From Empirical-Probabilistic to Entropy-Randomized Machine Learning

Yuri S. Popkov

Institute for Systems Analysis
Federal Research Center “Computer Science and Control”
Russian Academy of Sciences
44-2 Vavilova str., 119333 Moscow, Russia,
popkov@isa.ru

Abstract

New problems of machine learning theory named randomized machine learning are considered. They are based on the entropy maximization methods, that give the best solutions under maximum uncertainty. In respect to parameterized model we obtain entropy optimized probability density functions of parameters. In machine learning procedures the randomized model is a generator of stochastic ensemble of possible solutions. The problems of classification and dynamic regression are considered.

1 Introduction

Training a computer program to solve problems which are difficult or uninteresting for the human becomes a popular area of scientific and applied research [Bishop, 2006, Vorontsov, 2006]. There exists a number of problems in this area which differ in their mathematical models and solving methods. But, despite their difference, they have common mathematical basis — optimization and statistics [Merkov, 2014]. Its significant part is classical concepts of these disciplines. Problems which are often arise in *Machine Learning* lead to the need of new approaches, methods and information technologies. A fundamental feature of machine learning is the uncertainty of the environment in which corresponding procedures are implemented. In order to somehow estimate the uncertainty different models are used. The most common model is the stochastic one, which allows to give results of machine learning in probabilistic meaning based on data. The concept of *Empirical-probabilistic machine learning (EP-ML)* gives the answer with empirically

In: Yu. G. Evtushenko, M. Yu. Khachay, O. V. Khamisov, Yu. A. Kochetov, V.U. Malkova, M.A. Posypkin (eds.): Proceedings of the OPTIMA-2017 Conference, Petrovac, Montenegro, 02-Oct-2017, published at http://ceur-ws.org

computed probability, which is computed during the learning process [Hastie et al., 2001]. This result can be achieved by setting of prior probabilistic characteristics of undefined parameterized model followed by their estimation using arrays of real retrospective data. The most common approach of EP-ML is based on Bayes formula. It is known a fact that the method has high sensitivity to prior probabilistic characteristics which are set by experts. This feature can not be treated as positive characteristics of *EP-ML*, but there exists more important circumstance, which makes *EP-ML* methodologically flawed under uncertainty: its probabilistic characteristics are "set" and (often the only one) solutions that fit these conditions are generated [Zolotikh, 2013].

In general, the phenomena of the uncertainty in terms of its stochastic representation is much larger. It is a stochastic environment filled with random objects: vectors, trajectories, whose probabilistic characteristics are unknown. So it seems adequate to represent it as a special randomized model. Such a randomization is considered to be optimal under maximum uncertainty. Procedures in which the information entropy is used as the measure of uncertainty we will refer to procedures of *Entropy-Randomized Machine Learning (ER-ML)* [Popkov & Popkov, 2014].

2 Statement and Solution of the ER-ML Problems

Key blocks of ER-ML-procedure are the model (ERM-ML) whose parameters are randomized, and the algorithm (ERA-ML), which is a composition of mathematical formulation of the problem of estimation of probabilistic characteristics of the model.

Mathematical model is described by a nonrandom vector functional $\hat{\Omega}(\tilde{X}_{\varrho}^{(j)} \,|\, \mathbf{a}, P(\mathbf{a}))$ with random parameters $\mathbf{a}$. For each observation j, the input array (a matrix $\tilde{X}_{\varrho}^{(j)}$) consists of ϱ column vectors $\mathbf{x}(j-\varrho), \mathbf{x}(j-\varrho+1), \dots, \mathbf{x}(j)$. The model with the above-mentioned properties will be called *the randomized parameterized model (RPM)*. Consequently, the model output at observation j represents an ensemble $\hat{\mathcal{Y}}(j \,|\, P(\mathbf{a}))$ of the random vectors $\hat{\mathbf{y}}(j \,|\, P(\mathbf{a}))$ relating to the input data and random parameters through the vector functional $\hat{\Omega}(\tilde{X}_{\varrho}^{(j)} \,|\, \mathbf{a}, P(\mathbf{a}))$, i.e.,

$$\hat{\mathcal{Y}}(j \,|\, P(\mathbf{a})) = \hat{\Omega}(\tilde{X}_{\varrho}^{(j)} \,|\, \mathbf{a}, P(\mathbf{a})), \qquad j = \overline{1, s}. \tag{1}$$

