Nonparametric Density Estimation

Athanasios G. Georgiadis

Assistant Professor,
Trinity College of Dublin.

SCSS, May 12 2021

Contents

Motivation

Assume that a phenomenon under study is expressed by a random variable (rv) X distributed on some space $\mathcal{M}$.
For having a full understanding of X, we need to know its

Probability Density Function (PDF) $f_X(x) = f(x),\ x \in \mathcal{M}$.

In practice we rarely know f and we must "learn" it based on our data.

Parametric estimation

Methods of parametric estimation go back to Fisher.

If X belongs to a parametric class e.g $X \sim \mathcal{N}(\mu, \sigma^2)$, then it suffices to estimate the corresponding parameters.

But such an assumption may not be valid.

Nonparametric density estimation: Concept

- Let X be a rv, with an unknown density f.
- Target: Estimate $f(x)$, $x \in \mathcal{M}$.

- Assume that f belongs to a large function class $\mathbb{F}$ (continuous, differentiable, Lipschitz cont., Sobolev, Nikol'skii, Besov spaces).

- Let $X_1, \ldots, X_n$, $n \in \mathbb{N}$, be a random sample; independent rv with the same —unknown— f (iid).

Density estimators

- In STAT00 we estimated *parameters*; for example

$$\hat{\mu} = \frac{X_1 + \cdots + X_n}{n} = g(\boldsymbol{X}),$$

where $g : \mathbb{R}^n \to \mathbb{R}$, $g(x_1, \ldots, x_n) = \frac{x_1 + \cdots + x_n}{n}$.

- Set $\boldsymbol{X} = (X_1, \ldots, X_n)$. The joint density

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i), \quad \forall \boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{M}^n. \tag{1}$$

- *Density* Estimator: $\hat{f}_n(x, \boldsymbol{X})$, $x \in \mathcal{M}$, where

$$\hat{f}_n : \mathcal{M} \times \mathcal{M}^n \to \mathbb{R}.$$

- Measure the estimation in both a *stochastic* and a *functional* sense; a risk

$$\mathcal{R}(\hat{f}_n, f) = \mathbb{E}\left(\left\|\hat{f}_n - f\right\|_p\right), \quad \text{for } p \geq 1 \tag{2}$$

measures successfully the loss of such an estimation.

Target:

Construct estimators st the risk over all pdfs lying on a large function space $\mathbb{F}$ to be as small, as possible.

The solution on $\mathbb{R}^d$

(α) Smoothness is needed $\mathbb{F}$: e.g. Sobolev spaces W_p^s.

(β) Kernel Density Estimators $\hat{f}_n^K$

(γ) Giving an upper bound

$$\sup_{f \in \mathbb{F}} \mathcal{R}(\hat{f}_n^K, f) \leq C n^{-r}, \quad r = \frac{s}{2s+d} \tag{3}$$

(δ) And a Lower bound

$$\inf_{\hat{f}_n} \sup_{f \in \mathbb{F}} \mathcal{R}(\hat{f}_n, f) \geq c n^{-r}, \tag{4}$$

i.e. we cannot do better, the above r is the optimal one; *minimax estimation*.

On the rate

The rate for densities on $\mathbb{R}^d$, of smoothness s is n^{-r},

$$r = \frac{s}{2s + d} \tag{5}$$

- It depends both on the smoothness and the dimension.
- The smoother the density, the faster the estimation.
- The higher the dimension, the worst the estimation.

Kernel density estimation

- Murray Rosenblatt, Remarks on some nonparametric estimates of a density function. Ann. Math. Stat. 1956.
- Emanuel Parzen, On estimation of probability density function and mode Ann. Math. Stat. 1962.

- Alexander Tsybakov, Introduction to nonparametric estimation.

Norms in vector spaces

Let V be a vector space. The norm $\|v\|$ of any vector $v \in V$ is a way to count the size of v.

