Research Note:

Some Reflections on the Calculation of Reliable Gradient Estimates for RTO Schemes

Benoît CHACHUAT

(30th April 2008)

École Polytechnique Fédérale de Lausanne (EPFL), Laboratoire d'Automatique, Station 9, CH-1015 Lausanne, Switzerland.

1 Problem Statement

Throughout this note, we consider an optimization problem in the form of the following general NLP: **(minimize wrt π)**

$$\min_{\boldsymbol{\pi}} \quad \phi(\boldsymbol{\pi}, \mathbf{y}) \tag{P_p}$$
$$\text{s.t.} \quad \mathbf{y} = \mathscr{F}(\boldsymbol{\pi})$$
$$\mathbf{g}(\boldsymbol{\pi}, \mathbf{y}) \leq \mathbf{0},$$

where $\boldsymbol{\pi} \in \mathbb{R}^{n_\pi}$ denotes the decision (or input) variables; $\mathbf{y} \in \mathbb{R}^{n_y}$ the output variables; ϕ the scalar objective function; $\mathbf{g}$ the vector of n_g inequality constraints; and $\mathscr{F}$ the actual plant.

In practice, the map $\mathscr{F}$ is typically unknown, and only an approximate, finite-dimensional model is available to describe its behavior,

$$\mathbf{y} = \mathbf{f}(\boldsymbol{\pi}, \boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$ is a set of adjustable model parameters. Based on this model, one can then get an approximate solution to the original problem (P_p) by solving the optimization problem

$$\min_{\boldsymbol{\pi}} \quad \phi(\boldsymbol{\pi}, \mathbf{y}) \tag{P_m}$$
$$\text{s.t.} \quad \mathbf{y} = \mathbf{f}(\boldsymbol{\pi}, \boldsymbol{\theta})$$
$$\mathbf{g}(\boldsymbol{\pi}, \mathbf{y}) \leq \mathbf{0}.$$

Due to the presence of uncertainty in the form of model mismatch and process disturbances, however, the optimal solution to (P_m) –assuming it is unique– may be quite different from the optimal solution to (P_p).

The objective of real-time optimization (RTO) is to take advantage of the available measurements in order compensate for the uncertainty and adapt the model-based problem so as to get closer to the actual plant optimum. Our focus here is on modifier-adaptation methods [1] that use measurements to correct the values and the first-order derivatives of both the cost and constraint functions in (P_m) as

$$\min_{\boldsymbol{\pi}} \quad \phi(\boldsymbol{\pi}, \mathbf{y}) + \boldsymbol{\lambda}_\phi^\mathsf{T} \boldsymbol{\pi} \tag{P_a}$$
$$\text{s.t.} \quad \mathbf{y} = \mathbf{f}(\boldsymbol{\pi}, \boldsymbol{\theta})$$
$$\mathbf{g}(\boldsymbol{\pi}, \mathbf{y}) + \boldsymbol{\varepsilon}_g + \boldsymbol{\lambda}_g^\mathsf{T} \boldsymbol{\pi} \leq \mathbf{0},$$

where $\boldsymbol{\varepsilon}_g \in \mathbb{R}^{n_g}$ is the constraint-value modifier; $\boldsymbol{\lambda}_g \in \mathbb{R}^{n_\pi \times n_g}$ the constraint-gradient modifier; and $\boldsymbol{\lambda}_\phi \in \mathbb{R}^{n_\pi}$ the cost-gradient modifier.

