LMD mokslo darbai, III tomas, 384–391

Adaptive estimation: a brief review

Marijus RADAVIČIUS (MII)
e-mail: mrad@ktl.mii.lt

1. General remarks

The term 'adaptivity' has two different meanings even in mathematical statistics. For parametric and semiparametric statistical models with nuisance parameters, adaptivity of an estimator means that it is as efficient as the optimal estimator with *a priori* known true values of the nuisance parameters (see, e.g., Bickel (1982)). In this paper we shall discuss the adaptivity in the other (broader) sense (to some extent related to the previous meaning), namely, the adaptivity in nonparametric estimation of functions. In the widest sense the adaptive (statistical) estimators can be thought of as statistical procedures that perform a data-driven selection of the 'best' estimator among estimators known to be 'optimal' for some statistical model from a given class of models. Thus, adaptivity is closely related to the model selection problem, robust and distribution-free estimation. For original interpretation and comprehesive discussion of this subject we recommend the paper by Barron et al. (1999).

A classical nonparametric estimation problem is that of estimating probability density (p.d.). Let $X^N = (X_1, \ldots, X_N)$ be a sample of i.i.d. random variables having a p.d. $f(x), x \in K \subset \mathbf{R}^d$, with respect to some σ-finite measure μ. The problem is to find a 'good' estimator of f based on the sample X^N and prior information that $f \in W$. It is the nonparametric estimation problem we will use to illustrate main ideas of adaptive estimation in what follows. However, it should be emphasized that this problem, along with estimating regression function, spectral density of stationary sequence (process), signal observed in white Gaussian noise (so-called white noise model), ect., is a special case of general problem of estimating some unknown function (treated as infinite-dimensional parameter). The inspection of the existing literature on the topic shows that an idea or a method proposed in one of these areas sooner or later is realized in others as well. Recent advances enables one to formulate this statement in a precise form by means of Le Cam's deficiency pseudodistance Δ (see Le Cam (1986), Le Cam and Yang (1990)). Two statistical models are said to be asymptotically equivalent iff the pseudodistance Δ between them tends to zero together with some asymptotic parameter. Such asymptotic equivalence to the Gaussian white noise model is established for Gaussian regression (Brown and Low (1996)), p.d. (Nussbaum (1996)), and non-Gaussian regression (Grama and Nussbaum (1998)). The equivalence means that the two models are (asymptotically) equivalent for all purposes of statistical decisions with bounded loss functions. It should be emphasized, however, that this equivalence, as pointed out by Efromovich and Samarov (Efromovich and Samarov (1996)), has its limits. It holds only for classes of sufficiently smooth functions and, with exception of Gaussian regression, is nonconstructive (see Nussbaum (1996)).

Anyway, this provides some grounding to look at the nonparametric estimation problem of functions from a general standpoint without separating any special problem, say, p.d. estimation, from the others.

2. Two approaches

Let $\varrho(f, g)$ be some pseudodistance between two functions f and g (or loss function) and let $\delta_N(\hat{f}, f)$ denote the risk (i.e. expected losses) of an estimator $\hat{f} = \hat{f}(X^N)$ based on observations X^N, $N \to \infty$:

$$\delta_N(\hat{f}, f) = \mathbf{E}_f \varrho(\hat{f}, f).$$

For a given set of functions W and a given class of estimators $\mathcal{F}$, define the minimax risk

$$\delta_N(\mathcal{F}, W) = \inf_{\hat{f} \in \mathcal{F}} \sup_{f \in W} \delta_N(\hat{f}, f)$$

with the natural conventions $\delta_N(\mathcal{F}, f) = \delta_N(\mathcal{F}, \{f\})$ and $\delta_N(\hat{f}, W) = \delta_N(\{\hat{f}\}, f)$. We set $\delta_N(W) = \delta_N(\mathcal{F}, W)$ if the $\mathcal{F}$ is the set of all estimators of f based on X^N.

There are *two approaches* to the nonparametric estimation problem.

