Perspectives on Cross-Validation

Wenda Zhou

January 17, 2020

Estimators and Risk

Given loss $\mathcal{L}$ and hypothesis f, we are interested in its risk:

$$R(f) = \mathbb{E}_X \mathcal{L}(X, f).$$

Estimators and Risk

Given loss $\mathcal{L}$ and hypothesis f, we are interested in its risk:

$$R(f) = \mathbb{E}_X \mathcal{L}(X, f).$$

In general, we do not have access to the distribution of X, but rather samples $X_1, \ldots, X_n$. We may estimate the risk:

$$\hat{R}(f) = \sum_{i=1}^{n} \mathcal{L}(X_i, f).$$

Estimators and Risk

Given loss $\mathcal{L}$ and hypothesis f, we are interested in its risk:

$$R(f) = \mathbb{E}_X \mathcal{L}(X, f).$$

In general, we do not have access to the distribution of X, but rather samples $X_1, \ldots, X_n$. We may estimate the risk:

$$\hat{R}(f) = \sum_{i=1}^{n} \mathcal{L}(X_i, f).$$

We usually are not given f but estimate it from data $\hat{f}(X_1, \ldots, X_n)$:

$$R(\hat{f}) = \mathbb{E}_X \mathcal{L}(X, \hat{f}) = R_n.$$

Estimators and Risk

Given loss $\mathcal{L}$ and hypothesis f, we are interested in its risk:

$$R(f) = \mathbb{E}_X \mathcal{L}(X, f).$$

In general, we do not have access to the distribution of X, but rather samples $X_1, \ldots, X_n$. We may estimate the risk:

$$\hat{R}(f) = \sum_{i=1}^{n} \mathcal{L}(X_i, f).$$

We usually are not given f but estimate it from data $\hat{f}(X_1, \ldots, X_n)$:

$$R(\hat{f}) = \mathbb{E}_X \mathcal{L}(X, \hat{f}) = R_n.$$

The *insample* estimate of the risk is biased:

$$\hat{R}^{\text{in}}(\hat{f}) = \sum_{i=1}^{n} \mathcal{L}(X_i, \hat{f}(X_1, \ldots, X_n)).$$

Sample Splitting

Separate the training and testing sets (let $k = n/m$):

$$\hat{R}^{\text{split}}_{n,k} = \frac{1}{m} \sum_{i=n-m+1}^{n} \mathcal{L}(X_i, \hat{f}(X_1, \ldots, X_{n-m}))$$

Sample Splitting

Separate the training and testing sets (let $k = n/m$):

$$\hat{R}^{\text{split}}_{n,k} = \frac{1}{m} \sum_{i=n-m+1}^{n} \mathcal{L}(X_i, \hat{f}(X_1, \ldots, X_{n-m}))$$

It is an unbiased estimator of $R_{n-n/k} = R_{n,k}$.
If k is constant, then it is asymptotically unbiased for R_n when $\hat{f}$ is parametric.

Problem: part of the data is unused for learning.

Cross-Validation

$$\hat{R}^{\mathsf{cv}}_{n,k} = \frac{1}{k}\sum_{j=1}^{k}\sum_{i=(j-1)m+1}^{jm} \mathcal{L}(X_i, \hat{f}(X_{[\![n]\!]\setminus[\![(j-1)m+1,jm]\!]}))$$

$$\hat{R}^{\mathsf{cv}}_{n,k} = \frac{1}{k}\sum_{i=1}^{k} \hat{R}_i(\hat{f}_{/i}).$$

Cross-Validation

Is $\hat{R}^{\text{cv}}$ a better estimator than $\hat{R}^{\text{split}}$?

Note that we have: $\mathbb{E}\hat{R}^{\text{cv}} = \mathbb{E}\hat{R}^{\text{split}}$, hence it suffices to understand the variance.

Cross-Validation

Is $\hat{R}^{\mathrm{cv}}$ a better estimator than $\hat{R}^{\mathrm{split}}$?

Note that we have: $\mathbb{E}\hat{R}^{\mathrm{cv}} = \mathbb{E}\hat{R}^{\mathrm{split}}$, hence it suffices to understand the variance.

Our hope is that splits behave "independently":

$$\mathrm{Var}\,\hat{R}^{\mathrm{cv}} \approx \frac{1}{k}\,\mathrm{Var}\,\hat{R}^{\mathrm{split}}$$

Main difficulty: the splits are not actually independent, hence subtle analysis.

Cross-Validation: Some Previous Work

- Blum et al. (1999): $\operatorname{Var}\hat{R}^{\text{cv}} < \operatorname{Var}\hat{R}^{\text{split}}$.
- Kale et al. (2011): $\operatorname{Var}\hat{R}^{\text{cv}} \leq (1 + o(1))\frac{1}{k}\operatorname{Var}\hat{R}^{\text{split}}$ under stability conditions.
- Kumar et al. (2013): Further study the stability conditions in Kale et al.

Asymptotics of Cross-Validation

Joint work with Morgane Austern (MSR New England).

Asymptotics

To evaluate such problems, we will establish a central limit theorem.

$$n^{\alpha}(\hat{R}^{\mathsf{CV}}_{n,k} - R_{n,k}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$

A central limit theorem is powerful tool to understand the behaviour of a random quantity.

- Characterize the *rate* of convergence (i.e. α)
- Give sharp *constants* (i.e. σ^2)
- Full description of behaviour to that order / universality

Asymptotics for Cross-Validation

General Result

Suppose that $\hat{f}$ satisfies some stability conditions, and that $k = o(n)$, then we have that:

$$\sqrt{n}(\hat{R}^{\text{split}}_{n,k} - R_{n,k}) \to \mathcal{N}(0, \sigma_1^2 + \sigma_2^2)$$
$$\sqrt{k}\sqrt{n}(\hat{R}^{\text{cv}}_{n,k} - R_{n,k}) \to \mathcal{N}(0, \sigma_1^2 + \sigma_2^2 + 2\rho)$$

where we have:

$$\sigma_1^2 = \lim_n \mathbb{E}\,\mathrm{Var}(\mathcal{L}(X_1, \hat{f}) \mid \hat{f}),$$
$$\sigma_2^2 = \lim_n n(1 - 1/k)\,\mathrm{Var}\,\mathbb{E}[\mathcal{L}(X_1, \hat{f}) \mid \hat{f}],$$
$$\rho = \lim_n \mathrm{Cov}(\mathbb{E}[\mathcal{L}(X', \hat{f}(X_1, \dots, X_n)) \mid X'], \mathbb{E}[\mathcal{L}(\tilde{X}, \hat{f}(X', X_2, \dots, X_n)) \mid X']),$$

Asymptotics: Parametric M-estimator

Suppose that $\hat{f}$ is a parametric M-estimator for a loss Ψ:

$$\hat{f} = \arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^{n} \Psi(X_i, \theta),$$

and that Ψ and $\mathcal{L}$ are nice, then:

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^p} \mathbb{E}\Psi(X_1, \theta),$$

$$G_r = \partial_{\theta^*} R(\theta^*), \quad G_\Psi(X) = \partial_\theta \Psi(X, \theta^*), \quad H = \mathbb{E}[\partial_\theta^2 \Psi(X_1, \theta^*)]$$

$$\begin{aligned} \sigma_1^2 &= \operatorname{Var} \mathcal{L}(X_1, \theta^*), \\ \sigma_2^2 &= G_R^\top H^{-1} \operatorname{Cov}(G_\Psi) H^{-1} G_R, \\ \rho &= -G_R^\top H^{-1} \operatorname{Cov}(G_\Psi(X_1), \mathcal{L}(X_1, \theta^*)). \end{aligned}$$

