MERGING OF USER'S KNOWLEDGE INTO IDENTIFICATION PART OF SELF-TUNERS

Josef Böhm, Miroslav Kárný

Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
Pod vodárenskou věží 4, 18208, Prague 8
Czechoslovakia, Tel: (422)8152337,
E-mail: ADAPT@CSPGAS11.BITNET

Abstract. The problem of incorporating user's knowledge – possibly uncertain and/or contradictory – is inspected. Bayesian methodology together with a technique of generating fictitious data are used for computing appropriate initial conditions of recursive least squares for estimating parameters of Gaussian ARX model. Resulting algorithms respect different uncertainty of particular pieces of available information.
From engineer's view point, the paper presents algorithms which translate "technological" knowledge of the user into probabilistic language which is usually foreign to him.

Key words. Identification, prior information, expert knowledge, least squares, self-tuners.

Introduction

Standard self-tuning controllers are based on identification combined with an appropriate controller synthesis. The synthesis transforms estimates of the controlled-system parameters into controller parameters. Consequently, the control quality depends substantially on the identification results. This dependence is especially apparent at the beginning of adaptation process: the designed controller is immediately applied and the erroneous estimates may result in wild changes of the input signal. Such excitation is quite favorable for identification itself but it is undesirable both for the technology and users. The remedy is to incorporate as much available information about the system as possible into the estimation start-up.

Success of any identification is conditioned by the amount of the relevant information supplied about the controlled system. The informational content of data is case dependent and its variations influence substantially length of the parameter-learning period. Again, the initiation is important.

To summarize, the quality of closed-loop transients when using self-tuning controller depends directly on the identification results and any piece of information should be incorporated. No source of knowledge (personal experience, physical analysis, previous experiments, etc.) should be a priori omitted.

The problem of including prior knowledge in the parameter estimation by (recursive) least-squares ((R)LS) is discussed repeatedly. For instance, in [1] constrained LS and damped LS are mentioned and a new functionally constrained LS method proposed. They represent the methods which imbed available knowledge into recursive part of the estimator.

A systematic incorporation of user's knowledge into the RLS start-up has been almost unsupported. In this respect, some possibilities are treated in [3], [4], where the solution is based on a use of "fictitious" data. This paper continues in this line. A deeper understanding offers almost cook-book recipes for standard situations and serves as a source for mastering of this useful technique.

Theoretical background

Bayesian methodology describes uncertainties – irrespectively of their source – in probabilistic terms. In the treated cases, we can deal with probability density functions (p.d.f.). With a slight abuse of notation, $p(A|B)$ will denote the p.d.f. of an uncertain (random) variable A conditioned on B (the random variable, its realization and the corresponding p.d.f. argument will not be distinguished, as usual).

In the adopted framework, the uncertainty about a (multivariate) unknown parameter $\underline{\Theta}$ is fully described by a prior p.d.f. $p(\underline{\Theta}|D(0))$. The symbol $D(0)$ is a formal label for the information available. If this information *extends* to $D(t)$ – for instance, by measuring data at discrete time moments $1, 2 \dots, t$ – the complete information compresses into the posterior p.d.f. $p(\underline{\Theta}|D(t))$.

Proposition 1.[*Bayes rule*] The prior and posterior p.d.f.s are related by Bayes formula

$$p(\underline{\Theta}|D(t)) = \frac{p(D(t)|\underline{\Theta})p(\underline{\Theta})}{\int p(D(t)|\underline{\Theta})p(\underline{\Theta})\, d\underline{\Theta}} \propto p(D(t)|\underline{\Theta})p(\underline{\Theta}) \tag{1}$$

where $\propto$ denotes equality up to a $\underline{\Theta}$-independent normalizing factor. The p.d.f. $p(D(t)|\underline{\Theta})$ which links the

measured (uncertain) data to the unknown parameter has to result from the system model.
Proof. Elementary theorem of probability theory □

We shall use the Bayes rule specialized for the parameter having two components $\underline{\Theta} = (\Theta, r)$, regression coefficients (Θ) and a noise dispersion (r). The measured data are formed by the sequence of pairs $D(t) = \{(y(1), u(1)), \ldots, (y(t), u(t))\}$ where

$y(t)$ is the system output (here, the single-variate case is treated without loss of generality [6]);
$u(t)$ is an exogenous variable (possibly multivariate) fed into the identified system (usually, the system input) which is supposed to fulfill so called natural conditions of control [8] (met, loosely speaking, for any input generator for which $\underline{\Theta}$ is an unknown parameter).

