Supplementary Information for

Second-Order Induction in Prediction Problems

Rossella Argenziano, Itzhak Gilboa

Itzhak Gilboa
E-mail: tzachigilboa@gmail.com

This PDF file includes:

References for SI reference citations

Supporting Information: Proofs. **Proof of Observation 1**

Assume that $m = 1$, $n = 4$ and

i	x_i	y_i
1	0	0
2	1	0
3	3	1
4	4	1

In this example observations $1, 2$ are closer to each other than each is to any of observations $3, 4$ and vice versa. (That is, $|x_i - x_j| = 1$ for $i = 1, j = 2$ as well as for $i = 3, j = 4$, but $|x_i - x_j| \geq 2$ for $i \leq 2 < j$.) Moreover the values of y are the same for the "close" observations and different for "distant" ones. (That is, $y_i = y_j$ for $i = 1, j = 2$ as well as for $i = 3, j = 4$, but $|y_i - y_j| = 1$ for $i \leq 2 < j$.) If we choose a finite w, the estimated value for each i, $\bar{y}_i^{s_w}$, is a weighted average of the two distant observations and the single close one. In particular, for every $w < \infty$ we have $MSE\,(w) > 0$.

Observe that $w = w^1 = \infty$ doesn't provide a perfect fit either: if we set $w = w^1 = \infty$, each observation i is considered to be dissimilar to any other, and its y value is estimated to be the default value, $\bar{y}_i^{s_w} = y_0$. Regardless of the (arbitrary) choice of y_0, the MSE is bounded below by that obtained for $y = 0.5$ (which is the average y in the entire database). Thus, $MSE\,(\infty) \geq 0.25$.

Thus, $MSE\,(w) > 0$ for all $w \in [0, \infty]$. However, as $w \to \infty$ (but $w < \infty$), for each i the weight of the observation that is closest to i converges to 1 (and the weights of the distant ones – to zero), so that $\bar{y}_i^{s_w} \to y_i$. Hence, $MSE\,(w) \to_{w\to\infty} 0$. We thus conclude that $\inf_{w\in[0,\infty]} MSE\,(w) = 0$ but that there is no w that minimizes the MSE.

The same argument applied to the $AMSE\,(w, c)$ for any $c < c_0$ if we set $c_0 = 0.25$. $\square$

Proof of Proposition 1

We first wish to show that arbitrarily low values of the MSE can be obtained with probability that is arbitrarily close to 1, provided the weights w^j are all large enough. Let there be given $\nu > 0$ and $\xi > 0$. We wish to find N and W such that for every $n \geq N$, and every vector w such that $w^j \geq W$ but $w^j < \infty$ $(\forall j \leq m)$ we have

$$P\,(MSE\,(w) < \nu) \geq 1 - \xi.$$

Observe that a single j for which $w^j = \infty$ suffices to set the MSE at least as high as the variance of (y_i), as, with probability 1, each observation will be the unique one with the specific value of x^j.

We now define "proximity" of the x values that would guarantee "proximity" of the y values. Suppose that the latter is defined by $\nu/2$. As the function f is continuous on a compact set, it is uniformly continuous. Hence, there exists $\theta > 0$ such that, for any x, x' that satisfy $\|x - x'\| < \theta$ we have $[f\,(x) - f\,(x')]^2 < \nu/2$. Let us divide the set X into $\left(4K\sqrt{m}/\theta\right)^m$ equi-volume cubes, each with an edge of length $\frac{\theta}{2\sqrt{m}}$. Two points x, x' that belong to the same cube differ by at most $\frac{\theta}{2\sqrt{m}}$ in each coordinate and thus satisfy $\|x - x'\| < \theta/2$. Let us now choose N_1 such that, with probability of at least $(1 - \xi/2)$, each such cube contains at least two observations x_i $(i \leq N_1)$. This guarantees that, when observation i is taken out of the sample, there is another observation i' (in the same cube), with $[y_{i'} - f\,(x_i)]^2 < \nu/2$.

Next, we wish to bound the probability mass of each cube (defined by g). The volume of a cube is $\left(\frac{\theta}{2\sqrt{m}}\right)^m$ and the density function is bounded from below by η. Thus, the proportion of observations in the cube (out of all the n observations) converges (as $n \to \infty$) to a number that is bounded from below by $\zeta \equiv \eta \left(\frac{\theta}{2\sqrt{m}}\right)^m > 0$. Choose $N \geq N_1$ such that, with probability of at least $(1 - \xi/2)$, for each $n \geq N$ the proportion of the observations in the cube is at least $\zeta/2$. Note that this is a positive number which is independent of n.