Results: good news

Corollary: Parametric case with $\Psi = \mathcal{L}$

Suppose that $\hat{f}$ is a parametric estimator, and $\Psi = \mathcal{L}$.
Then, we have that:

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^p} R(\theta) \Rightarrow G_R = \partial_\theta R(\theta^*) = 0.$$

Which immediately implies:

$$\rho = -G_R^\top H^{-1} \operatorname{Cov}(G_\Psi(X_1), \mathcal{L}(X_1, \theta^*)) = 0.$$

Results: good news

Corollary: Parametric case with $\Psi = \mathcal{L}$

Suppose that $\hat{f}$ is a parametric estimator, and $\Psi = \mathcal{L}$.
Then, we have that:

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^p} R(\theta) \Rightarrow G_R = \partial_\theta R(\theta^*) = 0.$$

Which immediately implies:

$$\rho = -G_R^\top H^{-1} \operatorname{Cov}(G_\Psi(X_1), \mathcal{L}(X_1, \theta^*)) = 0.$$

Some surprises: ridge regression

Consider the ridge estimator:

$$\hat{\theta}_{\text{ridge}} = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y - x_i^\top \theta)^2 + \lambda \|\theta\|_2^2.$$

In this case, we have:

$$\mathcal{L}(x, y, \theta) = (y - x^\top \theta)^2,$$
$$\Psi(x, y, \theta) = (y - x^\top \theta)^2 + \lambda \|\theta\|_2^2.$$

Some surprises: ridge regression

Consider the ridge estimator:

$$\hat{\theta}_{\text{ridge}} = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y - x_i^\top \theta)^2 + \lambda \|\theta\|_2^2.$$

In this case, we have:

$$\mathcal{L}(x, y, \theta) = (y - x^\top \theta)^2,$$
$$\Psi(x, y, \theta) = (y - x^\top \theta)^2 + \lambda \|\theta\|_2^2.$$

Under a gaussian design $x \sim \mathcal{N}(0, S_x)$, $y = x^\top \theta_0 + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$, we have:

$$\rho = -4(h^\top S_x h + \sigma^2) h S_x (S_x + \lambda I)^{-1} S_x h < 0$$

where $h = \theta^*_{\text{ridge}} - \theta_0$.

Some surprises: ridge regression

For ridge with gaussian design, $\rho < 0$ implies that the reduction in variance is *larger* than k!

n	$\mathrm{Var}\,\hat{R}_{\text{split}}$	$\mathrm{Var}\,\hat{R}_{\text{cv}}$	Speedup
50	8.08 (0.06)	2.78 (0.02)	2.90 (0.03)
100	7.65 (0.05)	2.42 (0.02)	3.16 (0.03)
200	7.45 (0.05)	2.30 (0.01)	3.24 (0.03)
500	7.15 (0.05)	2.19 (0.01)	3.27 (0.03)
1000	7.23 (0.05)	2.14 (0.01)	3.38 (0.03)
∞	7.140	2.124	3.362

Table: Observed performance of 2-fold cross-validation for a ridge estimator.

$p = 3$, S_X toeplitz with increasing powers of $1/2$, 50000 replications.

Some surprises: impact of data distribution

The general formula indicates that ρ depends on the true distribution of the data. For example, we consider a binary classification problem:

$$Y \sim \text{Bernoulli}(0.5)$$
$$X \mid Y = 0 \sim d_1$$
$$X \mid Y = 1 \sim d_2$$

and consider the linear discriminant estimator:

$$\hat{\mu}_1 = \frac{2}{n}\sum_{i=1}^{n} X_i \mathbb{I}(Y_i = 0), \quad \hat{\mu}_2 = \frac{2}{n}\sum_{i=1}^{n} X_i \mathbb{I}(Y_i = 1)$$

We consider the $0 - 1$ loss (or accuracy):

$$\mathcal{L}(x, y, \mu_1, \mu_2) = \mathbb{I}\Big\{ y = \mathbb{I}\big(|x - \mu_1| > |x - \mu_2|\big) \Big\}.$$

Some surprises: impact of data distribution

- Slow setup: $d_1 = \Gamma(10, 0.15), \quad d_2 = \Gamma(1, 1)$
- Fast setup: $d_1 = \Gamma(1, 10), \quad d_2 = \Gamma(1, 1)$.

	Slow			Fast		
n	$\text{Var}\,\hat{R}_{\text{split}}$	$\text{Var}\,\hat{R}_{\text{CV}}$	Speedup	$\text{Var}\,\hat{R}_{\text{split}}$	$\text{Var}\,\hat{R}_{\text{CV}}$	Speedup
40	1.44	0.83	1.72	0.43	0.19	2.31
160	1.93	1.13	1.71	0.42	0.18	2.33
640	0.66	0.40	1.63	0.43	0.18	2.34
2560	0.53	0.33	1.62	0.44	0.18	2.37
∞	0.53	0.33	1.64	0.43	0.19	2.37

Table: Variance of train-test split and cross-validated accuracy for LDA.

20000 replications, standard errors shown in paper.

A few words on the proof technique

There are a couple of main strategies for central limit theorems. We use a strategy known as Stein's method.

Stein's Method

Fact: Z is normally distributed if and only if, for all absolutely continuous g where $\mathbb{E}|g'(Z)| < \infty$, we have:

$$\mathbb{E}[Zg(Z)] = \mathbb{E}[g'(Z)]$$

We can make this quantitative: for any r.v X:

$$d_W(X, \sigma Z) \leq \sup_{f \in \mathcal{H}} \left| \mathbb{E}[Xg(X) - \sigma^2 g'(X)] \right|$$

where $\mathcal{H} = \{f \in C^2 : \|g'\| \leq 1, \|g''\| \leq 1\}$.

To learn more: read Chatterjee's survey.

Some surprises: ridge regression

Consider the ridge estimator:

$$\hat{\theta}_{\text{ridge}} = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y - x_i^\top \theta)^2 + \lambda \|\theta\|_2^2.$$

In this case, we have:

$$\mathcal{L}(x, y, \theta) = (y - x^\top \theta)^2,$$
$$\Psi(x, y, \theta) = (y - x^\top \theta)^2 + \lambda \|\theta\|_2^2.$$

Under a gaussian design $x \sim \mathcal{N}(0, S_x)$, $y = x^\top \theta_0 + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$, we have:

$$\rho = -4(h^\top S_x h + \sigma^2) h S_x (S_x + \lambda I)^{-1} S_x h < 0$$

where $h = \theta^*_{\text{ridge}} - \theta_0$.

Asymptotics of Cross-Validation

Summary

- General theorem of estimators verifying stability conditions
- Formula for parametric M-estimators
- "Full" speedup for parametric models when $\Psi = \mathcal{L}$
- Surprising behaviour even for parametric models when $\Psi \neq \mathcal{L}$

Other ideas

- Some degenerate cases exist when $\sigma_1^2 = 0$: require careful handling
- Can we estimate $\operatorname{Var} \hat{R}^{\mathrm{cv}}$ from the data? Tricky when k is finite.
- High-dimensional asymptotics?

