Choosing a constant Kalman gain by nonlinear Chebyshev approximation

Dr.Jo
PAc @ F.R.Granaria AC

Nineteenth version, 24 November 2026

Abstract

For a linear time-invariant state-space model and a finite record of scalar measurements we treat the gain of a constant-gain predictor as a free parameter and choose it so that the largest absolute one-step prediction error is as small as possible. The predictions are polynomials in the gain, so this is a discrete nonlinear Chebyshev approximation problem. We give an exact recursion for the derivative of the prediction errors with respect to the gain, solve the problem by the generalised Gauss–Newton method, whose steps are linear programmes, and recall the classical local convergence result of Cromme and of Jittorntrum and Osborne: at a strongly unique solution the iteration converges quadratically, for every norm. The result therefore also covers a composite criterion that adds a multiple of the largest second difference of the predictions. With the maximum norm on the errors and on the parameters, the largest constant in the linearised strong-uniqueness condition can be computed after the iteration by $2p$ small linear programmes, where p is the number of parameters. In simulations with a constant-velocity model ($p = 2$) this check is passed by 74 to 82.5 per cent of the computed solutions, and there the iteration needs six or seven steps in the median. In almost all other data sets the solution has only two extremal points; the theory does not apply there, and a safeguarded iteration needs about six times as many steps. For these solutions we add a second stage, Newton's method on the optimality conditions of the extremal set with exact second derivatives; in the simulations it reduces the work to about six linear programmes and four Newton steps in the median. Its switching rule is heuristic. In these simulations the median stationary standard deviation of the one-step prediction errors was larger for the minimax gain than for the least-squares gain, under both noise distributions.

1 Introduction

The stationary Kalman filter for a time-invariant model is a linear recursion with a constant gain. If the noise covariances are known, the gain can be computed from an algebraic Riccati equation (under the usual detectability and stabilisability conditions; see, e.g., [1]); the filter goes back to Kalman [12]. If they are not known, a long-established alternative is to regard the gain itself as the unknown and to determine it from the measurements. Mehra [14] and Carew and Bélanger [4] obtain the stationary gain from the autocorrelation of the innovations of a suboptimal filter. In system identification the gain is a parameter of the innovations form of the model and is estimated by prediction-error methods [13]; the variant relevant here uses a quadratic criterion. Recent work studies the mean-square prediction error as a function of the gain: gradient methods with convergence guarantees for the stationary cost [17, 18, 2] (see also [3] for unknown singular covariances), and consistency of the local minimisers of the finite-sample prediction-error criterion [16].

All of these approaches rest on second-order quantities: correlations of the innovations or the mean-square prediction error. This paper keeps the parametrisation and changes the criterion: the gain is chosen to minimise the *largest* absolute prediction error on the record. The problem then belongs to nonlinear Chebyshev approximation, and the natural algorithm is the generalised Gauss–Newton method, in which every step solves the linear Chebyshev problem of the linearisation; an early algorithm for nonlinear minimax approximation is due to Osborne and Watson [15]. Cromme [5, Satz 1] proved that the method converges locally quadratically at strongly unique solutions with a regular derivative, in the maximum norm, and later placed the result in an abstract framework [6, Theorem 3.1 and Example 4]. Jittorntrum and Osborne [11] proved it for an arbitrary norm (Theorem 2.1), showed that strong uniqueness of the problem and of its linearisation at the solution are equivalent (Lemma 2.1), and characterised strong uniqueness by subgradients of the norm (Theorems 2.2 and 2.3). The theory used below is therefore known. A web search carried out in November 2026 found no publication that applies this method to the choice of a filter gain; the search did not cover the subject databases or the literature on set-membership (bounded-error) estimation, so we do not claim that none exists.

The contributions are modest and specific.

1. The prediction errors of a constant-gain predictor are polynomials in the gain, and their Jacobian is given exactly by a matrix recursion (Proposition 4). For fixed n its cost is, like that of the predictor, linear in the record length N; with dense matrices the predictor takes $O(Nn^2)$ and the recursion $O(Nn^3)$ operations.

2. For the reader's convenience we give a short self-contained proof of the convergence theorem of Jittorntrum and Osborne [11, Theorem 2.1] for an arbitrary norm, with the neighbourhood and the contraction factor expressed through a Lipschitz constant of the Jacobian (Theorem 7). Its hypothesis is strong uniqueness of the linearised problem, which is equivalent to local strong uniqueness of the nonlinear problem [11, Lemma 2.1]; Lemma 6 records the constants. This is not a new result.

3. Since the theorem holds for every norm, it applies to a composite criterion with a smoothness term (Corollary 8). Strong uniqueness for the composite criterion is a separate hypothesis; it is not inherited from the plain maximum norm.

4. For the maximum norm the known characterisation of strong uniqueness [11, Theorems 2.2 and 2.3] is made quantitative: the largest constant in the linearised condition, with the maximum norm also on the parameter space, is the value of a small optimisation problem that can be solved by $2p$ linear programmes (Proposition 9). This gives an a posteriori numerical check that a computed solution is consistent with the hypothesis of the theorem.

5. For solutions with fewer than $p+1$ extremal points, where the theorem does not apply, we give the exact second derivatives of the prediction errors by a further recursion and use them in a second stage: Newton's method on the optimality conditions of the extremal set, started from the Gauss–Newton iterates (Section 5). Two-stage methods of this kind were described by Hald and Madsen [7, 8], whose second stage is a quasi-Newton iteration; Newton's method with second derivatives on these equations is treated by Hettich and Zencke [10, Section 5.4], who refer to [9] for Chebyshev approximation. The switching rule used here is heuristic and is not that of [8].

6. Simulations show how often the computed solutions pass the check of item 4, what happens when they do not, and what the minimax gain costs in mean-square terms (Section 6).

We do not claim that the minimax gain predicts better in the mean-square sense than the least-squares gain. In the experiments of Section 6 the median stationary standard deviation of its one-step prediction errors was larger than that of the least-squares gain, under Gaussian and under uniform noise. The maximum norm is appropriate when a bound on the worst prediction error over the record is what is wanted. It depends only on the largest errors, so single outlying measurements can change the solution strongly.

2 The approximation problem

Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{C} \in \mathbb{R}^{1\times n}$ be known and consider

$$\boldsymbol{x}_{k+1} = \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{w}_k, \qquad y_k = \boldsymbol{C}\boldsymbol{x}_k + v_k, \qquad k = 0, 1, \ldots, \tag{1}$$

with unobserved disturbances $\boldsymbol{w}_k$ and v_k. No distributional assumption is needed for the approximation problem; the assumptions on the disturbances used in the simulations are stated in Section 6. For a gain $\boldsymbol{K} \in \mathbb{R}^n$ and an initial estimate $\hat{\boldsymbol{x}}_0$ the constant-gain predictor is

$$\hat{\boldsymbol{x}}_{k+1} = \boldsymbol{A}\hat{\boldsymbol{x}}_k + \boldsymbol{K}e_k, \qquad e_k = y_k - \boldsymbol{C}\hat{\boldsymbol{x}}_k. \tag{2}$$

The numbers $e_k = e_k(\boldsymbol{K})$ are the one-step prediction errors (innovations; for a gain other than the optimal one they are in general not white) and $\hat{y}_k = \boldsymbol{C}\hat{\boldsymbol{x}}_k$ the predictions. In filter form, with a gain $\boldsymbol{K}^{\mathrm{f}}$ applied to the measurement update, the same predictions arise for $\boldsymbol{K} = \boldsymbol{A}\boldsymbol{K}^{\mathrm{f}}$. Given a record $y_0, \ldots, y_{N-1}$ and a fixed $\hat{\boldsymbol{x}}_0$, put

$$\boldsymbol{e}(\boldsymbol{K}) = \big(e_0(\boldsymbol{K}), \ldots, e_{N-1}(\boldsymbol{K})\big)^{\mathsf{T}} \in \mathbb{R}^N.$$

With $\boldsymbol{A}_{\boldsymbol{K}} = \boldsymbol{A} - \boldsymbol{K}\boldsymbol{C}$ the recursion (2) reads $\hat{\boldsymbol{x}}_{k+1} = \boldsymbol{A}_{\boldsymbol{K}}\hat{\boldsymbol{x}}_k + \boldsymbol{K}y_k$, hence

$$\hat{\boldsymbol{x}}_k = \boldsymbol{A}_{\boldsymbol{K}}^{k}\,\hat{\boldsymbol{x}}_0 + \sum_{j=0}^{k-1} \boldsymbol{A}_{\boldsymbol{K}}^{k-1-j}\boldsymbol{K}\,y_j, \tag{3}$$

and e_k is a polynomial of degree at most k in the components of $\boldsymbol{K}$.

Definition 1. The *Chebyshev gain problem* is

$$\min_{\boldsymbol{K}\in\mathbb{R}^n} \ \varphi(\boldsymbol{K}), \qquad \varphi(\boldsymbol{K}) = \|\boldsymbol{e}(\boldsymbol{K})\|_\infty = \max_{0\le k<N} |e_k(\boldsymbol{K})|. \tag{4}$$

For $N \ge 3$ and $\lambda \ge 0$ the *composite problem* is

$$\min_{\boldsymbol{K}\in\mathbb{R}^n} \ \varphi_\lambda(\boldsymbol{K}), \qquad \varphi_\lambda(\boldsymbol{K}) = \|\boldsymbol{e}(\boldsymbol{K})\|_\infty + \lambda\,\|\boldsymbol{D}\hat{\boldsymbol{y}}(\boldsymbol{K})\|_\infty, \tag{5}$$

where $\hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{e}$ is the vector of predictions and $\boldsymbol{D} \in \mathbb{R}^{(N-2)\times N}$ is the second-difference matrix, $(\boldsymbol{D}\boldsymbol{z})_k = z_k - 2z_{k+1} + z_{k+2}$.