The use of modifiers is attractive in the sense that a KKT point $\boldsymbol{\pi}^*$ for the corrected model-based problem (P_a) is also a KKT point for the original problem (P_p), provided that the modifiers satisfy:

$$\boldsymbol{\varepsilon}_g = \mathbf{g}(\boldsymbol{\pi}^*, \mathscr{F}(\boldsymbol{\pi}^*)) - \mathbf{g}(\boldsymbol{\pi}^*, \mathbf{f}(\boldsymbol{\pi}^*, \boldsymbol{\theta}))$$

$$\boldsymbol{\lambda}_g^{\mathsf{T}} = \left.\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right|_{(\boldsymbol{\pi}^*, \mathscr{F}(\boldsymbol{\pi}^*))} \left.\frac{\partial \mathscr{F}}{\partial \boldsymbol{\pi}}\right|_{\boldsymbol{\pi}^*} - \left.\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right|_{(\boldsymbol{\pi}^*, \mathbf{f}(\boldsymbol{\pi}^*, \boldsymbol{\theta}))} \left.\frac{\partial \mathbf{f}}{\partial \boldsymbol{\pi}}\right|_{\boldsymbol{\pi}^*}$$

$$\boldsymbol{\lambda}_\phi^{\mathsf{T}} = \left.\frac{\partial \phi}{\partial \mathbf{y}}\right|_{(\boldsymbol{\pi}^*, \mathscr{F}(\boldsymbol{\pi}^*))} \left.\frac{\partial \mathscr{F}}{\partial \boldsymbol{\pi}}\right|_{\boldsymbol{\pi}^*} - \left.\frac{\partial \phi}{\partial \mathbf{y}}\right|_{(\boldsymbol{\pi}^*, \mathbf{f}(\boldsymbol{\pi}^*, \boldsymbol{\theta}))} \left.\frac{\partial \mathbf{f}}{\partial \boldsymbol{\pi}}\right|_{\boldsymbol{\pi}^*}.$$

An iterative scheme that adapts the modifiers so as to satisfy the foregoing conditions upon convergence can easily be devised. Perhaps the key issue in applying this approach, however, is tied to the fact that the gradient of the plant outputs with respect to the plant inputs, $\frac{\partial \mathscr{F}}{\partial \boldsymbol{\pi}}$, also called *experimental gradient*, must be available.

The focus in the remainder of this note is on the reliable and accurate estimation of the experimental gradient based on the knowledge of input/output measurements at previous operating points.

2 Experimental Gradient Estimation

Given n_π past operating points, $\boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(n_\pi)}$, and the corresponding measured outputs $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n_\pi)}$, the experimental gradient relative to the ith output at the new point $\boldsymbol{\pi}$ can be approximated as:

$$\left.\frac{\partial \mathscr{F}_i}{\partial \boldsymbol{\pi}}\right|_{\boldsymbol{\pi}} \approx \mathcal{G}_i(\boldsymbol{\pi}) := \mathcal{U}(\boldsymbol{\pi})^{-1} \mathcal{Y}_i(\mathscr{F}(\boldsymbol{\pi})) \tag{2}$$

with:

$$\mathcal{U}(\boldsymbol{\pi}) := \begin{pmatrix} \boldsymbol{\pi} - \boldsymbol{\pi}^{(n_\pi)} & \boldsymbol{\pi}^{(n_\pi)} - \boldsymbol{\pi}^{(n_\pi-1)} & \cdots & \boldsymbol{\pi}^{(2)} - \boldsymbol{\pi}^{(1)} \end{pmatrix}^{\mathsf{T}} \in \mathbb{R}^{n_\pi \times n_\pi} \tag{3}$$

$$\mathcal{Y}_i(\mathbf{y}) := \begin{pmatrix} y_i - y_i^{(n_\pi)} & y_i^{(n_\pi)} - y_i^{(n_\pi-1)} & \cdots & y_i^{(2)} - y_i^{(1)} \end{pmatrix}^{\mathsf{T}} \in \mathbb{R}^{n_\pi},$$

for each $i = 1, \ldots, n_y$.

