

# 1 Simple model

## Definitions

We find the best model for  $y$  over function class  $\mathcal{G}$ . Presume  $g^* \in \mathcal{G}$  is the true model and

$$y = g^*(X) + \epsilon$$

where  $\epsilon$  are sub-Gaussian errors for constants  $K$  and  $\sigma_0^2$

$$\max_{i=1:n} K^2 (E [\exp(|\epsilon_i|^2 K^2) - 1]) \leq \sigma_0^2$$

Given a training set  $T$ , We define the fitted models

$$\hat{g}_\lambda = \|y - g\|_T^2 + \lambda^2 I^v(g)$$

Given a validation set  $V$ , let the CV-fitted model be

$$\hat{g}_{\hat{\lambda}} = \arg \min_{\lambda} \|y - \hat{g}_\lambda\|_V^2$$

We will suppose  $I(g^*) > 0$ .

## Assumptions

Suppose the entropy of the class  $\mathcal{G}'$  is

$$H \left( \delta, \mathcal{G}' = \left\{ \frac{g - g^*}{I(g) + I(g^*)} : g \in \mathcal{G}, I(g) + I(g^*) > 0 \right\}, P_T \right) \leq \tilde{A} \delta^{-\alpha} \quad (1)$$

Suppose  $v > 2\alpha/(2 + \alpha)$ .

Suppose for all  $\lambda \in \Lambda$ ,  $I^v(\hat{g}_\lambda)$  is upper bounded by  $\|\hat{g}_\lambda\|_n^2 = \frac{1}{n} \sum_{i=1}^n \hat{g}_\lambda(x_i)$ . See Lemma 1 below for the specific assumption. This assumption includes Ridge, Lasso, Generalized Lasso, and the Group Lasso.

## Result 1: Single $\lambda$ , Single Penalty, cross-validation over general $X_T, X_V$

Suppose that the training and validation set are independently sampled, so the values  $X_i$  are not necessarily the same. Suppose the training and validation sets are both of size  $n$ . Suppose  $X$  is bounded s.t.  $|X| \leq R_X$  and the domain of  $g \in \mathcal{G}$  is over  $(-R_X, R_X)$ .

Suppose the same entropy bound (2) for both the training set  $P_T$  and validation set  $P_V$ .

Suppose for all  $\lambda \in \Lambda$ , there exists a compatibility constant  $M$  s.t.  $I^v(\hat{g}_\lambda)$  is upper bounded by its  $L_2$ -norm with some constant  $M$  (and  $M_0$ ) such that

$$I^v(\hat{g}_\lambda) \leq M \|\hat{g}_\lambda\|_n^2 + M_0$$

Suppose the entropy bound for both training set  $P_T$  and validation set  $P_V$ .

Suppose that

$$\sup_{g \in \mathcal{G}} \frac{\|g - g^*\|_\infty}{I(g) + I(g^*)} \leq K < \infty$$

Let  $\tilde{\lambda}$  be the optimal  $\lambda$  by Vandegeer. Then

$$\|\hat{g}_{\tilde{\lambda}} - \hat{g}_{\hat{\lambda}}\|_V = O_p \left( n^{-1/(2+\alpha)} \right) \left( I^{\alpha/(2+\alpha)}(g^*) + I(g^*) \right)$$

and  $\|\hat{g}_{\bar{\lambda}} - g^*\|_V$  is of the same order (differs by some constant).

**Proof:**

By the triangle inequality,

$$\|\hat{g}_{\bar{\lambda}} - g^*\|_V \leq \|\hat{g}_{\bar{\lambda}} - \hat{g}_{\hat{\lambda}}\|_V + \|\hat{g}_{\hat{\lambda}} - g^*\|_V$$

We bound each component on the RHS separately.

First bound  $\|\hat{g}_{\bar{\lambda}} - g^*\|_V$ . By Vandegeer Thrm 10.2 and Lemma 2,

$$\begin{aligned} \|\hat{g}_{\bar{\lambda}} - g^*\|_V &\leq \|\hat{g}_{\bar{\lambda}} - g^*\|_T + \left| \|\hat{g}_{\bar{\lambda}} - g^*\|_V - \|\hat{g}_{\bar{\lambda}} - g^*\|_T \right| \\ &\leq O_p\left(n^{-1/(2+\alpha)}\right) I^{\alpha/(2+\alpha)}(g^*) + O_p\left(n^{-1/(2+\alpha)}\right) (I(g^*) + I(\hat{g}_{\bar{\lambda}})) \\ &\leq O_p\left(n^{-1/(2+\alpha)}\right) \left(I^{\alpha/(2+\alpha)}(g^*) + I(g^*)\right) \end{aligned}$$

Next bound  $\|\hat{g}_{\bar{\lambda}} - \hat{g}_{\hat{\lambda}}\|_V$ . The basic inequality gives us

$$\|\hat{g}_{\bar{\lambda}} - \hat{g}_{\hat{\lambda}}\|_V^2 \leq 2 \left| \langle \epsilon, \hat{g}_{\bar{\lambda}} - \hat{g}_{\hat{\lambda}} \rangle_V \right| + 2 \left| \langle g^* - \hat{g}_{\bar{\lambda}}, \hat{g}_{\bar{\lambda}} - \hat{g}_{\hat{\lambda}} \rangle_V \right|$$

**Case a:**  $\left| \langle \epsilon, \hat{g}_{\bar{\lambda}} - \hat{g}_{\hat{\lambda}} \rangle_T \right|$  is the bigger term on the RHS

By Vandegeer (10.6),

$$\|\hat{g}_{\bar{\lambda}} - \hat{g}_{\hat{\lambda}}\|_V^2 \leq O_P(n^{-1/2}) \|\hat{g}_{\bar{\lambda}} - \hat{g}_{\hat{\lambda}}\|^{1-\alpha/2} (I(\hat{g}_{\bar{\lambda}}) + I(\hat{g}_{\hat{\lambda}}))^{\alpha/2}$$

If  $I(\hat{g}_{\bar{\lambda}}) > I(\hat{g}_{\hat{\lambda}})$ , then

$$\|\hat{g}_{\bar{\lambda}} - \hat{g}_{\hat{\lambda}}\|_V \leq O_P(n^{-1/(2+\alpha)}) I(g^*)^{\alpha/(2+\alpha)}$$

Otherwise, suppose  $I(\hat{g}_{\bar{\lambda}}) < I(\hat{g}_{\hat{\lambda}})$ . Since  $I$  is a pseudo-norm,

$$\begin{aligned} \|\hat{g}_{\bar{\lambda}} - \hat{g}_{\hat{\lambda}}\|_V &\leq O_P(n^{-1/(2+\alpha)}) I(\hat{g}_{\hat{\lambda}})^{\alpha/(2+\alpha)} \\ &\leq O_P(n^{-1/(2+\alpha)}) (I(\hat{g}_{\bar{\lambda}} - \hat{g}_{\hat{\lambda}}) + I(\hat{g}_{\hat{\lambda}}))^{\alpha/(2+\alpha)} \end{aligned}$$

If  $I(\hat{g}_{\bar{\lambda}} - \hat{g}_{\hat{\lambda}}) \leq I(\hat{g}_{\hat{\lambda}})$ , then we're done. Otherwise if  $I(\hat{g}_{\bar{\lambda}} - \hat{g}_{\hat{\lambda}}) \geq I(\hat{g}_{\hat{\lambda}})$ , by the assumption that  $I^V(\cdot)$  is bounded by the L2 norm,

$$\|\hat{g}_{\bar{\lambda}} - \hat{g}_{\hat{\lambda}}\|_V \leq O_P(n^{-1/(2+\alpha)}) (M \|\hat{g}_{\bar{\lambda}} - \hat{g}_{\hat{\lambda}}\|_V^2 + M_0)^{\alpha/v(2+\alpha)}$$