Both problems have the form $\min_{\boldsymbol{\theta}} \|\boldsymbol{F}(\boldsymbol{\theta})\|$ with a smooth map $\boldsymbol{F}$ and a norm: for (4), $\boldsymbol{F} = \boldsymbol{e}$ with the maximum norm; for (5),

$$\boldsymbol{F}_\lambda(\boldsymbol{K}) = \begin{pmatrix} \boldsymbol{e}(\boldsymbol{K}) \\ \boldsymbol{D}\hat{\boldsymbol{y}}(\boldsymbol{K}) \end{pmatrix} \in \mathbb{R}^N \times \mathbb{R}^{N-2}, \qquad \|(\boldsymbol{u}, \boldsymbol{v})\|_\lambda = \|\boldsymbol{u}\|_\infty + \lambda\|\boldsymbol{v}\|_\infty. \tag{6}$$

For $\lambda > 0$ this is a norm on $\mathbb{R}^{2N-2}$. The smoothness term acts on the predictions $\hat{\boldsymbol{y}}$, not on the errors, and the two terms are added after the maxima have been taken, so the criterion is not the maximum norm of the stacked vector $(\boldsymbol{u}, \lambda\boldsymbol{v})$. It can be written as a maximum norm of a longer vector, since $\|\boldsymbol{u}\|_\infty + \lambda\|\boldsymbol{v}\|_\infty = \max_{i,j} \max(|u_i + \lambda v_j|, |u_i - \lambda v_j|)$; this representation has $2N(N-2)$ components and is not used here.

Remark 2 (Why a maximum norm). If the errors e_k were independent and uniformly distributed on an interval $[-a, a]$ with unknown a, the likelihood of a record would be $(2a)^{-N}$ for $a \geq \max_k |e_k|$ and zero otherwise, and maximising it over the gain would be problem (4) (the map from the measurements to the errors is unit lower triangular, so no Jacobian factor appears). This is only a motivation: the innovations of a linear predictor driven by uniformly distributed disturbances are weighted sums of such variables and are not uniform. The criterion is best read as a bound: the minimiser is the gain with the smallest worst-case prediction error on the record.

Remark 3 (Stability). On a finite record φ is defined for every $\boldsymbol{K}$, including gains for which $\boldsymbol{A}_{\boldsymbol{K}}$ has spectral radius ≥ 1. A minimiser need not be stabilising. The spectral radius should be checked after the computation; in Sections 6.1 and 6.2 it is reported.

3 Derivatives

Proposition 4. *Let $\boldsymbol{S}_k \in \mathbb{R}^{n \times n}$ be defined by*

$$\boldsymbol{S}_0 = \boldsymbol{0}, \qquad \boldsymbol{S}_{k+1} = (\boldsymbol{A} - \boldsymbol{K}\boldsymbol{C})\, \boldsymbol{S}_k + e_k(\boldsymbol{K})\, \boldsymbol{I}_n. \tag{7}$$

Then $\partial \hat{\boldsymbol{x}}_k / \partial \boldsymbol{K} = \boldsymbol{S}_k$ and the gradient of e_k with respect to $\boldsymbol{K}$ is the row vector $-\boldsymbol{C}\boldsymbol{S}_k$. Thus the Jacobian $\boldsymbol{J}(\boldsymbol{K}) \in \mathbb{R}^{N \times n}$ of $\boldsymbol{e}$ has the rows $-\boldsymbol{C}\boldsymbol{S}_k$, $k = 0, \ldots, N-1$. If $\hat{\boldsymbol{x}}_0$ is also treated as a parameter, then $\partial \hat{\boldsymbol{x}}_k / \partial \hat{\boldsymbol{x}}_0 = (\boldsymbol{A} - \boldsymbol{K}\boldsymbol{C})^k$ and $\partial e_k / \partial \hat{\boldsymbol{x}}_0 = -\boldsymbol{C}(\boldsymbol{A} - \boldsymbol{K}\boldsymbol{C})^k$.

Proof. Differentiate $\hat{\boldsymbol{x}}_{k+1} = \boldsymbol{A}\hat{\boldsymbol{x}}_k + \boldsymbol{K} e_k$ in a direction $\boldsymbol{h} \in \mathbb{R}^n$ of the gain. With $\delta \hat{\boldsymbol{x}}_k = (\partial \hat{\boldsymbol{x}}_k / \partial \boldsymbol{K})\boldsymbol{h}$ and $\delta e_k = -\boldsymbol{C}\, \delta \hat{\boldsymbol{x}}_k$,

$$\delta \hat{\boldsymbol{x}}_{k+1} = \boldsymbol{A}\, \delta \hat{\boldsymbol{x}}_k + \boldsymbol{h}\, e_k + \boldsymbol{K}\, \delta e_k = (\boldsymbol{A} - \boldsymbol{K}\boldsymbol{C})\, \delta \hat{\boldsymbol{x}}_k + e_k\, \boldsymbol{h},$$

and $\delta \hat{\boldsymbol{x}}_0 = \boldsymbol{0}$ because $\hat{\boldsymbol{x}}_0$ is fixed. This is (7) applied to $\boldsymbol{h}$. For the initial state, $\boldsymbol{e}$ depends on $\hat{\boldsymbol{x}}_0$ only through the first term of (3). □

Since $\boldsymbol{e}$ is a polynomial map, $\boldsymbol{J}$ is Lipschitz continuous on every bounded set. The same holds for the Jacobian of $\boldsymbol{F}_\lambda$, which is obtained from $\boldsymbol{J}$ by $\partial(\boldsymbol{D}\hat{\boldsymbol{y}})/\partial \boldsymbol{K} = -\boldsymbol{D}\boldsymbol{J}$.

4 The Gauss–Newton method and its local convergence

In this section $\boldsymbol{F} : U \to \mathbb{R}^M$ is continuously differentiable on an open convex set $U \subset \mathbb{R}^p$, with Jacobian $\boldsymbol{J}$; $\|\cdot\|$ is an arbitrary norm on $\mathbb{R}^M$ and $|\cdot|$ an arbitrary norm on $\mathbb{R}^p$. We write $\varphi(\boldsymbol{\theta}) = \|\boldsymbol{F}(\boldsymbol{\theta})\|$ and assume, with a constant $L > 0$,

$$\|(\boldsymbol{J}(\boldsymbol{\theta}) - \boldsymbol{J}(\boldsymbol{\theta}'))\boldsymbol{d}\| \leq L\, |\boldsymbol{\theta} - \boldsymbol{\theta}'|\, |\boldsymbol{d}| \qquad (\boldsymbol{\theta}, \boldsymbol{\theta}' \in U,\ \boldsymbol{d} \in \mathbb{R}^p), \tag{8}$$

which implies the Taylor estimate

$$\|\boldsymbol{F}(\boldsymbol{\theta} + \boldsymbol{d}) - \boldsymbol{F}(\boldsymbol{\theta}) - \boldsymbol{J}(\boldsymbol{\theta})\boldsymbol{d}\| \leq \tfrac{L}{2} |\boldsymbol{d}|^2 \qquad (\boldsymbol{\theta},\ \boldsymbol{\theta} + \boldsymbol{d} \in U). \tag{9}$$

Method. Given $\boldsymbol{\theta}_j$, let $\boldsymbol{d}_j$ be a minimiser of the convex function

$$g_{\boldsymbol{\theta}_j}(\boldsymbol{d}) = \|\boldsymbol{F}(\boldsymbol{\theta}_j) + \boldsymbol{J}(\boldsymbol{\theta}_j)\boldsymbol{d}\| \tag{10}$$

and set $\boldsymbol{\theta}_{j+1} = \boldsymbol{\theta}_j + \boldsymbol{d}_j$. For the maximum norm, (10) is the linear programme

$$\min_{\boldsymbol{d}, t}\ t \quad \text{subject to} \quad -t \leq F_i(\boldsymbol{\theta}_j) + (\boldsymbol{J}(\boldsymbol{\theta}_j)\boldsymbol{d})_i \leq t \quad (i = 1, \ldots, M),$$

and for the norm (6) it is a linear programme with two bound variables, $\min t + \lambda s$, with the constraints $|\cdot| \leq t$ on the first block and $|\cdot| \leq s$ on the second.

Definition 5. A point $\boldsymbol{\theta}^* \in U$ satisfies *linearised strong uniqueness* with constant $\gamma > 0$ if

$$\|\boldsymbol{F}(\boldsymbol{\theta}^*) + \boldsymbol{J}(\boldsymbol{\theta}^*)\boldsymbol{d}\| \;\geq\; \varphi(\boldsymbol{\theta}^*) + \gamma\,|\boldsymbol{d}| \qquad \text{for all } \boldsymbol{d} \in \mathbb{R}^p. \tag{11}$$

A point $\boldsymbol{\theta}^* \in U$ is called a *strongly unique* local minimiser of φ if there is $\gamma > 0$ with $\varphi(\boldsymbol{\theta}) \geq \varphi(\boldsymbol{\theta}^*) + \gamma|\boldsymbol{\theta} - \boldsymbol{\theta}^*|$ for all $\boldsymbol{\theta}$ in a neighbourhood of $\boldsymbol{\theta}^*$. By Lemma 6 below this holds if and only if $\boldsymbol{\theta}^*$ satisfies linearised strong uniqueness with some constant.

Lemma 6. *Let $\boldsymbol{\theta}^* \in U$ and $\gamma > 0$.*

(i) If (11) holds, then $\varphi(\boldsymbol{\theta}) \geq \varphi(\boldsymbol{\theta}^*) + \frac{\gamma}{2}|\boldsymbol{\theta} - \boldsymbol{\theta}^*|$ for all $\boldsymbol{\theta} \in U$ with $|\boldsymbol{\theta} - \boldsymbol{\theta}^*| \leq \gamma/L$. In particular $\boldsymbol{\theta}^*$ is a strict local minimiser of φ.

(ii) If $\varphi(\boldsymbol{\theta}) \geq \varphi(\boldsymbol{\theta}^*) + \gamma|\boldsymbol{\theta} - \boldsymbol{\theta}^*|$ for all $\boldsymbol{\theta}$ in a neighbourhood of $\boldsymbol{\theta}^*$, then (11) holds with the same γ.

Proof. Write $g_*(\boldsymbol{d}) = \|\boldsymbol{F}(\boldsymbol{\theta}^*) + \boldsymbol{J}(\boldsymbol{\theta}^*)\boldsymbol{d}\|$ and $\varphi^* = \varphi(\boldsymbol{\theta}^*)$.

(i) Let $\boldsymbol{a} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$. By (9) and (11),

$$\varphi(\boldsymbol{\theta}) \;\geq\; g_*(\boldsymbol{a}) - \tfrac{L}{2}|\boldsymbol{a}|^2 \;\geq\; \varphi^* + \gamma|\boldsymbol{a}| - \tfrac{L}{2}|\boldsymbol{a}|^2,$$

and $\frac{L}{2}|\boldsymbol{a}|^2 \leq \frac{\gamma}{2}|\boldsymbol{a}|$ for $|\boldsymbol{a}| \leq \gamma/L$.

