09 Cross-validation

All of the methods we have explored thus far involve tuning parameters, which trade off bias against variance: number of neighbors k for kNN estimators, bandwidth h for kernel/local polynomial estimators, basis dimension k for series estimators, number of leaves ℓ for random forest estimators, etc. We have described choices of these tuning parameters based on balancing upper bounds on the (squared) bias and variance to minimize mean squared error, e.g., with

$$k = Cn^{2/(2+d)} \qquad h = Cn^{-1/(2s+d)} \qquad k = Cn^{d/(2s+d)} \qquad \ell = Cn^{1/(1+0.75/s\log 2)}.$$

However, in finite samples the choice of constant C can overwhelm the dependence on n and in practice the underlying smoothness s is often unknown. Further, our MSE bounds may be incorrect if smoothness or other structural assumptions fail to hold, in which case we might still want to pick the best estimator in some sense. Thus in these notes we pursue how to use the data directly to pick tuning parameters, via cross-validation.

1 Sample splitting

Here we consider splitting the sample into two parts, D_1^n and D_2^n of size n_1 and n_2, respectively, using D_1^n for training and using D_2^n as a test sample for estimating a best tuning parameter choice. Specifically, for a set $\mathcal{H}_n$ of tuning parameter values, first construct estimators $\widehat{\mu}_h(x)$ for every $h \in \mathcal{H}_n$ using the training sample D_1^n. Then move to the test data and pick the best $\widehat{h}$ satisfying, for example,

$$\widehat{h} = \operatorname*{arg\,min}_{h\in\mathcal{H}_n} \frac{1}{n_2} \sum_{i\in D_2^n} \{Y_i - \widehat{\mu}_h(X_i)\}^2.$$

Now one can construct the final estimator $\widehat{\mu}_{\widehat{h}}(x)$. A fundamental question is how the risk of this estimator compares to that of an oracle who had used the best (random) choice of tuning parameter

$$h^* = \operatorname*{arg\,min}_{h\in\mathcal{H}} \|\widehat{\mu}_h - \mu\|^2 = \operatorname*{arg\,min}_{h\in\mathcal{H}} \int \left\{\widehat{\mu}_h(x) - \mu(x)\right\}^2 d\mathbb{P}(x).$$

Note this h^* is called the best random choice since it is a random variable through its dependence on D_1^n, via the sequence of estimators $\widehat{\mu}_h$. However, also note that it is not itself a feasible choice in practice since it depends on the unknown distribution $\mathbb{P}$ and μ. There are two flavors of results for cross-validation: exact oracle inequalities with a larger error, and approximate oracle inequalities with a smaller error. The following from Györfi et al. [2002] illustrates both.

Theorem 1. *Assume* $|Y| \le B$ *and* $\max_{h\in\mathcal{H}_n} \|\widehat{\mu}_h\|_\infty \le B$. *Then*

$$\mathbb{E}\|\widehat{\mu}_{\widehat{h}} - \mu\|^2 \le \mathbb{E}\|\widehat{\mu}_{h^*} - \mu\|^2 + C\sqrt{\frac{\log|\mathcal{H}_n|}{n_2}}$$

where $C = 16B^2$, *and for any* $\delta > 0$

$$\mathbb{E}\|\widehat{\mu}_{\widehat{h}} - \mu\|^2 \le (1+\delta)\mathbb{E}\|\widehat{\mu}_{h^*} - \mu\|^2 + C_\delta\left(\frac{1+\log|\mathcal{H}_n|}{n_2}\right)$$

where $C_\delta = B^2(16/\delta + 35 + 19\delta)$.

Importantly, the above theorem shows that one can achieve an error rate matching, up to some multiplicative factor (possibly even one, matching even constants exactly), that of the estimator based on the optimally chosen tuning parameter. Further, this is not only true asymptotically, but at every finite sample size n. (Hamers & Kohler (2003) show how to further improve the constant in the approximate oracle inequality from $C_\delta = B^2(16/\delta + 35 + 19\delta)$ to $C'_\delta = B^2(4/\delta + 32/3 + 20\delta/3)$.)

In practice one can take $\mathcal{H}_n = \{2^k : k \in \{-n, -(n-1), ..., -1, 0, 1, ..., n-1, n\}\}$ for bandwidth tuning parameters, or $\mathcal{H}_n = \{1, 2, ..., n\}$ for numbers of neighbors or series dimension tuning parameters. In either case the gap between the achieved and oracle error for the second approximate oracle inequality bound is of the order $\log n/n$, which is of smaller order than $1/n^{1-\epsilon}$ for any $\epsilon > 0$ (or the square root of this for the first exact oracle inequality bound).

With the above choices of $\mathcal{H}_n$ this theorem immediately yields minimax optimal estimators, e.g., in smoothness classes, which do not require knowing the smoothness or the constant in the Holder condition. Note one can let $\mathcal{H}_n$ denote a set of pairs of (h, s) values, e.g., for local polynomial estimators. Going even further, nothing in the theorem requires the set of estimators

$$\Big\{\widehat{\mu}_h(x) : h \in \mathcal{H}_n\Big\}$$

to even be of the same form, so one could include local polynomial estimators, series estimators, additive model estimators, random forests, etc., and one would still have a guarantee on the expected L_2 error relative to that of an oracle that selected the best estimator by minimizing the true L_2 error. This is a very powerful and model-agnostic guarantee.

In what follows we prove the exact and approximate oracle inequalities separately, starting with the exact inequality.

Proof. We will bound $R_n \equiv \mathbb{E}\left(\|\widehat{\mu}_{\widehat{h}} - \mu\|^2 \;\middle|\; D_1^n\right)$, and then take an expectation outside to yield the result. First define

$$\begin{aligned} Z(h) = Z(h; D_1^n) &\equiv \Big\{Y - \widehat{\mu}_h(X)\Big\}^2 - \Big\{Y - \mu(X)\Big\}^2 \\ &= \Big\{\mu(X) - \widehat{\mu}_h(X)\Big\}\Big\{Y - \mu(X) + Y - \widehat{\mu}_h(X)\Big\} \in \pm(2B)(2B) = \pm 4B^2 \end{aligned}$$

and note that

$$R_n = \mathbb{E}\left(\|\widehat{\mu}_{\widehat{h}} - \mu\|^2 \;\middle|\; D_1^n\right) = \mathbb{E}\Big\{Z(\widehat{h}) \mid D_1^n\Big\},$$

since by a usual bias-variance decomposition we have

$$\begin{aligned} \mathbb{E}\left[\Big\{Y - \widehat{\mu}_{\widehat{h}}(X)\Big\}^2 \;\middle|\; D_1^n\right] &= \mathbb{E}\left[\Big\{Y - \mu(X) + \mu(X) - \widehat{\mu}_{\widehat{h}}(X)\Big\}^2 \;\middle|\; D_1^n\right] \\ &= \mathbb{E}\left[\Big\{Y - \mu(X)\Big\}^2\right] + \mathbb{E}\left[\Big\{\mu(X) - \widehat{\mu}_{\widehat{h}}(X)\Big\}^2 \;\middle|\; D_1^n\right]. \end{aligned}$$