If  $M_0$  is bigger, we're done. Otherwise,

$$\|\hat{g}_{\bar{\lambda}} - \hat{g}_{\hat{\lambda}}\|_V \leq O_P(n^{-v/(2v-2\alpha+\alpha v)}) < O_P(n^{-1/(2+\alpha)})$$

**Case b:**  $\left| \langle g^* - \hat{g}_{\bar{\lambda}}, \hat{g}_{\bar{\lambda}} - \hat{g}_{\hat{\lambda}} \rangle_V \right|$  is the bigger term on the RHS

By Cauchy Schwarz,

$$\|\hat{g}_{\bar{\lambda}} - \hat{g}_{\hat{\lambda}}\|_V \leq O_P(1) \|g^* - \hat{g}_{\bar{\lambda}}\|_V$$

## 2 General Additive Model

### Definitions

We find the best model for  $y$  over function classes  $\mathcal{G} = \left\{ \sum_{j=1}^J g_j : g_j \in \mathcal{G}_j \right\}$ . Suppose we observe:

$$y = \sum_{j=1}^J g_j^* + \epsilon$$

where  $\sum_{j=1}^J g_j^* \in \mathcal{G}$ . Suppose  $\epsilon$  are sub-Gaussian errors for constants  $K$  and  $\sigma_0^2$ :

$$\max_{i=1:n} K^2 (E [\exp(|\epsilon_i|^2 K^2) - 1]) \leq \sigma_0^2$$

Given a training set  $T$ , we fit models by least squares with multiple penalties

$$\{\hat{g}_{\lambda,j}\}_{j=1}^J = \arg \min_{\sum g_j \in \mathcal{G}} \|y - \sum_{j=1}^J g_j\|_T^2 + \lambda^2 \sum_{j=1}^J I_j^{v_j}(g_j)$$

Given a validation set  $V$ , let the CV-fitted model be

$$\{\hat{g}_{\hat{\lambda},j}\}_{j=1}^J = \arg \min_{\lambda} \|y - \sum_{j=1}^J \hat{g}_{\lambda,j}\|_V^2$$

### Reasonable assumption:

- The entropy bound (2) in result 2 comes from the assumptions in Lemma 3. The  $\alpha$  below is  $\alpha = \max_{j=1:J} \{\alpha_j\}$ , so convergence is only as fast as fitting the highest-entropy function class. The constant  $A$  must be appropriately inflated such that the entropy bound holds for all  $\delta \in (0, R]$ .

### “Special” assumptions:

- We assume exponents  $v_j = 1$ , whereas Vandegeer Thrm 10.2 only assumes  $v > 2\alpha/(2 + \alpha)$ . Without this assumption, I wasn’t able to form inequalities between  $\sum_{j=1}^J I_j(g_j) \leq \text{something} + \sum_{j=1}^J I_j^{v_j}(g_j)$ . Indeed, Remark 1 in “High-dimensional Additive Modeling” (Vandegeer 2009) notes the importance of using the semi-norm instead of the square of the semi-norm.
- We suppose the following incoherence condition, in the spirit of Vandegeer 2014 “The additive model with different smoothness for the components”: Let  $p_V(\vec{x})$  be the empirical density over the validation set. Let  $p_{V_j}$  be the marginal density of  $x_j$  for the empirical distribution of the validation set. Let

$$r_V(\vec{x}) = \frac{p_V(\vec{x})}{\prod_{j=1}^J p_{V_j}(x_j)}, \quad \gamma_V^2 = \int r_V(\vec{x}) \prod_{j=1}^J p_{V_j}(x_j) d\mu$$

Suppose that  $\gamma_V < 1/(J - 1)$ . Furthermore, we will suppose that  $\int g_j p_{V_j} d\mu = 0$  for  $j = 2, \dots, J$ .

### Result 2: Additive Model with multiple penalties, Single oracle $\lambda$ over $X_T$

Suppose there is some  $0 < \alpha < 2$  s.t. for all  $\delta \in (0, R]$ ,

$$H \left( \delta, \left\{ \frac{\sum_{j=1}^J g_j - g_j^*}{\sum_{j=1}^J I_j(g_j) + I_j(g_j^*)} : g_j \in \mathcal{G}_j, \sum_{j=1}^J I_j(g_j) + I_j(g_j^*) > 0 \right\}, \|\cdot\|_T \right) \leq A\delta^{-\alpha} \quad (2)$$

If  $\lambda$  is chosen s.t.

$$\tilde{\lambda}_T^{-1} = O_p \left( n^{1/(2+\alpha)} \right) \left( \sum_{j=1}^J I_j(g_j^*) \right)^{(2-\alpha)/2(2+\alpha)}$$

then

$$\left\| \sum_{j=1}^J \hat{g}_j - g_j^* \right\|_T = O_p \left( \tilde{\lambda}_T \right) \left( \sum_{j=1}^J I_j(g_j^*) \right)^{1/2}$$

and

$$\sum_{j=1}^J I_j(\hat{g}_j) = O_p(1) \sum_{j=1}^J I_j(g_j^*)$$

**Proof:**

The basic inequality gives us:

$$\left\| \sum_{j=1}^J \hat{g}_j - g_j^* \right\|_T^2 + \lambda^2 \sum_{j=1}^J I_j(\hat{g}_j) \leq 2 \left| \left( \epsilon_T, \sum_{j=1}^J \hat{g}_j - g_j^* \right) \right| + \lambda^2 \sum_{j=1}^J I_j(g_j^*)$$

**Case 1:**  $\left| \left( \epsilon_T, \sum_{j=1}^J \hat{g}_j - g_j^* \right) \right| \leq \lambda^2 \sum_{j=1}^J I_j(g_j^*)$

$$\left\| \sum_{j=1}^J \hat{g}_j - g_j^* \right\|_T \leq O_p(\lambda) \left( \sum_{j=1}^J I_j(g_j^*) \right)^{1/2}$$

**Case 2:**  $\left| \left( \epsilon_T, \sum_{j=1}^J \hat{g}_j - g_j^* \right) \right| \geq \lambda^2 \sum_{j=1}^J I_j(g_j^*)$

By Vandegeer (10.6), the basic inequality becomes

$$\left\| \sum_{j=1}^J \hat{g}_j - g_j^* \right\|_T^2 + \lambda^2 \sum_{j=1}^J I_j(\hat{g}_j) \leq O_p \left( n^{-1/2} \right) \left\| \sum_{j=1}^J \hat{g}_j - g_j^* \right\|_T^{1-\alpha/2} \left( \sum_{j=1}^J I_j(\hat{g}_j) + I_j(g_j^*) \right)^{\alpha/2}$$

**Case 2a:**  $\sum_{j=1}^J I_j(\hat{g}_j) \leq \sum_{j=1}^J I_j(g_j^*)$

Then

$$\left\| \sum_{j=1}^J \hat{g}_j - g_j^* \right\|_T \leq O_p \left( n^{-1/(2+\alpha)} \right) \left( \sum_{j=1}^J I_j(g_j^*) \right)^{\alpha/(2+\alpha)}$$

**Case 2b:**  $\sum_{j=1}^J I_j(\hat{g}_j) \geq \sum_{j=1}^J I_j(g_j^*)$

Then

$$\sum_{j=1}^J I_j(\hat{g}_j) \leq O_p \left( n^{-1/(2-\alpha)} \right) \lambda^{-4/(2-\alpha)} \left\| \sum_{j=1}^J \hat{g}_j - g_j^* \right\|_T$$

Hence

$$\left\| \sum_{j=1}^J \hat{g}_j - g_j^* \right\|_T \leq O_p \left( n^{-1/(2-\alpha)} \right) \lambda^{-2\alpha/(2-\alpha)}$$

**Result 3: Additive Model with multiple penalties, Single cross-validation  $\lambda$  over general  $X_T, X_V$**

Suppose that the training and validation set are independently sampled, so the values  $X_i$  are not necessarily the same. Suppose the training and validation sets are both of size  $n$ . Suppose  $X$  is bounded s.t.  $|X| \leq R_X$  and the domain of  $g \in \mathcal{G}$  is over  $(-R_X, R_X)$ .

