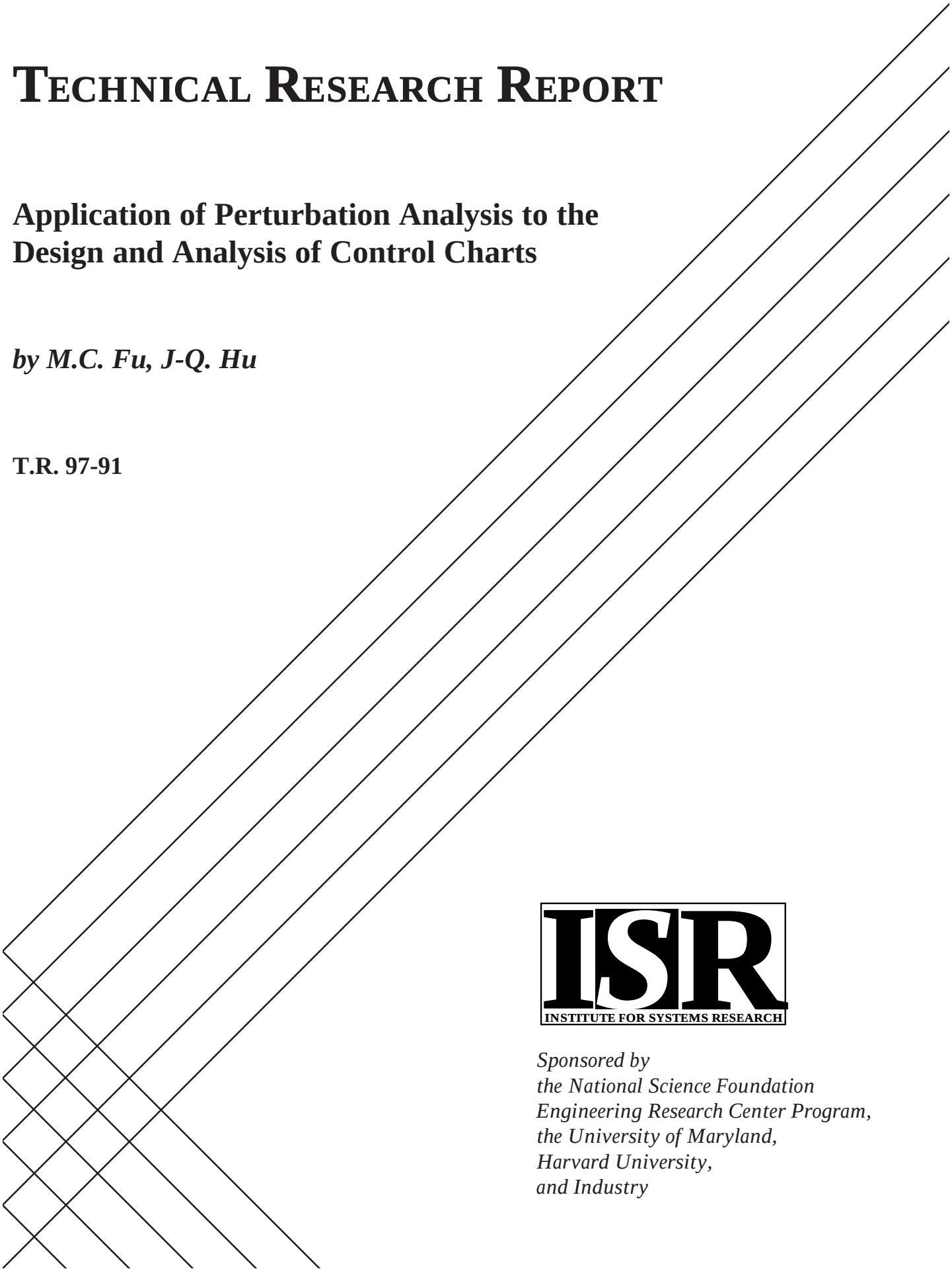
TECHNICAL RESEARCH REPORT

Application of Perturbation Analysis to the Design and Analysis of Control Charts

by M.C. Fu, J-Q. Hu

T.R. 97-91



Sponsored by
the National Science Foundation
Engineering Research Center Program,
the University of Maryland,
Harvard University,
and Industry

Application of Perturbation Analysis to the Design and Analysis of Control Charts

Michael C. Fu[1]

College of Business and Management
Institute for Systems Research
University of Maryland
College Park, MD 20742
mfu@umd5.umd.edu
fax: (301) 314-9157

Jian-Qiang Hu[2]

Department of Manufacturing Engineering
Boston University
Boston, MA 02215
hu@enga.bu.edu
fax: (617) 353-5548

March 1997

Abstract

The design of control charts in statistical quality control addresses the optimal selection of the design parameters such as the sampling frequency and the control limits; and includes sensitivity analysis with respect to system parameters such as the various process parameters and the economic costs of sampling. The advent of more complicated control chart schemes has necessitated the use of Monte Carlo simulation in the design process, particularly in the evaluation of performance measures such as average run length. In this paper, we apply perturbation analysis to derive gradient estimators that can be used in gradient-based optimization algorithms and in sensitivity analysis when Monte Carlo simulation is employed. We illustrate the technique on a simple Shewhart control chart and on a more complicated control chart that includes the exponentially-weighted moving average control chart as a special case.

Keywords: perturbation analysis, Monte Carlo simulation, statistical quality control, control charts, average run length, sensitivity analysis, economic design problem

[1]M.C. Fu is supported in part by the National Science Foundation under Grant No. NSF EEC-9402384 and by a 1996 Summer Research Grant from the Maryland Business School.

[2]J.Q. Hu is supported in part by the National Science Foundation under Grant Nos. EEC-9527422 and DDM-9212368.

1 Introduction

Two critical issues that must be addressed in the design of control charts in statistical quality control are the *optimal* selection of the design parameters such as sample size, sampling frequency, and control limits; and *sensitivity analysis* with respect to system parameters such as the economic costs of sampling and the characteristics of the potential process shifts (cf. Montgomery 1996). Depending on the design approach, the performance measures of interest fall into two main types: average run lengths or expected economic costs. The increasing complexity of many of the recently proposed control charts in the research literature has led to analytically intractable models, so Monte Carlo simulation is routinely used to estimate performance. Examples include Grimshaw and Alt (1997), where control charts for quantile function values are proposed; Albin et al. (1997), where a number of different control charts are compared in their average run length to false alarms and to detection of process mean and standard deviation shifts; Baxley (1995), where variable sampling interval control charts are applied; Seppala et al. (1995), where subgroup bootstrap and parametric methods for determining process control limits are compared; and Gan (1995), where the performance of control charts for joint monitoring of a process mean and variance are evaluated and compared.

The generality of Monte Carlo simulation makes it a popular tool, since it allows the modeller to be quite flexible. Among the clearest advantages are the following:

- Assumptions on process characteristics can be relaxed (e.g., normality, independence, and stationarity assumptions). For example, in Grimshaw and Alt (1997), a control chart is derived under a (customary in the literature) large sample approximation invoking the Central Limit Theorem. However, as they point out, in practice relatively small sample sizes are used. In their small test example, the simulation estimate of in-control average run length was 147.6, compared to a theoretical value, based on the large sample approximation, of 200.

- Any control chart can be handled, including Shewhart, Cumulative Sum (CUSUM), Exponentially Weighted Moving Averages (EWMA), and Bayesian. In comparing various control charts, Albin et al. (1997) "chose to use simulation. Essentially one program (with less than fifty lines of code) is used for all combinations of charts and run rules that we consider."

- An economic cost model can be made as general as desired. Barish and Hauser (1963) applied Monte Carlo simulation to test various combinations of parameters in an economic cost control chart design.

On the other hand, when it comes to sensitivity analysis and optimal design of control charts, the use of Monte Carlo simulation has been limited to "brute force" application. In other words, sensitivity analysis is conducted by changing the value of the parameter of interest and re-running the simulation, and optimization is carried out in a somewhat ad hoc trial-and-error manner, i.e., no formal optimization techniques are employed. The primary goal of this paper can be stated as follows:

> To introduce the use of Monte Carlo simulation *gradient estimation techniques* to the design and analysis of control charts.

The two most commonly used such techniques are perturbation analysis (PA) and the likelihood ratio method. Monographs for the former are Ho and Cao (1991), Glasserman (1991), Cao (1994), and Fu and Hu (1997), and for the latter is Rubinstein and Shapiro (1993). These methods have been applied predominantly to queueing and inventory models. This work represents the first application of PA to statistical quality control. Advantages inherent in applying these techniques when using Monte Carlo simulation include the following:

- the implementation of the estimators requires very little additional overhead in the simulation;
- the estimation is computationally efficient compared to the multiple runs that would be needed to construct finite difference estimates for each parameter of interest;
- the estimators have lower variance (generally) than naive finite difference estimates;
- the optimal design problem can be addressed using gradient-based algorithms (cf. Fu 1994).

Specifically, in this paper we derive sensitivity estimates for average run lengths with respect to different types of parameters: the control limits, the sampling frequency, and various process shift parameters. The rest of the paper is organized as follows. In Section 2, we introduce the problem setting with the requisite notation and briefly discuss the gradient estimation technique to be applied. In Section 3, we consider the standard in-control and out-of-control average run length performance measures and derive sensitivity estimators with respect to the control limits. Both the simple Shewhart chart and a more general chart that includes the EWMA control chart are addressed. In Section 4, we incorporate the dynamics of the process shift and derive sensitivity estimators of average run length. In addition to considering the control limit parameters, we also derive estimators with respect to the sampling frequency and various process shift parameters. Section 5 concludes with a summary, extensions to other performance measures such as average time to signal, and a discussion of avenues for further research, including the economic design problem.