In what follows we use the notation $P\{f(Z)\} = \mathbb{E}\{f(Z) \mid D_1^n\}$ and $\mathbb{P}_n\{f(Z)\} = \frac{1}{n_2}\sum_{i \in D_2^n} f(Z_i)$ as shorthand. Now consider the central decomposition

$$R_n = P\{Z(\widehat{h})\} = \mathbb{P}_n\{Z(\widehat{h})\} + (P - \mathbb{P}_n)Z(\widehat{h}) \equiv S_n + T_n,$$

and we bound the conditional mean of each term. By definition of $\widehat{h} = \arg\min_h \mathbb{P}_n\{Z(h)\}$ we have

$$\begin{aligned} S_n &= \mathbb{P}_n\{Z(\widehat{h})\} \le \mathbb{P}_n\{Z(h^*)\} \\ &= \frac{1}{n_2} \sum_{i \in D_2^n} \left[\left\{ \widehat{\mu}_{h^*}(X_i) - Y_i \right\}^2 - \left\{ \mu(X_i) - Y_i \right\}^2 \right]. \end{aligned}$$

Therefore

$$\begin{aligned} \mathbb{E}(S_n \mid D_1^n) &\le \int \left[\left\{ \widehat{\mu}_{h^*}(X) - \mu(X) + \mu(X) - Y \right\}^2 - \left\{ \mu(X) - Y \right\}^2 \right] \, d\mathbb{P}(X, Y) \\ &= \|\widehat{\mu}_{h^*} - \mu\|^2 \end{aligned}$$

whose expectation matches the first term in the desired bound. Now for the second term T_n note that $\mathbb{E}\{(P - \mathbb{P}_n)Z(h) \mid D_1^n\} = 0$ and $\sum_i \{Z_i(h) - PZ(h)\}$ is sub-Gaussian with parameter $\sqrt{n_2} 8B^2$ since each independent $Z_i(h)$ is sub-Gaussian with parameter $4B^2$. Therefore we can apply the maximal inequality for sub-Gaussians (from Lecture 8, also see last section of these notes) to deduce that

$$\mathbb{E}(T_n \mid D_1^n) \le \mathbb{E}\Big\{ \max_{h \in \mathcal{H}_n} |(P - \mathbb{P}_n)Z(h)| \;\Big|\; D_1^n \Big\} \le 16B^2 \sqrt{\frac{\log |\mathcal{H}_n|}{n_2}}.$$

This yields the result. A similar bound can alternatively be obtained as follows. First we use a union bound and Hoeffding's inequality to get

$$\mathbb{P}\Big\{ |(P - \mathbb{P}_n)Z(\widehat{h})| \ge t \mid D_1^n \Big\} \le |\mathcal{H}_n| \max_{h \in \mathcal{H}_n} \mathbb{P}\Big\{ |(P - \mathbb{P}_n)Z(h)| \ge t \mid D_1^n \Big\} \le |\mathcal{H}_n| 2 \exp\left\{ \frac{-n_2 t^2}{(8B^2)^2/2} \right\}.$$

Then for $C = 32B^4$ we have

$$\begin{aligned} \mathbb{E}(T_n \mid D_1^n) &\le \sqrt{\mathbb{E}(T_n^2 \mid D_1^n)} \le \sqrt{u + \int_u^\infty \mathbb{P}(T_n^2 \ge t \mid D_1^n) \, dt} \\ &= \sqrt{u + \int_u^\infty \mathbb{P}(|T_n| \ge \sqrt{t} \mid D_1^n) \, dt} \le \sqrt{u + \int_u^\infty |\mathcal{H}_n| 2 \exp\left(\frac{-n_2 t}{C} \right) \, dt} \\ &= \sqrt{u + 2|\mathcal{H}_n| \left\{ \frac{-C}{n_2} \exp\left(\frac{-n_2 t}{C} \right) \right\} \Big|_u^\infty} = \sqrt{u + \frac{2C|\mathcal{H}_n|}{n_2} \exp\left(\frac{-n_2 u}{C} \right)} \end{aligned}$$

To minimize over u we take $u = C \log 2|\mathcal{H}_n|/n_2$, which gives $\mathbb{E}(T_n \mid D_1^n) \le \sqrt{\frac{C(1+\log 2|\mathcal{H}_n|)}{n_2}}$. □

The approximate oracle inequality is given for example in Theorem 7.1 from Györfi et al. [2002]. Whereas the exact oracle inequality result uses a centered decomposition in the proof, the approximate oracle inequality uses a decentered decomposition, which pulls the variables a bit away from their expectation, so that the corresponding second term $\mathbb{P}(T_n \ge t \mid D_1^n)$ will be smaller.

Proof. As before we bound $R_n \equiv \mathbb{E}\left(\|\widehat{\mu}_{\widehat{h}} - \mu\|^2 \;\middle|\; D_1^n\right)$, and then take an expectation outside to yield the result. This time consider the decentered decomposition

$$R_n = P\{Z(\widehat{h})\} = (1+\delta)\mathbb{P}_n\{Z(\widehat{h})\} + \Big[(1+\delta)(P-\mathbb{P}_n)Z(\widehat{h}) - \delta P\{Z(\widehat{h})\}\Big] \equiv S_n + T_n.$$

As in the exact oracle inequality, by definition of $\widehat{h} = \arg\min_h \mathbb{P}_n\{Z(h)\}$ we have

$$\begin{aligned} S_n &= (1+\delta)\mathbb{P}_n\{Z(\widehat{h})\} \le (1+\delta)\mathbb{P}_n\{Z(h^*)\} \\ &= (1+\delta)\frac{1}{n_2}\sum_{i\in D_2^n}\left[\Big\{\widehat{\mu}_{h^*}(X_i) - Y_i\Big\}^2 - \Big\{\mu(X_i) - Y_i\Big\}^2\right]. \end{aligned}$$

so that

$$\mathbb{E}(S_n \mid D_1^n) = (1+\delta)\|\widehat{\mu}_{h^*} - \mu\|^2$$

whose expectation matches the first term in the desired bound. Now we consider T_n, and show that its conditional expectation given D_1^n is bounded above by the second term in the desired bound, $C(1+\log|\mathcal{H}_n|)/n_2$. We start with the following lemma.

