Cross-Validation Estimates IMSE

Mark Plutowski *
Department of Computer Science and Engineering
and Institute for Neural Computation
University of California, San Diego

Shinichi Sakata
Department of Economics
University of California, San Diego

Halbert White
Department of Economics
and Institute for Neural Computation
University of California, San Diego

Abstract

Integrated Mean Squared Error (IMSE) is a version of the usual mean squared error criterion, averaged over all possible training sets of a given size. If it could be observed, it could be used to determine optimal network complexity or optimal data subsets for efficient training. We show that two common methods of cross-validating average squared error deliver unbiased estimates of IMSE, converging to IMSE with probability one. These estimates thus make possible approximate IMSE-based choice of network complexity. We also show that two variants of cross validation measure provide unbiased IMSE-based estimates potentially useful for selecting optimal data subsets.

1 Summary

Suppose we are given a fixed network architecture. (We dispense with this assumption later.) Let z^N denote a given set of N training examples. Let $Q_N(z^N)$ denote the expected squared error (the expectation taken over all possible examples) of the network after being trained on z^N. This measures the quality of fit afforded by training on a given set of N examples.

Let $IMSE_N$ denote the Integrated Mean Squared Error for training sets of size N. Given reasonable assumptions, it is straightforward to show that $IMSE_N = E[Q_N(Z^N)] - \sigma^2$, where the expectation is now over all training sets of size N, Z^N is a random training set of size N, and σ^2 is the noise variance.

Let $C_N = C_N(z^N)$ denote the "delete-one cross-validation" squared error measure for a network trained on z^N. C_N is obtained by training networks on each of the N training sets of size $N-1$ obtained by deleting a single example; the measure follows by computing squared error for the corresponding deleted example and averaging the results. Let $G_{N,M} = G_{N,M}(z^N, \tilde{z}^M)$ denote the "generalization" measure obtained by separating the available data of size $N+M$ into a training set z^N of size N, and a validation ("test") set $\tilde{z}^M$ of size M; the measure follows by training on z^N and computing averaged squared error over $\tilde{z}^M$.

We show that C_N is an unbiased estimator of $E[Q_{N-1}(Z^N)]$, and hence, estimates $IMSE_{N-1}$ up to noise variance. Similarly, $G_{N,M}$ is an unbiased estimator of $E[Q_N(Z^N)]$. Given reasonable conditions on the estimator and on the data generating process we demonstrate convergence with probability 1 of $G_{N,M}$ and C_N to $E[Q_N(Z^N)]$ as N and M grow large.

A direct consequence of these results is that when choice is restricted to a set of network architectures whose complexity is bounded above a priori, then choosing the architecture for which either C_N (or $G_{N,M}$) is minimized leads to choice of the network for which $IMSE_N$ is nearly minimized for all N (respectively, N, M) sufficiently large.

*Contact Author: Plutowski. Email: pluto@cs.ucsd.edu.

We also provide results for training sets sampled at particular inputs. Conditional $IMSE$ is an appealing criterion for evaluating a particular choice of training set in the presence of noise. These results demonstrate that delete-one cross-validation estimates average MSE (the average taken over the given set of inputs,) and that hold-out set cross-validation gives an unbiased estimate of $E[Q_N(Z^N)|Z^N = (x^N, Y^N)]$, given a set of N input values x^N for which corresponding (random) output values Y^N are obtained. Either cross-validation measure can therefore be used to select a representative subset of the entire dataset that can be used for data compression, or for more efficient training (as training can be faster on smaller datasets) [2, 4].

2 Definitions

2.1 Learning Task

We consider the learning task of determining the relationship between a random vector X and a random scalar Y, where X takes values in a subset $\mathbf{X}$ of $\Re^r$, (r a positive integer) and Y takes values in a subset $\mathbf{Y}$ of $\Re$. (e.g. $\mathbf{X} = \Re^r$ and $\mathbf{Y} = \Re$). We refer to $\mathbf{X}$ as the *input space*. The learning task is thus one of training a neural network with r inputs and one output. It is straightforward to extend the following analysis to networks with multiple targets.

We make the following assumption on the observations to be used in the training of the networks.

Assumption 1 *$\mathbf{X}$ is a Borel subset of $\Re^r$ and $\mathbf{Y}$ is a Borel subset of $\Re$. Let $\mathbf{Z} \equiv \mathbf{X} \times \mathbf{Y}$ and $\Omega \equiv \mathbf{Z}^\infty \equiv \times_{i=1}^{\infty} \mathbf{Z}$. Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space with $\mathcal{F} = \mathcal{B}(\Omega)$, the Borel σ-field generated by the measurable finite dimensional product cylinders. The observations on $Z \equiv (X', Y)'$ to be used in the training of the network are a realization of an i.i.d. stochastic process $\{Z_i \equiv (X_i', Y_i)' : \Omega \to \mathbf{X} \times \mathbf{Y}\}$.*

When $\omega \in \Omega$ is fixed, we write $z_i \equiv Z_i(\omega)$ for each $i = 1, 2, \ldots$. Also write $Z^N \equiv (Z_1, \ldots, Z_N)$ and $z^n \equiv (z_1, \ldots, z_n)$.

Assumption 1 allows uncertainty caused by measurement errors of observations as well as a *probabilistic* relationship between X and Y . It, however, does not prevent a deterministic relationship between X and Y such that $Y = g(X)$ for some measurable mapping $g : \Re^r \to \Re$.

We suppose interest attaches to the conditional expectation of Y given X, written $g(X) = E(Y|X)$. The next assumption guarantees the existence of $E(Y_i|X_i)$ and $E(\varepsilon_i|X_i)$, $\varepsilon_i \equiv Y_i - E(Y_i|X_i)$, as well as the existence of the variance of ε_i.

Assumption 2 *$E(Y^2) < \infty$.*

For convenience, we next assume the conditional homoskedasticity of Y_i given X_i.

Assumption 3 *$E(\varepsilon_1^2|X_1) = \sigma^2$, where σ^2 is a strictly positive constant.*

2.2 Network Model

Let $f^p(\cdot, \cdot) : \mathbf{X} \times \mathbf{W}^p \to \mathbf{Y}$ be a network function with "weight space" $\mathbf{W}^p$, where p denotes the dimension of the "weight space" (the number of weights). We impose some mild conditions on the network architecture.

Assumption 4 *For each $p \in \{1, 2, \ldots, \bar{p}\}$, $\bar{p} \in \mathbf{N}$, $\mathbf{W}^p$ is a compact subset of $\Re^p$, and $f^p : \mathbf{X} \times \mathbf{W}^p \to \Re$ satisfies the following conditions:*

1. *$f^p(\cdot, w) : \mathbf{X} \to \mathbf{Y}$ is measurable for each $w \in \mathbf{W}^p$;*
2. *$f^p(x, \cdot) : \mathbf{W}^p \to \mathbf{Y}$ is continuous for all $x \in \mathbf{X}$.*

We further make a joint assumption on the underlying data generating process and the network architecture to assure that the training dataset and the networks behaves appropriately.

Assumption 5 *There exists a function* $D : \mathbf{X} \to \Re_+ \equiv [0, \infty)$ *such that for each* $x \in \mathbf{X}$ *and* $w \in \mathbf{W}^p$, $|f^p(x,w)| \le D(x)$, *and* $E\left[(D(X))^2\right] < \infty$.

Hence, $f^p(\cdot, w)$ is square integrable for each $w \in \mathbf{W}^p$.

We will measure network performance using mean squared error, which for weights w is given by

$$\lambda(w;p) \equiv E\left[(Y - f^p(X,w)^2\right]. \tag{1}$$

The optimal weights are those that minimize $\lambda(w;p)$. The set of all optimal weights is given by $\mathbf{W}^{p*} \equiv \{w^* \in \mathbf{W}^p : \lambda(w^*;p) \le \lambda(w;p)$ for any $w \in \mathbf{W}^p\}$. The index of the best network is p^*, given by the smallest p minimizing $\min_{w \in \mathbf{W}^p} \lambda(w;p)$, $p \in \{1, 2, \ldots, \bar{p}\}$.