E.g. 1. On $\mathbb{R}^2$: $\|(x, y)\| = \sqrt{x^2 + y^2}$, for every $(x, y) \in \mathbb{R}^2$.
A sequence on $\mathbb{R}^2$:

$$\vec{a}_n = \Big(\frac{1}{n} + 1, \frac{1}{n^2} + 2\Big), \quad n \in \mathbb{N}.$$

We say that $\vec{a}_n \to (1, 2) \in \mathbb{R}^2$, because

$$\|\vec{a}_n - (1, 2)\| = \Big\|\Big(\frac{1}{n}, \frac{1}{n^2}\Big)\Big\| = \sqrt{\frac{1}{n^2} + \frac{1}{n^4}} \to 0.$$

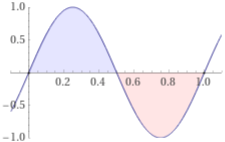
Norms for functions

- We need to count distances $d(f,g) = \|f-g\|$, between functions.
- Let $p \geq 1$ and $g : \mathcal{M} \to \mathbb{R}$, then $g \in L^p(\mathcal{M})$ (Lebesgue) if-f

$$\|g\|_p := \left(\int_{\mathcal{M}} |g(x)|^p dx \right)^{1/p} < \infty. \tag{6}$$

- When $p = 1$ and $\mathcal{M} = [0,1]$, then

$$\|g\|_1 = \int_0^1 |g(x)| dx = \text{Area plot } x\text{-axis}.$$



Remarks

- Clearly when f is a PDF; $\|f\|_1 = 1$.
- L^∞ is the space containing all the (essentially) bounded functions.
- Interpolation property of Lebesgue spaces: Let $1 \leq p_1 < p_2 \leq \infty$. Then

$$f \in L^{p_1} \cap L^{p_2} \Rightarrow f \in L^p, \quad \text{for every } p_1 < p < p_2. \tag{7}$$

- E.g. If f is a bounded PDF, then $f \in L^p$, for every $p \in [1, \infty]$.

Risk

Let f be unknown and $\hat{f}_n$ be an estimator of it. We define the L^p-risk as

$$\begin{aligned}\mathcal{R}(\hat{f}_n, f) &:= \left(\mathbb{E}(\|\hat{f}_n - f\|_p^p)\right)^{1/p} \qquad (8)\\ &= \left(\int_{\mathcal{M}} \cdots \int_{\mathcal{M}} \left(\int_{\mathcal{M}} |\hat{f}_n(x, \boldsymbol{x}) - f(x)|^p dx\right) \prod_{i=1}^{n} f(x_i) dx_i\right)^{\frac{1}{p}}.\end{aligned}$$

KDEs Parzen Annals Math Stat (62)

- X is distributed on $\mathcal{M} = \mathbb{R}$.
- CDF: $F(x) = \int_{-\infty}^{x} f(t)dt = \mathbb{P}(X \leq x)$.
- Let $x \in \mathbb{R}$. We define the empirical estimator of the CDF

$$\begin{aligned} \widehat{F}_n(x) &= \frac{\#\{X_i : X_i \leq x\}}{\#\{X_i\}} \\ &= \frac{1}{n}\sum_{i=1}^{n} I(\{X_i \leq x\}) \longrightarrow F(x) = \mathbb{P}(X \leq x), \quad n \to \infty \end{aligned}$$

by SLLN and where I the indicator function.

- On the other hand

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}. \tag{9}$$

Rosenblatt's kernel

Combining the above we set:

$$\hat{f}_n^R(x) := \frac{1}{2nh}\sum_{i=1}^{n} I\big(x-h < X_i \le x+h\big)$$
$$=: \frac{1}{nh}\sum_{i=1}^{n} K_0\left(\frac{X_i - x}{h}\right), \tag{10}$$

where $K_0(u) := \frac{1}{2}I(-1 < u \le 1)$ Rosenblatt's kernel (rectangular kernel). The positive number h is called bandwidth (and it is supposed to be arbitrary small).

- Triangular, Gaussian kernels etc.
- $n \to \infty$ and $h = h_n \to 0$, but we already observe that this has to be done carefully.

Kernel Density Estimators (KDEs)

- More generally $K : \mathbb{R} \to \mathbb{R}$ kernel;

$$\int_{\mathbb{R}} K(z)dz = 1. \tag{11}$$

Kernel density estimator (KDE)

$$\hat{f}_n(x) := \hat{f}_n^K(x; \boldsymbol{X}) := \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right). \tag{12}$$

- $h = h_n \to 0$, when $n \to \infty$; the *bandwidth*.