2 Problem Setting

We consider the standard control chart setting involving a single measurable process variable with two distinct states called "in control" and "out of control." Samples of the process are taken at regularly spaced intervals and a test statistic generated (possibly based on past samples, as well). The test statistic is compared with control limits (that may vary as a function of time, as well) to declare the process in control or out of control. We begin by defining the following notation:

$$
\begin{aligned}
h &= \text{sampling interval, i.e., samples are taken every } h \text{ time units,} \\
n &= \text{sample size,} \\
F_0 &= \text{sampling process c.d.f. (with p.d.f. } f_0\text{) when in control,} \\
F_1 &= \text{sampling process c.d.f. (with p.d.f. } f_1\text{) when out of control,} \\
\mu_0 &= \text{in-control process mean,} \\
\mu_1 &= \text{out-of-control process mean,} \\
X_i &= \text{output from the } i\text{th sample, i.i.d. } F_0 \text{ or } F_1, \\
Y_i &= \text{test statistic after } i\text{th sample,} \\
\mathrm{LCL}_i &= \text{lower control limit for the } i\text{th test statistic,} \\
\mathrm{UCL}_i &= \text{upper control limit for the } i\text{th test statistic.}
\end{aligned}
$$

In words, samples of size n are taken every h units of time to generate $\{X_i\}$, from which the test statistic sequence $\{Y_i\}$ is derived. An out-of-control signal is declared if the test statistic Y_i falls outside of the interval defined by the lower and upper control limits $[\mathrm{LCL}_i, \mathrm{UCL}_i]$. The underlying process has an in-control c.d.f. F_0 with mean μ_0 and an out-of-control c.d.f. F_1 with mean μ_1. The sampling distributions F_0 and F_1 can be quite general, with standard distributions assumed (invoking the central limit theorem) being the normal distribution or the chi-squared distribution.

In this paper, we will assume that the test statistic generated by the control chart at the jth sampling has the following general form:

$$Y_i = \psi(X_i, Y_{i-1}), \tag{1}$$

where ψ is a function independent of other system parameters. Usually, $Y_1 = X_1$ is specified as the initial condition.

Example 1. Shewhart chart:

$$Y_i = X_i \text{ for all } j \geq 1.$$

Example 2. EWMA chart:

$$Y_i = \alpha X_i + (1-\alpha) Y_{i-1} \text{ for all } j \geq 2, \quad 0 < \alpha < 1, \quad Y_1 = X_1.$$

As described above, an out-of-control signal is declared when the test statistic falls outside of the specified control limits. The corresponding sample number is defined as the run length, which is the performance measure of interest:

$$L = \min\{i : Y_i \notin [\mathrm{LCL}_i, \mathrm{UCL}_i]\}. \tag{2}$$

The expectation of this stopping time for $\{Y_i\}$ is what is commonly known as the average run length (ARL). We will consider three forms of the ARL performance measure:

- in-control ARL;
- out-of-control ARL;
- ARL under process shift dynamics.

The in-control (out-of-control) ARL assumes that the process is in control (respectively, out of control) the entire time. Ideally, one wants long in-control run lengths and short out-of-control run lengths. The third type of ARL assumes that the process starts in control, but goes out of control at some later time. To be more precise, we introduce the following notation that characterizes the process shift dynamics:

$$\begin{aligned} T &= \text{(r.v.) time to go from } F_0 \text{ to } F_1, \\ F &= \text{c.d.f. (with p.d.f. } f\text{) for } T\text{, parametrized by } \mu \text{ (e.g., the mean)} \\ G_i &= \text{c.d.f. (with p.d.f. } g_i\text{) for } X_i \in \{F_0, F_1\}. \end{aligned}$$

Starting from a new in-control state, the process will go out of control after T units of time, where T is a random variable independent of $\{X_i\}$ with c.d.f. F, p.d.f. f, and parameter μ (e.g., the mean). Clearly, the

run length depends on T, and the notation $L(T)$ will be used to denote this explicit dependence. The event $\{hL < T\}$ indicates a false alarm. The underlying dynamics that drive the process out of control may be quite general, since Monte Carlo simulation is to be employed, e.g., T need not be exponentially distributed, as is assumed in most analytical models. The sampling distribution sequence $\{G_i\}$ forms a discrete-time stochastic process which takes on the "value" F_0 or F_1, depending on whether or not the process is in control.

Now we can define the three performance measures of interest:

$$\begin{aligned} \mathrm{ARL}(T) &= E[L(T)], \\ \mathrm{ARL}_0 &= E[L|T = \infty], \\ \mathrm{ARL}_1 &= E[L|T = 0]. \end{aligned}$$

ARL_0 is the in-control (on-target) ARL, and ARL_1 is the out-of-control (off-target) ARL. $\mathrm{ARL}(T)$ is the ARL for a process that starts in control, but shifts out of control at time T. $\mathrm{ARL}(T)$ is useful for control chart design based on economic costs; when there is no confusion, the T argument (or subscript) will be dropped. Furthermore, except where specified otherwise, the initial condition $Y_1 = X_1$ is implicit throughout.

Using perturbation analysis, we will derive estimators for sensitivities of the average run length

$$\frac{d\mathrm{ARL}_0}{d\theta}, \quad \frac{d\mathrm{ARL}_1}{d\theta}, \quad \text{and} \quad \frac{d\mathrm{ARL}(T)}{d\theta},$$

where θ will represent different types of parameters: the sampling frequency, control limit parameters, and various process parameters.

2.1 Perturbation Analysis

Perturbation analysis (PA) is a technique for gradient estimation that is particulary useful whenever Monte Carlo simulation is employed (cf. Ho and Cao 1991, Glasserman 1991, and Fu and Hu 1997). The simplest and generally most efficient technique is infinitesimal perturbation analysis (IPA). Intuitively, an IPA estimator is derived under the assumption that small changes in the parameter cause small changes in the performance measure. Technically speaking, a sufficient condition for the IPA estimator to be unbiased is that the sample performance measure be a.s. continuous with respect to the parameter. IPA is not applicable to our problem, because L is a discrete random variable, taking on integer values. The resulting L as a function of the parameter will be piecewise constant with jumps, and thus the resulting IPA estimator will be 0. For example, if the control limits are perturbed, the sample number at which the out-of-control signal is declared will not change if the perturbation is small enough. As the perturbation is increased, at some point the sample number at which the out-of-control signal is declared will change.

For the cases where IPA does not apply, there are various PA alternatives/extensions (see Fu and Hu 1997 for illustration of various forms on a simple random variable example). We apply the technique introduced by Gong and Ho (1987) that employs conditional Monte Carlo, known as smoothed perturbation analysis (SPA). The particular approach taken here is based on the general framework of Fu and Hu (1992), as presented in Fu and Hu (1997). The general form of the estimator contains two parts:

- The first part is simply the IPA estimator.
- The second part consists of the sum of conditional contributions representing the product of a probability jump rate and the resulting jump in the performance measure that is computed as a conditional expected difference on two sample paths (based on the original sample path) that are called the degenerated nominal path (DNP) and perturbed path (PP). The original sample path is called the nominal path (NP).

The first part is zero for our problem, as already discussed. The analysis in the next two sections focus on deriving the second part. As is standard in perturbation analysis, unbiasedness of the estimators is established by invoking the dominated convergence theorem.

3 In-Control and Out-of-Control ARLs

For the in-control and out-of-control ARLs, we consider derivative estimates with respect to the control limits. Throughout this section, we treat the two cases simultaneously by defining the two constants

$$\tau_0 = \infty, \quad \tau_1 = 0,$$

to correspond to the in-control and out-of-control cases, respectively. We can then use

$$T = \tau_j, \quad j = 0, 1,$$

to refer to the two cases concurrently, where $\{X_i\}$ would have corresponding p.d.f. f_j and c.d.f. F_j. For example, using this notation, we have $\mathrm{ARL}_j = E[L|T = \tau_j]$.

We consider the constant control limits case first:

$$l = \mathrm{LCL}_i, \quad u = \mathrm{UCL}_i.$$

If θ is a parameter in both LCL and UCL, as is typically the case, then the chain rule can be applied to obtain the sensitivities

$$\frac{dE[L]}{d\theta} = \frac{dE[L]}{du}\frac{du}{d\theta} + \frac{dE[L]}{dl}\frac{dl}{d\theta}.$$

As discussed in the previous section, the critical point to note is that if a change in the control does cause a change in the sample number in which the out-of-control signal is declared, this change is *finite* (here, integer-valued) and not infinitesimal; hence, the IPA estimator is biased. By conditioning, we now obtain an unbiased estimator that consists of terms that are a product of a probability rate of the change in the sample number multiplied by the expected difference in the performance measure due to the change. As usual, we will refer to the original path as the nominal path, and to the path under the change as the perturbed path.

Under the framework of Fu and Hu (1997), we have the following:

$$\begin{aligned}\frac{dE[L]}{d\theta} = \lim_{\Delta\theta\to 0}\Bigg[&E\left[\frac{L(\theta+\Delta\theta)-L(\theta)}{\Delta\theta}\,\middle|\,\mathcal{B}^c(\Delta\theta)\right]P(\mathcal{B}^c(\Delta\theta)) \\ &+ E\left[\frac{L(\theta+\Delta\theta)-L(\theta)}{\Delta\theta}\,\middle|\,\mathcal{B}(\Delta\theta)\right]P(\mathcal{B}(\Delta\theta))\Bigg]\end{aligned}$$

$$= \frac{dE[L \cdot \mathbf{1}(\mathcal{B}^c)]}{d\theta} + E\left[\lim_{\Delta\theta\to 0} E\left[L(\theta+\Delta\theta) - L(\theta)\right| z, \mathcal{B}(\Delta\theta)] \lim_{\Delta\theta\to 0} \frac{P(\mathcal{B}(\Delta\theta)|z)}{\Delta\theta}\right]$$
$$= E\left[\frac{dL}{d\theta}\right] + E\left[\left(E_z[L^{PP}] - E_z[L^{DNP}]\right)\frac{dP_z}{d\theta}\right],$$

where

$$\begin{aligned} E_z[L^{PP}] &= \lim_{\Delta\theta\to 0} E[L(\theta+\Delta\theta)|z, B(\Delta\theta)], \\ E_z[L^{DNP}] &= \lim_{\Delta\theta\to 0} E[L(\theta)|z, B(\Delta\theta)], \\ \frac{dP_z}{d\theta} &= \lim_{\Delta\theta\to 0} \frac{P(\mathcal{B}(\Delta\theta)|z)}{\Delta\theta}, \\ B(\Delta\theta) &= \{\omega : L(\theta+\Delta\theta) \neq L(\theta)\}, \end{aligned}$$

i.e., $B(\Delta\theta)$ gives the set of sample paths on which a perturbation $\Delta\theta$ causes the run length to change, $dP_z/d\theta$ is the probability "rate" at which this change takes place, and $E_z[L^{PP}]$ is the corresponding expected run length on those sample paths, in the limiting case as the perturbation becomes infinitesimal. The subscript z indicates a conditional expectation or probability on set z, called the characterization, to be selected.

