Performance Measures and the Robust and Optimal Control Design

M. Huba∗

∗ STU FEI Bratislava (e-mail: mikulas.huba@stuba.sk)

Abstract: The paper discusses performance measures used dominantly in the robust and optimal control design. By simple examples of the first order time delayed (FOTD) system control it illustrates that the usually prescribed levels of the maximal and complementary sensitivity functions indeed define situations with interesting loop properties, but may not be universally applied to the robust and optimal design of systems with an uncertain feedback variable in a broader range. For this purpose the shape related measures based on deviations from monotonicity yield results matching the technological requirements of practice in a much more appropriate way. It is also shown that in an optimal nominal controller design the monotonicity based performance measures nearly coincide with the multiple real dominant pole (MRDP) method. For the loop optimization using a broader spectrum of different performance measures, the performance portrait method may be recommended. It avoids the problems of convergence to the absolute optimum and, once generated, the performance portrait may repeatedly be used with a limited effort for a broader spectrum of different cost function specifications.

Keywords: Robust control, optimal control, performance measures

1. INTRODUCTION

A key step in the optimal and robust controller design is to choose a cost function and a performance assessment method. As, for example, pointed out in Shinskey (1990), the speed of a control process may well be approximated to choose a cost function and a performance assessment method. As, for example, pointed out in Shinskey (1990), the speed of a control process may well be approximated

\[ IAE = \int_0^\infty |e(t)| \, dt \quad e = r - y \]  

Since in the controller optimization IAE usually yields a global optimum for transients with a not always acceptable overshooting, which may also display a reduced robustness, additional optimization constraints are frequently applied. In majority of contemporary publications, these are represented by peaks of the maximal sensitivity and complementary sensitivity functions. Additional constraints may be put on the noise, or on the load disturbance sensitivity (Mercader and Banos, 2017), etc. After a problem specification, the optimization may be carried out by a huge number of available mathematical packages.

Historically, the choice of the preferred performance assessment methods and measures varied. Ziegler and Nichols (1942) used in their pioneering and still frequently cited work a shape related “quarter amplitude damping”. During several decades of the minimum time control dominance, a number of (rectangular) pulses of an optimal relay control signal covered by the “Feldbaum theorem” have been broadly used (Feldbaum, 1965; Föllinger, 1993; Glattfelder and Schaufelberger, 2003). Similar shape related requirements may be found also in other practically oriented publications. For example, Visioli and Zhong (2011) cite a paper by Wang and Cluett (1997) which is based on specifying the desired control signal corresponding to a considered reference setpoint signal \( r \). For its step change and a damping factor 1 such a control initially increases to a maximum and then decays exponentially and monotonically to zero. Since such shape related requirements are explicitly, or implicitly followed by many authors, at the PID'12 conference in Brescia, a new method for the robust and optimal PI controller design (Huba, 2012) has been presented based on the shape related performance measures used for a “performance portrait” generation. It starts with checking the closed loop properties over a grid of normed (dimensionless) loop parameters. The collected information accumulated by using appropriately chosen performance measures may later be used in a controller design matching differently chosen technological specifications. The performance assessment and loop optimization may be carried out on an arbitrary set of performance measures, including all the above mentioned indexes. The design and in its extended version in Huba (2013) have been carried out by the shape related performance measures based on the concept of monotonicity.

Monotonicity (MO) represents one of the central concepts in mathematics applied in control. A measure for evaluating deviations from monotonicity may be found, for example, in Åström and Hägglund (2004). For evaluating the deviations from piecewise MO transients usual in FOTD control, two modified relative total variance measures (Skogestad, 2003) will be used in this paper. For the setpoint steps, the deviations of an output \( y_s(t) \) from the strict monotonicity may be expressed in terms of a

\[ TV_0(y_s) = \left( \sum |y_s(t) - y_s(t-1)| \right) \]  

FrCT4.4

© 2018 International Federation of Automatic Control
\[ TV_0(ys) = \sum_i |ys_{i+1} - ys_i| - |ys_\infty - ys_0| \] (2)