2.3 Least-Squares Estimator

When Assumptions 1 and 4 hold, the nonlinear least-squares estimator exists. Formally, we have

Lemma 1 *Suppose that Assumptions 1 and 4 hold. Then*

1. *For each* $N \in \mathbf{N}$, *there exists a measurable function* $l_N(\cdot;p) : \mathbf{Z}^N \to \mathbf{W}^p$ *such that* $l_N(Z^N;p)$ *solves the following problem with probability one:*

$$\min_{w \in \mathbf{W}^p} N^{-1} \sum_{i=1}^{N} (Y_i - f^p(X_i, w))^2.$$

Suppose in addition that Assumptions 2 and 5 hold. Then

2. $\lambda(\cdot;p) : \mathbf{W}^p \to \Re$ *is continuous on* $\mathbf{W}^p$, *and* $\mathbf{W}^{p*}$ *is not empty.*

For convenience, we also define $\hat{w}_N^p : \Omega \to \mathbf{W}^p$ by

$$\hat{w}_N^p(\omega) \equiv l_N(Z^N(\omega);p)$$

for each $\omega \in \Omega$.

Next let $i_1, i_2, \ldots, i_N$ be distinct natural numbers and let $\breve{Z}^N = (Z_{i_1}, \ldots, Z_{i_N})'$. Then $l_N(\breve{Z}^N)$ given above solves

$$\frac{1}{N} \sum_{j=1}^{N} (Y_{i_j} - f^p(X_{i_j}, w))^2$$

with probability one. In particular, we will consider the estimate using the dataset z_{-i}^N made by deleting the ith observation from z^N. Let Z_{-i}^N be a random matrix made by deleting the ith row from Z^N. Thus, $l_{N-1}(Z_{-i}^N;p)$ is a measurable least squares estimator and we can consider its probabilistic behavior.

3 Integrated Mean Squared Error

Integrated Mean Squared Error (IMSE) has been used to regulate network complexity [9]. Another (conditional) version of IMSE is used as a criterion for evaluating training examples [5, 6, 7, 8]. The first version depends only on the sample size, not the particular sample. The second (conditional) version depends additionally upon the observed location of the examples in the input space.

3.1 Unconditional IMSE

The (unconditional) mean squared error (MSE) of the network output at a particular input value x is

$$M_N(x;p) = E\left[\{g(x) - f^p(x, l_N(Z^N;p))\}^2\right]. \tag{2}$$

Evaluating the expectation of M_N over the input space gives the unconditional IMSE:

$$IMSE_N(p) \quad = \quad E\left[M_N(X;p)\right], \tag{3}$$

where μ is the input distribution.

3.2 Conditional IMSE

To evaluate exemplars obtained at inputs x^N, we modify Equation (2) by conditioning on x^N, giving

$$M_N(x|x^N;p) \quad = \quad E\left[\{g(x) - f^p(x, l_N(Z^N;p))\}^2 | X^N = x^N\right],$$

The conditional IMSE (given inputs x^N) is then

$$IMSE_N(x^N;p) \quad = \quad E\left[M_N(X|x^N;p)\right]. \tag{4}$$

4 Cross-Validation

Cross-validatory measures have been used successfully to assess the performance of a wide range of estimators [10, 11, 12, 13, 14, 15]. Cross-validatory measures have been derived for various performance criteria, including the Kullback-Liebler Information Criterion (KLIC) and the Integrated Squared Error (ISE, asymptotically equivalent to IMSE) [16]. Although provably inappropriate in certain applications [17, 18], optimality and consistency results for the cross-validatory measures have been obtained for several estimators, including linear regression, orthogonal series, splines, histograms, and kernel density estimators [16, 19, 20, 21, 22, 23, 24]. The authors are not aware of similar results applicable to neural networks, although two more general, but weaker results do apply [26]. A general result applicable to neural networks shows asymptotic equivalence between cross-validation and Akaike's Criterion for network selection [25, 30], as well as between cross-validation and Moody's Criterion [31, 30].

4.1 Expected Network Error

Given our assumptions, we can relate cross-validation to IMSE. For clarity and notational convenience, we first introduce a measure of network error closely related to IMSE.

For each weight $w \in \mathbf{W}^p$, we have defined the mean squared error $\lambda(w;p)$ in equation 1. We define Q_N to map each dataset to the mean squared error of the estimated network

$$Q_N(z^N;p) \equiv \lambda(l_N(z^N;p);p).$$

When Assumption 3 holds, we have

$$\begin{aligned}
\lambda(w;p) &= E\left[(g(X) - f^p(X,w))^2\right] + \sigma^2 \\
&= E\left[(g(X_{N+1}) - f^p(X_{N+1},w))^2\right] + \sigma^2
\end{aligned}$$

as is easily verified. We therefore have by independence that

$$Q_N(z^N;p) \quad = \quad E\left[(g(X_{N+1}) - f^p(X_{N+1}, l_N(Z^N;p)))^2 | Z^N = z^N\right] + \sigma^2.$$

Thus, by using the law of iterated expectations, we have

$$\begin{aligned}
E\left[Q_N(Z^N;p)\right] &= E\left[E\left[Q_N(Z^N;p)|Z^N\right]\right] \\
&= E\left[E\left[(g(X_{N+1}) - f^p(X_{N+1}, l_N(Z^N;p)))^2|Z^N\right] + \sigma^2\right] \\
&= E\left[(g(X_{N+1}) - f^p(X_{N+1}, l_N(Z^N;p)))^2\right] + \sigma^2 \\
&= E\left[E\left[(g(X_{N+1}) - f^p(X_{N+1}, l_N(Z^N;p)))^2|Z_{N+1}\right]\right] + \sigma^2 \\
&= E[M_N(X_{N+1};p)] + \sigma^2 \\
&= E[M_N(X;p)] + \sigma^2 \\
&= IMSE_N(p) + \sigma^2.
\end{aligned}$$

Likewise, given $x^N \in \mathbf{X}^N$,

$$E[Q_N(Z^N;\, p)|X^N \;=\; x^N] = IMSE(x^N;p) + \sigma^2. \tag{5}$$

4.2 Cross-Validatory Estimation of Error

In practice we work with observable quantities only. In particular, we must estimate the error of network p over novel data ("generalization") from a finite set of examples. Such an estimate is given by the delete-one cross-validation measure:

$$C_N(z^N;p) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - f^p(x_i, l_{N-1}(z_{-i}^N;p))\right)^2, \tag{6}$$

where z_{-i}^N denotes the training set obtained by deleting the ith example. Using z_{-i}^N instead of z^N avoids a downward bias due to testing upon examples used in training, as we show below (Theorem 3).

Another version of cross-validation is commonly used for evaluating "generalization" when an abundant supply of novel data is available for use as a "hold-out" set:

$$G_{N,M}(z^N, \tilde{z}^M;p) = \frac{1}{M}\sum_{i=1}^{M}\left(\tilde{y}_i - f^p(\tilde{x}_i, l_N(z^N;p))\right)^2, \tag{7}$$

where $\tilde{z}^M = (z_{N+1}, \ldots, z_{N+M})$.

5 Expectation of the Cross-Validation Measures

We now consider the relation between the cross-validation measure and IMSE. We examine delete-one cross-validation first.

Proposition 1 (Unbiasedness of C_N) *Let Assumptions 1 through 5 hold. Then for given N, C_N is an unbiased estimator of $IMSE_{N-1}(p) + \sigma^2$:*

$$E\left[C_N(Z^N;p)\right] = IMSE_{N-1}(p) + \sigma^2. \tag{8}$$

With hold-out set cross-validation, one obtains an assessment of potential network performance independent of the training set.