In theory, the smaller the difference between the operating points $\boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(n_\pi+1)}$, the more accurate the approximation $\mathcal{G}_i$ of the experimental gradient. In practice, however, having the past operating points too close to one another can lead to unreliable estimates because the plant outputs are inevitably corrupted by noise. To conduct the analysis, it shall be assumed throughout that the measurements y_i, $i = 1 \ldots, n_y$, at a given operating point $\boldsymbol{\pi}$ are independent Gaussian white noise processes with a mean of $\mathscr{F}_i(\boldsymbol{\pi})$ and a variance of $\sigma_{y_i}^2$,

$$y_i \sim \mathcal{N}(\mathscr{F}_i(\boldsymbol{\pi}), \sigma_{y_i}^2), \quad \forall i \in \{1, \ldots, n_y\}. \tag{4}$$

In previous work [2, 3], the effect of noise has been accounted for by requiring that the matrix $\mathcal{U}(\boldsymbol{\pi})$ at the new operating point $\boldsymbol{\pi}$ be sufficiently well conditioned. While a large condition number can indeed lead to considerable amplification of the measurement errors, thus leading to poor gradient estimates, a difficulty with this approach is that the relation between the condition number of $\mathcal{U}(\boldsymbol{\pi})$ and the effect of measurement noise on $\mathcal{G}(\boldsymbol{\pi})$ is not straightforward.

The novel approach proposed herein is based on the rather natural idea that the expected level of noise in $\mathcal{G}$, as induced by the noise in the output measurements, can be kept sufficiently small by ensuring a certain distance between successive operating points. From (2) and (4), the components of the gradient estimate

$\mathcal{G}_i(\boldsymbol{\pi})$ at a new point $\boldsymbol{\pi}$ are Gaussian white noise processes,

$$\mathcal{G}_{ij}(\boldsymbol{\pi}) \sim \mathcal{N}(\mu_{\mathcal{G}_{ij}}(\boldsymbol{\pi}), \sigma_{\mathcal{G}_{ij}}(\boldsymbol{\pi})), \quad \forall (i,j) \in \{1, \ldots, n_y\} \times \{1, \ldots, n_\pi\} \tag{5}$$

$$\text{with:} \quad \mu_{\mathcal{G}_{ij}}(\boldsymbol{\pi}) = \left[\mathcal{U}(\boldsymbol{\pi})^{-1}\right]_{j\cdot} \mathcal{Y}_i(\mathscr{F}(\boldsymbol{\pi})) \tag{6}$$

$$\sigma_{\mathcal{G}_{ij}}(\boldsymbol{\pi}) = \sigma_{y_i} \sqrt{2 \sum_{k=1}^{n_\pi} \left[\mathcal{U}(\boldsymbol{\pi})^{-1}\right]_{jk}^2}. \tag{7}$$

Observe that the variance of the estimated gradient components results from the measurement noise in the measured outputs $\mathbf{y}$ corresponding to the new point $\boldsymbol{\pi}$, but also on the measurement noise in the measured outputs $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n_\pi)}$ relative to the past operating points $\boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(n_\pi)}$.

Based on (5–7), different types of constraints can be defined in order to limit the effect of measurement noise on the elements of the estimated gradients when selecting future operating points. In particular, the variance (7) of the estimated gradient components being independent of the process model, it appears promising to define constraints in terms of $\sigma_{\mathcal{G}_{ij}}(\boldsymbol{\pi})$ since such constraints would remain valid no matter how inaccurate the process model is.

2.1   Worst-case Constraint on Gradient Standard Deviation

Defining an upper bound on the maximal standard deviation $\sigma_{\mathcal{G}_{ij}}(\boldsymbol{\pi})$ when selecting the new operating point $\boldsymbol{\pi}$, either output-wise or not,

$$\max_{1 \le j \le n_\pi} \sigma_{\mathcal{G}_{ij}}(\boldsymbol{\pi}) \le \bar{\sigma}_{\mathcal{G}_i}^{\text{abs}}, \quad i = 1, \ldots, n_y, \qquad \max_{\substack{1 \le i \le n_y \\ 1 \le j \le n_\pi}} \sigma_{\mathcal{G}_{ij}}(\boldsymbol{\pi}) \le \bar{\sigma}_{\mathcal{G}}^{\text{abs}},$$