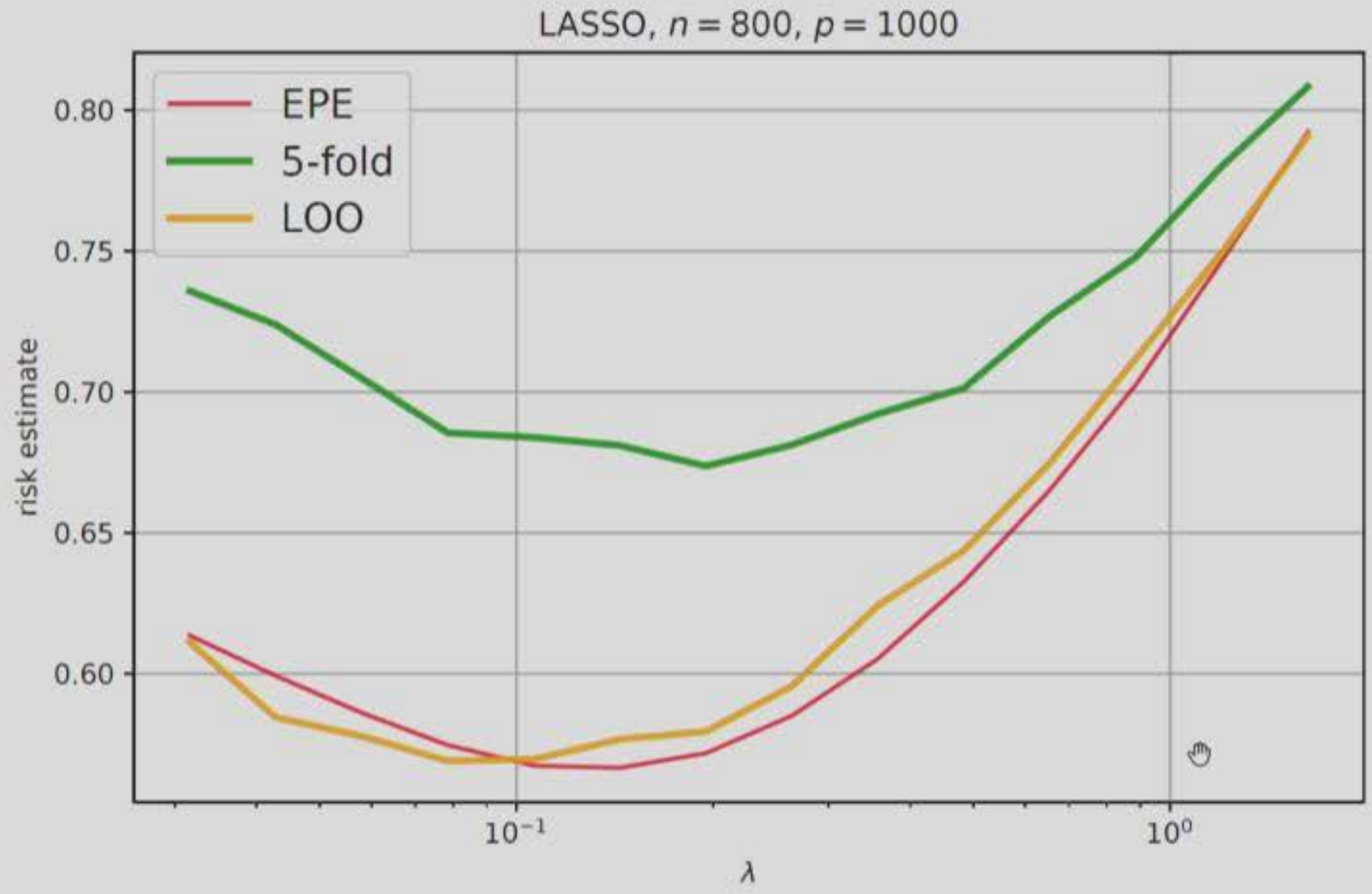
Cross-Validation in the High-Dimensional Regime

Joint work with Kamiar Rad (CUNY Baruch) and Arian Maleki (Columbia).

On the bias of cross-validation

- We often say that cross-validation (or data splitting) is *unbiased*.
- However, $\hat{R}^{\text{cv}}_{n,k}$ is unbiased for $R_{n,k}$, and not R_n.
- In high-dimensional problems, reducing the sample-size by a constant factor affects fundamentally the estimator.

On the bias of cross-validation



On the bias of cross-validation

- Bias reduces as number of folds increases: can we analyze the extreme case of leave-one-out cross-validation ($n = k$)?
- Not clear how variance behaves: large correlations between folds

Generalized Linear Models

Penalized Generalized linear models are a flexible class of models. Consider i.i.d. data $(y_i, x_i) \in \mathbb{R} \times \mathbb{R}^p$.

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \ell(y_i, x_i^\top \beta) + \lambda R(\beta)$$

- Contains in particular LASSO, SVM, matrix completion.
- Decouples the high-dimensional interaction $x_i^\top \beta$ with prediction loss ℓ.

Bounding the error of LOOCV

Theorem (Rad, Z., Maleki)

Assume that (y_i, x_i) is well-behaved, and that ℓ is smooth enough, then, we have that, as $n \to \infty$, $n/p = \delta$:

$$\mathbb{E}(\hat{R}^{\text{cv}}_{n,n} - R_n)^2 \leq \frac{C}{n}.$$

- Idea for proof: Taylor expansion / mean-value theorem.
- Tight rate, but no constants.

Bounding the error of LOOCV

Theorem (Rad, Z., Maleki)

Assume that (y_i, x_i) is well-behaved, and that ℓ is smooth enough, then, we have that, as $n \to \infty$, $n/p = \delta$:

$$\mathbb{E}(\hat{R}^{\mathsf{cv}}_{n,n} - R_n)^2 \leq \frac{C}{n}.$$

- Idea for proof: Taylor expansion / mean-value theorem.
- Tight rate, but no constants.

Approximate Leave-One-Out for Fast Parameter Tuning

Joint work with Shuaiwen Wang (Columbia), Peng Xu (Columbia), Haihao Lu (MIT), Vahab Mirrokni (Google), Arian Maleki (Columbia)

Approximate Computation for LOO

- LOOCV is statistically desirable
- LOOCV is computationally infeasible

Can we obtain a fast approximate estimate of the LOOCV risk?

Approximation through linearization

For linear smoothers, which are estimators which verify:

$$\hat{y} = S(X)y,$$

there exists a closed-form expression for leave-one-out estimates.
In particular, for OLS, we have a closed form expression in terms of the hat matrix:

$$\tilde{r}_i = \frac{\hat{r}_i}{1 - H_{ii}},$$

where $\hat{r}_i = \hat{y}_i - y_i$, $\tilde{r}_i = \hat{y}_i^{/i} - y_i$, and H is given by:

$$H = X(X^\top X)^{-1} X^\top$$

The Primal Approach

$\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{/i}$ respectively minimize:

$$L(\boldsymbol{\beta}) = \sum_{j=1}^{n} \ell(y_j; \boldsymbol{x}_j^\top \boldsymbol{\beta}) + r(\boldsymbol{\beta}),$$

$$L^{/i}(\boldsymbol{\beta}) = \sum_{j \neq i} \ell(y_j; \boldsymbol{x}_j^\top \boldsymbol{\beta}) + r(\boldsymbol{\beta}).$$

The Primal Approach

$\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{/i}$ respectively minimize:

$$L(\boldsymbol{\beta}) = \sum_{j=1}^{n} \ell(y_j; \boldsymbol{x}_j^\top \boldsymbol{\beta}) + r(\boldsymbol{\beta}),$$

$$L^{/i}(\boldsymbol{\beta}) = \sum_{j \neq i} \ell(y_j; \boldsymbol{x}_j^\top \boldsymbol{\beta}) + r(\boldsymbol{\beta}).$$

Idea: $\hat{\boldsymbol{\beta}}$ might be a good starting point to $\hat{\boldsymbol{\beta}}^{/i}$. Approximate $\hat{\boldsymbol{\beta}}^{/i}$ by a Newton step from $\hat{\boldsymbol{\beta}}$.

$$\tilde{\boldsymbol{\beta}}^{/i} = \hat{\boldsymbol{\beta}} + (H^{/i})^{-1} G^{/i}.$$

where $H^{/i} = \nabla^2 L^{/i}(\hat{\boldsymbol{\beta}})$ and $G^{/i} = \nabla L^{/i}(\hat{\boldsymbol{\beta}})$.