Proposition 2 (Unbiasedness of $G_{N,M}$) *Let Assumptions 1 through 5 hold. Let $\tilde{Z}^M \equiv (Z_{N+1}, \ldots, Z_{N+M})$. Then for given N and M, $G_{N,M}$ is an unbiased estimator of $IMSE_N(p) + \sigma^2$:*

$$E\left[G_{N,M}(Z^N, \tilde{Z}^M;p)\right] = IMSE_N(p) + \sigma^2. \tag{9}$$

The latter approach is appealing for large M, N. We expect delete-one cross-validation to be more appealing than hold-out set cross-validation, when training data are not abundant.

6 Expectation of Cross-Validation when Sampling at Selected Inputs

We obtain analogous results for training sets obtained by sampling at a given set of inputs x^N. We first consider the result for delete-one cross-validation.

Proposition 3 (Expectation of C_N given x^N) *Let Assumptions 1 through 5 hold. Then, given a particular set of inputs, x^N, C_N is an unbiased estimator of average $MSE_{N-1} + \sigma^2$, the average taken conditional on x^N:*

$$E\left[C_N(Z^N,p)\middle|\, X^N = x^N\right] \;=\; \frac{1}{N}\sum_{i=1}^{N} M_{N-1}(x_i|x_{-i}^N;p) \;+\; \sigma^2,$$

where x_{-i}^N is a matrix made by deleting the ith row of x^N.

This essentially gives an estimate of MSE_{N-1} limited to $x \in x^N$, losing a degree of freedom while providing no estimate of the MSE away from the training points. For this average to converge to $IMSE_{N-1}$, it will suffice for the empirical distribution of x^N, $\bar{\mu}_N$, to converge to μ, i.e., $\bar{\mu}_N \Rightarrow \mu$.

We obtain a stronger result for hold-out set cross-validation. The hold-out set gives independent information on MSE_N, resulting in an estimate of $IMSE_N$ for given x^N.

Proposition 4 (Expectation of $G_{N,M}$ given x^N) *Let Assumptions 1 through 5 hold. Let $\tilde{Z}^M \equiv (Z_{N+1}, \ldots, Z_{N+M})'$. Then, given a particular set of inputs, x^N, $G_{N,M}$ is an unbiased estimator of of $IMSE_N(x^N;p) + \sigma^2$:*

$$E\left[G_{N,M}(Z^N, \tilde{Z}^M; p)\middle| X^N = x^N\right] \quad = \quad IMSE_N(x^N;p) \ + \sigma^2.$$

7 Strong Convergence of Hold-Out Set Cross-Validation

Our conditions deliver not only unbiasedness, but also convergence of hold-out set cross-validation to $IMSE_N$, with probability 1.

Theorem 1 (Convergence of Hold-Out Set w.p. 1) *Let Assumptions 1 through 5 hold. Also let $\tilde{Z}^M \equiv (Z_{N+1}, \ldots, Z_{N+M})$. If for some $A > 0$ a sequence $\{M_N\}$ of natural numbers satisfies $M_N > AN$ for all $N = 1, 2, \ldots$, then*

$$G_{N,M_N}(Z^N, \tilde{Z}^{M_N}; p) \ - \ E[Q_N(Z^N;p)] \ \xrightarrow{a.s.} \ 0, \text{ as } N \to \infty.$$

8 Strong Convergence of Delete-One Cross-Validation

Given an additional condition (uniqueness of optimal weights) we can show strong convergence for delete-one cross-validation.

First we establish uniform convergence of the estimators $\hat{w}^p(Z^N_{-i})$ to optimal weights (uniformly over $1 \le i \le N$.) The following lemma is useful for this purpose.

Lemma 2 *Let Θ be a compact set. Suppose that $\{h_i : \Theta \to \Re\}$ is a sequence of non-stochastic functions such that*

$$\sup_{\theta \in \Theta} \left| N^{-1} \sum_{i=1}^{N} h_i(\theta) \right| \to 0 \text{ as } N \to \infty.$$

Then

$$\max_{k \in (1,\ldots,N)} \max_{\theta \in \Theta} \left| (N-1)^{-1} \sum_{1 \le i \le N, i \ne k} h_i(\theta) \right| \to 0 \text{ as } N \to \infty.$$

Using this lemma, we can show the next theorem about convergence of $l_{N-1}(Z^N_{-i};p)$ uniformly in i.

Theorem 2 *Let Assumptions 1 through 5 hold. Let Z^N_{-k} be the dataset made by deleting the kth observation from Z^N. Then*

$$\max_{1 \le i \le N} d\left(l_{N-1}(Z^N_{-i};p), \mathbf{W}^{p*}\right) \to 0 \ a.s.\text{-}\mathcal{P} \text{ as } N \to \infty, \tag{10}$$

where $d(w, \mathbf{W}^{p*}) \equiv \inf_{w^* \in \mathbf{W}^{p*}} \|w - w^*\|$.

This convergence result leads to the next result that the delete-one cross validation measure does not under-estimate the optimized MSE, namely, $\inf_{w \in \mathbf{W}^p} \lambda(w;p)$.

Theorem 3 *Let Assumptions 1 through 5 hold. Define*

$$C_N(z^N;p) \equiv N^{-1} \sum_{i=1}^{N} \left(y_i - f^p\left(x_i, l_{N-1}\left(z_{-i}^N;p\right)\right)\right)^2.$$

Then

$$\liminf_{N\to\infty} C_N(Z^N;p) \geq \min_{w\in\mathbf{W}^p} \lambda(w;p) \ \ a.s.\text{-}\mathcal{P}.$$

When the optimum weight w^* is unique, we have a stronger result about convergence of the delete-one cross validation measure.

Assumption 6 $\mathbf{W}^{p*}$ *is a singleton, i.e.,* $\mathbf{W}^{p*}$ *has only one element.*

Theorem 4 *Let Assumptions 1 through 6 hold. Then*

$$C_N\left(Z^N;p\right) - E\left[Q_N(Z^N;p)\right] \to 0 \ \ a.s. \ \ as\ N\to\infty.$$

9 Conclusion

Our results justify the intuition that cross-validation measures unbiasedly and consistently estimate the expected squared error of networks trained on finite training sets, therefore providing means of obtaining *IMSE*-approximate methods of selecting appropriate network architectures, or for evaluating a particular choice of training set.

Use of these cross-validation measures therefore permits us to avoid underfitting the data, asymptotically. Note, however, that although we also thereby avoid overfitting asymptotically, this avoidance is not necessarily accomplished by choosing a minimally complex architecture. The possibility exists that $IMSE_{N-1}(p) = IMSE_{N-1}(p+1)$. Because our cross-validated estimates of these quantities are random we may by chance observe $C_N(Z^N;p) > C_N(Z^N;p+1)$ and therefore select the more complex network, even though the less complex network is equally good. Of course, because the IMSE's are the same, no performance degradation (overfitting) will result in this solution.

Acknowledgements

This work was supported by NSF grant IRI 92-03532. We thank David Wolpert, Jan Larsen, Jeff Racine, Vjachislav Krushkal, and Patrick Fitzsimmons for valuable discussions.

References

[1] White, H. 1989. "Learning in Artificial Neural Networks: A Statistical Perspective." *Neural Computation*, **1** 4, pp.425-464. MIT Press, Cambridge, MA.

[2] Plutowski, Mark E., and Halbert White. 1991. "Active selection of training examples for network learning in noiseless environments." *Technical Report No. CS91-180,* Department of Computer Science and Engineering, The University of California, San Diego. La Jolla, California.

[3] Plutowski, Mark E., and Halbert White. 1993. "Selecting concise training sets from clean data." *IEEE Transactions on Neural Networks*. **4**, 3, pp.305-318.

[4] Plutowski, Mark E., Garrison Cottrell, and Halbert White. 1993. "Learning Mackey-Glass from 25 examples, Plus or Minus 2." Presented at 1992 Neural Information Processing Systems conference. To appear in Giles, C.L., Hanson, S.J., and Cowan, J.D. (eds.), *Advances in neural information processing systems 6*, San Mateo, CA: Morgan Kaufmann Publishers.

