Module 11

Shrinkage / Model Assessment

Kentaro Nakamura

GOV 2003

April 24th, 2026

Logistics

- Important dates
  - May 1st: Last problem set due
  - May 4th: Review Session (CGIS K354, 3pm-)
  - May 7th: Final Exam (Emerson Hall 104, 9am-)

- Today is the last section during the semester
  - Thank you so much everyone for coming!

Agenda

- Scaling issue for Shrinkage

- Lasso regression
  - Extra: Formal analysis of sparsity under Lasso
  - Optimization (Coordinate decent)
  - Extra: Statistical inference after Variable Selection

- Model Assessment
  - Optimism of training error
  - Mallow's C_p statistics and AIC
  - AIC and likelihood ratio test

Scaling issue for Shrinkage

- Consider the ridge regression

$$\hat{\beta} = \arg\min_{\beta} ||Y_i - \tilde{\boldsymbol{X}}_i^\top \beta||^2 + \lambda||\beta||_2^2$$

- Now, suppose that we rescale the j-th feature by $\tilde{X}_{ij}^{\text{new}} = c \cdot \tilde{X}_{ij}$ for all i
- Similar to OLS, this adjusts the coefficient $\beta_j^{\text{new}} = \frac{\beta_j}{c}$
- However, the scaling affects the penalty
  - Previously, the penalty is β_j^2
  - Now, the penalty is β_j^2/c^2
- **Takeaway**: The coefficient of ridge regression is sensitive to transformation
  - Always do transformation beforehand

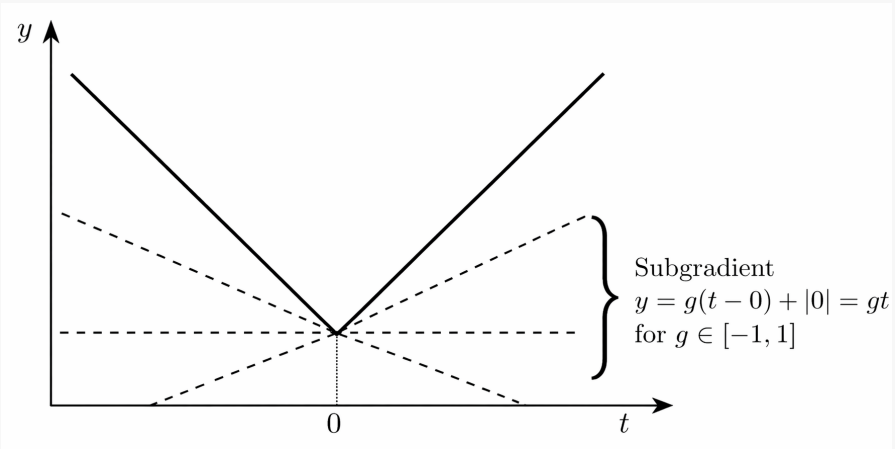
(Extra) Lasso regression: Sparsity (1)

- Let's formally analyze why Lasso leads to sparse solution. Write Lasso as

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \tilde{\boldsymbol{x}}_i^\top \beta)^2 + \lambda ||\beta||_1$$

  - Scaling is slightly different, just for the sake of explanation
  - This is general since you just need to rescale λ to account for $1/2n$

- Note that the regularization term (L^1 penalty) is not differentiable at $|\beta_j| = 0$
  - To account for this, we use the subgradient



(Extra) Lasso regression: Sparsity (2)

Definition (Subgradient)

A number g is subgradient of f at t if for all s

$$f(s) \geq f(t) + g(s - t)$$

- Recall that by the first-order Taylor approximation, gradient satisfies
$$f(s) = f(t) + \nabla f(t)(s - t)$$
  - Subgradient gives you the lower approximation
- Applying this to $f(t) = |t|$, we get $|s| \geq gs$ for the subgradient at $t = 0$
  - Solving for the case of $s > 0$ and $s < 0$, we get $-1 \leq g \leq 1$
- Note that subgradient is not required for final.
  - This is only to explain the sparsity mechanism formally