Agenda

Target: Well-estimate f by $\hat{f}_n$.

(α) Fix assumptions on the densities' class (determine $\mathbb{F}$).
(β) Construct some kernels K and therefore estimators $\hat{f}_n^*$
(γ) Giving a rate of convergence

$$\sup_{f \in \mathbb{F}} \mathcal{R}(\hat{f}_n^*, f) \leq Cn^{-r}, \tag{13}$$

with $r > 0$, to be determined.
(δ) Prove that

$$\inf_{\hat{f}_n} \sup_{f \in \mathbb{F}} \mathcal{R}(\hat{f}_n, f) \geq cn^{-r}, \tag{14}$$

i.e. we cannot do better, the above r is the optimal one; *minimax estimation*.

Approaching the problem: an elementary form

- Specify $p = 2$, refer to the corresponding risk as Mean Integrated Squared Error. $\text{MISE}(\hat{f}_n, f) := \mathcal{R}(\hat{f}_n, f)^2$.
- Target:

$$\text{MISE}(\hat{f}_n, f)^{1/2} \leq cn^{-r}, \quad \text{for some } r > 0. \tag{15}$$

- Decomposition

$$\text{MISE}(\hat{f}_n, f) = \|b\|_2^2 + \|\sigma^2\|_1, \tag{16}$$

where the two terms above are the bias and variance:

$$b(x) := \mathbb{E}\left[\hat{f}_n(x; \boldsymbol{X})\right] - f(x) \tag{17}$$

and

$$\sigma^2(x) := \mathbb{E}\left[\left(\hat{f}_n(x; \boldsymbol{X}) - \mathbb{E}\left[\hat{f}_n(x; \boldsymbol{X})\right]\right)^2\right]. \tag{18}$$

We study these two terms independently.

Bounding variance

Proposition

If $K \in L^2$, then for every pdf f

$$\|\sigma^2\|_1 \leq \frac{\|K\|_2^2}{nh}, \tag{19}$$

Remarks:
(α) No assumptions on f; just a pdf!
(β) $h = h_n \to 0$, but carefully!
E.g. for $h \sim n^{-(1-r)}$, $0 < r < 1$, we derive

$$\|\sigma^2\|_1 \leq cn^{-r}.$$

(γ) Proof: just Fubini-Tonelli Theorem and iid.

Bounding bias: Regularity assumption in f

- Since $\int K = 1$, we simplify

$$b(x) = \int_{\mathbb{R}} K\left(\frac{y-x}{h}\right)(f(y) - f(x))\frac{dy}{h} \qquad (20)$$
$$= \int_{\mathbb{R}} K(z)(f(x+hz) - f(x))dz.$$

- Where is the n? Inside h.
- Here we need some regularity for the pdf to be assumed:

Taylor

We assume that $f \in \mathcal{C}^s$, for some $s \in \mathbb{N}$. Then Taylor's formula asserts

$$f(x+hz) - f(x) = \sum_{\nu=1}^{s-1} \frac{f^{(\nu)}(x)}{\nu!}(hz)^{\nu} + R_s(f), \tag{21}$$

where

$$R_s(f) := \int_0^1 (hz)^s \frac{(1-t)^{s-1}}{(s-1)!} f^{(s)}(x+thz)dt \tag{22}$$

Bias

$$b(x) = \sum_{\nu=1}^{s-1} \frac{f^{(\nu)}(x)}{\nu!} h^{\nu} \int_{\mathbb{R}} z^{\nu} K(z) dz + h^s \int_{\mathbb{R}} \int_0^1 \frac{(1-t)^{s-1}}{(s-1)!} f^{(s)}(x+thz) z^s K(z) dt dz \tag{23}$$

Bounding bias: Moments for the kernel

- The decay is aligned in the powers of h.
- We put the following assumption on the kernel:

Zero moments up to the order $s-1$:

$$\int_{\mathbb{R}} z^{\nu} K(z)dz = 0, \quad \text{for every } \nu = 1, \dots, s-1. \tag{24}$$

Under these vanishing moments and (23):

$$b(x) = h^s \int_{\mathbb{R}} K(z) z^s \int_0^1 \frac{(1-t)^{s-1}}{(s-1)!} f^{(s)}(x + thz) dt dz. \tag{25}$$

