Active Set Algorithms for Estimating Shape-Constrained Density Ratios[⋆,⋆⋆]

Lutz Dümbgen[a,*], Alexandre Mösching[b] and Christof Strähl[a]

[a]University of Bern, Institute of Mathematical Statistics and Actuarial Science, Alpeneggstr. 22, 3012 Bern, Switzerland
[b]University of Göttingen, Institute for Mathematical Stochastics, Goldschmidtstr. 7, 37077 Göttingen, Germany

ARTICLE INFO

Keywords:
active set method
log-concave density
log-convex density
tail inflation

Abstract

In many instances, imposing a constraint on the shape of a density is a reasonable and flexible assumption. It offers an alternative to parametric models, which can be too rigid, and to other nonparametric methods, which require the choice of tuning parameters. The nonparametric estimation of log-concave or log-convex density ratios is treated by means of active set algorithms in a unified framework. In the setting of log-concave densities, the new algorithm is similar to, but substantially faster than, previously considered active set methods. Log-convexity, on the other hand, is a less common shape-constraint, described by some authors as "tail inflation". The active set method proposed here is novel in this context. As a by-product, new goodness-of-fit tests of single hypotheses are formulated and are shown to be more powerful than higher criticism tests in a simulation study.

1. Introduction

Suppose we observe independent random variables $X_1, X_2, \ldots, X_n$ with unknown distributions $P_1, P_2, \ldots, P_n$ on the real line. This paper discusses the estimation of the marginal (average) distribution $P := n^{-1} \sum_{i=1}^n P_i$ under certain shape-constraints. Of course, this framework includes the case of i.i.d. observations from a single distribution P.

Within the broad field of nonparametric statistics, inference about P under shape-constraints is a well-established alternative to the assumption of quantitative smoothness properties, e.g. certain bounds on the maximum modulus of some higher order derivative of the density of P (w.r.t. Lebesgue measure). While estimation under smoothness assumptions typically involves tuning parameters, such as bandwidths of kernel density estimators, maximum likelihood estimation under shape-constraints is often possible without any further specifications. For a thorough discussion on the benefits of shape-constraints, we refer to Groeneboom and Jongbloed (2014).

One particular example of a shape-constraint is log-concavity of the density of P. A broad overview of statistical methods with such densities, including the multivariate case, is given by Samworth (2018). A second example of a shape-constraint is convexity of the density of P on the positive half-line (see Groeneboom et al. (2001)). In the present paper, we reconsider the estimation of log-concave densities, and also examine a less familiar setting which is related to the estimation of convex densities:

Setting 1: Log-concave densities. We assume that P has a log-concave density f with respect to Lebesgue measure. That is, $\log f : \mathbb{R} \to [-\infty, \infty)$ is concave.

Setting 2: Tail inflation. For a given continuous reference distribution P_o on $\mathbb{R}$, we assume that P has a log-convex density f with respect to P_o. That is, $\log f : \mathbb{R} \to \mathbb{R}$ is convex.

The notion of tail inflation has been introduced by McCullagh and Polson (2012, 2017) to investigate statistical sparsity. They consider the case where the observations $X_i > 0$, the reference distribution P_o is the chi-squared distribution with one degree of freedom, and $\log f$ is convex and isotonic (non-decreasing). However, Setting 2 is also

⋆ Supplementary material is available with the online version of the paper.
⋆⋆ Computer code in R for the procedures described in this article can be found online at https://github.com/duembgen-lutz/LogConDens and https://github.com/duembgen-lutz/TailInflation.

*Corresponding author
duembgen@stat.unibe.ch (L. Dümbgen); alexandre.moesching@uni-goettingen.de (A. Mösching); christof.straehl@gmx.net (C. Strähl)
ORCID(s): 0000-0003-0172-9285 (L. Dümbgen); 0000-0002-8270-3724 (A. Mösching)

related other secenarios. For example, in multiple hypothesis testing the $X_1, X_2, \dots, X_n$ could represent test statistics for given null hypotheses $H_1, H_2, \dots, H_n$, where X_i has distribution P_o whenever H_i is true. In image analysis, the random variables X_i could be measured intensities at different pixels of a digital image, with P_o describing pure background noise or measurement errors.

The primary goals in these settings are to estimate P, or to test the null hypothesis that all P_i are equal to P_o. The assumption of log-convexity of $f = dP/dP_o$ may seem a bit arbitrary at first sight, but note, for instance, that the testing problems considered by Donoho and Jin (2004) may be viewed as a special case of Setting 2, with P_o the standard Gaussian distribution $\mathcal{N}(0,1)$. Indeed, the latter authors considered i.i.d. observations with distribution P a mixture $(1-\varepsilon)\mathcal{N}(0,1) + \varepsilon\mathcal{N}(\mu,1)$ with unknown parameters $\varepsilon \in [0,1]$ and $\mu \geq 0$. As shown later, if each P_i is a mixture of Gaussian distributions with standard deviation at least 1, then each P_i, as well as the marginal distribution P, has a log-convex density with respect to P_o. Consequently, if we estimate the log-density $\theta := \log f$ of P, this gives rise to a new likelihood ratio test statistic for the null hypothesis that all P_i are equal to P_o.

Outline of the paper. Our main goals are to establish existence and uniqueness of the nonparametric maximum likelihood estimator $\hat{\theta}$ of $\theta := \log f$ in Setting 2 and to devise explicit algorithms for its computation. Since Settings 1 and 2 are closely related, it is worthwhile to treat both of them simultaneously, highlighting similarities and differences. In Section 2, the specific estimation problems are described in more detail, and it is shown that under certain assumptions, the maximizer $\hat{\theta}$ exists and is unique.

In Section 3, we describe a general active set method for the computation of $\hat{\theta}$. The starting point is the active set method described by Dümbgen et al. (2007/2011) and Dümbgen and Rufibach (2011), which is similar to the support reduction algorithm of Groeneboom et al. (2008). The new version is more efficient in that all single Newton steps take shape-constraints on θ into account. We also adopt the proposal of Liu and Wang (2018) to occasionally deactivate more than one constraint in one step, but in contrast to the latter authors, we do not resort to quadratic programming routines within the algorithm. In Setting 2, we explore the full infinite-dimensional parameter space rather than using ad hoc finite-dimensional approximations.

Numerical examples illustrating the estimation method are given in Section 4. For Setting 1, we demonstrate the benefits of the new method in a small simulation study. We also show that our estimator for Setting 2 leads to a promising goodness-of-fit test, and simulations indicate that the power of this test can exceed the power of higher criticism methods proposed by Donoho and Jin (2004) and Gontscharuk et al. (2016).

Section 5 provides proofs for the existence, uniqueness and special properties of $\hat{\theta}$, while Appendix A provides technical details for specific applications, as well as a proof of convergence that generalises and simplifies a previous proof of Sommer-Simpson (2019). The algorithms have been implemented in the statistical langage R (R Core Team, 2016) and are publicly available.

2. General considerations, existence and uniqueness

In what follows, we consider an arbitrary discrete distribution

$$\hat{P} := \sum_{i=1}^{n} w_i \delta_{x_i}$$

with $n \geq 2$ probability weights $w_1, \dots, w_n > 0$ and real support points $x_1 < \cdots < x_n$. In Settings 1 and 2, these points $x_1, \dots, x_n$ are the order statistics of the observations $X_1, \dots, X_n$ while $w_i = n^{-1}$. The general form of $\hat{P}$ also covers the situation of $N \geq n$ raw observations from P that are recorded with rounding errors: In this case, $x_1, \dots, x_n$ are the different recorded values, and w_i is the relative frequency of x_i in the sample.

2.1. Parameter spaces and target functional

In general, we assume that $\hat{P}$ estimates an unknown distribution P that has a density f with respect to a given continuous measure M on $\mathbb{R}$. Precisely,

$$f(x) = f_\theta(x) := e^{\theta(x)}$$

with an unknown function parameter $\theta : \mathbb{R} \to [-\infty, \infty)$ in a given family Θ reflecting the particular shape-constraints to be specified later. Then θ is estimated by a function $\hat{\theta} \in \Theta$ maximizing the normalized log-likelihood

$$\ell(\theta) \;:=\; \int \theta \, d\hat{P} \;=\; \sum_{i=1}^{n} w_i \theta(x_i)$$

under the constraint that $\int e^\theta \, dM = 1$.

In the specific settings we have in mind, all functions $\theta \in \Theta$ satisfy $0 < \int e^\theta \, dM \le \infty$ and $\theta + c \in \Theta$ for arbitrary real constants c. Thus we may apply the Lagrange trick of Silverman (1982) and rewrite $\hat{\theta}$ as

$$\hat{\theta} \;=\; \operatorname*{arg\,max}_{\theta \in \Theta} \; L(\theta)$$

with

$$L(\theta) \;:=\; \int \theta \, d\hat{P} - \int e^\theta \, dM + 1 \;\in\; [-\infty, \infty).$$

Indeed, for $\theta \in \Theta$ with $L(\theta) > -\infty$ and $c \in \mathbb{R}$, the derivative $\partial L(\theta + c)/\partial c$ equals $1 - e^c \int e^\theta \, dM$. Hence, a function $\hat{\theta} \in \Theta$ with $L(\hat{\theta}) > -\infty$ maximizes $L(\theta)$ over all $\theta \in \Theta$ if and only if it maximises $\ell(\theta)$ under the constraint that $\int e^\theta \, dM = 1$. Note also that $L(\theta) = \ell(\theta)$ if and only if $\int e^\theta \, dM = 1$.

Setting 1. M is Lebesgue measure on $\mathbb{R}$, and the parameter space Θ consists of all concave, upper semicontinuous functions $\theta : \mathbb{R} \to [-\infty, \infty)$ such that $\int e^\theta \, dM > 0$.

For Setting 2 in the introduction, we distinguish between two versions, where the second covers the framework of McCullagh and Polson (2012).

Setting 2A. M stands for the reference distribution P_o. We assume that P_o is continuous with $P_o(B) > 0$ for any non-degenerate interval $B \subset \mathbb{R}$, and

$$\left\{ \lambda \in \mathbb{R} : \int e^{\lambda x} \, P_o(dx) < \infty \right\} \;=\; \big(\lambda_\ell(P_o), \lambda_r(P_o)\big)$$

for certain numbers $-\infty \le \lambda_\ell(P_o) < 0 < \lambda_r(P_o) \le \infty$. The extended parameter space Θ consists of all convex functions $\theta : \mathbb{R} \to \mathbb{R}$.

Example 2.1 (Gaussian mixtures). Let $P_o = \mathcal{N}(0,1)$. Suppose that P is a mixture of Gaussian distributions with standard deviation at least 1, i.e. $P = \int \mathcal{N}(\mu, \sigma^2) \, Q(d\mu, d\sigma)$ for some probability distribution Q on $\mathbb{R} \times [1, \infty)$. Then $\theta := \log dP/dP_o$ is given by

$$\theta(x) \;=\; \log \int e^{\theta(x,\mu,\sigma)} \, Q(d\mu, d\sigma)$$

with

$$\theta(x, \mu, \sigma) \;:=\; \log \frac{d\mathcal{N}(\mu, \sigma^2)}{d\mathcal{N}(0,1)}(x) \;=\; -\log \sigma + \frac{(\sigma^2 - 1)x^2 + 2\mu x - \mu^2}{2\sigma^2}.$$

Obviously, $\theta(\cdot, \mu, \sigma)$ is a convex function for arbitrary $\mu \in \mathbb{R}$ and $\sigma \ge 1$, so the log-mixture density θ is convex too. This can be deduced from Hölder's inequality or Artin's theorem (see Section D.4 of Marshall and Olkin (1979)).

Example 2.2 (Student distributions). Let $P_o = \mathcal{N}(0, \sigma^2)$ and $P = t_k$ with $\sigma, k > 0$. Tedious but elementary calculations show that $\theta = \log(dP/dP_o)$ is convex if and only if $\sigma^2 \le k/(k+1)$.

Example 2.3 (Logistic distributions). Let $P_o = \mathcal{N}(0,1)$, and let P be the logistic distribution with scale parameter $\sigma > 0$, i.e. with Lebesgue density $p(x) = \sigma^{-1}(e^{x/\sigma} + e^{-x/\sigma} + 2)^{-1}$. Here one can show that $\theta = \log(dP/dP_o)$ is convex if and only if $\sigma \ge 2^{-1/2}$.

Setting 2B. M stands for the reference distribution P_o. We assume that P_o is continuous with $P_o((-\infty, 0]) = 0$, $P_o(B) > 0$ for any non-degenerate interval $B \subset (0, \infty)$, and

$$\left\{ \lambda \in \mathbb{R} : \int e^{\lambda x} P_o(dx) < \infty \right\} = \big(-\infty, \lambda_r(P_o)\big)$$

for some number $\lambda_r(P_o) \in (0, \infty]$. Now the extended parameter space Θ consists of all convex functions $\theta : \mathbb{R} \to \mathbb{R}$ such that $\theta \equiv \theta(0)$ on $(-\infty, 0]$. In particular, all $\theta \in \Theta$ are isotonic.

Example 2.4 (Scale mixtures of Gamma distributions). Let $P_o = \text{Gamma}(\alpha, \beta)$, the gamma distribution with given shape parameter $\alpha > 0$ and rate parameter $\beta > 0$. Suppose that P is a scale mixture of gamma distributions with the same shape parameter, i.e. $P = \int \text{Gamma}(\alpha, \beta/s)\, Q(ds)$ for some probability measure Q on $(0, \infty)$. Then $\theta := \log dP/dP_o$ is given by

$$\theta(x) = \log \int e^{\theta(x,s)}\, Q(ds)$$

with $\theta(x, s) := \beta(1 - 1/s)x - \alpha \log s$. The latter expression is linear in x, whence θ is convex. If $Q([1, \infty)) = 1$, then θ is also isotonic.

A special instance of this setting are raw observations $\tilde{X}_i = S_i G_i$, $1 \le i \le n$, with independent random variables $S_1, \ldots, S_n \ge 1$ and $G_1, \ldots, G_n \sim \mathcal{N}(0, 1)$. With $P_o := \chi^2_1 = \text{Gamma}(1/2, 1/2)$, the marginal distribution P of the observations $X_i := \tilde{X}_i^2$ has the log-density $\theta = \log \int e^{\theta(\cdot,s)}\, Q(ds)$ with respect to P_o, where $Q := n^{-1} \sum_{i=1}^n \mathcal{L}(S_i)$.

2.2. Existence and uniqueness of the estimator

In Settings 1 and 2A-B, the target functional L is strictly concave on the convex set $\{\theta \in \Theta : L(\theta) > -\infty\}$. This follows easily from strict convexity of the exponential function. Precisely, there exists a unique maximizer $\hat{\theta} \in \Theta$ of L which is piecewise linear and satisfies further properties summarized in the following three lemmas. The first lemma has been proved by Walther (2002), see also Dümbgen et al. (2007/2011) or Cule et al. (2010):

Lemma 2.5. *In Setting 1, there exists a unique maximizer $\hat{\theta}$ of L over Θ. Precisely, there exist $m \ge 2$ points $\tau_1 < \cdots < \tau_m$ in $\{x_1, x_2, \ldots, x_n\}$ with $\tau_1 = x_1$, $\tau_m = x_n$, with the following properties:*

$$\hat{\theta} \begin{cases} \text{is linear on } [\tau_j, \tau_{j+1}],\ 1 \le j < m, \\ \text{equals } -\infty \text{ on } \mathbb{R} \setminus [x_1, x_n], \end{cases}$$

and the slope $\hat{\theta}'(\tau_j +) = \big(\hat{\theta}(\tau_{j+1}) - \hat{\theta}(\tau_j)\big)/(\tau_{j+1} - \tau_j)$ is strictly decreasing in $j \in \{1, \ldots, m-1\}$.

Lemma 2.6. *In Setting 2A, there exists a unique maximizer $\hat{\theta}$ of L over Θ. Precisely, either $\hat{\theta}$ is linear, or there exist $m \in \{1, \ldots, n-1\}$ points $\tau_1 < \cdots < \tau_m$ in $[x_1, x_n] \setminus \{x_1, \ldots, x_n\}$ with the following properties:*

$$\hat{\theta} \text{ is linear on } \begin{cases} \mathcal{X}_0 := (-\infty, \tau_1], \\ \mathcal{X}_j := [\tau_j, \tau_{j+1}],\ 1 \le j < m, \\ \mathcal{X}_m := [\tau_m, \infty), \end{cases}$$

and the sequence of slopes of $\hat{\theta}$ on these $m + 1$ intervals is strictly increasing. Furthermore, each interval (x_i, x_{i+1}), $1 \le i < n$, contains at most one point τ_j.

Lemma 2.7. *In Setting 2B, there exists a unique maximizer $\hat{\theta}$ of L over Θ. Precisely, either $\hat{\theta} \equiv 0$, or there exist $m \in \{1, \ldots, n-1\}$ points $\tau_1 < \cdots < \tau_m$ in $\{0\} \cup [x_1, x_n] \setminus \{x_1, \ldots, x_n\}$ with the following properties:*

$$\hat{\theta} \text{ is } \begin{cases} \text{constant on } (-\infty, \tau_1], \\ \text{linear on } \mathcal{X}_j := [\tau_j, \tau_{j+1}],\ 1 \le j < m-1, \\ \text{linear on } \mathcal{X}_m := [\tau_m, \infty), \end{cases}$$

and the slope $\hat{\theta}'(\tau_j +)$ is strictly positive and strictly increasing in $j \in \{1, \ldots, m\}$. Furthermore, each interval (x_i, x_{i+1}), $1 \le i < n$, contains at most one point τ_j.

Note that the number m in Lemma 2.7 could be 1, meaning that $\hat{\theta}$ is constant on $[0, \tau_1]$ and linear on $[\tau_1, \infty)$ with slope $\hat{\theta}'(\tau_1 +) \in (0, \lambda_r(P_o))$.

