Minimum error entropy principle for learning

Ding-Xuan Zhou

City University of Hong Kong

E-mail: mazhou@cityu.edu.hk

Supported in part by Research Grants Council of Hong Kong

Start

July 10, 2013

Outline of the Talk

I. Least squares regression and ERM

II. Regularized least squares regression and kernel PCA

III. Minimum error entropy principle and kernel approximation

IV. MEE algorithm with large parameter

V. MEE algorithm with small parameter

I. Least squares regression and ERM

I.1. Model for the least squares regression. Learn $f : \mathcal{X} \to \mathcal{Y}$ from random samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$

Take $\mathcal{X}$ to be a compact metric space and $\mathcal{Y} = \mathbf{R}$. $y \approx f(x)$ Due to noises or other uncertainty, we assume a (unknown) probability measure ρ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ governs the sampling.

marginal distribution ρ_X on $\mathcal{X}$: $\mathbf{x} = \{x_i\}_{i=1}^m$ drawn according to ρ_X

conditional distribution $\rho(\cdot|x)$ at $x \in \mathcal{X}$

Learning the **regression function**: $f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x)$

$y_i \approx f_\rho(x_i)$

I.2. Least squares generalization error

$\mathcal{E}^{ls}(f) = \int_{\mathcal{Z}} (f(x) - y)^2 d\rho$ minimized by f_ρ:

$$\mathcal{E}^{ls}(f) - \mathcal{E}^{ls}(f_\rho) = \|f - f_\rho\|^2_{L^2_{\rho_X}} =: \|f - f_\rho\|^2_{\rho_X} \geq 0.$$

Empirical Risk Minimization (ERM)

Let $\mathcal{H}$ be a compact subset of $C(\mathcal{X})$ called hypothesis space. The ERM algorithm is given by

$$f_{\mathbf{z}} = \arg\min_{f \in \mathcal{H}} \mathcal{E}^{ls}_{\mathbf{z}}(f), \qquad \mathcal{E}^{ls}_{\mathbf{z}}(f) = \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2.$$

Theory of uniform convergence: bound $\sup_{f \in \mathcal{H}} \left| \mathcal{E}^{ls}_{\mathbf{z}}(f) - \mathcal{E}^{ls}(f) \right|$ by capacity of $\mathcal{H}$.

Approximation error $\inf_{f \in \mathcal{H}} \left\{ \mathcal{E}^{ls}(f) - \mathcal{E}^{ls}(f_\rho) \right\} = \inf_{f \in \mathcal{H}} \|f - f_\rho\|^2_{\rho_X}$

I.3. Error analysis in statistics for the least squares regression

Special example: Let $\mathcal{X}$ be a bounded domain of $\mathbb{R}^n$ with Lipschitz boundary, and $\mathcal{H} = \{f \in H^s(\mathcal{X}) : \|f\|_{H^s(\mathcal{X})} \leq R\}$ be a ball of the Sobolev space $H^s(\mathcal{X})$ with index $s > \frac{n}{2}$.

Error bounds in the literature of statistics: If $f_\rho \in \mathcal{H}$, then with confidence $1 - \delta$,

$$\|f_{\mathbf{z}} - f_\rho\|_{\rho_X}^2 \leq \tilde{C} m^{-\frac{1}{1+n/(2s)}} \log \frac{2}{\delta},$$

provided that the output random variable Y satisfies $|Y| \leq M$ almost surely (standard), or exponential decay for Y in some form.

Heavy tailed noise: Y does not decay exponentially: $\mathbb{E}[|Y|^4] < \infty$ (Audibert-Catoni 2011)

Our work in minimum error entropy algorithm (Hu-Fan-Wu-Zhou): $\mathbb{E}[|Y|^q] < \infty$ for some $q > 2$

II. Regularized least squares regression and kernel PCA

II.1. Regularized least squares regression. Mercer kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ continuous, symmetric and positive semidefinite

Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_K$: completion of the span of the set of functions $\{K_t = K(t, \cdot) : t \in \mathcal{X}\}$

Regularized least squares regression to avoid over-fitting

$$f_{\mathbf{z},\lambda} := \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2 + \lambda \|f\|_K^2 \right\}, \qquad (1)$$

where $\lambda = \lambda(m) > 0$ is a regularization parameter.

Representer Theorem $\Rightarrow$ $f_{\mathbf{z},\lambda} = \sum_{i=1}^{m} c_{i,\mathbf{z}} K_{x_i}$ where $\{c_{i,\mathbf{z}}\}_{i=1}^{m}$ can be solved by a linear system. But there is no sparsity in this representation in general.