guarantees that the standard deviation of *any* of the estimated gradient elements remain lower than $\bar{\sigma}_{\mathcal{G}_i}^{\text{abs}}$ (or $\bar{\sigma}_{\mathcal{G}}^{\text{abs}}$). One could also decide to restrict the *relative* standard deviation of the elements in $\mathcal{G}$,

$$\max_{1 \le j \le n_\pi} \sigma_{\mathcal{G}_{ij}}(\boldsymbol{\pi}) \le \mathcal{G}_{ij}(\boldsymbol{\pi}) \bar{\sigma}_{\mathcal{G}_i}^{\text{rel}}, \quad i = 1, \ldots, n_y, \qquad \max_{\substack{1 \le i \le n_y \\ 1 \le j \le n_\pi}} \sigma_{\mathcal{G}_{ij}}(\boldsymbol{\pi}) \le \mathcal{G}_{ij}(\boldsymbol{\pi}) \bar{\sigma}_{\mathcal{G}}^{\text{rel}}.$$

By combining the foregoing absolute and relative restrictions, yet another constraint is obtained as

$$\max_{\substack{1 \le i \le n_y \\ 1 \le j \le n_\pi}} \sigma_{\mathcal{G}_{ij}(\boldsymbol{\pi})} \le \max\{\bar{\sigma}_{\mathcal{G}}^{\text{abs}}; \mathcal{G}_{ij}(\boldsymbol{\pi}) \bar{\sigma}_{\mathcal{G}}^{\text{rel}}\}. \tag{8}$$

2.2   Norm-based Constraint on Gradient Standard Deviation

Instead of considering the largest standard deviation of the elements of $\mathcal{G}(\boldsymbol{\pi})$ in selecting the new operating point $\boldsymbol{\pi}$, one could as well consider any p-norm of the vector $\boldsymbol{\sigma}_{\mathcal{G}_i}$ (or, alternatively, of the matrix $\boldsymbol{\sigma}_{\mathcal{G}}$). The resulting constraints read

$$\|\boldsymbol{\sigma}_{\mathcal{G}_i}(\boldsymbol{\pi})\|_p \le \bar{\sigma}_{\mathcal{G}_i}^p, \tag{9}$$

for each $i = 1, \ldots, n_y$.

Theorem 1 (Necessary Conditions I) *For the condition (9) to be satisfied it is necessary that the following disjunctive affine constraints hold:*

$$\left(\boldsymbol{\alpha}^\mathsf{T} \boldsymbol{\pi} \ge \beta_i^+\right) \vee \left(\boldsymbol{\alpha}^\mathsf{T} \boldsymbol{\pi} \le \beta_i^-\right), \tag{10}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{n_\pi}$ and $\beta_i^+, \beta_i^- \in \mathbb{R}$, $\beta_i^+ > \beta_i^-$, are given by

$$\alpha_k := (-1)^{1+k} \det \mathcal{U}_{(1,k)}, \quad k = 1, \ldots, n_\pi \tag{11}$$

$$\beta_i^\pm := \sum_{k=1}^{n_\pi} (-1)^{1+k} \pi_k^{(n_\pi)} \det \mathcal{U}_{(1,k)} \pm \frac{\sqrt{2}\sigma_{y_i}}{\bar{\sigma}_{\mathcal{G}_i}^p} \left\| \det \mathcal{U}_{(1,\cdot)} \right\|_p, \tag{12}$$

with $(-1)^{i+j} \det \mathcal{U}_{(i,j)}$ standing for the $(i,j)^{\text{th}}$ cofactor of $\mathcal{U}$.