For our problem, we have two simplifications:

1. As discussed earlier, the IPA term (first term) is zero.
2. $L^{DNP} = L$, i.e., the performance measure on DNP is equal to that found on the original sample path NP.

What remains to be carried out are the following:

1. Choosing the characterization z.
2. Calculating the probability rate term $\frac{dP_z}{d\theta}$.
3. Estimating $E_z[L^{PP}]$ (or $E_z[L^{PP} - L]$).

We derive the explicit estimator for the case where θ is u, as the case where θ is l is completely analogous. We consider both the left-hand and right-hand derivatives. For $\Delta u > 0$, it is obvious that $L(u+\Delta u) \neq L(u)$ if and only if $u < Y_{L(u)} \leq u + \Delta u$, i.e.,

$$E[L(u+\Delta u) - L(u)] = E[(L(u+\Delta u) - L(u))\mathbf{1}(u < Y_L \leq u + \Delta u)]$$

In this case, the perturbation Δu causes the test statistic to no longer signal out of control, and hence the run length is extended as a result (see Figure 1). In particular, we can think of the process as starting over with a new initial condition $\tilde{Y}_1 = \psi(X_1, Y_L)$, i.e., the additional length is equal to

$$E[L|T = \tau_j, \tilde{Y}_1 = \psi(X_1, Y_L)],$$

where the tilde notation is used to distinguish it from the original process.

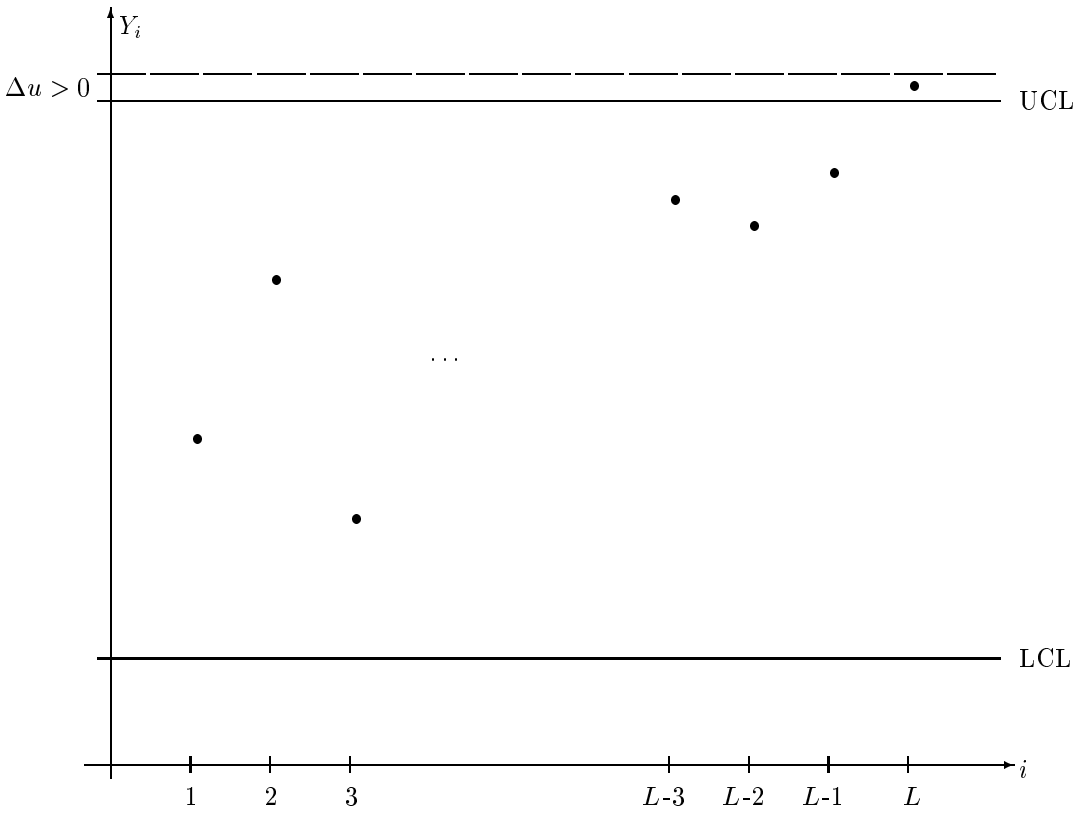


Figure 1: Extension of run length caused by positive perturbation in upper control limit.

On the other hand, letting $z = (L, X_1, \ldots, X_{L-1})$, we have

$$\begin{aligned} & E[(L(u+\Delta u) - L(u))\mathbf{1}(u < Y_L \le u + \Delta u)] \\ &= E[E[(L(u+\Delta u) - L(u))\mathbf{1}(u < Y_L \le u + \Delta u)|z]] \\ &= E\left[E[(L(u+\Delta u) - L(u))|z, u < Y_L \le u + \Delta u]\mathbf{1}(Y_L > u)P(u < Y_L \le u + \Delta u|z, Y_L > u)\right]. \end{aligned}$$

Note that in the last equation, we introduce the condition $Y_L > u$ so that our estimator will be simpler; otherwise, the implicit condition $Y_L \notin [l, u]$ would have to be used in calculating $P(u < Y_L \le u + \Delta u|z)$, resulting in a more complicated estimator.

Therefore, we have $P_z(\mathcal{B}(\Delta u)) = P(u < Y_L \le u + \Delta u|z, Y_L > u)$, so the probability rate term for the change is calculated via

$$\frac{dP_z}{d\theta} = \lim_{\Delta u \to 0^+} \frac{P(u < Y_L \le u + \Delta u|z, Y_L > u)}{\Delta u} = \frac{f_j(\psi^{-1}(u, Y_{L-1}))}{1 - F_j(\psi^{-1}(u, Y_{L-1}))} \frac{d\psi^{-1}(u, Y_{L-1})}{du}.$$

where $\psi^{-1}(\cdot, \cdot)$ denotes the inverse with respect to the first argument, so that

$$X_i = \psi^{-1}(Y_i, Y_{i-1}).$$

For example, we have for the EWMA control chart:

$$\psi(x, y) = \alpha x + (1 - \alpha)y \quad \Longrightarrow \quad \psi^{-1}(w, y) = (w - (1 - \alpha)y)/\alpha,$$

$$X_i = \frac{Y_i - (1-\alpha)Y_{i-1}}{\alpha}.$$

Also,

$$\begin{aligned}
&\lim_{\Delta u \to 0^+} E[(L(u+\Delta u) - L(u))|z, u < Y_L \le u + \Delta u] \\
&\qquad = \lim_{\Delta u \to 0^+} E[L|T = \tau_j, \tilde{Y_1} = \psi(X_1, Y_L), u < Y_L \le u + \Delta u] \\
&\qquad = E[L|T = \tau_j, \tilde{Y_1} = \psi(X_1, u)].
\end{aligned}$$

Thus, our final right-hand estimator for $d\mathrm{ARL}_j/du,\ j = 0, 1$, is the following:

$$\frac{f_j(\psi^{-1}(u, Y_{L-1}))}{1 - F_j(\psi^{-1}(u, Y_{L-1}))} \frac{d\psi^{-1}(u, Y_{L-1})}{du} E[L|T = \tau_j, \tilde{Y_1} = \psi(X_1, u)]\mathbf{1}\{Y_L > u\}. \tag{3}$$

For unbiasedness, we need

$$\begin{aligned}
&\lim_{\Delta u \to 0} \frac{1}{\Delta u} E[L(u+\Delta u) - L(u)] \\
&\quad = E\left[\lim_{\Delta u \to 0} E[(L(u+\Delta u) - L(u))|z, u < Y_L \le u + \Delta u]\mathbf{1}(Y_L > u) \frac{1}{\Delta u} P(u < Y_L \le u + \Delta u | z, Y_L > u)\right].
\end{aligned}$$

As usual, we use the dominated convergence theorem to establish this. Basically, to apply the dominated convergence theorem the key condition required is the bound

$$E\left[\sup_{0 \le \Delta u \le \epsilon} E[L|T = \tau_j, \tilde{Y_1} = \psi(X_1, u + \Delta u)]\mathbf{1}(Y_L > u) \frac{1}{\Delta u} P(u < Y_L \le u + \Delta u | z, Y_L > u)\right] < \infty, \tag{4}$$

for any $\epsilon > 0$. The following conditions suffice to establish (4):

(A1) $\psi(\cdot, \cdot)$ is continuously differentiable and strictly increasing with respect to its first argument.

(A2) $\psi^{-1}(\cdot, \cdot)$ is a decreasing function with respect to its second argument, $\left|\frac{d\psi^{-1}(x, \cdot)}{dx}\right| < K$ for all x, where $K > 0$ is a constant, and $F_j(\psi^{-1}(u, l)) < 1$, for $j = 0, 1$.

(A3) $|f_j(x)| < K$ for all $x,\ j = 0, 1$.

(A4) $E[L|T = \tau_j, \tilde{Y_1} = \psi(X_1, u + \Delta u)] < K$, for $0 \le \Delta u \le \epsilon,\ j = 0, 1$.

The conditions on ψ in (A1) and (A2) are mild. For example, it can easily be shown that the EWMA contol chart satisfies them. Condition (A3) is easily verifiable and holds for most of the well-known distributions. On the other hand, (A4) is a technical condition not as straightforward to verify as the other conditions, but viewed as a bound on average run length, it is reasonable to assume it holds for systems of practical interest.

Theorem 1. Under (A1)–(A4), (3) is an unbiased estimator for $d\mathrm{ARL}_j/du,\ j = 0, 1$.