Smoothness spaces naturally pop up

- Recall that we're estimating the square norm of $b(x)$.
- Choosing my kernel such that

$$[K]_s := \int_{\mathbb{R}} |z^s||K(z)|dz < \infty, \tag{26}$$

Minkowski's inequality for integrals
($\| \int g(\cdot, z)dz\|_p \leq \int \|g(\cdot, z)\|_p dz$) implies:

$$\|b\|_2 \leq \frac{[K]_s}{s!} \left\| f^{(s)} \right\|_2 h^s. \tag{27}$$

Sobolev spaces are present.

Sobolev spaces

Measure the smoothness and integrability of a function. Let $s \in \mathbb{N}$ and $p \geq 1$, the Sobolev space W_p^s, consists of all the functions such that

$$f \in W_p^s \Longleftrightarrow \|f\|_{W_p^s} := \sum_{\nu=0}^{s} \left\| f^{(\nu)} \right\|_p < \infty. \tag{28}$$

Of course $L^p = W_p^0 \supset W_p^1 \supset W_p^2 \supset \cdots$.
Let also $m > 0$. We denote by $W_p^s(m) := \{f : \|f\|_{W_p^s} \leq m\}$.

Proposition

Let $s \in \mathbb{N}$, $m > 0$ and K satisfying (11), (24) and (26), then there exists a constant $c = c(K, s, m) > 0$:

$$\|b\|_2 \leq ch^s, \quad \text{for every} \quad f \in W_p^s(m). \tag{29}$$

Kernels' choice

- A kernel $K : \mathbb{R} \to \mathbb{R}$ as above will be called a kernel of order s; $\mathcal{K}(s)$.
- Of course $\mathcal{K}(S) \subset \mathcal{K}(s)$, for $S > s$.

- Yes, there exist such kernels. A classical construction involves Legendre polynomials. Another option is by using the properties of the Fourier transform. Plenty of examples appropriate for applications.

Bandwidth selection

By all the previous steps:

$$\sup_{f \in W_2^s(m)} \mathrm{MISE}(\hat{f}_n, f) \leq \frac{\|K\|_2^2}{nh} + \left(\frac{[K]_s m}{s!}\right)^2 h^{2s}. \tag{30}$$

- We choose the bandwidth $h = h_n$ st the right hand side to be *minimized*.

Bretagnolle and Huber ('79) and Haminskii, Ibragimov ('80)

Theorem

Let $s \in \mathbb{N}$, $p \geq 2$ and $m > 0$. Then the KDE $\hat{f}_n$ associated with a kernel K of order s and $h \sim n^{-1/(2s+1)}$ satisfies:

$$\sup_{f \in W_p^s(m)} \mathcal{R}(\hat{f}_n, f) \leq c n^{-s/(2s+1)}. \quad (31)$$

Moreover the estimation is minimax.

Perspective

• Given that I don't know of course the smoothness level of f, what can I do?
• Use a kernel of order s, according to how fast is your PC and how much counts for you the accuracy.

• What type of a kernel could work for any smoothness level, optimal?
• Littlewood-Paley/ bump: Infinitely differentiable, Compactly supported, unit around the origin.

• I neither know the p for f...
• Adaptive estimation. Disconnect the integrability levels between the risk and the densities. Different rates. Wavelet estimators.

• Can we do anything with the dimension?
• The question finds answer in the geometry of the data's domain. Density estimation on spheres or manifolds.

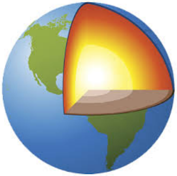
Progress in the area: key developments on $\mathbb{R}^d$