Lemma 1. *Under the assumptions of the theorem, and for T_n and C defined as above, we have*

$$\mathbb{P}(T_n \ge t \mid D_1^n) \le |\mathcal{H}_n|\exp(-n_2 t/C).$$

Proof. We have by a union bound that

$$\begin{aligned} \mathbb{P}(T_n \ge t \mid D_1^n) &= \mathbb{P}\Big[(1+\delta)(P-\mathbb{P}_n)Z(\widehat{h}) - \delta P\{Z(\widehat{h})\} \ge t \;\Big|\; D_1^n\Big] \\ &\le \mathbb{P}\left[\max_{h\in\mathcal{H}_n}\Big\{(1+\delta)(P-\mathbb{P}_n)Z(h) - \delta P\{Z(h)\}\Big\} \ge t \;\Big|\; D_1^n\right] \\ &= \mathbb{P}\left[\exists h \in \mathcal{H}_n : (1+\delta)(P-\mathbb{P}_n)Z(h) - \delta P\{Z(h)\} \ge t \;\Big|\; D_1^n\right] \\ &\le \sum_{h\in\mathcal{H}_n}\mathbb{P}\left[(P-\mathbb{P}_n)Z(h) \ge \frac{t+\delta P\{Z(h)\}}{1+\delta} \;\Big|\; D_1^n\right] \\ &\le |\mathcal{H}_n|\max_{h\in\mathcal{H}_n}\mathbb{P}\left[(P-\mathbb{P}_n)Z(h) \ge \frac{t+\delta P\{Z(h)\}}{1+\delta} \;\Big|\; D_1^n\right]. \end{aligned}$$

Now note

$$\begin{aligned} \sigma^2 &= \mathrm{var}\{Z(h) \mid D_1^n\} \le \mathbb{E}\{Z(h)^2 \mid D_1^n\} \\ &= \mathbb{E}\left[\Big\{\mu(X) - \widehat{\mu}_h(X)\Big\}^2\Big\{Y - \mu(X) + Y - \widehat{\mu}_h(X)\Big\}^2 \;\Big|\; D_1^n\right] \\ &\le (4B)^2\|\widehat{\mu}_h - \mu\|^2 = 16B^2 P\{Z(h)\}. \end{aligned}$$

Therefore using the above bounds and Bernstein's inequality

$$\begin{aligned} \mathbb{P}(T_n \ge t \mid D_1^n) &\le |\mathcal{H}_n|\max_{h\in\mathcal{H}_n}\mathbb{P}\left[(P-\mathbb{P}_n)Z(h) \ge \frac{t+\delta\sigma^2/(16B^2)}{1+\delta} \;\Big|\; D_1^n\right] \\ &\equiv |\mathcal{H}_n|\max_{h\in\mathcal{H}_n}\mathbb{P}\Big[(P-\mathbb{P}_n)Z(h) \ge M(\delta,t) \;\Big|\; D_1^n\Big] \\ &\le |\mathcal{H}_n|\exp\left\{\frac{-n_2 M(\delta,t)^2}{2\sigma^2 + 2M(\delta,t)(8B^2)/3}\right\}. \end{aligned}$$

Finally, Györfi et al. [2002] (Problem 7.1) show

$$\frac{M(\delta,t)^2}{2\sigma^2 + 2M(\delta,t)(8B^2)/3} \geq t/C$$

for $C = C_\delta$ as defined in the theorem, which yields the result. □

Applying the lemma, now for any $u \geq 0$ we have

$$\begin{aligned}\mathbb{E}(T_n \mid D_1^n) &\leq u + \int_u^\infty \mathbb{P}(T_n \geq t \mid D_1^n)\, dt \leq u + |\mathcal{H}_n| \int_u^\infty \exp(-n_2 t/C)\, dt \\ &= u + |\mathcal{H}_n| \frac{C \exp(-n_2 u/C)}{n_2}\end{aligned}$$

and setting $u = C \log|\mathcal{H}_n|/n_2$ gives

$$\mathbb{E}(T_n \mid D_1^n) \leq \frac{C \log|\mathcal{H}_n|}{n_2} + \frac{C|\mathcal{H}_n|}{n_2} \exp(-\log|\mathcal{H}_n|) = C\left(\frac{1 + \log|\mathcal{H}_n|}{n_2}\right)$$

yielding the desired second term in the bound. □

In addition to using cross-validation for model selection, it can also be of interest for directly estimating the risk of one or more estimators (e.g,. for understanding performance, or for testing whether one estimator has better risk than another, etc.). The following result illustrates how one can obtain for example a Berry-Esseen bound on how far away the distribution of the (centered and scaled) cross-validated risk estimate is from a Gaussian.

Theorem 2. *Define the estimated risk* $\widehat{R}(\widehat{\mu}) = \frac{1}{n_2}\sum_{i \in D_1^n} \left\{Y_i - \widehat{\mu}(X_i)\right\}^2$ *and its conditional expectation* $R(\widehat{\mu}) = \mathbb{E}\left[\left\{Y - \widehat{\mu}(X)\right\}^2 \,\middle|\, D_1^n\right]$, *as well as the quantities*

$$\begin{aligned}\sigma_n^2 &= var\left[\left\{Y - \widehat{\mu}(X)\right\}^2 \,\middle|\, D_1^n\right] \\ \rho_n &= \mathbb{E}\left[\left|\{Y - \widehat{\mu}(X)\}^2 - \mathbb{E}[\{Y - \widehat{\mu}(X)\}^2 \mid D_1^n]\right|^3 \,\middle|\, D_1^n\right].\end{aligned}$$

Then

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left\{ \frac{\widehat{R}(\widehat{\mu}) - R(\widehat{\mu})}{\sigma_n/\sqrt{n_2}} \leq t \,\middle|\, D_1^n \right\} - \Phi(t) \right| \leq \frac{C\rho_n/\sigma_n^3}{\sqrt{n_2}}$$

for some constant $C \in (0.4, 0.5)$.

Proof. Apply the Berry-Esseen inequality (conditionally, given D_1^n) to the variables $Z_i = \{Y_i - \widehat{\mu}(X_i)\}^2 - \mathbb{E}[\{Y_i - \widehat{\mu}(X_i)\}^2 \mid D_1^n]$. □

The above result motivates approximate 95% confidence intervals for the true error $R(\widehat{\mu}_h)$ of the form

$$\widehat{R}(\widehat{\mu}) \pm 1.96\widehat{\sigma}_n/\sqrt{n_2}$$

for $\widehat{\sigma}_n^2 = \widehat{\text{var}}[\{Y - \widehat{\mu}_h(X)\}^2 \mid D_1^n]$ the empirical variance estimator over the second sample.

2 K-fold cross-validation

In practice it is popular to take the sample splitting approach described above, and swap the roles of test and training samples, so that each observation gets to contribute to both a test and training set. This could be motivated by concerns about data scarcity (e.g., in sample-splitting we only use half the data for training/testing) or stability (sample-splitting can give somewhat different answers based on different splits, whereas K-fold cross-validation can be more stable).