Proposition 2.[*Bayes rule and likelihood function*] Under the natural conditions of control, it holds

$$p(\Theta, r|D(t)) \propto \mathcal{L}(\Theta, r|D(t))p(\Theta, r|D(0))$$

where the *likelihood function* $\mathcal{L}(\Theta, r|D(t))$ is defined

$$\mathcal{L}(\Theta, r|D(t)) = \prod_{i=1}^{t} p(y(i)|D(i-1), u(i), \Theta, r)$$

and understood as a function of the parameters Θ, r.
Proof. See e.g. [8] □

The conditional p.d.f.s $p(y(i)|D(i-1), u(i), \Theta, r)$ which link the observed data to unknown parameters must be defined by the system model [8].

Remarks

1. Notice the recursive nature of Bayes rule: If $D(t)$ extends to $D(\bar{t})$, $\bar{t} \geq t$, the posterior p.d.f. $p(\underline{\Theta}|D(t))$ plays the role of the prior p.d.f. for determining $p(\underline{\Theta}|D(\bar{t}))$.
2. The necessity to specify a prior p.d.f. is a frequent objection against Bayesian formalism. It is, however, always possible to use a noninformative (i.e. sufficiently flat) prior p.d.f. when there is real lack of prior information.
3. If some prior information is available Bayesian setup opens a gate for its systematic exploitation. The practical construction of the relevant p.d.f. is, however, non-trivial especially when diverse sources are combined. This paper provides the user with a systematic support for this purpose. It uses the direct consequence of the remarks 1 and 2:

If the available prior information can be formalized as a (fictitious) measurement of data on the inspected system, the prior p.d.f. can be generated according to Bayes rule applied to them, starting from the noninformative prior p.d.f.

Bayesian estimation of ARX model

We shall present Bayesian view on RLS by applying Bayesian estimation to ARX model.

The Gaussian ARX model gives $p(y(t)|D(t-1),\ u(t), \Theta, r) = \mathcal{N}(y(t)|\Theta'\psi(t), r)$ where

Θ is the i_ψ-vector of unknown regression coefficients, $'$ means transposition;
$r > 0$ is the unknown conditional dispersion of the output;
$\psi(t)$ is the regression i_ψ-vector, i.e. a known function of $D(t-1),\ u(t)$;
$\mathcal{N}(y|\hat{y}, r)$ denotes the Gaussian p.d.f. determined by the expected ($\mathcal{E}$) value $\hat{y}$ of y and by a dispersion r.

For the ARX model, the likelihood takes the form

$$\mathcal{L}(\Theta, r|D(t)) \equiv GiW(\Theta, r|V(t), \nu(t)) \propto$$
$$\propto r^{-0.5\nu(t)} \exp\left\{-\frac{1}{2r}\begin{bmatrix}-1\\ \Theta\end{bmatrix}' V(t)\begin{bmatrix}-1\\ \Theta\end{bmatrix}\right\}$$

where GiW is Gauss-inverse-Wishart p.d.f. and the data $D(t)$ are compressed into the sufficient statistics

$$\begin{aligned}\nu(t) &= \nu(t-1) + 1 \qquad (2)\\ V(t) &= \begin{bmatrix} V_y(t) & V'_{\psi y}(t)\\ V_{\psi y}(t) & V_\psi(t)\end{bmatrix} = V(t-1) + \Psi(t)\Psi'(t)\\ \nu(0) &= 0, \quad V(0) = 0.\end{aligned}$$

$V(t)$ is an $(i_\psi+1, i_\psi+1)$-symmetric positive semidefinite matrix which becomes positive definite if the data $D(t)$ have produced at least $i_\psi + 1$ linearly independent data vectors $\Psi'(\cdot) = [y(\cdot); \psi'(\cdot)]$.
Proof. See e.g. [8]. □

Proposition 3.[*Reproducing prior p.d.f. for the ARX model*] Suppose that the function $\mathcal{L}(\Theta, r|V_o, \nu_o) \propto GiW(\Theta, r|V_o, \nu_o)$ can be normalized to a p.d.f., i.e. the symmetric positive definite $(i_\psi + 1, i_\psi + 1)$ matrix V_o and the scalar ν_o guarantee $0 < \int GiW(\Theta, r|V_o, \nu_o)\, d\Theta\, dr < \infty$. Then, the prior p.d.f. $p(\Theta, r|D(0)) \propto \mathcal{L}(\Theta, r|V_o, \nu_o)$ reproduces, i.e. keeps the fixed functional form $p(\Theta, r|D(t)) \propto GiW(\Theta, r|V(t), \nu(t))$ with statistics evolving according to the recursions (2) *with zero initial conditions replaced by* $\nu(0) = \nu_o,\ V(0) = V_o$.
Proof. Straightforward consequence of Prop. 2. □

