

# Almost-Cross-Validation Theorem

July 1, 2017

We are interested in bounding the error of the selected model when tuning penalty parameters by a “modified averaged version of cross-validation”. Our result is an application of Mitchell’s result to penalized regression problems where the fitted functions are smooth with respect to the penalty parameters.

Suppose that the data is generated from the model

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

Suppose the errors are independent and bounded ( $\|\epsilon\|_\infty < \infty$ ).

The penalized regression model fitted on dataset  $D$  is denoted

$$\hat{g}_D(\cdot|\boldsymbol{\lambda}) = \arg \min_{g \in \mathcal{G}} L_D(g|\boldsymbol{\lambda})$$

Split data  $D$  into  $K$  folds, where each fold is  $D_k$  and  $D_{-k} = D \setminus D_k$ . Suppose  $D$  has size  $n$ ,  $D_k$  all have size  $n_V$ , and  $D_{-k}$  all have size  $n_T$ .

We select penalty parameters such that

$$\hat{\boldsymbol{\lambda}} = \arg \min_{\boldsymbol{\lambda}} \sum_{k=1}^K \|y - \hat{g}_{D_{-k}}(\cdot|\boldsymbol{\lambda})\|_k^2$$

We consider the behavior of the “modified averaged version of cross-validation”

$$\hat{g}_{MCV}(\cdot|D) = \frac{1}{K} \sum_{k=1}^K \hat{g}_{D_{-k}}(\cdot|\boldsymbol{\lambda})$$

Under sufficient entropy conditions, the error of the selected model will converge to the error of the oracle.

We are interested in bounding its generalization error

$$E_D \|\hat{g}_{MCV}(\cdot|D) - g^*\|^2 = E_D \left[ \int (\hat{g}_{MCV}(x|D) - g^*(x))^2 d\mu(x) \right]$$

## 1 Theorem 2

We will assume that  $\sup_{g \in \mathcal{G}} \|g\|_\infty \leq G$ . Suppose  $\|\epsilon\|_\infty \leq G$ .

Let  $\Lambda = [\lambda_{min}, \lambda_{max}]^J$ .

Suppose that for all  $k = 1, \dots, K$ , the following smoothness condition holds: For some constant  $C > 0$ , for all  $\lambda_1, \lambda_2 \in \Lambda$ , we have

$$\|\hat{g}_{D-k}(\cdot|\lambda_1) - \hat{g}_{D-k}(\cdot|\lambda_2)\|_\infty \leq C\|\lambda_1 - \lambda_2\|$$

Then there is an absolute constant  $c > 0$  such that for all  $a > 0$

$$E [\|g^* - \hat{g}_{MCV}(\cdot|D)\|^2] \leq (1+a) \min_{\lambda \in \Lambda} E [\|g^* - \hat{g}_{D(n_T)}(\cdot|\lambda)\|^2] + \frac{(1+a)^2}{a} \frac{cJ}{n_V} \left( C_\Lambda + \frac{1}{2} \log n_V + 4GC_\Lambda \log n_V \right)$$

where

$$C_\Lambda = 1 + \log(128GC(\lambda_{max} - \lambda_{min}))$$

### Proof

We apply Theorem 3.5 in Mitchell's paper. We consider the loss function  $Q(x, g) = (g(x) - y)^2$ . We will use the set of statistics

$$\mathcal{G}(T) = \{\hat{g}_T(\cdot|\lambda)\}$$

where  $T$  is some training data.

**1. Establish Assumptions A.1 and A.2 are satisfied:**

Theorem 3.5 relies on assumption A.1 and A.2 to be satisfied.