The Primal Approach

$$\begin{aligned}\nabla^2 L^{/i}(\boldsymbol{\beta}) &= \sum_{j \neq i} \ddot{\ell}(y_j, \boldsymbol{x}_j^\top \boldsymbol{\beta}) \boldsymbol{x}_j \boldsymbol{x}_j^\top + \nabla^2 R(\boldsymbol{\beta}) \\ &= \nabla^2 L(\boldsymbol{\beta}) - \ddot{\ell}(y_i, \boldsymbol{x}_i^\top \boldsymbol{\beta}) \boldsymbol{x}_i \boldsymbol{x}_i^\top .\end{aligned}$$

The Primal Approach

$$\nabla^2 L^{/i}(\boldsymbol{\beta}) = \sum_{j \neq i} \ddot{\ell}(y_j, \boldsymbol{x}_j^\top \boldsymbol{\beta}) \boldsymbol{x}_j \boldsymbol{x}_j^\top + \nabla^2 R(\boldsymbol{\beta})$$
$$= \nabla^2 L(\boldsymbol{\beta}) - \ddot{\ell}(y_i, \boldsymbol{x}_i^\top \boldsymbol{\beta}) \boldsymbol{x}_i \boldsymbol{x}_i^\top .$$

$\nabla^2 L^{/i}$ differs from $\nabla^2 L$ by a rank-1 matrix. Use rank-1 inverse formula:

$$(H^{/i})^{-1} = H^{-1} + \frac{H^{-1} \boldsymbol{x}_i \boldsymbol{x}_i^\top H^{-1}}{\ddot{\ell}_i^{-1} + \boldsymbol{x}_i^\top H^{-1} \boldsymbol{x}_i}.$$

Plug-in to Newton's formula to get:

$$\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{/i} = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \frac{H_{ii} \dot{\ell}(y_i; \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})}{1 - H_{ii} \ddot{\ell}(y_i; \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})}.$$

The Primal Approach

General formula for smooth problems:

$$\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{/i} = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \frac{H_{ii}\dot{\ell}(y_i; \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})}{1 - H_{ii}\ddot{\ell}(y_i; \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})}.$$

- General formula
- Provable accuracy (compared to LOO: Rad and Maleki, 2018)

Non-Smooth Estimators

- In high-dimensional setting, often wish to use non-smooth penalizers.
- Non-smooth penalizers can induce structure in the estimation (sparsity, low-rank).

Non-Smooth Estimators

- In high-dimensional setting, often wish to use non-smooth penalizers.
- Non-smooth penalizers can induce structure in the estimation (sparsity, low-rank).

Consider lasso estimator:

LASSO: $$\min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{j=1}^{n} (\boldsymbol{x}_j^\top \boldsymbol{\beta} - y_j)^2 + \lambda \|\boldsymbol{\beta}\|_1$$

Problem: R is not differentiable everywhere, and $\nabla^2 R(\hat{\boldsymbol{\beta}})$ very likely to be ill-defined.

Non-Smooth Estimators

- In high-dimensional setting, often wish to use non-smooth penalizers.
- Non-smooth penalizers can induce structure in the estimation (sparsity, low-rank).

Consider lasso estimator:

$$\text{LASSO:} \qquad \min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{j=1}^{n} (\boldsymbol{x}_j^\top \boldsymbol{\beta} - y_j)^2 + \lambda \|\boldsymbol{\beta}\|_1$$

Problem: R is not differentiable everywhere, and $\nabla^2 R(\hat{\boldsymbol{\beta}})$ very likely to be ill-defined.

ALO Examples: LASSO

LASSO: $$\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{x}_j^\top \boldsymbol{\beta} - y_j)^2 + \lambda\|\boldsymbol{\beta}\|_1$$

ALO Examples: LASSO

$$\text{LASSO:} \qquad \min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{x}_j^\top \boldsymbol{\beta} - y_j)^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Let $\hat{\boldsymbol{\beta}}$ be the estimator on the full dataset

$$\text{ALO:} \qquad \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}^{/i} \approx \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \frac{H_{ii}}{1 - H_{ii}}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} - y_i)$$

$$\boldsymbol{H} = \boldsymbol{X}_S(\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^\top, \quad S = \{j : \hat{\beta}_j \neq 0\}$$

ALO Examples: LASSO

LASSO: $$\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{x}_j^\top \boldsymbol{\beta} - y_j)^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Let $\hat{\boldsymbol{\beta}}$ be the estimator on the full dataset

ALO: $$\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}^{/i} \approx \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \frac{H_{ii}}{1 - H_{ii}}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} - y_i)$$

$$\boldsymbol{H} = \boldsymbol{X}_S(\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^\top, \quad S = \{j : \hat{\beta}_j \neq 0\}$$

Equivalently, we may write:

ALO residual ⟵ $$\tilde{r}_i = \frac{\hat{r}_i}{1 - H_{ii}}$$

$\hat{r}_i$ → In-sample residual

$1 - H_{ii}$ → leverage

with $\tilde{r}_i = y - \tilde{y}_i$ and $\hat{r}_i = y - \hat{y}_i$.