2.1. Target function approach.

According to this approach a class $\mathcal{F} = \{\hat{f}_\beta, \beta \in \mathcal{B}\}$ of estimators of $f \in W$ is given, where β is some (tuning) parameter. For example:

(1) for projection estimators (Cencov (1962, 1982)), β is the dimension of a projection space, for orthogonal series estimators β, is a sequence of weights (Hall (1987)),

(2) for kernel estimators, β is a bandwidth (smoothing parameter) (Stone (1984)),

(3) for spline estimators, β can be a number of knots, a set of knots, penalization or regularization parameters (Wahba (1990)), ect.

The task is to choose β so as to minimize the risk $\delta(\hat{f}_\beta, f)$ (or some its approximation or estimate). Clearly, the optimal value $\beta^* = \beta^*(f)$ of the tuning parameter β depends on the unknown function f, usually through some smoothness characteristics of f. Thus, the problem is to find a data-driven value $\hat{\beta}$ of β such that the ratio

$$r(\hat{\beta}) = \limsup_{N \to \infty} \frac{\delta_N(\hat{f}_{\hat{\beta}}, f)}{\delta_N(\mathcal{F}, f)}, \qquad f \in W,$$

would be as small as possible. This is called *adaptivity to the target function* (see Barron et al. (1999)).

This approach, based on the prespecified class of estimators $\mathcal{F}$, gives no way to compare different estimating techniques (simulation beeing an evident exception) and to evaluate the loss in adaptivity (efficiency) due to the restriction of estimating procedures to $\mathcal{F}$ instead of the set of all estimators.

2.2. Minimax approach.

This approach is based on the notion of the minimax risk (efficiency).

Definition 1. An estimator $\hat{f}_W$ is called (asymptotically) minimax (efficient) on W iff

$$\delta_N(\hat{f}_W, W) \sim \delta_N(W).$$

Again, a minimax estimator $\hat{f}_W$, if it exists and can be found, depends on the set W (may be through some simple its smoothness characteristics). Thus, we have a family of estimators $\{\hat{f}_W, W \in \mathcal{W}\}$ parametrized by W where $\mathcal{W}$ is some class (scale) of sets of (smooth) functions. Since W is frequently *a priori* unknown, we have the situation very similar to that in the first (target function) approach. However, the difference does exist: now we deal with the class of (asymtotically) optimal estimators. The estimators which are independent of W and share some minimaxity properties (see below) are called *minimax adaptive*.

3. Minimax estimation

Let

$$\Delta(\hat{f}, W) = \limsup_{N\to\infty} \frac{\delta_N(\hat{f}, W)}{\delta_N(W)}.$$

If $\Delta(\hat{f}, W) < \infty$ the estimator $\hat{f}$ is called *rate minimax* (variants: minimax in rate, near minimax).

Apparently, Chencov (1962, 1982) was the first who established a lower bound (nonasymptotic!) for the minimax risk $\delta(W)$ with L_2-losses. Under some (geometric) conditions on $W \in L_2$ he proved that

$$\delta_N(W) \geq \gamma(N)/N \quad \forall N,$$

where $\gamma(N) = \gamma(N, W) \to \infty$. He also introduced projection estimators and showed their rate minimaxity (for a properly chosen projection space). Rate minimax estimators are found for a wide class of losses (including L_p losses, $1 \leq p \leq \infty$) for ellipsoids and hyperrectangulars in L_2, large scale of subsets in Sobolev, Hölder, Besov spaces (see Ibragimov and Khas'minskii (1981), Brentagnole and Huber (1979), Stone (1982), Bentkus and Kazbaras (1981), Bentkus (1985), Donoho (1990), Kerkyacharian and Picard (1993), and references therein).

3.1. Minimax efficiency.

Up to now there are only three special cases where minimax estimators are known.

1. *The case of* L_2. In the pioneering paper Pinsker (1980) proposed the minimax estimator for the squared L_2 distance and W being an ellipsoid in L_2. This result was extended to other nonparametric models: distribution and spectral densities (Efroimovich and Pinsker (1981, 1982)) and regression (Nussbaum (1985), Golubev and Nussbaum (1990), Efromovich (1996)). Tsybakov (1997) has shown that the result holds also for non-squared losses, $\varrho(f, g) = w(\|f - g\|_2)$,

and can also be applied when estimating derivatives. Rudzkis and Radavičius (1993) built up minimax estimators for sets $W \subset L_2$ defined via accuracy of their approximation by the finite-dimensional linear subspaces in L_2 generated by a given orthobasis. This approach was proposed by Cencov (1962, 1982).