From an inversion of the first order plant dynamics follows (see Theorem 1 in Huba (2013)) that to a MO output corresponds a one-pulse (1P) optimal control consisting of two MO intervals separated with an extreme value \( u_m \notin (u_0, u_\infty) \). The control effort exceeding over these two intervals MO changes may be expressed in terms of

\[ TV_1(u) = \sum_i |u_{i+1} - u_i| - |u_\infty - u_0| \] (3)

In a disturbance step response of a delay-free plant, at least over the time interval \( T_d \), the output MO increases and thus an ideal MO return to the original output value will take the 1P shape evaluated in terms of \( TV_1(\dot{y}_d) \).

In total, the tolerable shape related deviations from ideal shapes at the plant input and output and for the setpoint and disturbance steps represent shape related constraints that may be formulated in form of inequalities

\[ TV_0(ys) \leq \varepsilon_{ys}; \quad TV_1(\dot{y}_d) \leq \varepsilon_{gd} \]
\[ TV_1(u) \leq \varepsilon_{ud}; \quad TV_1(\dot{u}) \leq \varepsilon_{ud} \] (4)

Based on the technology requirements, they may be specified by a vector of positive numbers \( \varepsilon_{ys}, \varepsilon_{gd}, \varepsilon_{us}, \varepsilon_{ud} \) with the index “s” corresponding to the setpoint and “d” corresponding to the disturbance response. For the sake of simplicity, as MO disturbance responses will be denoted those characterized by a MO return to the reference value, i.e. those with \( TV_1(\dot{y}_d) = 0 \).

This paper aims to discuss the pros and cons of the traditional and these newer alternative performance measures and by simple examples to illustrate situations when the traditional sensitivity constraints and optimization methods do not represent optimal solutions.

2. P CONTROLLER FOR FOTD PLANT CONTROL

In order to keep the task formulation and the corresponding discussion as simple as possible, in illustrating the considered control design, the first order time delayed (FOTD) plant models will be used

\[ S(s) = \frac{Y(s)}{U(s)} = \frac{Ke^{-Ta_s}}{s + a}; \quad G = \frac{K_s}{a} \] (5)

for \( a \neq 0 \); \[ S(s) = \frac{Ke^{-Ta_s}}{T_s + 1}; \quad T = \frac{1}{a}; \quad G = \frac{K_s}{a} \]

According to O’Dwyer (2009), in the PID control design they represent the most frequently used plant models.

2.1 P control for a delay-free plant

For a piecewise constant setpoints values \( r \), input and output disturbances \( d_i \), \( d_o \) and a delay-free plant, a control error decreasing with the time constant \( T_r \) specified by the setpoint-to-output relation

\[ F_{ry}(s) = \frac{Y(s)}{R(s)} = \frac{1}{T_r s + 1} \] (6)

may be achieved by a P control

\[ u = K_P e + u_\infty \]
\[ e = r - y \]
\[ K_P = (1/T_r - a)/K_s \]
\[ u_\infty = K_r (r - d_o) - d_i \] (7)

The static feedforward control \( u_\infty \) keeps the output at the required reference value \( r \) in steady states. A closed loop with the controller (7) remains stable up to the moment, when its pole \( s = -1/T_r \) remains negative and

\[ 1/T_r = K_P K_s + a > 0 \] (8)

For stable and marginally stable plants \( a \geq 0 \) this holds for any \( K_P K_s > 0 \) and stability will be satisfied by any \( 0 < T_r \). For unstable plants \( a < 0 \) \( K_P \) cannot be arbitrarily decreased (the time constant \( T_r \) in (7) cannot be arbitrarily increased), just to a value fulfilling (8).

