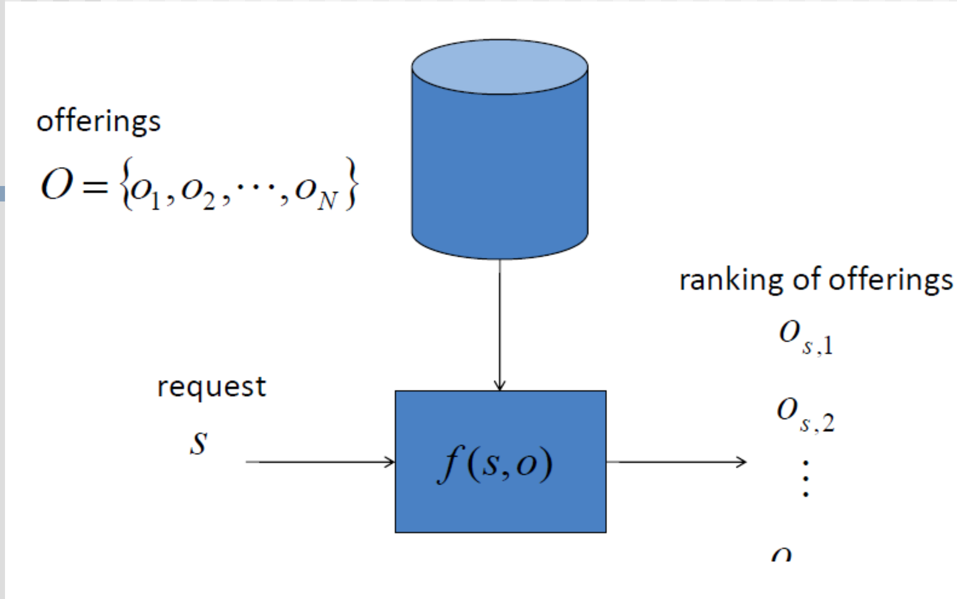
II The international scientific and practical forum "Science and Business"
Dnepr, Ukraine, 1 July, 2016

ADAPTIVE MODELS OF WEB DOCUMENTS RANKING DEVELOPMENT BASED ON DYNAMIC CLUSTERING

National Technical University
“Kharkiv Polytechnic Institute”