3. A general active set strategy

3.1. The space of relevant functions

In view of Lemmas 2.5, 2.6 and 2.7, it suffices to consider continuous, piecewise linear functions θ on

$$\mathcal{X} := \begin{cases} [x_1, x_n] & \text{in Setting 1} \\ \mathbb{R} & \text{in Setting 2A} \\ [0, \infty) & \text{in Setting 2B} \end{cases}$$

with changes of slope only in

$$\mathcal{D} := \begin{cases} \{x_i : 1 < i < n\} & \text{in Setting 1,} \\ (x_1, x_n) & \text{in Setting 2A,} \\ \{0\} \cup (x_1, x_n) & \text{in Setting 2B.} \end{cases}$$

In Setting 2B, a change of slope at 0 means that $\theta'(0+) \neq 0$. The linear space of all such functions θ is denoted by $\mathbb{V}$. One particular basis is given by the functions

$$x \mapsto 1,$$
$$x \mapsto x \quad \text{(in Settings 1 and 2A)}$$

and

$$x \mapsto V_\tau(x) := \xi(x - \tau)^+, \quad \tau \in \mathcal{D},$$

where

$$\xi := \begin{cases} -1 & \text{in Setting 1,} \\ +1 & \text{in Settings 2A-B.} \end{cases}$$

This means that $\dim(\mathbb{V})$ equals n in Setting 1 and ∞ in Settings 2A-B. Any $\theta \in \mathbb{V}$ may be written as

$$\theta(x) = \left\{ \begin{array}{l} \alpha_0 \\ \quad + \alpha_1 x \quad \text{(in Settings 1 and 2A)} \\ \quad\quad + \sum_{\tau \in \mathcal{D}} \beta_\tau V_\tau(x) \end{array} \right\} \tag{1}$$

with real coefficients $\alpha_0, \alpha_1, \beta_\tau$ such that $\beta_\tau \neq 0$ for at most finitely many $\tau \in \mathcal{D}$. Note that $\xi\beta_\tau$ is equal to the change of slope, $\theta'(\tau+) - \theta'(\tau-)$, whence

$$\theta \in \Theta \quad \text{if and only if} \quad \beta_\tau \geq 0 \ \text{ for all } \tau \in \mathcal{D}.$$

3.2. Properties of L

On the set $\mathbb{V}$, the functional L is continuous with respect to the norm

$$\|\theta\| := \begin{cases} \max_{x \in [x_1, x_n]} |\theta(x)| & \text{in Setting 1,} \\ \max_{x \in [x_1, x_n]} |\theta(x)| + |\theta'(x_1)| + |\theta'(x_n)| & \text{in Settings 2A-B.} \end{cases} \tag{2}$$

For Setting 1, $\|\cdot\|$ quantifies uniform convergence on $\mathcal{X}$. For Settings 2A-B, convergence with respect to $\|\cdot\|$ is equivalent to uniform convergence on arbitrary bounded subsets of $\mathcal{X}$. Moreover, in Setting 1, L is real-valued, whereas in Settings 2A-B it follows from our assumptions on P_o that

$$\{\theta \in \mathbb{V} : L(\theta) > -\infty\} = \begin{cases} \{\theta \in \mathbb{V} : \theta'(x_1) > \lambda_\ell(P_o) \text{ and } \theta'(x_n) < \lambda_r(P_o)\} & \text{in Setting 2A,} \\ \{\theta \in \mathbb{V} : \theta'(x_n) < \lambda_r(P_o)\} & \text{in Setting 2B.} \end{cases}$$

Finally, on the set $\{\theta \in \mathbb{V} : L(\theta) > -\infty\}$, the functional L is strictly concave. Precisely, for $\theta, v \in \mathbb{V}$ with $L(\theta) > -\infty$,

$$DL(\theta, v) := \frac{d}{dt}\Big|_{t=0} L(\theta + tv) = \int v \, d\hat{P} - \int_{\mathcal{X}} v e^{\theta} \, dM,$$
$$H(\theta, v) := -\frac{d^2}{dt^2}\Big|_{t=0} L(\theta + tv) = \int_{\mathcal{X}} v^2 e^{\theta} \, dM.$$

These derivatives $DL(\theta, v)$ and $H(\theta, v)$ are well-defined, because $\int_{\mathcal{X}} e^{\theta(x)+\varepsilon|x|} \, M(dx) < \infty$ for sufficiently small $\varepsilon > 0$. Note that $H(\theta, v) > 0$ unless $\|v\| = 0$.

3.3. Characterizing $\hat{\theta}$

The properties of L imply that a function $\theta \in \mathbb{V} \cap \Theta$ with $L(\theta) > -\infty$ equals $\hat{\theta}$ if and only if

$$DL(\theta, v) \le 0 \quad \text{for any } v \in \mathbb{V} \text{ such that } \theta + tv \in \Theta \text{ for some } t > 0. \tag{3}$$

Representing θ as in (1) and v analogously, one can easily verify that (3) is equivalent to the following four conditions:

$$\int_{\mathcal{X}} e^{\theta} \, dM = 1, \tag{4}$$
$$\int_{\mathcal{X}} x e^{\theta(x)} \, M(dx) = \hat{\mu} \quad \text{(in Settings 1 and 2A)}, \tag{5}$$
$$\int_{\mathcal{X}} V_\tau e^{\theta} \, dM = \int V_\tau \, d\hat{P} \quad \text{whenever } \beta_\tau > 0, \tag{6}$$
$$\int_{\mathcal{X}} V_\tau e^{\theta} \, dM \ge \int V_\tau \, d\hat{P} \quad \text{whenever } \beta_\tau = 0, \tag{7}$$

where $\hat{\mu}$ denotes the empirical mean $\hat{\mu} := \int x \, \hat{P}(dx) = \sum_{i=1}^n w_i x_i$.

Local optimality. Requirements (4–6) can be interpreted as follows: For $\theta \in \mathbb{V}$ let $D(\theta) \subset \mathcal{D}$ be the finite set of its "deactivated (equality) constraints". That means,

$$D(\theta) := \big\{\tau \in \mathcal{D} : \theta'(\tau-) \ne \theta'(\tau+)\big\}.$$

For an arbitrary finite set $D \subset \mathcal{D}$ we define

$$\mathbb{V}_D := \big\{\theta \in \mathbb{V} : D(\theta) \subset D\big\}.$$

This is a linear subspace of $\mathbb{V}$ with dimension $2 + \#D$ (in Settings 1 and 2A) or $1 + \#D$ (in Setting 2B). Requirements (4–6) are then equivalent to saying that $\int_{\mathcal{X}} v e^{\theta} \, dM = \int v \, d\hat{P}$ for all $v \in \mathbb{V}_{D(\theta)}$. That is,

$$DL(\theta, v) = 0 \quad \text{for all } v \in \mathbb{V}_{D(\theta)}. \tag{8}$$

In other words, θ is "locally optimal" in the sense that

$$\theta = \operatorname*{arg\,max}_{\eta \in \mathbb{V}_{D(\theta)}} L(\eta).$$

Checking global optimality. Requirement (7) is equivalent to

$$h_\theta(\tau) := DL(\theta, V_\tau) \le 0 \quad \text{for all } \tau \in \mathcal{D} \setminus D(\theta). \tag{9}$$

Thus, a function $\theta \in \mathbb{V} \cap \Theta$ with $L(\theta) > -\infty$ is equal to $\hat{\theta}$ if and only if it is locally optimal in the sense of (8), and it satisfies (9). As explained in Section A.1, for computational efficiency and numerical accuracy, it is advisable to replace the simple kink functions V_τ with localised versions $V_{\tau,\theta} = V_\tau - \eta_{\tau,\theta}$, where $\eta_{\tau,\theta} \in \mathbb{V}_{D(\theta)}$, though the general description of our methods is easier in terms of the V_τ.

3.4. Basic procedures

Our active set method involves a candidate $\theta \in \Theta \cap \mathbb{V}$ for the function $\hat{\theta}$ such that f_θ defines a probability density w.r.t. M and a finite set $D \subset \mathcal{D}$ such that $D(\theta) \subset D$.

Basic step 1: Obtaining a proposal θ_{new} via Newton's method

Recall that the functional L is continuous and concave on the finite-dimensional space $\mathbb{V}_D$. Moreover, on $\{\eta \in \mathbb{V}_D : L(\eta) > -\infty\}$ it is twice continuously differentiable with negative definite Hessian operator. Thus we may perform a standard Newton step to obtain a function $\theta_{\text{new}} \in \mathbb{V}_D$ such that

$$\delta := DL(\theta, \theta_{\text{new}} - \theta) \geq 0$$

with equality if and only if

$$\theta = \theta_{\text{new}} = \operatorname*{arg\,max}_{\eta \in \mathbb{V}_D} L(\eta).$$

Even in the case where $\delta > 0$, it may happen that $L(\theta_{\text{new}}) \leq L(\theta)$. To guarantee a real improvement, we apply a standard Armijo–Goldstein step size correction and replace θ_{new} with $\theta + 2^{-n}(\theta_{\text{new}} - \theta)$, where n is the smallest nonnegative integer such that

$$L(\theta + 2^{-n}(\theta_{\text{new}} - \theta)) - L(\theta) \geq 2^{-n} DL(\theta, \theta_{\text{new}} - \theta)/3.$$

(A theoretical justification of this step size correction can be found, for instance, in Dümbgen (2017).) In algorithmic language, as long as $L(\theta_{\text{new}}) < L(\theta) + \delta/3$, we replace $(\theta_{\text{new}}, \delta)$ with the pair $\big((\theta + \theta_{\text{new}})/2, \delta/2\big)$. After finitely many steps, the new pair $(\theta_{\text{new}}, \delta)$ will satisfy $L(\theta_{\text{new}}) \geq L(\theta) + \delta/3$ and $\delta = DL(\theta, \theta_{\text{new}} - \theta) > 0$. In the pseudocode provided later, this Newton–Armijo–Goldstein step is abbreviated as "$(\theta_{\text{new}}, \delta) \leftarrow \text{Newton}(\theta, D)$".

Basic step 2: Modification of θ or reduction of D

Having computed a new proposal θ_{new} as in basic step 1, where $\delta = DL(\theta, \theta_{\text{new}} - \theta) > 0$, we first check whether it belongs to Θ or at least satisfies

$$(1 - t)\theta + t\theta_{\text{new}} \in \Theta \quad \text{for some } t > 0.$$

If we represent θ and θ_{new} as in (1) with coefficients $\alpha_0, \alpha_1, \beta_\tau$ for θ and $\alpha_{0,\text{new}}, \alpha_{1,\text{new}}, \beta_{\tau,\text{new}}$ for θ_{new}, then the latter requirement is satisfied if

$$\beta_{\tau,\text{new}} > 0 \quad \text{whenever } \tau \in D \setminus D(\theta). \tag{10}$$

If (10) is violated, we leave θ unchanged, but we replace D with $D \setminus \{\tau_o\}$, where τ_o is an index in $D \setminus D(\theta)$ such that $\beta_{\tau_o,\text{new}}$ is minimal. If (10) is satisfied, we perform a second step size correction and replace θ with $(1 - t_o)\theta + t_o\theta_{\text{new}}$, where $t_o \in (0, 1]$ is the largest number such that the latter convex combination belongs to Θ. An explicit expression for t_o is given by

$$t_o := \max\big\{t \in (0, 1] : (1 - t)\theta + t\theta_{\text{new}} \in \Theta\big\} = \min\Big(\{1\} \cup \Big\{\frac{\beta_\tau}{\beta_\tau - \beta_{\tau,\text{new}}} : \tau \in D(\theta), \beta_{\tau,\text{new}} < 0\Big\}\Big).$$

In addition, we then replace θ with $\theta - c$ for some constant c such that f_θ defines a probability density. Finally, we replace D with $D(\theta)$ for the modified candidate θ. Note that $L(\theta)$ increases strictly, and in case of $t_o < 1$, the new set D is a proper subset of the previous set D.

All in all, we obtain a new pair (θ, D) such that $L(\theta)$ has increased strictly or D is a proper subset of the former set D. Moreover, the new θ differs from the previous one if and only if the new value $L(\theta)$ is strictly larger than the previous one. In the pseudocode provided later, this whole modification of (θ, D) is written as "$(\theta, D) \leftarrow \text{StepForward}(\theta, D, \theta_{\text{new}})$".

Local search

If we start from a pair (θ, D) with $\theta \in \Theta \cap \mathbb{V}$, $L(\theta) > -\infty$ and $D \supset D(\theta)$, a local search means to iterate basic steps 1 and 2 with a certain threshold $\delta_{\text{Newton}} \geq 0$ as follows:

$(\theta_{\text{new}}, \delta) \leftarrow \text{Newton}(\theta, D)$
while $\delta > \delta_{\text{Newton}}$ do
  $(\theta, D) \leftarrow \text{StepForward}(\theta, D, \theta_{\text{new}})$
  $(\theta_{\text{new}}, \delta) \leftarrow \text{Newton}(\theta, D, \theta_{\text{new}})$
end while

Imagine for the moment that $\delta_{\text{Newton}} = 0$. After finitely many iterations, the set D would remain unchanged and be equal to $D(\theta)$, while the first assignment within the while-loop would amount to $\theta \leftarrow \theta_{\text{new}}$. That means, eventually, a local search leads to a standard Newton procedure and a locally optimal function θ.

Note also that after finitely many steps, $L(\theta)$ is strictly larger than the original value unless the starting point θ was already locally optimal while the set $D \supsetneq D(\theta)$ has been chosen poorly in the sense that basic step 2 leaves θ unchanged and results in a stepwise reduction of D until $D = D(\theta)$ again.

In practice, of course, we run a local seach with a small threshold $\delta_{\text{Newton}} > 0$. The resulting θ is called *almost locally optimal*.

Basic step 3: Deactivating constraints

Suppose that $\theta \in \Theta \cap \mathbb{V}$ is (almost) locally optimal, but (9) is violated. More precisely, suppose that $\max_{\tau \in D} h_\theta(\tau)$ is strictly larger than a given threshold $\delta_{\text{Knot}} \geq 0$. Then we choose a nonempty finite set $D_o \subset \mathcal{D} \setminus D(\theta)$ such that

$$h_\theta(\tau_o) \;>\; \delta_{\text{Knot}} \quad \text{for all } \tau_o \in D_o. \tag{11}$$

Thereafter we start a new local search with $D = D(\theta) \cup D_o$.

The obvious question is whether such a choice of D is reasonable. It may happen that during the first iterations of the local search, θ remains unchanged while elements of the set D_o are removed again. But eventually, at least one of its elements will be retained and θ will be modified. To prove this claim, we write $\theta_{\text{new}} = \theta + v + \sum_{\tau_o \in D_o} \beta_{\tau_o,\text{new}} V_{\tau_o}$ with some function $v \in \mathbb{V}_{D(\theta)}$. Then it follows from (8) that

$$0 \;<\; DL(\theta, \theta_{\text{new}} - \theta) \;=\; DL(\theta, v) + \sum_{\tau_o \in D_o} \beta_{\tau_o,\text{new}} DL(\theta, V_{\tau_o}) \;=\; \sum_{\tau_o \in D_o} \beta_{\tau_o,\text{new}} h_\theta(\tau_o),$$

and because of (11), at least one coefficient $\beta_{\tau_o,\text{new}}$, $\tau_o \in D_o$, has to be strictly positive. Consequently, starting a local search with this choice of D yields a strict improvement of $L(\theta)$ after at most $\#D_o$ iterations.

Our explicit construction of D_o depends on the current set $D(\theta)$ and essentially follows the proposal of Liu and Wang (2018). Suppose first that $D(\theta) = \emptyset$. Then we choose $D_o = \{\tau_o\}$ with a point $\tau \in \mathcal{D}$ such that $h_\theta(\tau_o) = \max_{\tau \in \mathcal{D}} h_\theta(\tau)$. Otherwise, let $\tau_1 < \cdots < \tau_m$ be the $m \geq 1$ different elements of $D(\theta)$. With $\tau_0 := -\infty$ and $\tau_{m+1} := \infty$, we set $\mathcal{D}_j := \mathcal{D} \cap (\tau_j, \tau_{j+1})$. For each $0 \leq j \leq m$ with $\mathcal{D}_j \neq \emptyset$, we determine a point $\tau_o \in \arg\max_{\tau \in \mathcal{D}_j} h_\theta(\tau)$. If $h_\theta(\tau_o)$ is greater than both δ_{Knot} and $10^{-3} \max_{\tau \in D} h_\theta(\tau)$, then τ_o is added to D_o. The latter condition on $h_\theta(\tau_o)$ prevents us from deactivating too many constraints early on, which would increase the dimensionality unnecessarily.

All in all, basic step 3 amounts to a procedure "$(h_o, D_o) \leftarrow \text{NewKnots}(\theta)$". It returns $h_o := \max_{\tau \in \mathcal{D}} h_\theta(\tau)$ and, in case of $h_o > \delta_{\text{Knot}}$, a nonempty finite set $D_o \subset \mathcal{D}$ such that $DL(\theta, V_{\tau_o}) > \max(10^{-3} h_o, \delta_{\text{Knot}})$ for all $\tau_o \in D_o$.

Explicit maximisation of h_θ

In Setting 1, maximizing h_θ over subsets of $\mathcal{D}$ is straightforward, because $\mathcal{D}$ is finite. In Settings 2A-B, suppose that $\theta \in \Theta \cap \mathbb{V}$ is (almost) locally optimal, and that $P_\theta(dx) := e^{\theta(x)} P_o(dx)$ defines a probability measure on $\mathcal{X}$. Here,

$$h_\theta(\tau) \;=\; \int V_\tau \, d(\hat{P} - P_\theta) \;=\; \int (x - \tau)^+ \, (\hat{P} - P_\theta)(dx).$$

Note that for any probability measure Q on $\mathbb{R}$ with $\int |x| \, Q(dx) < \infty$ and $\tau \in \mathbb{R}$,

$$H_Q(\tau) \;:=\; \int (x - \tau)^+ \, Q(dx)$$

defines a convex and non-increasing function $H_Q : \mathbb{R} \to [0, \infty)$ with derivatives

$$H'_Q(\tau-) \;=\; -Q([\tau,\infty)) \;=\; Q((-\infty,\tau)) - 1,$$
$$H'_Q(\tau+) \;=\; -Q((\tau,\infty)) \;=\; Q((-\infty,\tau]) - 1.$$

Hence $h_\theta = H_{\hat{P}} - H_{P_\theta}$ is a Lipschitz-continuous function on $\mathbb{R}$ with derivatives

$$h'_\theta(\tau\pm) \;=\; \hat{F}(\tau\pm) - F_\theta(\tau),$$

where $\hat{F}$ and F_θ denote the cumulative distribution functions of $\hat{P}$ and P_θ, respectively. Note that $\hat{F}$ is constant on the intervals $(-\infty, x_1)$, $[x_1, x_2)$, …, $[x_{n-1}, x_n)$, $[x_n, \infty)$ whereas F_θ is continuous on $\mathbb{R}$ and strictly increasing on $\mathcal{X}$. Consequently,

(i) h_θ is strictly concave on each interval $[x_i, x_{i+1}]$, $1 \le i < n$,

(ii) h_θ is concave and non-increasing on $(-\infty, x_1]$,

(iii) h_θ is concave and non-decreasing on $[x_n, \infty)$ with $\lim_{\tau\to\infty} h_\theta(\tau) = 0 > h_\theta(x_n)$.

