy log Sobolev inequality
2. Log-concave distributions
3. Gaussian convolutions of distributions with bounded support

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,

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

Examples of distributions that satisfy T_2:

1. Distributions that satisfy log Sobolev inequality
2. Log-concave distributions
3. Gaussian convolutions of distributions with bounded support

T_2-inequality

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,

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

Examples of distributions that satisfy T_2:

1. Distributions that satisfy log Sobolev inequality
2. Log-concave distributions
3. Gaussian convolutions of distributions with bounded support

Dimension-free concentration

Lemma, (Van Handel, 2014)

Let μ be a probability measure, and $X_i \sim_{i.i.d.} \mu$. Then the following are equivalent:

1. μ satisfies T_2-inequality with constant σ
2. For every 1-Lipschitz function g,

$$\mathbb{P}\left(|g(X_1, \cdots, X_N) - \mathbb{E}[g(X_1, \cdots, X_N)]| \geq t\right) \leq C_0 e^{-t^2/2\sigma^2}$$

LOOCV consistency

Theorem

Assume all the aforementioned assumptions, then as $n, p \to \infty$,

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

LOOCV consistency

Theorem

Assume all the aforementioned assumptions, then as $n, p \to \infty$,

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

LOOCV consistency

Theorem

Assume all the aforementioned assumptions, also assume L is pseudo-Lipschitz, then as $n, p \to \infty$,

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

Generalized cross-validation (GCV)

- LOOCV is consistent, while in most cases computationally expensive
- For predictors that are linear smoothers, we can use GCV to approximate LOOCV [Golub et al., 1979]
- Suppose we have a predictor $\hat{f}$ that is a linear smoother: $\hat{f}(x) = s_x^\top y$, $s_x \in \mathbb{R}^n$ is a function of the training data X and the test point x
- GCV estimate of the prediction risk:

$$\hat{R}^{\text{gcv}}(\hat{f}) = \frac{\|y - Sy\|_2^2}{(1 - \mathrm{tr}[S]/n)^2}$$

  $S \in \mathbb{R}^{n \times n}$ has rows $s_{x_1}^\top, \cdots, s_{x_n}^\top$
- GD and ridge are linear smoothers

Generalized cross-validation (GCV)

- LOOCV is consistent, while in most cases computationally expensive
- For predictors that are linear smoothers, we can use GCV to approximate LOOCV [Golub et al., 1979]
- Suppose we have a predictor $\hat{f}$ that is a linear smoother: $\hat{f}(x) = s_x^\top y$, $s_x \in \mathbb{R}^n$ is a function of the training data X and the test point x
- GCV estimate of the prediction risk:

$$\hat{R}^{\text{gcv}}(\hat{f}) = \frac{\|y - Sy\|_2^2}{(1 - \text{tr}[S]/n)^2}$$

  $S \in \mathbb{R}^{n\times n}$ has rows $s_{x_1}^\top, \cdots, s_{x_n}^\top$
- GD and ridge are linear smoothers

Generalized cross-validation (GCV)

- LOOCV is consistent, while in most cases computationally expensive
- For predictors that are linear smoothers, we can use GCV to approximate LOOCV [Golub et al., 1979]
- Suppose we have a predictor $\hat{f}$ that is a linear smoother: $\hat{f}(x) = s_x^\top y$, $s_x \in \mathbb{R}^n$ is a function of the training data X and the test point x
- GCV estimate of the prediction risk:

$$\hat{R}^{\text{gcv}}(\hat{f}) = \frac{\|y - Sy\|_2^2}{(1 - \text{tr}[S]/n)^2}$$

  $S \in \mathbb{R}^{n \times n}$ has rows $s_{x_1}^\top, \cdots, s_{x_n}^\top$
- GD and ridge are linear smoothers

Generalized cross-validation (GCV)

- LOOCV is consistent, while in most cases computationally expensive
- For predictors that are linear smoothers, we can use GCV to approximate LOOCV [Golub et al., 1979]
- Suppose we have a predictor $\hat{f}$ that is a linear smoother: $\hat{f}(x) = s_x^\mathsf{T} y$, $s_x \in \mathbb{R}^n$ is a function of the training data X and the test point x
- GCV estimate of the prediction risk:

$$\hat{R}^{\text{gcv}}(\hat{f}) = \frac{\|y - Sy\|_2^2}{(1 - \text{tr}[S]/n)^2}$$

  $S \in \mathbb{R}^{n \times n}$ has rows $s_{x_1}^\mathsf{T}, \cdots, s_{x_n}^\mathsf{T}$
- GD and ridge are linear smoothers

Generalized cross-validation (GCV)

- LOOCV is consistent, while in most cases computationally expensive
- For predictors that are linear smoothers, we can use GCV to approximate LOOCV [Golub et al., 1979]
- Suppose we have a predictor $\hat{f}$ that is a linear smoother: $\hat{f}(x) = s_x^\intercal y$, $s_x \in \mathbb{R}^n$ is a function of the training data X and the test point x
- GCV estimate of the prediction risk:

$$\hat{R}^{\mathrm{gcv}}(\hat{f}) = \frac{\|y - Sy\|_2^2}{(1 - \mathrm{tr}[S]/n)^2}$$

  $S \in \mathbb{R}^{n \times n}$ has rows $s_{x_1}^\intercal, \cdots, s_{x_n}^\intercal$

- GD and ridge are linear smoothers

Generalized cross-validation (GCV)