Remarks

1. The ARX model is usually written in the "equation" form $y(t) - \Theta'\psi(t) = e(t)$ where the $e(t)$ is *white Gaussian noise*, i.e. $p(e(t)|D(t-1), u(t), \Theta, r) = \mathcal{N}(e(t)|0, r)$. This form stresses *time-invariance of the noise dispersion*, loosely speaking, a common uncertainty level is assumed. Only under this assumption, the data vectors $\Psi(t)$ sum into the statistic $V(\cdot)$ with a constant weight. Moreover, the *noise has to be sequence of uncorrelated random variables*, otherwise optimality of RLS is lost.
2. Proposition 3 illustrates the conclusion made at the end of previous paragraph: $V_o,\ \nu_o$ which shape prior p.d.f. can be thought as if they were found by measuring data on ARX system and using them for modification of a noninformative (pre)prior p.d.f.

Properties of GiW distribution

Let us summarize the facts relevant to the aim of the paper.

Proposition 4.[*LS form of the GiW p.d.f.*] The Gauss-inverse-Wishart p.d.f. of a real i_ψ-vector Θ and of a positive dispersion r reads

$$GiW(\Theta, r|\hat{\Theta}, P, \hat{r}, \kappa) \propto r^{-(\kappa+i_\psi+2)/2}$$
$$\exp\left\{-\frac{1}{2r}[(\Theta-\hat{\Theta})'P^{-1}(\Theta-\hat{\Theta}) + \kappa\hat{r}]\right\}$$
$$P > 0, \quad \hat{r} > 0, \quad \kappa > 0$$

where the statistics determining this form of the *GiW* p.d.f. are related to the statistics V, ν (2) as follows

$$\begin{aligned} P &= V_\psi^{-1}, \quad \hat{\Theta} = PV_{\psi y} \\ \hat{r} &= \frac{\lambda}{\kappa}, \qquad \lambda = V_y - V_{\psi y}' V_\psi^{-1} V_{\psi y}, \\ \kappa &= \nu - i_\psi - 2 > 0. \end{aligned} \tag{3}$$

Proof. By completing squares with respect to Θ and by using the partitioned form (2) of V. □

Proposition 5.[*Selected properties of the GiW distribution*] For the *GiW* distribution, it holds

$$\begin{aligned} p(\Theta) &\propto \left\{\lambda + (\Theta - \hat{\Theta})' P^{-1} (\Theta - \hat{\Theta})\right\}^{-(\kappa+1)/2} \\ \mathcal{E}[\Theta] &= \mathcal{E}[\Theta|r] = \hat{\Theta} \quad \mathrm{cov}[\Theta] = \frac{\lambda}{\kappa} P \\ \mathcal{E}[r] &= \hat{r} = \frac{\lambda}{\kappa}, \quad \mathrm{cov}[r] = \frac{\lambda^2}{(\kappa-2)(\kappa-4)} \end{aligned} \tag{4}$$

Proof. See e.g. [9] □

Remarks

1. The statistics $\hat{\Theta}, P, \hat{r}$ are well known least-squares quantities.

2. The definitions (3) and the recursions (2), together with the well known matrix-inversion lemma, lead to updating formulae for least-squares statistics, to recursive least squares (RLS).

If the time argument is suppressed and the updated quantities are distinguished by the subscript $_n$, the evolutions read

$$\begin{aligned} \hat{\Theta}_n &= \hat{\Theta} + \frac{P\psi}{1+\zeta}\hat{e}, \quad \hat{e} = y - \hat{\Theta}'\psi, \\ P_n &= P - \frac{P\psi\psi' P}{1+\zeta}, \quad \kappa_n = \kappa + 1, \\ \zeta &= \psi' P \psi, \quad \hat{r}_n = \hat{r} + \frac{1}{\kappa_n}\left[\frac{\hat{e}^2}{1+\zeta} - \hat{r}\right]. \end{aligned}$$

3. The prior p.d.f. is defined by all four variables $\hat{\Theta}, P, \lambda, \kappa$. Neglecting this fact leads easily to disappointing results. For instance, a lot of effort is often spent when selecting the initial point estimate $\hat{\Theta}$ of the regression coefficients Θ. Looking, however, at the above updating formula, it is clear that a single step can spoil even the exact estimates if there is a bit of noise and the gain (determined by P) is improperly chosen.