The errors in the output data are modeled by an ensemble $\mathcal{E}(j \,|\, Q_j(\xi^{(j)}))$ of the random vectors $\xi^{(j)}$ with the PDF $Q_j(\xi^{(j)})$, which is added to the ensemble of the RPM output:

$$\mathcal{V}(j \,|\, P(\mathbf{a}), Q_j(\xi^{(j)})) = \hat{\mathcal{Y}}(j \,|\, P(\mathbf{a})) + \mathcal{E}(j \,|\, Q_j(\xi^{(j)})), \qquad j = \overline{1, s}. \tag{2}$$

Thus the model is the generator of random vectors with given density.

ERA-ML is formulated as the functional entropy programming problem contained k-balances with real data:

$$\begin{aligned} \mathcal{H}[P(\mathbf{a}), Q(\xi)] = -\int_{\mathcal{A}} P(\mathbf{a}) \ln \frac{P(\mathbf{a})}{P^0(\mathbf{a})}\, d\mathbf{a} - \\ \sum_{j=1}^{s} \int_{\Xi_j} Q_j(\xi^{(j)}) \ln \frac{Q_j(\xi^{(j)})}{Q_j^0(\xi^{(j)})} d\xi^{(j)} \Rightarrow \max, \end{aligned} \tag{3}$$

under conditions normalized

$$\int_{\mathcal{A}} P(\mathbf{a}) d\mathbf{a} = 1, \quad \int_{\Xi_j} Q_j(\xi(j))\, d\xi^{(j)} = 1, \; j = \overline{1, s}. \tag{4}$$

and empirical balances

$$\mathbf{m}^{(k)}(j\,|\,P(\mathbf{a}), Q_j(\xi^{(j)})) = \mathbf{y}(j), \quad j=\overline{1,s}. \tag{5}$$

Here the vector $\mathbf{m}^{(k)}$ contains the components, that are k-roots of k-moments; $P^0(\mathbf{a}), Q_j^0(\xi^{(j)})$ denote the prior PDFs of the parameters and noises, respectively..

For $k=1$ this problem has an analytical solution parameterized by Lagrange multipliers:

$$\begin{aligned} P^*(\mathbf{a}) &= \frac{P^0(\mathbf{a})\exp\left[-\sum_{j=1}^{s}\langle\theta^{(j)}, \mathbf{v}^{(j)}(\mathbf{a})\rangle\right]}{\mathcal{P}(\theta)}, \\ Q_j^*(\xi^{(j)}) &= \frac{Q_j^0(\xi^{(j)})\exp\left[-\sum_{j=1}^{s}\langle\theta^{(j)}, \xi^{(j)}\rangle\right]}{\mathcal{Q}_j(\theta)}, \quad j=\overline{1,s}. \end{aligned} \tag{6}$$

In these equalities,

$$\begin{aligned} \mathcal{P}(\theta) &= \int_{\mathbb{A}} P^0(\mathbf{a})\exp\left[-\sum_{j=1}^{s}\langle\theta^{(j)}, \mathbf{v}^{(j)}(\mathbf{a})\rangle\right] d\mathbf{a}, \\ \mathcal{Q}_j(\theta) &= \int_{\Xi_j} Q_j^0(\xi^{(j)})\exp\left[-\sum_{j=1}^{s}\langle\theta^{(j)}, \xi^{(j)}\rangle\right] d\xi^{(j)}, \quad j=\overline{1,s}. \end{aligned} \tag{7}$$

Here $\theta = \{\theta^{(1)}, \dots, \theta^{(s)}\}$ are Lagrange multipliers. They provide considerably specific system of nonlinear equations consist of so-called integral components: multidimensional definite integrals of the parameters and noises. Nonlinearity of the equations and the availability of integral components lead to the need of exploit numerical methods based on Monte Carlo Method (MMC). We have developed the *GFS: generation, filtration, selection* algorithm targeted to solving such problems [Popkov et al., 2015].

The problem of applying MMC to solving of global optimization problems with analytically defined functions is studied in many publications, for instance, [Strongin & Sergeyev, 2000,Zhigliavsky, 2006,Sergeyev & Kvasov, 2008,Polyak & Gryasina, 2008]. *GFS*-algorithm is oriented to the problems with algorithmically defined functions.