II.2. Principle component analysis (PCA)

Let $X = [x_1|x_2|\cdots|x_m] \in \mathbb{R}^{n\times m}$ represent m sampling points in $\mathbb{R}^n$. We seek a k-dimensional affine space $\{\mu + U\beta : \beta \in \mathbb{R}^k\}$ in $\mathbb{R}^n$ to best approximate the data, where $U = [u_1|u_2|\cdots|u_k] \in \mathbb{R}^{n\times k}$ consists of k-columns of an orthogonal matrix.
Solution to PCA: $\widehat{\mu} = \overline{x} = \frac{1}{m}\sum_{i=1}^m x_i$. Define the sample covariance matrix

$$\Sigma = \frac{1}{m}\,[x_1 - \overline{x}|\cdots|x_m - \overline{x}]\,[x_1 - \overline{x}|\cdots|x_m - \overline{x}]^T \in \mathbb{R}^{n\times n}.$$

It has eigenvalues $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_n \geq 0$ corresponding to normalized eigenvectors $\widehat{u}_1, \cdots, \widehat{u}_n$. Its SVD $\Sigma = \sum_{i=1}^n \widehat{\lambda}_i \widehat{u}_i \widehat{u}_i^T$ gives

$$\widehat{U} = [\widehat{u}_1|\widehat{u}_2|\cdots|\widehat{u}_k] \in \mathbb{R}^{n\times k}$$

and a new sample representation

$$x \in \mathbb{R}^n \quad \approx \quad \overline{x} + \sum_{\ell=1}^{k} [(x - \overline{x})\cdot \widehat{u}_\ell]\,\widehat{u}_\ell.$$

II.3. Kernel principle components

Empirical features (kernel principle components): If the Gram matrix $\mathbb{K} := \frac{1}{m}[K(x_i, x_j)]_{i,j=1}^m$ (of rank d) has normalized eigenpairs $\{(\widehat{\lambda}_i^{\mathbf{x}}, \widehat{\mu}_i)\}_{i=1}^m$, the empirical features $\{(\lambda_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})\}$ are defined by $\phi_i^{\mathbf{x}} = \sum_{j=1}^m (\widehat{\mu}_i)_j K_{x_j} / \sqrt{m\widehat{\lambda}_i^{\mathbf{x}}}$ (for $i = 1, \ldots, d$).

Another view: Define the integral operator $L_K f = \int_X f(x) K_x d\rho_X$ on $\mathcal{H}_K$ and the empirical integral operator $L_K^{\mathbf{x}} : \mathcal{H}_K \to \mathcal{H}_K$ by

$$L_K^{\mathbf{x}} f = \frac{1}{m} \sum_{i=1}^m f(x_i) K_{x_i} = \frac{1}{m} \sum_{i=1}^m \langle f, K_{x_i} \rangle_K K_{x_i}.$$

Denote $\{(\lambda_i, \phi_i)\}$ the normalized eigenpairs of L_K, and $\{(\lambda_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})\}$ normalized eigenpairs of $L_K^{\mathbf{x}}$. Then we have the expressions $\lambda_i^{\mathbf{x}} = \widehat{\lambda}_i^{\mathbf{x}}$ and $\phi_i^{\mathbf{x}} = \sum_{j=1}^m (\widehat{\mu}_i)_j K_{x_j} / \sqrt{m\widehat{\lambda}_i^{\mathbf{x}}}$ for $i = 1, \ldots, d$, and $\phi_i^{\mathbf{x}}(x_j) = 0$ for $i > d$.

Kernel PCA: Schölkopf-Smola-Müller, ...

II.4. **Regularized kernel PCA** (Guo-Fan-Zhou)

output function $f^{\mathbf{Z}} = \sum_{i=1}^{\infty} c_i^{\mathbf{Z}} \phi_i^{\mathbf{X}}$ with $c^{\mathbf{Z}} = (c_i^{\mathbf{Z}})_{i=1}^{\infty}$ given by

$$\arg\min_{c \in \ell^2} \left\{ \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{j=1}^{\infty} c_j \phi_j^{\mathbf{X}}(x_i) - y_i \right)^2 + \gamma \sum_{j=1}^{\infty} \Omega(|c_j|) \right\}, \qquad (2)$$

where $\Omega : [0, \infty) \to \mathbb{R}_+$ is a univariate function.