For example, in $K = 2$-fold cross-validation we choose tuning parameters via

$$\widehat{h} = \underset{h \in \mathcal{H}_n}{\arg\min} \frac{1}{2} \left[\frac{1}{n_2} \sum_{i \in D_2^n} \{Y_i - \widehat{\mu}_{1h}(X_i)\}^2 + \frac{1}{n_1} \sum_{i \in D_1^n} \{Y_i - \widehat{\mu}_{2h}(X_i)\}^2 \right].$$

In general for K-fold cross-validation we take

$$\widehat{h} = \underset{h \in \mathcal{H}_n}{\arg\min} \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i \in D_k^n} \left\{ Y_i - \widehat{\mu}_{-\kappa(i),h}(X_i) \right\}^2$$

where $\kappa(i)$ is the fold for unit i and $\widehat{\mu}_{-k,h}$ is the estimator $\widehat{\mu}_h$ applied to all folds except k.

In practice most people use $K = 5$ or $K = 10$ folds (sometimes $K = n$ is used, which is called leave-one-out cross-validation - we will return to this later); however this choice is driven more by computational concerns and/or tradition, more than any rigorous theoretical justification.

Here it is natural to define the optimal index as

$$h^* = \underset{h \in \mathcal{H}_n}{\arg\min} \frac{1}{K} \sum_{k=1}^{K} \|\widehat{\mu}_{-k,h} - \mu\|^2$$

and so one goal is to try to bound the cross-validated risk

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\|\widehat{\mu}_{-k,\widehat{h}} - \mu\|^2$$

relative to that of the oracle

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\|\widehat{\mu}_{-k,h^*} - \mu\|^2.$$

It turns out a similar result to sample splitting can be proved here.

Theorem 3. *Assume* $|Y| \leq B$ *and* $\max_{h \in \mathcal{H}_n} \|\widehat{\mu}_h\|_\infty \leq B$. *Then for any* $\delta > 0$

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\|\widehat{\mu}_{-k,\widehat{h}} - \mu\|^2 \leq (1+\delta) \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\|\widehat{\mu}_{-k,h^*} - \mu\|^2 + C_\delta \left(\frac{1 + \log|\mathcal{H}_n|}{n/K} \right)$$

where $C_\delta = B^2(16/\delta + 35 + 19\delta)$.

Proof. Using the same decomposition as before, let

$$R_k \equiv \mathbb{E}\left(\|\widehat{\mu}_{-k,\widehat{h}} - \mu\|^2 \;\middle|\; D^n_{-k}\right) = \mathbb{E}\left[\left\{Y - \widehat{\mu}_{-k,\widehat{h}}(X)\right\}^2 \;\middle|\; D^n_{-k}\right] - \mathbb{E}\left[\left\{Y - \mu(X)\right\}^2\right]$$
$$\equiv P^k Z_k(\widehat{h})$$

where we let $P^k\{f(Z)\} = \mathbb{E}\{f(Z) \mid D^n_{-k}\}$ and

$$Z_k(h) = \left\{Y_i - \widehat{\mu}_{-k,h}(X_i)\right\}^2 - \left\{Y_i - \mu(X_i)\right\}^2$$

similar to before, and note

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}\|\widehat{\mu}_{-k,\widehat{h}} - \mu\|^2 = \mathbb{E}\left\{\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left(\|\widehat{\mu}_{-k,\widehat{h}} - \mu\|^2 \mid D^n_{-k}\right)\right\} = \mathbb{E}\left(\frac{1}{K}\sum_{k=1}^K R_k\right)$$

Now use the decomposition

$$R_k = \frac{1}{K}\sum_{k=1}^K P^k Z_k(\widehat{h}) = \frac{1}{K}\sum_{k=1}^K \mathbb{P}^k_n Z_k(\widehat{h}) + \frac{1}{K}\sum_{k=1}^K \left\{(1+\delta)(P^k - \mathbb{P}^k_n)Z_k(\widehat{h}) - \delta P^k Z_k(\widehat{h})\right\}$$
$$\equiv \frac{1}{K}\sum_{k=1}^K S_k + \frac{1}{K}\sum_{k=1}^K T_k$$

where $\mathbb{P}^k_n$ denotes a sample average over the kth fold. Now we consider the terms S_k and T_k in turn, as before. If $n_k = n/K$ we have

$$\frac{1}{K}\sum_{k=1}^K S_k = \frac{1}{K}\sum_{k=1}^K \frac{1}{n/K}\sum_{i \in D^n_k} Z_{ki}(\widehat{h})$$
$$= (1+\delta)\frac{1}{n}\sum_{i=1}^n \left[\left\{\widehat{\mu}_{-\kappa(i),\widehat{h}}(X_i) - Y_i\right\}^2 - \left\{\mu(X_i) - Y_i\right\}^2\right]$$
$$\leq (1+\delta)\frac{1}{n}\sum_{i=1}^n \left[\left\{\widehat{\mu}_{-\kappa(i),h^*}(X_i) - Y_i\right\}^2 - \left\{\mu(X_i) - Y_i\right\}^2\right]$$
$$= (1+\delta)\frac{1}{K}\sum_{k=1}^K \mathbb{P}^k_n Z_k(h^*) \equiv (1+\delta)\frac{1}{K}\sum_{k=1}^K S^*_k$$

so that

$$\mathbb{E}\left(\frac{1}{K}\sum_{k=1}^K S_k\right) \leq \mathbb{E}\left\{\frac{1}{K}\sum_{k=1}^K \mathbb{E}(S^*_k \mid D^n_{-k})\right\}$$
$$= (1+\delta)\frac{1}{K}\sum_{k=1}^K \mathbb{E}\|\widehat{\mu}_{-k,h^*} - \mu\|^2$$

which is the first term in the desired bound.

Now move to the terms $\frac{1}{K}\sum_k T_k$. Applying the same logic as before shows for each k that

$$\mathbb{E}(T_k \mid D^n_{-k}) \leq C\left(\frac{1+\log|\mathcal{H}_n|}{n_k}\right)$$

so that when $n_k = n/K$ we have

$$\mathbb{E}\left(\frac{1}{K}\sum_k T_k\right) = \mathbb{E}\left\{\frac{1}{K}\sum_{k=1}^K \mathbb{E}(T_k \mid D^n_{-k})\right\} \leq C\left(\frac{1+\log|\mathcal{H}_n|}{n/K}\right)$$

Therefore overall we have

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}\|\widehat{\mu}_{-k,\widehat{h}} - \mu\|^2 \leq (1+\delta)\frac{1}{K}\sum_{k=1}^K \mathbb{E}\|\widehat{\mu}_{-k,h^*} - \mu\|^2 + C\left(\frac{1+\log|\mathcal{H}_n|}{n/K}\right)$$

as desired. □

As with sample splitting, we can also use cross-validation purely for risk estimation, instead of explicit model selection. Berry-Esseen style approximate normality results are less natural / more challenging to apply for K-fold cross-validation, but we can obtain asymptotic normality results, as in the following, which is similar to Theorem 3 of Dudoit and van der Laan [2005].