We can now turn to choose W. For each i, the proportion of observations x_k with $[f\,(x_i) - f\,(x_k)]^2 > \nu$ is bounded above by $(1 - \zeta)$. Choose w such that $w^j = W$. Observe that, as $W \to \infty$,

$$\frac{\sum_{k\neq i, [f(x_i)-f(x_k)]^2 > \nu} s(x_i, x_k)}{\sum_{k\neq i, [f(x_i)-f(x_k)]^2 \leq \nu} s(x_i, x_k)} \to 0$$

and this convergence is uniform in n (as the definition of ζ is independent of n). Thus a sufficiently high W can be found so that, for all $n \geq N$, $MSE\,(w_0) < \nu$ with probability $(1 - \xi)$ or higher.

Next we prove the second part of the proposition. Assume that x^j is informative, so that there exist x, x' such that $x^l = x'^l$ for all $l \neq j$ but $f\,(x) - f\,(x') = \delta > 0$. Assume that, for some $W < \infty$, $w^j \leq W$. Similar arguments to those above yield an lower bound $\nu > 0$ such that, for large n, with very high probability, $MSE\,(w) > \nu$: points around x will have estimated y values that are affected by points around x', and the weight of these will not converge to zero (it is bounded from below by e^{-W}).

Finally, we wish to show that one can have a low enough cost c_0 such that all the vectors in ε-$\arg\min AMSE$ would use the informative variables, as well as a low enough ε so that they would not use the uninformative variables. This would mean that for appropriately chosen c_0 and ε, the supports of all vectors in ε-$\arg\min AMSE$ have to coincide with $I(f)$. Let there be given $\xi > 0$. For each $j \in I(f)$ we can use the second part of the proposition (corresponding to $W = 0$) to find $\nu_j > 0$ and N_j such that, for every $n \geq N_j$, with probability of at least $(1 - \xi/2m)$, $w^j = 0$ implies $MSE\,(w) > \nu_j$. Define $N_j = 0$ for $j \notin I(f)$.

Choose $c_0 = min_j(\nu_j)/2(m + 1)$ and let $c < c_0$. Using the first part of the proposition, let N_0 and W_0 be such that, for all $n \geq N_0$, with probability of at least $(1 - \xi/2)$, $MSE\,(w_0) < c$ for w_0 defined by $w_0^l = W_0$ for all l. Consider $N = max(N_l)_{l\geq 0}$.

For every $n \geq N$, with probability of at least $(1-\xi)$ we have that (i) there are w with $MSE(w) < c$; (ii) for these w's, $AMSE(w) < (m+1)c$; (iii) for any vector w whose support does not include $j \in I(f)$, $AMSE(w) > \nu_j > (m+1)c$. This means that for every w with $AMSE(w) < (m+1)c$, we must have $I(f) \subset supp(w)$. Thus, considering near-minimizers of the $AMSE$ we will only find vectors that use all the informative variables. On the other hand, we wish to show that in the (high-probability) event considered above, variables that are not informative will not be used. Observe that $\varepsilon < c/2$ is small enough so that for every $w \in \varepsilon$-$\arg\min AMSE$, $w^j = 0$ for every $j \notin I(f)$ (as the inclusion of such a variable in the support of w would incur a cost that is by itself enough to make the $AMSE$ of the vector larger than the argmin by more than $\varepsilon < c/2$). □

Proof of Proposition 2:

Non-uniqueness is obtained by showing that, with a high probability there will be two variables, each of which can provide an almost perfect fit on its own. To this end, we first need to make sure that each observation y_i has a close enough y_k. For this reason the result only holds for a relatively large n (making sure that, with a high probability, no y_i is "isolated"), and then, given such an n, for a large enough number of predictors, $M(n)$, so that we should think of this case as $m >> n >> 1$.