2. *Sup-norm in Hölder classes.* Korostelev (1993) found the minimax estimators for Hölder classes with L_∞ risk (see also Donoho (1994)).

3. *Analytical functions.* In the case where W is a certain class of analytical functions, the minimax estimators for general L_p risk are proposed in (Golubev et al. (1996), the case $p = \infty$) and in (Guerre and Tsybakov (1998), the case $1 \leq p < \infty$).

3.2. Locally minimax estimation.

It worth noting that the notion of the minimax efficiency of nonparametric estimators under consideration differs from the classical (asymptotic) minimax efficiency in the parametric situation (see, e.g., Ibragimov and Khas'miskii (1981)). The key point is that, in the first case, the supremum is taken over the whole set W, whereas in the latter case the supremum is taken over W_g^N, $g \in \text{interior}(W)$, where W_g^N is a contracting ($N \to \infty$) neighborhood of g. Let W° be some (everywhere dense) subset of W.

Definition 2. An estimator $\hat{f}$ is called *locally minimax* iff

$$\delta_N(\hat{f}, W_g^N) \sim \delta_N(W_g^N) \qquad \forall g \in W^\circ.$$

Asymptotic lower bounds for the locally minimax risk with quadratic losses were established by Golubev (1991) who exploited a condition of nonparametric local asymptotic normality. He also proposed locally minimax estimators (and adaptive!, see below) for distribution and spectral densities (see Golubev (1992, 1993)). However, Golubev's estimators are attached to the trigonometric basis and are not minimax if the nonparametric estimation problem is actually parametric (even this fact is known *a priori*).

Another approach to the locally minimax efficiency is proposed in Rudzkis and Radavičius (1993). It applies to any differentiable orthobasis and enables one to treat both nonparametric and parametric problems in a unified way. In some special cases (for example, fitting p.d. by orthogonal polynomials) the construction of the locally minimax estimator can be considerably simplified (Radavičius (1995)). For general asymptotic lower bound of the locally minimax risk, see (Radavičius (1997)).

4. Adaptive estimation

4.1. Adaptive estimation of target function.

If $r(\hat{\beta}, f) = 1$ (see Subsection 2.1), the estimator is called *adaptive*. However, more common is the situation where $r(\hat{\beta}, f) < \infty \; \forall f \in W$. Such an estimate it is natural to call *rate adaptive*, but usually the same term is retained. The widely used methods for selection of $\hat{\beta}$ are *plug-in*,

cross-validation, generalized cross-validation and *bootstrap* (see, for instance, Wahba (1990), Härdle (1990), Davison and Hinkley (1997), Franke and Härdle (1992)). For adaptive kernel estimators, we refer, e.g., to Hall (1983), Stone (1984) and Kazbaras (1986), for projection (orthogonal series) estimators, we cite Rudzkis (1985), Kazbaras (1987), Hall (1987) among others.

Since the smoothness of the function f is frequently subject to significant variations through K, definition region of f (as, for instance, in the case of mixture of distributions or piecewise constant regression), it is natural to assume that β depends on the location $x \in K$. Thus, taking $\beta = \beta(x)$, we come to the notion of *locally (or pointwise) adaptive* estimation, which have attracted the considerable interest among statistitians in the last two decades. See Cheng (1997), Brockman et al. (1993), Härdle and Marron (1985), View (1991) among others, for kernel methods, and Luo and Wahba (1997), Mammen and van de Geer (1997), and references therein, for spline estimation.

4.2. Minimax adaptive estimation.

Let $\mathcal{W} = \{W(b), b \in B\}$ be some scale of sets (in L_p, Hölder, Sobolev or Besov spaces) indexed by a parameter b which characterizes (in some sense) the smoothness of $W(b)$. The set of estimators which are independent of b (but possibly dependent on $\mathcal{W}$) is denoted by $\mathcal{F}_a$.