Zwald, Blanchard-Massart-Vert-Zwald: kernel projection machine with $\{\phi_j^{\mathbf{X}}\}_{j=1}^{J}$ and Ω the indicator function of $(0, \infty)$

Examples of Ω: (a) $\Omega(|c|) = |c|^2$ gives (1), the ridge regression
(b) $\Omega(|c|) = |c|^q$ with $0 < q \leq 1$: Fu-Knight, ...
(c) SCAD penalty given by $\Omega'(c) = 1$ for $0 < c < 1$, $\Omega'(c) = 0$ for $c > b$, Ω' continuous and linear on $(1, b)$, with a parameter $b > 2$: Fan-Li, ...

Denote

$$S_i^{\mathbf{z}} = \begin{cases} \frac{1}{m\lambda_i^{\mathbf{x}}} \sum_{j=1}^m y_j \phi_i^{\mathbf{x}}(x_j), & \text{if } \lambda_i^{\mathbf{x}} > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Theorem 1 *Let $\Omega : [0, \infty) \to [0, \infty)$, $\gamma > 0$ and $\mathbf{z} \in (X \times Y)^m$. Then a sequence $c^{\mathbf{z}} = (c_i^{\mathbf{z}})_{i=1}^{\infty}$ is a solution to (2) if and only if for each i, $c_i^{\mathbf{z}}$ is a minimizer of the univariate function defined by*

$$h_i(c) = h_{\lambda_i^{\mathbf{x}}, S_i^{\mathbf{z}}, \gamma, \Omega}(c) = \lambda_i^{\mathbf{x}}(c - S_i^{\mathbf{z}})^2 + \gamma\Omega(|c|), \qquad c \in \mathbb{R}. \quad (3)$$

In particular, if $\Omega(c) > 0$ for $c > 0$, we have $c_i^{\mathbf{z}} = 0$ for $i > m$.

Properties induced by concave penalties

Theorem 2 *If* $\Omega : [0,\infty) \to [0,\infty)$ *is a nonzero continuous concave function satisfying* $\Omega(0) = 0$*, then* $\Omega(1) > 0$ *and that* $\Omega(c) \geq \Omega(1)c$ *for* $c \in (0,1]$ *and* $\Omega(c) \leq \Omega(1)c$ *for* $c \in [1,\infty)$*.*

If $\Omega'_+(0) = 0$*, then for each* i*,* $c_i^{\mathbf{z}}$ *vanishes if and only if either* $\lambda_i^{\mathbf{x}} = 0$ *or* $S_i^{\mathbf{z}} = 0$*.*

So for $\Omega(|c|) = |c|^2$ corresponding to (1) (ridge regression), sparsity is hard to get.

Concave exponent $q \in [0,1]$: For our mathematical analysis, we assume that for some $q \in [0,1]$ and $C_\Omega^* > 0$ there holds

$$\Omega(c) \leq C_\Omega^* c^q, \qquad \forall c \in (0,1].$$

Theorem 3 *Assume* $f_\rho = L_K^r(g_\rho)$ *for some* $r > \frac{1}{2}$ *and* $g_\rho \in \mathcal{H}_K$,

$$D_1 i^{-\alpha} \leq \lambda_i \leq D_2 i^{-\alpha}, \qquad \forall i \in \mathbb{N} \tag{4}$$

for some positive constants D_1, D_2 *and* α *with* $2\alpha \max\{r, 1\} > 1$. *Let* $0 < \delta < 1$. *If we choose*

$$\gamma = C_1 (D_2/\lambda_1)^{r+1} \left(\log \frac{4m}{\delta}\right)^{1+2r} m^{-\frac{3}{4}}, \tag{5}$$

then with confidence $1 - \delta$ *we have*

$$c_i^{\mathbf{z}} = 0, \qquad \forall\ m^{\theta_{sp}} + 1 \leq i \leq m \quad \text{with } \theta_{sp} = \frac{1}{\alpha(1+2r)} < 1 \tag{6}$$

and

$$\|f^{\mathbf{z}} - f_\rho\|_K \leq C_2 \left(\log \frac{4m}{\delta}\right)^{1+2r} m^{-\theta_{rate}}$$

with $\theta_{rate} = \frac{\alpha \min\{6r-1, 4r(2-q)\} - 2(2-q)}{4(2r+1)(2-q)\alpha}$, *and constants* C_1 *and* C_2 *independent of* m *or* δ.

Simulation on MHC-peptide binding data

The quantitative Immune Epitope Database (IEDB) benchmark data set of human leukocyte antigen (HLA)–peptide binding affinities consists of 14 groups, each containing the affinities of a set $\mathcal{P}_a$ of peptides to a specific HLA allele a. For $p \in \mathcal{P}_a$, the affinity $y_p \in [0,1] \subset \mathcal{Y} = \mathbb{R}$ is a real number.