ALO Examples: LASSO

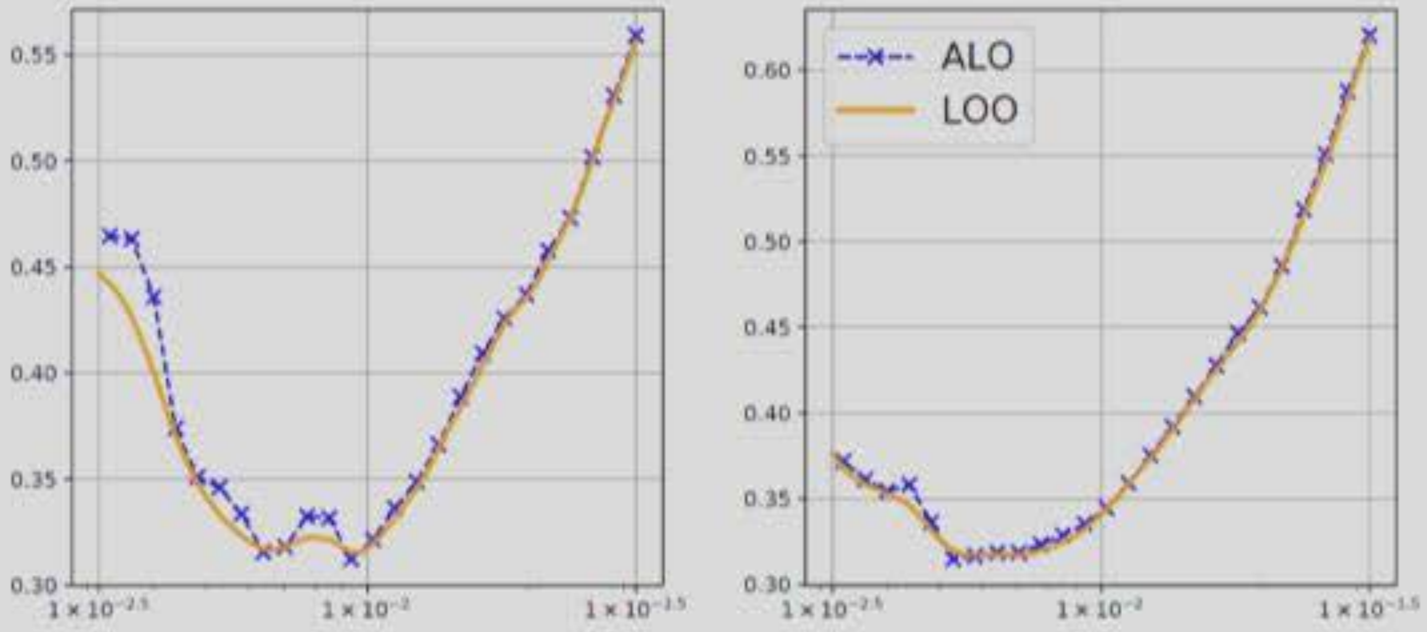


Figure: LOO vs ALO risk estimates for LASSO

p	200	400	1600
single fit	0.035	0.13	0.60
ALO	0.06	0.21	0.89
LOOCV	27	107	480

Table: Time (in s) for each procedure ($n = 800$)

ALO Examples: SVM, Nuclear Norm

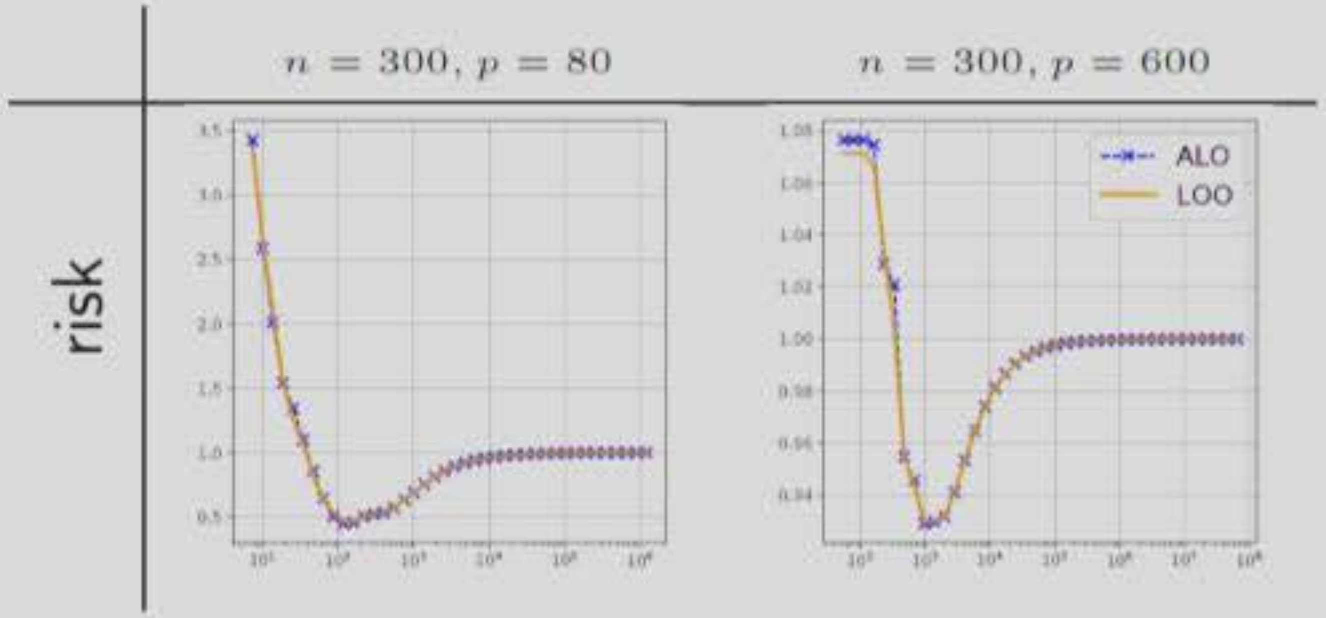


Figure: LOO vs. ALO risk estimates of SVM.

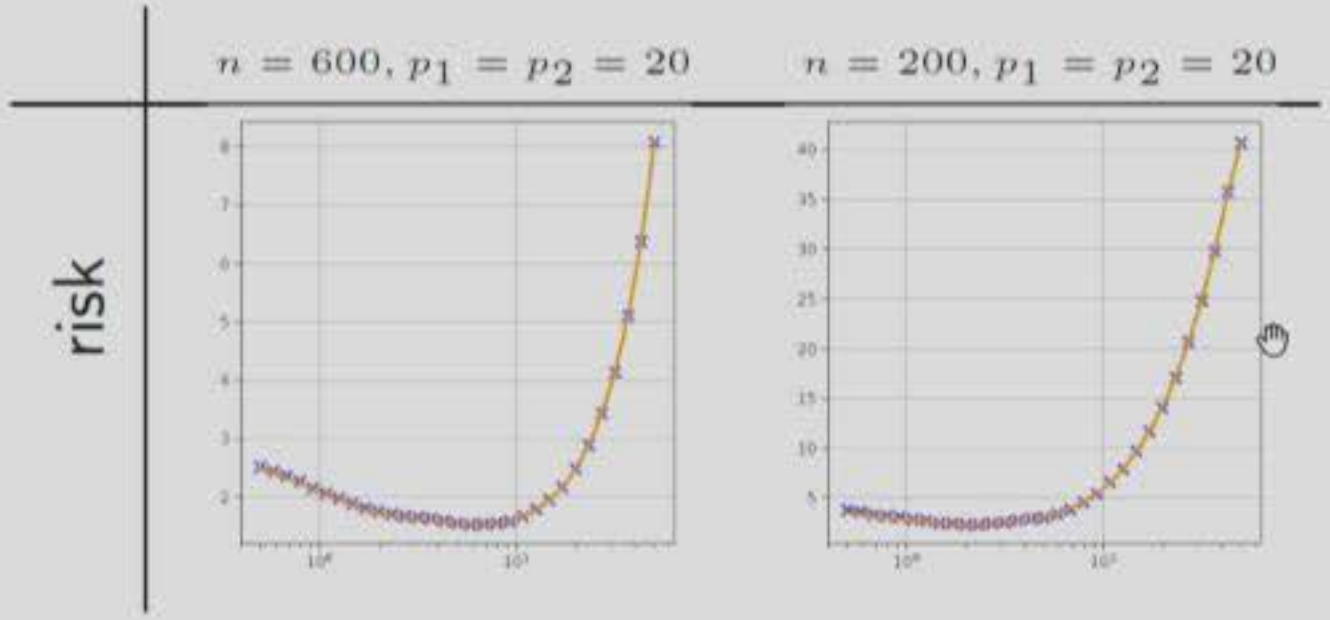


Figure: LOO vs. ALO risk estimates of nuclear norm minimization.

The Dual Approach - LASSO Example

primal: $\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$

$$\updownarrow$$

$\min_{\boldsymbol{\beta},\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$, s.t. $\boldsymbol{w} = \boldsymbol{X}^\top\boldsymbol{\beta}$

$$\updownarrow$$

dual: $\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{\theta}\|_2^2$, s.t. $\|\boldsymbol{X}^\top\boldsymbol{\theta}\|_\infty \leq \lambda$

$$\updownarrow$$

$$\hat{\boldsymbol{\theta}} = \mathrm{Proj}_{\boldsymbol{\Delta}_n}(\boldsymbol{y}), \quad \boldsymbol{\Delta}_n = \{\boldsymbol{\theta} : \|\boldsymbol{X}^\top\boldsymbol{\theta}\|_\infty \leq \lambda\}$$

Primal-Dual correspondence: $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\theta}}$.