The limit in (iii) follows from dominated convergence together with the fact that $(x - x_n)^+ \ge (x - \tau)^+ \to 0$ as $x_n \le \tau \to \infty$. The strict inequality for $h_\theta(x_n)$ follows from $\hat{P}((x_n,\infty)) = 0 < P_\theta((x_n,\infty))$. Hence any τ with $h_\theta(\tau) > 0$ has to satisfy $\tau < x_n$.

In Setting 2A one may even conclude from local optimality of θ that

(ii') h_θ is concave and non-increasing on $(-\infty, x_1]$ with limit $\lim_{\tau\to-\infty} h_\theta(\tau) = 0 > h_\theta(x_1)$,

because $\int (x-\tau)\,(\hat{P} - P_\theta)(dx) = 0$, so the equality $(x-\tau)^+ = x - \tau + (\tau - x)^+$ leads to the alternative representation $h_\theta(\tau) = \int (\tau - x)^+ \,(\hat{P} - P_\theta)(dx)$. Consequently, it suffices to search for local maximizers of h_θ on (x_1, x_n).

In Setting 2B, (ii) implies that the maximizer of h_θ on $[0, x_1]$ is 0. Hence it suffices to search for local maximizers of h_θ on $\{0\} \cup (x_1, x_n)$.

If we want to maximize $h = h_\theta$ on an interval $[a, b] = [x_i, x_{i+1}]$ for some $1 \le i < n$, we could first check whether $h'(a+) \le 0$ or $h'(b-) \ge 0$. In these cases, $h(a) = \max_{\tau\in[a,b]} h(\tau)$ or $h(b) = \max_{\tau\in[a,b]} h(\tau)$, respectively. In case of $h'(a+) > 0 > h'(b-)$, we determine the unique point $\tau \in (a, b)$ satisfying $h'_\theta(\tau) = 0$. In general, this leads to a numerical approximation of τ, but in our specific examples for Settings 2A-B, τ may be computed explicitly by means of the standard Gaussian or gamma quantile functions (see Sections A.3 and A.4).

Finding a starting point θ

One possibility to determine a starting point θ is to activate all constraints initially and find an optimal function in $\mathbb{V}_\emptyset \subset \Theta$. In Setting 2A, we are then looking for a function $\theta(x) = \hat{\kappa}x - c(\hat{\kappa})$ with $c(\kappa) := \log \int_{\mathcal{X}} e^{\kappa x}\, P_o(dx)$, and $\hat{\kappa} \in \mathbb{R}$ is the unique real number such that $c'(\hat{\kappa}) = \hat{\mu}$. Specifically, if $P_o = \mathcal{N}(0,1)$, then $c(\kappa) = \kappa^2/2$, whence $\hat{\kappa} = \hat{\mu}$.

In Setting 2B, activating all constraints would lead to the trivial space $\mathbb{V}_\emptyset = \{0\}$. Alternatively, one could determine an optimal function in $\mathbb{V}_{\{0\}} \cap \Theta$. With $\hat{\kappa}$ as before, i.e. $c'(\hat{\kappa}) = \hat{\mu}$, the optimal function θ is given by $\theta(x) = \hat{\kappa}^+ x - c(\hat{\kappa}^+)$. Specifically, if $P_o = \text{Gamma}(\alpha, \beta)$, then $c(\kappa) = -\alpha \log((1 - \kappa/\beta)^+)$, so that $\hat{\kappa} = \beta - \alpha/\hat{\mu}$.

All in all, for Settings 2A-B we obtain a starting point $\theta \in \Theta$ that is locally optimal and depends only on $\hat{\mu}$, indicated as "$\theta \leftarrow \text{Start}(\hat{\mu})$".

In Setting 1, finding an optimal function in $\mathbb{V}_\emptyset$ would amount to solving a nonlinear equation numerically. Alternatively, we start with the MLE θ of a Gaussian log-density up to an additive constant, i.e.

$$\theta_0(x) \;:=\; -(x - \hat{\mu})^2/(2\hat{\sigma}^2)$$

with $\hat{\mu} = \sum_{i=1}^n w_i x_i$ and $\hat{\sigma}^2 := \sum_{i=1}^n w_i (x_i - \hat{\mu})^2$. Next, we fix a nonempty set $D_0 \subset \mathcal{D}$ and replace θ_0 with the unique linear spline $\theta \in \mathbb{V}_{D_0}$ such that $\theta \equiv \theta_0$ on $D_0 \cup \{x_1, x_n\}$, before normalizing it via $\theta \leftarrow \theta - \log\big(\int_{x_1}^{x_n} e^{\theta(x)}\, dx\big)$. All of these operations are hidden behind "$\theta \leftarrow \text{Start}(\hat{\mu}, \hat{\sigma}, D_0)$" in the subsequent pseudocode. Note that this starting point θ is not locally optimal in general.

3.5. Complete algorithms

In Settings 2A-B, where a locally optimal starting point is easily found, our complete algorithm works as follows:

```
θ ← Start(μ̂)
(h_o, D_o) ← NewKnots(θ)
while h_o > δ_Knot do
   D ← D(θ) ∪ D_o
   # Local search:
   (θ_new, δ) ← Newton(θ, D)
   while δ > δ_Newton do
      (θ, D) ← StepForward(θ, D, θ_new)
      (θ_new, δ) ← Newton(θ, D)
   end while
   # Check global optimality:
   (h_o, D_o) ← NewKnots(θ)
end while
```

In Setting 1, our algorithm has a slightly different beginning, because the starting point θ is not locally optimal:

```
θ ← Start(μ̂, σ̂, D_0)
(D_o, h_o) ← (∅, ∞)
while h_o > δ_Knot do
   …
end while
```

Note that in Setting 1, an affine transformation $x \mapsto a + bx$ of our data with $b > 0$ would result in new directional derivatives $DL(\theta, V_{\tau_o})$ that differ from the original values by this factor b. By way of contrast, the output δ of Newton(θ, D) is invariant under such transformations. Hence it is advisable to distinguish the stopping thresholds δ_{Newton} and δ_{Knots}, where $\delta_{\text{Knot}} > 0$ is chosen to be a small constant times $\hat{\sigma}$. In Settings 2A-B the parameter δ_{Knot} should reflect the spread of the reference distribution P_o.

3.6. Convergence

After circulating a first version of the present paper, Sommer-Simpson (2019) provided a proof of convergence of our algorithm in Setting 2B. Lemma 3.1 below implies that in all three settings, the output of our algorithm is arbitrarily close to $\hat{\theta}$ if δ_{Knot} and δ_{Newton} are sufficiently small. Our proof of this generalizes and simplifies the arguments of Sommer-Simpson (2019).

To formulate the result, let $\theta \in \Theta \cap \mathbb{V}$ with $L(\theta) > -\infty$. To check local optimality of θ, we perform a Newton step for L on the parameter space $\mathbb{V}_{D(\theta)}$. This yields a function $\theta_{\text{new}} \in \mathbb{V}_{D(\theta)}$ maximizing a second order Taylor approximation of L on $\mathbb{V}_{D(\theta)}$, and the directional derivative

$$\delta_{\text{Newton}}(\theta) \;:=\; DL(\theta, \theta_{\text{new}} - \theta).$$

In our algorithm, θ is viewed as approximately locally optimal if $\delta_{\text{Newton}}(\theta)$ is smaller than a given number δ_{Newton}. If that is the case, we check whether

$$\delta_{\text{Knot}}(\theta) \;:=\; \max_{\tau \in D} DL(\theta, V_{\tau,\theta})$$

is smaller than a given number δ_{Knot}. Note also that during our algorithm, the value $L(\theta)$ never decreases.

Lemma 3.1. *In all Settings and for any constant $L_o \in (-\infty, L(\hat{\theta}))$, there exist constants C_{Newton} and C_{Knot} such that for all $\theta \in \Theta \cap \mathbb{V}$ with $L(\theta) \geq L_o$,*

$$L(\hat{\theta}) - L(\theta) \;\leq\; C_{\text{Newton}} \sqrt{\delta_{\text{Newton}}(\theta)} + C_{\text{Knot}} \delta_{\text{Knot}}(\theta).$$

Remark 3.2. For $\theta \in \Theta \cap \mathbb{V}$, it follows from $L(\theta) \to L(\hat{\theta})$ that $\|\theta - \hat{\theta}\| \to 0$, where $\|\cdot\|$ is the norm in (2).

Sample size	100	200	500	10^3	10^4	10^5
Rel. efficiency	2.003	1.988	2.467	2.749	3.615	6.067
Running time (s)	$4.425 \cdot 10^{-3}$	$6.310 \cdot 10^{-3}$	$8.450 \cdot 10^{-3}$	0.0115	0.1029	1.4805

Table 1
Mean relative efficiencies and running times of the new algorithm for Setting 1 with Gaussian samples.

4. Numerical examples, simulations and an application

4.1. Comparisons in Setting 1

An obvious question is how much better the new algorithm for Setting 1 is in comparison to the active set method of Dümbgen and Rufibach (2011). To enable a fair comparison, we implemented the latter method as follows:

$\theta \leftarrow \text{Start}(\hat{\mu}, \hat{\sigma}, D_0)$
$h_o \leftarrow \infty$
$D \leftarrow D_0$
while $h_o > \delta_{\text{Knot}}$ do
  $(\theta_{\text{new}}, \delta) \leftarrow \text{Newton}(\theta, D)$
  while $\delta > \delta_{\text{Newton}}$ do
    $\theta_{\text{new}} \leftarrow \text{Newton}(\theta, D)$
    $\theta_{\text{new}} \leftarrow \text{Normalize}(\theta_{\text{new}})$
  end while
  if $\theta_{\text{new}} \in \Theta$ then
    $\theta \leftarrow \theta_{\text{new}}$
    $(h_o, \tau_o) \leftarrow \text{NewKnot}(\theta)$
    $D \leftarrow D(\theta) \cup \{\tau_o\}$
  else
    $(\theta, D) \leftarrow \text{StepForward}(\theta, D, \theta_{\text{new}})$
  end if
end while

Here $\theta \leftarrow \text{Normalize}(\theta)$ stands for replacing θ with $\theta - c$ such that f_θ defines a probability density. And "$(h_o, \tau_o) \leftarrow \text{NewKnot}(\theta)$" returns only one point $\tau_o \in \mathcal{D}$ with maximal directional derivative $h_o = DL(\theta, \tau_o)$. This is the first main difference between the old and the new algorithm. The second main difference is that a full Newton procedure is run on $\mathbb{V}_D$ without checking and enforcing the shape-constraint that $\theta \in \Theta$. An advantage of omitting the shape-constraint is that the Newton search runs a bit faster. A disadvantage is that we sometimes iterate and optimize in a region far from Θ, whereas in the subsequent $\text{StepForward}(\theta, D, \theta_{\text{new}})$, only a rather small step is performed.

Concerning D_0, extensive numerical experiments showed that the choice $D_0 = \{x_{j(1)}, x_{j(2)}, x_{j(3)}\}$ with approximately equispaced indices $1 < j(1) < j(2) < j(3) < n$ is a good choice for a broad range of sample sizes n. With this choice, we simulated a random sample of size n from the standard Gaussian distribution 200 times and fitted a log-concave density with the old and the new method. Figure 1 shows boxplots of the running time with the old method divided by the running time with the new method. One sees clearly that the improvement is substantial, particularly for large sample sizes. It is similar in magnitude to the improvements reported by Wang (2018) for the algorithm of Liu and Wang (2018). Table 1 reports the means of these relative efficiencies as well as the mean absolute running times. The methods have been implemented in pure R code, and the simulations have been performed on a MacBook Pro (2.6 GHz 6-Core Intel Core i7), the stopping thresholds being $\delta_{\text{Newton}} = 10^{-7}/n$ and $\delta_{\text{Knot}} = 10^{-7}\hat{\sigma}/n$.

4.2. Numerical examples for Settings 2A-B

Setting 2A. Inspired by the testing problem described in Section 4.3, we simulated $n = 400$ independent observations X_i with distribution $P_i = \mathcal{N}(0, 1)$ for $i > 20$ and $P_i = \mathcal{N}(1.5, 1)$ for $i \leq 20$. With the reference distribution $P_o = \mathcal{N}(0, 1)$, the corresponding log-density ratio equals

$$\theta(x) \;=\; \log \frac{dP}{dP_o}(x) \;=\; \log(0.95 + 0.05\, e^{1.5x - 1.125}).$$

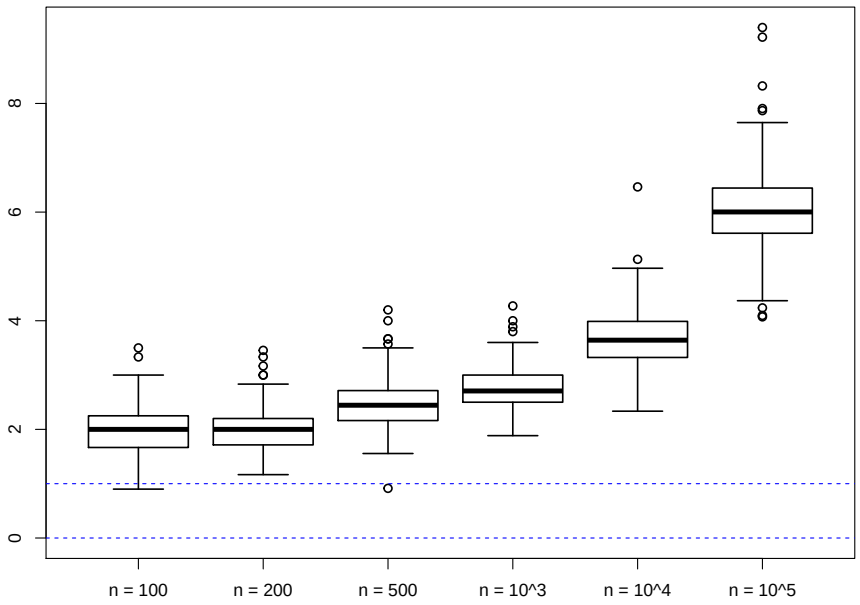

Figure 1: Relative efficiencies of the new algorithm for Setting 1 with Gaussian samples.

The resulting estimator $\hat{\theta}$ had $m = 5$ knots, and Figure 2 depicts the function

$$t \;\mapsto\; h(t) = DL(\hat{\theta}, V_t),$$

where the knots of $\hat{\theta}$ are indicated by vertical lines. As predicted by theory, $h(t) \leq 0$ for all t, with equality when $t \in D(\hat{\theta})$. Figure 3 depicts the true and estimated tail inflation functions θ and $\hat{\theta}$. Figure 4 shows the corresponding densities $p_o = \phi$, $p = e^{\theta} p_o$ and $\hat{p} = e^{\hat{\theta}} p_o$. Note that the estimator $\hat{p}$ captures the heavier right tail of p in comparison to p_o. Applying the goodness-of-fit test described in Section 4.3 to this particular data set yielded a Monte-Carlo p-value smaller than 10^{-3} (with $10^5 - 1$ simulations) for the null hypothesis that all 400 observations are standard Gaussian.

Setting 2B. We simulated $n = 1000$ independent observations X_i such that $X_i \sim \chi^2_1$ for $i > 200$, $X_i/1.4 \sim \chi^2_1$ for $100 < i \leq 200$ and $X_i/2 \sim \chi^2_1$ for $i \leq 100$. With the reference distribution $P_o = \chi^2_1$, the corresponding log-density ratio equals

$$\theta(x) \;=\; \log\bigl(8 + 1.4^{-1/2} e^{x/7} + 2^{-1/2} e^{x/4}\bigr) - \log 10.$$

The estimator $\hat{\theta}$ in this case had $m = 6$ knots, and Figures 5 and 6 are analogous to the displays for Setting 2A, showing the directional derivatives $h(\tau) = DL(\hat{\theta}, V_\tau)$ and the log-density ratios $\theta, \hat{\theta}$, respectively. Applying the goodness-of-fit test described in Section 4.3 to this particular data set yielded a Monte-Carlo p-value of 10^{-5} (with $10^5 - 1$ simulations) for the null hypothesis that all 400 observations have distribution χ^2_1.

4.3. Data-driven goodness-of-fit tests

With the estimator $\hat{\theta}$ at hand, one may use the likelihood ratio statistic

$$T_{LR}(X_1, \ldots, X_n) \;:=\; \sum_{i=1}^{n} \hat{\theta}(X_i)$$

to test the null hypothesis that all distributions P_i are equal to P_o versus the alternative hypothesis that the marginal P has a convex log-density $\theta \not\equiv 0$ with respect to P_o. Large values of T_{LR} indicate a violation of the null hypothesis. The distribution of this test statistic under the null hypothesis is unknown but can be easily estimated via Monte Carlo simulations.