Suppose the same entropy bound (2) for both the training set  $P_T$  and validation set  $P_V$ .

In addition to the assumptions in Result 4, suppose the infinity norm is also bounded

$$\sup_{g_j \in \mathcal{G}_j} \frac{\|\sum_{j=1}^J g_j - g_j^*\|_\infty}{\sum_{j=1}^J I_j(g_j) + I_j(g_j^*)} \leq K < \infty$$

Suppose there exist constants  $M, M_0$  s.t. for all  $j$  and all  $\lambda \in \Lambda$

$$I_j(\hat{g}_{\lambda,j}) \leq M \|\hat{g}_{\lambda,j}\|_V^2 + M_0$$

**Special assumption:** Suppose the incoherence condition  $\gamma_V < 1/(J-1)$ . We will also suppose  $\int g_j p_{V,j} d\mu = 0$  for  $j = 2, \dots, J$ .

Let  $\hat{\lambda}$  be the optimal  $\lambda$  as specified in Result 2. Then

$$\left\| \sum_{j=1}^J \hat{g}_{\hat{\lambda},j} - \hat{g}_{\hat{\lambda},j} \right\|_V = O_p \left( n^{-1/(2+\alpha)} \right) (1 - \gamma(J-1))^{\alpha/(2+\alpha)} \left( \left( \sum_{j=1}^J I_j(g_j^*) \right)^{\alpha/(2+\alpha)} + \sum_{j=1}^J I_j(g_j^*) + \left\| \sum_{j=1}^J g_j^* \right\|_V^{\alpha/2(2+\alpha)} \right)$$

and  $\left\| \sum_{j=1}^J g_j^* - \hat{g}_{\hat{\lambda},j} \right\|_V$  is on the same order (differs by a constant).

**Proof:**

By the triangle inequality,

$$\left\| \sum_{j=1}^J g_j^* - \hat{g}_{\hat{\lambda},j} \right\|_V \leq \left\| \sum_{j=1}^J \hat{g}_{\hat{\lambda},j} - \hat{g}_{\hat{\lambda},j} \right\|_V + \left\| \sum_{j=1}^J g_j^* - \hat{g}_{\hat{\lambda},j} \right\|_V$$

By Lemma 2 and Result 2, we can easily bound  $\left\| \sum_{j=1}^J g_j^* - \hat{g}_{\hat{\lambda},j} \right\|_V$ .

$$\begin{aligned} \left\| \sum_{j=1}^J g_j^* - \hat{g}_{\hat{\lambda},j} \right\|_V &\leq \left\| \sum_{j=1}^J g_j^* - \hat{g}_{\hat{\lambda},j} \right\|_T + \left| \left\| \sum_{j=1}^J g_j^* - \hat{g}_{\hat{\lambda},j} \right\|_T - \left\| \sum_{j=1}^J g_j^* - \hat{g}_{\hat{\lambda},j} \right\|_V \right| \\ &\leq O_p \left( n^{-1/(2+\alpha)} \right) \left( \left( \sum_{j=1}^J I_j(g_j^*) \right)^{\alpha/(2+\alpha)} + \sum_{j=1}^J I_j(g_j^*) \right) \end{aligned}$$

Next bound  $\left\| \sum_{j=1}^J \hat{g}_{\hat{\lambda},j} - \hat{g}_{\hat{\lambda},j} \right\|_V$ . By definition of  $\hat{\lambda}$ , we have the basic inequality

$$\left\| \sum_{j=1}^J \hat{g}_{\hat{\lambda},j} - \hat{g}_{\hat{\lambda},j} \right\|_V^2 \leq 2 \left| \left( \epsilon, \sum_{j=1}^J \hat{g}_{\hat{\lambda},j} - \hat{g}_{\hat{\lambda},j} \right)_V \right| + 2 \left| \left( \sum_{j=1}^J g_j^* - \hat{g}_{\hat{\lambda},j}, \sum_{j=1}^J \hat{g}_{\hat{\lambda},j} - \hat{g}_{\hat{\lambda},j} \right)_V \right|$$

**Case 1:**  $\left| \left( \epsilon, \sum_{j=1}^J \hat{g}_{\hat{\lambda},j} - \hat{g}_{\hat{\lambda},j} \right)_V \right|$  is bigger

By Vandegeer (10.6),

$$\left\| \sum_{j=1}^J \hat{g}_{\tilde{\lambda},j} - \hat{g}_{\hat{\lambda},j} \right\|_V^{1+\alpha/2} \leq O_p(n^{-1/2}) \left( \sum_{j=1}^J I_j(\hat{g}_{\tilde{\lambda},j}) + I_j(\hat{g}_{\hat{\lambda},j}) \right)^{\alpha/2}$$

If  $\sum_{j=1}^J I_j(\hat{g}_{\tilde{\lambda},j}) \geq \sum_{j=1}^J I_j(\hat{g}_{\hat{\lambda},j})$ , we're done.

Otherwise, suppose  $\sum_{j=1}^J I_j(\hat{g}_{\tilde{\lambda},j}) < \sum_{j=1}^J I_j(\hat{g}_{\hat{\lambda},j})$ .

Since all the penalties are bounded by the L2 norm,

$$\begin{aligned} \sum_{j=1}^J I_j(\hat{g}_{\tilde{\lambda},j}) &\leq M \sum_{j=1}^J \|\hat{g}_{\tilde{\lambda},j}\|_V^2 + M_0 J \\ &\leq M(1 - \gamma(J-1)) \left\| \sum_{j=1}^J \hat{g}_{\tilde{\lambda},j} \right\|_V^2 + M_0 J \end{aligned}$$

where the latter inequality is due to the incoherence assumption and Lemma 4.

Then

$$\left\| \sum_{j=1}^J \hat{g}_{\tilde{\lambda},j} - \hat{g}_{\hat{\lambda},j} \right\|_V^{1+\alpha/2} \leq O_p(n^{-1/2}) \left( M(1 - \gamma(J-1)) \left\| \sum_{j=1}^J \hat{g}_{\tilde{\lambda},j} \right\|_V^2 + M_0 J \right)^{\alpha/2}$$

If  $M_0 J$  is the biggest, we're done. Otherwise,

$$\begin{aligned} \left\| \sum_{j=1}^J \hat{g}_{\tilde{\lambda},j} - \hat{g}_{\hat{\lambda},j} \right\|_V^{1+\alpha/2} &\leq O_p(n^{-1/2}) (1 - \gamma(J-1))^{\alpha/2} \left\| \sum_{j=1}^J \hat{g}_{\tilde{\lambda},j} \right\|_V^\alpha \\ &\leq O_p(n^{-1/2}) (1 - \gamma(J-1))^{\alpha/2} \left( \left\| \sum_{j=1}^J \hat{g}_{\tilde{\lambda},j} - \hat{g}_{\hat{\lambda},j} \right\|_V + \left\| \sum_{j=1}^J \hat{g}_{\tilde{\lambda},j} - g_j^* \right\|_V + \left\| \sum_{j=1}^J g_j^* \right\|_V \right)^\alpha \end{aligned}$$

If  $\left\| \sum_{j=1}^J \hat{g}_{\tilde{\lambda},j} - \hat{g}_{\hat{\lambda},j} \right\|_V$  or  $\left\| \sum_{j=1}^J \hat{g}_{\tilde{\lambda},j} - g_j^* \right\|_V$  is the biggest on the RHS, then the rate is faster than  $O_p(n^{-1/(2+\alpha)})$ . If  $\left\| \sum_{j=1}^J g_j^* \right\|_V$  is the biggest, then

$$\left\| \sum_{j=1}^J \hat{g}_{\tilde{\lambda},j} - \hat{g}_{\hat{\lambda},j} \right\|_V \leq O_p(n^{-1/(2+\alpha)}) \left\| \sum_{j=1}^J g_j^* \right\|_V^{\alpha/2(2+\alpha)}$$