Assumption A.1 states that the Orlicz norm with  $\psi_1 = \exp(x) - 1$  is bounded for some constant  $K_0$ :

$$\left\| (\hat{g}_D(\cdot|\lambda) - y)^2 - (g^* - y)^2 \right\|_{L_{\psi_1}} \leq K_0$$

where

$$\|f\|_{\psi_1} = \inf \{C > 0 : E\psi(|f|/C) \leq 1\}$$

Since we have assumed that  $\|g\|_\infty \leq G$ , then by Lemma Orlicz-norm-properties (see Appendix)

$$\begin{aligned} \left\| (\hat{g}_D(\cdot|\lambda) - y)^2 - (g^* - y)^2 \right\|_{L_{\psi_1}} &\leq 2 \left\| (\hat{g}_D(\cdot|\lambda) - y)^2 - (g^* - y)^2 \right\|_\infty \\ &\leq 18G^2 \end{aligned}$$

Assumption A.2 states that

$$\left\| (\hat{g}_D(\cdot|\lambda) - y)^2 - (g^* - y)^2 \right\|_{L_2} \leq K_1 \|\hat{g}_D(\cdot|\lambda) - g^*\|_{L_2}$$

To see that this is satisfied, note that

$$\begin{aligned}
\left\| (\hat{g}_D(\cdot|\boldsymbol{\lambda}) - y)^2 - (g^* - y)^2 \right\|_{L_2}^2 &= \int \left( (\hat{g}_D(\cdot|\boldsymbol{\lambda}) - y)^2 - (g^* - y)^2 \right)^2 d\mu(x) \\
&= \int \left( (\hat{g}_D(\cdot|\boldsymbol{\lambda}) - g^*)^2 + 2(\hat{g}_D(\cdot|\boldsymbol{\lambda}) - g^*)(g^* - y) \right)^2 d\mu(x) \\
&= \int (\hat{g}_D(\cdot|\boldsymbol{\lambda}) - g^*)^2 ((\hat{g}_D(\cdot|\boldsymbol{\lambda}) - g^*) + 2(g^* - y))^2 d\mu(x) \\
&\leq \int (\hat{g}_D(\cdot|\boldsymbol{\lambda}) - g^*)^2 (2K + 2K)^2 d\mu(x) \\
&\leq 16G^2 \|\hat{g}_D(\cdot|\boldsymbol{\lambda}) - g^*\|_{L_2}^2
\end{aligned}$$

## 2. Calculate the $L_2$ and $\psi_1$ entropies

Theorem 3.5 requires calculating the entropies of the excess loss functions

$$\mathcal{Q}(T) = \left\{ (\hat{g}_T(x|\boldsymbol{\lambda}) - y)^2 - (g^*(x) - y)^2 : \lambda \in \Lambda \right\}$$

where  $T$  is any training dataset.

We are interested in calculating the  $\|\cdot\|_{\psi_1}$  and  $\|\cdot\|_{L_2}$  entropy of the function class

$$\mathcal{Q}_d^{L_2}(T) = \left\{ \tilde{Q} \in \mathcal{Q}(T) : \|\tilde{Q}\|_{L_2} \leq \sqrt{d} \right\}$$

To bound these two entropies, we'll actually bound  $H(u, \mathcal{Q}_d^{L_2}(T), \|\cdot\|_{\infty})$  since

$$H(u, \mathcal{Q}_d^{L_2}(T), \|\cdot\|_{\psi_1}) \leq H(u/2, \mathcal{Q}_d^{L_2}(T), \|\cdot\|_{\infty})$$

and

$$H(u, \mathcal{Q}_d^{L_2}(T), \|\cdot\|_{L_2}) \leq H(u, \mathcal{Q}_d^{L_2}(T), \|\cdot\|_{\infty})$$

We show that the excess log functions  $Q(x|\boldsymbol{\lambda})$  are smoothly parametric in  $\boldsymbol{\lambda}$ :

$$\begin{aligned}
\|Q(x|\boldsymbol{\lambda}_1) - Q(x|\boldsymbol{\lambda}_2)\|_{\infty} &= \left\| (\hat{g}_T(x|\boldsymbol{\lambda}_1) - y)^2 - (\hat{g}_T(x|\boldsymbol{\lambda}_2) - y)^2 \right\|_{\infty} \\
&= \|(\hat{g}_T(x|\boldsymbol{\lambda}_1) - \hat{g}_T(x|\boldsymbol{\lambda}_2))(\hat{g}_T(x|\boldsymbol{\lambda}_1) + \hat{g}_T(x|\boldsymbol{\lambda}_2) - 2y)\|_{\infty} \\
&\leq 6G \|\hat{g}_T(x|\boldsymbol{\lambda}_1) - \hat{g}_T(x|\boldsymbol{\lambda}_2)\|_{\infty}
\end{aligned}$$