Nielsen and Lund used an artificial neural network-based algorithm called NN-align and gave on this data set the state-of-the-art prediction in 2009. Shen-Wong-Xiao-Guo-Smale (2012) developed a string kernel $\widehat{K}^3$, and applied it with the regularized kernel least squares regression algorithm (RLS), which produced slightly better prediction than NN-align on the same data set. We use this $\widehat{K}^3$ in (2) with $\Omega(c) = |c|^q$, where q is set to be 1, 2/3, and 1/3, and with the SCAD penalty, and achieve some sparsity in addition to the comparable (slightly better) precision.

We divide $\mathcal{P}_a$ into 5 disjoint subsets for a 5-fold cross-validation. Within the training data, another 5-fold cross-validation is employed to select the regularization parameter to minimize the RMSE score. Then algorithm (2) is trained to predict the affinities. Each peptide $p \in \mathcal{P}_a$ has a predicted affinity $\tilde{y}_p$. We use

$$\mathcal{E}_{\mathsf{RMSE},a} = \left(\frac{1}{\#\mathcal{P}_a} \sum_{p \in \mathcal{P}_a} (\tilde{y}_p - y_p)^2 \right)^{1/2}$$

as the RMSE score. A lower RMSE score indicates a better performance.

The area under the receiver operating characteristic (ROC) curve (AUC), defined as

$$\mathcal{E}_{\mathsf{AUC},a} = \frac{\#\left\{(p,p') : p \in \mathcal{P}_{a,B}, p' \in \mathcal{P}_{a,N}, \tilde{y}_p > \tilde{y}_{p'}\right\}}{\left(\#\mathcal{P}_{a,B}\right)\left(\#\mathcal{P}_{a,N}\right)} \in [0,1],$$

is another performance index. Here $\mathcal{P}_{a,B} = \{p \in \mathcal{P}_a : y_p > 0.426\}$ and $\mathcal{P}_{a,N} = \mathcal{P}_a \backslash \mathcal{P}_{a,B}$ are the sets of binding peptides and non-binding ones, with the threshold 0.426. A higher AUC score indicates a better performance.

The simulation is summarized in the following able. Each cell consists of the average of proportions of the non-zero coefficients in the five rounds of test (the top percentage), RMSE (the middle number), and AUC (the bottom number).

Allele a	$\#\mathcal{P}_a$	NN-align	RLS	RKPCA $q=1$	RKPCA $q=2/3$	RKPCA $q=1/3$
DRB1*0101	5166	–	–	74.65%	59.30%	60.81%
		–	0.18660	0.18690	0.18746	0.18830
		0.836	**0.85707**	0.85651	0.85512	0.85306
DRB1*0301	1020	–	–	88.04%	71.84%	56.47%
		–	0.18497	0.18476	0.18495	0.18551
		0.816	0.82813	**0.82995**	0.82950	0.82714
DRB1*0401	1024	–	–	72.39%	60.16%	61.40%
		–	0.24055	0.24089	0.24202	0.24277
		0.771	**0.78431**	0.78023	0.77697	0.77505
DRB1*0404	663	–	–	70.55%	57.84%	57.88%
		–	0.20702	0.20797	0.20918	0.20878
		0.818	0.81425	0.81695	0.81134	0.80801
DRB1*0405	630	–	–	81.47%	69.56%	63.06%
		–	0.20069	0.20037	0.20017	0.20076
		0.781	0.79296	0.79837	**0.79929**	0.79791
DRB1*0701	853	–	–	98.65%	91.76%	86.96%
		–	0.21944	0.21826	0.21840	0.21849
		0.841	0.83440	0.83883	0.83918	0.83916
DRB1*0802	420	–	–	96.85%	93.75%	87.98%
		–	0.19666	0.19555	0.19557	0.19572
		0.832	0.83538	**0.83968**	0.83938	0.83749