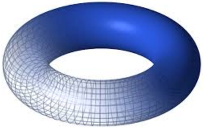
- Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., Picard, D., Density estimation by wavelet thresholding. *Ann. Stat.* 24, 508-539 (1996).
- Efroimovich, S.Yu., Non-parametric estimation of the density with unknown smoothness. *Ann. Stat.* 36, 1127-1155 (1986).
- Kerkyacharian, G., Lepski, O., Picard, D., Nonlinear estimation in anisotropic multiindex denoising Sparse case. *Theory Probab. Appl.* 52, 58–77 (2008)
- Goldenshluger, A., Lepski, O., Uniform bounds for norms of sums of independent random functions. *Ann. Probab.* 39, 2318-2384 (2011).
- Goldenshluger, A., Lepski, O., Bandwidth selection in kerrnel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Stat.* 39, 1608-1632 (2011).
- Goldenshluger A., Lepski O., Minimax estimation of norms of a probability density: I. Lower bounds -2020-
- Goldenshluger A., Lepski O., Minimax estimation of norms of a probability density: II. Rate-optimal estimation procedures -2020-

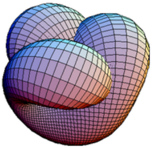
Motivation

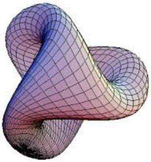
Geostatistics, Climatology, Environmental studies, Astrophysics, Oceanography, Seismology...

Manifolds



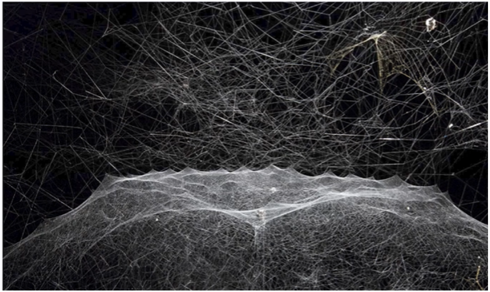




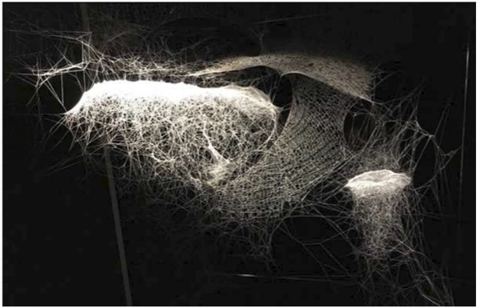


$$\mathbb{T}^m = \{x \in \mathbb{R}^m : \ x_i > 0, \ x_1 + \cdots + x_m < 1\}. \tag{32}$$

SPIDER WEB FROM EXHIBITION ON AIR TOMAS SARACENO



SPIDER WEB



The problem on the sphere.

- The problem has been solved by *needlet estimators* and has been used in applications in astrophysics.

- P. Baldi, G. Kerkyacharian, D. Marinucci, D. Picard, Adaptive density estimation for directional data using needlets. *Ann. Statist.* 37 (2009), no. 6A, 3362-3395.
- Subsampling needlet coefficients on the sphere. *Bernoulli* 15 (2009), no. 2, 438-463.

Preparation

- All the necessary *analysis' background* needed to be built for the *specific* manifold.

- F. Narcowich, P. Petrushev, J. D. Ward, Localized tight frames on spheres. *SIAM J. Math. Anal.* 38 (2006), no. 2, 574-594.
- F. Narcowich, P. Petrushev, J. Ward, Decomposition of Besov and Triebel-Lizorkin spaces on the sphere. *J. Funct. Anal.* 238 (2006), no. 2, 530-564.

Difficulties

- For studying density estimation on a new manifold we need:

(α) Well defined notion of regularity.
(β) Smoothness spaces.
(γ) Operational tools from Analysis and Geometry.

(δ) Kernels and/or wavelets or a substitute.
(ε) Extraction of the proper Statistical theorems with precise density estimators.

- When the data are located on another manifold, we have to re-face all these...

Challenge

Work on a general framework unifying as many examples as possible!

Develop the necessary background. Prove the proper statistical results and construct tools for immediate practical use in the most common examples.

- G. Kerkyacharian, P. Petrushev, Heat kernel based decomposition of spaces of distributions in the framework of Dirichlet spaces. Trans. Amer. Math. Soc. 367 (2015), 121–189.

The setting (Roughly speaking)

1. Let $(\mathcal{M}, \rho, \mu)$ a metric measure space:

$$\mu(B(x,r)) \sim r^d, \quad 0 < d := \text{homogeneous dimension}, \quad (33)$$

uniformly in $x \in \mathcal{M}$, $r > 0$ and $B(x,r) = \{y : \rho(x,y) < r\}$.