2.2 Closed loop with a dead time

The setpoint to output transfer function of the loop with a dead time in the feedforward path may be derived as

\[ F_{ry}(s) = \frac{Y(s)}{R(s)} = \frac{(K_P K_s + a)}{(s + a)e^{Ts} + K_P K_s} \] (9)

2.3 Loop stability

Borders of the stability area corresponding to the characteristic quasi-polynomial

\[ A(s) = (s + a)e^{Ts} + K_P K_s \] (10)

may be derived by the parameter space method (Ackermann, 2002) yielding the critical gains corresponding to permanent oscillations (with some critical poles at the imaginary axis). By substituting \( sT_d = j\omega T_d = j\tau_d \) into (10) and introducing dimensionless variables

\[ K = K_P K_s T_d; \quad \tau_a = \omega T_d; \quad A = a T_d \] (11)

two real equations

\[ Ke^{-j\tau_a} + j\tau_d + A = 0 \]

For \( \omega = 0 \), when \( \tau_d = 0 \), one gets the lower critical gain

\[ K_{min} = -A = -a T_d \] (12)

The upper critical gain may be derived in a form of parametric equations

\[ K_{max} = \frac{\tau_a}{\sin\tau_d}; \quad A = -\frac{a T_d}{\sin\tau_d} \] (15)

In the limit for \( a \to 0 \), i.e. for integral plants, one gets then

\[ K_{min} = 0; \quad K_{max} = \pi/2 \] (16)

2.4 The double-real-dominant-pole tuning for a delay \( T_d \)

For an optimal controller tuning, the MRDP may be used (Huba, 2013). From (10) the double real dominant pole fulfilling the requirements \( A(s_o) = 0 \) and \( A(s_o) = 0 \) is

\[ s_o = \frac{1}{2} \theta T_d/T \] (17)

The control loop remains stable just for \( s_o < 0 \), i.e. for

\[ A = a T_d > -1 \] (18)

The corresponding controller gain is

\[ K_P = e^{-(1+a T_d)/2}K_s T_d \] (19)

It is to note that for unstable plants this gain increases (Fig. 1) up to \( K \to 1 \) for \( A \to -1 \), whereas for stable plants it converges to zero. Already for \( A = a T_d \) \( T_d/T = 2 \) the normed proportional gain \( K = K_P K_s T_d \approx 0.0498 \) may be considered as negligible and other more effective controllers should be applied. It is also obvious that the P control may not be used up to \( A \to -1 \).
Fig. 1. Optimal, lower and upper critical gains and the maximal $M_s$ and $M_t$ levels

2.5 The sensitivity functions

The role of the sensitivity functions in describing impact of the loop uncertainties is well known. When enhancing the parameter space $(A, K)$ for $A = aT_d \in (-1, 1)$ by the sensitivity and complementary sensitivity functions

$$M_s = \max \left\{ \left| \frac{1 + L(s)\omega}{L(s)\omega} \right| : \omega \geq 0; \ L(s) = K_pS(s) \right\}$$

and

$$M_t = \max \left\{ \left| \frac{L(s)}{1 + L(s)} \right| : \omega \geq 0 \right\}$$

(20)

corresponding to (19) (Fig. 1), it is to see that close to the stability border the sensitivity peaks converge to infinity and over the stability area they may take values significantly exceeding the textbook recommendations.

2.6 Questions

At this moment we have to ask:

(1) How to interpret the assertion “typical/recommended values are within the range 1.2-2.0”? Does it mean that it is allowed to deal just with situations when $A \geq -0.3$?

(2) What will be the impact of some particular sensitivity constraints chosen, for example as in PI control by Mercader and Banos (2017)

$$M_s = 1.7, M_t = 1.3$$

(21)

Which arguments may justify such a choice?

(3) When they write “higher values of $M_s$ and $M_t$ lead to oscillatory responses to step disturbances and higher overshoots for step references” - is it really not possible to get MO responses for higher $M_s$ and $M_t$ values?

Is it, for example, possible to work with a controller gain corresponding to $M_s = 4$ and $A \approx -0.45$ and still achieve a smooth MO output and a IP input?

(4) How to solve the problem with a parameter $a$ varying in a broader range with possibly negative values?

(5) With respect to the minimum values of the sensitivity functions, could we expect the optimum position of a working point specified by the controller gain at the floor of the “valleys” in Fig. 1 which do not coincide?