The Dual Approach - LASSO Example

Leave-i-out problem:

primal: $\min_{\boldsymbol{\beta}} \sum_{j \neq i} \frac{1}{2}(y_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_1$

dual: $\hat{\boldsymbol{\theta}}^{/i} = \mathrm{Proj}_{\boldsymbol{\Delta}_{n-1}}(\boldsymbol{y}_{-i})$

$$\boldsymbol{y}_{-i} - \boldsymbol{X}_{-i}\hat{\boldsymbol{\beta}}^{/i} = \hat{\boldsymbol{\theta}}^{/i}$$

Dimension Mismatch. $\rightarrow$ Lift the Dimension.

The Dual Approach - LASSO Example

primal: $\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$

$$\updownarrow$$

$\min_{\boldsymbol{\beta},\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$, s.t. $\boldsymbol{w} = \boldsymbol{X}^\top\boldsymbol{\beta}$

$$\updownarrow$$

dual: $\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{\theta}\|_2^2$, s.t. $\|\boldsymbol{X}^\top\boldsymbol{\theta}\|_\infty \leq \lambda$

$$\updownarrow$$

$$\hat{\boldsymbol{\theta}} = \mathrm{Proj}_{\boldsymbol{\Delta}_n}(\boldsymbol{y}), \quad \boldsymbol{\Delta}_n = \{\boldsymbol{\theta} : \|\boldsymbol{X}^\top\boldsymbol{\theta}\|_\infty \leq \lambda\}$$

Primal-Dual correspondence: $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\theta}}$.

The Dual Approach - LASSO Example

Leave-i-out problem:

primal: $\min_{\boldsymbol{\beta}} \sum_{j \neq i} \frac{1}{2}(y_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_1$

dual: $\hat{\boldsymbol{\theta}}^{/i} = \mathrm{Proj}_{\boldsymbol{\Delta}_{n-1}}(\boldsymbol{y}_{-i})$

$$\boldsymbol{y}_{-i} - \boldsymbol{X}_{-i}\hat{\boldsymbol{\beta}}^{/i} = \hat{\boldsymbol{\theta}}^{/i}$$

Dimension Mismatch. $\rightarrow$ Lift the Dimension.

The Dual Approach - LASSO Example

Leave-i-out problem:

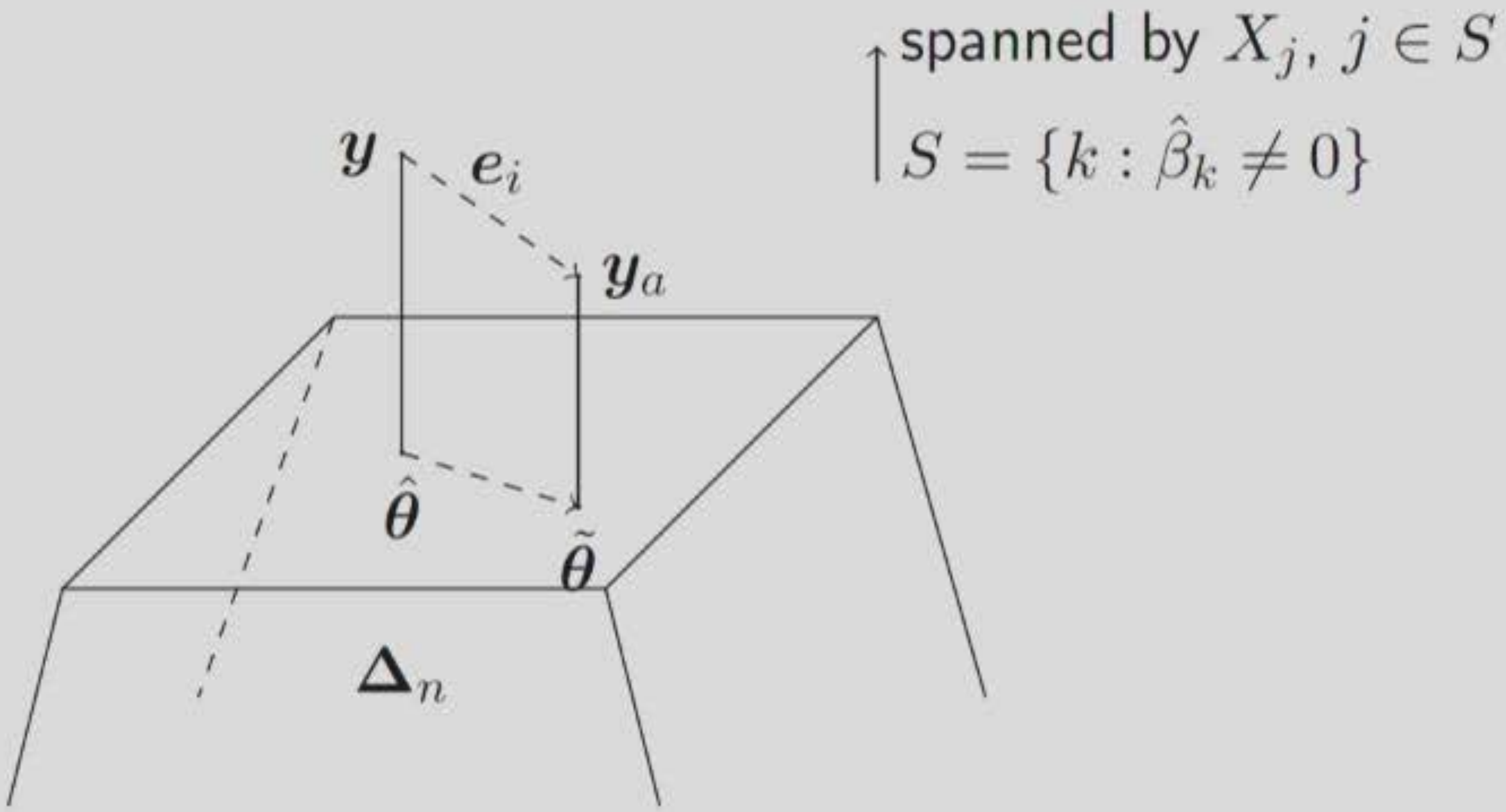
$$\text{primal: } \min_{\boldsymbol{\beta}} \sum_{j \neq i} \frac{1}{2}(\underset{\overset{\shortparallel}{y_{a,j}}}{y_j} - \boldsymbol{x}_j^\top \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_1 + \frac{1}{2}(\underset{\overset{\shortparallel}{y_{a,i}}}{\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}^{/i}} - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$$

dual: $\hat{\boldsymbol{\theta}}^{/i} = \mathrm{Proj}_{\boldsymbol{\Delta}_{n-1}}(\boldsymbol{y}_{-i})$ $\qquad$ $\tilde{\boldsymbol{\theta}} = \mathrm{Proj}_{\boldsymbol{\Delta}_n}(\boldsymbol{y}_a)$

$\boldsymbol{y}_{-i} - \boldsymbol{X}_{-i}\hat{\boldsymbol{\beta}}^{/i} = \hat{\boldsymbol{\theta}}^{/i}$ $\qquad$ $\boldsymbol{y}_a - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{/i} = \tilde{\boldsymbol{\theta}} \Rightarrow \tilde{\theta}_i = 0$

Dimension Mismatch. $\rightarrow$ Lift the Dimension.