Proof Observe first that the $(1,k)^{\text{th}}$ cofactor of $\mathcal{U}$ is independent of $\boldsymbol{\pi}$ since only the first row of $\mathcal{U}$ depends on $\boldsymbol{\pi}$. By contradiction, assume that (9) holds, but (10) does not. That is,

$$\left| \sum_{k=1}^{n_\pi} (-1)^{1+k} \left(\pi_k - \pi_k^{(n_\pi)} \right) \det \mathcal{U}_{(1,k)} \right| < \frac{\sqrt{2}\sigma_{y_i}}{\bar{\sigma}_{\mathcal{G}_i}^p} \left\| \det \mathcal{U}_{(1,\cdot)} \right\|_p. \tag{13}$$

Noting that the determinant of $\mathcal{U}(\boldsymbol{\pi})$ is given by

$$\det \mathcal{U}(\boldsymbol{\pi}) = \sum_{k=1}^{n_\pi} (-1)^{1+k} \left(\pi_k - \pi_k^{(n_\pi)} \right) \det \mathcal{U}_{(1,k)}, \tag{14}$$

and that

$$\left[\mathcal{U}(\boldsymbol{\pi})^{-1} \right]_{k1} = \frac{(-1)^{1+k} \det \mathcal{U}_{(1,k)}}{\det \mathcal{U}(\boldsymbol{\pi})},$$

for each $k = 1, \ldots, n_\pi$, one has that the inequality (13) is equivalent to

$$\sqrt{2}\sigma_{y_i} \left\| \left[\mathcal{U}(\boldsymbol{\pi})^{-1} \right]_{\cdot 1} \right\|_p > \bar{\sigma}_{\mathcal{G}_i}^p. \tag{15}$$

On the other hand, by the definition of $\sigma_{\mathcal{G}_{ij}}(\boldsymbol{\pi})$ in (7),

$$\|\boldsymbol{\sigma}_{\mathcal{G}_i}(\boldsymbol{\pi})\|_p = \sqrt{2}\sigma_{y_i} \left\| \sqrt{\sum_{k=1}^{n_\pi} [\mathcal{U}(\boldsymbol{\pi})^{-1}]_{\cdot k}^2} \right\|_p \geq \sqrt{2}\sigma_{y_i} \left\| [\mathcal{U}(\boldsymbol{\pi})^{-1}]_{\cdot 1} \right\|_p.$$

From (15), one finally obtains that $\|\boldsymbol{\sigma}_{\mathcal{G}_i}(\boldsymbol{\pi})\|_p > \bar{\sigma}_{\mathcal{G}_i}^p$, which contradicts (9). $\quad\square$

Corollary 1 (Minimal Distance to Previous Points) *Given n_π points $\boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(n_\pi)}$ in $\mathbb{R}^{n_\pi}$, such that a unique hyperplane, $\mathcal{H}^{(n_\pi)}$, passes through these points, the distance between any new point $\boldsymbol{\pi}$ satisfying the constraint (9) and $\mathcal{H}^{(n_\pi)}$ is greater than $\frac{\sqrt{2}\sigma_{y_i}}{\bar{\sigma}_{\mathcal{G}_i}^p}$.*

Proof From (14), the hyperplane $\mathcal{H}^{(n_\pi)}$ is defined by the equation

$$\sum_{k=1}^{n_\pi} (-1)^{1+k} \left(\pi_k - \pi_k^{(n_\pi)} \right) \det \mathcal{U}_{(1,k)} = 0.$$

In other words, $\mathcal{H}^{(n_\pi)}$ is such that $\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{\pi} = \gamma$, with $\gamma := \sum_{k=1}^{n_\pi} (-1)^{1+k} \pi_k^{(n_\pi)} \det \mathcal{U}_{(1,k)}$. Clearly, this hyperplane is parallel to the hyperplanes $\mathcal{H}_+^{(n_\pi)} : \boldsymbol{\alpha}^\mathsf{T}\boldsymbol{\pi} = \beta_i^+$ and $\mathcal{H}_-^{(n_\pi)} : \boldsymbol{\alpha}^\mathsf{T}\boldsymbol{\pi} = \beta_i^-$ defined in Theorem 1. Moreover, $\mathcal{H}^{(n_\pi)}$ is equidistant to $\mathcal{H}_+^{(n_\pi)}$ and $\mathcal{H}_-^{(n_\pi)}$ and one has