Under the assumption that

$$\|\hat{g}_T(x|\boldsymbol{\lambda}_1) - \hat{g}_T(x|\boldsymbol{\lambda}_2)\|_{\infty} \leq C \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|$$

then

$$\|Q(x|\boldsymbol{\lambda}_1) - Q(x|\boldsymbol{\lambda}_2)\|_\infty \leq 6GC \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|$$

Hence by lemma param covering cube

$$\begin{aligned} H(u, \mathcal{Q}_d^{L_2}(T), \|\cdot\|_\infty) &\leq H\left(\frac{u}{6GC}, \Lambda, \|\cdot\|_2\right) \\ &\leq \log \left[ \frac{1}{C_J} \left( \frac{36GC(\lambda_{max} - \lambda_{min}) + 2u}{u} \right)^J \right] \end{aligned}$$

Hence

$$H(u, \mathcal{Q}_d^{L_2}(T), \|\cdot\|_{\psi_1}) \leq \log \left[ \frac{1}{C_J} \left( \frac{72GC(\lambda_{max} - \lambda_{min}) + 2u}{u} \right)^J \right]$$

and

$$H(u, \mathcal{Q}_d^{L_2}(T), \|\cdot\|_{L_2}) \leq \log \left[ \frac{1}{C_J} \left( \frac{36GC(\lambda_{max} - \lambda_{min}) + 2u}{u} \right)^J \right]$$

We calculate each component of the complexity term  $J(d)$ :

$$\begin{aligned} \gamma_1(\mathcal{Q}_d^{L_2}(T), \|\cdot\|_{\psi_1}) &\leq \int_0^{2G} H(u, \mathcal{Q}_d^{L_2}(T), \|\cdot\|_{\psi_1}) du \\ &= \int_0^{2G} \log \left[ \frac{1}{C_J} \left( \frac{72GC(\lambda_{max} - \lambda_{min}) + 2u}{u} \right)^J \right] du \\ &= 2G \int_0^1 \left[ \log \left( \frac{1}{C_J} \right) + J \log \left( \frac{72GC(\lambda_{max} - \lambda_{min}) + 4Gv}{2Gv} \right) \right] dv \\ &\leq 2G \int_0^1 \left[ \log \left( \frac{1}{C_J} \right) + J \log \left( \frac{72GC(\lambda_{max} - \lambda_{min})}{v} \right) + J \log 4 \right] dv \\ &= 2G \left[ \log \left( \frac{1}{C_J} \right) + J + J \log (72GC(\lambda_{max} - \lambda_{min})) + J \log 4 \right] \\ &< 2GJ (1 + \log (288GC(\lambda_{max} - \lambda_{min}))) \end{aligned}$$

(since  $C_J > 1$  for all  $J$ )

and

$$\begin{aligned}
\gamma_2(\mathcal{Q}_d^{L_2}(T), \|\cdot\|_{L_2}) &= \int_0^{\sqrt{d}} \left[ H(u, \mathcal{Q}_d^{L_2}(T), \|\cdot\|_{L_2}) \right]^{1/2} du \\
&= \sqrt{d} \int_0^1 \left( \log \left[ \frac{1}{C_J} \left( \frac{36GC(\lambda_{max} - \lambda_{min}) + 2\sqrt{d}v}{\sqrt{d}v} \right)^J \right] \right)^{1/2} dv \\
&\leq \sqrt{d} \left[ \int_0^1 \left( \log \left( \frac{1}{C_J} \right) + J \log \left( \frac{72GC(\lambda_{max} - \lambda_{min})}{\sqrt{d}v} \right) + J \log 4 \right) dv \right]^{1/2} \\
&= \sqrt{d} \left[ \int_0^1 \left( \log \left( \frac{1}{C_J} \right) + J \log \left( \frac{72GC(\lambda_{max} - \lambda_{min})}{\sqrt{d}} \right) + J \log \frac{1}{v} + J \log 4 \right) dv \right]^{1/2} \\
&= \sqrt{d} \left[ \log \left( \frac{1}{C_J} \right) + J + J \log 4 + J \log \left( \frac{72GC(\lambda_{max} - \lambda_{min})}{\sqrt{d}} \right) \right]^{1/2} \\
&< \sqrt{d} \left[ J \left( 1 + \log \left( \frac{288GC(\lambda_{max} - \lambda_{min})}{\sqrt{d}} \right) \right) \right]^{1/2}
\end{aligned}$$