Definition 3. An estimator $\hat{f} \in \mathcal{F}_a$ is called

(a) *minimax adaptive* if

$$a(\hat{f}) \stackrel{\text{def}}{=} \limsup_{N\to\infty} \sup_{b\in B} \frac{\delta_N(\hat{f}, W(b))}{\delta_N(W(b))} = 1;$$

(b) *optimal rate (minimax) adaptive* if $a(\hat{f}) < \infty$;

(c) *(asymptotically) exact (minimax) adaptive* (Tsybakov (1998)) if the rate is optimal among all optimal rate adaptive estimators, i.e.,

$$a(\hat{f}) = a(\mathcal{F}_a) = \inf_{\tilde{f}\in\mathcal{F}_a} a(\tilde{f});$$

(d) *near (minimax) adaptive* (up to a factor $L(N)$) if the convergence rate of the adaptive estimator is within the factor $L(N)$ to the minimax rate, i.e.,

$$\sup_{b\in B} \frac{\delta_N(\hat{f}, W(b))}{\delta_N(W)} = O(L(N));$$

here $L(N)$ is a slowly varying function, usually a power of logarithm.

Minimax adaptive estimators are known only in the case of estimation in L_2 (Case 1 in Subsection 3.1). See, for instance, Rudzkis (1985), Efroimovich (1985, 1996), Kazbaras (1986, 1987), Efroimovich and Pinsker (1986), Golubev (1990, 1992, 1993).

For L_∞ risk (case 2) minimax adaptive estimators, in general, do not exist (see, e.g., Tsybakov (1998)). For optimal rate adaptive and near adaptive estimation, we refer to Lepskii (1991),

Golubev and Nussbaum (1992), Donoho and Johnstone (1995), Donoho et al. (1995) among others. Exact adaptive estimation is studied by Tsybakov (1998).

Locally (pointwise) adaptive minimax estimation.

Achievements in locally adaptive estimation inspired investigations of locally (pointwise) adaptive estimating procedures from the minimax point of view. There are known (at the present moment) two methods allowing one to attack the problem.

The first method is to describe the local smoothness in terms of balls of Besov spaces and then to build up an estimator by making use of a wavelet basis which is known to have nice approximation properties in Besov spaces (see Donoho and Johnstone (1994, 1995), Donoho et al. (1995, 1996) among others). The local adaptivity is achieved simply through thresholding of the empirical wavelet coefficients.

The second method, exploiting kernel estimators, is based on Lepskii's pointwise minimax results (Lepskii (1990-1992)) and is developed in (Lepskii and Spokoiny (1995, 1997), Lepskii et al. (1997)). Note that for the minimax poitwise risk the situation is quite different from that for global losses. As shown by Lepskii (1990) in this case there is no optimal rate adaptive estimator on the scale of Hölder classes: a logarithmic factor is a necessary payment for adaptation. The reason is that the pointwise losses are not asymptotically degenerate.

References

[1] A. Barron, L. Birge, P. Massart, Risk bounds for model selection via penalization, *Probab. Theory Relat. Fields*, **113**, 301–413 (1999).

[2] R.J. Bentkus, Asymptotics of the minimax mean square risk of statistical estimation of a spectral density. *Lith. Math. J.*, **24**, 93-98 (1985).

[3] R.J. Bentkus, A.R. Kazbaras, On optimal statistical estimation of a distribution density. *Soviet Math. Dokl.*, **23**, 487-490 (1985).

[4] P.J. Bickel, On adaptive estimation *Ann. Statist.*, **10**, 647-671 (1982).

[5] J. Bretangole, C. Huber, Estimation des densities: Risque minimax, *Z. Wahrsch. Verw. Gebiete*, **47**, 119–137 (1979).

[6] M. Brockman, T. Gasser, and E. Herrman, Locally adaptive bandwidth choice for regression estimators, *JASA*, **88**, 1302–1309 (1993).

[7] L.D. Brown, M.G. Low, Asymptotic equivalence of nonparametric regression and white noise, *Ann. Statist.*, **24**, 2384–2398 (1996).