(Extra) Lasso regression: Sparsity (3)

- By taking the derivative, the solution is optimal if and only if

$$0 = -\frac{1}{n}\tilde{\boldsymbol{X}}_i^\top (Y_i - \tilde{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}}) + \lambda\partial_\beta|\hat{\boldsymbol{\beta}}|$$
$$\Rightarrow \frac{1}{n}\tilde{\boldsymbol{X}}_i^\top (Y_i - \tilde{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}}) = \lambda\partial_\beta|\hat{\boldsymbol{\beta}}|$$

- **Case 1**: If $\hat{\beta}_j \neq 0$, then $\partial_\beta|\beta_j| = \text{sign}(\beta_j)$. So,

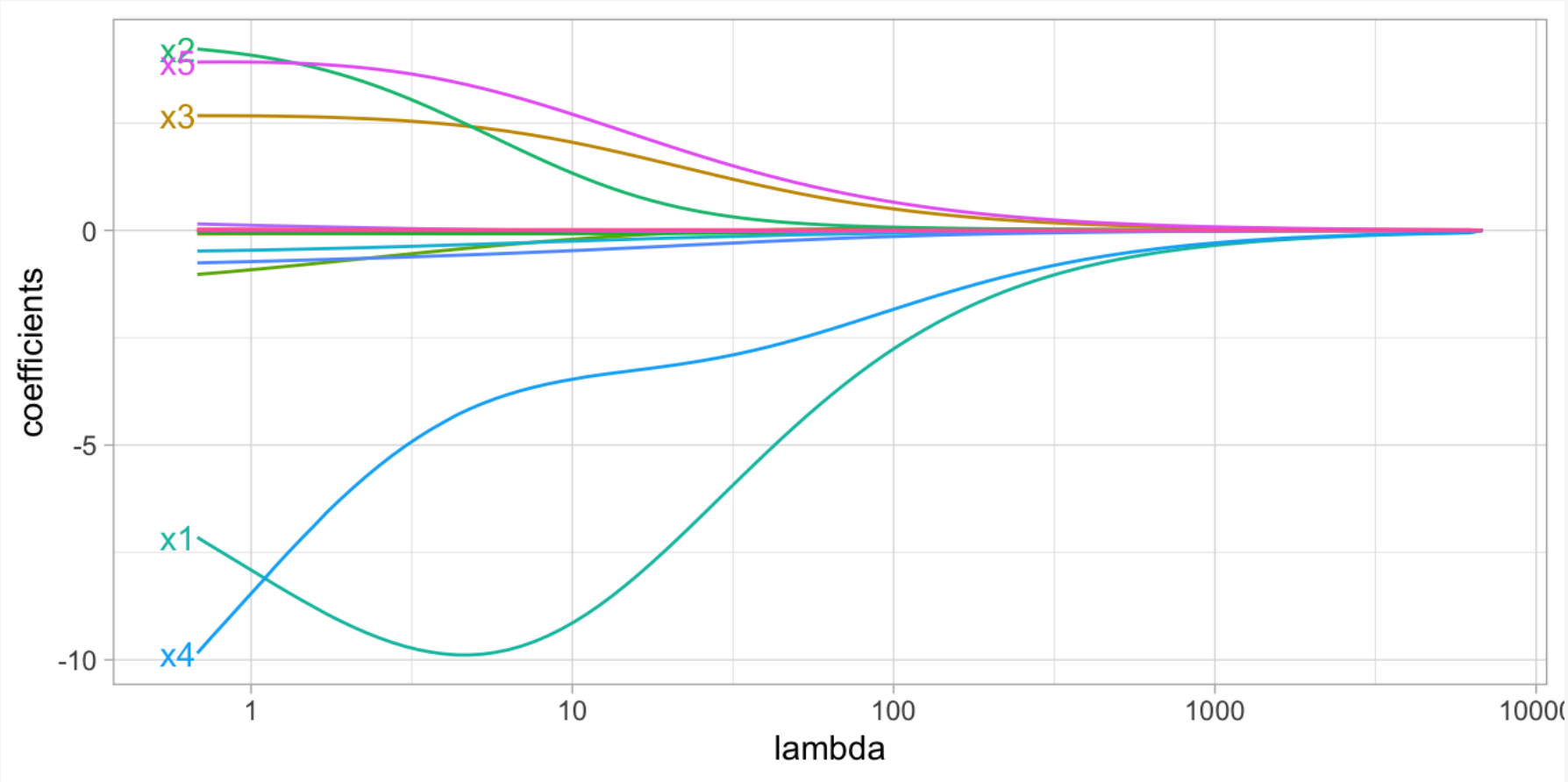
$$\left|\frac{1}{n}\tilde{\boldsymbol{X}}_i^\top (Y_i - \tilde{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}})\right| = \lambda$$