2. A suitable operator L determines the notion of smoothness and smoothness spaces.

The setting unifies the Euclidean space, the sphere, the ball, general Riemannian manifolds, spaces of matrices, and more.

Euclidean space

$\mathcal{M} = \mathbb{R}^d$ and

$$Lf = -\Delta f = -\left(\partial_1^2 + \cdots + \partial_d^2\right) f. \tag{34}$$

Sphere

$$\mathcal{M} = \mathbb{S}^d = \{x \in \mathbb{R}^{d+1} : \ \|x\| = 1\},$$

$$\rho(x, y) = \arccos\left(\langle x, y\rangle\right), \tag{35}$$

μ : the spherical measure
and L : the spherical Laplacian.

Ball

$$\mathcal{M} = \mathbb{B}^d = \left\{x \in \mathbb{R}^d : \ \|x\| < 1\right\},$$

$$\rho(x, y) = \arccos\left(\langle x, y\rangle + \sqrt{1 - \|x\|^2}\sqrt{1 - \|y\|^2}\right), \tag{36}$$

$$d\mu(x) = \left(1 - \|x\|^2\right)^{-1/2} dx \tag{37}$$

and

$$L = -\sum_{i=1}^{d}\left(1 - x_i^2\right)\partial_i^2 + 2\sum_{1\leq i<j\leq d} x_i x_j \partial_i \partial_j + d\sum_{i=1}^{d} x_i \partial_i. \tag{38}$$

Contributions

1. Start the research of Statistics on a general uniform framework.
2. Prepare objects ready for use in applications.
3. Kernel and Wavelet density estimators.
4. Adaptive upper bounds.
5. Optimal rate when restrict on the known examples.
6. Oracle inequalities.
7. General kernels (and simple in the computational sense).
8. Expression of the KDEs on several specific examples of common interest.
9. Lower bound: minimax density estimation.
10. Open problems.

Statistics and Probability on the general framework

1. I. Castillo, G. Kerkyacharian, D. Picard, Thomas Bayes' walk on manifolds. *Probab. Theory Related Fields* 158 (2014), no. 3-4, 665-710.
2. G. Kerkyacharian, S. Ogawa, S., P. Petrushev, D. Picard, Regularity of Gaussian processes on Dirichlet spaces. *Constr. Approx.* 47 (2018), no. 2, 277-320.
3. G. Cleanthous, g., G. Kerkyacharian, P. Petrushev, D. Picard, Kernel and wavelet density estimators on manifolds or more general metric spaces. *Bernoulli*. 26, No. 3, 1832-1862 (2020).
4. G. Cleanthous, g., E. Porcu, Oracle inequalities and upper bounds for kernel density estimators on manifolds or more general metric spaces. Submitted.

Kernel density estimators

- Let $K : [0, \infty) \to \mathbb{R}$ with rapid decay (symbol).
- Let $h > 0$ a microscopic quantity called “bandwidth". We denote by K_h the function $K_h(\lambda) = K(h\lambda)$, $\lambda \geq 0$. (dilation).
- Spectral theory gives rise to a function

$$K_h^L(x, y) \quad \text{for every} \ \ (x, y) \in \mathcal{M} \times \mathcal{M} \ \ \text{(kernel)}, \tag{39}$$

with convenient properties.

- We find the proper conditions on the symbols K and the KDE has an abstract form

$$\hat{f}_{n,h}(x) := \frac{1}{n} \sum_{i=1}^{n} K_h^L(x, X_i). \tag{40}$$

- Construct the above kernels in the examples of special interest.

Kernel density estimators on core examples

- $\mathcal{M} = \mathbb{R}^d$,

$$\hat{f}_{n,h}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K\Big(\frac{X_i - x}{h}\Big). \tag{41}$$

- $\mathcal{M} = \mathbb{S}^2$,

$$\hat{f}_{n,h}(x) = \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=0}^{\infty} \frac{2\ell + 1}{|\mathbb{S}^2|} K\Big(h\sqrt{\ell(\ell+1)}\Big) C_\ell^{1/2}(\langle X_i, x\rangle). \tag{42}$$

Thank you :)

Thank you very much for your attention!!!