Proof. In the estimator (3), the existence of $\psi^{-1}(\cdot, \cdot)$ and its differentiability are guaranteed by (A1). To establish (4), we note that due to (A4), we only need to establish a bound on

$$\left|\frac{1}{\Delta u} P(u < Y_L \le u + \Delta u | z, Y_L > u)\right|.$$

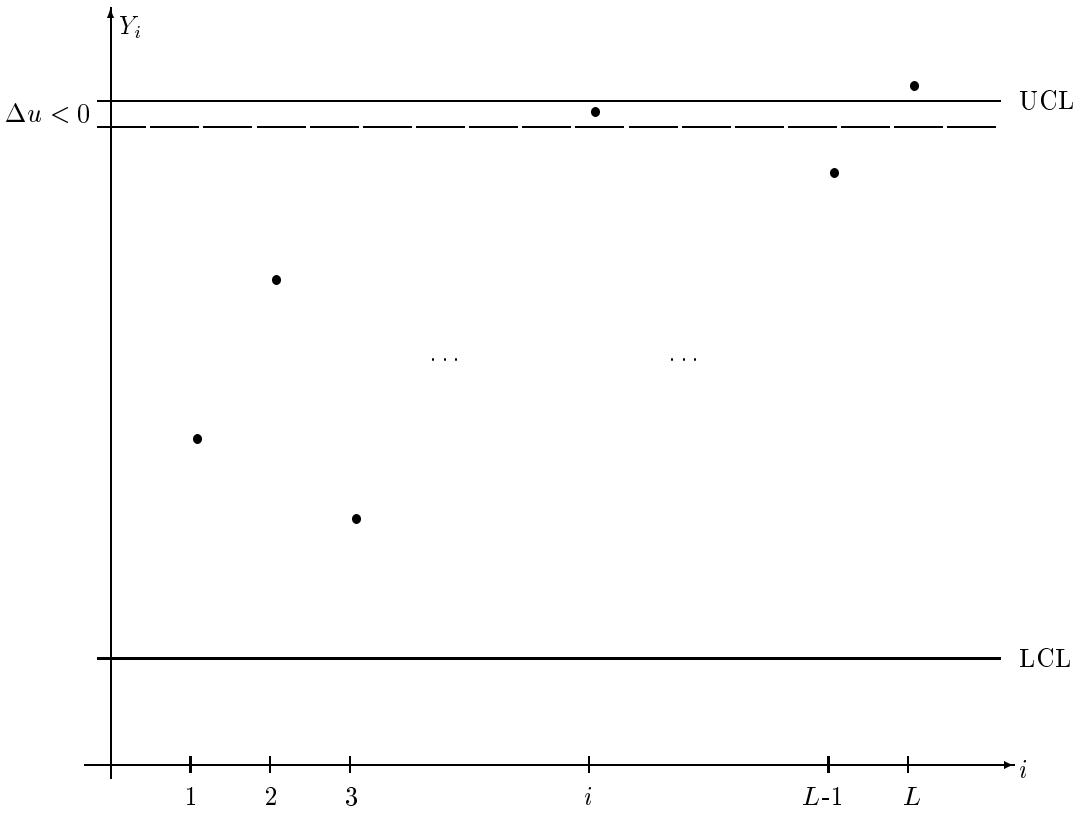


Figure 2: Shortening of run length caused by negative perturbation in upper control limit.

Since $\psi^{-1}(\cdot,\cdot)$ is decreasing with respect to its second argument due to (A2) and $l \le Y_{L-1} \le u$, we have

$$\begin{aligned}
&\left|\frac{1}{\Delta u}P(u < Y_L \le u+\Delta u | z, Y_L > u)\right| \\
&\quad = \left|\frac{1}{\Delta u}\frac{F_j(\psi^{-1}(u+\Delta u, Y_{L-1})) - F_j(\psi^{-1}(u, Y_{L-1}))}{1 - F_j(\psi^{-1}(u, Y_{L-1}))}\right| \\
&\quad \le \frac{1}{\Delta u}\frac{K|\psi^{-1}(u+\Delta u, Y_{L-1}) - \psi^{-1}(u, Y_{L-1})|}{1 - F_j(\psi^{-1}(u,l))} \quad \text{via (A2) and (A3)} \\
&\quad \le \frac{K^2}{1 - F_j(\psi^{-1}(u,l))} \quad \text{via (A2).}
\end{aligned}$$

This completes the proof. □

For the left-hand derivative $\Delta u < 0$, we know $L(u+\Delta u) \neq L(u)$ if $u + \Delta u < Y_i \le u$ for some $i < L$. In this case, we have a larger set of possible changes, in that any in-control signal prior to the first out-of-control may be altered to out of control, thus shortening the run length to that point (see Figure 2). If such a change occurs for sample i, the run length is reduced to i, and we have

$$\lim_{\Delta u \to 0^-} E[(L(u+\Delta u) - L(u))|z, u+\Delta u < Y_i \le u] = i - L.$$

For each term i, we define the characterization as the set of all sample information except X_i itself:

$$z_i = \{L, X_1, ..., X_L\} \backslash \{X_i\},$$

so that the probability rate term for the change is calculated via

$$\frac{dP_{z_i}}{d\theta} = \lim_{\Delta u \to 0^-} \frac{P(Y_i > u + \Delta u | l \le Y_i \le u)}{\Delta u} = -\frac{f_j(\psi^{-1}(u, Y_{i-1}))}{F_j(\psi^{-1}(u, Y_{i-1})) - F_j(\psi^{-1}(l, Y_{i-1}))} \frac{d\psi^{-1}(u, Y_{i-1})}{du},$$

and the final left-hand estimator for $d\mathrm{ARL}_j/du,\ j = 0, 1$, is the following:

$$\sum_{i<L} \frac{f_j(\psi^{-1}(u, Y_{i-1}))}{F_j(\psi^{-1}(u, Y_{i-1})) - F_j(\psi^{-1}(l, Y_{i-1}))} \frac{d\psi^{-1}(u, Y_{i-1})}{du} (L - i). \tag{5}$$

To establish unbiasedness, conditions (A2) and (A4) must be altered to the following:

(A2′) $\left|\frac{d\psi^{-1}(x, \cdot)}{dx}\right| < K$ for all x, where $K > 0$ is a constant, and $\inf_{l \le y \le u} F_j(\psi^{-1}(u, y)) - F_j(\psi^{-1}(l, y)) > 0$, for $j = 0, 1$.

(A4′) $E[L^2 | T = \tau_j, \tilde{Y_1} = \psi(X_1, u + \Delta u)] < K$, for $0 \le \Delta u \le \epsilon$.

The squared term in (A4′) arises from the fact that the estimator (5) has a summation containing on the order of L terms.

Theorem 2. Under (A1), (A2′), (A3) and (A4′), (5) is an unbiased estimator for $d\mathrm{ARL}_j/du,\ j = 0, 1$.

Since the proof proceeds essentially along the same lines as in Theorem 1, it is omitted here.

For our two examples, we have the following unbiased estimators for $d\mathrm{ARL}_j/du,\ j = 0, 1$:

Example 1. Shewhart chart:

$$\frac{f_j(u)}{1 - F_j(u)} \mathbf{1}\{X_L > u\} \cdot L,$$

$$\sum_{i<L} \frac{f_j(u)}{F_j(u) - F_j(l)} (L - i) = \frac{f_j(u)}{F_j(u) - F_j(l)} \frac{L(L-1)}{2}.$$

Example 2. EWMA chart:

$$\frac{f_j((u - (1-\alpha)Y_{L-1})/\alpha)}{1 - F_j((u - (1-\alpha)Y_{L-1})/\alpha)} \frac{1}{\alpha} E[L | T = \tau_j, \tilde{Y_1} = \psi(X_1, u)] \mathbf{1}\{Y_L > u\},$$

$$\sum_{i<L} \frac{f_j((u - (1-\alpha)Y_{i-1})/\alpha)}{F_j((u - (1-\alpha)Y_{i-1})/\alpha) - F_j((l - (1-\alpha)Y_{i-1})/\alpha)} \frac{1}{\alpha} (L - i).$$

Next we consider the general case where the control limits are also indexed by the sample number, i.e., we have a sequence $\{[l_i, u_i]\}$. This can be used to represent the most general of control charts such as CUSUM charts, where the control limits increase (usually linearly). We can use the chain rule to find sensitivities as follows:

$$\frac{\partial E[L]}{\partial \theta} = \sum_i \frac{\partial E[L]}{\partial u_i} \frac{\partial u_i}{\partial \theta} + \sum_i \frac{\partial E[L]}{\partial l_i} \frac{\partial l_i}{\partial \theta},$$

so for example this includes the constant control limit chart as a special case.

For $\Delta u_i > 0$, we can only have a change if $L = i$, so similar to the previous analysis, our estimator for $d\mathrm{ARL}_j/du_i,\ j = 0, 1$, is given by

$$\frac{f_j(\psi^{-1}(u_i, Y_{L-1}))}{1 - F_j(\psi^{-1}(u_i, Y_{L-1}))} \frac{d\psi^{-1}(u_i, Y_{L-1})}{du} E[L | T = \tau_j, \tilde{Y_1} = \psi(X_1, u)] \mathbf{1}\{Y_L > u\} \mathbf{1}\{L = i\}.$$

Similarly, for $\Delta u_i < 0$, we have the estimator

$$\frac{f_j(\psi^{-1}(u_i, Y_{i-1}))}{F_j(\psi^{-1}(u_i, Y_{i-1})) - F_j(\psi^{-1}(l_i, Y_{i-1}))} \frac{d\psi^{-1}(u_i, Y_{i-1})}{du} (L - i)^+.$$

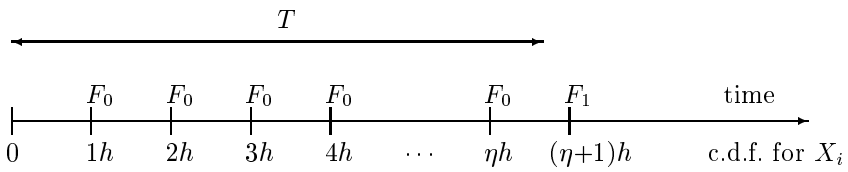


Figure 3: Relationship between η, T, and h.

4 Process Shift ARL

Now we consider the case where the process begins in control and goes out of control after a random time T. Thus, the distribution of each X_i depends on T, so that L also depends on T implicitly. The dependence of $\{X_i\}$ on T is not on the actual value of T, but just on which sampling interval it occurs. Therefore, we define the index random variable for the last in-control sample taken (refer to Figure 3)

$$\eta = \max\{j : hj < T\},$$

i.e.,

$$X_i \sim \begin{cases} F_0 & \text{for } i \leq \eta, \\ F_1 & \text{for } i > \eta. \end{cases}$$

In particular, if an i.i.d. sequence of random numbers $\{U_i\}$ is used to generate $\{X_i\}$ by inversion, we have

$$X_i = \begin{cases} F_0^{-1}(U_i) & \text{for } i \leq \eta, \\ F_1^{-1}(U_i) & \text{for } i > \eta. \end{cases}$$

There are two possibilities when an out-of-control signal is declared at epoch L:

1. the true state is out of control: $T \leq hL \iff \eta < L$;
2. the true state is actually in control (false alarm): $T > hL \iff \eta \geq L$.