This possibility is far of being academical as a large diagonal P is often recommended and used.

Even if the P-level is reasonable, the diagonal choice damages the carefully selected point estimates. The Bayesian interpretation of P implies it clearly: unrealistic independence of Θ entries is modelled by the diagonal matrix P.

The problem and its solution

From a formal point of view, a prior p.d.f. should be constructed that properly reflects the state of our knowledge. From an algorithmic view point, the sufficient statistics $\hat{\Theta}, P, \lambda, \kappa$ – i.e. initial conditions for RLS – reflecting our knowledge are searched for.

We shall restrict ourselves to the quantification of the information about the regression coefficients. The initial point estimate $\hat{r}$ of the noise dispersion r is assumed to be at disposal. Neither its construction nor a more detailed specification of uncertainty about r will be addressed in this paper.

The advocated procedure will be presented by describing cornerstones of the problem and its solution, namely,

- typical sources of prior information;
- generating of fictitious data;
- scaling of fictitious data;
- putting fictitious data together before continuing on real data.

Notation. The subscript $_f$ will be used at the discrete-time label in order to stress difference between the time of fictitious and real measurements.

Typical sources of prior information.

The assumed controlled system is usually a complex dynamic object. When identifying it for self-tuning control we try to estimate the coefficients of the ARX model of a fixed structure (order). It is not so easy to guess the values of its parameters, but some partial information about the system almost always exists. It could have the form of

- partial knowledge of the transfer function
  - some time constant (usually the largest),
  - approximate static gain;
- information about the frequency response
  - cut-off frequency,
  - gain (and phase) at a given frequency;
- lower and upper limits of a typical step response;
- a simpler regression model;
- any kind of "simulation" model, even nonlinear.

We call any of these information sources a *partial model*. The information contained in the partial models is of a different origin and/or precision. Some of them can be based on relatively precise measurements (frequency response at specific frequency or in a frequency range) others are based on a vague experience (time constants).

Clearly, the contribution of partial models to overall picture has to differ according to its precision specification: a partial model can be used appropriately only if its uncertainty (belief in its validity) is supplied. Data derived from the partial models for generating initial values of RLS have to reflect this uncertainty.

Idea of fictitious data

At least in some cases, it is easy to map knowledge reflected by partial models on the point estimate $\hat{\Theta}$, but a direct construction of the relevant covariance matrix P is far to be trivial. This makes quite appealing the outlined *idea of "fictitious" data*:

> Map a knowledge reflected in partial models onto such data $\Psi' = [y; \psi']$ that you cannot distinguish whether they were measured on your system or they are just a "fiction".

This vague and seemingly unnecessary idea – justified formally by the recursive nature of Bayesian estimation – helps surprisingly much in facing even quite complex situations. The strength of fictitious data stems from

- simplicity of their generating by partial models;

• ability to correlate the available knowledge to the signal values really measured;
• need to handle (hyper)planes only which are simpler objects than (hyper)ellipsoids induced by covariances.

To make the exposition transparent only single input ARX models linear-in-data are treated, i.e.

$$\Psi'(t) = [y(t); \ldots, y(t-l_y), u(t), \ldots, u(t-l_u)] \quad (5)$$

with some "orders" $l_y, l_u \geq 0$.

Unscaled fictitious data

By definition, fictitious data express affine relation among regression coefficients of ARX model. Consequently, scaling of these data is arbitrary until uncertainty level is taken into account. The choice of proper scaling is postponed to the next paragraph. Arbitrariness is stressed by introducing a scaling factor $a \neq 0$.

In some cases, the data vector Ψ can be found directly. An indirect construction is more frequent: Ψ is created by passing (mentally) appropriate input signal(s) through a model.

Static gain. The knowledge of the static gain G is an example when direct approach can be used. The fictitious data $\Psi' = [aG; aG, ..., aG, a, a, ...a]$ (for any $a \neq 0$) fit to systems with the gain G.