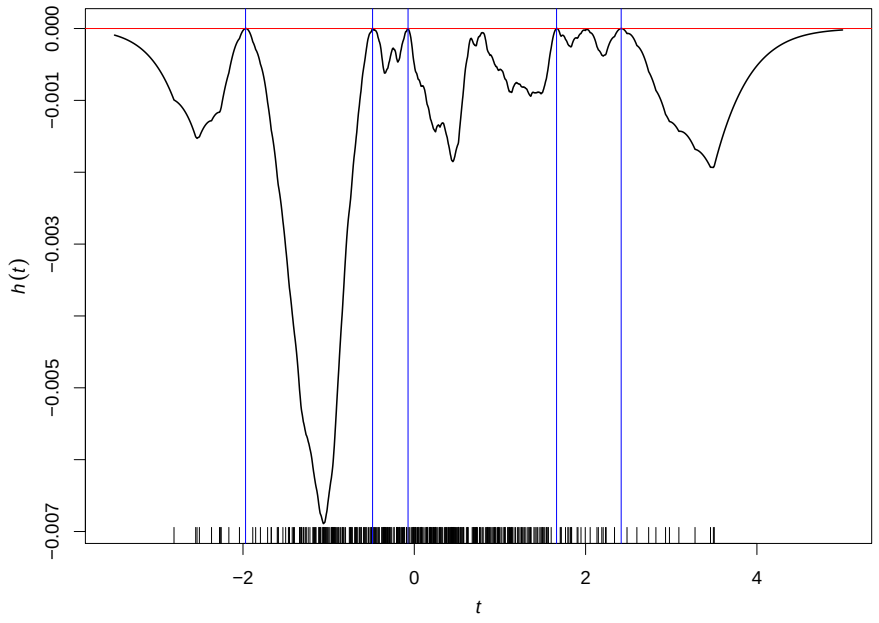

Figure 2: Directional derivatives $h(t) = DL(\hat{\theta}, V_t)$ for data example in Setting 2A.

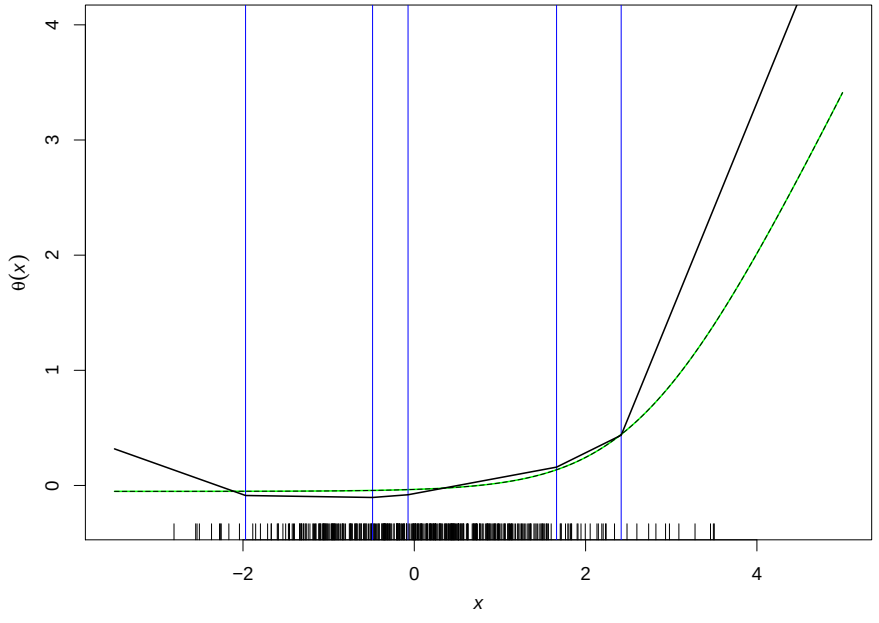

Figure 3: True (green, dashed) and estimated (black) tail inflation functions θ and $\hat{\theta}$ for data example in Setting 2A.

Specifically, consider Setting 2A with $P_o = \mathcal{N}(0,1)$. As mentioned before, if each distribution P_i, and thus the marginal P, is a mixture of Gaussian distributions with standard deviation at least 1, then $\theta = \log(dP/dP_o)$ is convex. This renders T_{LR} an interesting alternative to higher criticism statistics as introduced by Donoho and Jin (2004) and Gontscharuk et al. (2016). In the subsequent power simulations, we focus on a particular union-intersection test similar to those considered by the latter authors: With the order statistics $X_{(1)} < \cdots < X_{(n)}$ of the X_i, note that under H_o, the random variables $\Phi(X_{(1)}), \ldots, \Phi(X_{(n)})$ are distributed like the order statistics of a sample from the uniform distribution on $[0,1]$. In particular, $\Phi(X_{(i)})$ follows the beta distribution with parameters i and $n+1-i$. Denoting the corresponding

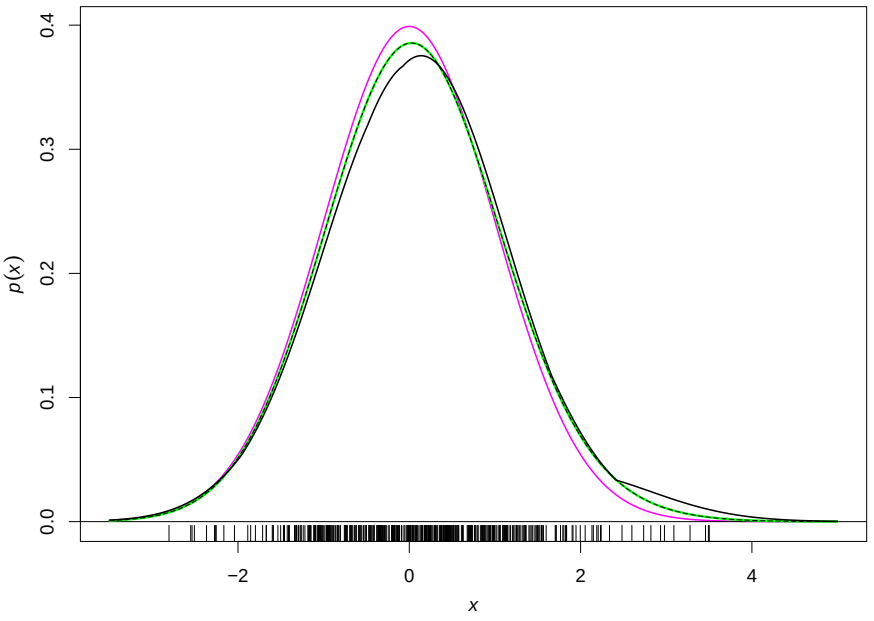

Figure 4: Lebesgue densities p_o (magenta), $p = e^{\theta} p_o$ (green, dashed) and $\hat{p} = e^{\hat{\theta}} p_o$ (black) for data example in Setting 2A.

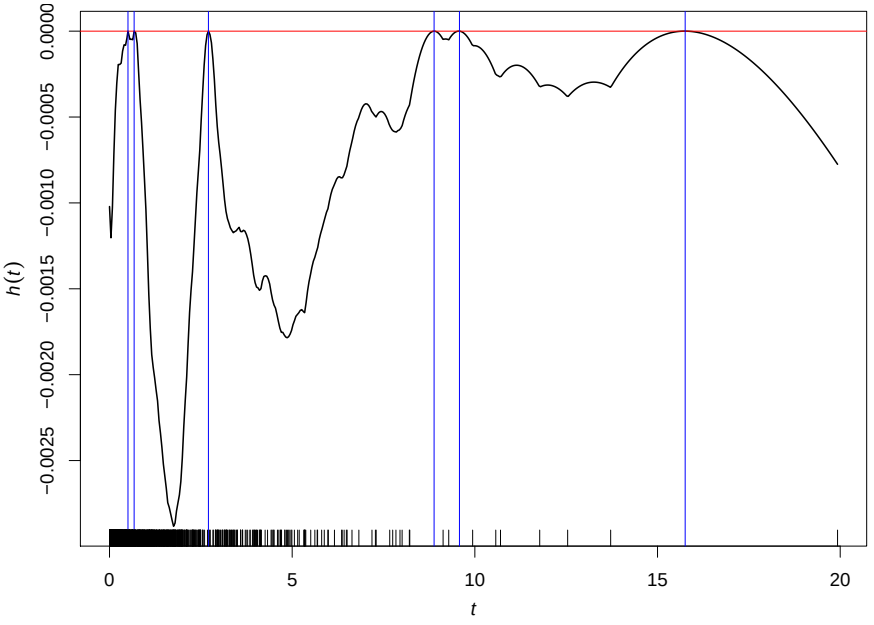

Figure 5: Directional derivatives $h(t) = DL(\hat{\theta}, V_t)$ for data example in Setting 2B.

distribution function with $B_{i,n+1-i}$, a union-intersection test statistic of H_o is given by

$$\begin{aligned} T_{UI}(X_1, \ldots, X_n) &:= \min\Big(\min_{i<(n+1)/2} B_{i,n+1-i}(\Phi(X_{(i)})), \min_{i>(n+1)/2} \big(1 - B_{i,n+1-i}(\Phi(X_{(i)}))\big) \Big) \\ &= \min\Big(\min_{i<(n+1)/2} B_{i,n+1-i}(\Phi(X_{(i)})), \min_{i>(n+1)/2} B_{n+1-i,i}(\Phi(-X_{(i)})) \Big), \end{aligned}$$

with small values indicating a violation of H_o. The rationale behind this test statistic is as follows: If H_o is violated

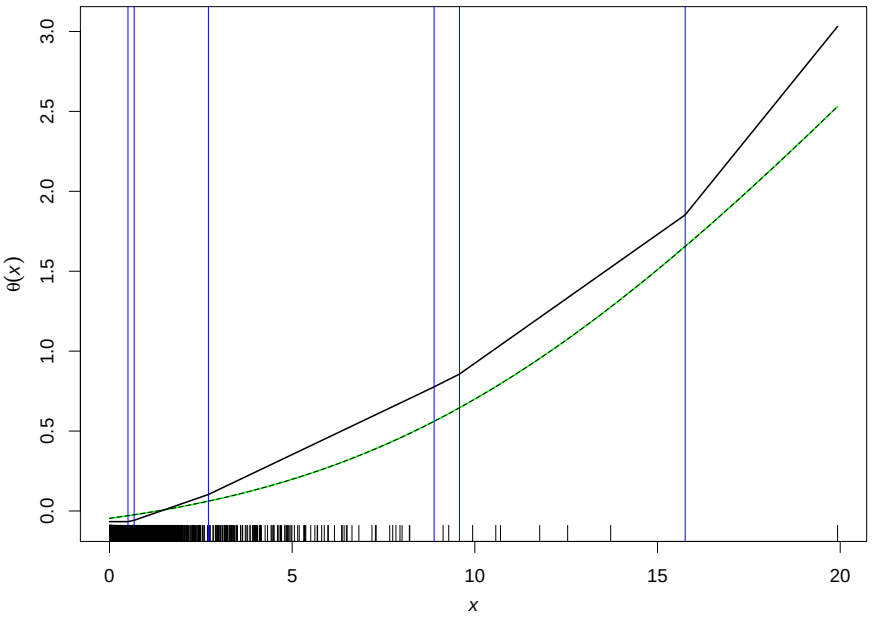

Figure 6: True (green, dashed) and estimated (black) tail inflation functions θ and $\hat{\theta}$ for data example in Setting 2B.

and θ is convex, then the left tail of P is heavier than that of P_o, leading to smaller order statistics $X_{(1)}, X_{(2)}, \ldots$, or the right tail of P is heavier than that of P_o, leading to larger order statistics $X_{(n)}, X_{(n-1)}, \ldots$. For numerical reasons, we also use the identity $1 - B_{i,n+1-i}(\Phi(x)) = B_{n+1-i,i}(\Phi(-x))$.

In a large simulation study involving different sample sizes n, we estimated the $(1-\alpha)$-quantile of the null distribution of $T_{LR}(X_1, \ldots, X_n)$ and the α-quantile of $T_{UI}(X_1, \ldots, X_n)$ in $10^5 - 1$ Monte Carlo simulations, where $\alpha = 1\%, 5\%$. With these critical values, we estimated the power of the two tests at level α under the following distribution of the sample: For a fixed distribution P_* on the real line and a subset $J \subset \{1, 2, \ldots, n\}$ with $k \geq 0$ elements, the distributions P_i of the random variables X_i are given by

$$P_i = \begin{cases} P_* & \text{if } i \in J, \\ P_o & \text{otherwise.} \end{cases}$$

Specifically, we used $P_* = \mathcal{N}(1.5, 1)$ and $P_* = \mathcal{N}(0, 3)$. This setting is similar to that of Donoho and Jin (2004) with $P_i = (1 - k/n)P_o + (k/n)P_*$ for all i. The latter setting corresponds to a random set J with $\#J$ having binomial distribution $\mathrm{Bin}(n, k/n)$.

For these two choices of P_*, Figures 7, 8 and 9 show the power $\mathbb{P}$(reject H_o at level α) of both tests as a function of $k = \#J$. Clearly, the test based on T_{LR} has higher power than that based on T_{UI}. The difference when $P_* = \mathcal{N}(0, 3)$ is stronger than in the case of the simple shift alternative $P_* = \mathcal{N}(1.5, 1)$.

Section A.6 contains further information about the null distribution of T_{LR} for different sample sizes and $P_o = \mathcal{N}(0, 1)$ or $P_o = \chi^2_1$. Note that the test described here, when implemented as a Monte-Carlo test, has exact test level α for any sample size n. Its (asymptotic) power properties are beyond the scope of this paper and a potential topic for future research.

5. Proofs

An essential ingredient for the proof of Lemmas 2.5, 2.6 and 2.7 is the following coercivity result.

Lemma 5.1. *Let M be a measure on $\mathbb{R}$, and let $L(\theta) := \int \theta \, d\hat{P} - \int e^\theta \, dM + 1$ for measurable functions $\theta : \mathbb{R} \to \mathbb{R}$.*
(a) *Suppose that $M(B) = \mathrm{Leb}(B \cap [x_1, x_n])$. Then for concave functions $\theta : \mathbb{R} \to \mathbb{R}$,*

$$L(\theta) \to -\infty \quad as \quad \max_{x \in [x_1, x_n]} |\theta(x)| \to \infty.$$

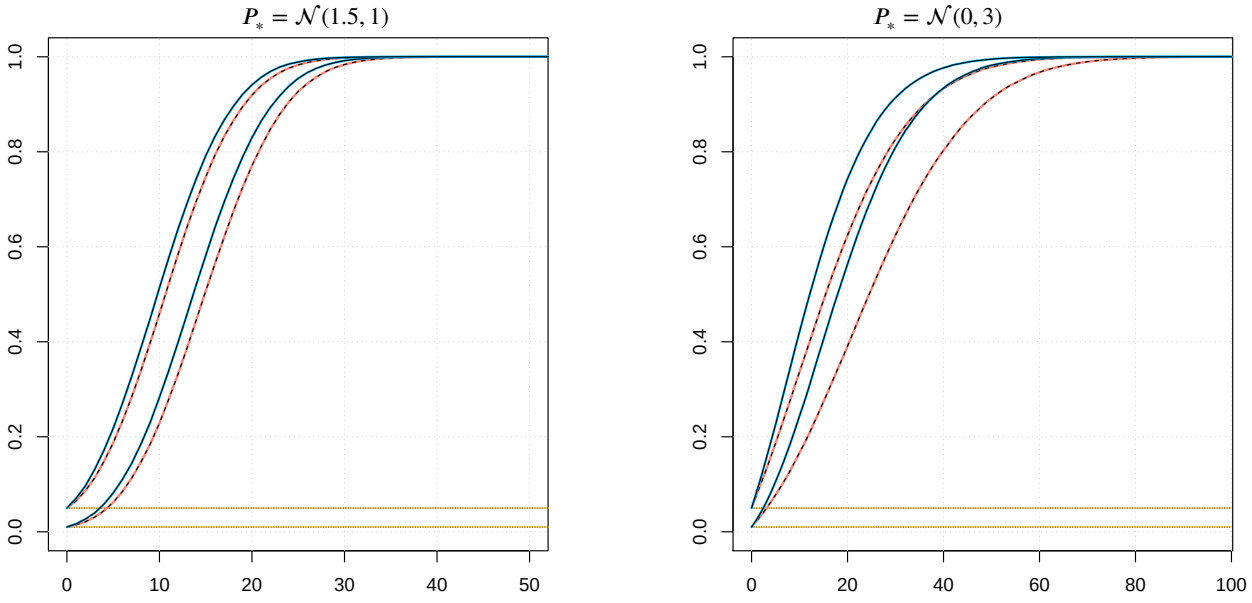

Figure 7: Power of goodness-of-fit tests based on T_{LR} (blue, solid) and T_{UI} (red, dashed) as a function of k for two distributions P_* and sample size $n = 100$. The test levels α are 5% and 1%.

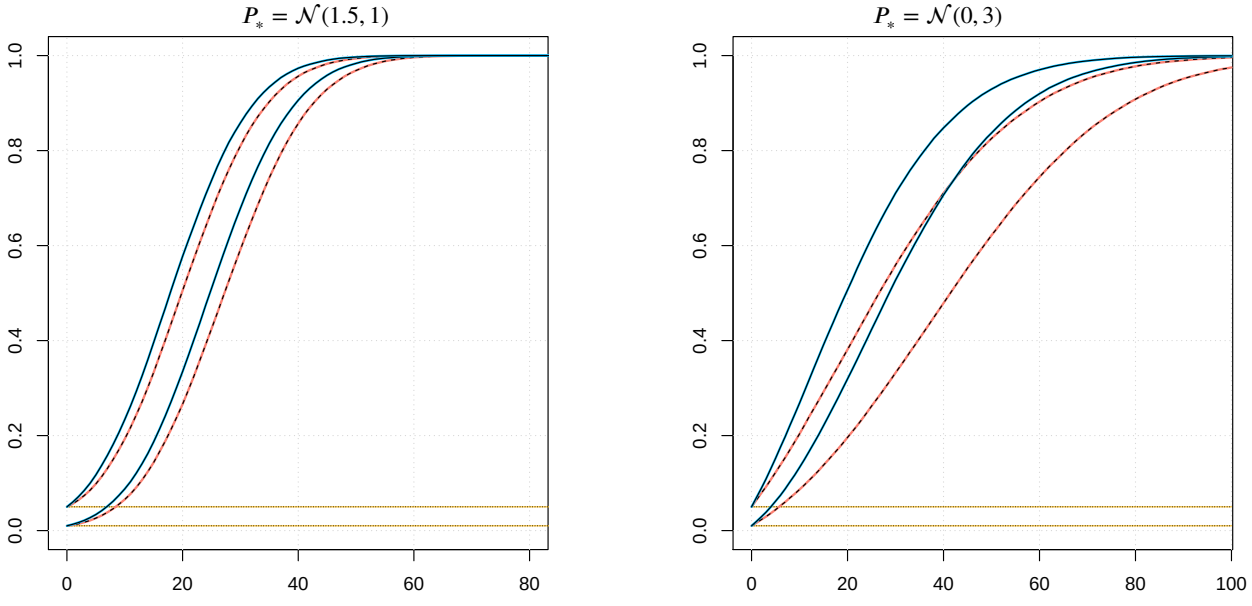

Figure 8: Power comparison for sample size $n = 400$.