**Case 2:**  $\left| \left( \sum_{j=1}^J g_j^* - \hat{g}_{\hat{\lambda},j}, \sum_{j=1}^J \hat{g}_{\tilde{\lambda},j} - \hat{g}_{\hat{\lambda},j} \right)_V \right|$  is bigger

By Cauchy Schwarz,

$$\left\| \sum_{j=1}^J \hat{g}_{\tilde{\lambda},j} - \hat{g}_{\hat{\lambda},j} \right\|_V \leq O_p(1) \left\| \sum_{j=1}^J g_j^* - \hat{g}_{\tilde{\lambda},j} \right\|_V$$

### 3 General Additive Model: Multiple Lambdas

#### Definitions

We find the best model for  $y$  over function classes  $\mathcal{G} = \left\{ \sum_{j=1}^J g_j : g_j \in \mathcal{G}_j \right\}$ . Suppose we observe:

$$y = \sum_{j=1}^J g_j^* + \epsilon$$

where  $\sum_{j=1}^J g_j^* \in \mathcal{G}$ . Suppose  $\epsilon$  are sub-Gaussian errors for constants  $K$  and  $\sigma_0^2$ :

$$\max_{i=1:n} K^2 \left( E \left[ \exp(|\epsilon_i|^2 K^2) - 1 \right] \right) \leq \sigma_0^2$$

Given a training set  $T$ , we fit models by least squares with multiple penalties and tuning parameters

$$\{\hat{g}_{\lambda,j}\}_{j=1}^J = \arg \min_{\sum g_j \in \mathcal{G}} \|y - \sum_{j=1}^J g_j\|_T^2 + \sum_{j=1}^J \lambda_j^2 I_j^{v_j}(g_j)$$

Suppose  $1 \leq v_j \leq 2$ .

Given a validation set  $V$ , let the CV-fitted model be

$$\{\hat{g}_{\hat{\lambda},j}\}_{j=1}^J = \arg \min_{\lambda} \|y - \sum_{j=1}^J \hat{g}_{\lambda,j}\|_V^2$$

#### Result 4: Additive Model, Oracle $\{\lambda_i\}$ given $X_T$

These results are implied by Vandegeer's paper "The additive model with different smoothness for the components."

Suppose for all  $j = 1 : J$

$$\mathcal{H} \left( \delta, \left\{ \frac{g_j - g_j^*}{I(g_j) + I(g_j^*)} \right\}, \|\cdot\|_n \right) \leq A_j \delta^{-\alpha_j} \forall \delta > 0$$

Let

$$\lambda_j = O_p(n^{-1/(2+\alpha_j)})$$

and

$$\left( \sum_{j=1}^J I_j^{q_j}(g_j^*) \right)^{1/2} \lambda_{\max} = O_P(1)R$$

There are some constants  $c_1, c_2$  s.t. for  $\lambda_j = O_p(n^{-1/(2+\alpha_j)})$ , we have

$$\left\| \sum g_j^* - \sum \hat{g}_{\hat{\lambda},j} \right\| \leq c_2 \lambda_{(j)}$$

where  $(j) = \arg \max \alpha_j$ . That is, the convergence rate depends on the highest-entropy function class (with respect to the penalty)

$$\left\| \sum_{j=1}^J g_j^* - \sum_{j=1}^J \hat{g}_j \right\|_T = O_p(n^{-1/(2+\alpha_{(j)})})$$

**Jean's version of the Proof for Vandegeer Thrm 3.1:**

Suppose for some constant  $R$ , we define the function class

$$\mathcal{M}(R) = \left\{ \{g_j\} : (\lambda_j/R)^{(1-q_j)/q_j} \lambda_j I_j(g_j - g_j^*) \leq R, \left\| \sum_{j=1}^J g_j - g_j^* \right\|_T \leq R \right\}$$

Recall that

$$\sup_{g_j \in \mathcal{G}_j} \frac{|(\epsilon^T, g_j - g_j^*)|}{(I_j(g_j) + I_j(g_j^*))^{\alpha_j/2} \|g_j - g_j^*\|^{1-\alpha_j/2}} = O_p(n^{-1/2})$$

By our choice of  $\lambda$ , we have that for function sets  $\{g_j - g_j^*\} \in \mathcal{M}(R)$ , the empirical process term decreases with  $n$ :

$$\begin{aligned} |(\epsilon^T, g_j - g_j^*)| &\leq O_P(n^{-1/2}) (I_j(g_j) + I_j(g_j^*))^{\alpha_j/2} \|g_j - g_j^*\|^{1-\alpha_j/2} \\ &\leq O_P(n^{-1/2}) \left( \lambda_j^{-1/q_j} R^{1/q_j} \right)^{\alpha_j/2} R^{1-\alpha_j/2} \\ &\leq O_P(n^{-1/(2+\alpha_j)}) R^2 \end{aligned}$$

Hence for sufficiently large  $n$ , Vandegeer Lemma's 5.4 (Jean's version below) states that the fitted functions  $\hat{g}_j$  are also within  $R$  of the truth:

$$\{\hat{g}_j - g_j^*\} \in \mathcal{M}(R) \implies \left\| \sum_{j=1}^J g_j^* - \hat{g}_j \right\|_T \leq R$$

Now we just need to determine the right value for  $R$ . Choose  $n$  sufficiently large s.t. the penalty term for function  $(j)$  is the highest (for the truth)

$$\lambda_j^2 I_j^{q_j}(g_j^*) \leq \lambda_{(j)}^2 I_{(j)}^{q_{(j)}}(g_{(j)}^*) \quad \forall j$$

Then choose  $R$  s.t.

$$\left( \lambda_{(j)}^2 I_{(j)}^{q_{(j)}}(g_{(j)}^*) \right)^{1/2} J^{1/2} = O_P(1) R$$

Hence

$$\left\| \sum_{j=1}^J g_j^* - \hat{g}_j \right\|_T \leq n^{-1/(2+\alpha_{(j)})} J^{1/2} I_{(j)}^{q_{(j)}/2}(g_{(j)}^*)$$

## Result 5: Additive Model, Cross-validated $\{\lambda_i\}$ over general $X_T, X_V$

Assume the same conditions as result 4, but also for the validation set.

**Condition 2.4:** Incoherence condition on the validation set. Let  $p_V(\vec{x})$  be the empirical density over the validation set. Let  $p_{Vj}$  be the marginal density of  $x_j$  for the empirical distribution of the validation set. Let

$$r_V(\vec{x}) = \frac{p_V(\vec{x})}{\prod_{j=1}^J p_{Vj}(x_j)}, \quad \gamma_V^2 = \int r_V(\vec{x}) \prod_{j=1}^J p_{Vj}(x_j) d\mu$$

Suppose that  $\gamma_V < 1/(J-1)$ . Furthermore, we will suppose that  $\int g_j p_{Vj} d\mu = 0$  for  $j = 2, \dots, J$ .

Additionally, suppose there exist constants  $M, M_0$  s.t. for all  $j$  and all  $\lambda \in \Lambda$

$$I_j(\hat{g}_{\lambda,j}) \leq M \|\hat{g}_{\lambda,j}\|_V^2 + M_0$$

Let  $\tilde{\lambda}$  be the optimal  $\{\lambda_i\}$  as specified in Result 4. Then



$$\left\| \sum_{j=1}^J \hat{g}_{\hat{\lambda},j} - \hat{g}_{\tilde{\lambda},j} \right\|_V = O_p \left( n^{-1/(2+\alpha_{(j)})} \right) (1 - \gamma(J-1))^{\alpha_{(j)}/(2+\alpha_{(j)})} \left( \left( \sum_{j=1}^J I_j(g_j^*) \right)^{\alpha_{(j)}/(2+\alpha_{(j)})} + \sum_{j=1}^J I_j(g_j^*) + \left\| \sum_{j=1}^J g_j^* \right\|_V^{\alpha_{(j)}/2} \right)$$

and  $\left\| \sum_{j=1}^J g_j^* - \hat{g}_{\hat{\lambda},j} \right\|_V = O_p(1) \left\| \sum_{j=1}^J g_j^* - \hat{g}_{\tilde{\lambda},j} \right\|_T$ .