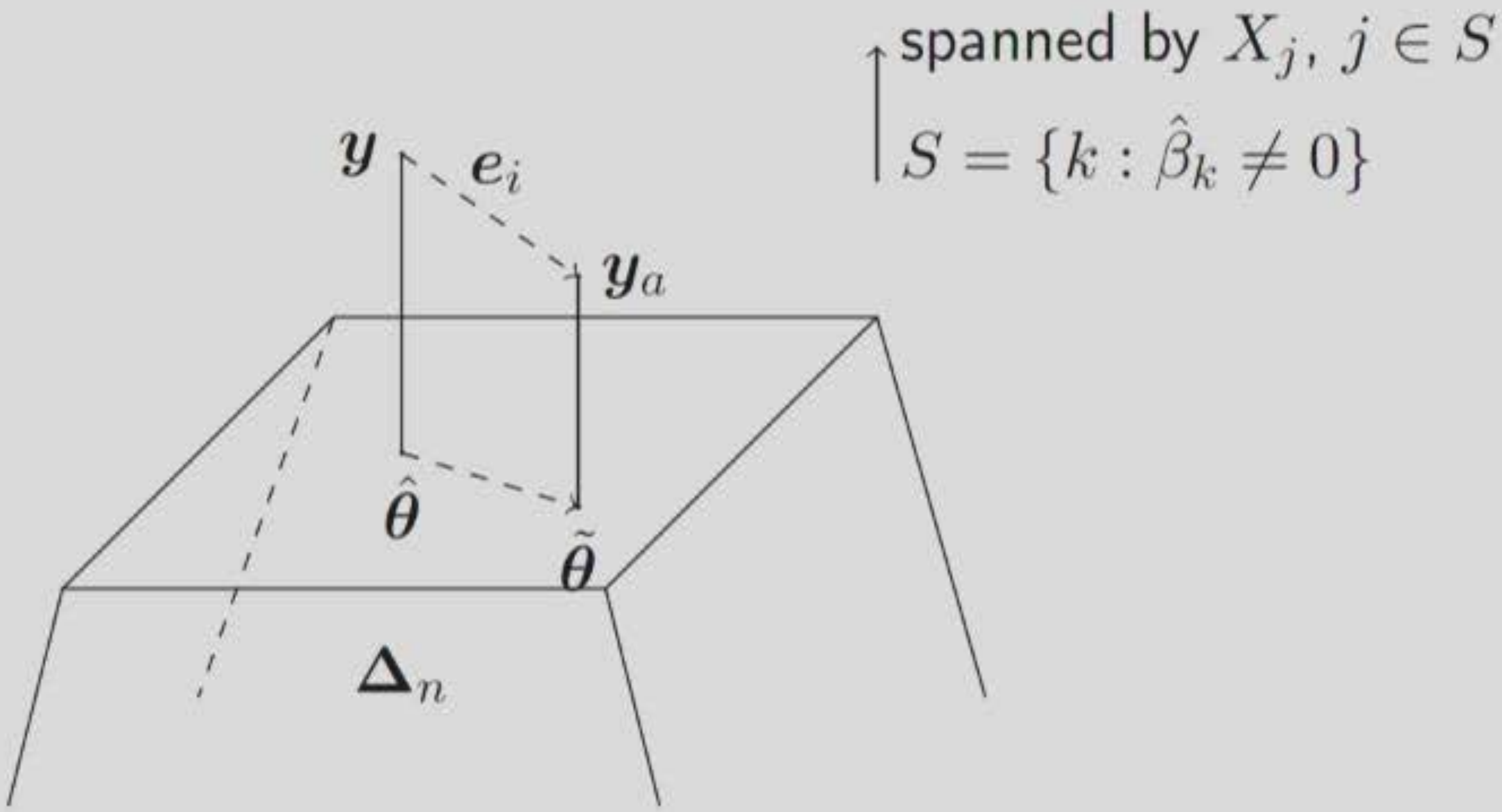
The Dual Approach - LASSO Example cont'



$$\hat{\theta} - \tilde{\theta} = (\boldsymbol{I} - \overbrace{\boldsymbol{X}_S(\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^\top}^{\boldsymbol{H}})(\boldsymbol{y} - \boldsymbol{y}_a)$$

$$\hat{\theta}_i = \left[(\boldsymbol{I} - \boldsymbol{X}_S(\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^\top)(\boldsymbol{y} - \boldsymbol{y}_a)\right]_i$$

The Dual Approach - LASSO Example cont'



$$\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} = (\boldsymbol{I} - \overbrace{\boldsymbol{X}_S(\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^\top}^{\boldsymbol{H}})(\boldsymbol{y} - \boldsymbol{y}_a)$$

$$y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} = \hat{\theta}_i = \left[(\boldsymbol{I} - \boldsymbol{X}_S(\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^\top)(\boldsymbol{y} - \boldsymbol{y}_a)\right]_i$$
$$= (1 - H_{ii})(y_i - y_{a,i})$$

$$y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}^{/i} = y_i - y_{a,i} = \frac{y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}}{1 - H_{ii}}$$

The Dual Approach

- Can be generalized beyond LASSO
- Very useful for norm-type regularizers (e.g. generalized LASSO, SLOPE)

The Dual Approach

- Can be generalized beyond LASSO
- Very useful for norm-type regularizers (e.g. generalized LASSO, SLOPE)

Equivalence between primal and dual approach.

- Formulated in terms of partial quadratics
- Holds even for non-smooth problems (see paper for details)

Non-Smooth Estimators

- In high-dimensional setting, often wish to use non-smooth penalizers.
- Non-smooth penalizers can induce structure in the estimation (sparsity, low-rank).

Consider lasso estimator:

$$\text{LASSO:} \qquad \min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{j=1}^{n} (\boldsymbol{x}_j^\top \boldsymbol{\beta} - y_j)^2 + \lambda \|\boldsymbol{\beta}\|_1$$

Problem: R is not differentiable everywhere, and $\nabla^2 R(\hat{\boldsymbol{\beta}})$ very likely to be ill-defined.

The Dual Approach - LASSO Example

primal: $\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$

$$\updownarrow$$

$\min_{\boldsymbol{\beta},\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1,$ s.t. $\boldsymbol{w} = \boldsymbol{X}^\top\boldsymbol{\beta}$