(b) *Suppose that the three numbers* $M((-\infty, x_1))$, $M([x_1, x_n])$ *and* $M((x_n, \infty))$ *are strictly positive. Then for convex functions* θ,

$$L(\theta) \to -\infty \quad as \quad \max_{x \in [x_1, x_n]} |\theta(x)| + \max\{-\theta'(x_1-), \theta'(x_n+)\} \to \infty.$$

Part (a) is known from Dümbgen et al. (2007/2011), but for the reader's convenience and later reference, a simplified argument is also given here.

Proof of Lemma 5.1. Let $i(\theta) := \min_{x \in [x_1, x_n]} \theta(x)$, $s(\theta) := \max_{x \in [x_1, x_n]} \theta(x)$ and $r(\theta) := s(\theta) - i(\theta)$.

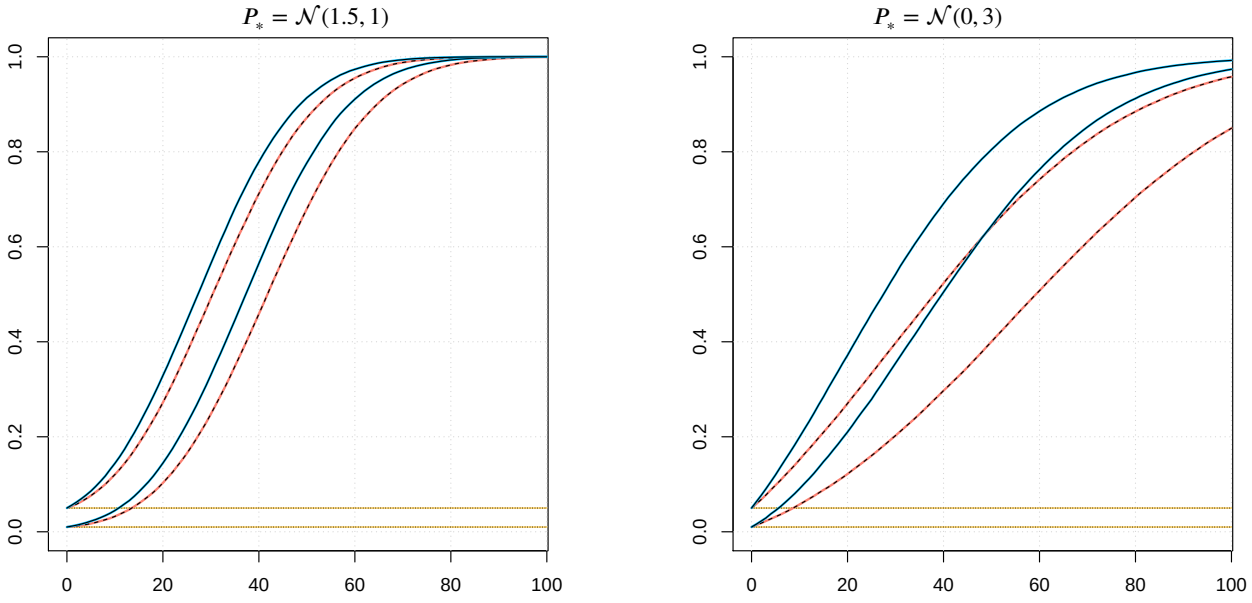

Figure 9: Power comparison for sample size $n = 1000$.

As to part (a), note first that

$$L(\theta) \;\le\; s(\theta) - e^{i(\theta)}(x_n - x_1) + 1 \;=\; i(\theta) - e^{i(\theta)}(x_n - x_1) + r(\theta) + 1.$$

The right-hand side converges to $-\infty$ if either $s(\theta) \to -\infty$ or $i(\theta) \to \infty$ while $r(\theta)$ stays bounded. Thus it suffices to show that $L(\theta) \to -\infty$ as $r(\theta) \to \infty$. By concavity of θ, the difference $\theta - i(\theta)$ is bounded from below on $[x_1, x_n]$ by a piecewise linear function with values in $[0, r(\theta)]$, and the value 0 is attained at x_1 or at x_n. Hence, with $w_{\min} := \min(w_1, w_n)$, we may conclude that

$$\begin{aligned} L(\theta) \;&\le\; i(\theta) + (1 - w_{\min})r(\theta) - e^{i(\theta)} \int_{x_1}^{x_n} e^{\theta(x) - i(\theta)}\, dx + 1 \\ &\le\; i(\theta) + (1 - w_{\min})r(\theta) - e^{i(\theta)}(x_n - x_1) \int_0^1 e^{r(\theta)t}\, dt + 1 \\ &\le\; i(\theta) + (1 - w_{\min})r(\theta) - e^{i(\theta)}(x_n - x_1)(e^{r(\theta)} - 1)/r(\theta) + 1. \end{aligned}$$

For fixed $r(\theta)$, the maximum of the latter bound with respect to $i(\theta)$ equals

$$-\log(x_n - x_1) - \log(1 - e^{-r(\theta)}) + \log r(\theta) - w_{\min} r(\theta),$$

and this converges to $-\infty$ as $r(\theta) \to \infty$.

As to part (b), convexity of θ implies that either

$$s(\theta) \;=\; \theta(x_1) \;>\; \theta(x_n), \quad -\theta'(x_1-) \;\ge\; \frac{r(\theta)}{x_n - x_1} \quad \text{and} \quad \theta(x) \;\ge\; s(\theta) + \theta'(x_1-)(x - x_1) \;\text{ for } x \le x_1, \tag{12}$$

or

$$s(\theta) \;=\; \theta(x_n) \;\ge\; \theta(x_1), \quad \theta'(x_n+) \;\ge\; \frac{r(\theta)}{x_n - x_1} \quad \text{and} \quad \theta(x) \;\ge\; s(\theta) + \theta'(x_n+)(x - x_n) \;\text{ for } x \ge x_n. \tag{13}$$

Hence with $\mathcal{X}_\ell := (-\infty, x_1)$ and $\mathcal{X}_r := (x_n, \infty)$,

$$L(\theta) \;\le\; s(\theta) - e^{s(\theta)} \min\bigl\{M(\mathcal{X}_\ell), M(\mathcal{X}_r)\bigr\} + 1 \;\to\; -\infty \quad \text{as } |s(\theta)| \to \infty,$$

because $M(\mathcal{X}_\ell), M(\mathcal{X}_r) > 0$. Moreover,

$$L(\theta) \;\le\; s(\theta) - e^{s(\theta)} \int e^{\theta - s(\theta)}\, d\boldsymbol{M} + 1 \;\le\; \sup_{s\in\mathbb{R}}\Big(s - e^s \int e^{\theta - s(\theta)}\, d\boldsymbol{M}\Big) + 1 \;=\; -\log \int e^{\theta - s(\theta)}\, d\boldsymbol{M},$$

and the right-hand side is not larger than

$$\begin{cases} -\log \displaystyle\int_{\mathcal{X}_\ell} e^{\theta'(x_1-)(x-x_1)}\, d\boldsymbol{M} - 1 & \text{in case of (12)} \\ -\log \displaystyle\int_{\mathcal{X}_r} e^{\theta'(x_n+)(x-x_n)}\, d\boldsymbol{M} - 1 & \text{in case of (13)} \end{cases}$$
$$\le\; -\min\left\{ \log \int_{\mathcal{X}_\ell} e^{-r(\theta)(x-x_1)/(x_n-x_1)}\, d\boldsymbol{M},\; \log \int_{\mathcal{X}_r} e^{r(\theta)(x-x_n)/(x_n-x_1)}\, d\boldsymbol{M} \right\} - 1.$$

Hence these inequalities show that

$$L(\theta) \;\to\; -\infty \quad \text{as } r(\theta) + \max\big\{-\theta'(x_1-), \theta'(x_n+)\big\} \to \infty. \qquad \square$$

Proof of Lemmas 2.6 and 2.7. We first consider Setting 2A. For an arbitrary function $\theta \in \Theta$ let

$$\tilde\theta(x) \;:=\; \begin{cases} \theta(x_1) + (x - x_1)\theta'(x_1+) & \text{if } x \le x_1, \\ \theta(x) & \text{if } x \in [x_1, x_n], \\ \theta(x_n) + (x - x_n)\theta'(x_n-) & \text{if } x \ge x_n. \end{cases}$$

Then $\tilde\theta \le \theta$, $\tilde\theta \equiv \theta$ on $[x_1, x_n]$, and $L(\tilde\theta) \ge L(\theta)$ with equality if, and only if $\tilde\theta \equiv \theta$. Thus we may restrict our attention to convex functions θ such that $\theta' \equiv \theta'(x_1+)$ on $(-\infty, x_1]$ and $\theta' \equiv \theta'(x_n-)$ on $[x_n, \infty)$.

Let $(\theta_k)_k$ be a sequence of such functions such that $\lim_{k\to\infty} L(\theta_k) = \sup_{\theta\in\Theta} L(\theta)$. By Lemma 5.1,

$$\sup_k \Big(\sup_{x\in[x_1,x_n]} |\theta_k(x)| + \max\big\{-\theta_k'(x_1), \theta_k'(x_n)\big\}\Big) \;<\; \infty.$$

Consequently, the sequence $(\theta_k)_k$ is uniformly bounded on $[x_1, x_n]$ and uniformly Lipschitz continuous on $\mathbb{R}$. Hence we may apply the theorem of Arzela–Ascoli and replace $(\theta_k)_k$ with a subsequence, if necessary, such that $\theta_k \to \theta \in \Theta$ pointwise and uniformly on any compact set as $k \to \infty$. By Fatou's lemma, $L(\theta) \ge \lim_{k\to\infty} L(\theta_k)$, so θ is a maximizer of L over Θ.

One can easily deduce from strict convexity of $\exp(\cdot)$ that L is strictly concave on Θ. Hence there exists a unique maximizer $\hat\theta$ of L over Θ.

Let

$$\check\theta(x) \;:=\; \max_{i=1,\dots,n} \big(\hat\theta(x_i) + \hat\theta'(x_i)(x - x_i)\big)$$

with $\hat\theta'(x_i-) \le \hat\theta'(x_i) \le \hat\theta'(x_i+)$ for $2 \le i < n$. This defines another function $\check\theta \in \Theta$ such that $(\check\theta(x_i))_{i=1}^n = (\hat\theta(x_i))_{i=1}^n$ and $\check\theta \le \hat\theta$. Thus we may conclude that $\hat\theta \equiv \check\theta$, a function with at most $n-1$ changes of slope, all of which are within (x_1, x_n).

Suppose that $\hat\theta$ changes slope at two points $\tau_1 < \tau_2$ but (τ_1, τ_2) contains no observation x_i. Then we could redefine

$$\hat\theta(x) \;:=\; \max\big(\hat\theta(\tau_1) + \hat\theta'(\tau_1-)(x - \tau_1), \hat\theta(\tau_2) + \hat\theta'(\tau_2+)(x - \tau_2)\big)$$

for $x \in (\tau_1, \tau_2)$. This modification would not change $(\hat\theta(x_i))_{i=1}^n$ but decrease strictly the integral $\int e^{\hat\theta(x)}\, P_o(dx)$, a contradiction to optimality of $\hat\theta$. Hence any interval $[x_i, x_{i+1}]$, $1 \le i < n$, contains at most one point τ such that $\hat\theta'(\tau-) < \hat\theta'(\tau+)$.

Finally, as argued in Section 3.3, $\hat\theta$ satisfies the (in)equalities

$$h(\tau) \;:=\; \int (x - \tau)^+\, (\hat P - P_{\hat\theta})(dx) \;\begin{cases} \le\; 0 & \text{for all } \tau \in (x_1, x_n), \\ =\; 0 & \text{if } \hat\theta'(\tau-) < \hat\theta'(\tau+). \end{cases}$$

But $h(\cdot)$ itself is continuous with one-sided derivatives

$$h'(\tau\pm) \;=\; \hat{F}(\tau\pm) - F_{\hat{\theta}}(\tau),$$

where $\hat{F}$ and $F_{\hat{\theta}}$ are the distribution functions of $\hat{P}$ and $P_{\hat{\theta}}$, respectively. If $\hat{\theta}$ changes slope at some point τ, then it follows from $h \le 0 = h(\tau)$ that $h'(\tau-) \ge 0 \ge h'(\tau+)$, so

$$0 \;\ge\; h'(\tau+) - h'(\tau-) \;=\; \hat{P}(\{\tau\}).$$

Hence τ cannot be an observation x_i.

These arguments prove Lemma 2.6. The same arguments apply to Setting 2B without essential changes, because the functions $\tilde{\theta}, \check{\theta}$ and $\theta = \lim_{k\to\infty} \theta_k$ above are automatically isotonic. The only difference, merely notational, is that in case of $\hat{\theta}'(0+) > 0$ we interpret 0 as a first knot τ_1. Hence Lemma 2.7 is also true. □

Proof of Lemma 3.1. We prove the lemma for Setting 2A. The arguments for Setting 2B and Setting 1 are very similar, see Section A.5. Let Θ_o be the set of all functions $\theta \in \Theta \cap \mathbb{V}$ such that $L(\theta) \ge L_o$. Obviously, the target function $\hat{\theta}$ belongs to Θ_o. It follows from Lemma 5.1 that

$$C_o \;:=\; \sup_{\theta\in\Theta_o} \sup_{x\in[x_1,x_n]} |\theta(x)| \;<\; \infty,$$

and

$$C_\ell \;:=\; \inf_{\theta\in\Theta_o} \theta'(x_1-) \;>\; \lambda_\ell(P_o), \quad C_r \;:=\; \sup_{\theta\in\Theta_o} \theta'(x_n+) \;<\; \lambda_r(P_o).$$

For arbitrary $\theta \in \Theta_o$, let $\theta_{\text{new}} \in \mathbb{V}_{D(\theta)}$ be the subsequent Newton proposal. Precisely, $\theta_{\text{new}} - \theta$ maximizes the second order Taylor approximation

$$L(\theta) + DL(\theta, \upsilon) - 2^{-1} H(\theta, \upsilon)$$

of $L(\theta + \upsilon)$ over all $\upsilon \in \mathbb{V}_{D(\theta)}$, and elementary considerations show that

$$DL(\theta, \theta_{\text{new}} - \theta) \;=\; \max_{\upsilon\in\mathbb{V}_{D(\theta)}\setminus\{0\}} \frac{DL(\theta,\upsilon)^2}{H(\theta,\upsilon)}.$$

Now let $\mathcal{V}$ be the set of basis functions $\upsilon_0(x) := 1$, $\upsilon_1(x) := x - x_1$ and $V_\tau(x) = (x-\tau)^+$, $\tau \in D$. Then for any $\upsilon \in \mathcal{V}$,

$$H(\theta,\upsilon) \;=\; \int \upsilon^2 e^\theta \, dP_o \;\le\; C_{\text{N}} \;:=\; \int \upsilon_{\max}(x)^2 e^{\theta_{\max}(x)} \, P_o(dx) \;<\; \infty,$$

where $\upsilon_{\max}(x) := \max(1, |x-x_1|)$ is an upper bound for $|\upsilon(x)|$, $\upsilon \in \mathcal{V}$, and $\theta_{\max}(x) := C_o - C_\ell(x-x_1)^- + C_r(x-x_n)^+$ is an upper bound for $\theta(x)$, $\theta \in \Theta_o$. That C_{N} is finite follows from the fact that $\int e^{\theta_{\max}(x)+\varepsilon|x|} \, P_o(dx) < \infty$ for sufficiently small $\varepsilon > 0$. Consequently,

$$DL(\theta,\upsilon) \;\le\; \sqrt{C_{\text{N}} \delta_{\text{Newton}}(\theta)} \quad \text{for all } \upsilon \in \mathcal{V} \cap \mathbb{V}_{D(\theta)}.$$

After these preparations, let us compare θ with $\hat{\theta}$. By concavity of $L(\cdot)$,

$$L(\hat{\theta}) - L(\theta) \;\le\; DL(\theta, \hat{\theta} - \theta).$$

Now we write $\hat{\theta} - \theta = \alpha_0\upsilon_0 + \alpha_1\upsilon_1 + \sum_{\tau\in D} \beta_\tau V_\tau$ with parameters satisfying

$$|\alpha_0| \;=\; \left|\hat{\theta}(x_1) - \theta(x_1)\right| \;\le\; 2C_o,$$
$$|\alpha_1| \;=\; \left|\hat{\theta}'(x_1) - \theta'(x_1)\right| \le\; C_r - C_\ell \quad \text{and}$$

$$\beta_\tau \;=\; \hat\theta'(\tau+) - \hat\theta'(\tau-) - \big(\theta'(\tau+) - \theta'(\tau-)\big) \;\begin{cases} \le\; \hat\theta'(\tau+) - \hat\theta'(\tau-), \\ \ge\; -\big(\theta'(\tau+) - \theta'(\tau-)\big). \end{cases}$$

In particular,

$$\sum_{\tau\in D} \beta_\tau^+ \;\le\; \hat\theta'(x_n) - \hat\theta'(x_1) \;\le\; C_r - C_\ell, \quad \sum_{\tau\in D} \beta_\tau^- \;\le\; \theta'(x_n) - \theta'(x_1) \;\le\; C_r - C_\ell.$$

If $\beta_\tau^- > 0$, then $\tau \in D(\theta)$. And if $\tau \in D(\theta)$, then $V_\tau \in \mathbb{V}_{D(\theta)}$ and $|DL(\theta, V_\tau)| \le \sqrt{C_N \delta_{\text{Newton}}(\theta)}$. For $\tau \in \mathcal{D} \setminus D(\theta)$, we know that $\beta_\tau = \beta_\tau^+$ and

$$DL(\theta, V_\tau) \;=\; DL(\theta, V_{\tau,\theta}) + DL(\theta, \eta_{\tau,\theta}) \;\le\; \delta_{\text{Knot}}(\theta) + (1 + x_n - x_1)\sqrt{C_N \delta_{\text{Newton}}(\theta)}.$$

Here $V_{\tau,\theta} = V_\tau - \eta_{\tau,\theta}$ is the localised kink function with $\eta_{\tau,\theta} \in \mathbb{V}_{D(\theta)}$ as described in Section A.1. The explicit construction of $\eta_{\tau,\theta}$ shows that it is a linear combination of at most two basis functions in $\mathcal{V} \cap \mathbb{V}_{D(\theta)}$ with coefficients whose absolute values sum to less than $1 + x_n - x_1$. This explains the upper bound $(1 + x_n - x_1)\sqrt{C_N \delta_{\text{Newton}}(\theta)}$ for $DL(\theta, \eta_{\tau,\theta})$. Consequently,

$$\begin{aligned} DL(\theta, \hat\theta - \theta) \;&\le \alpha_0 DL(\theta, v_0) + \alpha_1 DL(\theta, v_1) + \sum_{\tau\in D} \beta_\tau^+ DL(\theta, V_\tau)^+ + \sum_{\tau\in D} \beta_\tau^- DL(\theta, V_\tau)^- \\ &\le 2C_o\sqrt{C_N \delta_{\text{Newton}}(\theta)} + (C_r - C_\ell)\sqrt{C_N \delta_{\text{Newton}}(\theta)} \\ &\quad + \;(C_r - C_\ell)\big(\delta_{\text{Knot}}(\theta) + (1 + x_n - x_1)\sqrt{C_N \delta_{\text{Newton}}(\theta)}\big) + (C_r - C_\ell)\sqrt{C_N \delta_{\text{Newton}}(\theta)}, \end{aligned}$$

so the assertion is true with $C_{\text{Newton}} = \big(2C_o + (C_r - C_\ell)(3 + x_n - x_1)\big)\sqrt{C_N}$ and $C_{\text{Knot}} = C_r - C_\ell$. □

Acknowledgements

This work was supported by Swiss National Science Foundation. We owe thanks to Peter McCullagh for drawing our attention to the nonparametric tail inflation model of McCullagh and Polson (2012), to Jon Wellner for the hint to Artin's theorem and Gaussian mixtures, and to Jasha Sommer-Simpson for sharing his MSc thesis. Constructive comments of two referees and editorial support from Sam Allen are gratefully acknowledged.