Theorem 4. *Denote the cross-validated estimated and true risk by*

$$\widehat{R}(\widehat{\mu}) = \frac{1}{K}\sum_{k=1}^K \frac{1}{n_k}\sum_{i\in D_k^n}\left\{Y_i - \widehat{\mu}_{-k}(X_i)\right\}^2$$

$$R(\widehat{\mu}) = \frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\left\{Y - \widehat{\mu}_{-k}(X)\right\}^2 \;\middle|\; D^n_{-k}\right].$$

Assume the number of folds K is constant, and

1. *$|Y| \leq B$ and $\|\widehat{\mu}\|_\infty \leq B$, and*
2. *there exists a function $\overline{\mu}(x)$ such that $\|\overline{\mu}\|_\infty \leq B$ and $\|\widehat{\mu}_{-k} - \overline{\mu}\| = o_{\mathbb{P}}(1)$.*

Then

$$\frac{\widehat{R}(\widehat{\mu}) - R(\widehat{\mu})}{\sigma/\sqrt{n}} \rightsquigarrow N(0,1)$$

where $\sigma^2 = var\{(Y - \overline{\mu}(X))^2\}$.

Proof. Let $\mathbb{P}_n^k$ denote the empirical measure over the kth fold, and let $\mathbb{P}\{f(Z)\} = \int f(z)\, d\mathbb{P}(z)$. Then we have

$$\begin{aligned}
\widehat{R}(\widehat{\mu}) - R(\widehat{\mu}) &= \frac{1}{K}\sum_{k=1}^K (\mathbb{P}_n^k - \mathbb{P})\left\{(Y - \widehat{\mu}_{-k})^2\right\} \\
&= \frac{1}{K}\sum_{k=1}^K (\mathbb{P}_n^k - \mathbb{P})\left\{(Y - \widehat{\mu}_{-k})^2 - (Y - \overline{\mu})^2\right\} + \frac{1}{K}\sum_{k=1}^K (\mathbb{P}_n^k - \mathbb{P})\left\{(Y - \overline{\mu})^2\right\} \\
&\equiv \frac{1}{K}\sum_{k=1}^K T_k + (\mathbb{P}_n - \mathbb{P})\left\{(Y - \overline{\mu})^2\right\}
\end{aligned}$$

Now we show that

$$T_k = O_{\mathbb{P}}\left(\frac{\|\widehat{\mu}_{-k} - \overline{\mu}\|}{\sqrt{n_k}}\right)$$

by bounding the mean and variance and applying Markov's inequality. First note the mean is

$$\mathbb{E}(T_k \mid D^n_{-k}) = \mathbb{E}\left[\mathbb{P}^k_n\Big\{(Y - \widehat{\mu}_{-k})^2 - (Y - \overline{\mu})^2\Big\} \;\Big|\; D^n_{-k}\right] - \mathbb{P}\Big\{(Y - \widehat{\mu}_{-k})^2 - (Y - \overline{\mu})^2\Big\} = 0.$$

For the variance we have

$$\begin{aligned}
\mathrm{var}(T_k \mid D^n_{-k}) &= \mathrm{var}\left[\mathbb{P}^k_n\Big\{(Y - \widehat{\mu}_{-k})^2 - (Y - \overline{\mu})^2\Big\} \;\Big|\; D^n_{-k}\right] \\
&= \left(\frac{1}{n_k}\right) \mathrm{var}\left\{(Y - \widehat{\mu}_{-k})^2 - (Y - \overline{\mu})^2 \;\Big|\; D^n_{-k}\right\} \\
&\leq \left(\frac{1}{n_k}\right) \mathbb{E}\left[\Big\{(Y - \widehat{\mu}_{-k})^2 - (Y - \overline{\mu})^2\Big\}^2 \;\Big|\; D^n_{-k}\right] \\
&= \left(\frac{1}{n_k}\right) \mathbb{E}\left\{\Big(\overline{\mu} - \widehat{\mu}_{-k}\Big)^2\Big(Y - \overline{\mu} + Y - \widehat{\mu}_{-k}\Big)^2 \;\Big|\; D^n_{-k}\right\} \\
&\leq \frac{16B^2\|\widehat{\mu}_{-k} - \overline{\mu}\|^2}{n_k}
\end{aligned}$$

where we used the fact that $(Y - \widehat{\mu}_{-k})^2 - (Y - \overline{\mu})^2 = (\overline{\mu} - \widehat{\mu}_{-k})(Y - \overline{\mu} + Y - \widehat{\mu}_{-k})$ and the boundedness of Y, $\widehat{\mu}$, and $\overline{\mu}$. Now by iterated expectation and Markov's inequality

$$\begin{aligned}
\mathbb{P}\left(\frac{\sqrt{n_k}T_k}{\|\widehat{\mu}_{-k} - \overline{\mu}\|} \geq M_\epsilon\right) &= \mathbb{P}\left(\frac{n_k T_k^2}{\|\widehat{\mu}_{-k} - \overline{\mu}\|^2} \geq M_\epsilon^2\right) \\
&\leq \mathbb{E}\left\{\frac{n_k}{\|\widehat{\mu}_{-k} - \overline{\mu}\|^2 M_\epsilon^2}\mathbb{E}(T_k^2 \mid D^n_{-k})\right\} \leq \frac{16B^2}{M_\epsilon^2}
\end{aligned}$$

Thus we can always take $M_\epsilon = 4B/\sqrt{\epsilon}$ to ensure the probability is no more than any $\epsilon > 0$. Now since $n_k = n/K$ and $\|\widehat{\mu}_{-k} - \overline{\mu}\| = o_{\mathbb{P}}(1)$ we have $T_k = o_{\mathbb{P}}(1/\sqrt{n/K})$. And since K is finite we have $\frac{1}{K}\sum_k T_k = o_{\mathbb{P}}(1/\sqrt{n/K}) = o_{\mathbb{P}}(1/\sqrt{n})$ as well. This implies

$$\widehat{R}(\widehat{\mu}) - R(\widehat{\mu}) = (\mathbb{P}_n - \mathbb{P})\Big\{(Y - \overline{\mu})^2\Big\} + o_{\mathbb{P}}(1/\sqrt{n})$$

which yields the result, by Slutsky's and the CLT. □

The above asymptotic normality result motivates 95% confidence intervals for the true error $R(\widehat{\mu})$ of the form

$$\widehat{R}(\widehat{\mu}) \pm 1.96\widehat{\sigma}/\sqrt{n}$$

where $\widehat{\sigma}^2$ is an appropriate estimator of the variance $\mathrm{var}\{(Y - \overline{\mu})^2\}$, e.g.,

$$\widehat{\mathrm{var}}\left[\Big\{Y_i - \widehat{\mu}_{-\kappa(i)}(X_i)\Big\}^2\right],$$

i.e., the empirical variance of the out-of-sample residuals $\{Y_i - \widehat{\mu}_{-\kappa(i)}(X_i)\}^2$.