Fig. 2. $TV_s(u_s), K=1$ and $TV_s(y_s), K=1$ identified for different $\epsilon$ by the PP method and the analytical borders (14), (19) (dotted); below the corresponding $IAE(y_s)$ values
Do the “optimal” sensitivity values depend on the plant and controller type?

Remark 1. Besides these questions related to sensitivity constraints there are also questions regarding the MRDP method. Does the gain (19) derived by the double real dominant pole method represent values optimal with respect to the speed and the shape related constraints (4) specified, for example, with \( \epsilon = \epsilon_{ys} = \epsilon_{as} \rightarrow 0 \)?

3. PERFORMANCE PORTRAIT APPLICATION

Verification of the impact of the above derived analytical results with respect to the cost function (1) and the shape related constraints (4) may be accomplished by the performance portrait method (PPM), which consists of

- generating a performance portrait (PP) of the given loop over a grid of dimensionless parameters and
- choosing the controller parameters satisfying the particular specification on the transient responses.

Mercader and Banos (2017) denoted this method as a “brute force approach”. Although an inspection of a controller performance in all possible situations should be compulsory for all design methods, they have forgotten two other important facts:

- Once generated, the PP may be repeatedly used for any covered optimization constraints.
- This method does not face convergence problems typically occurring in the search for the absolute extreme by the traditional optimization methods.

Furthermore, also the optimization by the well-known methods as the M-constrained integral gain optimization (MIGO) requires “straightforward but tedious calculations”. And once the optimization constraints change, it has to be completely repeated, while the PPM requires only a new data search with changed criteria.

3.1 PP specification

By the dimensionless parameters (11) the closed loop (9) may be expressed as

\[
F_{rs}(p) = \frac{(K + A)}{(p + A)e^p + K}
\]

(22)

To demonstrate relations of the analytically derived curves (14) and (19) to the shape related deviations at the input and output, the PP will be generated for \( A \in [-0.5, 0.5] \) and \( K \in [0,1] \) over grid of 41 x 41 points. To consider influence of the feedforward coefficient \( K_r \), the third dimension may be added defined by the set of considered values \( K_{r,k} \), \( k = 1,2,... \). However, because we wish to limit this analysis just to the shape related deviations at the input and output, whereas a steady state error will not be considered, it is enough to calculate the PP for a single value of \( K_r \).

In all considered points of the grid, the setpoint step responses will be calculated with an integration step \( T_s = 0.001 \) and with the simulation time \( t_{sim} = 100 \). Thereby, for the sake of simplicity, the choice of the considered performance measures will be limited to \( IAE_y, TV_1(u_s) \) and \( TV_0(y_s) \). Obviously, this is not the method limitation (Mercader and Banos, 2017). When required, this choice may be extended by such performance measures, as the settling time, maximal overshooting, peak time, etc. Although, a certain estimate of the maximum overshooting is already provided by the shape deviations themselves.

Evaluation of the shape related deviations at the plant input and output, which always correspond to some \( \epsilon > 0 \), shows a slightly broader MO areas than in case of the analytical derivation (Huba, 2013). However, especially for unstable plants with \( A \rightarrow -1 \), due to the numerical imperfections of the standard solvers in Matlab/Simulink, the calculated MO areas may be slightly narrower than the analytically derived ones (Fig. 2). Thus, the answer to the question from Remark 1 is that the curves corresponding to (19) and (14) may be considered as the MO borders. Thus, we may note the first contradiction to the \( M_r \)-constrained optimization, when we get for each \( A \) a unique optimal \( K_P \) yielding also a unique pair \( M_s, M_r \).