4.1 Control Limits

This analysis proceeds very similarly to the in-control and out-of-control cases, with the following exceptions:

- the initial condition on the additional run length involves the residual time of T at the time of the out-of-control signal;
- the probability rate term is replaced with a random distribution G_i, equal to either F_1 or F_0 depending on whether the out-of-control signal is a true or false alarm, respectively.

We again consider both the left- and right-hand derivatives. For $\Delta u > 0$, we only need to consider the case $u < Y_L \leq u + \Delta u$, as before. Proceeding in the same manner, we obtain the estimator

$$\frac{g_L(\psi^{-1}(u, Y_{L-1}))}{1 - G_L(\psi^{-1}(u, Y_{L-1}))} \frac{d\psi^{-1}(u, Y_{L-1})}{du} E[L | T = \tau_{res}, \tilde{Y_1} = \psi(X_1, u)] \mathbf{1}\{Y_L > u\}.$$

where $\tau_{res} = (T - hL)^+$ is the residual time until the system actually goes out of control from the epoch at which an out-of-control signal is declared. Thus, if $T \leq hL$, the system is already out of control, and hence the residual time is zero.

For $\Delta u < 0$, considering $u + \Delta u < Y_i \leq u$ for each $i < L$ as before, we obtain

$$\sum_{i<L} \frac{g_i(\psi^{-1}(u, Y_{i-1}))}{G_i(\psi^{-1}(u, Y_{i-1})) - G_i(\psi^{-1}(l, Y_{i-1}))} \frac{d\psi^{-1}(u, Y_{i-1})}{du}(L - i).$$

Example 1. Shewhart chart:

$$\frac{g_L(u)}{1 - G_L(u)} E[L|T = \tau_{res}]\mathbf{1}\{X_L > u\},$$
$$\sum_{i<L} \frac{g_i(u)}{G_i(u) - G_i(l)}(L - i).$$

Example 2. EWMA chart:

$$\frac{g_L((u - (1-\alpha)Y_{L-1})/\alpha)}{1 - G_L((u - (1-\alpha)Y_{L-1})/\alpha)} \frac{1}{\alpha} E[L|T = \tau_{res}, \tilde{Y_1} = \psi(X_1, u)]\mathbf{1}\{Y_L > u\},$$
$$\sum_{i<L} \frac{g_i((u - (1-\alpha)Y_{i-1})/\alpha)}{G_i((u - (1-\alpha)Y_{i-1})/\alpha) - G_i((l - (1-\alpha)Y_{i-1})/\alpha)} \frac{1}{\alpha}(L - i).$$

Again, we can consider the general case where the control limits are also indexed by the sample number, $\{[l_i, u_i]\}$. Doing so, we obtain the following right-hand and left-hand estimators for $d\mathrm{ARL}_j/du_i$ $j = 0, 1$:

$$\frac{g_i(\psi^{-1}(u_i, Y_{L-1}))}{1 - G_i(\psi^{-1}(u_i, Y_{L-1}))} \frac{d\psi^{-1}(u_i, Y_{L-1})}{du} E[L|T = \tau_{res}, \tilde{Y_1} = \psi(X_1, u)]\mathbf{1}\{Y_L > u\}\mathbf{1}\{L = i\},$$

$$\frac{g_i(\psi^{-1}(u_i, Y_{i-1}))}{G_i(\psi^{-1}(u_i, Y_{i-1})) - G_i(\psi^{-1}(l_i, Y_{i-1}))} \frac{d\psi^{-1}(u, Y_{i-1})}{du}(L - i)^+.$$

One thing we note is that in this case there is a difference between simulation and on-line estimation. In particular, the random variable T would not be observable in the real system, implying that g_i and G_i would be unobservable; thus, the estimators can only be implemented in a simulation.

Finally, we point out that Theorems 1 and 2 still apply if the boundedness conditions (A4) and (A4′) hold for all $T = \tau_{res} \in (0, \infty)$.

4.2 Sampling Frequency

We now consider $\theta = h$, the sampling interval. As before, the critical point to note is that if a change in the sampling interval does cause a change in the sample number in which the out-of-control signal is declared this change is *finite* and not infinitesimal. We first consider the right-hand estimator $\Delta h > 0$. Writing $\eta(h)$ to denote the dependence of η on the sampling interval, we observe increasing the sampling interval may cause the interval in which the process actually goes out of control to decrease, specifically $\eta(h + \Delta h) = \eta(h) - 1$, which would imply that for $h + \Delta h$, we would have

$$X_i \sim \begin{cases} F_0 & \text{for } i \leq \eta - 1, \\ F_1 & \text{for } i > \eta - 1, \end{cases}$$

so that the distribution of X_η would change. This is the critical event order change. Then by conditioning on $z = \{\eta\} \cup \{X_i\}_{i=1}^{\infty}$, we will derive our estimator.

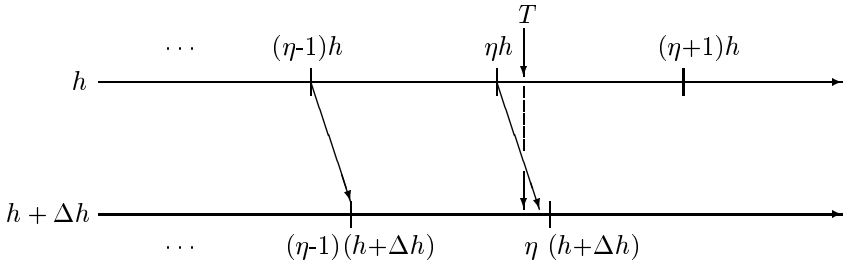


Figure 4: Potential change caused by perturbation $\Delta h > 0$.

We condition on $\eta(h) = i$. By defintion of $\eta(h)$, we have $hi < T < h(i+1)$. Based on the possible values of $\eta(h + \Delta h)$, we consider four cases:

$$\begin{aligned}
&\eta(h+\Delta h) > i && \mathbf{1}\{T > (h+\Delta h)(i+1)\} = 0 \text{ since } \Delta h > 0;\\
&\eta(h+\Delta h) < i-1 && \mathbf{1}\{T \le (h+\Delta h)(i-1)\} = 0 \text{ for } \Delta h < h/(i-1);\\
&\eta(h+\Delta h) = i && L(h+\Delta h) = L(h) \text{ if } (h+\Delta h)i < T \le (h+\Delta h)(i+1);\\
&\eta(h+\Delta h) = i-1 && \Longrightarrow (h+\Delta h)(i-1) \le T < (h+\Delta h)i.
\end{aligned}$$

Thus, we need only consider further the case $(h + \Delta h)(i - 1) \le T < (h + \Delta h)i$, depicted in Figure 4. In particular, we have shown

Lemma 1. For $0 < \Delta h < h/(\eta - 1)$,

$$\begin{aligned}
&E[(L(h+\Delta h) - L(h))|z]\\
&\quad = E[(L(h+\Delta h) - L(h)\mathbf{1}\{(h+\Delta h)(\eta-1) \le T < (h+\Delta h)\eta\}|z]\\
&\quad = E[(L(h+\Delta h) - L(h)|z, \mathbf{1}\{(h+\Delta h)(\eta-1) \le T < (h+\Delta h)\eta\}]\\
&\qquad\qquad \times E[\mathbf{1}\{(\eta-1)(h+\Delta h) \le T < \eta(h+\Delta h)\}|z].
\end{aligned}$$

The probability rate term in this case is calculated in the usual way by

$$\begin{aligned}
&P(\eta(h+\Delta h) = i-1|\eta(h) = i)\\
&\quad = P((i-1)(h+\Delta h) < T \le i(h+\Delta h)|ih < T \le (i+1)h)\\
&\quad = P(ih < T \le i(h+\Delta h)|ih < T \le (i+1)h) \qquad \text{for } 0 < \Delta h < h/i\\
&\quad = \frac{P(ih < T \le i(h+\Delta h)}{P(ih < T \le (i+1)h)}\\
&\quad = \frac{F(i(h+\Delta h)) - F(ih)}{F((i+1)h) - F(ih)}.
\end{aligned}$$

Thus, we have

$$\lim_{\Delta h \to 0} \frac{P(\eta(h+\Delta h) = i-1|\eta(h) = i)}{\Delta h} = \frac{i \cdot f(ih)}{F((i+1)h) - F(ih)},$$

and the final estimator is given by

$$\frac{\eta \cdot f(\eta h)}{F((\eta+1)h) - F(\eta h)} E_z\left[L^{PP} - L\right], \tag{6}$$

where the three paths are defined as follows:

- NP, the original sample path: $\eta h < T \leq (\eta+1)h$;
- DNP: $T = (\eta h)_+ \Longrightarrow \eta^{DNP} = \eta$;
- PP: $T = (\eta h)_- \Longrightarrow \eta^{PP} = \eta - 1$.

Since the run-length performance measure does not depend on the actual value of T but just on η, we again have $L^{DNP} = L$.

To establish the unbiasedness of our estimators, we impose the following assumptions:

(A5) $f(t)/(F(t_1) - F(t)) \leq K$, $\forall t_1 \geq t$ s.t. $F(t_1) > F(t)$.

(A6) $E[\sup_{h\in(h_{\min},h_{\max})} L(h)\eta] < \infty$, where $0 < h_{\min} < h_{\max} < \infty$.

First, we have

Lemma 2. Under (A5),

$$\frac{F(t+\Delta t) - F(t)}{F(t_1) - F(t)} \leq K\Delta t,$$

where $\Delta t \leq t_1 - t$.

Proof. Based on (A5), we have

$$\frac{d}{dt}(e^{Kt}F(t)) \leq Ke^{Kt}F(t_1).$$

Taking integration on the both sides of the above equation from t to $t+\Delta t$, we obtain

$$e^{K\Delta t}F(t+\Delta t) - F(t) \leq F(t_1)(e^{K\Delta t} - 1),$$

which leads to

$$\frac{F(t+\Delta t) - F(t)}{F(t_1) - F(t)} \leq 1 - e^{-K\Delta t} \leq K\Delta t.$$

□

Now we are ready to prove

Theorem 3. Under (A5)–(A6), (6) is an unbiased estimator for $dE[L]/dh$.