3 Applications

ER-ML procedure is applied to the problem of the *soft-binary classification.* The randomized model (decision rule) bases on a single-layer neural network with random parameters $\mathbf{a}$ is used to solve this problem:

$$\hat{y}^{(i)}(\mathbf{a}) = \text{sigm}\,\left(\langle\mathbf{e}^{(i)}, \mathbf{a}\rangle\right), \qquad i=\overline{1,m}, \tag{8}$$

where sigm is

$$\text{sigm}(x) = \frac{1}{1+\exp[-\alpha(x-\Delta)]}, \tag{9}$$

with fixed parameters α, Δ. This function has a random argument, as the parameters $\mathbf{a}$ of the randomized model are random. The values of sigm(x) from the interval $[1/2, 1]$ correspond to class 1, while the values from the open interval $[0, 1/2)$ to class 2.

Table 1

i	$e_1^{(i)}$	$e_2^{(i)}$	$e_3^{(i)}$	$e_4^{(i)}$
1	0.11	0.75	0.08	0.21
2	0.91	0.65	0.11	0.81
3	0.57	0.17	0.31	0.91

Thus, the "soft-binary" classification problem in terms of ER-ML is stated as

$$\mathcal{H}[P(\mathbf{a})] = -\int_{\mathcal{A}} P(\mathbf{a}) \ln P(\mathbf{a}) d\mathbf{a} \Rightarrow \max, \tag{10}$$

subject to the conditions

$$\int_{\mathcal{A}} P(\mathbf{a}) d\mathbf{a} = 1, \tag{11}$$

$$\int_{\mathcal{A}} P(\mathbf{a}) \text{sigm} \left(\langle \mathbf{e}^{(i)}, \mathbf{a} \rangle\right) d\mathbf{a} = y^{(i)}, \qquad i = \overline{1, m}. \tag{12}$$

The solution of this problem has the form

$$P^*(\mathbf{a}) = \frac{W^*(\mathbf{a})}{\mathcal{P}(\theta)}, \tag{13}$$

where

$$W^*(\mathbf{a}) = \exp\left(-\langle \theta, \hat{\mathbf{y}}(\mathbf{a}) \rangle\right), \tag{14}$$

$$\mathcal{P}(\theta) = \int_{\mathcal{A}} \exp\left[-\langle \theta, \hat{\mathbf{y}}(\mathbf{a}) \rangle\right] d\mathbf{a}. \tag{15}$$

Consider classification procedure for an arbitrary document $\mathbf{t}^{(j)}$.

Step 1-i. Generate an ensemble $\hat{\mathcal{Y}}^{(i)}$ of the randomized model output (decision rules) (8) with the function $P^*(\mathbf{a})$ (13). The ensemble contains N random values from the interval $[0, 1]$.

Step 2-i. If a random value from this ensemble exceeds 1/2, then document $\mathbf{t}^{(i)}$ is assigned class 1; otherwise, class 2.

Step 3-i Suppose that N_1 values are assigned class 1 and N_2 values class 2. Since the number of trials N is sufficiently large, the quantities $p_1^{(i)} = N_1/N$ and $p_2^{(i)} = N_2/N$ yield the empirical probabilities of assigning appropriate classes to document $\mathbf{t}^{(i)}$.

By repeating steps *2-i*, *3-i* for the whole collection $\mathbb{T}$, we obtain the probability distribution of assigning class 1 or 2 to the document.

Example 1. Let us consider a problem of "soft-binary" classification of 3 documents, each of which is characterized by 4 weights.

The dimension of RML-algorithm is 4, the learning collection consists of three documents each described by four weights, see Table 1.

The randomized model (8) has the parameters $\alpha = 1.0$ and $\Delta = 0$. The "learner" responses are $\mathbf{y} = \{0.18;\ 0.81;\ 0.43\}$ ($y_i < 0.5$ corresponds to class 2, $y_i \geq 0.5$ to class 1). The parameters belong to the ranges $a_i \in [-10, 10]$, $i = \overline{1, 4}$. For this learning collection, the entropy-optimal function $W^*(\mathbf{a})$ (14) takes the form