Frequency response. Let the input considered be $u(t_f) = cos(\omega t_f)$, specified by a fixed frequency ω. Then – in steady state – the output is expected to be $y(t_f) = Y(\omega)cos(\omega t_f + \phi(\omega))$. The inclusion of all pairs $\Psi'(t_f) = [y(t_f); \psi'(t_f)]$ for infinite amount of t_f can be shown to be equivalent to the following fictitious data vectors [5]

$$\begin{aligned}\Psi'_c(\omega) &= a[Y(\omega)\cos(\phi); Y(\omega)\cos(\omega+\phi), .., \\ &\quad Y(\omega)\cos(l_y\omega+\phi), 1, \cos(\omega), .., \cos(l_u\omega)] \\ \Psi'_s(\omega) &= a[Y(\omega)sin(\phi); Y(\omega)\sin(\omega+\phi), .., \\ &\quad Y(\omega)\sin(l_y\omega+\phi), 0, \sin(\omega), .., \sin(l_u\omega)].\end{aligned}$$

The derivation rests on the identity
$\Psi(t_f) = \cos(\omega t_f)\Psi_c - \sin(\omega t_f)\Psi_s$.

Time constant. The information about a time constant T can be represented in various ways. The knowledge of the frequency response described above is one possibility. Knowledge of a set of measured (thought of or designed when constructing the process to be controlled) step responses is other possibility. For it, the corresponding input (unit step) and output data pairs are natural entries of the fictitious data vectors.

If there is a true *dominating* time-constant, then it is reasonable (for the given purpose) to approximate the system by the first order model which responds by $y(t_f) = 1 - \exp\{-t_f/T\}$ on the unit-step input (unit gain G is assumed for simplicity). These data have to be appropriately paired into data vectors (5).

Simulation model. Experiments with simulation models are natural sources of input-output data, i.e. of the fictitious data. Typically, quite complex and/or non-linear models are at disposal. Simulation runs are preferable to the often applied numerical/analytical model reductions/approximations as they preserve (in the generated data) the information about the discrepancies of the simulated and ARX models. The approximation takes place at the moment when the data are used in RLS. Note that the discrepancy will increase the uncertainty of this partial model.

Scaling of fictitious data

We have shown how to generate – up to scaling – the data representing the prior knowledge. The scaling ambiguity will be removed here and appropriate examples will be given.

Essentially, scaling factors of respective fictitious data vectors are used for balancing the respective uncertainties attributed to them. We rely on:

Proposition 6.[*Scaling of fictitious data with uncertainty in fictitious noise*] Let us assign to fictitious data vectors $\Psi(t_f)$ the fictitious noise $e(t_f) = -[-1, \Theta']\Psi(t_f)$ with uncertainty given by the dispersions $\gamma^2(t_f)$, $\gamma(t_f) \neq 0$. Then, the scaled data

$$\tilde{\Psi}(t_f) = \gamma\Psi(t_f)/\gamma(t_f), \text{ with a } t_f\text{-invariant } \gamma \neq 0$$

have to be used in RLS.

For notational simplicity, we select $\gamma = 1$.

Proof. As recalled above, (fictitious) data vectors may enter into LS with a common weight only if the noise has a constant dispersion, say γ^2. Elementary properties of moments imply that $\tilde{e}(t_f) = \gamma e(t_f)/\gamma(t_f)$ have the required dispersion γ^2. This scaling and the definition of the fictitious noise implies the rest $\tilde{e}(t_f) = \gamma e(t_f)/\gamma(t_f) = -\gamma[-1, \Theta']\Psi(t_f)/\gamma(t_f) = -[-1, \Theta']\tilde{\Psi}(t_f)$.

□

Remarks

1. The requirement to know (time-variant) fictitious noise dispersions seems to be demanding. Often, however, the intervals $[y_0(t_f) - \gamma(t_f), y_0(t_f) + \gamma(t_f)]$ are available in which we expect (with high confidence) values of $y(t_f)$ for given $\psi(t_f)$. For Gaussian distribution, which is implicitly attributed to the fictitious noise, the standard deviation is proportional to the interval length. Thus, the data $[\tilde{y}(t_f); \tilde{\psi}(t_f)] = [y(t_f); \psi'(t_f)]/\gamma(t_f)$ lead to the fictitious noise with a common dispersion. Specific value of this dispersion depends on the numerical meaning of "high confidence". As stated in Proposition 6, this ambiguity makes no harm. The meaning itself, however, should not vary within the fictitious data set.