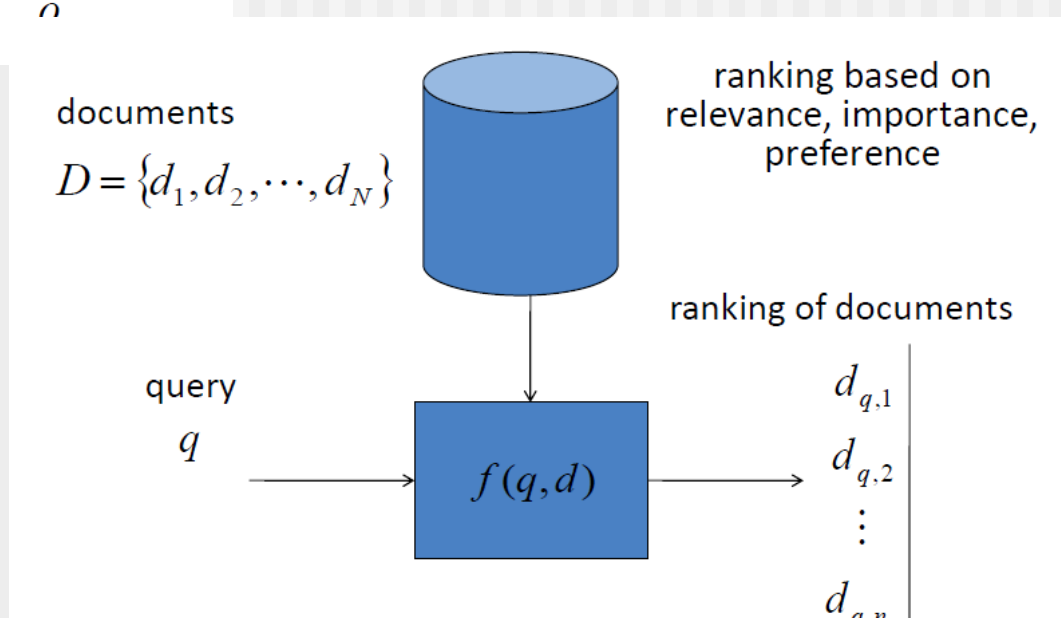
Prof. L. M. Lyubchyk

Ranking Problem

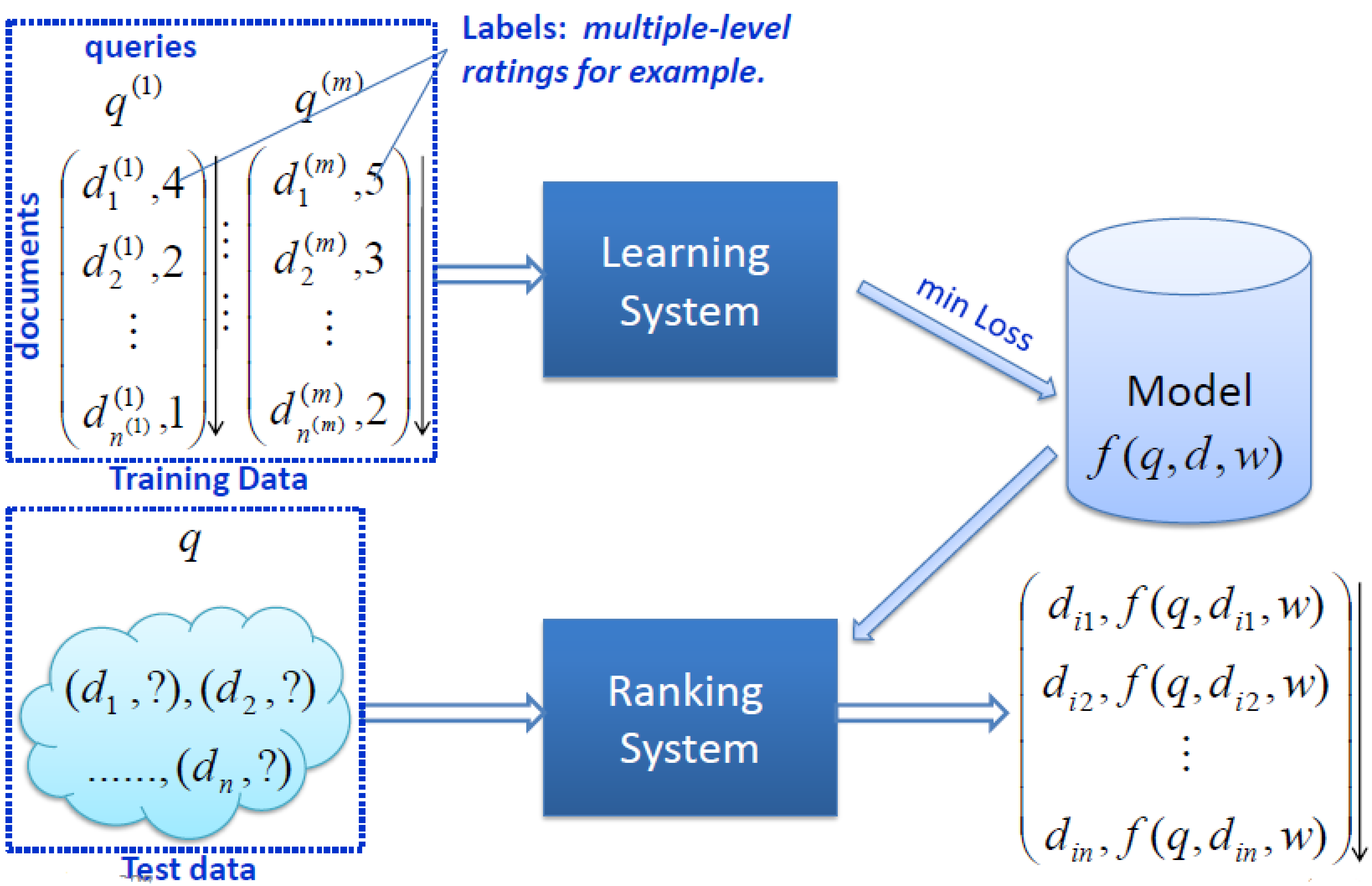


- Machine learning technologies
- For information presentation
- Learning ranking function $f(s,o)$
- Using labeled data (supervised learning)
- Ranking function is feature based