$$W^*(\mathbf{a}) = \exp\left(-\sum_{i=1}^{3} \theta_i y_i(\mathbf{a})\right),$$

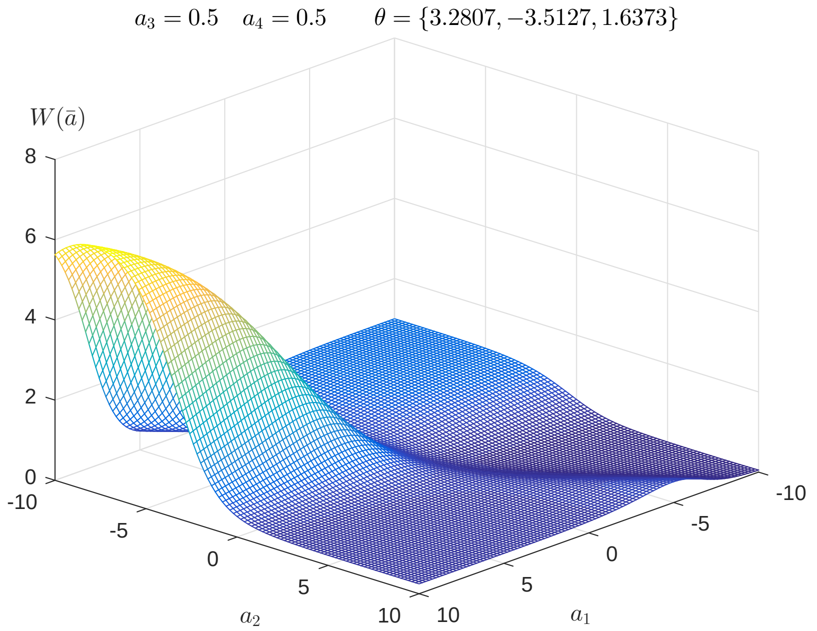


Figure 1: Two-dimensional section of the joint PDF for Example 1.

$$y_i(\mathbf{a}) = \left(1 + \exp(-\sum_{k=1}^{4} e(i)_k, a_k)\right)^{(-1)}.$$

Figure 1 shows the two-dimensional section of the function (14) under $a_3 = 0.5; a_4 = 0.5$.

For classification we use a collection of 500 documents represented by an array of the four-dimensional random vectors $\mathbf{t}^{(i)}$, $i = \overline{1,500}$ with independent components obeying the uniform distribution on the interval [0,1]. For each element of this sample, generate the random parameters of the model (9) according to the PDF $W(\mathbf{a})$ $(N = 1000)$.

Figures 2a–2b demonstrates the empirical probabilities $p_1^{(i)}, p_2^{(i)}$ of assigning class 1 and 2 to document t_i. For different documents their assigning probabilities vary from 15 to 85 percent.

Consider the *dynamic regression problem* in respect to the World population forecasting. We use the simple discrete-form exponential randomized model of population dynamics with measurement errors

$$v[ih] = E_i(b, m|E_0) + \xi[ih], \qquad i \in [0, I], \tag{16}$$

with the function

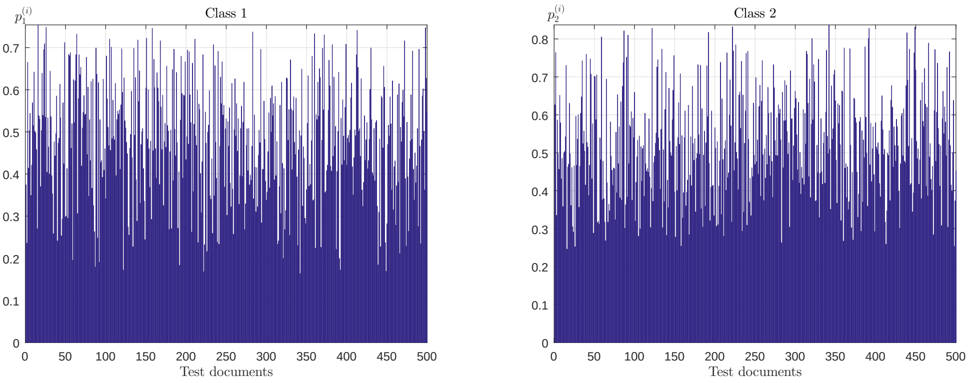
$$E_i(r, u_r \mid E_0) = E_0 \exp[(r + u_r i)ih], \quad i \in [0, I]. \tag{17}$$

where r means reproduction rate, u_r is the velocity of its changing.