For other recent distributional results for cross-validation, see:

1. Austern and Zhou [2020]
2. Bayle et al. [2020]

3 Leave-one-out cross-validation

3.1 Closed-form expression

Theorem 5. *Let $\widehat{\mu}(x) = \sum_i w_i(x; X^n) Y_i$ be a linear smoother and let $\widehat{\mu}_{-i}(x) = \sum_{j \neq i} w_j(x; X^n_{-i}) Y_j$ be the leave-one-out version. If the weights satisfy*

$$1 - \frac{w_j(X_i; X^n)}{w_j(X_i; X^n_{-i})} = w_i(X_i; X^n)$$

then there is a closed-form expression for the leave-one-out estimate given by

$$\widehat{\mu}_{-i}(X_i) = \frac{\widehat{\mu}(X_i) - w_i(X_i; X^n) Y_i}{1 - w_i(X_i; X^n)},$$

which depends only on the full data estimate $\widehat{\mu}(x)$, the weight $w_i(X_i; X^n)$, and the outcome Y_i.

Proof. Note the condition equivalently implies

$$w_j(X_i; X^n_{-i}) = \frac{w_j(X_i; X^n)}{1 - w_i(X_i; X^n)}$$

so that by definition of $\widehat{\mu}_{-i}$ we have

$$\begin{aligned}
\widehat{\mu}_{-i}(X_i) &= \sum_{j \neq i} w_j(X_i; X^n_{-i}) Y_j = \frac{1}{1 - w_i(X_i; X^n)} \sum_{j \neq i} w_j(X_i; X^n) Y_j \\
&= \frac{1}{1 - w_i(X_i; X^n)} \left\{ \sum_{j=1}^{n} w_j(X_i; X^n) Y_j - w_i(X_i; X^n) Y_i \right\} \\
&= \frac{\widehat{\mu}(X_i) - w_i(X_i; X^n) Y_i}{1 - w_i(X_i; X^n)}.
\end{aligned}$$

□

Proposition 1. *The condition*

$$1 - \frac{w_j(X_i; X^n)}{w_j(X_i; X^n_{-i})} = w_i(X_i; X^n)$$

holds for kNN, kernel, local polynomial, series, and centered random forest estimators.

Proof. For kNN estimators we can omit points with zero distance, yielding

$$w_j(x; X^n) = \frac{1}{k} \mathbb{1}\Big\{ 0 < \|X_j - x\| \leq D_k(x; X^n) \Big\} = \frac{1}{k} \mathbb{1}\Big\{ X_j \in \mathcal{N}_k(x) \Big\}$$

so that $w_i(X_i; X^n) = 0$, and the estimators $\widehat{\mu}$ and $\widehat{\mu}_{-i}$ are the same, i.e.,

$$\widehat{\mu}_{-i}(X_i) = \frac{1}{k} \sum_{j \neq i} \mathbb{1}\Big\{ X_j \in \mathcal{N}_k(X_i; X^n_{-i}) \Big\} = \frac{1}{k} \sum_{j=1}^{n} \mathbb{1}\Big\{ X_j \in \mathcal{N}_k(X_i; X^n) \Big\} = \widehat{\mu}(X_i).$$

For kernel estimators we have

$$w_j(x;X^n) = \frac{\frac{1}{nh^d}K\left(\frac{X_j-x}{h}\right)}{\frac{1}{n}\sum_{k=1}^n \frac{1}{h^d}K\left(\frac{X_k-x}{h}\right)} = \frac{K\left(\frac{X_j-x}{h}\right)}{\sum_{k=1}^n K\left(\frac{X_k-x}{h}\right)}$$

so that $w_i(X_i;X^n) = K(0)/\sum_k K\left(\frac{X_k-X_i}{h}\right)$ and

$$1 - \frac{w_j(X_i;X^n)}{w_j(X_i;X^n_{-i})} = 1 - \frac{\sum_{k\neq i}^n K\left(\frac{X_k-X_i}{h}\right)}{\sum_{k=1}^n K\left(\frac{X_k-X_i}{h}\right)} = \frac{K(0)}{\sum_{k=1}^n K\left(\frac{X_k-X_i}{h}\right)} = w_i(X_i;X^n).$$

Note if we had used x instead of X_i we would get

$$1 - \frac{w_j(x;X^n)}{w_j(x;X^n_{-i})} = 1 - \frac{\sum_{k\neq i}^n K\left(\frac{X_k-x}{h}\right)}{\sum_{k=1}^n K\left(\frac{X_k-x}{h}\right)} = \frac{K\left(\frac{X_i-x}{h}\right)}{\sum_{k=1}^n K\left(\frac{X_k-x}{h}\right)} = w_i(x;X^n)$$

which implies $\widehat{\mu}_{-i}(x) = \frac{\widehat{\mu}(x)-w_i(x;X^n)Y_i}{1-w_i(x;X^n)}$ more generally (this doesn't hold for series estimators).

For series estimators we have

$$w_j(x;X^n) = \frac{1}{n}b(x)^{\mathrm{T}}\mathbb{P}_n\{b(X)b(X)^{\mathrm{T}}\}^{-1}b(X_j)$$

so that $w_i(X_i;X^n) = b(X_i)^{\mathrm{T}}\widehat{Q}^{-1}b(X_i)/n$. First note that

$$\widehat{Q}_{-i} = \frac{1}{n-1}\sum_{j\neq i} b(X_j)b(X_j)^{\mathrm{T}} = \left(\frac{n}{n-1}\right)\left\{\widehat{Q} - \frac{1}{n}b(X_i)b(X_i)^{\mathrm{T}}\right\}$$

so that by the Sherman-Morrison formula

$$\widehat{Q}_{-i}^{-1} = \left(\frac{n-1}{n}\right)\left\{\widehat{Q}^{-1} + \frac{\widehat{Q}^{-1}b(X_i)b(X_i)^{\mathrm{T}}\widehat{Q}^{-1}/n}{1-b(X_i)^{\mathrm{T}}\widehat{Q}^{-1}b(X_i)/n}\right\}.$$