- Data Labeling
- Feature Extraction
- Evaluation Measure
- Learning Method Algorithm



Framework of Learning to Rank for Information Retrieval



Pointwise Approach

Transforming ranking to regression, classification, or ordinal regression

	Learning		
	Regression	**Classification**	**Ordinal Regression**
Input	Feature vector x		
Output	Real number $y = f(x)$	Category $y = \text{sign}(f(x))$	Ordered category $y = \text{thresold}(f(x))$
Model	Ranking function $f(x)$		
Loss Function	Regression loss	Classification loss	Ordinal regression loss

Pairwise Approach

Transforming ranking to pairwise classification

	Learning	Ranking
Input	Ordered feature vector pair (x_i, x_j)	Feature vectors $\mathbf{x} = \{x_i\}_{i=1}^{n}$
Output	Classification on order of vector pair $y_{i,j} = \text{sign}(f(x_i - x_j))$	Permutation on vectors $\boldsymbol{\pi} = \text{sort}(\{f(x_i)\}_{i=1}^{n})$
Model	Ranking function $f(x)$	
Loss Function	Pairwise classification loss	Ranking evaluation measure

Listwise Approach

List as instance, query-document group structure is used

	Learning & Ranking
Input	Feature vectors $\mathbf{x} = \{x_i\}_{i=1}^{n}$
Output	Permutation on feature vectors $\boldsymbol{\pi} = \mathrm{sort}(\{f(x_i)\}_{i=1}^{n})$
Model	Ranking function $f(x)$
Loss Function	Listwise loss function (ranking evaluation measure)

PROBLEM STATEMENT

Consider a set of objects $x \in X$, with feature vector $x^{\mathrm{T}} = (x^1, x^2, \ldots, x^N)$.

Training data consists from feature vectors $\{x_i\}_{i=1}^{n}$, $x_i^{\mathrm{T}} = (x_i^1, x_i^2, \ldots, x_i^N)$,

a) Dataset as a data matrix $\mathbf{X}_n = \{x_i^j\}_{i,j=1}^{n,N}$.

b) Observation sequence data stream $\mathbf{x}_{n,s(n)} = \{x_n, x_{n-1}, \ldots, x_{n-s(n)}\}$,

Relative objects rank $r_{n,s(n)}(x_i)$, $n - s(n) \le i \le n$ defined by a ranking system.

Ranking mechanism is unknown, and described by a preference function

$$f(x), \text{ such that } f(x_i) > f(x_j) \text{ if } r_{n,s(n)}(x_i) < r_{n,s(n)}(x_j).$$

Preference learning to rank is to recover preferences function $\hat{f}(x)$ regarding the observations

$$(\mathbf{x}_{n,s(n)}, \mathbf{r}_{n,s(n)}(x_i)),\ i = \overline{n - s(n), n},\ \mathbf{r}_{n,s(n)}(x_i) = (r_{n,s(n)}(x_{n-s(n)}), \ldots, r_{n,s(n)}(x_n)).$$

Dataset is divided into a set of M clusters $\{X_m\}_{m=1}^{M}$ in feature space with parameters $(\bar{x}^m, \bar{r}^m)$, $m = \overline{1,M}$, $\bar{x}^m$ – center vector of m- th cluster,

$\bar{r}^m$ – average rank of objects of the same cluster

$$\bar{x}^m = \frac{1}{J^m} \sum_{i \in I_m} x^i, \quad \bar{r}^m = \frac{1}{|J^m|} \sum_{i \in I_m} r(x^i), \quad J^m = \{i \,|\, x_i \in \mathrm{X}_m\}. \qquad (1)$$

Preference function model in the quasilinear identification form $f(x) = \varphi^{\mathrm{T}}(x)c$,

$c^{\mathrm{T}} = (c_1, \ldots, c_L)$ – vector of unknown model parameters,

$\varphi^{\mathrm{T}}(x) = (\varphi_1(x), \ldots, \varphi_L(x))$ – vector of coordinate functions, L- model dimension.