(ii) Fix $\boldsymbol{d}$. For small $t \in (0,1]$ the point $\boldsymbol{\theta}^* + t\boldsymbol{d}$ lies in the neighbourhood, so by (9)

$$g_*(t\boldsymbol{d}) \;\geq\; \varphi(\boldsymbol{\theta}^* + t\boldsymbol{d}) - \tfrac{L}{2}t^2|\boldsymbol{d}|^2 \;\geq\; \varphi^* + \gamma t|\boldsymbol{d}| - \tfrac{L}{2}t^2|\boldsymbol{d}|^2.$$

The function g_* is convex and $g_*(\boldsymbol{0}) = \varphi^*$, so $g_*(t\boldsymbol{d}) \leq (1-t)\varphi^* + t\,g_*(\boldsymbol{d})$. Combining the two inequalities and dividing by t gives $g_*(\boldsymbol{d}) \geq \varphi^* + \gamma|\boldsymbol{d}| - \frac{L}{2}t|\boldsymbol{d}|^2$; let $t \to 0$. □

Theorem 7 (local quadratic convergence)**.** *Let $\boldsymbol{\theta}^* \in U$ satisfy (11) with constant γ, let (8) hold, and let the ball $B = \{\boldsymbol{\theta} : |\boldsymbol{\theta} - \boldsymbol{\theta}^*| \leq \gamma/(7L)\}$ be contained in U. Then for every $\boldsymbol{\theta}_0 \in B$ the Gauss–Newton iteration is well defined for every choice of the minimisers $\boldsymbol{d}_j$, stays in B, and satisfies*

$$|\boldsymbol{\theta}_{j+1} - \boldsymbol{\theta}^*| \;\leq\; \frac{7L}{2\gamma}\,|\boldsymbol{\theta}_j - \boldsymbol{\theta}^*|^2 \;\leq\; \tfrac{1}{2}\,|\boldsymbol{\theta}_j - \boldsymbol{\theta}^*|. \tag{12}$$

Proof. Let $\boldsymbol{\theta} \in B$, $\boldsymbol{a} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$, and $g_{\boldsymbol{\theta}}(\boldsymbol{d}) = \|\boldsymbol{F}(\boldsymbol{\theta}) + \boldsymbol{J}(\boldsymbol{\theta})\boldsymbol{d}\|$, g_* as before.

Existence of a minimiser. From (11) and $g_*(\boldsymbol{d}) \leq \varphi^* + \|\boldsymbol{J}(\boldsymbol{\theta}^*)\boldsymbol{d}\|$ we get $\|\boldsymbol{J}(\boldsymbol{\theta}^*)\boldsymbol{d}\| \geq \gamma|\boldsymbol{d}|$, and with (8) $\|\boldsymbol{J}(\boldsymbol{\theta})\boldsymbol{d}\| \geq (\gamma - L|\boldsymbol{a}|)|\boldsymbol{d}| \geq \frac{6}{7}\gamma|\boldsymbol{d}|$. Hence $g_{\boldsymbol{\theta}}(\boldsymbol{d}) \to \infty$ as $|\boldsymbol{d}| \to \infty$, and the continuous function $g_{\boldsymbol{\theta}}$ attains its minimum.

Comparison of the two linearisations. For every $\boldsymbol{d} \in \mathbb{R}^p$,

$$\boldsymbol{F}(\boldsymbol{\theta}) + \boldsymbol{J}(\boldsymbol{\theta})\boldsymbol{d} - \big[\boldsymbol{F}(\boldsymbol{\theta}^*) + \boldsymbol{J}(\boldsymbol{\theta}^*)(\boldsymbol{a} + \boldsymbol{d})\big] = \big[\boldsymbol{F}(\boldsymbol{\theta}) - \boldsymbol{F}(\boldsymbol{\theta}^*) - \boldsymbol{J}(\boldsymbol{\theta}^*)\boldsymbol{a}\big] + \big[\boldsymbol{J}(\boldsymbol{\theta}) - \boldsymbol{J}(\boldsymbol{\theta}^*)\big]\boldsymbol{d},$$

so by (9) and (8)

$$\big|g_{\boldsymbol{\theta}}(\boldsymbol{d}) - g_*(\boldsymbol{a} + \boldsymbol{d})\big| \leq \tfrac{L}{2}|\boldsymbol{a}|^2 + L|\boldsymbol{a}||\boldsymbol{d}|. \tag{13}$$

Error estimate. Let $\boldsymbol{d}$ minimise $g_{\boldsymbol{\theta}}$ and put $\boldsymbol{a}^+ = \boldsymbol{a} + \boldsymbol{d}$. Using (11), then (13), then the minimality of $\boldsymbol{d}$ against the competitor $-\boldsymbol{a}$, then (13) once more with $-\boldsymbol{a}$ in place of $\boldsymbol{d}$, and $g_*(\boldsymbol{0}) = \varphi^*$,

$$\begin{aligned}
\varphi^* + \gamma|\boldsymbol{a}^+| &\leq g_*(\boldsymbol{a} + \boldsymbol{d}) \leq g_{\boldsymbol{\theta}}(\boldsymbol{d}) + \tfrac{L}{2}|\boldsymbol{a}|^2 + L|\boldsymbol{a}||\boldsymbol{d}| \\
&\leq g_{\boldsymbol{\theta}}(-\boldsymbol{a}) + \tfrac{L}{2}|\boldsymbol{a}|^2 + L|\boldsymbol{a}||\boldsymbol{d}| \\
&\leq \varphi^* + \tfrac{L}{2}|\boldsymbol{a}|^2 + L|\boldsymbol{a}|^2 + \tfrac{L}{2}|\boldsymbol{a}|^2 + L|\boldsymbol{a}||\boldsymbol{d}|.
\end{aligned}$$

With $|\boldsymbol{d}| \le |\boldsymbol{a}^+| + |\boldsymbol{a}|$ this gives $\gamma|\boldsymbol{a}^+| \le 3L|\boldsymbol{a}|^2 + L|\boldsymbol{a}||\boldsymbol{a}^+|$, that is

$$|\boldsymbol{a}^+| \le \frac{3L|\boldsymbol{a}|^2}{\gamma - L|\boldsymbol{a}|} \le \frac{7L}{2\gamma}|\boldsymbol{a}|^2 \le \tfrac{1}{2}|\boldsymbol{a}|,$$

because $\gamma - L|\boldsymbol{a}| \ge \frac{6}{7}\gamma$ and $|\boldsymbol{a}| \le \gamma/(7L)$. In particular $\boldsymbol{\theta} + \boldsymbol{d} \in B$, and the argument can be repeated. □

Theorem 7 is Theorem 2.1 of Jittorntrum and Osborne [11], stated with a Lipschitz constant of the Jacobian in place of their constants. Their proof, like the one above, compares the linearisations at the iterate and at the solution and uses the solution as a competitor in the linear subproblem; they also note that the step need not be an exact minimiser, as long as it is no worse than that competitor. They describe the theorem as essentially the main result of Cromme [6]; for the maximum norm on a space of continuous functions on a compact set and a regular derivative it is Satz 1 of [5]. Lemma 6 corresponds to [11, Lemma 2.1]. The proof uses only the triangle inequality, (9) and (8), and therefore no property of the particular norm.

The condition $B \subset U$ is no restriction when (8) holds on a ball U of radius r around $\boldsymbol{\theta}^*$: a Lipschitz constant may be enlarged, and with $L > \gamma/(7r)$ the closed ball B lies in the open ball U. The constants are not sharp. The theorem concerns the undamped iteration and says nothing about the size of the neighbourhood in a given application, since L is rarely known.

Corollary 8 (composite criterion)**.** *Let $\lambda > 0$ and let $\boldsymbol{K}^*$ satisfy linearised strong uniqueness* (11) *for the map $\boldsymbol{F}_\lambda$ and the norm $\|\cdot\|_\lambda$ of* (6)*. Then the Gauss–Newton iteration for the composite problem* (5) *converges quadratically to $\boldsymbol{K}^*$ from every starting point in a neighbourhood of $\boldsymbol{K}^*$, with the constants of Theorem 7.*

Proof. $\boldsymbol{F}_\lambda$ is a polynomial map, so (8) holds on the unit ball U around $\boldsymbol{K}^*$ with some constant L_U. Theorem 7 applies with $L = \max(L_U, 2\gamma/7)$, for which B has radius at most $\frac{1}{2}$ and $B \subset U$. □

The hypothesis of the corollary concerns the composite problem itself. A minimiser of (5) is in general different from a minimiser of (4), and strong uniqueness of one does not imply strong uniqueness of the other.

4.1 A computable certificate for the maximum norm

Let $\|\cdot\| = \|\cdot\|_\infty$, let $\varphi^* = \varphi(\boldsymbol{\theta}^*) > 0$, and let

$$E = \{i : |F_i(\boldsymbol{\theta}^*)| = \varphi^*\}, \qquad \sigma_i = \operatorname{sign} F_i(\boldsymbol{\theta}^*) \quad (i \in E)$$

be the extremal set and its signs. Denote by $\boldsymbol{j}_i$ the i-th row of $\boldsymbol{J}(\boldsymbol{\theta}^*)$.

Proposition 9. *Define*

$$\gamma_E = \min_{|\boldsymbol{d}|=1} \max_{i \in E} \sigma_i \boldsymbol{j}_i \boldsymbol{d}. \tag{14}$$

Then (11) *holds for some $\gamma > 0$ if and only if $\gamma_E > 0$, and in that case γ_E is the largest admissible constant. A necessary condition for $\gamma_E > 0$ is $|E| \ge p + 1$. If $|\cdot| = \|\cdot\|_\infty$ on $\mathbb{R}^p$, then γ_E is the smallest of the optimal values of the $2p$ linear programmes*

$$\min_{\boldsymbol{d},t} t \quad \text{subject to} \quad \sigma_i \boldsymbol{j}_i \boldsymbol{d} \le t \ (i \in E), \quad d_l = \pm 1, \quad -1 \le d_m \le 1 \ (m \ne l),$$

for $l = 1, \dots, p$ and both signs.