**Proof:**

Exactly the same as Result 3

## Lemmas

### Lemma 1:

Suppose for all  $\lambda \in \Lambda$ , the penalty function  $I^v(g_\lambda)$  is upper-bounded by  $\|g_\lambda\|_n^2 = \frac{1}{n} \sum_{i=1}^n g_\lambda^2(x_i)$  with constants  $M_0$  and  $M$ :

$$I^v(g_\lambda) \leq M\|g_\lambda\|_n^2 + M_0$$

Suppose there is some function  $g \in \mathcal{G}$  such that

$$\|g - g_\lambda\|_n^{1+\alpha/2} \leq O_p(n^{-1/2})I^{\alpha/2}(g_\lambda)$$

Then

$$\|g - g_\lambda\|_n \leq O_p(n^{-1/(2+\alpha)})M^{\alpha v/(2+\alpha)}\|g\|_n^{2\alpha/v(2+\alpha)}$$

### Proof:

From the assumptions, we have

$$\|g - g_\lambda\|_n^{1+\alpha/2} \leq O_p(n^{-1/2}) (M\|g_\lambda\|_n^2 + M_0)^{\alpha/2v}$$

If  $M_0 > \|g_\lambda\|_n^2$ , we're done. Otherwise,

$$\begin{aligned} \|g - g_\lambda\|_n^{1+\alpha/2} &\leq O_p(n^{-1/2})M^{\alpha/2v}\|g_\lambda\|_n^{\alpha/v} \\ &\leq O_p(n^{-1/2})M^{\alpha/2v}(\|g_\lambda - g\|_n + \|g\|_n)^{\alpha/v} \end{aligned}$$

**Case 1:**  $\|g_\lambda - g\|_n \geq \|g\|_n$

Then

$$\|g - g_\lambda\|_n \leq O_p(n^{-v/(2v+\alpha v-2\alpha)})M^{\alpha v^2/(2v+\alpha v-2\alpha)}$$

Note that  $\sup_v -\frac{v}{2v+\alpha v-2\alpha} = -\frac{1}{2+\alpha}$ , so this rate is faster than  $O_p(n^{-\frac{1}{2+\alpha}})$ .

**Case 2:**  $\|g_\lambda - g\|_n \leq \|g\|_n$

Then

$$\|g - g_\lambda\|_n \leq O_p(n^{-1/(2+\alpha)})M^{\alpha v/(2+\alpha)}\|g\|_n^{2\alpha/v(2+\alpha)}$$

I believe we can often provide a good estimate of  $M$  for the entire class  $\mathcal{G}$ , which means that we can always estimate the sample size needed to ensure this case never occurs. That is, I believe we can often estimate  $M$  s.t.

$$I^v(g) \leq M\|g\|_n^2 + M_0 \forall g \in \mathcal{G}$$

### Lemma 2:

Let  $P_{n'}$  and  $P_{n''}$  be empirical distributions over  $\{X_i'\}_{i=1}^n, \{X_i''\}_{i=1}^n$ . Let  $P_{2n} = \frac{1}{2}(P_{n'} + P_{n''})$ . Suppose  $X$  is bounded s.t.  $|X| < R_X$ .

Let  $\mathcal{G}' = \left\{ \frac{g-g^*}{I(g)+I(g^*)} : g \in \mathcal{G}, I(g) + I(g^*) > 0 \right\}$ . Suppose  $g$  is defined over the domain over  $X$  (and zero otherwise). Suppose

$$\sup_{f \in \mathcal{G}'} \|f\|_{P_{2n}} \leq R < \infty, \quad \sup_{f \in \mathcal{G}'} \|f\|_\infty \leq K < \infty$$

and

$$H(\delta, \mathcal{G}', P_{n'}) \leq \tilde{A}\delta^{-\alpha}, \quad H(\delta, \mathcal{G}', P_{n''}) \leq \tilde{A}\delta^{-\alpha}$$

Then

$$Pr \left( \sup_{g \in \mathcal{G}} \frac{|\|g^* - g\|_{P_{n'}} - \|g^* - g\|_{P_{n''}}|}{I(g^*) + I(g)} \geq 6\delta \right) \leq 2 \exp \left( 2\tilde{A}\delta^{-\alpha} - \frac{4\delta^2 n}{K^2} \right)$$

**Proof:** The proof is very similar to that in Pollard 1984 (page 32), so some details below are omitted. First note that for any function  $f$  and  $h$ , we have

$$\|f\|_{P_{n'}} - \|h\|_{P_{n'}} \leq \|f - h\|_{P_{n'}} \leq \sqrt{2}\|f - h\|_{P_{2n}}$$

Similarly for  $P_{n''}$ .

Let  $\{h_j\}_{j=1}^N$  be the  $\sqrt{2}\delta$ -cover for  $\mathcal{G}'$  (where  $N = N(\sqrt{2}\delta, \mathcal{G}', P_{2n})$ ). Let  $h_j$  be the closest function (in terms of  $\|\cdot\|_{P_{2n}}$ ) to some  $f \in \mathcal{G}'$ . Then

$$\begin{aligned} \|f\|_{P_{n'}} - \|f\|_{P_{n''}} &\leq \|f - h_j\|_{P_{n'}} + |\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}| + \|f - h_j\|_{P_{n''}} \\ &\leq 4\delta + |\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}| \end{aligned}$$

Therefore for  $f = \frac{g^* - g}{I(g^*) + I(g)}$ , we have

$$\begin{aligned} Pr \left( \sup_{g \in \mathcal{G}} \frac{|\|g^* - g\|_{P_n} - \|g^* - g\|_{P_{n''}}|}{I(g^*) + I(g)} \geq 6\delta \right) &\leq Pr \left( \sup_{j \in 1:N} |\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}| \geq 2\delta \right) \\ &\leq N \max_{j \in 1:N} Pr \left( |\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}| \geq 2\delta \right) \end{aligned}$$

Now note that

$$\begin{aligned} |\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}| &= \frac{|\|h_j\|_{P_{n'}}^2 - \|h_j\|_{P_{n''}}^2|}{\|h_j\|_{P_{n'}} + \|h_j\|_{P_{n''}}} \\ &\leq \frac{|\|h_j\|_{P_{n'}}^2 - \|h_j\|_{P_{n''}}^2|}{\sqrt{2}\|h_j\|_{P_{2n}}} \end{aligned}$$

By Hoeffding's inequality,

$$\begin{aligned} Pr \left( |\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}| \geq 2\delta \right) &\leq Pr \left( |\|h_j\|_{P_{n'}}^2 - \|h_j\|_{P_{n''}}^2| \geq 2\sqrt{2}\delta\|h_j\|_{P_{2n}} \right) \\ &= Pr \left( \left| \sum_{i=1}^n W_i (h_j^2(x'_i) - h_j^2(x''_i)) \right| \geq 2\sqrt{2}n\delta\|h_j\|_{P_{2n}} \right) \\ &\leq 2 \exp \left( - \frac{16\delta^2 n^2 \|h_j\|_{P_{2n}}^2}{4 \sum_{i=1}^n (h_j^2(x'_i) - h_j^2(x''_i))^2} \right) \end{aligned}$$

Since  $\|h_j\|_\infty < K$ , then

$$\begin{aligned} \sum_{i=1}^n (h_j^2(x'_i) - h_j^2(x''_i))^2 &\leq \sum_{i=1}^n h_j^4(x'_i) + h_j^4(x''_i) \\ &\leq nK^2 \|h_j\|_{P_{2n}}^2 \end{aligned}$$