The measurement errors are modeled by a random vector $\xi = \{\xi[0], \ldots, \xi[Ih]\}$ with independent interval-type components and a PDF $Q(\xi)$ defined on the set

$$\Xi = \bigcup_{j=0}^{I} \Xi_j, \qquad \Xi_j = [\xi_j^-, \xi_j^+], \tag{18}$$

The *ER-ML* algorithm yields the following entropy-optimal PDFs:



(a) (b)

Figure 2: Empirical probabilities of assigning class 1 and 2 for Example 1.

- model parameters

$$\begin{aligned} P^*(r, u_r) &= \frac{1}{\mathcal{R}(\theta|E_0)} \prod_{i=0}^{I} p_i^*(r, u_r|\theta_i), \\ p_i^*(r, u_r|\theta_i) &= \exp\left(-\theta_i E_i(r, u_r|E_0)\right); \end{aligned} \tag{19}$$

- noise

$$\begin{aligned} Q^*(\xi) &= \frac{1}{\mathcal{Q}(\theta)} \prod_{j=0}^{I} q_j^*(\xi[jh]|\theta_j), \\ q_j^*(\xi[jh]|\theta_j) &= \exp\left(-\theta_j \xi[jh]\right). \end{aligned} \tag{20}$$

where

$$\mathcal{R}(\theta|E_0) = \int_{\mathcal{I}} \prod_{i=0}^{I} \exp\left(-\theta_i E_i(r, u_r|E_0)\right) dr du_r \tag{21}$$

and

$$\begin{aligned} \mathcal{Q}(\theta) &= \prod_{j=0}^{I} \int_{\xi_j^-}^{\xi_j^+} \exp(-\theta_j \xi[jh]) d\xi[jh] = \\ &= \prod_{j=0}^{I} \frac{1}{\theta_j} \left(\exp(-\theta_j \xi_j^-) - \exp(-\theta_j \xi_j^+)\right). \end{aligned} \tag{22}$$

Here θ are Lagrange multipliers.

Example 2. Find the entropy-optimal PDFs of the model parameters and noises for the retrospective data corresponding to the period from 1960 to 1995 with step $h = 5$ *years* (see Table 2). Using $E_0 = E_{real}^{ml}[0]$ in (19-22) we obtain required PDFs (see Fig. 3–4).

The RPM is used for comparison of UN- and RPM- prognoses for interval 1995–2015 on the base UN-prognosis made at 1985. The relative mean-square deviation between the real trajectory and the ensemble-average one is 0.3%. The relative mean-square deviation for the UN prognosis is 0.8%.

Table 2: World population in *billion people*

i	0	1	2	3	4	5	6	7
	1960	1965	1970	1975	1980	1985	1990	1995
$E^{ml}_{real}[i]$	3.026	3.358	3.691	4.070	4.449	4.884	5.320	5.724

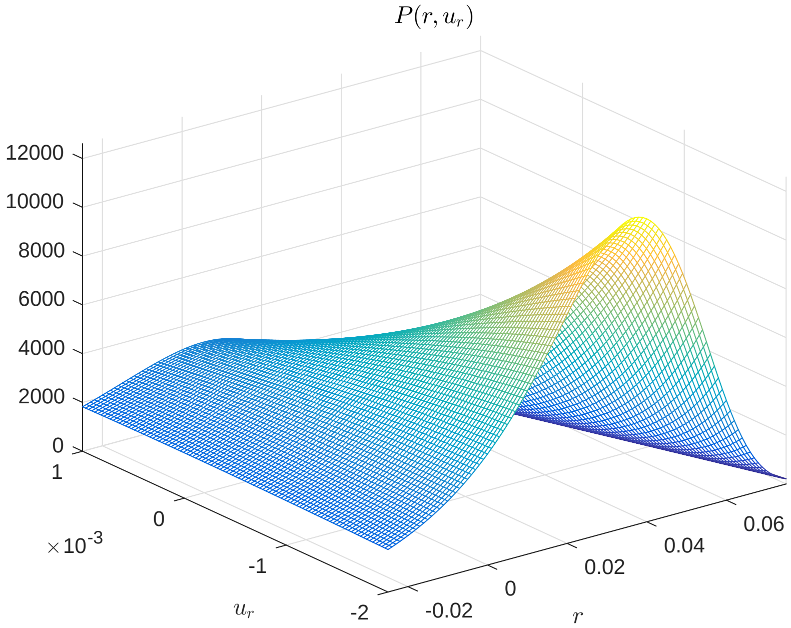


Figure 3: Joint PDF of r u_r in interval $\mathcal{I}_r \bigcup \mathcal{I}_{u_r}$.