Proof. For $i \in E$ and every $\boldsymbol{d}$, $g_*(\boldsymbol{d}) \geq \sigma_i\big(F_i(\boldsymbol{\theta}^*)+\boldsymbol{j}_i\boldsymbol{d}\big) = \varphi^*+\sigma_i\boldsymbol{j}_i\boldsymbol{d}$, hence $g_*(\boldsymbol{d}) \geq \varphi^*+\gamma_E|\boldsymbol{d}|$ by homogeneity. Conversely, for fixed $\boldsymbol{d}$ and all sufficiently small $t > 0$ the maximum in $g_*(t\boldsymbol{d})$ is attained on E with the signs σ_i, so $g_*(t\boldsymbol{d}) = \varphi^* + t\max_{i\in E}\sigma_i\boldsymbol{j}_i\boldsymbol{d}$; if (11) holds with γ, this gives $\max_{i\in E}\sigma_i\boldsymbol{j}_i\boldsymbol{d} \geq \gamma|\boldsymbol{d}|$ and thus $\gamma_E \geq \gamma$.

Let $\boldsymbol{G}$ be the matrix with rows $\sigma_i\boldsymbol{j}_i$, $i \in E$, and suppose $|E| \leq p$. If $\operatorname{rank}\boldsymbol{G} < p$, there is $\boldsymbol{d} \neq \boldsymbol{0}$ with $\boldsymbol{G}\boldsymbol{d} = \boldsymbol{0}$, so $\gamma_E \leq 0$. Otherwise $|E| = p$ and $\boldsymbol{G}$ is invertible; the solution of $\boldsymbol{G}\boldsymbol{d} = -(1,\dots,1)^{\mathsf{T}}$ gives a negative maximum, so $\gamma_E < 0$.

The unit sphere of $\|\cdot\|_\infty$ is the union of the $2p$ faces $\{d_l = \pm 1,\ |d_m| \leq 1\}$, and on each face the inner maximum is minimised by the stated linear programme. □

The equivalence is the maximum-norm case of the subgradient characterisation of strong uniqueness in [11, Theorem 2.2], and the necessary condition $|E| \geq p+1$ is the maximum-norm case of [11, Theorem 2.3]. What is added here is the identification of γ_E as the largest constant and its computation by linear programmes.

If $\varphi^* = 0$, condition (11) reads $\|\boldsymbol{J}(\boldsymbol{\theta}^*)\boldsymbol{d}\| \geq \gamma|\boldsymbol{d}|$ and holds for some $\gamma > 0$ exactly when $\boldsymbol{J}(\boldsymbol{\theta}^*)$ is injective.

The certificate is a numerical check, not a proof. It is evaluated at a computed point, not at $\boldsymbol{\theta}^*$; the set E is determined with a tolerance; the linear programmes are solved in floating-point arithmetic with the tolerances of the solver, and no error bound for the computed value of γ_E is available; and γ_E does not depend continuously on the point, because E can change. A positive value says that the computed solution is consistent with the hypothesis of Theorem 7. It does not show that the iterates were inside the ball B.

Remark 10 (Solutions with fewer extremal points). For a nonlinear map a local minimiser of φ can have fewer than $p+1$ extremal points. With $p = 2$, for instance, φ can have a minimum on the curve $|e_i| = |e_k|$ at a point where the two signed gradients are antiparallel; whether this point is a minimum is then decided by second derivatives. Such a solution is not strongly unique [11, Theorem 2.3], and Theorem 7 does not apply. Without strong uniqueness the undamped Gauss–Newton method can still converge quadratically in special cases [11, Section 3], but in general it can be slow, or fail to converge unless a line search is added [11, remark on p. 454, for polyhedral norms]; see also [6, Example 5]. Methods that use second-derivative information are a known remedy [6, Example 6 and Lemma 4.2], [8, p. 49]; Section 5 applies one to the present problem. Section 6 shows that this case is not rare in the present application.

Remark 11 (Initial state as a parameter). By Proposition 4, $\hat{\boldsymbol{x}}_0$ can be appended to the parameter vector at the cost of one further matrix recursion, and $\boldsymbol{e}$ is affine in $\hat{\boldsymbol{x}}_0$ for fixed $\boldsymbol{K}$. For a stabilising gain, however, $\partial e_k/\partial\hat{\boldsymbol{x}}_0$ decays geometrically in k, so extremal points at late times contribute almost nothing in these directions. In the experiments (Section 6) the extremal set typically consisted of one or two of the earliest times, whose errors depend directly on $\hat{\boldsymbol{x}}_0$, and of late times, whose errors hardly depend on it; the early errors can then be changed through $\hat{\boldsymbol{x}}_0$ with little effect on the late ones, and this can lead to a very small computed constant γ_E; whether several early extremal errors can be lowered together depends on their signs and derivatives. We therefore keep $\hat{\boldsymbol{x}}_0$ fixed in the main experiments.

5 Solutions with fewer than $p+1$ extremal points

This section treats the case left open by Theorem 7: the maximum norm, $\boldsymbol{F} = \boldsymbol{e}$, and a local minimiser $\boldsymbol{\theta}^*$ of φ with $\varphi^* > 0$ whose extremal set E has $m = |E| \leq p$ elements. Nothing in it is new in substance. Second-order conditions for minimax problems are standard; see [10, Theorem 3.3.14] for the semi-infinite case. Cromme [6, Example 6 and Lemma 4.2] treats solutions that are unique only to second order by a method with second derivatives. Newton's

method on the Kuhn–Tucker equations of the active set, with second derivatives of the Lagrangian, is Method 5.4.1 of Hettich and Zencke [10], stated there for semi-infinite problems and traced to [9] for Chebyshev approximation; their Theorem 5.4.2 gives local superlinear convergence under a second-order sufficient condition and a rank condition; the regularity of the matrix of the system is taken there from their Theorem 3.3.8 (Remark 3.3.16), and Corollary 3.3.15 gives local uniqueness of the solution. Proposition 15 below is the special case of a finite index set with strictly positive multipliers; the quadratic rate stated there is that of the standard local convergence theorem for Newton's method. The combination of a linear-programming stage with a second stage on these equations is the two-stage method of Hald and Madsen [7, 8]; the following description is taken from [8]. Their second stage does not use second derivatives: the corresponding block of the Jacobian is started by finite differences and then updated by a quasi-Newton formula. They switch to the second stage when the step of the first stage lies on the bound of the trust region, the set of nearly active functions has been the same at the latest three different first-stage iterates, and least-squares multipliers are nonnegative with a small residual in the stationarity equation; they switch back when the residual of the system is not reduced by a fixed factor, a function outside the active set becomes the largest, or a multiplier becomes negative. For these rules they state that a convergent sequence has a stationary limit (their Theorem 1, proved in [7]) and prove that, at a non-degenerate limit, for suitably chosen parameters, a sufficiently accurate difference approximation and a locally linearly convergent quasi-Newton update, the iteration ends either in the first stage with quadratic convergence or in the second stage. Non-degeneracy in their sense implies (S1) and (S2) below [8, Proposition 1]. We use exact second derivatives instead of quasi-Newton estimates and a different, heuristic switching rule, so that their convergence results do not carry over to the method below. What this section contributes is the recursion for the second derivatives of the innovations, stated explicitly for this predictor without a claim of novelty, a simple heuristic switching rule, and a test on the simulated records. In this section λ_i and $\boldsymbol{\lambda}$ denote multipliers; the weight λ of (5) does not occur.

Throughout this section $|\cdot|$ is the Euclidean norm on $\mathbb{R}^p$. We write $\boldsymbol{g}_i = \sigma_i \nabla e_i(\boldsymbol{\theta}^*)$ for $i \in E$, $\boldsymbol{G} \in \mathbb{R}^{m \times p}$ for the matrix with the rows $\boldsymbol{g}_i^{\mathsf{T}}$, and $\boldsymbol{H}_i = \nabla^2 e_i(\boldsymbol{\theta}^*)$.

Definition 12. The point $\boldsymbol{\theta}^*$ satisfies the *second-order conditions* if

(S1) there are multipliers $\lambda_i > 0$ $(i \in E)$ with $\sum_{i \in E} \lambda_i = 1$ and $\sum_{i \in E} \lambda_i \boldsymbol{g}_i = \boldsymbol{0}$;

(S2) the vectors $(\boldsymbol{g}_i, -1) \in \mathbb{R}^{p+1}$, $i \in E$, are linearly independent;

(S3) with $\boldsymbol{W} = \sum_{i \in E} \lambda_i \sigma_i \boldsymbol{H}_i$ there is $\mu > 0$ such that $\boldsymbol{d}^{\mathsf{T}} \boldsymbol{W} \boldsymbol{d} \geq \mu |\boldsymbol{d}|^2$ for all $\boldsymbol{d}$ with $\boldsymbol{G}\boldsymbol{d} = \boldsymbol{0}$; the largest such μ is called the *reduced curvature*.

By (S1) the rows of $\boldsymbol{G}$ are linearly dependent, so $\operatorname{rank} \boldsymbol{G} \leq m - 1 \leq p - 1$ and the null space of $\boldsymbol{G}$ is not trivial. For $p = m = 2$ the conditions say that the two signed gradients are antiparallel and not zero, and that $\boldsymbol{z}^{\mathsf{T}} \boldsymbol{W} \boldsymbol{z} > 0$ for the direction $\boldsymbol{z}$ orthogonal to them.

Proposition 13 (second derivatives)**.** *Let $\boldsymbol{s}_{i,k}$ be the i-th column of $\boldsymbol{S}_k$ in* (7)*, let $\boldsymbol{u}_i$ be the i-th unit vector of $\mathbb{R}^n$, and define $\boldsymbol{h}_{ij,k} \in \mathbb{R}^n$ by $\boldsymbol{h}_{ij,0} = \boldsymbol{0}$ and*

$$\boldsymbol{h}_{ij,k+1} = (\boldsymbol{A} - \boldsymbol{K}\boldsymbol{C})\, \boldsymbol{h}_{ij,k} - (\boldsymbol{C}\boldsymbol{s}_{i,k})\, \boldsymbol{u}_j - (\boldsymbol{C}\boldsymbol{s}_{j,k})\, \boldsymbol{u}_i. \tag{15}$$

Then $\partial^2 \hat{\boldsymbol{x}}_k / \partial K_i \partial K_j = \boldsymbol{h}_{ij,k}$ and $\partial^2 e_k / \partial K_i \partial K_j = -\boldsymbol{C}\boldsymbol{h}_{ij,k}$.