  - I.e., non-zero coefficient must satisfy the exact equality
- **Case 2**: If $\hat{\beta}_j = 0$, then $\partial_\beta|\beta_j| \in [-1, 1]$ by the definition of subgradient, so

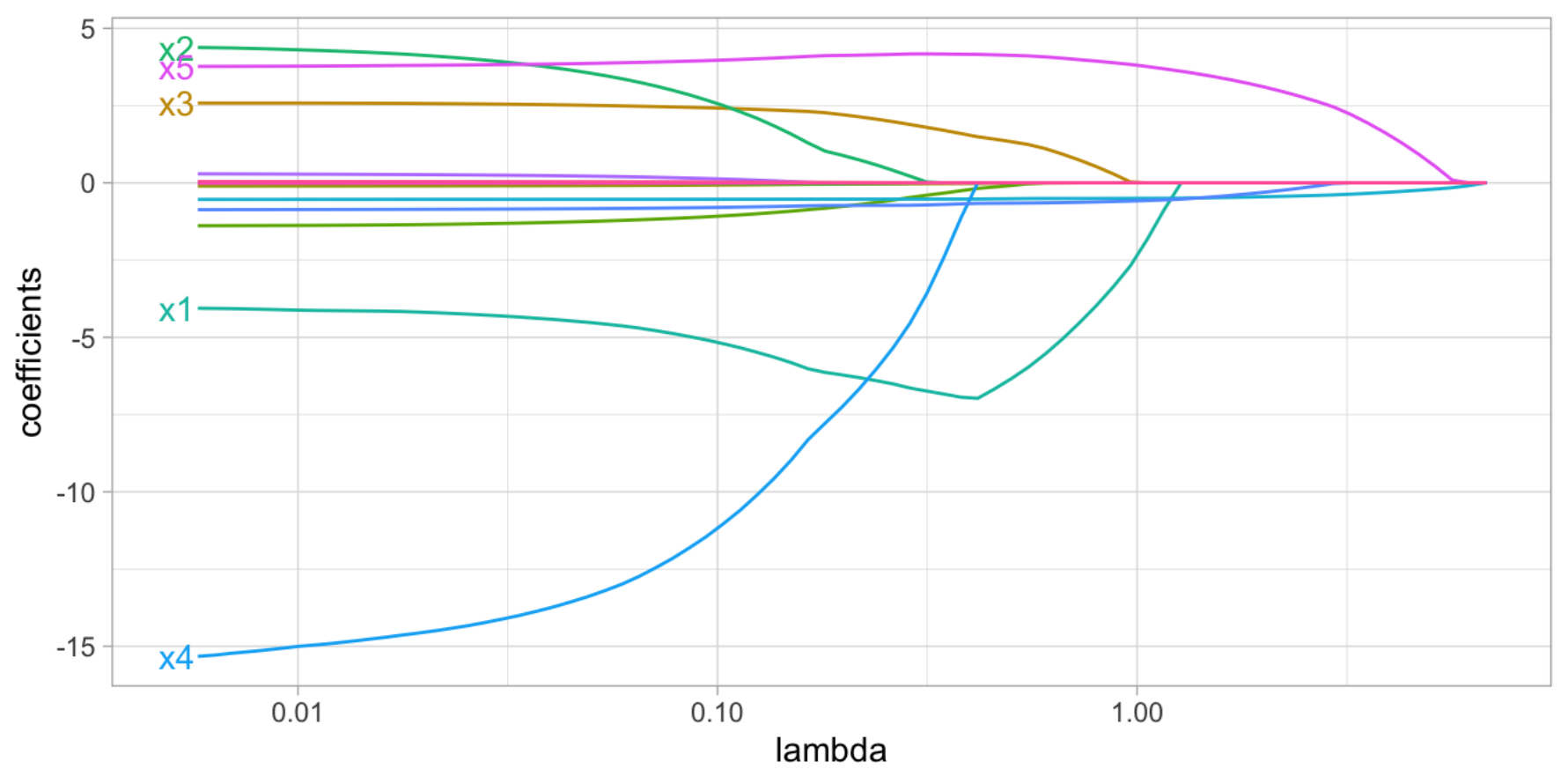
$$\left|\frac{1}{n}\tilde{\boldsymbol{X}}_i^\top (Y_i - \tilde{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}})\right| \leq \lambda$$