Preference learning to rank on clusters is optimization problem for empirical risk

$$R(c) = \sum_{m=1}^{M} (\bar{r}^m - \varphi^{\mathrm{T}}(\bar{x}^m)c)^2 + \Omega(c, w_0) \to \min_{c} \qquad (2)$$

Coordinate functions are taken hereby that its scalar products in will be positive definite kernel functions $\varphi^{T}(x_i)\varphi(x_j)=\kappa(x_i,x_j),\ i,j=\overline{1,M}$, Gaussian kernel $\kappa(x,x')=\exp\{-\mu(x-x')^2\}$.

Preference model as a linear combination of kernel function, located in the centers of the clusters

$$\hat{f}(x)=\sum_{m=1}^{M} d_m\cdot\kappa(x,\bar{x}^m), \qquad (3)$$

$d_m,\ m=\overline{1,M}$ – kernel-based preference model parameters determined by average rank of cluster .

Depending on the type of training are the following options preferences learning problem statement:

- *Non recurrent preference learning* using a dataset of fixed volume $\mathbf{X}_n=\{x_i^j\}_{i,j=1}^{n,N},\ \{r(x_i)\}_{i=1}^{n}$.
- *Recurrent preference learning* by data stream $\{(x_n,r_n)\},\ n=1,2,...$ using to update the model.

A. *Preference Function Kernel-Based Recovery*

In the kernel framework the measurement equation for preference function identification

$$\bar{r}_n^m = \hat{f}(\bar{x}_n^m) = \varphi^{\mathrm{T}}(\bar{x}_n^m)c + e_m, \quad m = \overline{0, M}, \quad \bar{r}_n = \mathbf{\Phi}_n^{\mathrm{T}} c + e_n, \qquad (4)$$

$\bar{r}_n = (\bar{r}_n^1 \;\; \bar{r}_n^2 \; \ldots \; \bar{r}_n^M)^{\mathrm{T}}$ – observation vector composed from average rank estimates,

$\mathbf{\Phi}_n = \left(\varphi(\bar{x}_n^1) \quad \varphi(\bar{x}_n^2) \quad \ldots \quad \varphi(\bar{x}_n^M)\right)$ – feature matrix from dataset $\mathbf{X}_n = \{x_i^j\}_{i,j=1}^{n,N}, \; \{r(x_i)\}_{i=1}^n,$

$e_n = (e_n^1 \; e_n^2 \ldots e_n^M)^{\mathrm{T}}$ – vector of average rank estimation errors.

Kernel matrix of ($M \times M$) dimension $\mathbf{K}_n = \mathbf{\Phi}_n^{\mathrm{T}} \mathbf{\Phi}_n, \; \mathbf{K}_n = \left\| k_{q,s} \right\|, \; k_{q,s} = \kappa(\bar{x}_n^q, \bar{x}_n^s), \; q, s = \overline{1, M}.$

Empirical risk optimization problem $R(c) = \frac{1}{2}\gamma \left\| e_n \right\|^2 + \frac{1}{2}\left\| (c - c_0) \right\|^2, \quad e_n = \bar{r}_n - \mathbf{\Phi}_n^{\mathrm{T}} c, \qquad (5)$

c_0 – vector of *a priory* values of preference model parameters, γ – regularization parameter.

An equivalent conjugate optimization problem stated using Lagrange function

$$L(c, e_n, \lambda) = R(c) + \lambda^{\mathrm{T}} (\bar{r}_n - \mathbf{\Phi}_n^{\mathrm{T}} c - e_n), \quad (6)$$

$\lambda^{\mathrm{T}} = (\lambda_1, \ldots, \lambda_M)$ – vector of Lagrange multipliers (conjugate variables).