Proof. By Proposition 4, $\boldsymbol{s}_{i,k+1} = (\boldsymbol{A} - \boldsymbol{K}\boldsymbol{C})\boldsymbol{s}_{i,k} + e_k \boldsymbol{u}_i$. Differentiating with respect to K_j and using $\partial(\boldsymbol{K}\boldsymbol{C})/\partial K_j = \boldsymbol{u}_j \boldsymbol{C}$ and $\partial e_k / \partial K_j = -\boldsymbol{C}\boldsymbol{s}_{j,k}$ gives (15). □

The recursion costs $O(n^4)$ operations per time step for all pairs (i,j) together. It is stated for $\boldsymbol{\theta} = \boldsymbol{K}$, so $p = n$; the initial state is kept fixed in this section.

Proposition 14 (growth of second order)**.** *Let $\boldsymbol{e}$ be twice continuously differentiable near $\boldsymbol{\theta}^*$ and let* (S1) *and* (S3) *hold. Then there is $r > 0$ such that*

$$\varphi(\boldsymbol{\theta}^* + \boldsymbol{d}) \;\geq\; \varphi^* + \tfrac{\mu}{4}|\boldsymbol{d}|^2 \qquad (|\boldsymbol{d}| \leq r).$$

In particular $\boldsymbol{\theta}^*$ is a strict local minimiser. It is not a strongly unique local minimiser in the sense of Definition 5, because the growth is not linear: $\varphi(\boldsymbol{\theta}^* + t\boldsymbol{z}) - \varphi^* = O(t^2)$ for every $\boldsymbol{z}$ with $\boldsymbol{G}\boldsymbol{z} = \boldsymbol{0}$.

Proof. For $i \in E$ and small $\boldsymbol{d}$,

$$\varphi(\boldsymbol{\theta}^* + \boldsymbol{d}) \;\geq\; \sigma_i e_i(\boldsymbol{\theta}^* + \boldsymbol{d}) = \varphi^* + \boldsymbol{g}_i^{\mathsf{T}}\boldsymbol{d} + \tfrac{1}{2}\boldsymbol{d}^{\mathsf{T}}\sigma_i\boldsymbol{H}_i\boldsymbol{d} + \rho_i(\boldsymbol{d}), \qquad |\rho_i(\boldsymbol{d})| \leq \omega(|\boldsymbol{d}|)|\boldsymbol{d}|^2, \tag{16}$$

with $\omega(t) \to 0$ as $t \to 0$. Let M bound the spectral norms of the $\boldsymbol{H}_i$ and of $\boldsymbol{W}$. Write $\boldsymbol{d} = \boldsymbol{u} + \boldsymbol{v}$ with $\boldsymbol{G}\boldsymbol{u} = \boldsymbol{0}$ and $\boldsymbol{v}$ orthogonal to the null space of $\boldsymbol{G}$. The number

$$c_1 = \min\{\max_{i\in E} \boldsymbol{g}_i^{\mathsf{T}}\boldsymbol{v} :\; \boldsymbol{v} \perp \ker \boldsymbol{G},\; |\boldsymbol{v}| = 1\}$$

is positive (if $\boldsymbol{G} = \boldsymbol{0}$ there is no such $\boldsymbol{v}$ and (18) alone gives the claim): if $\max_i \boldsymbol{g}_i^{\mathsf{T}}\boldsymbol{v} \leq 0$, then $\sum_i \lambda_i \boldsymbol{g}_i^{\mathsf{T}}\boldsymbol{v} = 0$ and $\lambda_i > 0$ give $\boldsymbol{g}_i^{\mathsf{T}}\boldsymbol{v} = 0$ for all i, so $\boldsymbol{v} \in \ker \boldsymbol{G}$ and $\boldsymbol{v} = \boldsymbol{0}$. Taking the maximum over i in (16) gives

$$\varphi(\boldsymbol{\theta}^* + \boldsymbol{d}) \;\geq\; \varphi^* + c_1|\boldsymbol{v}| - \big(\tfrac{M}{2} + \omega\big)|\boldsymbol{d}|^2, \tag{17}$$

and the combination of (16) with the weights λ_i gives, by (S1),

$$\varphi(\boldsymbol{\theta}^* + \boldsymbol{d}) \;\geq\; \varphi^* + \tfrac{1}{2}\boldsymbol{d}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{d} - \omega|\boldsymbol{d}|^2. \tag{18}$$

If $c_1|\boldsymbol{v}| \geq (\frac{M}{2} + \omega + \frac{\mu}{4})|\boldsymbol{d}|^2$, the claim follows from (17). Otherwise $|\boldsymbol{v}| \leq c_2|\boldsymbol{d}|^2$ with a constant c_2, hence $|\boldsymbol{u}|^2 \geq |\boldsymbol{d}|^2(1 - c_2^2|\boldsymbol{d}|^2)$ and, by (S3),

$$\boldsymbol{d}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{d} \;\geq\; \mu|\boldsymbol{u}|^2 - 2M|\boldsymbol{u}||\boldsymbol{v}| - M|\boldsymbol{v}|^2 \;\geq\; \mu|\boldsymbol{d}|^2 - c_3|\boldsymbol{d}|^3$$

for small $\boldsymbol{d}$. With (18) this gives $\varphi(\boldsymbol{\theta}^* + \boldsymbol{d}) - \varphi^* \geq (\frac{\mu}{2} - \omega - \frac{c_3}{2}|\boldsymbol{d}|)|\boldsymbol{d}|^2 \geq \frac{\mu}{4}|\boldsymbol{d}|^2$ for $|\boldsymbol{d}| \leq r$ with r small enough.

For the last statement, $\boldsymbol{g}_i^{\mathsf{T}}\boldsymbol{z} = 0$ for $i \in E$, so $\sigma_i e_i(\boldsymbol{\theta}^* + t\boldsymbol{z}) = \varphi^* + O(t^2)$ for these i, while $|e_k(\boldsymbol{\theta}^* + t\boldsymbol{z})| < \varphi^*$ for $k \notin E$ and small t by continuity. □

5.1 Newton's method on the extremal set

Fix the index set E and the signs σ_i. A point satisfying (S1) is a zero of the map

$$\boldsymbol{\Phi}(\boldsymbol{\theta}, \phi, \boldsymbol{\lambda}) = \begin{pmatrix} \sum_{i\in E} \lambda_i \sigma_i \nabla e_i(\boldsymbol{\theta}) \\ \big(\sigma_i e_i(\boldsymbol{\theta}) - \phi\big)_{i\in E} \\ \sum_{i\in E} \lambda_i - 1 \end{pmatrix} \in \mathbb{R}^{p+m+1}, \tag{19}$$

which has as many components as unknowns.

Proposition 15. *Let $\boldsymbol{e}$ be three times continuously differentiable near $\boldsymbol{\theta}^*$ and let* (S1)–(S3) *hold. Then the Jacobian of $\boldsymbol{\Phi}$ at $(\boldsymbol{\theta}^*, \varphi^*, \boldsymbol{\lambda})$ is nonsingular, and Newton's method for $\boldsymbol{\Phi} = \boldsymbol{0}$ converges quadratically to this point from every starting point sufficiently close to it.*

Proof. The Jacobian is

$$\begin{pmatrix} \boldsymbol{W} & \boldsymbol{0} & \boldsymbol{G}^{\mathsf{T}} \\ \boldsymbol{G} & -\boldsymbol{1} & \boldsymbol{0} \\ \boldsymbol{0}^{\mathsf{T}} & 0 & \boldsymbol{1}^{\mathsf{T}} \end{pmatrix}.$$

Let $(\boldsymbol{d}, \delta, \boldsymbol{\eta})$ be in its null space: $\boldsymbol{W}\boldsymbol{d} + \boldsymbol{G}^{\mathsf{T}}\boldsymbol{\eta} = \boldsymbol{0}$, $\boldsymbol{G}\boldsymbol{d} = \delta\boldsymbol{1}$, $\boldsymbol{1}^{\mathsf{T}}\boldsymbol{\eta} = 0$. Multiplying the second equation by $\boldsymbol{\lambda}^{\mathsf{T}}$ and using (S1) gives $\delta = 0$, so $\boldsymbol{G}\boldsymbol{d} = \boldsymbol{0}$. Multiplying the first by $\boldsymbol{d}^{\mathsf{T}}$ gives $\boldsymbol{d}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{d} = 0$, hence $\boldsymbol{d} = \boldsymbol{0}$ by (S3). Then $\sum_i \eta_i(\boldsymbol{g}_i, -1) = \boldsymbol{0}$, and $\boldsymbol{\eta} = \boldsymbol{0}$ by (S2). The second statement is the local convergence theorem for Newton's method; $\boldsymbol{\Phi}$ has a Lipschitz continuous derivative near the solution. □

Each Newton step needs $\boldsymbol{e}$, $\boldsymbol{J}$ and the second derivatives on E, that is, one pass of the recursions (7) and (15), and the solution of a linear system of order $p + m + 1$. No linear programme is solved. The proposition says nothing about how E is found.

5.2 A two-stage method

The following combination was used in the computations. The first stage is the safeguarded Gauss–Newton iteration of Section 6. In every iteration except the first (iterations are counted by linear programmes, as in Section 6), if the Gauss–Newton step satisfies $\|\boldsymbol{d}\|_\infty \leq 0.1$ (this is the step of the safeguarded method; the condition holds, up to the feasibility tolerance of the solver, whenever its trust-region radius satisfies $\Delta \leq 0.1$), the second stage is considered, unless a stopping test is met, before it is decided whether that step is accepted, for $m = p, p-1, \ldots, 2$:

1. *Candidate.* E consists of the m largest values $|e_i|$, with the signs of these e_i; $\boldsymbol{\lambda}$ is the least-squares solution of $\boldsymbol{G}^{\mathsf{T}}\boldsymbol{\lambda} = \boldsymbol{0}$, $\boldsymbol{1}^{\mathsf{T}}\boldsymbol{\lambda} = 1$.

2. *Pretest.* The candidate is dropped unless $\boldsymbol{\lambda} > \boldsymbol{0}$, $|\boldsymbol{G}^{\mathsf{T}}\boldsymbol{\lambda}| \leq 0.2\max_i|\boldsymbol{g}_i|$, and $\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{Z}$ is positive definite, where the columns of $\boldsymbol{Z}$ are the last $p - m + 1$ columns of $\boldsymbol{V}$ in a full singular value decomposition $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathsf{T}}$ with a $p \times p$ matrix $\boldsymbol{V}$ and the singular values in decreasing order; if $\boldsymbol{G}$ has rank m, these are the right singular vector of the smallest singular value and, for $p > m$, a basis of the kernel of $\boldsymbol{G}$; if $\boldsymbol{G}$ has rank $m-1$, as (S1) and (S2) imply at a solution, they form an orthonormal basis of the kernel of $\boldsymbol{G}$, the subspace of (S3). A reduced decomposition does not provide them for $p > m$.