[5] Fedorov, V.V. 1972. **Theory of Optimal Experiments**. Academic Press, New York.

[6] Box,G., and N.Draper. 1987. **Empirical Model-Building and Response Surfaces.** Wiley, New York.

[7] Khuri, A.I., and J.A.Cornell. 1987. **Response Surfaces (Designs and Analyses)**. Marcel Dekker, Inc., New York.

[8] Faraway, Julian J. 1990. "Sequential design for the nonparametric regression of curves and surfaces." Technical Report #177, Department of Statistics, The University of Michigan.

[9] Geman, Stuart, Elie Bienenstock, and René Doursat. 1992. "Neural networks and the bias/variance dilemma." *Neural Computation*. **4**, 1, 1-58.

[10] Stone, M. 1959. "Application of a measure of information to the design and comparison of regression experiments." *Annals Math. Stat.* **30** 55-69

[11] "Cross-validatory choice and assessment of statistical predictions." *J.R. Statist. Soc. B.* **36**, 2, 111-47.

[12] Bowman, Adrian W. 1984. "An alternative method of cross-validation for the smoothing of density estimates." *Biometrika* (1984), **71**, 2, pp. 353-60.

[13] Bowman, Adrian W., Peter Hall, D.M. Titterington. 1984. "Cross-validation in nonparametric estimation of probabilities and probability densities." *Biometrika* (1984), **71**, 2, pp. 341-51.

[14] Marron, M. 1987. "A comparison of cross-validation techniques in density estimation." *The Annals of Statistics*. **15**, 1, 152-162.

[15] Wahba, Grace. 1990. **Spline Models for Observational Data**. v. 59 in the CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, Philadelphia, PA, March 1990. Soft-cover, 169 pages, bibliography, author index. ISBN 0-89871-244-0

[16] Hall, Peter. 1983. "Large sample optimality of least squares cross-validation in density estimation." *The Annals of Statistics*. **11**, 4, 1156-1174.

[17] Schuster, E.F., and G.G. Gregory. 1981. "On the nonconsistency of maximum likelihood nonparametric density estimators". *Comp.Sci. & Statistics: Proc. 13th Symp. on the Interface.* W.F. Eddy ed. 295-298. Springer-Verlag.

[18] Chow, Y.-S. and S.Geman, L.-D. Wu. 1983. "Consistent cross-validated density estimation." *The Annals of Statistics*. **11**, 1, 25-38.

[19] Bowman, Adrian W. 1980. "A note on consistency of the kernel method for the analysis of categorical data." *Biometrika* (1980), **67**, 3, pp. 682-4.

[20] Stone, Charles J. 1984 "An asymptotically optimal window selection rule for kernel density estimates." *The Annals of Statistics*. **12**, 4, 1285-1297.

[21] Li, Ker-Chau. 1986. "Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing." *The Annals of Statistics*. **14**, 3, 1101-1112.

[22] Li, Ker-Chau. 1987. "Asymptotic optimality for C_p, C_L, cross-validation, and generalized cross-validation: discrete index set." *The Annals of Statistics*. **15**, 3, 958-975.

[23] Utreras, Florencio I. 1987. "On generalized cross-validation for multivariate smoothing spline functions." *SIAM J. Sci. Stat. Comput.* **8**, 4, July 1987.

[24] Andrews, Donald W.K. 1991. "Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors." *Journal of Econometrics*. **47** (1991) 359-377. North-Holland.

[25] Stone, M. 1977. "An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion." *J. Roy. Stat. Soc. Ser B*, **39**, 1, 44-47.

[26] Stone, M. 1977. "Asymptotics for and against cross-validation." *Biometrika*. **64**, 1, 29-35.

[27] Folland, Gerald B. 1984. **Real Analysis**. Wiley. New York.

[28] Billingsley, Patrick. 1986. **Probability and Measure.** Wiley, New York.

[29] Jennrich, R. 1969. "Asymptotic properties of nonlinear least squares estimators." *Ann. Math. Stat.* **40**, 633-643.

[30] Liu, Yong. 1993. "Neural network model selection using asymptotic jackknife estimator and cross-validation method." In Giles, C.L., Hanson, S.J., and Cowan, J.D. (eds.), *Advances in neural information processing systems 5*, San Mateo, CA: Morgan Kaufmann Publishers.

[31] Moody, John E. 1992. "The effective number of parameters, an analysis of generalization and regularization in nonlinear learning system." In Moody, J.E., Hanson, S.J., and Lippmann, R.P., (eds.), *Advances in neural information processing systems 4*, San Mateo, CA: Morgan Kaufmann Publishers.

[32] Bailey, Timothy L. and Charles Elkan. 1993. "Estimating the accuracy of learned concepts." *To appear in Proc. International Joint Conference on Artificial Intelligence.*

[33] White, Halbert. 1993. **Estimation, Inference, and Specification Analysis**. *Manuscript.*

[34] White, Halbert. 1984. **Asymptotic Theory for Econometricians**. Academic Press.

[35] Sussman, H.J. 1992. "Uniqueness of the weights for minimal feedforward nets with a given input-output map." *Neural Networks* **5**, 589-593.

[36] Albertini, Francesca, Eduardo D. Sontag, and Vincent Maillot. "Uniqueness of weights for neural networks." 1993. To appear in *Artificial Neural Networks with Applications in Speech and Vision* (R. Mammone, ed.), Chapman and Hall, London.

[37] Klein, Erwin and Anthony C. Thompson. 1984 **Theory of correspondences : including applications to mathematical economics.** Wiley.

Mathematical Appendix

Proof of Lemma 1 Part 1 follows from Lemma 2 of Jennrich [29] and a measurable selection theorem. Part 2 follows from the Lebesgue dominated convergence theorem with Assumptions 1 through 5. That $\mathbf{W}^{p*}$ is non-empty follows because λ is continuous and $\mathbf{W}^p$ is compact. $\square$

Proof of Proposition 1 By the law of iterated expectations and the results of subsection 4.1, we have

$$\begin{aligned} E\left[C_N(Z^N;p)\right] &= \frac{1}{N}\sum_{i=1}^{N} E\left[\left(Y_i - f^p(X_i, l_{N-1}(Z_{-i}^N;p))\right)^2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} E\left[E\left[\left(Y_i - f^p(X_i, l_{N-1}(Z_{-i}^N;p))\right)^2 |Z_{-i}^N\right]\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} E\left[Q_{N-1}(Z_{-i}^N;p)\right] \\ &= IMSE_{N-1}(p) + \sigma^2. \ \square \end{aligned}$$

Proof of Proposition 2 The proof is similar to that for Proposition 1:

$$\begin{aligned} E\left[G_{N,M}(Z^N, Z^M;p)\right] &= \frac{1}{M}\sum_{i=N+1}^{N+M} E\left[\left(Y_i - f^p(X_i, l_N(Z^N;p))\right)^2\right] \\ &= \frac{1}{M}\sum_{i=N+1}^{N+M} E\left[E\left[\left(Y_i - f^p(X_i, l_N(Z^N;p))\right)^2 |Z^N\right]\right] \\ &= \frac{1}{M}\sum_{i=N+1}^{N+M} E\left[Q_N(Z^N;p)\right] \\ &= IMSE_N(p) + \sigma^2. \ \square \end{aligned}$$

Proof of Proposition 3 The proof is similar to Proposition 1:

$$\begin{aligned} & E\left[C_N(Z^N,p)\middle| X^N = x^N\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} E\left[\left(Y_i - f^p(X_i, l_{N-1}(Z_{-i}^N;p))\right)^2 \middle| X^N = x^N\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} E\left[E\left[\left(Y_i - f^p(X_i, l_{N-1}(Z_{-i}^N;p))\right)^2 \middle| Z_{-i}^N, X^N = x^N\right]\middle| X^N = x^N\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} E\left[\left(g_i(X_i) - f^p(X_i, l_{N-1}(Z_{-i}^N;p))\right)^2 + \sigma^2 \middle| X^N = x^N\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} E\left[\left(g_i(X_i) - f^p(X_i, l_{N-1}(Z_{-i}^N;p))\right)^2 \middle| X^N = x^N\right] + \sigma^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} M_N(x_i|x_{-i}^N;p) + \sigma^2. \ \square \end{aligned}$$