[8] N.N. Cencov, Estimation of the unknown density function, *Dokl. Akad. Nauk SSSR*, **147**, 45–48 (1962), (in Russian).

[9] N.N. Cencov, *Statistical decision rules and optimal inference*, Translations of Mathematical Monographs 53, American Math. Society, Province (1982), (from 1972 in Russian).

[10] M.-Y. Cheng, A bandwidth selector for local linear density estimators, *Ann. Statist.*, **25**, 1001–1013 (1997).

[11] D.L. Donoho, Asymptotic minimax risk for sup-norm loss: solution via optimal recovery. *Probab. Theory Relat. Fields*, **99**, 145–170 (1994).

[12] D.L. Donoho, I.M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, *Biometrika*, **81**, 425–455 (1994).

[13] D.L. Donoho, I.M. Johnstone, Adapting to unknown smoothness via wavelet shrinkage, *JASA*, **90**(432), 1200–1225 (1995).

[14] D.L. Donoho, I.M. Johnstone, Minimax estimation via wavelet shrinkage, *Ann. Statist.*, **26**, 879–921 (1998).

[15] D.L. Donoho, I.M. Johnstone, G. Keryacharian, D. Picard, Density estimation by wavelet thresholding, *Ann. Statist.*, **24**, 508–539 (1996).

[16] S.Yu. Efroimovich, Nonparametric estimation of a density of unknown smoothness, *Theory Probab. Appl.*, **30**, 557–568 (1985).

[17] S. Efromovich, On nonparametric regression for iid observations in a general setting, *Ann. Statist.*, **24**, 1126–1144 (1996).
[18] S.Yu. Efroimovich, M.S. Pinsker, Estimation of square-integrable spectral density from a sequence of observations, *Problems of Information Transmission*, **17**, 50–68 (1981).
[19] S.Yu. Efroimovich, M.S. Pinsker, Estimation of square-integrable probability density of a random variable, *Problems of Information Transmission*, **18**, 175–182 (1982).
[20] S.Yu. Efroimovich, M.S. Pinsker, Self-tuning algorithm for minimax nonparametric estimation of spectral density, *Probl. Inf. Transm.*, **20**, 209–221 (1986).
[21] S. Efromovich, A. Samarov, *Statist. Probab. Letters*, **28**, 143–145 (1996).
[22] G.K. Golubev, Quasilinear estimates of signal in L2, *Problems of Information Transmission*, **26**, 15–20 (1990).
[23] G.K. Golubev, LAN in problems of nonparametric estimations of functions and lower bounds for quadratic risks, *Theory Probab. Appl.*, **36**, 152–157 (1991).
[24] G.K. Golubev, Nonparametric estimation of smooth probability densities in L2, *Problems of Information Transmission*, **28**, 44–54 (1992).
[25] G.K. Golubev, Nonparametric estimation of smooth spectral densities of Gaussian stationary sequence, *Theory Probab. Appl.*, **38**, 630–639 (1993).
[26] G.K. Golubev, M. Nussbaum, Adaptive spline estimates for nonparametric regression models, *Theory Probab.Appl.*, **37**, 521–529 (1992).
[27] I. Grama, M. Nussbaum, Asymptotic equivalence for generalized linear models, *Probab. Theory Relat. Fields*, **111**, 167–214 (1998).
[28] E. Guerre, A.B. Tsybakov, Exact asymptotic minimax constants for the estimation of analytical functions in Lp, *Probab. Theory Relat. Fields*, **112**, 33–51 (1998).
[29] P. Hall, Cross-validation and the smoothing of arthogonal series density estimators, *J. Multiv. Analysis*, **21**, 207–237 (1987).
[30] P. Hall, Large-sample optimality of least-squares cross-validation in density estimation, *Ann. Statist.*, **11**, 1256–1174 (1993).
[31] W. Härdle, J.S. Marron, *Applied nonparametric regression*, Cambridge University Press, Cambridge (1985).
[32] W. Härdle, J.S. Marron, Optimal bandwidth selection in nonparametric regression function estimation, *Ann. Statist.*, **13**, 1466–1481 (1985).
[33] I.A. Ibragimov, R.Z. Khasmin'skii, *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, Berlin, New York (1981).
[34] A. Kazbaras, An adaptive kernel-type estimator of distribution density, *Lith. Math. J.*, **26**, 318–324 (1986).
[35] A. Kazbaras, On an adaptive projective estimator of distribution density, *Liet. Matem. Rink. (Lith. Math. J.)*, **27**, 688–698 (1987) (in Russian).
[36] A.P. Korostelev, Asymptotically minimax regression estimator in the uniform norm up to exact constant, *Theory Probab. Appl.*, **38**, 737–743 (1993).
[37] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York etc. (1986).
[38] L. Le Cam, G. Yang, *Asymptotics in Statistics: Some Basic Concepts*, Springer-Verlag, New York etc.(1990).
[39] O.V. Lepskii, One problem of adaptive estimation in Gaussian white noise, *Theory Probab. Appl.*, **35**, 459–470.
[40] O.V. Lepskii, Asymptotic minimax adaptive estimation. 1. Upper bounds, *Theory Probab. Appl.*, **36**, 645–659.
[41] O.V. Lepskii, Asymptotic minimax adaptive estimation. 2. Statistical model without optimal adaptation. Adaptive estimators, *Theory Probab. Appl.*, **37**, 468–481.
[42] O.V. Lepskii, V.G. Spokoiny, Local adaptation to inhomogeneous smoothness: resolution level, *Mathematical methods of statistics*, **3**, 239–258.
[43] O.V. Lepskii, V.G. Spokoiny, Optimal pointwise adaptive methods in nonparametric estimation, *Ann. Statist.*, **25**, 2512–2546 (1997).
[44] Z. Luo, G. Wahba, Hybrid adaptive splines, *JASA*, **92**(437), 107–116 (1997).
[45] E. Mammen, S. van de Geer, Locally adaptive regression splines, *Ann. Statist.*, **25**, 387–413 (1997).
[46] M. Nussbaum, Asymptotic equivalence of density estimation and Gaussian white noise, *Ann. Statist.*, **24**, 2399–2430 (1996).
[47] M.S. Pinsker, Optimal filtration of square integrable signal on the Gaussian noise background, *Problems of Information Transmission*, **16**, 52–68 (1980).
[48] M. Radavičius, Efficient nonparametric estimation of distribution density in the basis of algebraic polynomials, *Acta Applicandae Mathematicae*, **38**, 13–35 (1995).