References

Cule, M., Samworth, R., Stewart, M., 2010. Maximum likelihood estimation of a multi-dimensional log-concave density. J. R. Stat. Soc. Ser. B Stat. Methodol. 72, 545–607. URL: http://dx.doi.org/10.1111/j.1467-9868.2010.00753.x.

Donoho, D., Jin, J., 2004. Higher criticism for detecting sparse heterogeneous mixtures. Ann. Statist. 32, 962–994.

Dümbgen, L., 2017. Optimization Methods – With Applications in Statistics. Lecture notes. University of Bern.

Dümbgen, L., Hüsler, A., Rufibach, K., 2007/2011. Active Set and EM Algorithms for Log-Concave Densities based on Complete and Censored Data. Technical report 61. University of Bern. URL: https://arxiv.org/abs/0707.4643.

Dümbgen, L., Rufibach, K., 2011. logcondens: Computations related to univariate log-concave density estimation. J. Statist. Software 39, 1–28. URL: http://www.jstatsoft.org/v39/i06, doi:10.18637/jss.v039.i06.

Gontscharuk, V., Landwehr, S., Finner, H., 2016. Goodness of fit tests in terms of local levels with special emphasis on higher criticism tests. Bernoulli 22, 1331–1363.

Groeneboom, P., Jongbloed, G., 2014. Nonparametric estimation under shape constraints. volume 38 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, New York. Estimators, algorithms and asymptotics.

Groeneboom, P., Jongbloed, G., Wellner, J.A., 2001. Estimation of a convex function: Characterizations and asymptotic theory. Ann. Statist. 29, 1653–1698. URL: http://dx.doi.org/10.1214/aos/1015345958, doi:10.1214/aos/1015345958.

Groeneboom, P., Jongbloed, G., Wellner, J.A., 2008. The support reduction algorithm for computing nonparametric function estimates in mixture models. Scand. J. Statist. 35, 385–399.

Liu, Y., Wang, Y., 2018. A fast algorithm for univariate log-concave density estimation. Aust. N. Z. J. Stat. 60, 258–275.

Marshall, A.W., Olkin, I., 1979. Inequalities: theory of majorization and its applications. volume 143 of *Mathematics in Science and Engineering*. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London.

McCullagh, P., Polson, N.G., 2012. Tail inflation. Preprint.

McCullagh, P., Polson, N.G., 2017. Statistical sparsity. Biometrika 105, 797–814.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Samworth, R.J., 2018. Recent progress in log-concave density estimation. Statist. Sci. 33, 493–509. URL: https://doi.org/10.1214/18-STS666, doi:10.1214/18-STS666.

Silverman, B.W., 1982. On the estimation of a probability density function by the maximum penalized likelihood method. Ann. Statist. 10, 795–810. URL: http://dx.doi.org/10.1214/aos/1176345872, doi:10.1214/aos/1176345872.

Sommer-Simpson, J., 2019. Convergence of Dümbgen's algorithm for estimation of tail inflation. Master's thesis. Department of Statistics, Univ. of Chicago. Arxiv:1906.04544.

von Neumann, J., 1951. Various techniques used in connection with random digits. J. Res. Nat. Bur. Stand. Appl. Math. Series 3, 36–38.

Walther, G., 2002. Detecting the presence of mixing with multiscale maximum likelihood. J. Amer. Statist. Assoc. 97, 508–513. doi:10.1198/016214502760047032.

Wang, Y., 2018. Computation of the nonparametric maximum likelihood estimate of a univariate log-concave density. WIREs Computational Statistics 11, e1452.

Supplementary material for

Active Set Algorithms for Estimating Shape-Constrained Density Ratios

Lutz Dümbgen, Alexandre Mösching, Christof Strähl

A. Technical details

A.1. Localised kink functions

As mentioned at the end of Section 3.3, working with the kink functions $V_\tau(x) = \xi(x-\tau)^+$ may be computationally inefficient and numerically problematic. For instance, by means of local search we obtain functions θ satisfying (8) approximately, but not perfectly. As a result it may happen that $DL(\theta, V_\tau) > 0$ for some $\tau \in D(\theta)$ although this contradicts (8). Furthermore, the support of V_τ may contain several points $\sigma \in D(\theta)$, so the evaluation of $DL(\theta, V_\tau)$ would involve several integrals of an affine function times a log-affine function with respect to P_o. Hence we propose to replace the simple kink functions V_τ in (9) with localised kink functions $V_{\tau,\theta} = V_\tau - \eta_{\tau,\theta}$ for some $\eta_{\tau,\theta} \in \mathbb{V}_{D(\theta)}$ such that

(i) θ is affine on $\{x \in \mathcal{X} : V_{\tau,\theta}(x) \neq 0\}$,
(ii) $\tau \mapsto V_{\tau,\theta}(x)$ is Lipschitz-continuous with constant 1 for any $x \in \mathcal{X}$,
(iii) $V_{\tau,\theta} \equiv 0$ if $\tau \in D(\theta)$.

Then we redefine the auxiliary function h_θ and replace (9) with

$$h_\theta(\tau) := DL(\theta, V_{\tau,\theta}) \le 0 \quad \text{for all } \tau \in \mathcal{D} \setminus D(\theta). \tag{14}$$

Note that in case of (8), the two requirements (9) and (14) are equivalent, because then $DL(\theta, V_{\tau,\theta}) = DL(\theta, V_\tau)$. We do assume that P_θ is a probability measure, even if (8) is not satisfied perfectly.

To simplify subsequent explicit formulae, let us introduce the following auxiliary functions: For real numbers $a < b$ let

$$j_{10}(x; a, b) := 1_{[a<x\le b]} \frac{b-x}{b-a} \quad \text{and} \quad j_{01}(x; a, b) := 1_{[a<x\le b]} \frac{x-a}{b-a},$$

so $j_{10}(x; a, b) + j_{01}(x; a, b) = 1_{[a<x\le b]}$. In addition we set $j_{01}(x; a, a) := j_{10}(x; a, a) := 0$.

In Setting 1 let $D(\theta) \cup \{x_1, x_n\} = \{\tau_1, \ldots, \tau_m\}$ with $m \ge 2$ points $\tau_1 < \cdots < \tau_m$ in $\{x_1, \ldots, x_n\}$. Then for $\tau_j \le \tau \le \tau_{j+1}$ with $1 \le j < m$,

$$V_{\tau,\theta}(x) := V_\tau(x) - \frac{\tau_{j+1}-\tau}{\tau_{j+1}-\tau_j} V_{\tau_j}(x) - \frac{\tau-\tau_j}{\tau_{j+1}-\tau_j} V_{\tau_{j+1}}(x) = \begin{cases} 0 & \text{for } x \notin [\tau_j, \tau_{j+1}] \\ \dfrac{(x-\tau_j)(\tau_{j+1}-\tau)}{\tau_{j+1}-\tau_j} & \text{for } x \in [\tau_j, \tau] \\ \dfrac{(\tau-\tau_j)(\tau_{j+1}-x)}{\tau_{j+1}-\tau_j} & \text{for } x \in [\tau, \tau_{j+1}] \end{cases}$$

$$= \frac{(\tau-\tau_j)(\tau_{j+1}-\tau)}{\tau_{j+1}-\tau_j} \big(j_{01}(x; \tau_j, \tau) + j_{10}(x; \tau, \tau_{j+1})\big).$$

Figure 10 illustrates these localised kink functions $V_{\tau,\theta}$.

Now we consider Settings 2A-B. If $D(\theta) = \emptyset$, we set $V_{\tau,\theta} := V_\tau = (\cdot - \tau)^+$ and note that $\partial V_\tau(x)/\partial\tau = -1_{[x>\tau]}$ for $x \neq \tau$. Otherwise, let $D(\theta) = \{\tau_1, \ldots, \tau_m\}$ with $m \ge 1$ points $\tau_1 < \cdots < \tau_m < x_n$, where $\tau_1 > x_1$ in Setting 2A and $\tau_1 \in \{0\} \cup (x_1, x_n)$ in Setting 2B. For $\tau \le \tau_1$ we define

$$V_{\tau,\theta}(x) := V_\tau(x) - (\tau_1 - \tau) - V_{\tau_1}(x) = \begin{cases} \tau - \tau_1 & \text{for } x \le \tau \\ x - \tau_1 & \text{for } x \in [\tau, \tau_1] \\ 0 & \text{for } x \ge \tau_1 \end{cases}$$

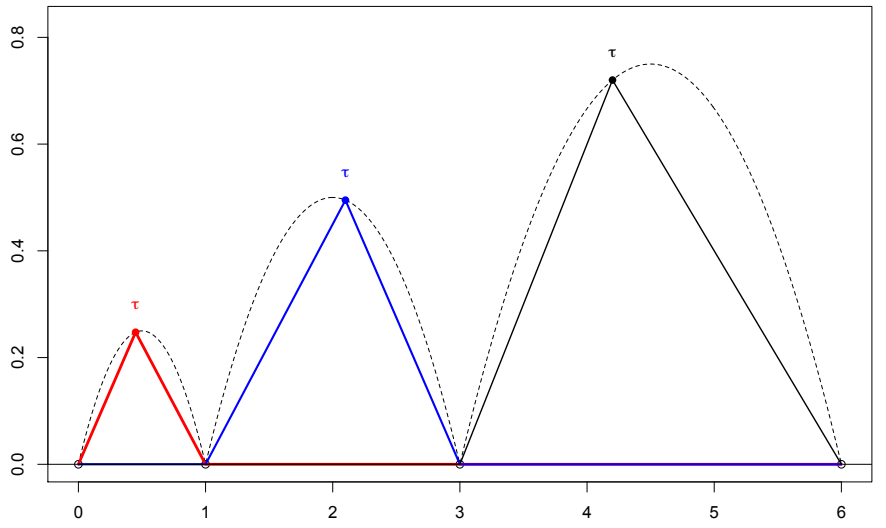

Figure 10: Localised kink functions in Setting 1: For $D(\theta) \cup \{x_1, x_n\} = \{0, 1, 3, 6\}$ one sees $V_{\tau,\theta}$ for three different values of τ.

$$= (\tau - \tau_1)\big(1_{[x \le \tau]} + j_{10}(x; \tau, \tau_1)\big) \tag{15}$$

and note that

$$\partial V_{\tau,\theta}(x)/\partial\tau \;=\; 1_{[x \le \tau]} \quad \text{for } x \neq \tau. \tag{16}$$

For $\tau_j \le \tau \le \tau_{j+1}$ with $1 \le j < m$ we set

$$V_{\tau,\theta}(x) \;:= V_\tau(x) - \frac{\tau_{j+1} - \tau}{\tau_{j+1} - \tau_j} V_{\tau_j}(x) - \frac{\tau - \tau_j}{\tau_{j+1} - \tau_j} V_{\tau_{j+1}}(x) \;=\; \begin{cases} 0 & \text{for } x \notin [\tau_j, \tau_{j+1}] \\ -\dfrac{(x - \tau_j)(\tau_{j+1} - \tau)}{\tau_{j+1} - \tau_j} & \text{for } x \in [\tau_j, \tau] \\ -\dfrac{(\tau - \tau_j)(\tau_{j+1} - x)}{\tau_{j+1} - \tau_j} & \text{for } x \in [\tau, \tau_{j+1}] \end{cases}$$

$$= (\tau - \tau_j)\big(1_{[\tau_j < x \le \tau]} - j_{10}(x; \tau_j, \tau_{j+1}) - j_{01}(x; \tau_j, \tau)\big). \tag{17}$$

and note that

$$\partial V_{\tau,\theta}(x)/\partial\tau \;=\; 1_{[\tau_j < x \le \tau]} - j_{10}(x; \tau_j, \tau_{j+1}) \quad \text{for } x \neq \tau, \tag{18}$$

because $1_{[\tau_j < x \le \tau]}$ and $(\tau - \tau_j) j_{01}(x; \tau_j, \tau) = 1_{[\tau_j < x \le \tau]}(x - \tau_j)$ are locally constant in $\tau \neq x$. Finally, for $\tau > \tau_m$ we define

$$V_{\tau,\theta}(x) \;:= V_\tau(x) - V_{\tau_m}(x) \;=\; \begin{cases} 0 & \text{for } x \le \tau_m \\ \tau_m - x & \text{for } x \in [\tau_m, \tau] \\ \tau_m - \tau & \text{for } x \ge \tau \end{cases}$$

$$= (\tau - \tau_m)\big(-1_{[x > \tau]} - j_{01}(x; \tau_m, \tau)\big) \tag{19}$$

and note that

$$\partial V_{\tau,\theta}(x) \;=\; -1_{[x > \tau]} \quad \text{for } x \neq \tau. \tag{20}$$

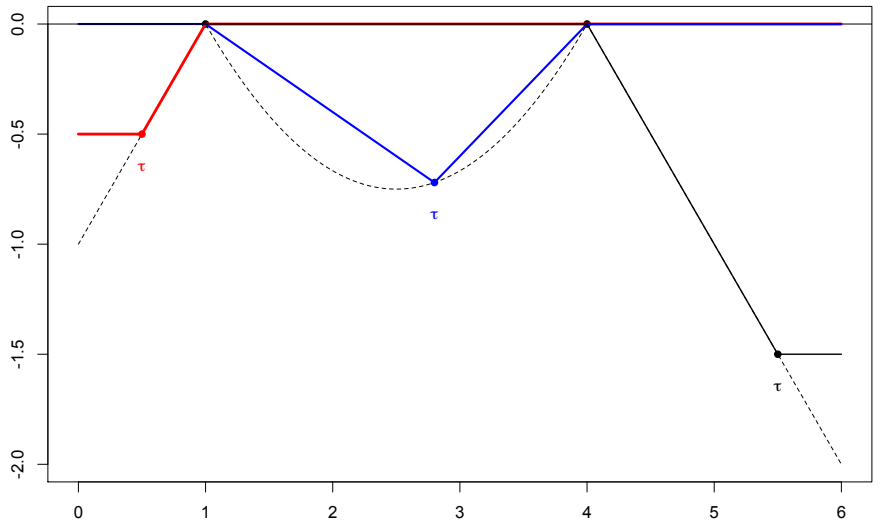

Figure 11: Localised kink functions in Settings 2A-B: For $D(\theta) = \{1, 4\}$ one sees $V_{\tau,\theta}$ for three different values of τ.

Figure 11 illustrates these localised kink functions $V_{\tau,\theta}$.

When searching for local maxima of $h_\theta(\tau) := DL(\theta, V_{\tau,\theta})$ in case of $D(\theta) = \{\tau_1, \dots, \tau_m\}$ as above, one should treat the $m+1$ intervals $(-\infty, \tau_1]$, $[\tau_j, \tau_{j+1}]$ with $1 \le j < m$ and $[\tau_m, \infty)$ separately, because h_θ equals 0 but could be non-differentiable at points in $D(\theta)$. Hence one should look for maximizers of h_θ on the $\tilde{n} - 1$ intervals $[t_i, t_{i+1}]$, $1 \le i < \tilde{n}$, where $t_1 < \cdots < t_{\tilde{n}}$ are the different elements of $\{x_1, \dots, x_n\} \cup \{\tau_1, \dots, \tau_m\}$.

Now we provide explicit formulae for h_θ and its one-sided derivatives. One can easily derive from (15) and (16) that for $\tau < \tau_1$,

$$\begin{aligned} h'_\theta(\tau+) &= (\hat{F} - F_\theta)(\tau) \quad \text{and} \\ h_\theta(\tau) &= (\tau - \tau_1)\Big(h'_\theta(\tau+) + \int j_{10}(x; \tau, \tau_1)\,(\hat{P} - P_\theta)(dx)\Big). \end{aligned}$$

For $1 \le j < m$ and $\tau_j \le \tau < \tau_{j+1}$, equations (17) and (18) lead to

$$\begin{aligned} h'_\theta(\tau+) &= (\hat{P} - P_\theta)((\tau_j, \tau]) - \int j_{10}(x; \tau_j, \tau_{j+1})\,(\hat{P} - P_\theta)(dx) \quad \text{and} \\ h_\theta(\tau) &= (\tau - \tau_j)\Big(h'_\theta(\tau+) - \int j_{01}(x; \tau_j, \tau)\,(\hat{P} - P_\theta)(dx)\Big). \end{aligned}$$

Finally, for $\tau \ge \tau_m$, it follows from (19) and (20) that

$$\begin{aligned} h'_\theta(\tau+) &= (\hat{F} - F_\theta)(\tau) = -(\hat{P} - P_\theta)((\tau, \infty)) \quad \text{and} \\ h_\theta(\tau) &= (\tau - \tau_m)\Big(h'_\theta(\tau+) - \int j_{01}(x; \tau_m, \tau)\,(\hat{P} - P_\theta)(dx)\Big). \end{aligned}$$

The representation of $h_\theta(\tau)$ in terms of $h'_\theta(\tau+)$ is particularly convenient, because h_θ is evaluated only at its local maximizers, i.e. zeros of h'_θ.