DRB1*0901	530	– – 0.616	– 0.25398 **0.66591**	73.11% 0.25563 0.66293	53.35% 0.25653 0.66273	50.94% 0.25784 0.66163
DRB1*1101	950	– – 0.823	– 0.20776 0.83703	94.61% 0.20799 0.83679	83.82% 0.20802 0.83680	80.21% 0.20780 **0.83706**
DRB1*1302	498	– – **0.831**	– 0.22569 0.80410	84.99% 0.22518 0.80479	72.64% 0.22540 0.80439	62.25% 0.22578 0.80303
DRB1*1501	934	– – 0.758	– 0.23268 **0.76436**	75.80% 0.23318 0.76258	64.94% 0.23401 0.76086	74.79% 0.23419 0.76058
DRB3*0101	549	– – **0.844**	– 0.15945 0.80228	92.94% 0.15932 0.80504	89.57% 0.15916 0.80546	87.52% 0.15911 0.80622
DRB4*0101	446	– – **0.811**	– 0.20809 0.81057	96.75% 0.20765 0.81096	81.28% 0.20838 0.80791	76.18% 0.20834 0.80713
DRB5*0101	924	– – 0.797	– 0.23038 **0.80568**	100.00% 0.23045 0.80549	99.95% 0.23045 0.80550	98.76% 0.23046 0.80557
Average		– – 0.7982	– 0.21100 0.80260	85.77% 0.21101 **0.80351**	74.98% 0.21141 0.80246	71.80% 0.21170 0.80136

We make some observations from the simulation:

(a) In terms of AUC on this real data set, RLS has better performance than NN-align. The improvement is 0.55% in average, with better AUC scores for 9 out of 14 test groups while the score difference is always at the second significant figure. Algorithm (2) with $\Omega(c) = |c|$ has even slightly better performance, giving an improvement of 0.11% in average, and better AUC scores for 8 out of 14 test groups with the score difference always at the third significant figure only. Improvements in Shen-Wong-Xiao-Guo-Smale (2012) and in our simulation seem to be small, but this data set has been well investigated in the immunological literature and any improvement is difficult. In particular, the dissimilarity metric BLOSUM62-2 among the 20 basic amino-acids, based on which the string kernel $\hat{K}^3$ is constructed was obtained in a very tight form after long-term effort and a vast medical literature (Henikoff and Henikoff 1992).

(b) As the parameter q (the same as the concave exponent) in the ℓ^q-regularizer $\Omega(|c|) = |c|^q$ decreases from 1 to $\frac{2}{3}$ and $\frac{1}{3}$, the sparsity improves for 10 out of 14 test groups while the AUC worsens for 9 out of 14 test groups. The result in terms of the AUC error is consistent with our theoretical analysis for the error bound stated in terms of the concave exponent q.

(c) Sparsity and error bounds in terms of both AUC and root-mean-square error for the simulation with the SCAD penalty is almost the same on this real data set as that with $\Omega(|c|) = |c|$, verifying again the role of the concave exponent $q = 1$.

III. Minimum error entropy (MEE) principle and kernel approximation

Principe, Erdogmus-Principe, Suykens, ...

Applications: adaptive system training, blind source separation, maximally informative subspace projections, clustering, feature selection, blind deconvolution, ...
Idea: extract from data as much information as possible about the data generating systems by minimizing error entropies in various ways.

Shannon's entropy of a random variable E with pdf p_E is

$$H_S(E) = -\mathbb{E}[\log p_E] = -\int p_E(e) \log p_E(e) de$$

and Rényi's entropy of order α ($\alpha > 0$ but $\alpha \neq 1$) is

$$H_{R,\alpha}(E) = \frac{1}{1-\alpha} \log \mathbb{E}[p_E^{\alpha-1}] = \frac{1}{1-\alpha} \log \left(\int (p_E(e))^\alpha de \right)$$

satisfying $\lim_{\alpha \to 1} H_{R,\alpha}(E) = H_S(E)$.

In supervised learning, the random variable E is $E = Y - f(X)$ when a predictor $f(X)$ is used.

MEE principle: search for a predictor $f(X)$ that contains the most information of the response variable Y by minimizing information entropies of the error variable $E = Y - f(X)$.

The classical least squares error $\mathbb{E}[Y - f(X)]^2 = \int e^2 p_E(e) de$ minimizes the variance of E involving the first two moments and is perfect to deal with Gaussian noise. But it does not work necessarily well for problems involving heavy tailed non-Gaussian noise. For such problems, MEE might still perform very well in principle since moments of all orders of the error variable are taken into account by entropies.

Here we only consider Rényi's entropy of order 2: $H_R(f) = H_{R,2}(f(X) - Y) = -\log \int (p_E(e))^2 de = -\log \int p_E(e) p_E(e) de$.

Kernel approximation of the pdf p_E by Parzen windowing

$$\widehat{p}_E(e) = \frac{1}{mh} \sum_{i=1}^{m} G(\frac{(e - e_i)^2}{2h^2}),$$

where $e_i = y_i - f(x_i)$, $h > 0$ is an MEE scaling parameter, and G is a windowing function (example: $G(t) = \exp\{-t\}$ Gaussian windowing).