  - Zero-coefficients satisfy a loose inequality, encouraging sparsity

Regularization and Parameter: Ridge



Regularization and Parameter: Lasso



Coordinate decent (1)

- For lasso, we cannot have the gradient at 0
  - We only have the subgradient
  - As a result, the gradient decent algorithm is hard to be used
- For Lasso, we use **coordinate decent**
  - **Key Idea**: Only optimize one dimension β_j while fixing the remaining part $\beta_{(-j)}$ (then repeat)
- Recall that Lasso optimization problem is

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \tilde{\boldsymbol{X}}_i^\top \beta)^2 + \lambda ||\beta||_1$$

- If you optimize only j-th coordinate, then

$$\min_{\beta_j} \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \tilde{\boldsymbol{X}}_{i,-j}^\top \beta_{-j} - X_j \beta_j)^2 + \lambda |\beta_j|$$

  - This is much simpler

Coordinate decent (2)

- If you simplify, you get

$$\min_{\beta_j} \frac{1}{2n} \boldsymbol{X}_j^\top \boldsymbol{X}_j \beta_j^2 - \frac{1}{n} \boldsymbol{X}_j^\top (\boldsymbol{Y} - \boldsymbol{X}_{-j} \boldsymbol{\beta}_{-j}) \beta_j + \lambda |\beta_j|$$

- Case 1: If $\beta_j \neq 0$, then by taking derivative

$$\frac{1}{n} \boldsymbol{X}_j^\top \boldsymbol{X}_j \beta_j - \frac{1}{n} \boldsymbol{X}_j^\top (\boldsymbol{Y} - \boldsymbol{X}_{-j} \boldsymbol{\beta}_{-j}) + \lambda \text{sign}(\beta_j) = 0$$

  which gives us

$$\beta_j = \frac{z_j - \lambda \cdot \text{sign}(\beta_j)}{c_j}$$

  where $c_j = \frac{1}{n} \boldsymbol{X}_j^\top \boldsymbol{X}_j$ and $z_j = \frac{1}{n} \boldsymbol{X}_j^\top (\boldsymbol{Y} - \boldsymbol{X}_{-j} \boldsymbol{\beta}_{-j})$

- Case 2: If $\beta_j = 0$, then by subgradient,

$$-z_j \in \lambda[-1, 1] \Rightarrow |z_j| \leq \lambda$$

Coordinate decent (3)