A.2. Details for Setting 1

Auxiliary functions. For real numbers $x_1 < x_2$ and a linear function θ on $[x_1, x_2]$,

$$\int_{x_1}^{x_2} e^{\theta(x)}\,dx = (x_2 - x_1) J\big(\theta(x_1), \theta(x_2)\big)$$

with

$$J(r,s) \;:=\; \int_0^1 e^{(1-v)r+vs}\,dv \;=\; \begin{cases} \dfrac{e^s-e^r}{s-r} & \text{if } r\neq s,\\ e^s & \text{if } r=s. \end{cases} \tag{21}$$

In general, for integers $a, b \geq 0$,

$$J_{ab}(r,s) := \frac{\partial^{a+b}}{\partial r^a \partial s^b} J(r,s) \;=\; \int_0^1 (1-v)^a v^b e^{(1-v)r+vs}\,dv.$$

Let $m := (r+s)/2$ and $\delta := (s-r)/2$, so $r = m-\delta$, $s = m+\delta$ and $s-r = 2\delta$. In case of $\delta \neq 0$ we may write

$$J(r,s) \;=\; e^m \sinh(\delta)/\delta.$$

Moreover, with $\Delta := s-r = 2\delta$, partial integration leads to the formulae

$$\begin{aligned}
J_{10}(r,s) \;&=\; e^r \int_0^1 (1-v)e^{\Delta v}\,dv \;=\; e^r\Big(-\frac{1}{\Delta}+\frac{e^\Delta-1}{\Delta^2}\Big) &&= e^m\big(\sinh(\delta)-\delta e^{-\delta}\big)/(2\delta^2),\\
J_{20}(r,s) \;&=\; e^r \int_0^1 (1-v)^2 e^{\Delta v}\,dv \;=\; e^r\Big(-\frac{1}{\Delta}-\frac{2}{\Delta^2}+\frac{2(e^\Delta-1)}{\Delta^3}\Big) &&= e^m\big(\sinh(\delta)/\delta-(1+\delta)e^{-\delta}\big)/(2\delta^2),\\
J_{11}(r,s) \;&=\; e^r \int_0^1 (1-v)ve^{\Delta v}\,dv \;=\; e^r\Big(\frac{e^\Delta+1}{\Delta^2}-\frac{2(e^\Delta-1)}{\Delta^3}\Big) &&= e^m\big(\cosh(\delta)-\sinh(\delta)/\delta\big)/(2\delta^2).
\end{aligned}$$

If $|\delta|$ is close to 0, the formulae above get problematic. Here is a reasonable approximation for small values of $|\delta|$: For integers $a, b \geq 0$ let $B_{ab} := \int_0^1 u^a(1-u)^b\,du = a!b!/(a+b+1)!$, and let U_{ab} be a random variable with distribution $\mathrm{Beta}(a+1, b+1)$, so

$$\begin{aligned}
\mu_{ab} := \mathbb{E}\,U_{ab} \;&=\; \frac{a+1}{a+b+2},\\
\sigma_{ab}^2 := \mathrm{Var}(U_{ab}) \;&=\; \frac{(a+1)(b+1)}{(a+b+2)^2(a+b+3)},\\
\gamma_{ab} := \mathbb{E}\big((U_{ab}-\mu_{ab})^3\big) \;&=\; \frac{2(a+1)(b+1)(b-a)}{(a+b+2)^3(a+b+3)(a+b+4)}.
\end{aligned}$$

Then

$$J_{ab}(r,s) \;=\; B_{ab}\,\mathbb{E}\exp\big(U_{ab}r+(1-U_{ab})s\big) \;=\; B_{ab}\exp\big(\mu_{ab}r+(1-\mu_{ab})s\big)\,\mathbb{E}\exp\big((U_{ab}-\mu_{ab})(r-s)\big),$$

and

$$\log \mathbb{E}\exp\big((U_{ab}-\mu_{ab})(r-s)\big) \;=\; \frac{\sigma_{ab}^2(r-s)^2}{2}+\frac{\gamma_{ab}(r-s)^3}{6}+O(|r-s|^4)$$

as $|r-s| \to 0$. Hence

$$\begin{aligned}
J_{ab}(r,s) \;=\; &\frac{a!b!}{(a+b)!(a+b+1)}\cdot\exp\Big(\frac{(a+1)r+(b+1)s}{a+b+2}\\
&+\frac{(a+1)(b+1)(r-s)^2}{2(a+b+2)^2(a+b+3)}+\frac{(a+1)(b+1)(b-a)(r-s)^3}{3(a+b+2)^3(a+b+3)(a+b+4)}\Big)\cdot\big(1+O(|r-s|^4)\big)
\end{aligned}$$

as $|r-s| \to 0$. Specifically,

$$\begin{aligned}
J(r,s) \;&\approx\; \exp\big((r+s)/2+(r-s)^2/24\big),\\
J_{10}(r,s) \;&\approx\; 2^{-1}\exp\big((2r+s)/3+(r-s)^2/36-(r-s)^3/810\big),\\
J_{20}(r,s) \;&\approx\; 3^{-1}\exp\big((3r+s)/4+3(r-s)^2/160-(r-s)^3/960\big),\\
J_{11}(r,s) \;&\approx\; 6^{-1}\exp\big((r+s)/2+(r-s)^2/40\big).
\end{aligned}$$

Numerical experiments show that the relative error of these approximations is less than 10^{-10} for $|r-s| \leq 0.01$.

Local parametrizations. Let us fix arbitrary points $\tau_1 < \cdots < \tau_m$ in $\{x_1, \ldots, x_n\}$ with $\tau_1 = x_1$ and $\tau_m = x_n$. Any function $\theta : \mathbb{R} \to [-\infty, \infty)$ which is linear on each interval $[\tau_j, \tau_{j+1}]$, $1 \le j < m$, and satisfies $\theta \equiv -\infty$ of $\mathbb{R} \setminus [\tau_1, \tau_m]$ is uniquely determined by the vector $\boldsymbol{\theta} = (\theta_j)_{j=1}^m := (\theta(\tau_j))_{j=1}^m \in \mathbb{R}^m$. Then $L(\theta) = L(\boldsymbol{\tau}, \boldsymbol{\theta})$ with $L(\boldsymbol{\tau}, \cdot) : \mathbb{R}^m \to \mathbb{R}$ given by

$$L(\boldsymbol{\tau}, \boldsymbol{\theta}) \;:=\; \sum_{i=1}^n w_i \theta(x_i) - \sum_{j=1}^{m-1} (\tau_{j+1} - \tau_j) J(\theta_j, \theta_{j+1}) + 1 \;=\; \sum_{j=1}^m \tilde{w}_j \theta_j - \sum_{j=1}^{m-1} (\tau_{j+1} - \tau_j) J(\theta_j, \theta_{j+1}) + 1 \tag{22}$$

with the auxiliary function $J(\cdot, \cdot)$ defined in (21) and the weights

$$\tilde{w}_j \;:=\; 1_{[j=1]} w_1 + \sum_{i=1}^n \Big(1_{[j>1,\, x_i \le \tau_j]} \frac{(x_i - \tau_{j-1})^+}{\tau_j - \tau_{j-1}} + 1_{[j<m,\, x_i > \tau_j]} \frac{(\tau_{j+1} - x_i)^+}{\tau_{j+1} - \tau_j} \Big) w_i.$$

The function $L(\boldsymbol{\tau}, \cdot)$ on $\mathbb{R}^m$ is twice continuously differentiable with negative definite Hessian matrix, see the next paragraph.

Gradient vector and Hessian matrix of $L(\boldsymbol{\tau}, \boldsymbol{\theta})$ in (22). For fixed $\boldsymbol{\tau}$ and as a function of $\boldsymbol{\theta} \in \mathbb{R}^m$, $L(\boldsymbol{\tau}, \boldsymbol{\theta})$ has gradient vector $\nabla L(\boldsymbol{\tau}, \boldsymbol{\theta}) =: \boldsymbol{g}(\boldsymbol{\tau}, \boldsymbol{\theta})$ with components

$$g_j(\boldsymbol{\tau}, \boldsymbol{\theta}) \;=\; \tilde{w}_j - 1_{[j<m]} (\tau_{j+1} - \tau_j) J_{10}(\theta_j, \theta_{j+1}) - 1_{[j>1]} (\tau_j - \tau_{j-1}) J_{10}(\theta_j, \theta_{j-1})$$

and negative Hessian matrix $-D^2 L(\boldsymbol{\tau}, \boldsymbol{\theta}) =: \boldsymbol{H}(\boldsymbol{\tau}, \boldsymbol{\theta})$ with components

$$\begin{aligned} H_{jj}(\boldsymbol{\tau}, \boldsymbol{\theta}) \;&=\; 1_{[j<m]} (\tau_{j+1} - \tau_j) J_{20}(\theta_j, \theta_{j+1}) + 1_{[j>1]} (\tau_j - \tau_{j-1}) J_{20}(\theta_j, \theta_{j-1}), \\ H_{j,j+1}(\boldsymbol{\tau}, \boldsymbol{\theta}) = H_{j+1,j}(\boldsymbol{\tau}, \boldsymbol{\theta}) \;&=\; (\tau_{j+1} - \tau_j) J_{11}(\theta_j, \theta_{j+1}), \\ H_{jk}(\boldsymbol{\tau}, \boldsymbol{\theta}) \;&=\; 0 \quad \text{if } |k - j| \ge 2. \end{aligned}$$

Note also that

$$\boldsymbol{g}(\boldsymbol{\tau}, \boldsymbol{\theta})^\top \boldsymbol{\delta} \;=\; \int_{[x_1, x_n]} \delta(x) \, (\hat{P}(dx) - e^{\theta(x)} \, dx) \quad \text{and} \quad \boldsymbol{\delta}^\top \boldsymbol{H}(\boldsymbol{\tau}, \boldsymbol{\theta}) \boldsymbol{\delta} \;=\; \int_{[x_1, x_n]} \delta(x)^2 e^{\theta(x)} \, dx,$$

the last equality showing positive definiteness of $\boldsymbol{H}(\boldsymbol{\tau}, \boldsymbol{\theta})$.

Evaluating the directional derivative $DL(\theta, V_{\tau,\theta})$. If $\theta \in \mathbb{V}$ with $\{x_1, x_n\} \cup D(\theta)$ having elements $\tau_1 < \cdots < \tau_m$, then for $1 \le j < m$ and $\tau_j \le \tau \le \tau_{j+1}$,

$$\begin{aligned} DL(\theta, V_{\tau,\theta}) \;&=\; \sum_{i=1}^n V_{\tau,\theta}(x_i) w_i - \frac{(\tau - \tau_j)(\tau_{j+1} - \tau)}{\tau_{j+1} - \tau_j} \int_{\tau_j}^{\tau_{j+1}} \big(j_{01}(x; \tau_j, \tau) + j_{10}(x; \tau, \tau_{j+1}) \big) e^{\theta(x)} \, dx \\ &=\; \sum_{i=1}^n V_{\tau,\theta}(x_i) w_i - \frac{(\tau - \tau_j)(\tau_{j+1} - \tau)}{\tau_{j+1} - \tau_j} \big((\tau - \tau_j) J_{10}(\theta_*, \theta_j) + (\tau_{j+1} - \tau) J_{10}(\theta_*, \theta_{j+1}) \big) \end{aligned}$$

with

$$\theta_* \;:=\; \theta(\tau) \;=\; \frac{(\tau_{j+1} - \tau)\theta_j + (\tau - \tau_j)\theta_{j+1}}{\tau_{j+1} - \tau_j}.$$

Activating one constraint. Suppose that $m \ge 3$ in (22). If we activate the constraint at τ_{j_o}, where $1 < j_o < m$, this amounts to replacing $(\tilde{w}_{j_o - 1}, \tilde{w}_{j_o}, \tilde{w}_{j_o + 1})$ with

$$\Big(\tilde{w}_{j_o - 1} + \frac{\tau_{j_o+1} - \tau_{j_o}}{\tau_{j_o+1} - \tau_{j_o-1}} \tilde{w}_{j_o}, \; 0, \; \tilde{w}_{j_o+1} + \frac{\tau_{j_o} - \tau_{j_o-1}}{\tau_{j_o+1} - \tau_{j_o-1}} \tilde{w}_{j_o} \Big)$$

and then removing the j_o-th components of $\boldsymbol{\tau}$ and $(\tilde{w}_j)_{j=1}^m$.

A.3. Details for Setting 2A

We provide explicit formulae for the special case of $P_o = \mathcal{N}(0,1)$ with Lebesgue density ϕ and distribution function Φ.

Auxiliary functions. The subsequent formulae follow from tedious but elementary algebra, the essential ingredients being

$$e^{\theta x}\phi(x) \;=\; e^{\theta^2/2}\phi(x-\theta) \quad \text{for } x, \theta \in \mathbb{R}$$

and

$$\begin{aligned}
\int \phi(z)\,dz &= C + \Phi(z),\\
\int z\phi(z)\,dz &= C - \phi(z),\\
\int z^2\phi(z)\,dz &= C - z\phi(z) + \Phi(z).
\end{aligned}$$

On the one hand, for a fixed number $a \in \mathbb{R}$ let

$$K(\theta_0,\theta_1) = K(\theta_0,\theta_1;a) \;:=\; \int_a^\infty e^{\theta_0+\theta_1(x-a)}\phi(x)\,dx. \tag{23}$$

Then

$$K(\theta_0,\theta_1) \;=\; e^{\theta_0-\theta_1 a+\theta_1^2/2}\,\Phi(\theta_1-a) \;=\; \frac{\partial K(\theta_0,\theta_1)}{\partial\theta_0},$$

and explicit expressions for

$$K_\ell(\theta_0,\theta_1) \;:=\; \frac{\partial^\ell K(\theta_0,\theta_1)}{\partial\theta_1^\ell} \;=\; \int_a^\infty (x-a)^\ell e^{\theta_0+\theta_1(x-a)}\phi(x)\,dx$$

are given by

$$\begin{aligned}
K_1(\theta_0,\theta_1) &= e^{\theta_0-\theta_1 a+\theta_1^2/2}\big((\theta_1-a)\Phi(\theta_1-a)+\phi(\theta_1-a)\big),\\
K_2(\theta_0,\theta_1) &= e^{\theta_0-\theta_1 a+\theta_1^2/2}\Big(\big(1+(\theta_1-a)^2\big)\Phi(\theta_1-a)+(\theta_1-a)\phi(\theta_1-a)\Big).
\end{aligned}$$

Moreover,

$$\int_{-\infty}^a e^{\theta_0+\theta_1(x-a)}\phi(x)\,dx \;=\; K(\theta_0,-\theta_1;-a).$$

On the other hand, for fixed real numbers $a < b$ let

$$J(\theta_0,\theta_1) = J(\theta_0,\theta_1;a,b) \;:=\; \int_a^b \exp\Big(\frac{b-x}{b-a}\,\theta_0 + \frac{x-a}{b-a}\,\theta_1\Big)\phi(x)\,dx. \tag{24}$$

With

$$\tilde\theta_0 \;:=\; \frac{b\theta_0-a\theta_1}{b-a}, \quad \tilde\theta_1 \;:=\; \frac{\theta_1-\theta_0}{b-a} \quad \text{and} \quad \tilde b \;:=\; b-\tilde\theta_1, \quad \tilde a \;:=\; a-\tilde\theta_1$$

we may write

$$J(\theta_0,\theta_1) \;=\; e^{\tilde\theta_0+\tilde\theta_1^2/2}\big(\Phi(\tilde b)-\Phi(\tilde a)\big).$$

Furthermore, explicit expressions for

$$J_{\ell m}(\theta_0,\theta_1) \;:=\; \frac{\partial^{\ell+m} J(\theta_0,\theta_1)}{\partial\theta_0^\ell\,\partial\theta_1^m} \;=\; \int_a^b \frac{(b-x)^\ell (x-a)^m}{(b-a)^{\ell+m}} \exp\Big(\frac{b-x}{b-a}\,\theta_0 + \frac{x-a}{b-a}\,\theta_1\Big)\phi(x)\,dx$$

for $\ell, m \in \{0,1,2\}$ with $1 \le \ell + m \le 2$ are given by

$$\begin{aligned}
J_{10}(\theta_0,\theta_1) &= e^{\tilde\theta_0+\tilde\theta_1^2/2}\,\frac{\tilde b\big(\Phi(\tilde b)-\Phi(\tilde a)\big)+\phi(\tilde b)-\phi(\tilde a)}{b-a},\\
J_{01}(\theta_0,\theta_1) &= J_{10}(\theta_1,\theta_0;-b,-a),\\
J_{20}(\theta_0,\theta_1) &= e^{\tilde\theta_0+\tilde\theta_1^2/2}\,\frac{(1+\tilde b^2)\big(\Phi(\tilde b)-\Phi(\tilde a)\big)+(\tilde a-2\tilde b)\phi(\tilde a)+\tilde b\phi(\tilde b)}{(b-a)^2},\\
J_{11}(\theta_0,\theta_1) &= e^{\tilde\theta_0+\tilde\theta_1^2/2}\,\frac{-(1+\tilde a\tilde b)\big(\Phi(\tilde b)-\Phi(\tilde a)\big)+\tilde b\phi(\tilde a)-\tilde a\phi(\tilde b)}{(b-a)^2},\\
J_{02}(\theta_0,\theta_1) &= e^{\tilde\theta_0+\tilde\theta_1^2/2}\,\frac{(1+\tilde a^2)\big(\Phi(\tilde b)-\Phi(\tilde a)\big)+(2\tilde a-\tilde b)\phi(\tilde b)-\tilde a\phi(\tilde a)}{(b-a)^2}.
\end{aligned}$$

In case of $\tilde a > 0$, the right hand side of the equation

$$\Phi(\tilde b)-\Phi(\tilde a) \;=\; \Phi(-\tilde a)-\Phi(-\tilde b)$$

is numerically more accurate than its left-hand side. In connection with $J(\theta_0,\theta_1)$ we also use the the lower bound

$$\log(\Phi(\tilde b)-\Phi(\tilde a)) \;=\; -\frac{\tilde m^2}{2}+\log\int_{-\tilde d}^{\tilde d}\exp(\tilde m z)\phi(z)\,dz \;\ge\; -\frac{\tilde m^2}{2}+\log\big(\Phi(\tilde d)-\Phi(-\tilde d)\big)$$

with $\tilde m := (\tilde a+\tilde b)/2$ and $\tilde d := (\tilde b-\tilde a)/2$. The bound follows from $\exp(\tilde m z) \ge 1+\tilde m z$.