Approximations of Rényi's entropy is $-\log(\frac{1}{m}\sum_{i=1}^{m} \widehat{p}_E(e_i))$:

$$\widehat{H_R} = -\log \frac{1}{m^2 h} \sum_{i=1}^{m} \sum_{j=1}^{m} G\left(\frac{(e_i - e_j)^2}{2h^2}\right).$$

The MEE learning algorithm associated with $\mathcal{H}$ and G is

$$f_{\mathbf{z}} = \arg\min_{f \in \mathcal{H}} \left\{ -\log \frac{1}{m^2 h} \sum_{i=1}^{m} \sum_{j=1}^{m} G\left(\frac{\left[(y_i - f(x_i)) - \left(y_j - f(x_j)\right)\right]^2}{2h^2}\right)\right\}. \tag{7}$$

IV. MEE algorithm with large parameter: $h \geq 1$

Hu-Fan-Wu-Zhou (2013)

Assumption on the output variable Y:

$$\mathbb{E}[|Y|^q] < \infty \text{ for some } q > 2, \text{ and } f_\rho \in L^\infty_{\rho_X}. \tag{8}$$

Assumption on the windowing function G: $G \in C^2[0,\infty)$, $G'_+(0) = -1$, and

$$C_G := \sup_{t\in(0,\infty)} \Big\{ |(1+t)G'(t)| + |(1+t)G''(t)| \Big\} < \infty. \tag{9}$$

Assumption on the covering number $\mathcal{N}(\mathcal{H},\varepsilon)$ of $\mathcal{H}$: for some constants $p > 0$ and $A_p > 0$, there holds

$$\log \mathcal{N}(\mathcal{H},\varepsilon) \leq A_p \varepsilon^{-p}, \qquad \forall \varepsilon > 0. \tag{10}$$

Theorem 4 *For any* $0 < \delta < 1$, *with confidence* $1 - \delta$ *we have*

$$\mathbf{var}[f_{\mathbf{z}}(X) - f_\rho(X)] \leq \widetilde{C}_{\mathcal{H}} m^{-\frac{\min\{q-2,2\}}{(1+p)(1+\min\{q-2,2\})}} \log \frac{2}{\delta} + 2 \inf_{f \in \mathcal{H}} \mathbf{var}[f(X) - f_\rho(X)] \quad (11)$$

by taking $h = m^{\frac{1}{(1+p)\min\{q-1,3\}}}$.

If $|Y| \leq M$ *almost surely for some* $M > 0$, *then by taking* $h = m^{\frac{1}{2(1+p)}}$, *with confidence* $1 - \delta$ *we have*

$$\mathbf{var}[f_{\mathbf{z}}(X) - f_\rho(X)] \leq \widetilde{C}_{\mathcal{H}} m^{-\frac{1}{1+p}} \log \frac{2}{\delta} + 2 \inf_{f \in \mathcal{H}} \mathbf{var}[f(X) - f_\rho(X)]. \quad (12)$$

Here $\widetilde{C}_{\mathcal{H}}$ *is a constant independent of* m, δ *or* h.

Key feature for large parameter

The generalization error or information error is

$$\mathcal{E}^{(h)}(f) = \int_{\mathcal{Z}} \int_{\mathcal{Z}} -h^2 G\left(\frac{[(y - f(x)) - (y' - f(x'))]^2}{2h^2}\right) d\rho(x, y) d\rho(x', y').$$

An essential barrier: f_ρ may not be a minimizer of $\mathcal{E}^{(h)}$. This is different from $\mathcal{E}^{ls}(f) = \int_{\mathcal{Z}} (f(x) - y)^2 d\rho$.

Key observation for large h: Under assumptions (8) and (9), for $f \in L^\infty$, we have

$$\begin{aligned} &\left|\mathcal{E}^{(h)}(f) + h^2 G(0) - \mathcal{E}^{ls}(f_\rho) - \mathbf{var}[f(X) - f_\rho(X)]\right| \\ &\leq 5 \cdot 2^7 C_G \left((\mathbb{E}[|Y|^q])^{\frac{\min\{q,4\}}{q}} + \|f\|_\infty^{\min\{q,4\}}\right) h^{-\min\{q-2,2\}}. \end{aligned}$$

So $\mathbf{var}[f(X) - f_\rho(X)]$ can be bounded by $\left|\mathcal{E}^{(h)}(f) - \mathcal{E}^{(h)}(f_\rho)\right|$.