- Now, notie that the sign of β_j and z_j matches.
  - Suppose that $\beta_j > 0$. Then,

$$\beta_j = \frac{z_j - \lambda}{c_j} > 0 \Rightarrow z_j > \lambda > 0$$

    which means that z_j is positive
  - On the other hand, if $\beta_j < 0$, then

$$\beta_j = \frac{z_j + \lambda}{c_j} < 0 \Rightarrow z_j < -\lambda < 0$$

    which means that z_j is negative

- Combining them, we get the closed-form solution

$$\beta_j = \frac{1}{c_j} \cdot \underbrace{\text{sign}(z_j) \cdot \max\{|z_j| - \lambda, 0\}}_{\text{Soft Thresholding}}$$

  - If $|z_j| < \lambda$, we set the coefficient to be zero
  - If $|z_j| > \lambda$, we shrink coefficient toward zero by λ

Extra: Statistical inference after Variable Selection

- Lasso enables you to select variables
  - You use data to select the variables

- **Question**: How can we perform the statistical inference with the selected variable?
  - You want to (i) select the variable and (ii) run regression to perform the statistical inference
  - However, you use the data TWICE!

- This can be seen as a variant of multiple hypothesis testing (or called **selective inference**)
  - You need to condition on the event (i.e., the event that you selected the particular variables)
  - This is what `fixedLassoInf` package is doing

Optimism of Training Error (1)

- Recall that we derived the bias-variance decomposition

$$\begin{aligned}
\mathrm{MSE}(x_{\text{new}}, y_{\text{new}}) &:= \mathbb{E}[(y_{\text{new}} - \hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D}))^2] \\
&= \underbrace{\mathbb{E}[(y_{\text{new}} - f(\boldsymbol{x}_{\text{new}}))^2]}_{\text{Irreducible Error}} \\
&\quad + \underbrace{\left(f(\boldsymbol{x}_{\text{new}}) - \mathbb{E}[\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})]\right)^2}_{\text{Bias}(x_{\text{new}})^2} \\
&\quad + \underbrace{\mathbb{E}\left[\left(\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D}) - \mathbb{E}[\hat{f}(\boldsymbol{x}_{\text{new}}; \mathcal{D})]\right)^2\right]}_{\text{Variance}(x_{\text{new}})}
\end{aligned}$$

- The above definition requires the calculating MSE on the test data
  - I.e., training data $\mathcal{D}$ is different from test data
- Can we calculate test error from the training data?

Optimism of Training Error (2)

- **Setup**: We have training data $\mathcal{D} = \{X_i, Y_i\}_{i=1}^n$ and train the model $\hat{f}$ using the training data

- Training error is defined as

$$\mathrm{Err}_{\mathrm{train}}(\mathcal{D}) := \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}(X_i))^2$$

- Consider evaluating the sample prediction error

$$\mathrm{Err}_{\mathrm{sample}}(\mathcal{D}) := \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_{\mathrm{new}} - \hat{f}(\boldsymbol{X}_{\mathrm{new}}; \mathcal{D}))^2 \mid \boldsymbol{X}_{\mathrm{new}} = \boldsymbol{X}_i, \mathcal{D}]$$

  - This is the test error conditional on the observed $\boldsymbol{X}_i$
  - Indeed, $\mathbb{E}[\mathrm{Err}_{\mathrm{sample}}(\mathcal{D}) \mid \mathcal{D}]$ becomes test error

Optimism of Training Error (3)

- **Goal**: Quantify the gap between sample prediction error and training error

$$\text{Optimism} := \mathbb{E}\left[\text{Err}_{\text{sample}}(\mathcal{D}) - \text{Err}_{\text{train}}(\mathcal{D}) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right]$$

- We will show that

$$\text{Optimism} = \frac{2}{n} \sum_{i=1}^{n} \text{Cov}(\hat{Y}_i, Y_i \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$$

  - where $\hat{Y}_i := \hat{f}(\boldsymbol{X}_i; \mathcal{D})$ is the fitted value