3. *Newton stage.* At most ten Newton steps for (19) are taken from the current iterate, with this $\boldsymbol{\lambda}$ and with ϕ equal to the mean of the $\sigma_i e_i$, $i \in E$, as starting values.

4. *Acceptance.* The result is accepted, and the method stops, if the residual of (19) is below 10^{-11} (the first block relative to $\max(1, \max_{i,l}|G_{il}|)$, the second relative to ϕ, the third absolutely), all $\lambda_i > 0$, $\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{Z}$ is positive definite at the new point, the e_i with $i \in E$ have kept their signs, no $|e_k|$ with $k \notin E$ exceeds ϕ, and φ is not larger than at the current Gauss–Newton iterate (the last two up to a relative tolerance of 10^{-9}), and the smallest singular value of the matrix with the rows $(\boldsymbol{g}_i^{\mathsf{T}}, -1)$ exceeds 10^{-8}. Otherwise the result is discarded, the next smaller m is tried, and after $m = 2$ the first stage continues.

An accepted point therefore satisfies (S1)–(S3) within these tolerances on an active set E. The test admits $|e_k| = \phi$ for $k \notin E$, so E need not be the whole extremal set. For such a subset the inequality of Proposition 14 remains valid, because its proof uses, besides (S1) and (S3) on E, only $\varphi(\boldsymbol{\theta}^*) = \sigma_i e_i(\boldsymbol{\theta}^*)$ for $i \in E$ and $\varphi \geq \max_{i\in E}\sigma_i e_i$. The last statement of the proposition, that $\boldsymbol{\theta}^*$ is not strongly unique, does not carry over: its proof needs $|e_k(\boldsymbol{\theta}^*)| < \varphi^*$

for $k \notin E$. In the runs of Section 6.3 every accepted set was the whole extremal set. The objective value is not larger, within the tolerance, than that of the iterate from which the second stage started. This is a numerical statement, with the reservations made for γ_E. If no candidate is ever accepted, the method is the first stage unchanged. The thresholds 0.1, 0.2 and ten and the choice of the candidate set are heuristics. We prove no global convergence, and nothing guarantees that the second stage is started when a solution with $m \le p$ extremal points is near. Solutions with a single extremal point ($m = 1$, a stationary point of one e_i) and the composite criterion (5) are not treated.

After the computation, (S1)–(S3) can be checked at the computed point in the same way as γ_E: one computes the extremal set with a tolerance, the least-squares multipliers, the smallest singular value of the matrix with the rows $(\boldsymbol{g}_i^{\mathsf{T}}, -1)$, and the smallest eigenvalue μ of $\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{Z}$. The reservations stated after Proposition 9 apply here as well.

6 Numerical experiments

All computations were done in Python 3.12.16 with NumPy 2.5.5 and SciPy 1.18.1; the linear programmes are solved by the dual simplex method of HiGHS through the SciPy function `linprog`. The programmes and the seeds are part of the supplementary files. The results below are single simulation studies with one model family. They illustrate the statements of Sections 4 and 5; they prove nothing beyond the data sets used.

Model. We use the constant-velocity model with sampling interval one,

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad \boldsymbol{C} = \begin{pmatrix} 1 & 0 \end{pmatrix}, \qquad \boldsymbol{Q} = \operatorname{Cov} \boldsymbol{w}_k = q \begin{pmatrix} 1/3 & 1/2 \\ 1/2 & 1 \end{pmatrix}, \qquad R = \operatorname{Var} v_k = 1,$$

with $\boldsymbol{x}_0 = \hat{\boldsymbol{x}}_0 = \boldsymbol{0}$, so $p = n = 2$. The sequences $(\boldsymbol{w}_k)$ and (v_k) are independent of each other, and each consists of independent, identically distributed variables with mean zero and the covariances above. They are either Gaussian or, in the second case, generated from independent variables that are uniformly distributed on $[-\sqrt{3}, \sqrt{3}]$: v_k is such a variable and $\boldsymbol{w}_k$ is the Cholesky factor of $\boldsymbol{Q}$ applied to a vector of two of them. The stationary variance below uses only the first and second moments. Three gains are compared: the stationary gain from the Riccati equation with the true covariances, the least-squares gain minimising $\sum_k e_k(\boldsymbol{K})^2$, and the minimax gain of (4). For a stabilising gain $\boldsymbol{K}$ the stationary variance of the innovations is $\boldsymbol{C}\boldsymbol{P}_{\boldsymbol{K}}\boldsymbol{C}^{\mathsf{T}} + 1$, where $\boldsymbol{P}_{\boldsymbol{K}}$ solves the Lyapunov equation $\boldsymbol{P} = \boldsymbol{A}_{\boldsymbol{K}}\boldsymbol{P}\boldsymbol{A}_{\boldsymbol{K}}^{\mathsf{T}} + \operatorname{Cov} \boldsymbol{w}_k + \boldsymbol{K}\boldsymbol{K}^{\mathsf{T}}$. We report the square root of this variance relative to its minimum, which is attained by the Riccati gain. This quantity is exact for the model and involves no further simulation. It measures how well a gain predicts in the stationary regime, in the mean-square sense; it is not a measure of the accuracy with which a gain estimates the Riccati gain.

Globalisation. Far from a solution the full Gauss–Newton step can increase φ. In the computations the step is restricted to $\|\boldsymbol{d}\|_\infty \le \Delta$, which keeps the subproblem a linear programme; Δ is replaced by $\frac{1}{2}\min(\Delta, \|\boldsymbol{d}\|_\infty)$ when the actual decrease $\varphi(\boldsymbol{\theta}) - \varphi(\boldsymbol{\theta}+\boldsymbol{d})$ is less than one tenth of the decrease predicted by the linearisation, $\text{pred} = \varphi(\boldsymbol{\theta}) - \|\boldsymbol{F}(\boldsymbol{\theta}) + \boldsymbol{J}(\boldsymbol{\theta})\boldsymbol{d}\|$, and doubled after a step on the boundary whose actual decrease is at least three quarters of the predicted one. A step is taken only if the actual decrease is at least one tenth of the predicted one. The iteration stops when $\text{pred} \le 10^{-12}\max(1, \varphi)$, when $\|\boldsymbol{d}\|_\infty \le 10^{-12}$, when Δ falls below 10^{-14}, and after 200 linear programmes at the latest. The first test indicates that $\boldsymbol{d} = \boldsymbol{0}$ nearly minimises the linearised problem, that is, that the iterate is nearly a

stationary point of φ; so does the second when the step does not lie on the boundary of the trust region. The third test was never met in the runs of Sections 6.1 to 6.5; the smallest radius that occurred was $8.9 \cdot 10^{-8}$ (supplementary programme `stop_check.py`). If the step respects the bound exactly, the test cannot be met: Δ is reduced only after a rejected step, whose length exceeds 10^{-12}, to half the smaller of Δ and that length, so Δ stays above $5 \cdot 10^{-13}$. The solver respects the bound only up to its feasibility tolerance, however: in 4 of the 42 291 linear programmes of these runs $\|\boldsymbol{d}\|_\infty$ exceeded Δ, by at most $8.1 \cdot 10^{-8}$. The test is therefore kept as a safeguard. A stop at the limit indicates nothing. Every linear programme is counted, whether its step is accepted or rejected. The *iteration count* of a run, also called its number of steps, is the number of linear programmes solved minus one, the last programme being the one at which a stopping test is met; it equals the number of accepted steps when no step is rejected. When the bound is inactive and the step is accepted, the step is that of Section 4. The safeguard is standard and is not analysed here; Theorem 7 concerns the iteration without it, and all iteration counts below are counts of the safeguarded method. All runs start from $\boldsymbol{K}_0 = (0.5,\, 0.1)^{\mathsf{T}}$ with $\Delta = 0.5$.

6.1 One data set

For $q = 1$, $N = 200$ and Gaussian disturbances (seed 2), the iteration stops after seven steps (Table 1); the bound was never active, so in this run every step is a pure Gauss–Newton step. The solution has the three extremal indices $k = 9, 19, 59$ with signs $-, -, +$, and the certificate of Proposition 9 gives $\gamma_E = 0.214$ in the maximum norm of the parameter space. The errors of the last steps, $7.6 \cdot 10^{-3}$, $7.9 \cdot 10^{-5}$, $2.0 \cdot 10^{-8}$, show the quadratic rate. A general sequential quadratic programming code (SLSQP in SciPy, the *reference solver* below) applied to the equivalent problem $\min t$ subject to $-t \le e_k(\boldsymbol{K}) \le t$ returns the same gain to eleven digits. Figure 1 shows the level lines of φ and the iterates.