We now turn to prove the result formally. Let there be given $c > 0$. Choose $\bar{\varepsilon} = c/3$. We wish it to be the case that if $MSE(w) \leq \varepsilon$ with $\#supp(w) = 1$, then $w \in \varepsilon$-$\arg\min AMSE$, but for no $w \in \varepsilon$-$\arg\min AMSE$ is it the case that $\#supp(w) > 1$. Clearly, the choice $\bar{\varepsilon} = c/3$ guarantees that for every $\varepsilon \in (0, \bar{\varepsilon})$, the second part of the claim holds: if a vector w satisfies $MSE(w) \leq \varepsilon$, no further reduction in the MSE can justify the cost of additional variables, which is at least c. Conversely, because $c < v/2$ (the variance of y), a single variable j that obtains a near-zero MSE would have a lower $AMSE$ than the empty set.

Let there now be given $\varepsilon \in (0, \bar{\varepsilon})$ and every $\delta > 0$. We need to find N and, for every $n \geq N$, $M(n)$, such that for every $n \geq N$ and $m \geq M(n)$,

$$P\left(supp(\varepsilon\text{-}\arg\min AMSE) \text{ is not closed under union}\right) \geq 1 - \delta.$$

Let N be large enough so that, with probability $(1-\delta/2)$, for all $n \geq N$,

$$\max_i \min_{k \neq i} \left[y_i - y_j\right] < \varepsilon/2.$$

(To see that such an n can be found, one may divide the $[-K, K]$ interval of values to intervals of length $\varepsilon/2$ and choose N to be large enough so that, with the desired probability, there are at least two observations in each such interval.)

Given such $n \geq N$ and the realizations of $(y_i)_{i \leq n}$, consider the realizations of x^j. Assume that, for some j, it so happens that $\left|x_i^j - y_i\right| < \varepsilon/4$ for all $i \leq n$. In this case, by setting w^j to be sufficiently high, and $w^l = 0$ for $l \neq j$, one would obtain $MSE(w) \leq \varepsilon$ and $AMSE(w, c) \leq \varepsilon + c$.[1] For each j, however, the probability that this will be the case is bounded below by some $\xi > 0$, independent of n and j. Let $M_1(n)$ be a number such that, for any $m \geq M_1(n)$, the probability that at least one such j satisfies $\left|x_i^j - y_i\right| < \varepsilon/4$ is $(1-\delta/4)$, and let $M(n) > M_1(n)$ be a number such that, for any $m \geq M(n)$, the probability that at least one more such $j' > j$ satisfies $\left|x_i^{j'} - y_i\right| < \varepsilon/4$ is $(1-\delta/8)$.

Thus, for every $n \geq N$, and every $m \geq M(n)$, with probability $1-\delta$ there are two vectors, w^j with support $\{j\}$ and $w^{j'}$ with support $\{j'\}$, each of which obtaining $MSE(w) \leq \varepsilon$ and thus, both belonging to ε-$\arg\min AMSE$. To see that in this case the $supp(\varepsilon$-$\arg\min AMSE)$ is not closed under union, it suffices to note that no w with support greater than a singleton, nor a w with an empty support (that is, $w \equiv 0$) can be in the ε-$\arg\min AMSE$. □

Proof of Theorem 1

We first verify that the problem is in NP. Given a database and a vector of extended rational weights $w^j \in [0, \infty]$, the calculation of the $AMSE$ takes $O\left(n^2 m\right)$ steps. Specifically, the calculation of the similarity function $s(x, x')$ is done by first checking whether there exists a j such that $w^j = \infty$ and $x^j \neq x'^j$ (in which case $s(x, x')$ is set to 0), and, if not – by ignoring the j's for which $w^j = \infty$.

The proof is by reduction of the SET-COVER problem to EMPIRICAL-SIMILARITY. The former, which is known to be NPC (see (1)), is defined as

Problem 1 *SET-COVER: Given a set P, $r \geq 1$ subsets thereof, $T_1, ..., T_r \subseteq P$, and an integer k $(1 \leq k \leq r)$, are there k of the subsets that cover P? (That is, are there indices $1 \leq i_1 \leq i_2 \leq ... \leq i_k \leq r$ such that $\cup_{j \leq k} T_{i_j} = P$?)*

Given an instance of SET-COVER, we construct, in polynomial time, an instance of EMPIRICAL-SIMILARITY such that the former has a set cover iff the latter has a similarity function that obtains the desired AMSE. Let there be given P, $r \geq 1$ subsets thereof, $T_1, ..., T_r \subseteq P$, and an integer k. Assume without loss of generality that $P = \{1, ..., p\}$, that $\cup_{i \leq r} T_i = P$, and that $z_{uv} \in \{0, 1\}$ is the incidence matrix of the subsets, that is, that for $u \leq p$ and $v \leq r$, $z_{uv} = 1$ iff $u \in T_v$.