$$d(\mathcal{H}^{(n_\pi)}, \mathcal{H}_\pm^{(n_\pi)}) = \frac{|\gamma - \beta_i^\pm|}{\|\boldsymbol{\alpha}\|_p} = \frac{\sqrt{2}\sigma_{y_i}}{\bar{\sigma}_{\mathcal{G}}^p} \frac{\|\det \mathcal{U}_{1\cdot}\|_p}{\|\boldsymbol{\alpha}\|_p}. \tag{16}$$

The result follows by noting that $\|\boldsymbol{\alpha}\|_p = \|\det \mathcal{U}_{1\cdot}\|_p$. $\quad\square$

Remark 1 (Independence of Minimal Distance to Previous Points) *Corollary 1 shows that the minimal distance between a new point $\boldsymbol{\pi}$ and the previous points $\boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(n_\pi)}$ as imposed by (9) depends only on the standard deviation of the measurement noise and the specified bound $\bar{\sigma}_{\mathcal{G}}^p$. In particular, this distance remains unchanged no matter where the previous points $\boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(n_\pi)}$ are located in $\mathbb{R}^{n_\pi}$.*

Defining affine constraints such as (10) in selecting a new operating point is necessary for the estimated gradient components to not be affected too strongly by the measurement noise. However, (10) is not a sufficient condition for the constraint (9) to hold. Additional necessary conditions are derived in the following theorem.

Theorem 2 (Necessary Conditions II) *For the condition (9) to be satisfied it is necessary that the following disjunctive convex/concave constraints hold for each $\ell = 2, \ldots, n_\pi$, in addition to the disjunctive affine constraints (10):*

$$\left(\boldsymbol{\alpha}^\mathsf{T} \boldsymbol{\pi} \geq \mu_i^{\ell,+}(\boldsymbol{\pi}) \right) \vee \left(\boldsymbol{\alpha}^\mathsf{T} \boldsymbol{\pi} \leq \mu_i^{\ell,-}(\boldsymbol{\pi}) \right), \tag{17}$$

with $\boldsymbol{\alpha}$ given by (11); and where $\mu_i^{\ell,+}, \mu_i^{\ell,-} : \mathbb{R}^{n_\pi} \to \mathbb{R}$ are a convex function on $\mathbb{R}^{n_\pi}$ and a concave function on $\mathbb{R}^{n_\pi}$, respectively, satisfying $\mu_i^{\ell,+}(\boldsymbol{\pi}) \geq \mu_i^{\ell,-}(\boldsymbol{\pi}), \forall \boldsymbol{\pi} \in \mathbb{R}^{n_\pi}$, and given by

$$\mu_i^{\ell,\pm}(\boldsymbol{\pi}) := \sum_{k=1}^{n_\pi} (-1)^{1+k} \pi_k^{(n_\pi)} \det \mathcal{U}_{(1,k)} \pm \frac{\sqrt{2}\sigma_{y_i}}{\bar{\sigma}_{\mathcal{G}_i}^p} \left\| \det \mathcal{U}_{(\ell,\cdot)}(\boldsymbol{\pi}) \right\|_p. \tag{18}$$

Proof Note first that $\mu_i^{\ell,+}, \mu_i^{\ell,-} : \mathbb{R}^{n_\pi} \to \mathbb{R}$ are a convex function on $\mathbb{R}^{n_\pi}$ and a concave function on $\mathbb{R}^{n_\pi}$, respectively, because the function $\boldsymbol{\pi} \mapsto \|\boldsymbol{\pi}\|_p$ is convex on $\mathbb{R}^{n_\pi}$ and the $(\ell,k)^{\text{th}}$ cofactor of $\mathcal{U}$ is an affine function of $\boldsymbol{\pi}$, for each $k = 1, \ldots, n_\pi$ and each $\ell = 2, \ldots, n_\pi$. By contradiction, assume that (9) holds, but (17) does not for some $\ell \in \{2, \ldots, n_\pi\}$. That is,