Conditions optimality $c - c_0 - \mathbf{\Phi}_n \lambda = 0, \;\; \gamma e_n = \lambda, \;\; \bar{r}_n - \mathbf{\Phi}_n^{\mathrm{T}} c = e_n,$ (7)

Model parameters and conjugate variables optimal estimations

$$\begin{aligned} \hat{\lambda}_n &= \mathbf{A}_n^{-1}(\gamma)(\bar{\mathbf{r}}_n - \mathbf{\Phi}_n^{\mathrm{T}} c_0), \\ \hat{c}_n &= c_0 + \mathbf{\Phi}_n \mathbf{A}_n^{-1}(\gamma)(\bar{\mathbf{r}}_n - \mathbf{\Phi}_n^{\mathrm{T}} c_0), \\ \mathbf{A}_n(\gamma) &= \gamma^{-1} \mathbf{I}_n + \mathbf{K}_n. \end{aligned} \quad (8)$$

The preference function model a parameters estimate (7) depends of it's *a priory* value c_0.

B. Regularisation via expert knowledge integration

Preference function as a linear combination of features $f^0(x) = x^{\mathrm{T}} w_0$ with expert weights w_0.

Optimal *a priory* value of preference function model parameters c_0 from the condition of best approximation of average rank vector $\bar{r}_0 = \mathbf{\Phi}_n^{\mathrm{T}} c_0$ by linear preference model $\tilde{r}_0 = \mathbf{X}_n w_0$.

A priory value of preference function parameters c_0, which are optimally concordant with expert estimations of feature weights w_0, is found from the optimization problem

$$Q_0(c_0) = \frac{1}{2}\omega \cdot \|\zeta\|^2 + \frac{1}{2}\|c_0\|^2, \quad \zeta = \tilde{r}_0 - \bar{r}_0 = \mathbf{X}_n w_0 - \mathbf{\Phi}_0^{\mathrm{T}} c_0. \tag{9}$$

Using the appropriate Lagrange function with Lagrange multipliers $\nu^{\mathrm{T}} = (\nu_1, \ldots, \nu_n)$

$$L(c_0, \zeta, \nu) = Q_0(c_0) + \nu^{\mathrm{T}}(\mathbf{X}_n w_0 - \mathbf{\Phi}_0^{\mathrm{T}} c_0 - \zeta). \tag{10}$$

Optimal *a priory* value of preference function model parameters c_0

$$c_0^* = \mathbf{\Phi}_n \mathbf{B}_n^{-1}(\omega)\mathbf{X}_n w_0, \ \mathbf{B}_n(\omega) = \omega^{-1}\mathbf{I}_n + \mathbf{K}_n. \quad (11)$$

Optimal preference function estimate $\hat{f}_n(x) = \varphi^{\mathrm{T}}(x) c_n^*$, regularized by features weights

$$\begin{aligned}
&\hat{f}_n(x) = \chi_n^{\mathrm{T}}(x) \cdot \hat{d}_n, \\
&\chi_n^{\mathrm{T}}(x) = \varphi^{\mathrm{T}}(x)\mathbf{\Phi}_n = \left(\kappa(x, \bar{x}_n^1), \kappa(x, \bar{x}_n^2), \ldots, \kappa(x, \bar{x}_n^M)\right)^{\mathrm{T}} \\
&\hat{d}_n = \mathbf{A}_n^{-1}(\gamma)\bar{r}_n + \mathbf{D}_n(\omega)\mathbf{X}_n w_0, \\
&\mathbf{D}_n(\omega) = [\mathbf{I}_n - \mathbf{A}_n^{-1}(\gamma)\mathbf{K}_n]\mathbf{B}_n^{-1}(\omega).
\end{aligned} \quad (13)$$

coefficients $\hat{d}_n^m = d^m(\bar{x}_n, \bar{r}_n), \ m = \overline{1, M}$ of linear kernel function combination (3) depends

from available dataset $\mathbf{X}_n = \{x_i^j\}_{i,j=1}^{n,N}, \ \{r(x_i)\}_{i=1}^{n}$ and expert feature weights w_0.

C. *Clusters Psrameters Updating*

Updating the cluster's $\{X_m\}_{m=1}^{M}$ parameters $(\bar{x}_n^m, \bar{r}_n^m)$, $m=\overline{1,M}$ using data stream $\{(x_n, r_n)\}$, $n=1,2,\ldots$

The next element of data stream x_{n+1} included in the cluster $l = \underset{m}{\operatorname{argmin}} \left\| x_{n+1} - \bar{x}^m \right\|^2$.