[49] M. Radavičius, Lower bound for quadratic losses of infinite-dimensional parameter, *Lith. Math. J.*, **37**, 71–86 (1997).
[50] R. Rudzkis, On an estimate of the spectral density, *Lith. Math. J.*, **25**, 273–280 (1995).
[51] R. Rudzkis, M. Radavičius, Locally minimax efficiency of nonparametric estimates of square-integrable densities, *Lith. Math. J.*, **33**, 56–75 (1993).
[52] C.J. Stone, Optimal global rates of convergence for nonparametric regression, *Ann. Statist.*, **10**, 1040–1053 (1982).
[53] C.J. Stone, An asymptotically optimal window selection rule for kernel density estimates, *Ann. Statist.*, **12**, 1285–1297 (1984).
[54] A.B. Tsybakov, Asymptotically efficient signal estimation in L2 with general loss functions, *Problems of Information Transmission*, **33**, 78–88 (1997).
[55] A.B. Tsybakov, Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes, *Ann. Statist.*, **26**, 2420–2469 (1998).
[56] P. View, Nonparametric regression: optimal local bandwidth choice, *J. Roy. Statist. Soc.*, ser. B, **53**, 453–464 (1991).
[57] G. Wahba, *Spline models for observational data*, S.I.A.M., Phyladelphya (1990).

Adaptyvus vertinimas: trumpa apžvalga

M. Radavičius

Pateikta trumpa funkcijų adaptyvaus neparametrinio vertinimo apžvalga. Iliustracijai paimtas nežinomo tikimybinio tankio neparametrinio vertinimo uždavinys.