$$\left| \sum_{k=1}^{n_\pi} (-1)^{1+k} \left(\pi_k - \pi_k^{(n_\pi)} \right) \det \mathcal{U}_{(1,k)} \right| < \frac{\sqrt{2}\sigma_{y_i}}{\bar{\sigma}_{\mathcal{G}_i}^p} \left\| \det \mathcal{U}_{(\ell,\cdot)}(\boldsymbol{\pi}) \right\|_p. \tag{19}$$

From (14) and noting

$$\left[\mathcal{U}(\boldsymbol{\pi})^{-1} \right]_{k\ell} = \frac{(-1)^{\ell+k} \det \mathcal{U}_{(\ell,k)}(\boldsymbol{\pi})}{\det \mathcal{U}(\boldsymbol{\pi})},$$

for each $k = 1, \ldots, n_\pi$, one has that the inequality (19) is equivalent to

$$\sqrt{2}\sigma_{y_i} \left\| \left[\mathcal{U}(\boldsymbol{\pi})^{-1} \right]_{\cdot\ell} \right\|_p > \bar{\sigma}_{\mathcal{G}_i}^p. \tag{20}$$

On the other hand, by the definition of $\sigma_{\mathcal{G}_{ij}}(\boldsymbol{\pi})$ in (7),

$$\left\| \boldsymbol{\sigma}_{\mathcal{G}_i}(\boldsymbol{\pi}) \right\|_p = \sqrt{2}\sigma_{y_i} \left\| \sqrt{\sum_{k=1}^{n_\pi} \left[\mathcal{U}(\boldsymbol{\pi})^{-1} \right]_{\cdot k}^2} \right\|_p \geq \sqrt{2}\sigma_{y_i} \left\| \left[\mathcal{U}(\boldsymbol{\pi})^{-1} \right]_{\cdot\ell} \right\|_p.$$

From (20), one then obtains that $\left\| \boldsymbol{\sigma}_{\mathcal{G}_i}(\boldsymbol{\pi}) \right\|_p > \bar{\sigma}_{\mathcal{G}_i}^p$, which contradicts (9). $\square$

2.3 Discussion

Clearly, adding such a constraint as (8) or (9) in the modified optimization problem (P_a) guarantees that the gradient estimates are not affected too strongly by the measurement noise at the new operating point $\boldsymbol{\pi}$. On the other hand, these constraints do not tell anything about how close to (or far form!) the actual

experimental gradient $\frac{\partial \mathscr{F}}{\partial \boldsymbol{\pi}}$ the gradient estimates $\mathcal{G}$ are at that point. As already mentioned, the only guarantee that $\mathcal{G}$ is a good approximation of $\frac{\partial \mathscr{F}}{\partial \boldsymbol{\pi}}$ consists of taking $\boldsymbol{\pi}$ close enough to the past operating points. Note that these two objectives of (i) reducing the sensitivity of $\mathcal{G}$ to noise, and (ii) getting accurate gradient estimates are conflicting. For some problems, it may happen that obtaining a gradient estimate that would be both accurate and reliable is simply not possible by using the finite difference scheme (2). This is the case, e.g., when the level of noise is very high.

References

[1] B. Chachuat, B. Srinivasan, and D. Bonvin. Model parameterization taylored to real-time optimization. In *Proc. 18th European Symposium on Computer Aided Process Engineering (ESCAPE-18)*, Lyon, France, 2008.

[2] M. Brdyś and P. Tatjewski. An algorithm for steady-state optimising dual control of uncertain plants. In *Proc. 1st IFAC Workshop on New Tends in Design of Control Systems*, pages 249–254, 1994.

[3] W. Gao and S. Engell. Iterative set-point optimization of batch chromatography. *Comput. Chem. Eng.*, 29:1401–1409, 2005.

[4] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, New York, 1994.