Estimates of corresponding cluster parameters and the average rank of its elements is updating

$$\bar{x}_{n+1}^l = \frac{1}{J_{n+1}^l} \sum_{i \in J_{n+1}^l} x_i = \frac{\left|J_n^l\right|}{\left|J_n^l\right| + 1} \bar{x}_n^l + \frac{1}{\left|J_n^l\right| + 1} x_{n+1}, \tag{14a}$$

$$\bar{x}_{n+1}^m = \bar{x}_n^m, \quad m \neq l,$$

$$\bar{r}_{n+1}^l = \frac{1}{\left|J_{n+1}^l\right|} \sum_{i \in J_{n+1}^l} r(x_i) = \frac{\left|J_n^l\right|}{\left|J_n^l\right| + 1} \bar{r}_{n+1}^l + \frac{1}{\left|J_n^l\right| + 1} r(x_{n+1}),$$

$$\bar{r}_{n+1}^m = \bar{r}_n^m, \quad m \neq l, \tag{14b}$$

D. Recursive Preference Learning

Updated clusters parameters $(\bar{x}^m_{n+1}, \bar{r}^m_{n+1}), m = \overline{1, M}$ gives preference function estimate

$$\hat{f}_{n+1}(x) = \chi^{\mathrm{T}}_{n+1}(x) \cdot \hat{d}_{n+1},$$
$$\hat{d}_{n+1} = \mathbf{A}^{-1}_{n+1}(\gamma)\bar{r}_{n+1} + \mathbf{D}_{n+1}(\omega)\mathbf{X}_{n+1}w_0. \quad (15)$$

Recurrent estimation of function preferences model parameters

$$\hat{c}_{n+1} = \hat{c}_n + \mathbf{\Phi}_{n+1}\mathbf{A}^{-1}_{n+1}(\gamma)(\bar{\mathbf{r}}_{n+1} - \mathbf{\Phi}^{\mathrm{T}}_{n+1}\hat{c}_n). \quad (16)$$

Nonparametric recurrent equation for preference function estimate by data stream

$$\hat{f}_{n+1}(x) = \hat{f}_n(x) + \chi^{\mathrm{T}}_{n+1}(x)\mathbf{A}^{-1}_{n+1}(\gamma)(\bar{r}_{n+1} - \hat{r}_n(\bar{x}_{n+1})),$$
$$\hat{r}_n(\bar{x}_{n+1}) = (\hat{f}_n(\bar{x}^1_{n+1}) \ \hat{f}_n(\bar{x}^2_{n+1}) \ \ldots \ \hat{f}_n(\bar{x}^M_{n+1}))^{\mathrm{T}}, \quad (17)$$
$$\hat{f}_n(x) = \varphi^{\mathrm{T}}(x)\hat{c}_n$$

Equivalent "measurement equation" for conjugate variables at instant $n+1$

$$\bar{r}_{n+1} = \hat{r}_n(\bar{x}_{n+1}) + \mathbf{K}_{n+1}\lambda_{n+1} + e_{n+1}. \quad (18)$$

Local identification criterion at instant $n+1$ as a moving estimation cost function

$$R_{n+1}(\lambda_{N+1}) = \left\| \bar{r}_{n+1} - \hat{r}_n(\bar{x}_{n+1}) - \mathbf{K}_{n+1}\lambda_{n+1} \right\|^2 + \\ + \gamma^{-1}(\lambda - \hat{\lambda}_n)^{\mathrm{T}} \mathbf{K}_{n+1}(\lambda - \hat{\lambda}_n) \to \min_{\boldsymbol{\lambda}_{n+1}}. \quad (19)$$

Recurrent estimation algorithm for conjugate variables

$$\hat{\lambda}_{n+1} = \mathbf{A}_{n+1}^{-1}(\gamma)\left(\gamma^{-1}\hat{\lambda}_n + \bar{r}_{n+1} - \hat{r}_n(\bar{x}_{n+1})\right) \quad (20)$$

Recursive nonparametric estimation of preference function

$$\hat{f}_{n+1}(x) = \hat{f}_n(x) + \chi_{n+1}^{\mathrm{T}}(x)\hat{\lambda}_{n+1}. \quad (21)$$

THANK YOU FOR YOUR ATTENTION