Proof. As usual, the proof is based on the dominated convergence theorem. Note that

$$\left|E[(L(h+\Delta h) - L(h))|z, \mathbf{1}\{(i-1)(h+\Delta h) \leq T < i(h+\Delta h)\}]\right| \leq 2 \sup_{h\in(h_{\min},h_{\max})} L(h),$$

and

$$\begin{aligned}
&\frac{1}{\Delta h}E[\mathbf{1}\{(h+\Delta h)(\eta-1) \leq T < (h+\Delta h)\eta\}|z] \\
&= \frac{1}{\Delta h}P((h+\Delta h)(h+\Delta h) \leq T < (h+\Delta h)\eta|h\eta < T \leq h(\eta+1)) \\
&= \frac{1}{\Delta h}P(\eta h < T \leq (h+\Delta h)|\eta h < T \leq (\eta+1)h)\mathbf{1}\{\Delta h < h/\eta\} \\
&\quad + \frac{1}{\Delta h}P(\eta h < T \leq (h+\Delta h)|\eta h < T \leq (\eta+1)h)\mathbf{1}\{\Delta h \geq h/\eta\}
\end{aligned}$$

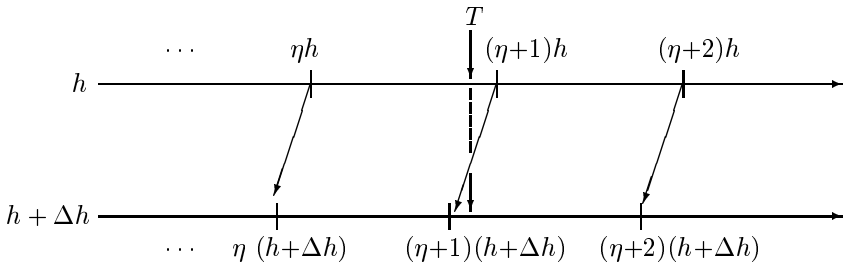


Figure 5: Potential change caused by perturbation $\Delta h < 0$.

$$\begin{aligned} &\leq \frac{1}{\Delta h}\frac{F(\eta(h+\Delta h)) - F(\eta h)}{F((\eta+1)h) - F(\eta h)} + \frac{1}{\Delta h}\mathbf{1}\{\Delta h \geq h/\eta\} \\ &\leq K + \eta/h \\ &\leq (K+1/h)\eta. \end{aligned}$$

Applying Lemma 1 and using (A6), we have

$$E\left[\sup_{\Delta h}\frac{1}{\Delta h}E[(L(h+\Delta h) - L(h))|z]\right] < \infty.$$

Then, applying the dominated convergence theorem,

$$\lim_{\Delta h\to 0}\frac{E[L(h+\Delta h)] - E[L(h)]}{\Delta h} = E\left[\lim_{\Delta h\to 0}\frac{1}{\Delta h}E[(L(h+\Delta h) - L(h))|z]\right].$$

□

We now turn to the left-hand derivative $\Delta h < 0$, where we consider $P(\eta(h+\Delta h) = i+1|\eta(h) = i)$, i.e., as depicted in Figure 5, we have

$$\begin{aligned} &P(\eta(h+\Delta h) = i+1|\eta(h) = i) \\ &\quad = P((i+1)(h+\Delta h) < T \leq (i+2)(h+\Delta h)|ih < T \leq (i+1)h) \\ &\quad = P((i+1)(h+\Delta h) < T \leq (i+1)h|ih < T \leq (i+1)h) \\ &\qquad\qquad \text{for } -h(i+2) < \Delta h < 0 \\ &\quad = \frac{P((i+1)(h+\Delta h) < T \leq (i+1)h}{P(ih < T \leq (i+1)h)} \\ &\quad = \frac{F((i+1)h) - F((i+1)(h+\Delta h))}{F((i+1)h) - F(ih)}. \end{aligned}$$

Thus, we have

$$\lim_{\Delta h\to 0}\frac{P(\eta(h+\Delta h) = i+1|\eta(h) = i)}{\Delta h} = -\frac{(i+1)\cdot f((i+1)h)}{F((i+1)h) - F(ih)},$$

and the final estimator is given by

$$\frac{(\eta+1)\cdot f((\eta+1)h)}{F((\eta+1)h) - F(\eta h)}E_z\left[L - L^{PP}\right], \tag{7}$$

where the three paths are defined as follows:

- NP, the original sample path, on which $\eta h < T \leq (\eta+1)h$;
- DNP: $T = ((\eta+1)h)_- \Longrightarrow \eta^{DNP} = \eta$;
- PP: $T = ((\eta+1)h)_+ \Longrightarrow \eta^{PP} = \eta + 1$.

Unbiasedness would again be established by applying the dominated convergence theorem. Since the proof proceeds essentially along the same lines as in Theorem 3, it is omitted here.

Theorem 4. Under (A5)–(A6), (7) is an unbiased estimator for $dE[L]/dh$.

4.3 Process "Drift" Parameters

Next, we consider sensitivities to parameters that enter the dynamics of the process going from in control to out of control. In particular, we consider the derivative with respect to μ, a parameter in the distribution of T. A little thought reveals that the resulting estimator is similar to the previous one, since the change occurs only in the timing of the samplings. The only differences are the following:

- $\Delta T < 0$ corresponds to $\Delta h > 0$ and $\Delta T > 0$ corresponds to $\Delta h < 0$;
- the probability rate term is slightly different, as computed below.

For $\Delta T < 0$, we have

$$\begin{aligned} P(\eta(T+\Delta T) = i-1 \mid \eta(T) = i) & \\ &= P((i-1)h < T + \Delta T \leq ih \mid ih < T \leq (i+1)h) \\ &= P(ih < T \leq ih - \Delta T \mid ih < T \leq (i+1)h) \\ &= \frac{F(ih - \Delta T) - F(ih)}{F((i+1)h) - F(ih)}. \end{aligned}$$

Thus, the left-hand derivative is given by

$$\frac{-f(\eta h)}{F((\eta+1)h) - F(\eta h)} \frac{dT}{d\mu} \left[L^{PP} - L\right], \tag{8}$$

where PP is the sample path on which $T = (\eta h)_- \Longrightarrow \eta^{PP} = \eta - 1$.

For $\Delta T > 0$, we have

$$\begin{aligned} P(\eta(T+\Delta T) = i+1 | \eta(T) = i) & \\ &= P((i+1)h < T + \Delta T \leq (i+2)h | ih < T \leq (i+1)h) \\ &= P(ih - \Delta T < T \leq (i+1)h | ih < T \leq (i+1)h) \\ &= \frac{F((i+1)h) - F((i+1)h - \Delta T)}{F((i+1)h) - F(ih)}, \end{aligned}$$

Thus, the right-hand derivative is given by

$$\frac{-f((\eta+1)h)}{F((\eta+1)h) - F(\eta h)} \frac{dT}{d\mu} \left[L^{PP} - L\right], \tag{9}$$

where PP is the sample path on which $T = ((\eta+1)h)_+ \Longrightarrow \eta^{PP} = \eta + 1$.

Similar to Theorems 3 and 4, we have the following unbiasedness result. Its proof would again be similar to that of Theorem 3, and hence is omitted here. The only additional fact we need to use is $T \le (\eta+1)h$, along with the following technical conditions:

(A7) $|dT/d\mu| \le K_1 T + K_2$, where $K_1, K_2 > 0$ are two constants.

(A8) $E[\sup_{h\in(h_{\min},h_{\max})} L^2\eta] < \infty$, where $0 < h_{\min} < h_{\max} < \infty$.

Theorem 5. Under (A5), (A7), and (A8), (8) and (9) are unbiased estimators for $dE[L]/d\mu$.

4.4 In-Control and Out-of-Control Process Parameters

We now consider the case where the parameter is one of the process (in-control or out-of-control) means, so that $\{X_i\}$ changes. In particular, we have

$$\frac{dX_i}{d\mu_0} = 0 \text{ for } i \ge T/h, \tag{10}$$

and

$$\frac{dX_i}{d\mu_1} = 0 \text{ for } i < T/h. \tag{11}$$

Note that the random variable T is independent of μ_0 and μ_1.

The effect is similar to shifting the control limits. For the Shewhart chart, the estimator for $dE[L]/d\mu_j$ $(j = 0,\ 1)$ is given by the following:

$$\begin{aligned}
&\left(\frac{dX_L}{d\mu_j}\right)^- \mathbf{1}\{X_L > u\}\frac{g_L(u)}{1-G_L(u)}E[L|T=\tau_{res}] \\
&\quad + \sum_{i<L}\frac{g_i(l)}{G_i(u)-G_i(l)}\left(\frac{dX_i}{d\mu_j}\right)^-(L-i) \\
&\quad + \left(\frac{dX_L}{d\mu_j}\right)^+ \mathbf{1}\{X_L < l\}\frac{g_L(l)}{G_L(l)}E[L|T=\tau_{res}] \\
&\quad + \sum_{i<L}\frac{g_i(u)}{G_i(u)-G_i(l)}\left(\frac{dX_i}{d\mu_j}\right)^+(L-i).
\end{aligned} \tag{12}$$

When the control chart has dependence on previous data such as in the more general case, we use the relationship between Y_i and X_i and previous statistics to "propagate" perturbations. Differentiating, we have

$$\frac{dY_i}{d\theta} = \frac{d\psi}{dX_i}\frac{dX_i}{d\theta} + \frac{d\psi}{dY_{i-1}}\frac{dY_{i-1}}{d\theta}. \tag{13}$$

In particular, for the EWMA control chart, we have

$$\frac{dY_i}{d\mu_j} = \alpha\frac{dX_i}{d\mu_j} + (1-\alpha)\frac{dY_{i-1}}{d\mu_j}. \tag{14}$$

Then, the final estimator is similar to the Shewhart case, with X_i replaced by Y_i:

$$\left(\frac{dY_L}{d\mu_j}\right)^- \mathbf{1}\{Y_L > u\}\frac{g_L(\psi^{-1}(u,Y_{L-1}))}{1-G_L(\psi^{-1}(u,Y_{L-1}))}\frac{d\psi^{-1}(u,Y_{L-1})}{du}E[L|T=\tau_{res},\tilde{Y_1}=\psi(X_1,u)]$$

$$+\sum_{i<L}\frac{g_i(\psi^{-1}(l,Y_{i-1}))}{G_i(\psi^{-1}(u,Y_{i-1}))-G_i(\psi^{-1}(l,Y_{i-1}))}\frac{d\psi^{-1}(u,Y_{i-1})}{du}\left(\frac{dY_i}{d\mu_j}\right)^{-}(L-i)$$
$$+\left(\frac{dY_L}{d\mu_j}\right)^{+}\mathbf{1}\{Y_L<l\}\frac{g_L(\psi^{-1}(l,Y_{L-1}))}{G_L(\psi^{-1}(l,Y_{L-1}))}\frac{d\psi^{-1}(u,Y_{L-1})}{du}E[L|T=\tau_{res},\tilde{Y}_1=\psi(X_1,u)]$$
$$+\sum_{i<L}\frac{g_i(\psi^{-1}(u,Y_{i-1}))}{G_i(\psi^{-1}(u,Y_{i-1}))-G_i(\psi^{-1}(l,Y_{i-1}))}\frac{d\psi^{-1}(u,Y_{i-1})}{du}\left(\frac{dY_i}{d\mu_j}\right)^{+}(L-i). \tag{15}$$

Next, we note that we can also consider derivatives with respect to parameters in the relationship ψ itself, as well. For example, in the case of the EWMA control chart, one might be interested in sensitivities with respect to the smoothing constant α. Again, we simply differentiate the relationship to obtain a propagation rule:

$$\frac{dY_i}{d\alpha}=X_i-Y_{i-1}, \tag{16}$$

which is used in the estimator (15) above.