3.2 MO/1P Uncertainty Box

From (14) and (19) and from the requirement of zero shape related deviations at the plant input and output, for the controller tuning an inequality follows

\[-aT_d < K_P K_s T_d < e^{-(1+aT_d)}\]

(23)

In case of interval plant parameters

\[
K_s \in [K_{s,\min}, K_{s,\max}]
\]

\[
a \in [a_{\min}, a_{\max}]
\]

\[
T_d \in [T_{d,\min}, T_{d,\max}]
\]

such an inequality has to be fulfilled for any triplet \( K_s, a, T_d \). Since \( K = K_P K_s T_d \) represents an increasing function of \( K_s \) and \( T_d \) and the constraints (14) and (19) are decreasing functions of \( A = aT_d \), the limit situations correspond to \( a_{\min} T_{d,\max} \) and \( a_{\max} T_{d,\max} \), when (23) turns to

\[-a_{\min} T_{d,\max} < K_P K_{s,\min} T_{d,\max} < e^{-(1+a_{\max} T_{d,\max})}\]

(25)

It means that in the plane of the parameters \((aT_{d,\max}, K_P K_s T_{d,\max})\) the required properties hold in all points of an Uncertainty Box (UB)
\[ UB = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \]

\[ B_{11} = (a_{\text{min}T_d,\text{max}}, KP_k, K_{s,\text{max}} T_{d,\text{max}}) \]

\[ B_{12} = (a_{\text{max}T_d,\text{max}}, KP_k, K_{s,\text{max}} T_{d,\text{max}}) \]

\[ B_{21} = (a_{\text{min}T_d,\text{max}}, KP_k, K_{s,\text{min}} T_{d,\text{max}}) \]

\[ B_{22} = (a_{\text{max}T_d,\text{max}}, KP_k, K_{s,\text{min}} T_{d,\text{max}}) \]

From a comparison with Fig. 1 it is then obvious that

- in tasks with a dominant uncertainty of the parameter \( a \) the points \( B_{12} \) and \( B_{21} \) correspond to significantly different values of the sensitivity functions and
- in case of unstable systems, the sensitivity levels have to take much higher values than usually recommended, but without necessarily leading to oscillatory transients.

This makes the \( M_s \)-constrained optimization questionable.

### 3.3 More relaxed UB allocation

In control applications, a strict requirement of an input 1P-output MO behavior is not always advantageous. Such processes may become unnecessarily sluggish and, with respect to the stability area, such an UB is located asymmetrically close to the lower stability border. From this point of view it may become much more appropriate to shift the vertex \( B_{12} \) of the curve \( K_{\text{opt}} \) (to the right). Whereas Víteková and Vítek (2014) solve such a relaxed task (for a nominal tuning) by use of tables, PP methods allows to find an optimal UB location by a specification of \( c \).

**Remark 2.** Let us note that for the P control and \( A = 0 \) the double real dominant pole tuning yields \( M_s = 1.4 \).

Comparison of this frequently used figure stresses the importance of questions 2 and 6 in Section 2.6.

### 4. PI AND PID CONTROLLER APPLICATION

The example with the P controller has been motivated by the simplicity and transparency of all relevant considerations. An extension to a much more frequent situation with PI and PID controllers will be illustrated by the analytically tuned 2DOF PI and PID control derived by the MRDP (Víteková and Vítek, 2016; Huba, 2015; Huba and Bisták, 2016). It consists of a 1DOF controller \( C(s) \) and a prefilter \( F_p(s) \) with weighting coefficients \( b, c \).

\[
R(s) = K_c \frac{1 + T_s + T_D s^2}{T_s} ; \quad F_p(s) = \frac{cT_D s^2 + bT_s + 1}{T_D s^2 + T_s + 1} \quad (27)
\]

The increased number of loop coefficients makes this problem much more complex, which is especially to feel in visualization of the important results.