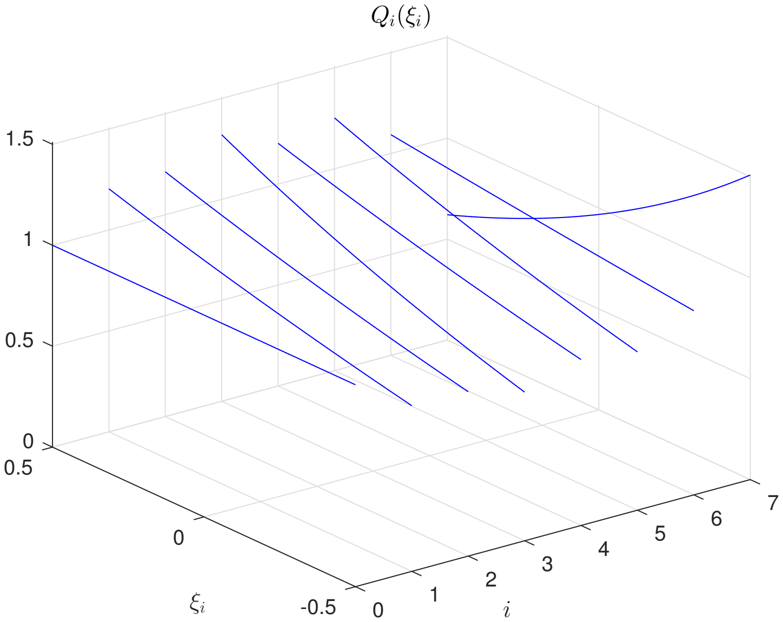


Figure 4: Ensemble of PDFs of noise ξ_i, $i \in [0, 7]$.

Acknowledgments

This work was supported by the Russian Foundation for Basic Research (project no. 16-07-00743).

References

[Bishop, 2006] *Bishop C.M.* Pattern Recognition and Machine Learning. Springer, Series: Information Theory and Statistics, 2006.

[Vorontsov, 2006] *Vorontsov K.V.* Matematicheskie metody obuchenia po precedentam (in Russian) — MIPT Lectures, 2006.

[Merkov, 2014] *Merkov A.B.* Raspoznavanie obrazov. Postroenit i obuchenie veroiatnostnikh modelei (in Russian) — Moscow, LENAND, 2014.

[Hastie et al., 2001] *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. Springer, 2001. http://www-stat.stanford.edu/ tibs/ElemStatLearn.

[Zolotikh, 2013] *Zolotikh N.Y.* Mashinnoe obuchenie i analiz dannih (in Russian), 2013. http://www.uic.unn.ru/ zny/ml.

[Popkov & Popkov, 2014] Popkov Y., Popkov A. New Method of Entropy-Robust Estimation for Randomized Models under Limited Data // Entropy, 2014, v.16, p. 675-698.

[Popkov et al., 2015] *Popkov Y.S., Popkov A.Y., Darkhovskii B.S.* Parallelnii Monte Carlo dlia postroenia entropiino-robastnikh ocenok (in Russian) // Matematicheskoe modelirovanie, 2015, Vol.27, No.6, p.14-32.

[Strongin & Sergeyev, 2000] *Strongin R.G., Sergeyev Ya.D.* Global Optimization with Non-Convex Constraints. Sequential and Parallel Algorithms. Kluwer Academic Publishers, Dordrecht, 2000.

[Sergeyev & Kvasov, 2008] Sergeyev, Ya.D. and Kvasov, D.E., Diagonal'nye metody global'noi optimizatsii (Diagonal Methods of Global Optimization). Moscow: Fizmatlit (2008).

[Zhigliavsky, 2006] *Zhigliavsky A., Žilinskas A.* Stochastic Global Optimization. Springer, 2006.

[Polyak & Gryasina, 2008] *Polyak B., Gryasina E.* Hit-and-Run: New design technique for stabilization, robusness and optimization of linear systems, In: Proc. of the IFAC World Congress. 2008, pp. 376-380.