Hence

$$Pr \left( |\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}| \geq 2\delta \right) \leq 2 \exp \left( - \frac{4\delta^2 n}{K^2} \right)$$

Since (Pollard and Vandegeer say that)

$$N(\sqrt{2}\delta, \mathcal{G}', P_{2n}) \leq N(\delta, \mathcal{G}', P_{n''}) + N(\delta, \mathcal{G}', P_{n''})$$

then

$$Pr \left( \sup_{g \in \mathcal{G}} \frac{|\|g^* - g\|_{P_n} - \|g^* - g\|_{P_{n''}}|}{I(g^*) + I(g)} \geq 6\delta \right) \leq 2 \exp \left( 2\tilde{A}\delta^{-\alpha} - \frac{4\delta^2 n}{K^2} \right)$$

Using shorthand, we can write

$$\sup_{g \in \mathcal{G}} \frac{|\|g^* - g\|_{P_n} - \|g^* - g\|_{P_{n''}}|}{I(g^*) + I(g)} = O_p(n^{-1/(2+\alpha)})$$

**Lemma 3:**

Suppose the function classes  $\mathcal{F}_j$  is a cone and  $I_j : \mathcal{F}_j \mapsto [0, \infty)$  is a psuedonorm. Furthermore, suppose

$$H(\delta, \{f_j \in \mathcal{F}_j : I_j(f_j) \leq 1\}, \|\cdot\|_n) \leq A_j \delta^{-\alpha_j}$$

Then if  $f_j^* \in \mathcal{F}_j$ , then

$$H \left( \delta, \left\{ \frac{\sum_{j=1}^J f_j - f_j^*}{\sum_{j=1}^J I_j(f_j) + I_j(f_j^*)} : f_j \in \mathcal{F}_j, I_j(f_j) + I_j(f_j^*) > 0 \right\}, \|\cdot\|_n \right) \leq 2 \sum_{j=1}^J A_j \left( \frac{\delta}{2J} \right)^{-\alpha_j}$$

**Proof:** Let  $\tilde{f}_j = \frac{f_j}{\sum_{j=1}^J I_j(f_j) + I_j(f_j^*)}$ . Then  $\tilde{f}_j \in \mathcal{F}_j$  and  $I_j(\tilde{f}_j) \leq 1$ . Let  $h_{(j)}$  be the closest function to  $\tilde{f}_j$  in the  $\delta$  cover of  $\mathcal{F}_j$ . Similarly, let  $h_{(j)}^*$  be the closest function to  $\tilde{f}_j^*$  in the  $\delta$  cover of  $\mathcal{F}_j$ . Then

$$\begin{aligned} \left\| \frac{\sum_{j=1}^J f_j - f_j^*}{\sum_{j=1}^J I_j(f_j) + I_j(f_j^*)} - \left( \sum_{j=1}^J h_{(j)} - h_{(j)}^* \right) \right\| &\leq \sum_{j=1}^J \left\| \frac{f_j - f_j^*}{\sum_{j=1}^J I_j(f_j) + I_j(f_j^*)} - (h_{(j)} - h_{(j)}^*) \right\| \\ &\leq \sum_{j=1}^J \left\| \frac{f_j}{\sum_{j=1}^J I_j(f_j) + I_j(f_j^*)} - h_{(j)} \right\| + \left\| \frac{f_j^*}{\sum_{j=1}^J I_j(f_j) + I_j(f_j^*)} - h_{(j)}^* \right\| \\ &\leq 2J\delta \end{aligned}$$

Hence

$$H \left( 2J\delta, \left\{ \frac{\sum_{j=1}^J f_j - f_j^*}{\sum_{j=1}^J I_j(f_j) + I_j(f_j^*)} : f_j \in \mathcal{F}_j, I_j(f_j) + I_j(f_j^*) > 0 \right\}, \|\cdot\|_n \right) \leq 2 \sum_{j=1}^J A_j \delta^{-\alpha_j}$$

**Lemma 4:**

Let  $p_n(\vec{x})$  be some empirical density and let  $p_{nj}$  be the corresponding empirical marginal density of  $x_j$ . Let

$$r(\vec{x}) = \frac{p_n(\vec{x})}{\prod_{j=1}^J p_{nj}(x_j)}, \quad \gamma^2 = \int (r(\vec{x}) - 1)^2 \prod_{j=1}^J p_{nj}(x_j) d\mu$$

Suppose  $\gamma < 1/(J-1)$ . Furthermore, suppose  $\int g_j p_{nj} d\mu = 0$  for  $j = 2, \dots, J$ . Then

$$\left\| \sum_{j=1}^J g_j \right\|_n^2 \geq (1 - \gamma(J-1)) \left( \sum_{j=1}^J \|g_j\|_n^2 \right)$$

**Proof:** The proof is very similar to Lemma 5.1 in Vandegeer 2014 “The additive model with different smoothness for the components.”

$$\left\| \sum_{j=1}^J g_j \right\|_n^2 = \sum_{j=1}^J \|g_j\|_n^2 + \sum_{j \neq k} \int g_j g_k p_n(\vec{x}) d\mu$$

We bound the latter term:

$$\begin{aligned} \left| \int g_j g_k p_n(\vec{x}) d\mu \right| &= \left| \int g_j g_k (r(\vec{x}) - 1) \Pi_{j=1}^J p_{n_j}(x_j) d\mu \right| \\ &\leq \gamma \left| \int g_j^2 g_k^2 \Pi_{j=1}^J p_{n_j}(x_j) d\mu \right|^{1/2} \\ &= \gamma \|g_j\|_n \|g_k\|_n \end{aligned}$$

Hence

$$\begin{aligned} \left\| \sum_{j=1}^J g_j \right\|_n^2 &\geq \sum_{j=1}^J \|g_j\|_n^2 - \gamma \sum_{j \neq k} \|g_j\|_n \|g_k\|_n \\ &\geq (1 - \gamma(J-1)) \sum_{j=1}^J \|g_j\|_n^2 + \gamma \sum_{j < k} (\|g_j\|_n - \|g_k\|_n)^2 \\ &\geq (1 - \gamma(J-1)) \sum_{j=1}^J \|g_j\|_n^2 \end{aligned}$$

#### Vandegeer’s Lemma 5.4 (Jean’s version)

Let

$$\tau_R(\{f_j\}) = \left\| \sum f_j \right\|_T + \sum_{j=1}^J (\lambda_j/R)^{(1-q_j)/q_j} \lambda_j I_j(f_j)$$

Suppose

$$\sum_{j=1}^J \lambda_j^2 I_j^{q_j}(f_j^*) \leq \delta_0^2 R^2$$

and for all function sets  $\{f_j\}$  s.t.  $\tau_R(\{f_j\}) \leq R$ , suppose

$$\sup_{f_j} |(\epsilon_T, f_j)| \leq \delta_0^2 R^2$$

Let

$$\hat{f}_j = \arg \min \|y - \sum_{j=1}^J f_j\|_T^2 + \sum_{j=1}^J \lambda_j^2 I_j^{q_j}(f_j)$$

Then  $\tau_R(\{\hat{f}_{\lambda,j} - f_j^*\}) \leq R$ .