#### 4.1 2DOF PI control

The “optimal” triple real dominant poles \( s_o \) and the corresponding parameters \( K_{co} \) and \( T_{io} \) are given by dimensionless values \( K_o, A_o \), and \( \tau_o \) as

\[
s_o = \frac{-A_o(4 + S^2)}{24}; \quad A = aT_d; \quad S = \sqrt{A^2 + 8} \]

\[
K_o = K_{co}K_T = (S - 2)e^{(S - A - 4)/2} \]

\[
\tau_o = \frac{T_{io}}{T_d} = \frac{A^2 + 2A + 28 - (A + 10)S}{2} \quad (28)
\]

### 4.2 2DOF PID control

The “optimal” PID parameters \( K_{co}, T_{Do} \) and \( T_{io} \) are

\[
s_o = (S - 6 - A)/(2T_d); \quad A = aT_d; \quad S = \sqrt{A^2 + 12} \]

\[
K_o = K_{co}K_T = 0.5[S(A + 12) - (A^2 + 2A + 36)e^{(S - A - 6)/2}] \]

\[
T_{io} = \frac{2(36 + 2A + A^2 - (A + 12)S)}{A^3 + 12A^2 + 36A + 288 - (A^2 + 12A + 84)S} \]

\[
T_{Do} = \frac{A^3 + 12A^2 + 36A + 288 - (A^2 + 12A + 84)S}{2} \quad (30)
\]
Table 1. Sensitivity peaks of controllers tuned by the MRDP method for $A = 0$

<table>
<thead>
<tr>
<th>Controller</th>
<th>$M_s$</th>
<th>$M_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>1.4</td>
<td>1</td>
</tr>
<tr>
<td>PI</td>
<td>1.7</td>
<td>1.44</td>
</tr>
<tr>
<td>PID</td>
<td>2</td>
<td>1.6</td>
</tr>
</tbody>
</table>

The dominant pole $s_o$ remains negative for $A \geq -2$. Optimal $b, c$ cancel two of the dominant poles $s_o$

$$b_o = \frac{2A^2 + 12A^2 + 36A + 288 - (A^2 + 12A + 84)s}{[A^2 + 2A + 36 - S(A + 12)](A + 6 - S)}$$

$$c_o = \frac{A^2 + 12A^2 + 36A + 288 - (A^2 + 12A + 84)s}{S - 2)(A + 6 - S)} \quad (31)$$

5. DISCUSSION

Comparison of Tab. 1 including $M_s$ and $M_t$ corresponding to controllers tuned by the MRDP method for $A = 0$ with the usual textbook recommendations shows that there are no optimal sensitivity levels, just levels corresponding to optimally tuned controllers with $A = 0$. With an increasing controller complexity also the sensitivity peaks increase. However, it does not mean that the PID control for unstable plants will cause deeper problems than, for example, an application of the PI control. Rather vice versa - the PID control doubles the achievable stability range from $A = -1$ for PI control up to $A = -2$.

By redrawing Fig. 4 into 5 we may show another interesting property of the sensitivity based tuning and a new interpretation of the whole matter: The sensitivity based robust control yields results reliably locally around $A \approx 0$. Until the ratio $T_d/T$ does not exceed a value from the range of $0.05 - 0.1$, similar sensitivities correspond both to the “precisely” tuned controllers with $A \neq 0$, and their simpler approximations based on the IPDT model corresponding to $A = 0$. This shows limitations of the sensitivity based tuning and, at the same time, also its relation to numerous other approaches based on IPDT models as, for example, the method by Ziegler and Nichols (1942), the model free control (Fliess and Join, 2013), ADRC (Gao, 2014), or the method considered in Mercader and Banos (2017).

6. CONCLUSIONS

The paper results may be summarized as follows:

1. The levels of the sensitivity functions peaks in the considered FOTD plant control dominantly depend on the parameter $A = aT_d$. Especially when dealing with unstable plants the usual textbook requirements on the choice of $M_s$ and $M_t$ seem to be not acceptable.
2. In PI control, the limits (21) preferred in Mercader and Banos (2017) may be shown to correspond also to the triple real dominant pole tuning for $A = 0$.
3. Similar properties may be shown also for other controllers. However, the “optimal” sensitivity values corresponding to $A = 0$, which may then be applied also to some $A \neq 0$, depend on the controller type.
4. Reliable results for higher uncertainties of $A$, especially when dealing with unstable plants, may be derived by the performance portrait method using the shape related deviations from ideal transients at the plant input and output. This method shows to be not only preciser, but also more flexible and cost effective than the traditional optimization methods.

REFERENCES