2. If our *uncertainty is related to other variables then to the fictitious noise* the uncertainty have to be recalculated, usually with a help of the partial model itself. A typical exploitation of "extreme" realizations to this purpose can be seen on the following example:

Let a dominant time constant T be expected in the range $[\underline{T}, \bar{T}]$. For unit step on input, we get (for $G = 1$) probable ranges of the outputs

$$y(t_f) \in \left[1 - \exp\{-t_f/\bar{T}\}, 1 - \exp\{-t_f/\underline{T}\}\right].$$

These intervals can be expressed in terms of the mean trajectory $y_0(t) = 1 - \left(\exp\{-t/\bar{T}\} + \exp\{-t/\underline{T}\}\right)/2$ and of the corresponding (time-variant) half-width (proportional to the standard deviation of the noise) $\gamma(t_f) = \left(\exp\{-t_f/\bar{T}\} - \exp\{-t_f/\underline{T}\}\right)/2$, i.e. in the style discussed in the remark 1. It leads directly to the scaled fictitious data vectors

$$\begin{aligned}&[\tilde{y}(t_f); \ldots, \tilde{y}(t_f - l_y), \tilde{u}(t_f), \ldots, \tilde{u}(t_f - l_u)] = \\ &= [y_0(t_f); \ldots, y_0(t_f - l_y), 1, \ldots, 1]/\gamma(t_f).\end{aligned}$$

3. There are partial models for which the above reduction of uncertainty to the noise part is hardly possible. *Nonlinear stochastic simulation models* are typical examples of this type. For these cases, no constructive cook-book exists, just a general guideline: try to keep the noise level as constant as possible by a suitable data re-normalization.

Putting fictitious data together

Before applying RLS to the fictitious data for correcting a vague prior information by the information contained in partial models *we should check* the remaining RLS applicability condition, i.e. *whiteness of the fictitious noise*.

As the particular pieces of information are usually quantified independently (for instance, by several experts) the whiteness violation is more the rule than an exception. The problem is faced by the two-stage procedure:

• application of RLS irrespectively of the whiteness violation gaining the RLS statistics $\tilde{\Theta}, \tilde{P}, \tilde{\lambda}, \tilde{\kappa}$;
• modification of the statistics $\tilde{\Theta}, \tilde{P}, \tilde{\lambda}, \tilde{\kappa}$ to the initial conditions of RLS $\hat{\Theta}(0), P(0), \lambda(0), \kappa(0)$.

Loosely speaking, the modification is based on an optimal approximation of the statistics $\tilde{\Theta}, \tilde{P}, \tilde{\lambda}, \tilde{\kappa}$ by another set which could result from a correctly applied RLS. A precise formulation can be found in [5]. It leads to the following solution

$$\begin{aligned}\hat{\Theta}(0) &= \tilde{\Theta}, \quad \kappa(0) = \min(\tilde{\kappa}, \dim(\Theta)) \\ \lambda(0) &= \text{the best estimate of noise level, cf. Prop. 5.,} \\ P(0) &= \frac{\tilde{\lambda}}{\lambda(0)}\tilde{P}. \end{aligned} \tag{6}$$

Instead of describing technical details of the approximation we give:

Explanatory remarks

1. The equality $\hat{\Theta}(0) = \tilde{\Theta}$ is intuitively appealing and no other alternative can be expected.
2. The need to choose $\lambda(0)$ is enforced by our assumption that the exploited partial models bring no information about it.
3. The LS remainder $\tilde{\lambda}$ is the most important by-product gained in the first stage of putting the fictitious data together.

At interpretation level, non-whiteness means that the mixture of the following cases occur:

∗ *the fictitious data are repetitive*: the remainder stops to grow after repetition, however, $\tilde{\kappa}$ increases irrespectively of it;
∗ *the fictitious data are contradictory*: equation errors are greater than expected as the partial pieces of information are insufficiently mutually compatible.

To summarize, the more fictitious noise differs from the ideal non-repetitive and non-contradictory case the more the estimate of the fictitious-noise dispersion differ from unity.
4. The clipping the κ-value at $\dim(\Theta)$ (resulting from the approximation) is intuitively appealing. The statistic κ can be interpreted as the counter of the fictitious data vectors used in RLS (the number of equations). If the number of these equations is smaller than the number of estimated parameters then the fictitious noise (equation errors) can always be taken as white.
5. The important re-normalization of the covariance matrix (implied by the referred approximation, too) can be interpreted as invariance of the coefficient covariance (cf. (4)) when passing from the fictitious noise level to the real noise level.

$$\mathrm{cov}(\Theta) = \frac{\tilde{\lambda}}{\kappa(0)}\tilde{P} = \frac{\lambda(0)}{\kappa(0)}P(0) \Rightarrow P(0) = \frac{\tilde{\lambda}}{\lambda(0)}\tilde{P}.$$