**Proof** We use the convexity of the penalties and the least squares function. Consider  $\tilde{f}_j = t\hat{f}_j + (1-t)f_j^*$  where

$$t = \frac{R}{R + \tau_R(\{\hat{f}_j - f_j^*\})}$$

First note that by convexity,

$$\tau_R(\{\tilde{f}_j - f_j^*\}) = \frac{R}{R + \tau_R(\{\hat{f}_j - f_j^*\})} \tau_R(\{\hat{f}_j - f_j^*\}) \leq R$$

Hence

$$\sup_{f_j} \left| \left( \epsilon_T, f_j^* - \tilde{f}_j \right) \right| \leq \delta_0^2 R^2$$

So by the basic inequality,

$$\left\| \sum_{j=1}^J f_j^* - \sum_{j=1}^J \tilde{f}_j \right\|_T^2 + \sum_{j=1}^J \lambda_j^2 I_j^{q_j}(\tilde{f}_j) \leq \sum_{j=1}^J \left| \left( \epsilon_T, f_j^* - \tilde{f}_j \right) \right| + \sum_{j=1}^J \lambda_j^2 I_j^{q_j}(f_j^*)$$

and with gross algebra, we can show that

$$(\lambda_j/R)^{(1-q_j)/q_j} \lambda_j I_j(\tilde{f}_j - f_j^*) \leq 4\delta_0 R$$

Then

$$\begin{aligned} \frac{R}{R + \tau_R(\{\hat{f}_j - f_j^*\})} \tau_R(\{\hat{f}_j - f_j^*\}) &= \tau_R(\{\tilde{f}_j - f_j^*\}) \\ &= \left\| \sum_{j=1}^J \tilde{f}_j - f_j^* \right\|_T + \sum_{j=1}^J (\lambda_j/R)^{(1-q_j)/q_j} \lambda_j I_j(\tilde{f}_j - f_j^*) \\ &\leq O_P(1) J \delta_0 R \end{aligned}$$

So for small enough  $\delta_0$ , we have

$$\tau_R(\{\hat{f}_j - f_j^*\}) \leq R$$

## 4 What if we can't bound the penalty?

Now suppose the problem does not satisfy the assumption that

$$I(g_\lambda) \leq M \|g_\lambda\|^2 + M_0$$

Then let's consider a modified way of choosing  $\lambda$ .  
Select  $\lambda$  s.t.

$$\hat{\lambda} = \arg \min_{\lambda \in \Lambda} \|y - g_\lambda\|_V^2 \text{ where } \Lambda = \{\lambda : n^{-\tau} I(\hat{g}_\lambda) \leq \|y - \hat{g}_\lambda\|_V\}$$

To get the optimal convergence rate for  $\|g^* - \hat{g}_\lambda\|_V$ , choose  $\tau = \frac{1}{2(1+\alpha)}$ . Then we get the rate

$$\|g^* - \hat{g}_\lambda\|_V = O_P(n^{-1/2(1+\alpha)})$$

**Proof:**

For sufficiently large  $n$ ,  $\tilde{\lambda}$  will be in the set  $\Lambda$  with high probability. To see this, note that

$$\begin{aligned} n^{-\tau} I(\hat{g}_{\tilde{\lambda}}) &= O_P(n^{-\tau}) I(g^*) \\ &\leq \|y - g^*\|_V - \|g^* - \hat{g}_{\tilde{\lambda}}\|_V \\ &\leq \|y - \hat{g}_{\tilde{\lambda}}\|_V \end{aligned}$$

where the first inequality comes from the fact that  $\|y - g^*\|_V = O_P(\sigma)$  with high probability and  $\|g^* - \hat{g}_{\tilde{\lambda}}\|_V = O_P(n^{-1/(2+\alpha)})$ .

Now proceed with the basic inequality. We know that

$$\|\hat{g}_{\tilde{\lambda}} - \hat{g}_\lambda\|_V^2 \leq 2 |(\epsilon, \hat{g}_{\tilde{\lambda}} - \hat{g}_\lambda)_V| + 2 |(g^* - \hat{g}_{\tilde{\lambda}}, \hat{g}_{\tilde{\lambda}} - \hat{g}_\lambda)_V|$$

The problematic case is when  $|(\epsilon, \hat{g}_{\tilde{\lambda}} - \hat{g}_\lambda)_V|$  is the bigger term on the RHS.

By Vandegeer (10.6),

$$\|\hat{g}_{\tilde{\lambda}} - \hat{g}_\lambda\|_V^2 \leq O_P(n^{-1/2}) \|\hat{g}_{\tilde{\lambda}} - \hat{g}_\lambda\|^{1-\alpha/2} (I(\hat{g}_{\tilde{\lambda}}) + I(\hat{g}_\lambda))^{\alpha/2}$$

We're done if  $I(\hat{g}_{\tilde{\lambda}}) > I(\hat{g}_\lambda)$ . Otherwise, suppose  $I(\hat{g}_{\tilde{\lambda}}) < I(\hat{g}_\lambda)$ . By definition of  $\Lambda$ , we have

$$\begin{aligned} \|\hat{g}_{\tilde{\lambda}} - \hat{g}_\lambda\|_V^{1+\alpha/2} &\leq O_P(n^{-1/2}) I(\hat{g}_{\tilde{\lambda}})^{\alpha/2} \\ &\leq O_P(n^{(-1+\tau\alpha)/2}) \|y - \hat{g}_{\tilde{\lambda}}\|_V^{\alpha/2} \\ &\leq O_P(n^{(-1+\tau\alpha)/2}) (\|y - g^*\|_V + \|\hat{g}_{\tilde{\lambda}} - g^*\|_V + \|\hat{g}_{\tilde{\lambda}} - \hat{g}_\lambda\|_V)^{\alpha/2} \end{aligned}$$

The slowest case is when  $\|y - g^*\|_V$  is the largest among the three terms. We have the rate

$$\|\hat{g}_{\tilde{\lambda}} - \hat{g}_\lambda\|_V \leq O_P(n^{(-1+\tau\alpha)/(2+\alpha)})$$

The optimal convergence rate is attained if we choose

$$-\tau = \frac{-1 + \tau\alpha}{2 + \alpha}$$

That is, we get

$$\tau = \frac{1}{2(1+\alpha)}$$

## 5 Examples

Our goal here is to show that the assumptions hold for various examples.

### 5.1 Penalties that are compatible with the L2 norm

Lasso, Fused Lasso, Generalized Lasso, Ridge, Elastic Net

To see this works for the Generalized lasso:

Let  $D$  be the fixed penalty matrix. Let  $D_{max}$  be its maximum eigenvalue. Suppose the smallest eigenvalue of  $X^T X$  stays away from zero. Then for some constants  $M_0$  and  $M$ ,

$$\begin{aligned}\|D\beta\|_1 &\leq D_{max}\|\beta\|_1 \\ &\leq D_{max}M\|X^T\beta\|_2^2 + M_0\end{aligned}$$

### 5.2 Sobolev Norm

Suppose  $\mathcal{G}$  is the class of smooth functions  $g : [0, 1] \mapsto \mathbb{R}$  s.t.  $I_{(k)}(g) = \sqrt{\int_0^1 g(t)^2 dt} + \sqrt{\int_0^1 g^{(k)}(t)^2 dt} < \infty$ .

$$\arg \min_{g \in \mathcal{G}} \|y - g(x_1)\|_T^2 + \lambda_g^2 I_{(k)}^2(g)$$

If we reformulate this using the Reproducing Kernel Hilbert space  $\mathcal{H}$ , the criterion becomes

$$\arg \min_{g_H \in \mathcal{H}, g_\perp \in \mathcal{H}^\perp} \|y - g_H(x_1) + g_\perp(x_1)\|_T^2 + \lambda_g^2 I_{(k)}^2(g_H)$$

If  $\Omega$  is the kernel for  $\mathcal{H}$  (with respect to this dataset), then

$$g_H = \Omega_n \alpha, \quad I_{(k)}^2(g_H) = \alpha^T \Omega_n \alpha$$

and for some other matrix  $\Sigma$ , we have

$$g_\perp = \Sigma \beta$$

**Assumption 1:** Show for some constant  $K$ ,

$$\frac{\|g\|_\infty}{I_{(k)}(g)} \leq K$$

**Proof:**

Let  $c$  be some value s.t.  $\|g\|_2 = g(c)$ . This must exist since  $g$  is continuous.

$$\begin{aligned}g(x) &\leq g(c) + \int_c^x g'(u) du \\ &\leq \|g\|_2 + \int_0^1 g^{(m)}(u) du \\ &\leq \|g\|_2 + \int_0^1 |g^{(m)}(u)| du \\ &\leq \|g\|_2 + C \sqrt{\int_0^1 |g^{(m)}(u)|^2 du} \\ &\leq CI_{(k)}(g)\end{aligned}$$

**Lemma:** The entropy is bounded for the fitted model:



$$\arg \min_{g_H \in \mathcal{H}, g_\perp \in \mathcal{H}^\perp} \|y - g_H(x_1) + g_\perp(x_1)\|_T^2 + \lambda_g^2 I_{(k)}^2(g_H)$$

Using the Hilbert Space reproducing kernel, we find that an equivalent optimization problem is

$$\hat{\alpha}_{\lambda^2}, \hat{\beta} = \arg \min_{\alpha, \beta} \|y - Y\beta - \Omega\alpha\|_T^2 + \lambda^2 \alpha^T \Omega \alpha$$

where  $Y^T \Omega = 0$ . Suppose the  $Y^T Y$  has its minimum eigenvalue bounded away from zero and the minimum eigenvalue of  $\Omega$  is on the order of  $O_P(n^{-\tau})$ .