Proof of Proposition 4

$$E\left[G_{N,M}(Z^N, \tilde{Z}^M;p)\middle| X^N = x^N\right]$$

$$
\begin{aligned}
&= \frac{1}{M}\sum_{i=N+1}^{N+M} E\left[\left(Y_i - f^p(X_i, l_N(Z^N;p))\right)^2 \Big| X^N = x^N\right] \\
&= \frac{1}{M}\sum_{i=N+1}^{N+M} E\left[E\left[\left(Y_i - f^p(X_i, l_N(Z^N;p))\right)^2 \Big| Z^N, X^N = x^N\right] \Big| X^N = x^N\right] \\
&= \frac{1}{M}\sum_{i=N+1}^{N+M} E\left[\left(g(X_i) - f^p(X_i, l_N(Z^N;p))\right)^2 + \sigma^2 \Big| X^N = x^N\right] \\
&= \frac{1}{M}\sum_{i=N+1}^{N+M} E\left[\left(g(X_i) - f^p(X_i, l_N(Z^N;p))\right)^2 \Big| X^N = x^N\right] + \sigma^2 \\
&= \frac{1}{M}\sum_{i=N+1}^{N+M} E\left[E\left[\left(g(X_i) - f^p(X_i, l_N(Z^N;p))\right)^2 \Big| X_i\right] \Big| X^N = x^N\right] + \sigma^2 \\
&= \frac{1}{M}\sum_{i=N+1}^{N+M} E\left[M_N(X_i|X^N;p) \big| X^N = x^N\right] + \sigma^2 \\
&= \frac{1}{M}\sum_{i=N+1}^{N+M} IMSE_N(x^N;p) + \sigma^2. \ \square
\end{aligned}
$$

Proof of Theorem 1 By Assumption 1, 4 and 5, $(y - f^p(x,w))^2$ is continuous in w and dominated by $2|y|^2 + D(x)^2$ for each $x \in \mathbf{X}$ and $y \in \mathbf{Y}$ with $E\left[2|Y|^2 + 2D(X)^2\right] < \infty$. This allows us to apply Jennrich's uniform law of large numbers (ULLN, Jennrich [29, Theorem 2]) to $\left\{(Y_i - f^p(X_i,\cdot))^2 : \mathbf{W}^p \to \mathbf{Y}\right\}_{i \in \mathbf{N}}$. Therefore $\lambda(\cdot;p)$ is continuous on $\mathbf{W}^p$ and there exists $F_1 \in \mathcal{F}$ with $\mathcal{P}(F_1) = 1$ such that for each $\omega \in F$

$$
\sup_{w \in \mathbf{W}^p}\left| N^{-1}\sum_{i=1}^{N}(Y_i(\omega) - f^p(X_i(\omega), w))^2 - \lambda(w;p)\right| \to 0
$$

as $N \to \infty$. In addition, for each N, there exists $F_{2N} \in \mathcal{F}$ with $\mathcal{P}(F_{2N}) = 1$ such that for each $\omega \in F_{2N}$ $l_N(Z^N(\omega);p)$ solves

$$
\min_{w \in \mathbf{W}^p} N^{-1}\sum_{i=1}^{N}(Y_i(\omega) - f^p(X_i(\omega), w))^2.
$$

Let $F \equiv F_1 \cap (\cap_{N=1}^{\infty} F_{2N})$. Then $\mathcal{P}(F) = 1$.

Let ω be arbitrary in F and $\epsilon > 0$ be arbitrary. Then there exists N_0 such that for any $N > N_0$

$$
\sup_{w \in \mathbf{W}^p}\left| N^{-1}\sum_{i=1}^{N}(Y_i(\omega) - f^p(X_i(\omega), w))^2 - \lambda(w;p)\right| < \frac{\epsilon}{1 + (2/A)}
$$

and

$$
\sup_{w \in \mathbf{W}^p}\left| (N + M_N)^{-1}\sum_{i=1}^{N+M_N}(Y_i(\omega) - f^p(X_i(\omega), w))^2 - \lambda(w;p)\right| < \frac{\epsilon}{1 + (2/A)}.
$$

This implies that for any $N > N_0$

$$
\begin{aligned}
&\sup_{w \in \mathbf{W}^p}\left| M_N^{-1}\sum_{i=N+1}^{N+M_N}(Y_i(\omega) - f^p(X_i(\omega), w))^2 - \lambda(w;p)\right| \\
\leq\ &\sup_{w \in \mathbf{W}^p}\left| \frac{N + M_N}{M_N}(N + M_N)^{-1}\sum_{i=1}^{N+M_N}\{(Y_i(\omega) - f^p(X_i(\omega), w))^2 - \lambda(w;p)\} - \right.
\end{aligned}
$$

$$
\begin{aligned}
& \frac{N}{M_N} N^{-1} \sum_{i=1}^{N} \{(Y_i(\omega) - f^p(X_i(\omega), w))^2 - \lambda(w;p)\} \Bigg| \\
\leq\ & \frac{N+M_N}{M_N} \sup_{w \in \mathbf{W}^p} \left| (N+M_N)^{-1} \sum_{i=1}^{N+M_N} \{(Y_i(\omega) - f^p(X_i(\omega), w))^2 - \lambda(w;p)\} \right| + \\
& \frac{N}{M_N} \sup_{w \in \mathbf{W}^p} \left| N^{-1} \sum_{i=1}^{N} \{(Y_i(\omega) - f^p(X_i(\omega), w))^2 - \lambda(w;p)\} \right| \\
\leq\ & \frac{N+M_N}{M_N} \frac{\epsilon}{1+(2/A)} + \frac{N}{M_N} \frac{\epsilon}{1+(2/A)} \\
\leq\ & \left(\frac{N+M_N}{M_N} + \frac{N}{M_N} \right) \frac{\epsilon}{1+(2/A)} \\
\leq\ & \left(1 + \frac{2N}{M_N} \right) \frac{\epsilon}{1+(2/A)} \\
\leq\ & \left(1 + \frac{2N}{AN} \right) \frac{\epsilon}{1+(2/A)} \\
=\ & \epsilon.
\end{aligned}
$$

Thus for any $N > N_0$, we have

$$
\sup_{w \in \mathbf{W}^p} \left| N^{-1} \sum_{i=1}^{N} (Y_i(\omega) - f^p(X_i(\omega), w))^2 - \lambda(w;p) \right| < \epsilon.
$$

and

$$
\sup_{w \in \mathbf{W}^p} \left| M_N^{-1} \sum_{i=N+1}^{N+M_N} (Y_i(\omega) - f^p(X_i(\omega), w))^2 - \lambda(w;p) \right| < \epsilon.
$$

Because $l_N(Z^N(\omega);p)$ minimizes the sum of the squared residuals, the first inequality implies that

$$
|\lambda(l_N(Z^N(\omega);p);p) - \lambda(w^{p*};p)| < \epsilon
$$

as easily verified, where w^{p*} is an arbitrary element of $\mathbf{W}^{p*}$. By using the second inequality with this relationship, we have

$$
\begin{aligned}
& \left| G_{N,M_N}(Z^N(\omega), \tilde{Z}^{M_N}(\omega);p) - \lambda(w^{p*};p) \right| \\
& = \left| M_N^{-1} \sum_{i=N+1}^{N+M_N} (Y_i(\omega) - f^p(X_i(\omega), l_N(Z^N(\omega);p)))^2 - \lambda(w^{p*};p) \right| \\
& \leq \left| M_N^{-1} \sum_{i=N+1}^{N+M_N} (Y_i(\omega) - f^p(X_i(\omega), l_N(Z^N(\omega);p)))^2 - \lambda(l_N(Z^N(\omega);p)) \right| + \\
& \quad \left| \lambda(l_N(Z^N(\omega);p)) - \lambda(w^{p*};p) \right| \\
& \leq 2\epsilon.
\end{aligned}
$$