The unbiasedness of the estimators (12) and (15) can also be established as we did before. The boundedness condition required depends on the parameter considered, involving either $dY_i/d\mu_j$, $dX_i/d\mu_j$ or $dY_i/d\alpha$ in a manner analogous to previous theorems, as one would expect.

4.5 Implementation of the Estimators

For the sampling frequency and process drift parameters, implementation of the estimators involves finding a method to estimate the expectation of the difference term $L^{PP}-L$. In this section, we describe two methods for constructing the perturbed path (PP) from the nominal path for the Shewhart chart for the right-hand derivative. In the Shewhart chart, the control limits are fixed and not a function of j, and all decisions are based only on the just sampled value of X_i (i.e., $Y_i=X_i$), where X_i is the sample mean for the jth sample and the $\{X_i\}$ are independent.

We recall that

$$X_i^{PP}\stackrel{d}{=}X_i, i\neq\eta,\quad X_\eta^{PP}\sim F_1, X_\eta\sim F_0,$$

where $\stackrel{d}{=}$ denotes equality in distribution, i.e., the two paths have the same distribution everywhere except at $i=\eta$. We propose the following two coupling constructions of $\mathbf{X}^{PP}$, giving the corresponding values of L^{PP} to be derived.

Coupling 1 (one-to-one):

$$\begin{aligned}X_i^{PP}&=X_i \text{ for } i\neq\eta,\\ X_\eta^{PP}&=F_1^{-1}(F_0(X_\eta)),\end{aligned}$$

$$L^{PP}=\begin{cases}\eta & \text{if } hL\geq T \text{ and } X_\eta^{PP}\notin[\mathrm{LCL},\mathrm{UCL}],\\ L+ARL_1 & \text{if } hL\in[T-h,T) \text{ and } X_\eta^{PP}\in[\mathrm{LCL},\mathrm{UCL}],\\ L & \text{otherwise.}\end{cases}$$

Coupling 2 (cut-and-paste):

$$X_i^{PP}=\begin{cases}X_i & \text{for } i\leq\eta-1,\\ X_{i+1} & \text{for } i\geq\eta,\end{cases}$$

$$L^{PP} = \begin{cases} L-1 & \text{if } hL \geq T, \\ L + ARL_1 & \text{if } hL \in [T-h, T) \text{ and } X_\eta^{PP} \in [\text{LCL}, \text{UCL}], \\ L & \text{otherwise.} \end{cases}$$

Coupling 1 uses a one-to-one correspondence between the two paths. In simulation terms, it can be thought of as the "common random numbers" construction, with the same value used for every sample except $i = \eta$. In the case of $i = \eta$, a transformation, using the fact that if $X_i \sim F_0$, then $F_1^{-1}(F_0(X_i)) \sim F_1$. Coupling 2 "cuts" out the "extra" F_0-distributed random variable, and "pastes" the remaining set to construct the perturbed path.

We divide our analysis into three different cases, based on the value of L in the following regions:

(I) $hL < T - h \quad (\eta > L)$;

(II) $hL \in [T-h, T) \quad (\eta = L)$;

(III) $hL \geq T \quad (\eta < L)$.

In both constructions, we note that if $L < T - h$ (Region I), then $L^{PP} = L$, since both paths are identical up to the stopping time L in this case, i.e.,

$$X_i^{PP} = X_i, \quad i < \eta$$

$$\Longrightarrow \quad \inf\{j : X_i^{PP} \notin [\text{LCL}, \text{UCL}]\} = \inf\{j : X_i \notin [\text{LCL}, \text{UCL}]\} = L < \eta.$$

We consider Region II next. For the nominal path, since $\eta = L$, by definition, X_η stops the process, i.e., $X_\eta \notin [\text{LCL}, \text{UCL}]$. However, in the perturbed path, $X_\eta^{PP} \neq X_\eta$. If $X_\eta^{PP} \notin [\text{LCL}, \text{UCL}]$, then we again have $L^{PP} = L$; otherwise, the process will continue until an out-of-control signal is given. Since the process is actually in an out-of-control state for $i > \eta$, the expectation for the additional run length is simply ARL_1. Hence, we take $L^{PP} = L + ARL_1 \cdot \mathbf{1}\{X_\eta^{PP} \in [\text{LCL}, \text{UCL}]\}$. In this case, the two constructions again yield the same L^{PP}, differing only in how X_η^{PP} is generated.

Lastly, we consider Region III ($L \geq T, \eta < L$), for which the two constructions yield distinctly different contributions. In Coupling 1, the perturbed path stops earlier at $L^{PP} = \eta < L$ if $X_\eta^{PP} \notin [\text{LCL}, \text{UCL}]$; otherwise, the rest of the path is the same, so $L^{PP} = L$. For Coupling 2, in Region III, the perturbed path is simply the nominal path shifted by one, so $L^{PP} = L - 1$. Combining all this, we have our two estimators for $dE[L]/dh$:

$$\frac{\eta \cdot f(\eta h)}{F((\eta+1)h) - F(\eta h)} \Big[(\eta - L)\mathbf{1}\{\eta < L, F_1^{-1}(F_0(X_\eta)) \notin [\text{LCL}, \text{UCL}]\} + ARL_1 \cdot \mathbf{1}\{\eta = L, F_1^{-1}(F_0(X_\eta)) \in [\text{LCL}, \text{UCL}]\} \Big],$$

$$\frac{\eta \cdot f(\eta h)}{F((\eta+1)h) - F(\eta h)} \Big[(-1)\mathbf{1}\{\eta < L\} + ARL_1 \cdot \mathbf{1}\{\eta = L, X_{\eta+1} \in [\text{LCL}, \text{UCL}]\} \Big].$$

For the left-hand derivative ($\Delta h < 0$), we have

$$X_i^{PP} \stackrel{d}{=} X_i, i \neq \eta + 1, \quad X_{\eta+1}^{PP} \sim F_0, \quad X_{\eta+1} \sim F_1,$$

i.e., the two paths have the same distribution everywhere except at $i = \eta + 1$. Similar to before, we propose the following two coupling constructions of $\mathbf{X}^{PP}$, giving the corresponding values of L^{PP} to be derived.

Coupling 1 (one-to-one):

$$X_i^{PP} = \begin{cases} X_i & \text{if } i \neq \eta + 1, \\ F_0^{-1}(F_1(X_{\eta+1})) & \text{if } i = \eta + 1, \end{cases}$$

$$L^{PP} = \begin{cases} \eta + 1 & \text{if } hL \geq T + h \text{ and } X_{\eta+1}^{PP} \notin [\text{LCL}, \text{UCL}], \\ L + ARL_1 & \text{if } hL \in [T, T + h) \text{ and } X_{\eta+1}^{PP} \in [\text{LCL}, \text{UCL}], \\ L & \text{otherwise.} \end{cases}$$

Coupling 2 (insert-and-paste):

$$X_i^{PP} = \begin{cases} X_i & \text{if } i \leq \eta, \\ \tilde{X} \sim F_0 \text{ (inserted)} & \text{if } i = \eta + 1, \\ X_{i-1} & \text{if } i \geq \eta + 2, \end{cases}$$

$$L^{PP} = \begin{cases} \eta + 1 & \text{if } hL \geq T + h \text{ and } X_{\eta+1}^{PP} \notin [\text{LCL}, \text{UCL}], \\ L + 1 & \text{if } hL \geq T + h \text{ and } X_{\eta+1}^{PP} \in [\text{LCL}, \text{UCL}], \\ L + ARL_1 & \text{if } hL \in [T, T + h) \text{ and } X_{\eta+1}^{PP} \in [\text{LCL}, \text{UCL}], \\ L & \text{otherwise.} \end{cases}$$

Coupling 1 uses a one-to-one correspondence between the two paths as before, whereas Coupling 2 "inserts" an "extra" F_0-distributed random variable $\tilde{X}$ for $X_{\eta+1}^{PP}$, and "pastes" the remaining set to construct the perturbed path.

We again divide our analysis into three different cases, based on the value of L:

(I) $hL < T \quad (\eta > L - 1)$;

(II) $hL \in [T, T + h) \quad (\eta = L - 1)$;

(III) $hL \geq T + h \quad (\eta < L - 1)$.

For the most part, the analysis is similar to that for the right-hand derivative, except that $X_{\eta+1}$ is the focus instead of X_η. Again, in both constructions, we have $L^{PP} = L$ in Region I ($hL < T$). Similarly in Region II, both constructions yield $L^{PP} = L + ARL_1 \cdot \mathbf{1}\{X_{\eta+1}^{PP} \in [\text{LCL}, \text{UCL}]\}$, differing only in how $X_{\eta+1}^{PP}$ is generated.