V. MEE algorithm with small parameter

Two types of consistency with $G(t) = \exp\{-t\}$

Entropy consistency:

$$\lim_{m\to\infty} \text{Prob}\Big(H_R(f_{\mathbf{z}}) - \inf_f H_R(f) > \varepsilon\Big) = 0, \quad \forall \varepsilon > 0.$$

Observation from $H_R(f) = -\log \int (p_E(e))^2 de$: define $V(f) = -\int (p_E(e))^2 de$, then there are two positive constants c_1, c_2 such that for every $f : \mathcal{X} \to \mathbb{R}$,

$$c_1 \left(V(f) - \inf_g V(g)\right) \le H_R(f) - \inf_g H_R(g) \le c_2 \left(V(f) - \inf_g V(g)\right).$$

Regression consistency: there is a constant $b_{\mathbf{z}}$ such that $f_{\mathbf{z}} + b_{\mathbf{z}}$ converges to f_ρ in probability, i.e.,

$$\lim_{m\to\infty} \text{Prob}\Big(\|f_{\mathbf{z}} + b_{\mathbf{z}} - f_\rho\|_{\rho_X}^2 > \varepsilon\Big) = 0, \quad \forall \varepsilon > 0.$$

Positive result on entropy consistency

Assumption on ρ: the density function $p_{\epsilon|X}$ of the noise variable $\epsilon = Y - f_\rho(X)$ for given $X = x$ exists and is uniformly bounded.

Assumption on $\mathcal{H}$: a minimizer of the Rényi's entropy $H_R(f)$ and the regression function f_ρ are in $\mathcal{H}$.

Theorem 5 *If $h = h(m)$ is chosen to satisfy*

$$\lim_{m\to\infty} h(m) = 0, \qquad \lim_{m\to\infty} h^2\sqrt{m} = +\infty, \tag{13}$$

then the entropy consistency holds true.

If, in addition, $p'_{\epsilon|X}$ exists and is uniformly bounded by a constant M independent of X, a convergence rate of order $O(m^{-\frac{1}{6}})$ can be obtained by choosing $h(m) \sim m^{-\frac{1}{6}}$.

Regression consistency for homoskedastic models

The regression mode $Y = f_\rho(X) + \epsilon$ is homoskedastic if the noise ϵ is independent of X. Otherwise it is said to be heteroskedastic.

Theorem 6 *If the regression model is homoskedastic, then f_ρ is a minimizer of the Rényi's entropy $H_R(f)$. Moreover, there is an absolute constant $\widetilde{C}$ such that*

$$\|f - \mathbb{E}(f - f_\rho) - f_\rho\|_{\rho_X}^2 \leq \widetilde{C}\left(H_R(f) - H_R(f_\rho)\right), \quad \forall f : \mathcal{X} \to \mathbb{R}.$$

Hence the regression consistency holds true.

Fourier analysis for homoskedastic models

Since the noise ϵ is independent of X, denote the pdf as p_ϵ. Then the pdf of the random variable $E = Y - f(X)$ is

$$p_E(e) = \int_{\mathcal{X}} p_\epsilon(e + f(x) - f_\rho(x)) d\rho_X(x)$$

and $\int_{\mathbb{R}} (p_E(e))^2 de = -V(f)$ can be expressed as

$$\int_{\mathcal{X}} \int_{\mathcal{X}} \int_{\mathbb{R}} p_\epsilon(e + f(x) - f_\rho(x)) p_\epsilon(e + f(u) - f_\rho(u)) de d\rho_X(x) d\rho_X(u).$$

By the Planchel formula, the integral on $\mathbb{R}$ equals

$$\frac{1}{2\pi} \int_{\mathbb{R}} \widehat{p_\epsilon}(\xi) \mathbf{e}^{i\xi(f(x) - f_\rho(x))} \overline{\widehat{p_\epsilon}(\xi) \mathbf{e}^{i\xi(f(u) - f_\rho(u))}} d\xi.$$

It follows that f_ρ minimizes $V(f)$ since $|\mathbf{e}^{i\xi t}| \leq 1$ and $-V(f)$ equals

$$\frac{1}{2\pi} \int_{\mathcal{X}} \int_{\mathcal{X}} \int_{\mathbb{R}} |\widehat{p_\epsilon}(\xi)|^2 \mathbf{e}^{i\xi(f(x) - f_\rho(x) - f(u) + f_\rho(u))} d\xi d\rho_X(x) d\rho_X(u).$$