Table 1: Gauss–Newton iteration for one data set ($q = 1$, $N = 200$). $\boldsymbol{K}^*$ is the final iterate.

j	$\varphi(\boldsymbol{K}_j)$	$\|\boldsymbol{K}_j - \boldsymbol{K}^*\|_\infty$	$\|\boldsymbol{d}_j\|_\infty$	bound active
0	8.9768444188	$9.0 \cdot 10^{-1}$	$4.3 \cdot 10^{-1}$	no
1	5.6824524193	$4.7 \cdot 10^{-1}$	$3.6 \cdot 10^{-1}$	no
2	5.0997456778	$1.9 \cdot 10^{-1}$	$1.5 \cdot 10^{-1}$	no
3	4.8392804944	$1.2 \cdot 10^{-1}$	$1.2 \cdot 10^{-1}$	no
4	4.7873530413	$7.6 \cdot 10^{-3}$	$7.5 \cdot 10^{-3}$	no
5	4.7772105010	$7.9 \cdot 10^{-5}$	$7.9 \cdot 10^{-5}$	no
6	4.7770840340	$2.0 \cdot 10^{-8}$	$2.0 \cdot 10^{-8}$	no
7	4.7770840117	0	$1.0 \cdot 10^{-14}$	no

Table 2: The three gains for the data set of Table 1. "rel. std" is the exact stationary standard deviation of the innovations relative to the optimum; ρ is the spectral radius of $\boldsymbol{A} - \boldsymbol{K}\boldsymbol{C}$.

gain	K_1	K_2	$\max_k \|e_k\|$	r.m.s. e_k	rel. std	ρ
Riccati (true $\boldsymbol{Q}$, R)	1.2500	0.4932	5.0318	1.9391	1.0000	0.493
least squares	1.2477	0.4374	4.8430	1.9358	1.0026	0.436
minimax	1.4012	0.4751	4.7771	1.9629	1.0166	0.425

On the record the minimax gain lowers the largest error from 5.03 (Riccati gain) and

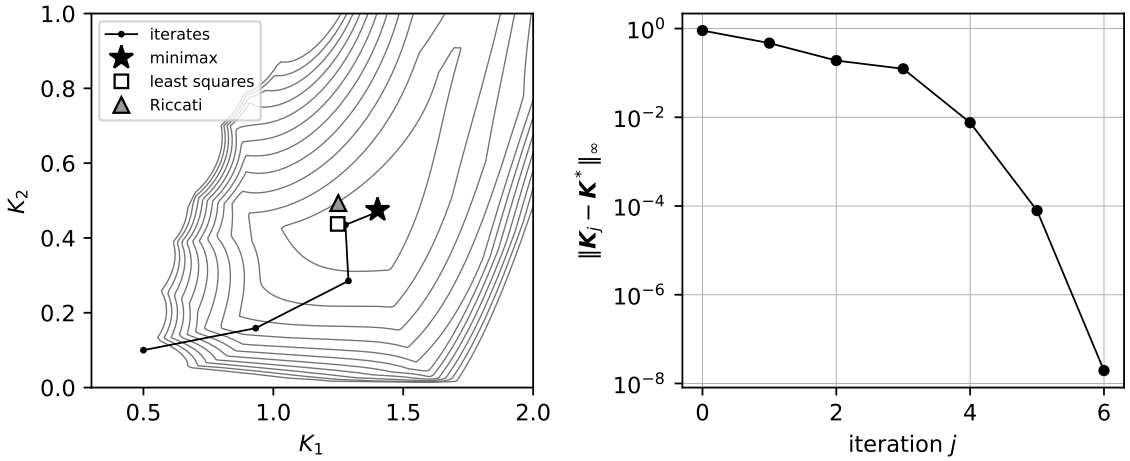

Figure 1: Left: level lines of φ over the gain plane for the data set of Table 1, with the Gauss–Newton iterates and the three gains of Table 2. Right: distance of the iterates to the final iterate.

4.84 (least squares) to 4.78, and its stationary standard deviation of the innovations is 1.7 per cent above the optimum and 1.4 per cent above that of the least-squares gain (Table 2).

6.2 Two hundred data sets per configuration

Table 3 summarises 200 independent records for each of eight configurations. A solution is counted as certified if it has at least three extremal points (relative tolerance 10^{-7}) and $\gamma_E > 10^{-6}$ and the iteration stopped before the limit of 200 linear programmes; the last condition did not change the counts. The threshold 10^{-6} is a convention, not an error bound; the smallest γ_E among the certified solutions was $1.7 \cdot 10^{-4}$. Of the 1600 runs, 1593 met the test on the predicted decrease (1183 of them also the test on the step length), five met only the test on the step length (all five certified, none with the step on the boundary of the trust region), and two stopped at the limit (supplementary programme `replacement_check.py`).

The columns "cert.", "2-pt", "it." and γ_E of Table 3 refer to the output of the Gauss–Newton method. The two prediction columns for the minimax gain are a different matter; the column for the least-squares gain is not affected. For a record whose solution is not certified, the reference solver is started at the Gauss–Newton output, and its gain is used for these two columns whenever its objective value is smaller, by a strict comparison without tolerance. This happened in 335 of the 344 records that are not certified (47, 41, 33, 38, 47, 41, 51 and 37 in the rows of the table). In 333 of them the two objective values differ by less than 10^{-9} relatively and the relative standard deviation changes by less than 10^{-6}; the other two are the records that reached the limit, where it changes by 0.80 and by $1.2 \cdot 10^{-4}$. The medians and upper quartiles in the table are the same to six decimal places with and without the replacement, so these columns also describe the gains returned by the Gauss–Newton method.

Table 3: Simulation study, 200 records per row. "cert.": solutions with at least three extremal points and $\gamma_E > 10^{-6}$, stopped before the limit; "2-pt": solutions with two extremal points; "it.": median iteration count among certified solutions; γ_E: median of γ_E among certified solutions; last three columns: stationary standard deviation of the innovations relative to its minimum, median for the least-squares gain, median and upper quartile for the minimax gain, with the replacement rule described in the text.

q	N	noise	cert.	2-pt	it.	γ_E	LS	minimax	(75 %)
1	200	Gaussian	153	47	7	0.53	1.004	1.065	1.121
1	200	uniform	157	43	7	0.47	1.003	1.049	1.088
1	1000	Gaussian	165	35	7	0.65	1.001	1.042	1.079
1	1000	uniform	160	40	7	0.47	1.001	1.029	1.067
0.1	200	Gaussian	153	46	6	0.63	1.003	1.049	1.106
0.1	200	uniform	159	41	6	0.36	1.003	1.051	1.091
0.1	1000	Gaussian	148	52	6	0.50	1.001	1.029	1.062
0.1	1000	uniform	161	38	6	0.34	1.001	1.050	1.093

Three observations follow.

The computed solutions are often, not always, consistent with the hypothesis of Theorem 7. Between 148 and 165 of 200 solutions (74 to 82.5 per cent) are certified. There the iteration takes six or seven steps in the median; the largest count was 70. On every certified solution the objective value agrees with that of the reference solver to a relative difference below $6 \cdot 10^{-8}$.

Two-point solutions. Almost all other solutions have exactly two extremal points (Remark 10); in two of the 1600 records the computed point had a different number of extremal points. For the configuration $q = 1$, $N = 200$, Gaussian, the 47 two-point cases needed 44 iterations in the median (quartiles 41 and 48.5, maximum 63), against seven for the certified cases. The iteration still arrived at the value found by the reference solver (relative difference

below 10^{-12} in all 47 cases). For these solutions the Gauss–Newton method is usable but slow, and the theory of Section 4 says nothing. Section 6.3 applies the two-stage method of Section 5 to them.

Stationary prediction error. The median excess of the stationary standard deviation over the Riccati optimum ranges from 0.05 to 0.4 per cent for the least-squares gain and from 2.9 to 6.5 per cent for the minimax gain, whose upper quartile reaches 12 per cent. In every row of Table 3, under Gaussian and under uniform disturbances, the median for the minimax gain is the larger: in this study the minimax gain did not predict better in the mean-square sense under either noise distribution. These are statements on medians over the records, not on every record. All computed minimax gains were stabilising.

6.3 Two-stage method

The two-stage method of Section 5 was run on the same 1600 records from the same starting point. The second stage was considered only at iterates whose Gauss–Newton step satisfied $\|\boldsymbol{d}\|_\infty \le 0.1$. Work is counted in linear programmes solved (one more than the iteration count used above) and in Newton steps; a Newton step costs one pass of the recursions (7) and (15) and no linear programme.

Table 4: Two-stage method on the records of Table 3. "2-pt": computed solutions with two extremal points; "s.o.": those that pass the numerical check of (S1)–(S3); "LP (GN)" and "LP": median number of linear programmes for the Gauss–Newton method alone and for the two-stage method on these records; "Newton": median number of accepted Newton steps; μ: median of the reduced curvature; "rej.": median number of discarded Newton steps over all 200 records.

q	N	noise	2-pt	s.o.	LP (GN)	LP	Newton	μ	rej.
1	200	Gaussian	47	47	45	7	4	6.4	9
1	200	uniform	43	43	43	8	4	4.4	7
1	1000	Gaussian	35	35	45	7	4	4.4	3
1	1000	uniform	40	40	44.5	7	4	4.3	2
0.1	200	Gaussian	46	46	45	6	4	7.8	9
0.1	200	uniform	42	42	43	6	4	4.9	10
0.1	1000	Gaussian	52	52	43	5	4	11.6	3.5
0.1	1000	uniform	38	38	42	6	4	5.8	1

Two-point solutions. The method ended at a solution with two extremal points in 343 records, in every case through the Newton stage (Table 4). On these records the Gauss–Newton method alone solved 44 linear programmes in the median (quartiles 40 and 47, maximum 71; the maximum belongs to the record discussed below, where the Gauss–Newton method ended at a different point, and is 68 without it); the two-stage method solved 6 (quartiles 5 and 8, maximum 28) and then took 4 Newton steps in the median (maximum 10). All 343 points pass the check of (S1)–(S3): the smallest multiplier was $1.9 \cdot 10^{-3}$, the smallest reduced curvature $\mu = 0.11$ (median 5.9), the smallest singular value of the matrix with the rows $(\boldsymbol{g}_i^{\mathsf{T}}, -1)$ was 0.63, and the largest non-extremal $|e_k|$ stayed below φ by at least $4 \cdot 10^{-4}$. The objective values agree with those of the reference solver, started at the computed point, to a relative difference below $5 \cdot 10^{-12}$. Table 5 shows one record.

Three-point solutions. In 1255 records the two-stage method ended in the first stage at a solution with three extremal points, with the same iterates as the Gauss–Newton method alone; all of them have $\gamma_E > 10^{-6}$. The second stage was nevertheless tried on many of them

Table 5: A record with a two-point solution ($q = 1$, $N = 200$, Gaussian, repetition 10): distance $\|\boldsymbol{K}_j - \boldsymbol{K}^*\|_\infty$ to the final point for the Gauss–Newton method alone and for the two-stage method (G: Gauss–Newton iterate, N: Newton step). The Gauss–Newton method stops after 47 iterations. Extremal points $k = 31, 53$, both with positive sign, multipliers 0.32 and 0.68, $\mu = 13.5$.

j	Gauss–Newton	stage	two-stage
0	$4.1 \cdot 10^{-1}$	G	$4.1 \cdot 10^{-1}$
1	$2.8 \cdot 10^{-1}$	G	$2.8 \cdot 10^{-1}$
2	$1.2 \cdot 10^{-1}$	G	$1.2 \cdot 10^{-1}$
3	$1.4 \cdot 10^{-1}$	G	$1.4 \cdot 10^{-1}$
4	$1.9 \cdot 10^{-2}$	G	$1.9 \cdot 10^{-2}$
5	$1.9 \cdot 10^{-2}$	G	$1.9 \cdot 10^{-2}$
6	$6.2 \cdot 10^{-2}$	N	$1.7 \cdot 10^{-4}$
7	$6.2 \cdot 10^{-3}$	N	$3.5 \cdot 10^{-7}$
8	$6.2 \cdot 10^{-3}$	N	0
20	$9.2 \cdot 10^{-4}$		
30	$6.4 \cdot 10^{-5}$		
40	$1.4 \cdot 10^{-6}$		
47	$8.4 \cdot 10^{-8}$		

and its result discarded: 8 discarded Newton steps in the median, 18 in the upper quartile, 188 at most. This is the price of the simple switching rule.