- **Setup for new draw**: $Y_{i,\text{new}}$ is an independent copy satisfying
  - $Y_{i,\text{new}} \mid \boldsymbol{X}_i \sim Y_i \mid \boldsymbol{X}_i$ (same conditional distribution)
  - $Y_{i,\text{new}} \perp \mathcal{D} \mid \boldsymbol{X}_i$ (independent of training data)

Optimism of Training Error (4)

- **Step 1**: Expand sample prediction error. Let $\mu_i := \mathbb{E}[Y_i \mid \boldsymbol{X}_i]$

$$\mathbb{E}[(Y_{i,\text{new}} - \hat{Y}_i)^2 \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, \mathcal{D}] = \mathbb{E}[Y_i^2 \mid \boldsymbol{X}_i] - 2\mu_i \hat{Y}_i + \hat{Y}_i^2$$

  - Uses $\mathbb{E}[Y_{i,\text{new}} \mid \boldsymbol{X}_i] = \mu_i$ and $\mathbb{E}[Y_{i,\text{new}}^2 \mid \boldsymbol{X}_i] = \mathbb{E}[Y_i^2 \mid \boldsymbol{X}_i]$

- Taking the outer expectation over training responses:

$$\begin{aligned}
&\mathbb{E}[\text{Err}_{\text{sample}}(\mathcal{D}) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n] \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \Big\{ \mathbb{E}[Y_i^2 \mid \boldsymbol{X}_i] - 2\mu_i \mathbb{E}[\hat{Y}_i \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n] \\
&\qquad\qquad + \mathbb{E}[\hat{Y}_i^2 \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n] \Big\}
\end{aligned}$$

Optimism of Training Error (5)

- **Step 2**: Expand training error. Note $(Y_i - \hat{Y}_i)^2 = Y_i^2 - 2Y_i\hat{Y}_i + \hat{Y}_i^2$

$$\begin{aligned}
&\mathbb{E}[\mathrm{Err}_{\mathrm{train}}(\mathcal{D}) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n] \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\Big\{\mathbb{E}[Y_i^2 \mid \boldsymbol{X}_i] \\
&\qquad - 2\mathbb{E}[Y_i\hat{Y}_i \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n] + \mathbb{E}[\hat{Y}_i^2 \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]\Big\}
\end{aligned}$$

- Notice that the $\mathbb{E}[Y_i^2]$ and $\mathbb{E}[\hat{Y}_i^2]$ terms are identical in both expressions
  - Only the cross-terms differ

Optimism of Training Error (6)

- **Step 3**: Subtract. Matching terms cancel, leaving

$$\begin{aligned}
&\mathbb{E}[\mathrm{Err}_{\mathrm{sample}}(\mathcal{D}) - \mathrm{Err}_{\mathrm{train}}(\mathcal{D}) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n] \\
&= \frac{2}{n} \sum_{i=1}^{n} \Big\{ \mathbb{E}[Y_i \hat{Y}_i \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n] - \mu_i \mathbb{E}[\hat{Y}_i \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n] \Big\} \\
&= \frac{2}{n} \sum_{i=1}^{n} \mathrm{Cov}(\hat{Y}_i, Y_i \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)
\end{aligned}$$

- **Takeaway**: Training error is optimistically biased downward
  - Typically $\mathrm{Cov}(\hat{Y}_i, Y_i) > 0$ since $\hat{Y}_i$ is fit using Y_i
  - The more flexible the model, the larger the optimism

Optimism of Training Error (7)

- **Special case**: Linear smoother $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$ with $\mathrm{Var}(\boldsymbol{Y} \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n$
  - Case 1: OLS model $Y_i = \boldsymbol{X}_i^\top \beta + \epsilon_i$ with homoskedasticity
  - Case 2: Generalized Additive Model (GAM) $Y_i = f(\boldsymbol{X}_i) + \epsilon_i$ with homoskedasticity
- Covariance simplifies to

$$\mathrm{Cov}(\hat{\boldsymbol{Y}}, \boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{H} \cdot \mathrm{Var}(\boldsymbol{Y} \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{H}$$