### 3. Apply Theorem 3.5

Now we must select an increasing function  $\mathcal{J}$  such that  $\mathcal{J}^{-1}$  is strictly convex and

$$\mathcal{J}(d) \geq \gamma_2(\mathcal{Q}_d^{L_2}(T), \|\cdot\|_{L_2}) + \frac{(\log n_V) \gamma_1(\mathcal{Q}_d^{L_2}(T), \|\cdot\|_{\psi_1})}{\sqrt{n_V}}, \forall d \geq d_{min}$$

Let us choose  $d_{min} = 1/n_V$ . Then

$$\begin{aligned}
\gamma_2(\mathcal{Q}_d^{L_2}(T), \|\cdot\|_{L_2}) &\leq \sqrt{d} \left[ J \left( 1 + \log \left( \frac{128GC(\lambda_{max} - \lambda_{min})}{\sqrt{d}} \right) \right) \right]^{1/2} \\
&\leq \sqrt{d} [J (1 + \log (128GC(\lambda_{max} - \lambda_{min})\sqrt{n_V}))]^{1/2}
\end{aligned}$$

Let

$$K_{n,1} = [J (1 + \log (288GC(\lambda_{max} - \lambda_{min})\sqrt{n_V}))]^{1/2}$$

and

$$K_{n,2} = 2GJ (1 + \log (288GC(\lambda_{max} - \lambda_{min}))) (\log n_V)$$

We define

$$Q(d) := \sqrt{d}K_{n,1} + \frac{1}{\sqrt{n_V}}K_{n,2}$$

Then  $\mathcal{J}^{-1}(b)$  is the strictly convex function

$$\mathcal{J}^{-1}(b) = \left( \frac{b - \frac{1}{\sqrt{n_V}}K_{n,2}}{K_{n,1}} \right)^2$$

The convex conjugate of  $\mathcal{J}^{-1}(b)$  is

$$\begin{aligned} \psi(z) &= \sup_x xz - \mathcal{J}^{-1}(x) \\ &= \sup_x xz - \left( \frac{x - \frac{1}{\sqrt{n_V}}K_{n,2}}{K_{n,1}} \right)^2 \\ &= \frac{K_{n,1}^2 z^2}{4} + \frac{1}{\sqrt{n_V}}K_{n,2}z \end{aligned}$$

We check the condition that Then  $\psi(z)/z^r$  is a decreasing function in  $z$  for  $r = 3 \geq 1$ . Also  $\lim_{z \rightarrow \infty} \psi(z) = \infty$ . Theorem 3.5 states that for all  $a > 0, q > 1$ , we have

$$E [\|y - \hat{g}_{MCV}(\cdot|D)\|^2] - E [\|y - g^*\|^2] \leq (1+a) \min_{\lambda \in \Lambda} (E_{D^{(n_T)}} [\|y - \hat{g}_{D^{(n_T)}}(\cdot|\lambda)\|^2] - E [\|y - g^*\|^2]) + \frac{ac\epsilon_q(1/q)}{q}$$

where  $\epsilon_q(u) = \psi\left(\frac{2q^{r+1}(1+a)u}{a\sqrt{n_V}}\right) \vee \frac{1}{n_V} \forall u > 0$ .

We calculate  $\epsilon_q\left(\frac{1}{q}\right)$ :

$$\epsilon_q\left(\frac{1}{q}\right) = \psi\left(\frac{2q^4(1+a)\frac{1}{q}}{a\sqrt{n_V}}\right) = \frac{K_{n,1}^2}{4} \left(\frac{2q^3(1+a)}{a\sqrt{n_V}}\right)^2 + \frac{1}{\sqrt{n_V}}K_{n,2} \left(\frac{2q^3(1+a)}{a\sqrt{n_V}}\right)$$