$$\updownarrow$$

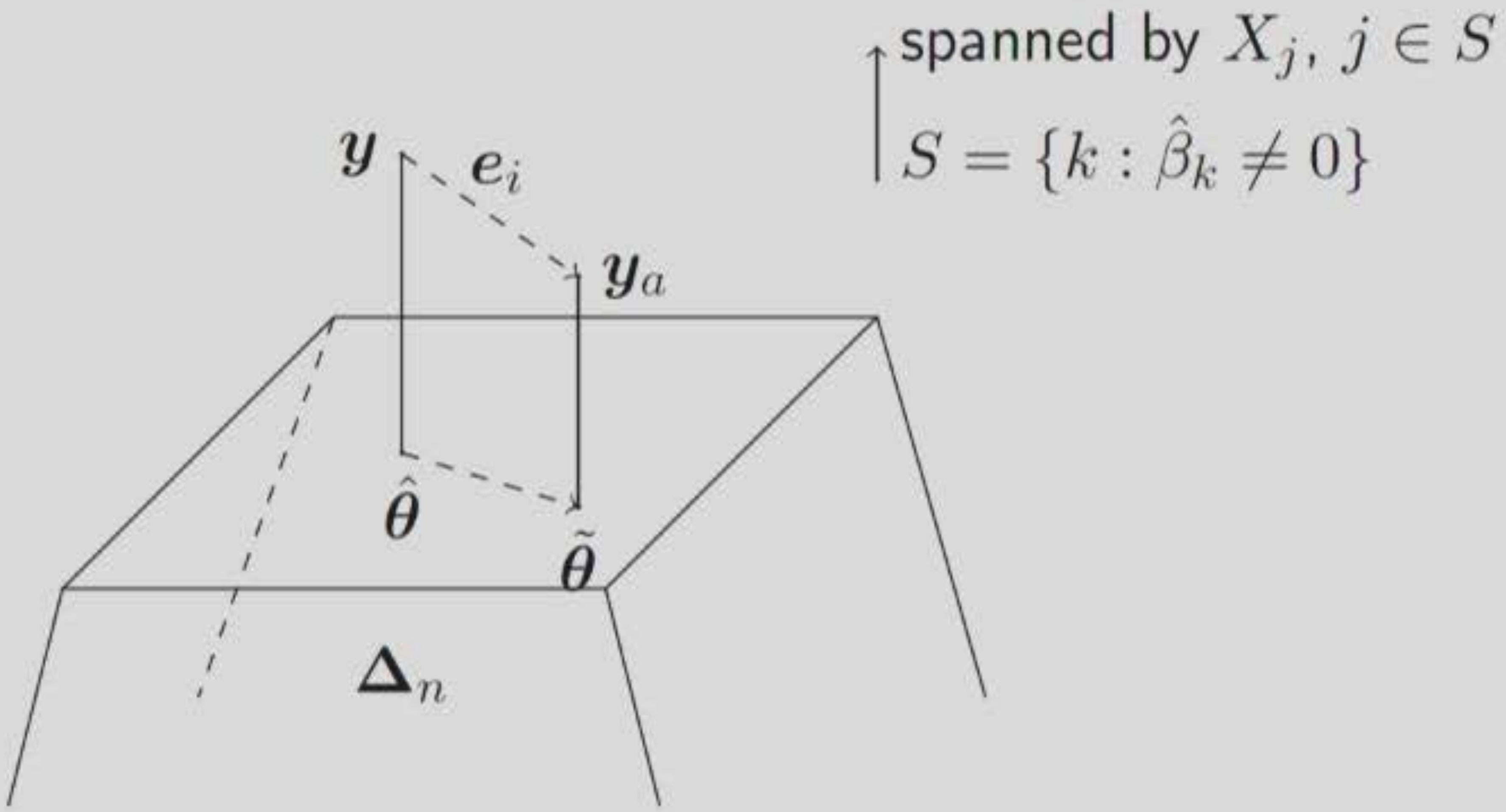
dual: $\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{\theta}\|_2^2,$ s.t. $\|\boldsymbol{X}^\top\boldsymbol{\theta}\|_\infty \leq \lambda$

$$\updownarrow$$

$$\hat{\boldsymbol{\theta}} = \mathrm{Proj}_{\boldsymbol{\Delta}_n}(\boldsymbol{y}), \quad \boldsymbol{\Delta}_n = \{\boldsymbol{\theta} : \|\boldsymbol{X}^\top\boldsymbol{\theta}\|_\infty \leq \lambda\}$$

Primal-Dual correspondence: $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\theta}}.$

The Dual Approach - LASSO Example cont'



$$\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} = (\boldsymbol{I} - \overbrace{\boldsymbol{X}_S(\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^\top}^{\boldsymbol{H}})(\boldsymbol{y} - \boldsymbol{y}_a)$$

$$y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} = \hat{\theta}_i = \left[(\boldsymbol{I} - \boldsymbol{X}_S(\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^\top)(\boldsymbol{y} - \boldsymbol{y}_a)\right]_i$$
$$= (1 - H_{ii})(y_i - y_{a,i})$$

$$y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}^{/i} = y_i - y_{a,i} = \frac{y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}}{1 - H_{ii}}$$

The Dual Approach

- Can be generalized beyond LASSO
- Very useful for norm-type regularizers (e.g. generalized LASSO, SLOPE)

Equivalence between primal and dual approach.

- Formulated in terms of partial quadratics
- Holds even for non-smooth problems (see paper for details)

Approximate Leave-One-Out

- Generic framework for obtaining risk estimators in the high-dimensional regime. In the scenario considered, compares favorably against alternatives:

  Compared to SURE cross-validation is model free, and estimates the out of sample risk. $\operatorname{tr} H$ is related to the degrees of freedom.

  Compared to IJ (Giordano et al. 2019) : ALO has better behavior when p is large compared to n. However, IJ is more flexible.

Approximate Leave-One-Out

- Generic framework for obtaining risk estimators in the high-dimensional regime. In the scenario considered, compares favorably against alternatives:

  Compared to SURE cross-validation is model free, and estimates the out of sample risk. $\operatorname{tr} H$ is related to the degrees of freedom.

  Compared to IJ (Giordano et al. 2019) : ALO has better behavior when p is large compared to n. However, IJ is more flexible.

- Work in progress: applications in neuroscience.
- Many unanswered questions: e.g. in the interpolating regime (when $\hat{y}_i = y_i$), nearly all linearization strategies (ALO, IJ) break down. How can we produce fast estimates of the risk in that regime?

Thanks!

ALO Examples: LASSO

$$\text{LASSO:} \qquad \min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{x}_j^\top \boldsymbol{\beta} - y_j)^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Let $\hat{\boldsymbol{\beta}}$ be the estimator on the full dataset

$$\text{ALO:} \qquad \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}^{/i} \approx \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \frac{H_{ii}}{1 - H_{ii}}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} - y_i)$$

$$\boldsymbol{H} = \boldsymbol{X}_S(\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^\top, \quad S = \{j : \hat{\beta}_j \neq 0\}$$

ALO Examples: LASSO

LASSO: $$\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{x}_j^\top \boldsymbol{\beta} - y_j)^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Let $\hat{\boldsymbol{\beta}}$ be the estimator on the full dataset

ALO: $$\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}^{/i} \approx \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \frac{H_{ii}}{1 - H_{ii}}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} - y_i)$$

$$\boldsymbol{H} = \boldsymbol{X}_S(\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^\top, \quad S = \{j : \hat{\beta}_j \neq 0\}$$

Equivalently, we may write:

$$\text{ALO residual} \longleftarrow \tilde{r}_i = \frac{\hat{r}_i}{1 - H_{ii}}$$

($\hat{r}_i$ → In-sample residual; $1 - H_{ii}$ → leverage)

with $\tilde{r}_i = y - \tilde{y}_i$ and $\hat{r}_i = y - \hat{y}_i$.

ALO Examples: LASSO

LASSO: $$\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{x}_j^\top \boldsymbol{\beta} - y_j)^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Let $\hat{\boldsymbol{\beta}}$ be the estimator on the full dataset

ALO: $$\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}^{/i} \approx \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \frac{H_{ii}}{1-H_{ii}}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} - y_i)$$

$$\boldsymbol{H} = \boldsymbol{X}_S(\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^\top, \quad S = \{j : \hat{\beta}_j \neq 0\}$$

Equivalently, we may write:

ALO residual $\longleftarrow$ $$\tilde{r}_i = \frac{\hat{r}_i}{1-H_{ii}}$$

$\hat{r}_i \rightarrow$ In-sample residual

$H_{ii} \rightarrow$ leverage

with $\tilde{r}_i = y - \tilde{y}_i$ and $\hat{r}_i = y - \hat{y}_i$.