In Region III ($hL \geq T + h$), we have a slight difference for Coupling 2. In Coupling 1, as before, the perturbed path stops earlier at $L^{PP} = \eta + 1 < L$ if $X_{\eta+1}^{PP} \notin [\text{LCL}, \text{UCL}]$; otherwise, $L^{PP} = L$. For Coupling 2, in Region III, the perturbed path has the extra $\tilde{X}$ inserted for $X_{\eta+1}^{PP}$. If it signals out of control, then the process stops there to give $L^{PP} = \eta + 1$; otherwise, the perturbed path is simply the nominal path shifted forward by one, so $L^{PP} = L + 1$. Combining all this, we have our two estimators for $dE[L]/dh$:

$$-\frac{(\eta+1) \cdot f((\eta+1)h)}{F((\eta+1)h) - F(\eta h)} \Big[((\eta+1) - L)\mathbf{1}\{\eta + 1 < L, F_0^{-1}(F_1(X_\eta)) \notin [\text{LCL}, \text{UCL}]\} \\ + ARL_1 \cdot \mathbf{1}\{(\eta+1)h = L, F_0^{-1}(F_1(X_\eta)) \in [\text{LCL}, \text{UCL}]\} \Big],$$

$$-\frac{(\eta+1) \cdot f((\eta+1)h)}{F((\eta+1)h) - F(\eta h)} \Big[((\eta+1) - L)\mathbf{1}\{(\eta+1) < L, \tilde{X} \notin [\text{LCL}, \text{UCL}]\} \\ + \mathbf{1}\{\eta + 1 < L, \tilde{X} \in [\text{LCL}, \text{UCL}]\} + ARL_1 \cdot \mathbf{1}\{\eta + 1 = L, \tilde{X} \in [\text{LCL}, \text{UCL}]\} \Big].$$

5 Summary and Avenues for Further Research

We have derived sensitivity estimates for control charts that can be efficiently implemented when Monte Carlo simulation is used for performance evaluation. Such estimators are useful for sensitivity analysis and optimization in the design of the control chart. We considered the average run length performance measure and two types of control charts. Although ARL performance measures are the most commonly used, and thus are addressed explicitly in this paper, in cases where the sample size and/or sampling interval are variable, other appropriate performance measures such as average time to signal and average number of observations to signal can also be handled. If n_i and h_i denote the ith sample size and sampling interval, respectively, then the time to signal is given by

$$\sum_{i=1}^{\infty} h_i \mathbf{1}\{i \leq L\},$$

and the number of observations to signal is given by

$$\sum_{i=1}^{\infty} n_i \mathbf{1}\{i \leq L\}.$$

In the remainder of this section, we briefly describe the formulation of the economic design problem and outline how the estimators would be used in optimizing the design.

5.1 Economic Design Problem

For the problem of designing control charts for statistical process control applications, there are basically three general approaches (Saniga 1989):

- *heuristics*, such as Shewhart himself originally suggested;
- *statistical design*, for which determination is made purely on the basis of statistical factors such as Type I error and the power;
- *economic design*, for which costs and profits are attached to various actions such as sampling and testing, investigation and correction, good production and nonconformance.

Saniga (1989) actually combines the latter two approaches. The optimal economic design problem was first formulated by Duncan (1956); see also Goel and Wu (1973) for CUSUM charts. Montgomery (1996) devotes a chapter to the problem of economic design. The focus has been on deriving analytical expressions for the time-average cost, and then using numerical analysis techniques to search for the optimum, as in general the resulting expressions cannot be analytically solved for the optimum.

The design parameters for the control chart are usually the sample size, the sampling frequency, and the control limits.

In order to formulate an economic design problem, cost parameters must be specified. The usual costs include costs on sampling and testing; a cost on investigating an out-of-control signal; a cost on correcting an out-of-control state; and a "penalty" for the production of nonconforming units ("failure" costs). Then

the objective is to select the values of the design parameters so as to minimize long-run average costs. Let θ denote the vector of design parameters, and $J(\theta)$ the expected cost. We then wish to find

$$\theta^* = \arg\min_{\theta \in \Theta} J(\theta),$$

where Θ represents the feasible region. By using the appropriate gradient estimates, the proposed solution technique is to apply stochastic approximation to perform the optimization via simulation. We will not give all the details here, but merely outline the setting in which the gradient estimates would be employed.

In terms of simulation, we can view the evolution over time as an interaction between two fundamental underlying processes: a process failure mechanism, and a process sampling mechanism. In the former, the state of the system is either in-control or out-of-control. In the latter, sampling takes place at regular intervals until an out-of-control signal occurs. Then an investigation is undertaken to see if the signal is true or specious. If the signal is a false alarm, then sampling resumes; if the signal is real, then corrective action is taken that returns the system back to the in-control state. Thus, the system state goes from in-control to out-of-control through some random mechanism, but can only return to in-control through the sampling mechanism. A production cycle will be defined as an (in-control, out-of-control) sequence partitioned into three periods: the time to go out of control, the time to detect that the system is out of control, and the time to find the assignable cause and return the system to the in-control state. For simplicity, we will consider a single out-of-control state, and define the following parameters:

a = fixed cost of taking a sample,
b = per-unit cost of sample,
w = cost of finding (and correcting) an assignable cause,
y = cost of investigating a false alarm,
c_0 = quality cost (per unit time) when in control,
c_1 = quality cost (per unit time) when out of control,
C = total cost in a production cycle,
T^* = length of a production cycle,
T = time to go out of control,
T_1 = time to detect out of control after it has occurred,
T_2 = time to interpret sample, investigate, assign and correct (assignable) cause,
γ_j = index of sample giving jth out-of-control sample,
= $\inf\{i > \gamma_{j-1} : X_i \notin [l, u]\}, j \geq 1 \ (\gamma_0 = 0)$,
$L^{(j)}$ = time between jth out-of-control sample and last investigation
= $h(\gamma_j - \gamma_{j-1})$,
$L^{(*)}$ = time from the last false alarm investigation to the time when the process goes out of control,
$T_0^{(j)}$ = time to investigate out-of-control signal false alarm,
N = number of samplings taken in a production cycle,
N^* = number of out-of-control signals in a production cycle.

Assuming that no samplings are taken during $T_0^{(j)}$ and T_2, we have

$$N = \frac{1}{h} \sum_{j=1}^{N^*-1} L^{(j)} + L^{(*)} + T_1.$$

The total production cycle length is given by

$$T^* = \sum_{j=1}^{N^*-1} \left(L^{(j)} + T_0^{(j)} \right) + L^{(*)} + T_1 + T_2,$$

and the expected cost in the production cycle is given by

$$C = (a + bn)N + c_0 T + c_1 (T^* - T) + y(N^* - 1) + w,$$

with the long-run average cost is given by

$$J(\theta) = \frac{E[C(\theta)]}{E[T^*(\theta)]}.$$

We note that

$$\frac{\partial E[T_0^{(j)}]}{\partial \theta} = \frac{\partial E[T_2]}{\partial \theta} = 0,$$

so that the only derivatives that are necessary are those for $L^{(j)}$, $L^{(*)}$, T_1, and N^*. We have already shown how such estimators are derived for $L^{(j)}$; similar analysis can be used to derive estimators for the others.

This simulation-based framework for optimal economic design problems in statistical quality control is very general and can be used in cases where analytical models cannot be easily applied. The approach uses gradient estimators of the objective function with respect to the design parameters to search for the optimum in a recursive stochastic approximation algorithm.

References

[1] Albin, S.A., Kang, L., and Shea, G., "An X and EWMA Chart for Individual Observations," *Journal of Quality Technology*, Vol.29, No.1, 41-48, 1997.

[2] Barish, N.N. and Hauser, N., "Economic Design for Control Decisions," *The Journal of Industrial Engineering*, Vol.14, 125-134, 1963.

[3] Baxley, R.V., "An Application of Variable Sampling Inverval Control Charts," *Journal of Quality Technology*, Vol.27, No.4, 275-282, 1995.

[4] Cao, X.R., *Realization Probabilities: The Dynamics of Queueing Systems*, Springer Lecture Notes in Control and Optimization, **194**, Springer-Verlag, New York, 1994.

[5] Duncan, A.J., "The Economic Design of $\bar{X}$-Charts Used to Maintain Current Control of a Process," *Journal of the American Statistical Association*, Vol.51, 228-242, 1956.

[6] Fu, M.C., "Optimization via Simulation: A Review," *Annals of Operations Research*, Vol. 53, pp.199-248, 1994.

[7] Fu, M.C. and Hu, J.Q., "Extensions and Generalizations of Smoothed Perturbation Analysis in a Generalized Semi-Markov Process Framework," *IEEE Transactions Automatic Control*, **37**, pp.1483-1500, 1992.

[8] Fu, M.C. and Hu, J.Q., *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*, Kluwer Academic Publishers, 1997.

[9] Gan, F.F., "Joint Monitoring of Process Mean and Variance Using Exponentially Weighted Moving Average Control Charts," *Technometrics*, Vol.37, No.4, 446-453, 1995.

[10] Glasserman, P., *Gradient Estimation Via Perturbation Analysis*, Kluwer Academic Publishers, 1991.

[11] Goel, A.L. and Wu, S.M., "Economically Optimum Design of CuSum Charts," *Management Science*, Vol.19, No.11, 1271-1282, 1973.

[12] Gong, W.B. and Ho, Y.C., "Smoothed Perturbation Analysis of Discrete-Event Dynamic Systems," *IEEE Transactions on Automatic Control*, **32**, 858-867, 1987.

[13] Grimshaw, S.D. and Alt, F.B., "Control Charts for Quantile Function Values," *Journal of Quality Technology*, Vol.29, No.1, 1-7, 1997.

[14] Ho, Y.C. and Cao, X.R., *Discrete Event Dynamic Systems and Perturbation Analysis,* Kluwer Academic Publishers, 1991.

[15] Ho, C. and Case, K.E., "Eceonomic Design of Control Charts: A Literature Review for 1981-1991," *Journal of Quality Technology*, Vol.26, No.1, 39-53, 1994.

[16] Kushner, H.J. and Clark, D.C., *Stochastic Approximation Methods for Constrained and Unconstrained Systems,* Springer-Verlag, New York, 1978.

[17] Montgomery, D.C., *Introduction to Statistical Quality Control*, 3rd edition, John Wiley & Sons, 1996.

[18] Rubinstein, R.Y. and Shapiro, A., *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*, John Wiley & Sons, 1993.

[19] Saniga, E.M., "Economic Statistical Control-Chart Designs With an Application to $\bar{X}$ and R Charts," *Technometrics*, Vol.31, 313-320, 1989.

[20] Seppala, T., Moskowitz, H., Plante, R., and Tang, J., "Statistical Process Control via the Subgroup Bootstrap," *Journal of Quality Technology*, Vol.27, No.2, 139-153, 1995.