Finally, we get

$$\begin{aligned} & E [\|y - \hat{g}_{MCV}(\cdot|D)\|^2] - E [\|y - g^*\|^2] \\ & \leq (1+a) \min_{\lambda \in \Lambda} (E_{D^{(n_T)}} [\|y - \hat{g}_{D^{(n_T)}}(\cdot|\lambda)\|^2] - E [\|y - g^*\|^2]) + ac \left( \frac{K_{n,1}^2}{4} q^5 \left( \frac{2(1+a)}{a\sqrt{n_V}} \right)^2 + K_{n,2} \frac{1}{\sqrt{n_V}} \left( \frac{2q^2(1+a)}{a\sqrt{n_V}} \right) \right) \\ & = (1+a) \min_{\lambda \in \Lambda} (E_{D^{(n_T)}} [\|y - \hat{g}_{D^{(n_T)}}(\cdot|\lambda)\|^2] - E [\|y - g^*\|^2]) + \frac{c}{n_V} \left( K_{n,1}^2 q^5 \frac{(1+a)^2}{a} + 2K_{n,2} q^2 (1+a) \right) \end{aligned}$$

As  $q \rightarrow 1$ , we get

$$E [\|y - \hat{g}_{MCV}(\cdot|D)\|^2] - E [\|y - g^*\|^2] \leq (1+a) \left[ \min_{\lambda \in \Lambda} [(E_{D^{(n_T)}} [\|y - \hat{g}_{D^{(n_T)}}(\cdot|\lambda)\|^2] - E [\|y - g^*\|^2])] \right] + \frac{c}{n_V} \left( K_{n,1}^2 \frac{(1+a)}{a} + 2K_{n,2} \right)$$

Let  $\mathcal{E}(g) = [(E_{D^{(n_T)}} [\|y - g\|^2] - E [\|y - g^*\|^2])]$

#### 4. Massaging to make a pretty theorem

We can write the above as

$$\mathcal{E}(\hat{g}_{MCV}(\cdot|D)) \leq (1+a) \min_{\lambda \in \Lambda} \mathcal{E}(\hat{g}_{D^{(n_T)}}(\cdot|\lambda)) + (1+a) \frac{cJ}{n_V} \left( \frac{(1+a)}{a} \left( \tilde{C} + \frac{1}{2} \log n_V \right) + 4G\tilde{C}(\log n_V) \right)$$

where

$$\begin{aligned} \tilde{C} &= 1 + \log(288GC(\lambda_{max} - \lambda_{min})) \\ &\leq 5 + \log(GC\Delta_\lambda) \end{aligned}$$

Moreover, since  $(1+a)/a > 1$ , we can write

$$\begin{aligned} \mathcal{E}(\hat{g}_{MCV}(\cdot|D)) &\leq (1+a) \min_{\lambda \in \Lambda} \mathcal{E}(\hat{g}_{D^{(n_T)}}(\cdot|\lambda)) + \frac{(1+a)^2}{a} \frac{cJ}{n_V} \left( \tilde{C} + \frac{1}{2} \log n_V + 4G\tilde{C} \log n_V \right) \\ &= (1+a) \min_{\lambda \in \Lambda} \mathcal{E}(\hat{g}_{D^{(n_T)}}(\cdot|\lambda)) + \frac{(1+a)^2}{a} \frac{cJ}{n_V} \left( 5 + \log(GC\Delta_\lambda) + \frac{1}{2} \log n_V + 4G(5 + \log(GC\Delta_\lambda)) \log n_V \right) \\ &\leq (1+a) \min_{\lambda \in \Lambda} \mathcal{E}(\hat{g}_{D^{(n_T)}}(\cdot|\lambda)) + \frac{(1+a)^2}{a} \frac{cJ}{n_V} (16G(5 + \log(GC\Delta_\lambda)) \log n_V) \end{aligned}$$

as long as  $4G(5 + \log(GC\Delta_\lambda)) \geq \frac{1}{2}$  and  $\log(GC\Delta_\lambda) \leq 4G(5 + \log(GC\Delta_\lambda)) \log n_V$  and  $5 \leq 4G(5 + \log(GC\Delta_\lambda)) \log n_V$ .