Let $n = 2(p+1)$ and $m = r$. Define the database $B = ((x_i, y_i))_{i \leq n}$ as follows. (In the database each observation is repeated twice to avoid bins of size 1.)

For $u \leq p$ define two observations, $i = 2u-1, 2u$ by

$$x_i^j = z_{uj} \qquad y_i = 1$$

[1] The fact that x_i^j is close to y_i is immaterial, of course, as the variables x_i^j are not used to predict y_i directly, but only to identify the y_k that would. If x_i^j is close to some monotone function of y_i the same argument would apply.

and add two more observations, $i = 2p+1, 2p+2$ defined by

$$x_i^j = 0 \qquad y_i = 0.$$

Next, choose c to be such that $0 < c < \frac{1}{mn^3}$, say, $c = \left(mn^3\right)^{-1}/2$ and $R = kc$.[2] This construction can obviously be done in polynomial time.

We claim that there exists a vector w with $AMSE(w,c) \le R$ iff a cover of size k exists for the given instance of SET-COVER.[3] For the "if" part, assume that such a cover exists, corresponding to $J \subseteq M$. Setting the weights

$$w^j = \begin{cases} \infty & j \in J \\ 0 & j \notin J \end{cases}$$

one obtains $AMSE(w,c) \le R$.

Conversely, for the "only if" part, assume that a vector of rational weights $w = \left(w^j\right)_j$ ($w^j \in [0,\infty]$) obtains $AMSE(w,c) \le R$. Let $J \subseteq M$ be the set of indices of predictors that have a positive w^j (∞ included). By the definition of R (as equal to ck), it has to be the case that $|J| \le k$. We argue that J defines a cover (that is, that $\{T_v\}_{v \in J}$ is a cover of P).

Observe that, if we knew that $|J| = k$, the inequality

$$AMSE(w,c) = MSE\,(w) + c\,|J| \le R = ck$$

could only hold if $MSE\,(w) = 0$, from which it would follow that w provides a perfect fit. In particular, for every $i \le 2p$ there exists $j \in J$ such that $x_i^j \ne x_{2p+1}^j$ that is, $x_i^j = 1$, and J defines a cover of P.

However, it is still possible that $|J| < k$ and $0 < MSE\,(w) \le c\,(k - |J|)$. Yet, even in this case, J defines a cover. To see this, assume that this is not the case. Then there exists $i \le 2p$ such that for all j, either $w^j = 0$ ($j \notin J$) or $x_i^j = 0 = x_{2p+1}^j$. This means that $s\,(x_i, x_{2p+1}) = s\,(x_i, x_{2p+2}) = 1$. In particular, $y_{2p+1} = y_{2p+2} = 0$ take part (with positive weights) in the computation of $\overline{y}_i^{sw}$ and we have $\overline{y}_i^{sw} < 1 = y_i$. The cases $2p+1, 2p+2$ obtain maximal similarity to i ($s\,(x_i, x_{2p+1}) = s\,(x_i, x_{2p+2}) = 1$), because $x_{2p+1}^j = x_{2p+2}^j = x_i^j (= 0)$ for all j with $w^j > 0$. (It is possible that for other observations $l \le 2p$ we have $s\,(x_i, x_{2p+1}) \in (0,1)$,but the weights of these observations are evidently smaller than that of $2p+1, 2p+2$.) Thus we obtain that the error $|\overline{y}_i^{sw} - y_i|$ must be at least $\frac{1}{n}$, from which $SSE\,(w) \ge \frac{1}{n^2}$ and $MSE\,(w) \ge \frac{1}{n^3}$ follow. This implies $AMSE(w,c) > R$ and concludes the proof. □

References

1. Garey MR, Johnson DS (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness.* (San-Francisco, CA:W. Freeman and Co.).

[2] As will be clarified shortly, the power of n in the constant c reflects the choice of the quadratic loss function. Different loss functions would require a corresponding cost c. For example, for an absolute value $c = \left(mn^2\right)^{-1}/2$ would suffice.

[3] This proof uses values of x and of y that are in $\{0,1\}$. However, if we consider the same problem in which the input is restricted to be positive-length ranges of the variables, one can prove a similar result with sufficiently small ranges and a value of R that is accordingly adjusted.