Different local minimisers. In one record ($q = 0.1$, $N = 200$, uniform) the two methods ended at different points: the Gauss–Newton method at a three-point solution after 71 linear programmes, the two-stage method at a two-point solution that passes (S1)–(S3) and whose objective value is 2.2 per cent larger. Both points pass their numerical check for a local minimiser; neither method controls which one is found. The restriction $\|\boldsymbol{d}\|_\infty \le 0.1$ matters here: in a run without it, the two-stage method ended in seven records, this one included, at a different point from the Gauss–Newton method (gains differing by more than 10^{-4}), in six of them with a larger objective value (by up to 21 per cent of the Gauss–Newton value; supplementary programme `near_check.py`).

Not covered. In the two remaining records both methods stopped at the limit of 200 linear programmes at a point with a single extremal point; in one of them the reference solver, started there, found a value 5.8 per cent lower. The second stage does not treat this case, and these two records account for the largest numbers of discarded Newton steps (927 and 902).

6.4 Composite criterion

For the data set of Table 1 the composite problem (5) was solved for increasing λ, each run started at the previous solution (Table 6). The smoothness term lowers the gain, reduces the largest second difference of the predictions and increases the largest error, as expected. For $\lambda = 1, 2, 4$ the solution has four active constraints in total ($3 + 1$ or $2 + 2$ in the two blocks), and the last steps decrease quadratically; for $\lambda = 0.25$ and 0.5 it has three ($2 + 1$), and the iteration needs 40 and 47 steps with a linear rate. This is the composite analogue of the two-point case. A certificate corresponding to Proposition 9 for the norm (6) was not implemented. At each computed point, 4000 random perturbations of the gain of size 10^{-7} to 10^{-1} did not decrease the objective.

Table 6: Composite criterion (5) for the data set of Table 1.

λ	K_1	K_2	$\|\boldsymbol{e}\|_\infty$	$\|\boldsymbol{D}\hat{\boldsymbol{y}}\|_\infty$	rel. std	iterations
0	1.4012	0.4751	4.7771	8.1903	1.017	0
0.25	1.1292	0.4270	4.9249	5.9604	1.009	40
0.5	0.9209	0.2382	5.3951	4.7466	1.118	47
1	0.7382	0.1553	6.0599	3.7085	1.335	7
2	0.5854	0.1318	7.3145	2.8205	1.562	6
4	0.4546	0.1274	8.7713	2.2715	1.821	4

6.5 Initial state as a parameter

With $\boldsymbol{\theta} = (\boldsymbol{K}, \hat{\boldsymbol{x}}_0)$, $p = 4$, and 100 records ($q = 1$, $N = 200$, Gaussian), 71 computed solutions had five extremal points; of the other 29 records, 24 stopped at points with two to four extremal points and five reached the limit of 200 linear programmes. For a positive optimal value, at least five extremal points are necessary for (11), but they are not sufficient. The computed constant γ_E was exactly zero for 29 of the 71 solutions; its median was $5{\cdot}10^{-10}$, its maximum $4.7 \cdot 10^{-3}$, and it exceeded the threshold 10^{-6} of Section 6.2 for 17 solutions. The feasibility tolerance of the linear-programme solver (10^{-7} by default) is not an error bound for γ_E, and we have none; values of the size of the median are therefore no evidence of a positive constant. A computed zero may be caused by ill-conditioning, but it does not show positivity, so for these records it is not established that the theory applies. This is the effect described in Remark 11: in 68 of the 71 solutions one or two of the five extremal points lie at $k \le 5$ (three in the other three solutions), and in the 29 cases with a computed $\gamma_E = 0$, among the linear programmes of Proposition 9 only those for faces $\{d_l = \pm 1\}$ whose index l belongs to a component of $\hat{\boldsymbol{x}}_0$ returned a value of at most 10^{-12} (two of them in 26 of the 29 cases); none for a face belonging to a component of $\boldsymbol{K}$ did, and the smallest computed value among the latter was 0.08. These are computed values without an error bound, and on a face with $d_l = \pm 1$ the other components of $\boldsymbol{d}$, those of the gain included, remain free, so such a face does not single out a direction that changes the initial state alone. As a measure of the sensitivity we use, for each solution, the maximum of $|\partial e_k / \partial \hat{x}_{0,l}|$ over the extremal points with $k > 5$ and over the two components l of $\hat{\boldsymbol{x}}_0$. This componentwise maximum had a median of $4 \cdot 10^{-8}$; it was below 10^{-3} in 55 of the 71 solutions and reached 0.44, so "typically" in Remark 11 is meant in this sense. The counts of this subsection are produced by the supplementary programme `initial_state_check.py`. No statement on the size of the neighbourhood of Theorem 7 follows from these numbers: its radius is $\gamma/(7L)$, and L was not determined.

7 Limits and open points

- The convergence statement is local and conditional. Strong uniqueness is checked after the computation, numerically and at the computed point, and the check failed in 17.5 to 26 per cent of the records of Section 6.2, depending on the configuration; with the initial state as a parameter (Section 6.5, $p = 4$) it was passed by only 17 of the 100 records: γ_E exceeded 10^{-6} for 17 of the 71 solutions with five extremal points, and the other 29 records had fewer extremal points or reached the limit. For the solutions with two extremal points the two-stage method of Section 5 gives fast local convergence and a second-order check, but its switching rule is heuristic, it has no global convergence theory, it can end at a different local minimiser than the Gauss–Newton method, and

it was tested on one model with $p = 2$ only. Solutions with a single extremal point and the composite criterion are not covered by it.

- The criterion is a worst-case criterion on the given record. It gives no statistical optimality; in the simulations the median stationary standard deviation of the prediction errors was larger for the minimax gain than for the least-squares gain under both noise distributions. Because the criterion depends only on the largest errors, a single outlying measurement can determine the solution.
- Consistency of the minimax gain as $N \to \infty$ was not studied. The largest error over a record grows with N for unbounded noise, so the question needs its own analysis.
- Only scalar measurements, a time-invariant model with known $\boldsymbol{A}$ and $\boldsymbol{C}$, and a single model family were treated. For vector measurements the maximum norm over all components gives the same structure, with a gain matrix as parameter; this was not tested.
- Stability of $\boldsymbol{A} - \boldsymbol{K}\boldsymbol{C}$ is not enforced (Remark 3).
- The convergence theory is that of [5, 6, 11]; two-stage methods were described in [7, 8], and Newton's method on the active-set equations is in [10]; the switching rule of Section 5 is ours and heuristic. What is specific to this paper is the application, the computed constant γ_E and the experiments; the derivative recursions are stated explicitly for this predictor, without a claim of novelty.

References

[1] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, 1979.

[2] M. A. Belabbas and A. Olshevsky, Interpretable gradient descent for Kalman gain, arXiv:2507.14354, 2025.

[3] L. Bier and S. Talebi, Learning Kalman policy for singular unknown covariances via Riemannian regularization, arXiv:2604.05071, 2026.

[4] B. Carew and P. R. Bélanger, Identification of optimum filter steady-state gain for systems with unknown noise covariances, *IEEE Trans. Automat. Control* 18 (1973), 582–587.

[5] L. Cromme, Eine Klasse von Verfahren zur Ermittlung bester nichtlinearer Tschebyscheff-Approximationen, *Numer. Math.* 25 (1976), 447–459.

[6] L. Cromme, Strong uniqueness. A far-reaching criterion for the convergence analysis of iterative procedures, *Numer. Math.* 29 (1978), 179–193.

[7] J. Hald and K. Madsen, A 2-stage algorithm for minimax optimization, in: *International Symposium on Systems Optimization and Analysis*, Lecture Notes in Control and Information Sciences 14, Springer, Berlin–Heidelberg, 225–239, doi:10.1007/BFb0002657; the year is given as 1978 in [8] and as 1979 in the Crossref record of the chapter under doi:10.1007/978-3-662-39921-7_21.

[8] J. Hald and K. Madsen, Combined LP and quasi-Newton methods for minimax optimization, *Math. Programming* 20 (1981), 49–62.

[9] R. Hettich, A Newton-method for nonlinear Chebyshev approximation, in: R. Schaback and K. Scherer (eds.), *Approximation Theory*, Lecture Notes in Mathematics 556, Springer, Berlin, 1976, 222–236.

[10] R. Hettich and P. Zencke, *Numerische Methoden der Approximation und semi-infiniten Optimierung*, Teubner, Stuttgart, 1982.

[11] K. Jittorntrum and M. R. Osborne, Strong uniqueness and second order convergence in nonlinear discrete approximation, *Numer. Math.* 34 (1980), 439–455.

[12] R. E. Kalman, A new approach to linear filtering and prediction problems, *Trans. ASME J. Basic Eng.* 82 (1960), 35–45.

[13] L. Ljung, *System Identification: Theory for the User*, 2nd ed., Prentice Hall, Upper Saddle River, 1999.

[14] R. K. Mehra, On the identification of variances and adaptive Kalman filtering, *IEEE Trans. Automat. Control* 15 (1970), 175–184.

[15] M. R. Osborne and G. A. Watson, An algorithm for minimax approximation in the nonlinear case, *Comput. J.* 12 (1969), 63–68.

[16] L. Simpson and M. Diehl, Identification of a Kalman filter: consistency of local solutions, arXiv:2601.04198, 2026.

[17] S. Talebi, A. Taghvaei and M. Mesbahi, Duality-based stochastic policy optimization for estimation with unknown noise covariances, arXiv:2210.14878, 2022.

[18] S. Talebi, A. Taghvaei and M. Mesbahi, Data-driven optimal filtering for linear systems with unknown noise covariances, arXiv:2305.17836, 2023.