6. The re-normalization of the matrix $\tilde{P}(0)$ can be understood also as an additional normalization of the whole set of the fictitious data: *the more contradictory the data set is, the smaller weight it gets.*

Algorithmic Summary

1. Start with non-informative initial conditions:
$P = 1/\varepsilon I$, $\varepsilon > 0$ very small, $\lambda = 1$, $\hat{\Theta} = 0, \kappa = 0$.
2. Create pairs $\Psi(t_f) = [y(t_f); \psi'(t_f)]$ and scale them by the standard deviation $\gamma(t_f)$ covering the expected variations of the output $y(t_f)$ to $\tilde{\Psi}(t_f) = \Psi(t_f)/\gamma(t_f)$.
3. Process all data $\Psi(\cdot)$ by RLS algorithm without forgetting.
4. Transform the gained statistics $\tilde{\Theta}, \tilde{P}, \tilde{\lambda}, \tilde{\kappa}$ into the RLS initial conditions $\hat{\Theta}(0), P(0), \lambda(0), \kappa(0)$ according to the formulae (6).

Simulation examples

Second order system with the transfer function $F(s) = 1/(s+1)^2$ is the controlled system *used in all examples.* A colored noise – gained by passing white noise of unit intensity through the same transfer function – is added to the system output. The sampling period $t_s = 0.026$sec with zero order hold gives rise to the regression model

$$y(t) = \sum_{i=1}^{2} a_i y(t-i) + \sum_{i=0}^{2} b_i u(t-i) + e(t)$$

with the coefficients $\Theta' = [a_1, a_2, b_0, b_1, b_2] = [1.9487, -.9493, .0003322, .0003265, 0]$.

The *self-tuning controller* approximately minimizes stationary quadratic criterion with the unit output penalty and the weight on u^2 equal to 1e-4. It uses a certainty-equivalence version of a strategy called iterations spread in time [2].

The influence of the incorporated information is judged according to the behavior of the input signal on the time interval corresponding to 50 sampling instants. The *loss* $\sum_{i=1}^{50} u_i^2$ is evaluated. This loss reflects overall differences in the tuning quality as the changes in the output behavior are negligible for the system assumed. The loss values presented in tables should be compared to the *ideal loss* 52 (reached under complete knowledge) and to the *loss* 10 900 *accumulated when a non-informative prior p.d.f. is used.*

The tables contain also the gained initial estimates $\hat{\Theta}(0)$. The initial values of $P(0)$ are not presented as they cannot be grasped by human beings. They are, however, even "optically" far from the textbook standard $1/\varepsilon I$.

Thus, a t_f-th row of tables contains the estimate gained after including the fictitious data $\Psi(1), \ldots, \Psi(t_f)$ and the loss reached when starting from the corresponding RLS initial conditions.

Example 1. It illustrates the influence of a successive inclusion of non-contradictory data representing correct information about the gain and phase at frequencies $\omega = \{0.1, 5\}\text{sec}^{-1}$. Uncertainties assigned to these values are constant, given by $\gamma = 0.1$.

According to the presented theory, the above information converts into four vectors of fictitious data $\Psi_c(\omega = 0.1)$, $\Psi_s(\omega = 0.1)$, $\Psi_c(\omega = 5)$, $\Psi_s(\omega = 5)$ which are fed in RLS as $\Psi(t_f)$, $t_f = 1, 2, 3, 4$.

t_f	$\hat{a}_1$	$\hat{a}_2$	$\hat{b}_0$	$\hat{b}_1$	$\hat{b}_2$	Loss
1	0.195	.195	.197	.197	.197	130
2	0.499	.499	–.004	–.001	.006	128
3	1.554	–.558	.414	–.753	.343	118
4	1.555	–.560	.371	–.753	.386	175

Example 2. It illustrates a positive influence even (reasonably) biased prior information. It presents results gained when two different regression models (differing from the true one) were used as the information sources. They correspond to the systems

$$F_a = 1./(s + .7)^2 \quad \text{and} \quad F_b = 1/(s + 1.3)^2$$

sampled with the period 0.02sec. Three points taken from different part of the step response (at time moments $\{t_1, t_2, t_3\} = \{0.2, 2, 5\}$sec) are used for both models. Again, the common uncertainty $\gamma = 0.3$ is assigned to all data. RLS modify the noninformative priors by the data $\{\Psi(t_f), t_f = 1, \ldots, 6\} = \Psi_a(t_1), \Psi_a(t_2), \Psi_a(t_3), \Psi_b(t_1), \Psi_b(t_2), \Psi_b(t_3)$ (the subscript $a(b)$ corresponds to the underlying system). When looking at the table we should recall that fictitious input are constant in this case, thus, the observed equality of $\hat{b}$s is natural.