Because ϵ is arbitrary, this implies that

$$
|G_{N,M_N}(Z^N(\omega), \tilde{Z}^{M_N}(\omega);p) - \lambda(w^{p*};p)| \rightarrow 0
$$

as $N \rightarrow \infty$. Further this holds for all $\omega \in F$. The result therefore follows. □

Proof of Lemma 2 Suppose that the claim is not true. Then there must exist a constant $M > 0$, a strictly increasing sequence $\{N_j\}_{j \in \mathbf{N}}$ in $\mathbf{N}$, a sequence $\{\theta^j\}_{j \in \mathbf{N}}$ in Θ and a sequence $\{k_j\}_{j \in \mathbf{N}}$ in

$\mathbf{N}$ such that $k_j \leq N_j$ and

$$\left|(N_j-1)^{-1}\sum_{1\leq i\leq N_j, i\neq k_j} h_i(\theta^j)\right| \geq M$$

for each $j \in \mathbf{N}$. This implies that

$$\begin{aligned}\frac{M}{2} &\leq \left|N_j^{-1}\sum_{1\leq i\leq N_j, i\neq k_j} h_i(\theta^j)\right| \\ &= \left|N_j^{-1}\sum_{i=1}^{N_j} h_i(\theta^j) - N_j^{-1}h_{k_j}(\theta)\right| \\ &\leq \left|N_j^{-1}\sum_{i=1}^{N_j} h_i(\theta^j)\right| + \left|N_j^{-1}h_{k_j}(\theta^j)\right|\end{aligned}$$

By rearranging this relation, we have

$$N_j^{-1}\left|h_{k_j}(\theta^j)\right| \geq \frac{M}{2} - \left|N_j^{-1}\sum_{i=1}^{N_j} h_i(\theta^j)\right|.$$

The second term of the RHS converges to zero as $j \to \infty$. Therefore for all j sufficiently large, it holds that

$$N_j^{-1}\left|h_{k_j}(\theta^j)\right| \geq \frac{M}{4}. \tag{11}$$

On the other hand, if there exists $N_0 \in \mathbf{N}$ such that $k_j \leq N_0$ for all $j \in \mathbf{N}$, we have

$$N_j^{-1}\left|h_{k_j}(\theta^j)\right| \leq N_j^{-1} \max_{j\in\{1,\ldots,N_0\}} \max_{\theta\in\Theta} h_i(\theta) \to 0 \text{ as } j \to \infty.$$

Hence the left hand side of inequality (11) converges to zero. It therefore must hold that $k_j \to \infty$ as $j \to \infty$. By assumption and this fact, we have

$$\left|k_j^{-1}\sum_{i=1}^{k_j} h_i(\theta^j)\right| < \frac{M}{8}$$

and

$$\left|(k_j-1)^{-1}\sum_{i=1}^{k_j-1} h_i(\theta^j)\right| < \frac{M}{8}$$

for all j sufficiently large. This implies that

$$\begin{aligned}\left|k_j^{-1}h_{k_j}(\theta^j)\right| &= \left|k_j^{-1}\sum_{i=1}^{k_j} h_i(\theta^j) - k_j^{-1}\sum_{i=1}^{k_j-1} h_i(\theta^j)\right| \\ &\leq \left|k_j^{-1}\sum_{i=1}^{k_j} h_i(\theta^j)\right| + \left|k_j^{-1}\sum_{i=1}^{k_j-1} h_i(\theta^j)\right| \\ &\leq \left|k_j^{-1}\sum_{i=1}^{k_j} h_i(\theta^j)\right| + \left|(k_j-1)^{-1}\sum_{i=1}^{k_j-1} h_i(\theta^j)\right| \\ &< \frac{M}{4}. \end{aligned} \tag{12}$$

Equations (11) and (12) imply that

$$\frac{M}{4} \leq N_j^{-1} \left| h_{k_j}(\theta^j) \right| \leq k_j^{-1} \left| h_{k_j}(\theta^j) \right| < \frac{M}{4},$$

for all j sufficiently large, which is obviously impossible because $M > 0$. The result therefore follows. $\square$

Proof of Theorem 2 By Assumption 1, 4 and 5, $(y - f^p(x, w))^2$ is continuous in w and dominated by $(D(x, y))^2$ for each $x \in \mathbf{X}$ and $y \in \mathbf{Y}$ with $E\left[(D(X, Y))^2\right] < \infty$, given Assumptions 2 and 5. This allows us to apply Jennrich's uniform law of large numbers (ULLN, Jennrich [29, Theorem 2]) to $\left\{(Y_i - f^p(X_i, \cdot))^2 : \mathbf{W}^p \to \Re\right\}_{i \in \mathbf{N}}$. Thus there exists $F_1 \in \mathcal{F}$ such that for each $\omega \in F_1$

$$\sup_{w \in \mathbf{W}^p} \left| N^{-1} \sum_{i=1}^{N} (Y_i(\omega) - f^p(X_i(\omega), w))^2 - \lambda(w; p) \right| \to 0 \text{ as } N \to \infty. \tag{13}$$

By Lemma 2 this implies that for each $\omega \in F_1$

$$\max_{k \in (1, \ldots, N)} \sup_{w \in \mathbf{W}^p} \left| (N-1)^{-1} \sum_{1 \leq i \leq N, i \neq k} (Y_i(\omega) - f^p(X_i(\omega), w))^2 - \lambda(w; p) \right| \to 0 \text{ as } n \to \infty. \tag{14}$$

By Lemma 1, for each $N \in \mathbf{N}$ and $i \in \{1, \ldots, N\}$, there exists $F_{2Ni} \in \mathcal{F}$ with $\mathcal{P}(F_{2Ni}) = 1$ such that for each $\omega \in F_{2Ni}$, $l(Z_{-i}^{N-1}(\omega); p)$ solves

$$\min_{w \in \mathbf{W}^p} (N-1)^{-1} \sum_{1 \leq i \leq N, i \neq k} (Y_i(\omega) - f^p(X_i(\omega), w))^2 . \tag{15}$$

Let $F_2 \equiv \cap_{N \in \mathbf{N}} \cap_{i=1}^{N} F_{2Ni}$ and $F_0 \equiv F_1 \cap F_2$. Then $\mathcal{P}(F_0) = 1$. Now we show that for each $\omega \in F_0$,

$$\max_{i \in (1, \ldots, N)} \inf_{w^* \in \mathbf{W}^{p*}} \left\| l_{N-1}\left(Z_{-i}^{N}(\omega); p\right) - w^* \right\| \to 0 \text{ as } N \to \infty.$$

Suppose that this does not hold for some $\omega_0 \in F_0$. Then there must exist a constant $M > 0$, a strictly increasing sequence $\{N_j\}_{j \in \mathbf{N}}$ in $\mathbf{N}$ and a sequence $\{k_j\}_{j \in \mathbf{N}}$ in $\mathbf{N}$ such that $1 \leq k_j \leq N_j$ and

$$\inf_{w^* \in \mathbf{W}^{p*}} \left\| l_{N_j - 1}\left(Z_{-k_j}^{N_j}(\omega_0); p\right) - w^* \right\| \geq M$$

for any $j \in \mathbf{N}$. Let

$$\eta_c = \{w \in \mathbf{W}^p : \|w - w^*\| \geq M \text{ for any } w^* \in \mathbf{W}^{p*}\}.$$

Then $l_{N_j-1}\left(Z_{-k_j}^{N_j}(\omega_0); p\right) \in \eta^c$. Obviously η_c is compact and $\lambda(\cdot; p)$ attains a minimum on η_c, i.e., there exists $w^c \in \eta^c$ such that $\lambda(w^c; p) \leq \lambda(w; p)$ for any $w \in \eta_c$. Because $w^c \notin \mathbf{W}^{p*}$, for any $w^* \in \mathbf{W}^{p*}$,

$$\epsilon \equiv \lambda(w^c; p) - \lambda(w^*; p) > 0.$$

Thus

$$\lambda\left(l_{N_j-1}\left(Z_{-k_j}^{N_j}(\omega_0); p\right); p\right) - \lambda(w^*; p) \geq \epsilon > 0 \tag{16}$$

for any $j \in \mathbf{N}$ and $w^* \in \mathbf{W}^{p*}$.