Therefore

$$\begin{aligned} b(X_i)^{\mathrm{T}}\widehat{Q}_{-i}^{-1}b(X_j) &= \left(\frac{n-1}{n}\right)b(X_i)^{\mathrm{T}}\widehat{Q}^{-1}b(X_j)\left\{1 + \frac{b(X_i)^{\mathrm{T}}\widehat{Q}^{-1}b(X_i)/n}{1-b(X_i)^{\mathrm{T}}\widehat{Q}^{-1}b(X_i)/n}\right\} \\ &= \left(\frac{n-1}{n}\right)\frac{b(X_i)^{\mathrm{T}}\widehat{Q}^{-1}b(X_j)}{1-b(X_i)^{\mathrm{T}}\widehat{Q}^{-1}b(X_i)/n} \end{aligned}$$

and

$$\begin{aligned} 1 - \frac{w_j(X_i;X^n)}{w_j(X_i;X^n_{-i})} &= 1 - \frac{b(X_i)^{\mathrm{T}}\widehat{Q}^{-1}b(X_j)/n}{b(X_i)^{\mathrm{T}}\widehat{Q}_{-i}^{-1}b(X_j)/(n-1)} \\ &= 1 - \left\{1 - b(X_i)^{\mathrm{T}}\widehat{Q}^{-1}b(X_i)/n\right\} = w_i(X_i;X^n) \end{aligned}$$

A similar argument works for local polynomial estimators. □

Crucially, the above result lets us compute leave-one-out cross validated errors very quickly and easily. Since it implies

$$Y_i - \widehat{\mu}_{-i}(X_i) = \frac{Y_i - \widehat{\mu}(X_i)}{1 - w_i(X_i; X^n)}$$

we have for example that

$$\widehat{R}_{-i}(\widehat{\mu}) = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \widehat{\mu}_{-i}(X_i)\right\}^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Y_i - \widehat{\mu}(X_i)}{1 - w_i(X_i; X^n)}\right\}^2.$$

3.2 Asymptotic normality via stability

Here we adapt an argument from Chen et al. [2022] to show how to obtain the asymptotic distribution of the leave-one-out cross-validated risk for regression estimators that are sufficiently stable.

Theorem 6. *Denote the leave-one-out cross-validated estimated and true risk by*

$$\widehat{R}(\widehat{\mu}) = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \widehat{\mu}_{-i}(X_i)\right\}^2$$
$$R(\widehat{\mu}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\{Y_i - \widehat{\mu}_{-i}(X_i)\right\}^2 \,\middle|\, D^n_{-i}\right].$$

Assume the same conditions as before:

1. $|Y| \leq B$ *and* $\|\widehat{\mu}\|_\infty \leq B$, *and*
2. *there exists a function* $\overline{\mu}(x)$ *such that* $\|\overline{\mu}\|_\infty \leq B$ *and* $\max_i \mathbb{E}\|\widehat{\mu}_{-i} - \overline{\mu}\| = o(1)$

and in addition the stability condition:

3. $\max_i \sqrt{\mathbb{E}\|\widehat{\mu} - \widehat{\mu}_{-i}\|^2} = o\left(\frac{1}{\sqrt{n}}\right)$

Then

$$\frac{\widehat{R}(\widehat{\mu}) - R(\widehat{\mu})}{\sigma/\sqrt{n}} \rightsquigarrow N(0,1)$$

where $\sigma^2 = var\{(Y - \overline{\mu}(X))^2\}$.

Proof. Defining $\widehat{f}_{-i} = \{Y_i - \widehat{\mu}_{-i}(X_i)\}^2$ and $f = \{Y_i - \overline{\mu}(X_i)\}^2$, we have the decomposition

$$\begin{aligned}
\widehat{R}(\widehat{\mu}) - R(\widehat{\mu}) &= \frac{1}{n}\sum_{i=1}^{n}\left(\left\{Y_i - \widehat{\mu}_{-i}(X_i)\right\}^2 - \mathbb{E}\left[\left\{Y_i - \widehat{\mu}_{-i}(X_i)\right\}^2 \,\middle|\, D^n_{-i}\right]\right) \\
&= \mathbb{P}_n\left(\widehat{f}_{-i} - \mathbb{P}\widehat{f}_{-i}\right) \\
&= \mathbb{P}_n\Big\{(\widehat{f}_{-i} - f) - \mathbb{P}(\widehat{f}_{-i} - f)\Big\} + (\mathbb{P}_n - \mathbb{P})f.
\end{aligned}$$

Now let $\varepsilon_i = \widehat{f}_{-i}(Z_i) - f(Z_i) - \int \left\{ \widehat{f}_{-i}(z) - f(z) \right\} d\mathbb{P}(z)$, so that the first term above (call it T_1) equals $T_1 \equiv (\mathbb{P}_n - \mathbb{P})(\widehat{f}_{-i} - f) = \mathbb{P}_n(\varepsilon)$.

We will now show that $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$ by showing that, after scaling by $\sqrt{n}$, its squared mean and variance tend to zero, i.e., that $\{\mathbb{E}(T_1)\}^2 = o(1/n)$ and $\text{var}(T_1) = o(1/n)$.

First note that $\mathbb{E}(\varepsilon_i \mid D^n_{-i}) = 0$ so that $\mathbb{E}(T_1) = 0$. Now the variance $\text{var}(T_1)$ equals

$$\begin{aligned}
\mathbb{E}\Big\{ (\mathbb{P}_n \varepsilon)^2 \Big\} &= \frac{1}{n^2} \left\{ \sum_{i=1}^n \mathbb{E}\left(\varepsilon_i^2\right) + \sum_{i \neq j} \mathbb{E}(\varepsilon_i \varepsilon_j) \right\} \\
&\leq \frac{1}{n} \max_i \text{var}(\varepsilon_i) + \max_{i,j} \mathbb{E}(\varepsilon_i \epsilon_j) \\
&= \max_i \frac{\mathbb{E}\left\{ \text{var}(\epsilon_i \mid D^n_{-i}) \right\}}{n} + \max_{i,j} \mathbb{E}(\varepsilon_i \epsilon_j) \\
&\leq \max_i \frac{\mathbb{E}\|\widehat{f}_{-i} - f\|^2}{n} + \max_{i,j} \mathbb{E}(\varepsilon_i \epsilon_j)
\end{aligned}$$

where the second and third lines use the fact that $\mathbb{E}(\varepsilon_i \mid D^n_{-i}) = 0$, and the last that $\text{var}(\epsilon_i \mid D^n_{-i}) = \text{var}(\widehat{f}_{-i} - f \mid D^n_{-i}) \leq \mathbb{E}\{(\widehat{f}_{-i} - f)^2 \mid D^n_{-i}\} = \|\widehat{f}_{-i} - f\|^2$. Now recall

$$\begin{aligned}
(\widehat{f}_{-i} - f)^2 &= \left[\Big\{ Y - \widehat{\mu}_{-i}(X) \Big\}^2 - \Big\{ Y - \overline{\mu}(X) \Big\}^2 \right]^2 \\
&= \Big\{ \overline{\mu}(X) - \widehat{\mu}_{-i}(X) \Big\}^2 \Big\{ Y - \overline{\mu}(X) + Y - \widehat{\mu}_{-i}(X) \Big\}^2 \\
&\leq (4B)^2 \Big\{ \overline{\mu}(X) - \widehat{\mu}_{-i}(X) \Big\}^2
\end{aligned}$$

so that

$$\|\widehat{f}_{-i} - f\|^2 = \int (\widehat{f}_{-i} - f)^2 \, d\mathbb{P} \leq (4B)^2 \|\widehat{\mu}_{-i} - \overline{\mu}\|^2.$$