Excess quantity: for any $f \in \mathcal{H}$, $2\pi[V(f) - V(f_\rho)]$ equals

$$\begin{aligned}
&\int_{\mathcal{X}}\int_{\mathcal{X}}\int_{\mathbb{R}} |\widehat{p_\epsilon}(\xi)|^2 \left(1 - \mathbf{e}^{i\xi(f(x)-f_\rho(x)-f(u)+f_\rho(u))}\right) d\xi d\rho_X d\rho_X \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}}\int_{\mathbb{R}} |\widehat{p_\epsilon}(\xi)|^2 2\sin^2 \frac{\xi(f(x) - f_\rho(x) - f(u) + f_\rho(u))}{2} d\xi d\rho_X d\rho_X \\
&\geq \int_{\mathcal{X}}\int_{\mathcal{X}}\int_{|\xi|\leq\frac{\pi}{4M}} |\widehat{p_\epsilon}(\xi)|^2 2(\frac{1}{\pi}|\xi(f(x) - f_\rho(x) - f(u) + f_\rho(u))|)^2 d\xi d\rho_X \\
&\geq \widetilde{c}\int_{\mathcal{X}}\int_{\mathcal{X}} (f(x) - f_\rho(x) - f(u) + f_\rho(u))^2 \rho_X(x) d\rho_X(u) \\
&= 2\widetilde{c}\mathbf{var}[f(X) - f_\rho(X)].
\end{aligned}$$

So there holds

$$\mathbf{var}[f_{\mathbf{z}}(X) - f_\rho(X)] \leq \frac{\pi}{\widetilde{c}}(V(f_{\mathbf{z}}) - V(f_\rho)) \leq \frac{\pi}{c_1\widetilde{c}}\left(H_R(f_{\mathbf{z}}) - \inf_g H_R(g)\right).$$

Fourier analysis for heteroskedastic models:

One example: Let $\mathcal{X} = \mathcal{X}_1 \bigcup \mathcal{X}_2 = [0, \frac{1}{2}] \bigcup [1, \frac{3}{2}]$ and ρ_X be uniform on $\mathcal{X}$ (so that $d\rho_X = dx$). The conditional distribution of $\epsilon | X$ is uniform on $[-\frac{1}{2}, \frac{1}{2}]$ if $x \in [0, \frac{1}{2}]$ and uniform on $[-\frac{3}{2}, -\frac{1}{2}] \bigcup [\frac{1}{2}, \frac{3}{2}]$ if $x \in [1, \frac{3}{2}]$. Then we have the following statements.

(1) A function $f^* : \mathcal{X} \to \mathbb{R}$ is a minimizer of $H_R(f)$ if and only if there are two constant f_1, f_2 with $|f_1 - f_2| = 1$ such that $f^* = f_1 \mathbf{1}_{\mathcal{X}_1} + f_2 \mathbf{1}_{\mathcal{X}_2}$.

(2) $\inf_{g:\mathcal{X}\to\mathbb{R}} H_R(g) = -\log(\frac{5}{8})$ and $H_R(f_\rho) = -\log(\frac{3}{8})$. So the regression function f_ρ is not a minimizer of the entropy $H_R(f)$.

(3) Let $\mathcal{F}$ denote the set of all minimizers of $H_R(f)$. There is an absolute constant $\widehat{C}$ such that

$$\min_{g \in \mathcal{F}} \|f - g\|_{\rho_X}^2 \leq \widehat{C} \Big(H_R(f) - \inf_g H_R(g) \Big), \quad \forall f : \mathcal{X} \to [-M, M].$$

(4) If the entropy consistency is true, there holds

$$\min_{g\in\mathcal{F}} \|f_{\mathbf{z}} - g\|_{\rho_X}^2 \longrightarrow 0 \quad \text{and} \quad \min_{b\in\mathbb{R}} \|f_{\mathbf{z}} + b - f_\rho\|_{\rho_X}^2 \longrightarrow \frac{1}{2}$$

in probability. As a result, the regression consistency cannot be true.

Method of analysis by bracket products: analysis

$$\sum_{\ell\in\mathbb{Z}} \widehat{p^*}(\xi + 2\ell\pi)\overline{\widehat{p^*(\cdot - b)}(\xi + 2\ell\pi)}$$
$$= \sum_{\ell\in\mathbb{Z}} \langle p^*(\cdot - \ell), p^*(\cdot - b)\rangle_{L^2(\mathbb{R})} \mathbf{e}^{i\ell\xi}.$$

THANK YOU!