On the other hand, it follows from equation (14) that there exists $j_0 \in \mathbf{N}$ such that for each $j > j_0$

$$\left| (N_j - 1)^{-1} \sum_{1 \leq i \leq N_j, i \neq k} \left(Y_i(\omega) - f^p\left(X_i(\omega), l_{N_j-1}\left(Z_{-k_j}^{N_j-1}(\omega_0); p\right)\right)\right)^2 - \lambda\left(l_{N_j-1}\left(Z_{-k_j}^{N_j-1}(\omega_0); p\right); p\right) \right| < \frac{\epsilon}{2}$$

and

$$\left|(N_j-1)^{-1}\sum_{1\le i\le N_j, i\ne k}(Y_i(\omega)-f^p(X_i(\omega),w^*)))^2-\lambda(w^*;p)\right|<\frac{\epsilon}{2}.$$

These two inequalities and inequality (16) imply that for any $j \geq j_0$,

$$\begin{aligned}
&(N_j-1)^{-1}\sum_{1\le i\le N_j, i\ne k}\left(Y_i(\omega)-f^p\left(X_i(\omega), l_{N_j-1}\left(Z_{-k_j}^{N_j-1}(\omega_0);p\right)\right)\right)^2 \\
&\quad -(N_j-1)^{-1}\sum_{1\le i\le N_j, i\ne k}(Y_i(\omega)-f^p(X_i(\omega),w^*))^2 \\
&= \left\{(N_j-1)^{-1}\sum_{1\le i\le N_j, i\ne k}\left(Y_i(\omega)-f^p\left(X_i(\omega), l_{N_j-1}\left(Z_{-k_j}^{N_j-1}(\omega_0);p\right)\right)\right)^2-\lambda\left(l_{N_j-1}\left(Z_{-k_j}^{N_j-1}(\omega_0);p\right);p\right)\right\} \\
&\quad +\left\{\lambda\left(l_{N_j-1}\left(Z_{-k_j}^{N_j-1}(\omega_0);p\right);p\right)-\lambda(w^*;p)\right\} \\
&\quad +\left\{\lambda(w^*;p)-(N_j-1)^{-1}\sum_{1\le i\le N_j, i\ne k}(Y_i(\omega)-f^p(X_i(\omega),w^*))^2\right\} \\
&> -\frac{\epsilon}{2}+\epsilon-\frac{\epsilon}{2}=0.
\end{aligned}$$

Therefore we have

$$\begin{aligned}
&(N_j-1)^{-1}\sum_{1\le i\le N_j-1, i\ne k_j}\left(Y_i(\omega)-f^p\left(X_i(\omega), l_{N_j-1}\left(Z_{-k_j}^{N_j-1}(\omega_0);p\right)\right)\right)^2 \\
&\quad > (N_j-1)^{-1}\sum_{1\le i\le N_j, i\ne k_j}(Y_i(\omega)-f^p(X_i(\omega),w^*))^2.
\end{aligned}$$

for all $j > j_0$. But this cannot happen because $\omega_0 \in F_0 \subset F_2 \subset F_{2N_jK_j}$ for each $j \in \mathbf{N}$. The desired result therefore must hold. $\square$

<u>Proof of Theorem 3</u> Let F_0 be the same set as in the proof of Theorem 2. Then for each $\omega \in F_0$, it holds that for each $N \in \mathbf{N}$, $k \in (1, \ldots, N)$,

$$\sum_{1\le i\le N, i\ne k}\left(Y_i(\omega)-f^p\left(X_i(\omega), l_{N-1}\left(Z_{-k}^N(\omega);p\right)\right)\right)^2 \le \sum_{1\le i\le N, i\ne k}\left(Y_i(\omega)-f^p\left(X_i(\omega), l_N\left(Z^N(\omega);p\right)\right)\right)^2$$

and

$$\sum_{i=1}^N\left(Y_i(\omega)-f^p\left(X_i(\omega), l_{N-1}\left(Z_{-k}^N(\omega);p\right)\right)\right)^2 \ge \sum_{i=1}^N\left(Y_i(\omega)-f^p\left(X_i(\omega), l_N\left(Z^N(\omega);p\right)\right)\right)^2.$$

These inequalities imply that for each $\omega \in F_0$, $N \in \mathbf{N}$ and $k \in (1, \ldots, N)$,

$$\begin{aligned}
&\left(Y_k(\omega)-f^p\left(X_k(\omega), l_{N-1}\left(Z_{-k}^N(\omega);p\right)\right)\right)^2 \\
&= \sum_{i=1}^N\left(Y_i(\omega)-f^p\left(X_i(\omega), l_{N-1}\left(Z_{-k}^N(\omega);p\right)\right)\right)^2-\sum_{1\le i\le N, i\ne k}\left(Y_i(\omega)-f^p\left(X_i(\omega), l_{N-1}\left(Z_{-k}^N(\omega);p\right)\right)\right)^2 \\
&\ge \sum_{i=1}^N\left(Y_i(\omega)-f^p\left(X_i(\omega), l_N\left(Z^N(\omega);p\right)\right)\right)^2-\sum_{1\le i\le N, i\ne k}\left(Y_i(\omega)-f^p\left(X_i(\omega), l_N\left(Z^N(\omega);p\right)\right)\right)^2 \\
&= \left(Y_k(\omega)-f^p\left(X_k(\omega), l_N\left(Z^N(\omega);p\right)\right)\right)^2.
\end{aligned}$$

It follows that for each $\omega \in F_0$, $N \in \mathbf{N}$

$$C_N\left(Z^N(\omega);p\right) \geq \sum_{i=1}^{N}\left(Y_i(\omega) - f^p\left(X_i(\omega), l_N\left(Z^N(\omega);p\right)\right)\right)^2.$$

Thus it holds that for each $\omega \in F_0$

$$\liminf_{N\to\infty} C_N\left(Z^N(\omega);p\right) \geq \liminf_{N\to\infty}\sum_{i=1}^{N}\left(Y_i(\omega) - f^p\left(X_i(\omega), l_N\left(Z^N(\omega);p\right)\right)\right)^2,$$

and we know that the RHS of this inequality is $\min_{w\in\mathbf{W}^p}\lambda(w;p)$ by equation (13) and the fact that

$$d\left(l_N\left(Z^N(\omega);p\right), \mathbf{W}^{p*}\right) = d\left(l_N\left(Z^{N+1}_{-(N+1)}(\omega);p\right), \mathbf{W}^{p*}\right) \to 0 \text{ as } N\to\infty$$

which we have established in Theorem 2. □

<u>Proof of Theorem 4</u> Let w^* be the unique maximum assumed in Assumption 6. Then

$$\left\|l_N\left(Z^N;p\right) - w^*\right\| = \left\|l_N\left(Z^{N+1}_{-(N+1)};p\right) - w^*\right\| \to 0 \text{ a.s.-}\mathcal{P} \text{ as } N\to\infty.$$

by Theorem 2. Because $\lambda(\cdot;p)$ is bounded on $\mathbf{W}^p$, it follows from the Lebesgue dominated convergence theorem that

$$\begin{aligned}\lim_{N\to\infty} E\left(Q_N(Z^N;p)\right) &= \lim_{N\to\infty} E\left[\lambda\left(l_N\left(Z^N;p\right);p\right)\right] \\ &= E\left[\lim_{N\to\infty}\lambda\left(l_N\left(Z^N;p\right);p\right)\right] = E\left[\lambda(w^*;p)\right] = \lambda(w^*;p).\end{aligned}$$