Therefore the first term above is $o(1/n)$ by consistency of $\widehat{\mu}_{-i}$ for $\overline{\mu}$. For the second term, we use Chen et al.'s double centering argument. First let

$$\varepsilon_i^{-j} = \widehat{f}_{-i,j}(Z_i) - f(Z_i) - \int \left\{ \widehat{f}_{-i,j}(z) - f(z) \right\} d\mathbb{P}(z)$$

where $\widehat{f}_{-i,j}(z)$ is the estimator trained on all observations except (i, j). Note ϵ_i^{-j} is only a function of $(Z_i, D^n_{-i,j}) = D^n_{-j}$, and has mean zero given $D^n_{-i,j}$, so that

$$\begin{aligned}
\mathbb{E}\left(\epsilon_i^{-j} \epsilon_j \right) &= \mathbb{E}\left\{ \mathbb{E}\left(\epsilon_i^{-j} \epsilon_j \mid D^n_{-j} \right) \right\} = \mathbb{E}\left\{ \epsilon_i^{-j} \mathbb{E}\left(\epsilon_j \mid D^n_{-j} \right) \right\} = 0 \\
\mathbb{E}\left(\epsilon_i^{-j} \epsilon_j^{-i} \right) &= \mathbb{E}\left\{ \mathbb{E}\left(\epsilon_i^{-j} \epsilon_j^{-i} \mid D^n_{-i,j} \right) \right\} = \mathbb{E}\left\{ \mathbb{E}\left(\epsilon_j^{-i} \mid D^n_{-i,j} \right) \mathbb{E}\left(\epsilon_i^{-j} \mid D^n_{-i,j} \right) \right\} = 0.
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{E}(\epsilon_i \epsilon_j) &= \mathbb{E}\left\{ \left(\epsilon_i - \epsilon_i^{-j} \right) \left(\epsilon_j - \epsilon_j^{-i} \right) \right\} \\
&\leq \sqrt{\mathbb{E}\|\epsilon_i - \epsilon_i^{-j}\|^2} \sqrt{\mathbb{E}\|\epsilon_j - \epsilon_j^{-i}\|^2} = o\left(\frac{1}{n} \right)
\end{aligned}$$

where we used Cauchy-Schwarz and in the last line that

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left\{\left(\epsilon_i - \epsilon_i^{-j}\right)^2 \mid D_{-i}^n\right\}\right] &= \mathbb{E}\left\{\text{var}\left(\epsilon_i - \epsilon_i^{-j} \mid D_{-i}^n\right)\right\} \\ &= \mathbb{E}\left\{\text{var}\left(\widehat{f}_{-i}(Z_i) - \widehat{f}_{-i,j}(Z_i) \mid D_{-i}^n\right)\right\} \\ &\leq \mathbb{E}\|\widehat{f}_{-i} - \widehat{f}_{-i,j}\|^2 \leq (4B)^2\mathbb{E}\|\widehat{\mu}_{-i} - \widehat{\mu}_{-i,j}\|^2\end{aligned}$$

together with the stability condition (3). □

Remark 1. It is natural to ask what kinds of estimators satisfy the stability condition above, i.e., that

$$\mathbb{E}\|\widehat{\mu} - \widehat{\mu}_{-i}\|^2 = o(1/n).$$

First note this is a somewhat weaker stability condition than that from Soloff et al. [2024], which we explored when discussing bagging. Chen et al. [2022] show that generic bagged estimators satisfy the condition if the number of observations per bag satisfies $m = o(\sqrt{n})$ and if the number of bags B is large enough.

4 Technical lemmas

Lemma 2 (Maximal inequality for sub-Gaussians)**.** *Let Z_j, $j = 1, ..., d$ be mean-zero random variables, possibly dependent, but sub-Gaussian with parameter σ, i.e., $\mathbb{E}\{\exp(tZ_j)\} \leq \exp(t^2\sigma^2/2)$ for all t. Then*

$$\mathbb{E}\left(\max_j |Z_j|\right) \leq 2\sigma\sqrt{\log d}.$$

Lemma 3 (Hoeffding's inequality)**.** *Let $Z_1, ..., Z_n$ be independent random variables, with $Z_i \in [a, b]$ with probability one. Then for all $t > 0$*

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n (Z_i - \mathbb{E}(Z_i))\right| > t\right\} \leq 2\exp\left\{\frac{-nt^2}{(b-a)^2/2}\right\}.$$

Lemma 4 (Bernstein's inequality)**.** *Let $Z_1, ..., Z_n$ be independent random variables, with $Z_i \in [a, b]$ and $\frac{1}{n}\sum_i var(Z_i) = \sigma^2 > 0$. Then for all $t > 0$*

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n (Z_i - \mathbb{E}(Z_i))\right| > t\right\} \leq 2\exp\left\{\frac{-nt^2}{2\sigma^2 + 2t(b-a)/3}\right\}.$$

Lemma 5 (Berry-Esseen inequality)**.** *Let $Z_1, ..., Z_n$ be iid random variables with $var(Z_i) = \sigma^2$ and $\mathbb{E}(|Z_i|^3) = \rho$. Then*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\frac{1}{n}\sum_i (Z_i - \mathbb{E}(Z_i))}{\sigma/\sqrt{n}} \leq t\right\} - \Phi(t)\right| \leq \frac{C\rho/\sigma^3}{\sqrt{n}}$$

for some $C \in (0.4, 0.5)$ and Φ the cdf of the standard normal distribution.

References

M. Austern and W. Zhou. Asymptotics of cross-validation. *arXiv preprint arXiv:2001.11111*, 2020.

P. Bayle, A. Bayle, L. Janson, and L. Mackey. Cross-validation confidence intervals for test error. *Advances in Neural Information Processing Systems*, 33:16339–16350, 2020.

Q. Chen, V. Syrgkanis, and M. Austern. Debiased machine learning without sample-splitting for stable estimators. *Advances in Neural Information Processing Systems*, 35:3096–3109, 2022.

S. Dudoit and M. J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical methodology*, 2(2):131–154, 2005.

L. Györfi, M. Kohler, A. Krzykaz, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

J. A. Soloff, R. F. Barber, and R. Willett. Bagging provides assumption-free stability. *Journal of Machine Learning Research*, 25(131):1–35, 2024.