- LOOCV is consistent, while in most cases computationally expensive
- For predictors that are linear smoothers, we can use GCV to approximate LOOCV [Golub et al., 1979]
- Suppose we have a predictor $\hat{f}$ that is a linear smoother: $\hat{f}(x) = s_x^\mathsf{T} y$, $s_x \in \mathbb{R}^n$ is a function of the training data X and the test point x
- GCV estimate of the prediction risk:

$$\hat{R}^{\mathrm{gcv}}(\hat{f}) = \frac{\|y - Sy\|_2^2}{(1 - \mathrm{tr}[S]/n)^2}$$

  $S \in \mathbb{R}^{n \times n}$ has rows $s_{x_1}^\mathsf{T}, \cdots, s_{x_n}^\mathsf{T}$
- GD and ridge are linear smoothers

GCV inconsistency

- GCV is consistent for high-dimensional ridge regression under mild data assumptions [Patil et al., 2021]
- *Question: Is GCV also consistent for gradient descent?*
- The answer is no: simple counterexample with Gaussian isotropic features

GCV inconsistency

- GCV is consistent for high-dimensional ridge regression under mild data assumptions [Patil et al., 2021]
- *Question: Is GCV also consistent for gradient descent?*
- The answer is no: simple counterexample with Gaussian isotropic features

GCV inconsistency

- GCV is consistent for high-dimensional ridge regression under mild data assumptions [Patil et al., 2021]
- *Question: Is GCV also consistent for gradient descent?*
- The answer is no: simple counterexample with Gaussian isotropic features

GCV inconsistency

- GCV is consistent for high-dimensional ridge regression under mild data assumptions [Patil et al., 2021]
- *Question: Is GCV also consistent for gradient descent?*
- The answer is no: simple counterexample with Gaussian isotropic features

Discussion and future directions

Summary

- GD for high-dimensional least squares regression
- LOOCV is uniformly consistent along the GD path under mild assumptions
- GCV is inconsistent in even simple examples
- Propose shortcut implementation of LOOCV to reduce computational cost (check our paper)

Future directions

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

Summary

- GD for high-dimensional least squares regression
- LOOCV is uniformly consistent along the GD path under mild assumptions
- GCV is inconsistent in even simple examples
- Propose shortcut implementation of LOOCV to reduce computational cost (check our paper)

Future directions

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

Discussion and future directions

Summary

- GD for high-dimensional least squares regression
- LOOCV is uniformly consistent along the GD path under mild assumptions
- GCV is inconsistent in even simple examples
- Propose shortcut implementation of LOOCV to reduce computational cost (check our paper)

Future directions

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

Discussion and future directions

Summary

- GD for high-dimensional least squares regression
- LOOCV is uniformly consistent along the GD path under mild assumptions
- GCV is inconsistent in even simple examples
- Propose shortcut implementation of LOOCV to reduce computational cost (check our paper)

Future directions

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

Discussion and future directions

Summary

- GD for high-dimensional least squares regression
- LOOCV is uniformly consistent along the GD path under mild assumptions
- GCV is inconsistent in even simple examples
- Propose shortcut implementation of LOOCV to reduce computational cost (check our paper)

Future directions

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

Discussion and future directions

Summary

- GD for high-dimensional least squares regression
- LOOCV is uniformly consistent along the GD path under mild assumptions
- GCV is inconsistent in even simple examples
- Propose shortcut implementation of LOOCV to reduce computational cost (check our paper)

Future directions

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

Discussion and future directions

Summary

- GD for high-dimensional least squares regression
- LOOCV is uniformly consistent along the GD path under mild assumptions
- GCV is inconsistent in even simple examples
- Propose shortcut implementation of LOOCV to reduce computational cost (check our paper)

Future directions

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

Discussion and future directions

Summary

- GD for high-dimensional least squares regression
- LOOCV is uniformly consistent along the GD path under mild assumptions
- GCV is inconsistent in even simple examples
- Propose shortcut implementation of LOOCV to reduce computational cost (check our paper)

Future directions

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

Discussion and future directions

Summary

- GD for high-dimensional least squares regression
- LOOCV is uniformly consistent along the GD path under mild assumptions
- GCV is inconsistent in even simple examples
- Propose shortcut implementation of LOOCV to reduce computational cost (check our paper)

Future directions

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

Discussion and future directions

Summary

- GD for high-dimensional least squares regression
- LOOCV is uniformly consistent along the GD path under mild assumptions
- GCV is inconsistent in even simple examples
- Propose shortcut implementation of LOOCV to reduce computational cost (check our paper)

Future directions

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

Discussion and future directions

Summary

- GD for high-dimensional least squares regression
- LOOCV is uniformly consistent along the GD path under mild assumptions
- GCV is inconsistent in even simple examples
- Propose shortcut implementation of LOOCV to reduce computational cost (check our paper)

Future directions

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

The End

Thank you!

References

- Ali, A., Dobriban, E., & Tibshirani, R. (2020). **The implicit regularization of stochastic gradient flow for least squares.** *International conference on machine learning*, 233–244.
- Ali, A., Kolter, J. Z., & Tibshirani, R. J. (2019). **A continuous-time view of early stopping for least squares regression.** *The 22nd international conference on artificial intelligence and statistics*, 1370–1378.
- Golub, G. H., Heath, M., & Wahba, G. (1979). **Generalized cross-validation as a method for choosing a good ridge parameter.** *Technometrics*, *21*(2), 215–223.
- Neu, G., & Rosasco, L. (2018). **Iterate averaging as regularization for stochastic gradient descent.** *Conference On Learning Theory*, 3222–3242.
- Patil, P., Wei, Y., Rinaldo, A., & Tibshirani, R. (2021). **Uniform consistency of cross-validation estimators for high-dimensional ridge regression.** *International Conference on Artificial Intelligence and Statistics*, 3178–3186.