Thus it suffices to show that

$$\left|C_N\left(Z^N;p\right) - \lambda(w^*;p)\right| \to 0 \text{ a.s.-}\mathcal{P} \text{ as } N\to\infty.$$

We further know that

$$\left|N^{-1}\sum_{i=1}^{N}\left(Y_i - f^p(X_i,w^*)\right)^2 - \lambda(w^*;p)\right| \to 0 \text{ a.s.-}\mathcal{P} \text{ as } N\to\infty.$$

Therefore it is sufficient to show that

$$\left|N^{-1}\sum_{i=1}^{N}\left(Y_i - f^p\left(X_i, l_{N-1}\left(Z^N_{-i};p\right)\right)\right)^2 - N^{-1}\sum_{i=1}^{N}\left(Y_i - f^p(X_i,w^*)\right)^2\right| \to 0 \text{ a.s.-}\mathcal{P} \text{ as } N\to\infty.$$

Define a correspondence $\bar{B}$ mapping each point in $[0,1]$ to a subset of $\mathbf{W}^p$ by

$$\bar{B}(e) \equiv \{w \in \mathbf{W}^p : \|w - w^*\| \leq e\}$$

for each $e \in [0,1]$. Then it is easy to verify that $\bar{B}(\cdot)$ is a compact-valued continuous correspondence, namely, $\bar{B}(e)$ is a compact subset of $\mathbf{W}^p$ for each $e \in [0,1]$ and $\bar{B}(\cdot)$ is upper and lower hemi-continuous at every $e \in [0,1]$. Also define $V : [0,1]\times\mathbf{X}\times\mathbf{Y} \to [0,\infty)$ by

$$V(e,x,y) \equiv \max_{w\in\bar{B}(e)}\left|(y - f^p(x,w))^2 - (y - f^p(x,w^*))^2\right|$$

for each $e \in [0,1]$, $x \in \mathbf{X}$ and $y \in \mathbf{Y}$.

By Lemma 2 of Jennrich [29] (See also White [33, 2.11]) and Lemma 13.2.3 of Klein and Thompson [37], $V(e,X_i,Y_i) : \Omega \to [0,\infty)$ is measurable for each $i \in \mathbf{N}$ and $e \in [0,1]$. We can also verify by Assumption 4 and the Theorem of Maximum that $V(\cdot,X_i,Y_i) : [0,1] \to$ is continuous a.s.-$\mathcal{P}$. In addition, we have

$$0 \leq V(e,x,y) \leq 2(D(x,y))^2$$

for each $x \in \mathbf{X}$ and $y \in \mathbf{Y}$ with

$$E\left[(D(X,Y))^2\right] < \infty.$$

Given Assumptions 2 and 5, these properties allow us to apply Jennrich's ULLN (Jennrich [29, Theorem 2]) to $\{V(e, X_i, Y_i)\}$. Thus $\bar{V} : [0,1] \to [0,\infty)$ defined by

$$\bar{V}(e) \equiv E\left[V(e, X, Y)\right]$$

for each $e \in [0,1]$ is continuous on $[0,1]$ and there exists $F_1 \in \mathcal{F}$ with $\mathcal{P}(F_1) = 1$ such that for each $\omega \in F_1$

$$\sup_{e \in [0,1]} \left| N^{-1} \sum_{i=1}^{N} V(e, X_i(\omega), Y_i(\omega)) - \bar{V}(e) \right| \to 0.$$

In Theorem 2, we have established that there exists $F_2 \in \mathcal{F}$ with $\mathcal{P}(F_2) = 1$ such that for each $\omega \in F_2$,

$$\max_{i \in (1,\ldots,N)} \left\| l_{N-1}\left(Z_{-i}^N(\omega); p\right) - w^* \right\| \to 0.$$

Now let $F_0 \equiv F_1 \cap F_2$; pick an arbitrary point $\epsilon \in (0, \infty)$ and an arbitrary point $\omega \in F_0$. Then the following properties follow from what we have established.

- There exists $\delta_\epsilon \in (0,1]$ such that for any $e \in [0, \delta_\epsilon)$,
$$\bar{V}(e) = \bar{V}(e) - \bar{V}(0) < \frac{\epsilon}{2};$$
- There exists $N_1 \in \mathbf{N}$ such that for any $N \geq N_1$,
$$\sup_{e \in [0,1]} \left| N^{-1} \sum_{i=1}^{N} V(e, X_i(\omega), Y_i(\omega)) - \bar{V}(e) \right| < \frac{\epsilon}{2};$$
- There exists $N_2 \in \mathbf{N}$ such that for any $N \geq N_2$,
$$\max_{i \in (1,\ldots,N)} \left\| l_{N-1}\left(Z_{-i}^N(\omega); p\right) - w^* \right\| < \frac{\delta_\epsilon}{2}.$$

Therefore, for any $N \geq N_0 \equiv \max\{N_1, N_2\}$,

$$\begin{aligned}
&\left| N^{-1} \sum_{i=1}^{N} \left(Y_i(\omega) - f^p\left(X_i(\omega), l_{N-1}\left(Z_{-i}^N(\omega); p\right)\right)\right)^2 - N^{-1} \sum_{i=1}^{N} \left(Y_i(\omega) - f^p(X_i(\omega), w^*)\right)^2 \right| \\
\leq\;& N^{-1} \sum_{i=1}^{N} \left| \left(Y_i(\omega) - f^p\left(X_i(\omega), l_{N-1}\left(Z_{-i}^N(\omega); p\right)\right)\right)^2 - \left(Y_i(\omega) - f^p(X_i(\omega), w^*)\right)^2 \right| \\
\leq\;& N^{-1} \sum_{i=1}^{N} V\left(\frac{\delta_\epsilon}{2}, X_i(\omega), Y_i(\omega)\right) \\
\leq\;& \left| N^{-1} \sum_{i=1}^{N} V\left(\frac{\delta_\epsilon}{2}, X_i(\omega), Y_i(\omega)\right) - \bar{V}\left(\frac{\delta_\epsilon}{2}\right) \right| + \left| \bar{V}\left(\frac{\delta_\epsilon}{2}\right) \right| \\
<\;& \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
\end{aligned}$$

Thus, for any $N \geq N_0$,

$$\left| N^{-1} \sum_{i=1}^{N} \left(Y_i(\omega) - f^p\left(X_i(\omega), l_{N-1}\left(Z_{-i}^N(\omega); p\right)\right)\right)^2 - N^{-1} \sum_{i=1}^{N} \left(Y_i(\omega) - f^p(X_i(\omega), w^*)\right)^2 \right| < \epsilon.$$

Because ϵ is an arbitrary strictly positive real number, this implies that

$$\left| N^{-1} \sum_{i=1}^{N} \left(Y_i(\omega) - f^p\left(X_i(\omega), l_{N-1}\left(Z_{-i}^N(\omega); p\right)\right)\right)^2 - N^{-1} \sum_{i=1}^{N} \left(Y_i(\omega) - f^p(X_i(\omega), w^*)\right)^2 \right| \to 0$$

as $N \to \infty$. Because this holds for all $\omega \in F_0$ and $\mathcal{P}(F_0) = 1$, we conclude that

$$\left| N^{-1} \sum_{i=1}^{N} \left(Y_i - f^p\left(X_i, l_{N-1}\left(Z_{-i}^N; p\right)\right)\right)^2 - N^{-1} \sum_{i=1}^{N} \left(Y_i - f^p(X_i, w^*)\right)^2 \right| \to 0 \text{ a.s.-}\mathcal{P}$$

as $N \to \infty$. $\square$