Suppose the maximum  $\lambda \in \Lambda$  is 1. Then the entropy is

$$H(\delta, \{\Omega \hat{\alpha}_\lambda : \lambda \in \Lambda\}, \|\cdot\|_T) \leq \log \left( \frac{K}{\delta} \right) + (\tau + 1) \log n$$

where  $K$  is a constant s.t.  $\|y - Y\hat{\beta}\| \leq K$ .

**Proof:**

First note that we know that

$$\hat{\beta} = (Y^T Y)^{-1} Y^T y_T$$

Given that  $Y^T Y$  has a minimum eigenvalue value bounded away from zero, there is some  $K$  s.t.  $\|y - Y\hat{\beta}\| \leq K$ .

Given  $\hat{\beta}$ , we have

$$\hat{\alpha}_{\lambda^2} = \frac{1}{n} \left( \frac{1}{n} \Omega^T \Omega + \lambda^2 \Omega \right)^{-1} \Omega^T (y - Y\hat{\beta})$$

We know that  $\Omega$  is a PD matrix. We express  $\Omega = Q D Q^T$  where  $Q$  is orthogonal and  $D$  is a diagonal matrix with entries  $\omega_i$ . Then

$$\Omega \hat{\alpha}_{\lambda^2} = Q (D + n\lambda^2)^{-1} D Q^T (y - Y\hat{\beta})$$

Hence we can measure the difference between models fitted with different  $\lambda$  values:

$$\begin{aligned} \|\Omega \hat{\alpha}_{\lambda^2} - \Omega \hat{\alpha}_{\lambda^2 + \delta}\| &= \left\| Q \left( (D + n\lambda^2)^{-1} - (D + n\lambda^2 + n\delta)^{-1} \right) D Q^T (y - Y\hat{\beta}) \right\| \\ &\leq \max_{i=1:n} \left| \frac{n\delta\omega_i}{(\omega_i + n\lambda^2)(\omega_i + n\lambda^2 + n\delta)} \right| \|y - Y\hat{\beta}\| \\ &\leq \delta n \max_{i=1:n} \left| \frac{\omega_i}{(\omega_i + n\lambda^2)^2} \right| \|y - Y\hat{\beta}\| \\ &\leq \delta n \max_{i=1:n} |\omega_i^{-1}| \|y - Y\hat{\beta}\| \\ &= \delta n O_P(n^\tau) \|y - Y\hat{\beta}\| \end{aligned}$$

Suppose we have  $\lambda_{max}^2 = 1$ .

Hence

$$\|\Omega \hat{\alpha}_{\lambda^2} - \Omega \hat{\alpha}_{\lambda^2 + \delta}\| \leq \delta O_P(n^{\tau+1}) K$$

So we can define a  $\tilde{\delta}$ -covering set for  $\{\Omega \hat{\alpha}_{\lambda^2}\}$  using  $\{\Omega \hat{\alpha}_{i\delta} : i = 0, 1, \dots, \frac{1}{\tilde{\delta}}\}$  where  $\delta = \tilde{\delta} O_P(n^{-(\tau+1)})/K$ . So the covering set contains a total of  $\frac{K}{\tilde{\delta}} O_P(n^{\tau+1})$  functions. Hence the entropy is

$$\log \left( \frac{K}{\tilde{\delta}} \right) + (\tau + 1) \log n$$

**Corrollary:**

Now suppose we have a training and validation sets  $(y_T, X_T)$  and  $(y_V, X_V)$ . We rewrite the joint optimization problem with two kernel matrices  $\Omega_T$  and  $\Omega_V$

$$\arg \min_{\lambda} \|y - Y\hat{\beta} - \Omega_V \hat{\alpha}_{\lambda^2}\|_V^2$$

$$\hat{\alpha}_{\lambda^2}, \hat{\beta} = \arg \min_{\alpha, \beta} \|y - Y\beta - \Omega_T \alpha\|_T^2 + \lambda^2 \alpha^T \Omega_T \alpha$$

where  $Y^T \Omega_T = 0$  and  $Y^T \Omega_V = 0$ . Suppose the  $Y^T Y$  has its minimum eigenvalue bounded away from zero.

Let  $\omega_T^2$  and  $\omega_V^2$  be the eigenvalues of  $\Omega_V$  and  $\Omega_T$ . Suppose that

$$\frac{\max_{i=1:n} |\omega_{Vi}|}{\min_{i=1:n} |\omega_{Ti}|^2} = O_p(n^\tau)$$

Suppose the maximum  $\lambda \in \Lambda$  is 1.

Then the entropy is

$$H(\delta, \{\Omega_V \hat{\alpha}_{\lambda} : \lambda \in \Lambda\}, \|\cdot\|_T) \leq \log \left( \frac{K}{\delta} \right) + (\tau + 1) \log n$$

where  $K$  is a constant s.t.  $\|y - Y\hat{\beta}\| \leq K$ .

**Proof:**

Proof structure is similar to above.

$$\begin{aligned} \Omega_V \hat{\alpha}_{\lambda^2} &= \Omega_V \Omega^{-1} (\Omega + n\lambda^2 I)^{-1} \Omega^T (y - Y\hat{\beta}) \\ &= \Omega_V Q_T D_T^{-1} (D_T + n\lambda^2 I)^{-1} D_T Q_T^T (y - Y\hat{\beta}) \end{aligned}$$

Hence we can measure the difference between models fitted with different  $\lambda$  values:

$$\begin{aligned} \|\Omega_V \hat{\alpha}_{\lambda^2} - \Omega_V \hat{\alpha}_{\lambda^2 + \delta}\| &= \left\| \Omega_V Q_T D_T^{-1} \left( (D_T + n\lambda^2 I)^{-1} - (D_T + n(\lambda^2 + \delta) I)^{-1} \right) D_T Q_T^T (y - Y\hat{\beta}) \right\| \\ &\leq n\delta \max |\omega_{Vi}| \max_{i=1:n} \left| \frac{1}{(\omega_{Ti} + n\lambda^2)(\omega_{Ti} + n\lambda^2 + n\delta)} \right| \|y - Y\hat{\beta}\| \\ &\leq \delta n \frac{\max_{i=1:n} |\omega_{Vi}|}{\min_{i=1:n} |\omega_{Ti}^2|} K \\ &= \delta O_P(n^{\tau+1}) K \end{aligned}$$

So we can define a  $\tilde{\delta}$ -covering set for  $\{\Omega_V \hat{\alpha}_{\lambda^2}\}$  using  $\{\Omega_V \hat{\alpha}_{i\delta} : i = 0, 1, \dots, \frac{1}{\tilde{\delta}}\}$  where  $\delta = \tilde{\delta} O_P(n^{-(\tau+1)})/K$ . So the covering set contains a total of  $\frac{K}{\tilde{\delta}} O_P(n^{\tau+1})$  functions. Hence the entropy is

$$\log \left( \frac{K}{\tilde{\delta}} \right) + (\tau + 1) \log n$$