t_f	$\hat{a}_1$	$\hat{a}_2$	$\hat{b}_0$	$\hat{b}_1$	$\hat{b}_2$	Loss
1	0.097	0.098	7.21e-3	7.21e-3	7.21e-3	411
2	0.499	0.500	1.98e-3	1.98e-3	1.98e-3	149
3	1.949	–0.949	1.65e-4	1.65e-4	1.65e-4	143
4	2.003	–1.003	6.20e-5	6.20e-5	6.20e-5	129
5	1.991	–.9916	6.61e-5	6.61e-5	6.61e-5	133
6	1.989	–.9903	6.72e-5	6.72e-5	6.72e-5	135

Example 3. Results with two partial models of different uncertainty are presented. The first data $\Psi(1), \Psi(2)$ comprise the information about the response of the true model $F(s)$ on the frequency $\omega = 0.1\text{sec}^{-1}$. The assigned uncertainty is $\gamma_1 = 0.01$. The data $\Psi(3), \Psi(4)$ reflect step response of the above system F_b at two extreme points 0.2, 5 sec. The uncertainty assigned to $\Psi(3), \Psi(4)$ has the common value $\gamma_2 = 1.0$.

t_f	$\hat{a}_1$	$\hat{a}_2$	$\hat{b}_0$	$\hat{b}_1$	$\hat{b}_2$	Loss
1	0.195	.195	.197	.197	.197	130
2	0.490	.495	9.5e-5	0.502	.010	134
3	0.579	–.418	.329	.003	–.324	120
4	1.761	–.764	.002	.007	–.007	167

Conclusions

The paper describes and motivates a way of incorporating user's knowledge of a different origin and nature into the initial conditions of RLS. The described algorithm is quite general with respect to information sources (theory, experience, simulation models) and their mutual relations (contradictions, repetitions and uncertainties in data are allowed). Plausible consequences of including such an information into a start-up of self-tuners are illustrated on simulated examples.

The presented theory has Bayesian motivation, but we have tried to keep in touch with RLS framework as much as possible in order to:

∗ demonstrate that Bayesian view-point substantiates subtle but important algorithmic steps which can be found sensible even without the Bayesian framework but which are difficult to invent ad hoc;
∗ provide a methodology-independent cook-book for a satisfactory start up of RLS.

It could be objected that it has little sense to care much about initial conditions of RLS as they are mostly applied with a sort of (e.g. exponential) forgetting and their influence is gradually lost. The improvements of the transients of closed loops with self-tuners make this objection a bit weaker but the direct possibility to use the constructed p.d.f. permanently as the reference for restricted forgetting [7] refutes the objection, hopefully, completely.

References

[1] Fraser A. (1991). Inclusion of prior knowledge in parameter estimation via weighted parameter functions. European Control Conference, Grenoble, free copy of the presented paper.

[2] Kárný M., A. Halousková, J. Böhm, R. Kulhavý and P. Nedoma (1985). Design of linear quadratic adaptive control: theory and algorithms for practice. Supplement to Kybernetika, **21**, No. 3-6.

[3] Kárný M., J. Böhm (1991). Probabilistic modelling of imprecisely known systems for robust LQ design. European Control Conference, Grenoble,bf 1, 426-431.

[4] Kárný M., (1991). A note on feeding uncertain knowledge into recursive least squares. Preprints 30th CDC, Brighton, U.K., **1**, 975-976.

[5] Kárný M. (1991). Soft prior information for recursive least squares. Submitted to Int. Journal of Adaptive Control and Signal Processing.

[6] Kárný M. (1992). Parametrization of multi-output autoregressive regressive model for self-tuning control. Kybernetika, in print.

[7] Kulhavý R. (1987). Restricted exponential forgetting in real-time identification. Automatica **23**, 5.

[8] Peterka V. (1981). Bayesian approach to system identification. In: Trends and Progress in System Identification (P. Eykhoff, ed.). Oxford, Pergamon Press 1981, Chap. 8. Translated into Russian: Mir, Moskva 1983.

[9] Zellner A. (1971). Introduction to Bayesian Inference in Econometrics. John Wiley and Sons, New York.