- Optimism becomes

$$\frac{2}{n} \sum_{i=1}^{n} \mathrm{Cov}(\hat{Y}_i, Y_i \mid \boldsymbol{X}) = \frac{2}{n} \sum_{i=1}^{n} [\mathrm{Cov}(\hat{\boldsymbol{Y}}, \boldsymbol{Y} \mid \boldsymbol{X})]_{ii} = \frac{2\sigma^2}{n} \mathrm{tr}(\boldsymbol{H})$$

- For OLS with p parameters, $\mathrm{tr}(\boldsymbol{H}) = p$, yielding **Mallows' C_p**:

$$C_p = \mathrm{Err}_{\mathrm{train}}(\mathcal{D}) + \frac{2p}{n} \sigma^2$$

  - This is an unbiased measure of prediction error for linear model
  - For example, it is used as a stopping criterion for variable selection method

Akaike Information Criterion (AIC)

Definition (Akaike Information Criterion (AIC))

AIC is defined as

$$\mathrm{AIC} = -2 \sum_{i=1}^{n} \log \text{ likelihood}_i + 2p$$

where p is the number of parameters

- AIC is also an estimator for prediction error
  - Lower AIC, the better
  - Intuitively, it incorporates the trade-off of overfitting (high p) and underfitting (low likelihood)

- AIC and likelihood ratio test is actually equivalent
  - This is where 2 is coming from.

Akaike Information Criterion (AIC): Example

- Consider the Gaussian linear model

$$Y_i \sim \mathcal{N}(X_i^\top \beta, \sigma^2)$$

- The log-likelihood is given by

$$\log L(\beta, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - X_i^\top \beta)^2$$

- Plug in the MLE $\hat{\beta}$ (and treat σ^2 as fixed for now)

$$-2 \log L(\hat{\beta}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \text{constant}$$

- Therefore, AIC is

$$\text{AIC} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + 2p + \text{constant}$$

- **Takeaway**: AIC = Training error + penalty for complexity

Equivalence of AIC and C_p in Linear Model

- Recall Mallows' C_p

$$C_p = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \frac{2p}{n}\sigma^2$$

- Multiply both sides by $\frac{n}{\hat{\sigma}^2}$

$$\frac{n}{\hat{\sigma}^2}C_p = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + 2p$$

- Rearranging gives

$$\frac{n}{\hat{\sigma}^2}C_p - n = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 - (n - 2p)$$

- Compare with AIC

$$\mathrm{AIC} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + 2p + \text{constant}$$

- The two criteria differ only by constants and scaling

AIC and Likelihood Ratio Test

- Consider two nested models:
  - Smaller model $\mathcal{M}_0$ with p_0 parameters
  - Larger model $\mathcal{M}_1$ with $p_1 > p_0$

- Likelihood ratio statistic is

$$\Lambda = -2 \log \frac{L_0}{L_1} = 2(\log L_1 - \log L_0)$$

- AIC for each model:

$$\begin{aligned} \mathrm{AIC}_0 &= -2 \log L_0 + 2p_0 \\ \mathrm{AIC}_1 &= -2 \log L_1 + 2p_1 \end{aligned}$$

- Take the difference:

$$\mathrm{AIC}_1 - \mathrm{AIC}_0 = -\Lambda + 2(p_1 - p_0)$$

AIC and Likelihood Ratio Test: Equivalence

- AIC prefers the larger model $\mathcal{M}_1$ if

$$\mathrm{AIC}_1 < \mathrm{AIC}_0 \iff \Lambda > 2(p_1 - p_0)$$

- Likelihood ratio test (LRT) rejects $\mathcal{M}_0$ if

$$\Lambda > \chi^2_{p_1 - p_0, \alpha}$$

- Takeaway:
  - AIC = LRT with a fixed threshold $(p_1 - p_0)$
  - LRT = data-dependent threshold via reference distribution