First, note that if  $G \geq 1/2$ , then

$$\log(GC\Delta_\lambda) \leq 2G(5 + \log(GC\Delta_\lambda)) \leq 4G(5 + \log(GC\Delta_\lambda)) \log n_V$$

and

$$5 \leq 2G(5 + \log(GC\Delta_\lambda)) \leq 4G(5 + \log(GC\Delta_\lambda)) \log n_V$$

Also

$$\frac{1}{8} \leq \frac{1}{2} (5 + \log(GC\Delta_\lambda)) \leq 4G(5 + \log(GC\Delta_\lambda))$$

Also, I want to write

$$5 \leq \log (GC\Delta_\lambda) \implies \exp(5)/G\Delta_\lambda \leq C$$

Then

$$\mathcal{E}(\hat{g}_{MCV}(\cdot|D)) \leq (1+a) \min_{\lambda \in \Lambda} \mathcal{E}(\hat{g}_{D^{(n_T)}}(\cdot|\lambda)) + 32 \frac{(1+a)^2}{a} \frac{cJ}{n_V} G \log(GC\Delta_\lambda) \log n_V$$

## 2 Lemma 2 for $\lambda$ that changes with $n$

We will assume that  $\sup_{g \in \mathcal{G}} \|g\|_\infty \leq G$ .

Suppose  $\Lambda = [n_T^{-t_{min}}, n_T^{t_{max}}]^J$ .

Suppose that for all  $k = 1, \dots, K$ , the following smoothness condition holds: For some constant  $C, \kappa > 0$ , for all  $\lambda_1, \lambda_2 \in \Lambda$ , we have

$$\|\hat{g}_{D_{-k}}(\cdot|\lambda_1) - \hat{g}_{D_{-k}}(\cdot|\lambda_2)\|_\infty \leq C n_T^\kappa \|\lambda_1 - \lambda_2\|$$

Then there is an absolute constant  $c > 0$  and a constant  $c_G$  that only depends on  $G$  such that for all  $a > 0$

$$E [\|g^* - \hat{g}_{MCV}(\cdot|D)\|^2] \leq (1+a) \min_{\lambda \in \Lambda} E [\|g^* - \hat{g}_{D^{(n_T)}}(\cdot|\lambda)\|^2] + \frac{(1+a)^2}{a} \frac{c^2 J}{n_V} [(c_G \log n_V + 1) ((\kappa + t_{max}) \log n_T + 1) + c_G]$$

## 3 Appendix

### 3.0.1 Lemma: Orlicz Norm Properties

For any function  $f$ , we have

$$\|f\|_{\psi_1} \leq 2\|f\|_\infty$$

Also

$$\|Kf\|_\psi = K\|f\|_\psi$$

Also suppose that  $\psi$  is a monotone function. Then

$$\|gf\|_\psi \leq \|g\|_\infty \|f\|_\psi$$

### Proof

To prove the bound:

$$E \left[ \exp \left( \frac{f}{2\|f\|_\infty} \right) - 1 \right] \leq \exp \frac{1}{2} - 1 < 1$$



To prove the norm scaling property:

$$\begin{aligned}
\|Kf\|_\psi &= \inf \{C > 0 : E\psi(|Kf|/C) \leq 1\} \\
&= \inf \{C > 0 : E\psi(|f|/(C/K)) \leq 1\} \\
&= \inf \{KC > 0 : E\psi(|f|/C) \leq 1\} \\
&= K \inf \{C > 0 : E\psi(|f|/C) \leq 1\} \\
&= K\|f\|_\psi
\end{aligned}$$

To prove the last bound, note that under the assumption that  $\psi$  is a monotone function, then

$$E\psi(|gf|/C) \leq E\psi(|\|g\|_\infty f|/C)$$

Therefore

$$\inf_C E\psi(|gf|/C) \leq \inf_C E\psi(|\|g\|_\infty f|/C)$$

Therefore

$$\begin{aligned}
\|gf\|_\psi &= \inf \{C > 0 : E\psi(|gf|/C) \leq 1\} \\
&\leq \inf \{C > 0 : E\psi(|\|g\|_\infty f|/C) \leq 1\} \\
&= \|g\|_\infty \inf \{C > 0 : E\psi(|f|/C) \leq 1\} \\
&= \|g\|_\infty \|f\|_\psi
\end{aligned}$$