Local parametrizations. Let us fix any vector $\boldsymbol\tau$ with $m \ge 1$ components $\tau_1 < \cdots < \tau_m$ in (x_1, x_n). Any function θ which is linear on the intervals $\mathcal{X}_0, \mathcal{X}_1, \ldots, \mathcal{X}_m$ specified in Lemma 2.6 is uniquely determined by the vector

$$\boldsymbol\theta = (\theta_j)_{j=0}^{m+1} \;:=\; \big(\theta'(\tau_1-), \theta(\tau_1), \ldots, \theta(\tau_m), \theta'(\tau_m+)\big)^\top \;\in\; \mathbb{R}^{m+2}.$$

Then $L(\theta)$ is given by

$$\begin{aligned}
L(\boldsymbol\tau,\boldsymbol\theta) \;&:=\; \sum_{i=1}^n w_i\theta(x_i) - \int_{\mathcal{X}_0} e^{\theta(x)}\,P_o(dx) - \sum_{j=1}^m \int_{\mathcal{X}_j} e^{\theta(x)}\,P_o(dx) + 1\\
&=\; \sum_{j=0}^{m+1} \tilde w_j\theta_j - K(\theta_1,-\theta_0;-\tau_1) - \sum_{1\le j<m} J(\theta_j,\theta_{j+1};\tau_j,\tau_{j+1}) - K(\theta_m,\theta_{m+1};\tau_m) + 1. \qquad (25)
\end{aligned}$$

with the auxiliary functions $K(\cdot,\cdot;\cdot)$ and $J(\cdot,\cdot;\cdot,\cdot)$ introduced in (23) and (24) and the 'weights'

$$\begin{aligned}
\tilde w_0 \;&:=\; -\sum_{i=1}^n (\tau_1 - x_i)^+ w_i,\\
\tilde w_1 \;&:=\; \sum_{i=1}^n \min\Big(1, \frac{(\tau_2-x_i)^+}{\tau_2-\tau_1}\Big)\,w_i,\\
\tilde w_j \;&:=\; \sum_{i=1}^n \Big(1_{[x_i\le\tau_j]}\,\frac{(x_i-\tau_{j-1})^+}{\tau_j-\tau_{j-1}} + 1_{[x_i>\tau_j]}\,\frac{(\tau_{j+1}-x_i)^+}{\tau_{j+1}-\tau_j}\Big)w_i \quad \text{for } 1<j<m,\\
\tilde w_m \;&:=\; \sum_{i=1}^n \min\Big(1, \frac{(x_i-\tau_{m-1})^+}{\tau_m-\tau_{m-1}}\Big)\,w_i,
\end{aligned}$$

$$\tilde{w}_{m+1} \;:=\; \sum_{i=1}^{n} (x_i - \tau_m)^+ \, w_i.$$

In case of $m = 1$, the weight $\tilde{w}_1$ is just given by $\tilde{w}_1 = 1$.

The function $L(\boldsymbol{\tau}, \cdot) : \mathbb{R}^{m+2} \to \mathbb{R}$ is twice continuously differentiable with negative definite Hessian matrix, see the next paragraph.

Gradient vector and Hessian matrix for $L(\boldsymbol{\tau}, \cdot)$ in (25). In case of $m \geq 2$, the gradient $\boldsymbol{g}(\boldsymbol{\tau}, \boldsymbol{\theta}) = \big(g_j(\boldsymbol{\tau}, \boldsymbol{\theta})\big)_{j=0}^{m+1}$ of $L(\boldsymbol{\tau}, \cdot)$ equals

$$g_j(\boldsymbol{\tau}, \boldsymbol{\theta}) \;=\; \tilde{w}_j - \begin{cases} -K_1(\theta_1, -\theta_0; -\tau_1) & \text{if } j = 0, \\ K(\theta_1, -\theta_0; -\tau_1) + J_{10}(\theta_1, \theta_2; \tau_1, \tau_2) & \text{if } j = 1, \\ J_{01}(\theta_{j-1}, \theta_j; \tau_{j-1}, \tau_j) + J_{10}(\theta_j, \theta_{j+1}; \tau_j, \tau_{j+1}) & \text{if } 2 < j < m, \\ J_{01}(\theta_{m-1}, \theta_m; \tau_{m-1}, \tau_m) + K(\theta_m, \theta_{m+1}; \tau_m) & \text{if } j = m, \\ K_1(\theta_m, \theta_{m+1}; \tau_m) & \text{if } j = m+1, \end{cases}$$

while its negative Hessian matrix $\boldsymbol{H}(\boldsymbol{\tau}, \boldsymbol{\theta}) = \big(H_{jk}(\boldsymbol{\tau}, \boldsymbol{\theta})\big)_{j,k=0}^{m+1}$ is given by

$$\begin{aligned} H_{00}(\boldsymbol{\tau}, \boldsymbol{\theta}) \;&=\; K_2(\theta_1, -\theta_0; -\tau_1), \\ H_{01}(\boldsymbol{\tau}, \boldsymbol{\theta}) = H_{10}(\boldsymbol{\tau}, \boldsymbol{\theta}) \;&=\; -K_1(\theta_1, -\theta_0; -\tau_1), \\ H_{11}(\boldsymbol{\tau}, \boldsymbol{\theta}) \;&=\; K(\theta_1, -\theta_0; -\tau_1) + J_{20}(\theta_1, \theta_2; \tau_1, \tau_2), \\ H_{j,j+1}(\boldsymbol{\tau}, \boldsymbol{\theta}) = H_{j+1,j}(\boldsymbol{\tau}, \boldsymbol{\theta}) \;&=\; J_{11}(\theta_j, \theta_{j+1}; \tau_j, \tau_{j+1}) \text{ for } 1 \leq j < m, \\ H_{jj}(\boldsymbol{\tau}, \boldsymbol{\theta}) \;&=\; J_{02}(\theta_{j-1}, \theta_j; \tau_{j-1}, \tau_j) + J_{20}(\theta_j, \theta_{j+1}; \tau_j, \tau_{j+1}) \text{ for } 1 < j < m, \\ H_{mm}(\boldsymbol{\tau}, \boldsymbol{\theta}) \;&=\; J_{02}(\theta_{m-1}, \theta_m; \tau_{m-1}, \tau_m) + K(\theta_m, \theta_{m+1}; \tau_m) \\ H_{m,m+1}(\boldsymbol{\tau}, \boldsymbol{\theta}) = H_{m+1,m}(\boldsymbol{\tau}, \boldsymbol{\theta}) \;&=\; K_1(\theta_m, \theta_{m+1}; \tau_m), \\ H_{m+1,m+1}(\boldsymbol{\tau}, \boldsymbol{\theta}) \;&=\; K_2(\theta_m, \theta_{m+1}; \tau_m), \\ H_{jk}(\boldsymbol{\tau}, \boldsymbol{\theta}) \;&=\; 0 \text{ if } |j - k| \geq 2. \end{aligned}$$

In case of $m = 1$ we get the simplified formulae

$$L(\boldsymbol{\tau}, \boldsymbol{\theta}) \;=\; \sum_{j=0}^{2} \tilde{w}_j \theta_j - K(\theta_1, -\theta_0; -\tau_1) - K(\theta_1, \theta_2; \tau_1) + 1,$$

$$g_j(\boldsymbol{\tau}, \boldsymbol{\theta}) \;=\; \tilde{w}_j - \begin{cases} -K_1(\theta_1, -\theta_0; -\tau_1) & \text{if } j = 0, \\ K(\theta_1, -\theta_0; -\tau_1) + K(\theta_1, \theta_2; \tau_1) & \text{if } j = 1, \\ K_1(\theta_1, \theta_2; \tau_1) & \text{if } j = 2, \end{cases}$$

and

$$\begin{aligned} H_{00}(\boldsymbol{\tau}, \boldsymbol{\theta}) \;&=\; K_2(\theta_1, -\theta_0; -\tau_1), \\ H_{01}(\boldsymbol{\tau}, \boldsymbol{\theta}) = H_{10}(\boldsymbol{\tau}, \boldsymbol{\theta}) \;&=\; -K_1(\theta_1, -\theta_0; -\tau_1), \\ H_{11}(\boldsymbol{\tau}, \boldsymbol{\theta}) \;&=\; K(\theta_1, -\theta_0; -\tau_1) + K(\theta_1, \theta_2; \tau_2) \\ H_{12}(\boldsymbol{\tau}, \boldsymbol{\theta}) = H_{21}(\boldsymbol{\tau}, \boldsymbol{\theta}) \;&=\; K_1(\theta_1, \theta_2; \tau_1), \\ H_{22}(\boldsymbol{\tau}, \boldsymbol{\theta}) \;&=\; K_2(\theta_1, \theta_2; \tau_1). \end{aligned}$$

Evaluating $h_\theta(\tau) := DL(\theta, V_{\tau,\theta})$ and $h'_\theta(\tau+)$. Suppose first that $\theta(x) = \hat{\mu}x - \hat{\mu}^2/2$, so $P_\theta = \mathcal{N}(\hat{\mu}, 1)$ and $D(\theta) = \emptyset$. Then one can show that

$$h'_\theta(\tau+) \;=\; \hat{F}(\tau) - \Phi(\tau - \hat{\mu}),$$

$$h_\theta(\tau) \;=\; \tau h'_\theta(\tau+) - \int_{(-\infty,\tau]} x\,\hat{P}(dx) + \hat{\mu}\Phi(\tau-\hat{\mu}) - \phi(\tau-\hat{\mu}).$$

Now suppose that θ is given by a vector $\boldsymbol{\tau}$ of $m \geq 1$ points $\tau_1 < \cdots < \tau_m$ and a vector $\boldsymbol{\theta} = (\theta_j)_{j=0}^{m+1}$ as in (25). Then for $\tau < \tau_1$,

$$\begin{aligned} h'_\theta(\tau+) \;&=\; \hat{F}(\tau) - K(\theta_*, -\theta_0; -\tau), \\ h_\theta(\tau) \;&=\; (\tau-\tau_1)\big(h'_\theta(\tau+) - J_{10}(\theta_*, \theta_1; \tau, \tau_1)\big) - \int 1_{[\tau<x\leq\tau_1]}(\tau_1 - x)\hat{P}(dx), \end{aligned}$$

where $\theta_* := \theta(\tau) = \theta_1 + (\tau-\tau_1)\theta_0$. For $1 \leq j < m$ and $\tau \in [\tau_j, \tau_{j+1})$,

$$\begin{aligned} h'_\theta(\tau+) \;&=\; \hat{P}((\tau_j,\tau]) - J(\theta_j, \theta_*; \tau_j, \tau) - \int\; j_{10}(x; \tau_j, \tau_{j+1})\,\hat{P}(dx) + J_{10}(\theta_j, \theta_{j+1}; \tau_j, \tau_{j+1}), \\ h_\theta(\tau) \;&=\; (\tau-\tau_j)\big(h'_\theta(\tau+) + J_{01}(\theta_j, \theta_*; \tau_j, \tau)\big) - \int 1_{[\tau_j<x\leq\tau]}(x-\tau_j)\,\hat{P}(dx), \end{aligned}$$

where $\theta_* := \theta(\tau) = (\tau_{j+1}-\tau_j)^{-1}\big((\tau_{j+1}-\tau)\theta_j + (\tau-\tau_j)\theta_{j+1}\big) = \theta_j + (\tau-\tau_j)\theta'_j$. Finally, for $\tau > \tau_m$,

$$\begin{aligned} h'_\theta(\tau+) \;&=\; K(\theta_*, \theta_{m+1}; \tau) - \hat{P}((\tau,\infty)), \\ h_\theta(\tau) \;&=\; (\tau-\tau_m)\big(h'_\theta(\tau+) + J_{01}(\theta_m, \theta_*; \tau_m, \tau)\big) - \int 1_{[\tau_m<x\leq\tau]}(x-\tau_m)\hat{P}(dx), \end{aligned}$$

where $\theta_* := \theta_m + (\tau-\tau_m)\theta_{m+1}$.

If τ is restricted to some interval I not containing any observations x_i or knots τ_j, the latter expressions for $h'_\theta(\tau+)$ are constant in τ except for one term $K(\theta_*, -\theta_0; -\tau)$, $J(\theta_j, \theta_*; \tau_j, \tau)$ or $K(\theta_*, \theta_{m+1}; \tau)$. Hence finding τ such that $h'_\theta(\tau+) = 0$ leads to equations of the following type: For given real numbers $\theta_0, \theta_1, \tau_0$ and c, find $\tau \in \mathbb{R}$ such that

$$K\big(\theta_0 + \theta_1(\tau-\tau_0), \pm\theta_1; \pm\tau\big) \;=\; c, \tag{26}$$

$$J\big(\theta_0, \theta_0 + \theta_1(\tau-\tau_0); \tau_0, \tau\big) \;=\; c, \tag{27}$$

and check whether $\tau \in I$. Since $K\big(\theta_0 + \theta_1(\tau-\tau_0), \pm\theta_1; \pm\tau\big)$ equals $e^{\theta_0-\theta_1\tau_0+\theta_1^2/2}\Phi(\mp(\tau-\theta_1))$, the unique solution of (26) is given by

$$\tau \;=\; \theta_1 \mp \Phi^{-1}(e^{-\theta_0+\theta_1\tau_0-\theta_1^2/2}c),$$

provided that $c > 0$ and $ce^{-\theta_0+\theta_1\tau_0-\theta_1^2/2} < 1$; otherwise no solution exists. Likewise, since $J\big(\theta_0, \theta_0+\theta_1(\tau-\tau_0); \tau_0, \tau\big)$ equals $e^{\theta_0-\theta_1\tau_0+\theta_1^2/2}\big(\Phi(\tau-\theta_1) - \Phi(\tau_0-\theta_1)\big)$, the unique solution of (27) is given by

$$\tau \;=\; \theta_1 + \Phi^{-1}\big(\Phi(\tau_0-\theta_1) + e^{-\theta_0+\theta_1\tau_0-\theta_1^2/2}c\big),$$

provided that $0 < \Phi(\tau_0-\theta_1) + ce^{-\theta_0+\theta_1\tau_0-\theta_1^2/2} < 1$; otherwise no solution exists.

Activating one constraint. Suppose that $m \geq 2$ in (25). If even $m \geq 3$, and if we activate the constraint at τ_{j_o}, where $1 < j_o < m$, the update of $\boldsymbol{\tau}$ and $(\tilde{w}_j)_{j=0}^{m+1}$ is essentially the same as in Setting 1. If we activate the constraint at τ_1, this amounts to replacing (τ_1, τ_2) and $(\tilde{w}_0, \tilde{w}_1, \tilde{w}_{j_o+1})$ with

$$(\tau_2) \quad \text{and} \quad \big(\tilde{w}_0 - (\tau_2-\tau_1)\tilde{w}_1,\; \tilde{w}_1 + \tilde{w}_2\big),$$

respectively. Similary, activating the constraint at τ_m amounts to replacing (τ_{m-1}, τ_m) with

$$(\tau_{m-1}) \quad \text{and} \quad \big(\tilde{w}_{m-1} + \tilde{w}_m,\; \tilde{w}_{m+1} + (\tau_m-\tau_{m-1})\tilde{w}_m\big),$$

respectively.

A.4. Details for Setting 2B

We provide explicit formulae for the special case of P_o being a gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\beta = 1$, i.e. P_o has density

$$p_o(x) \;=\; \Gamma(\alpha)^{-1} x^{\alpha-1} e^{-x}, \qquad x > 0.$$

Note that the case of a gamma distribution with rate parameter $\beta \neq 1$ may be reduced to the case $\beta = 1$ by multiplying all observations with β, then estimating the function θ by $\hat{\theta}_{\text{temp}}$ and finally setting $\hat{\theta}(x) := \hat{\theta}_{\text{temp}}(x/\beta)$.

Auxiliary functions. For $s > 0$, the c.d.f. of a gamma distribution with shape s and rate 1 is the function $G_s : [0, \infty] \to [0, 1]$ defined by

$$G_s(x) \;:=\; \Gamma(s)^{-1} \int_0^x z^{s-1} e^{-z}\, dz,$$

and, for $0 \leq a < b \leq \infty$, we define the partial integral

$$G_s(a, b) \;:=\; \Gamma(s)^{-1} \int_a^b z^{s-1} e^{-z}\, dz \;=\; G_s(b) - G_s(a).$$

On the one hand, for a fixed number $c \in \mathbb{R}$ let

$$K(\theta_0, \theta_1) \;=\; K(\theta_0, \theta_1; c) \;:=\; \int_c^\infty e^{\theta_0 + \theta_1 (x-c)} p_o(x)\, dx.$$

This is equal to ∞ in case of $\theta_1 \geq 1$. Otherwise, when $\theta_1 < 1$, let $\tilde{c} := (1 - \theta_1)c$. Then

$$K(\theta_0, \theta_1) \;=\; \frac{e^{\theta_0 - \theta_1 c}}{(1-\theta_1)^\alpha}\, G_\alpha(\tilde{c}, \infty) \;=\; \frac{\partial K(\theta_0, \theta_1)}{\partial \theta_0},$$

and explicit expressions for

$$K_\ell(\theta_0, \theta_1) \;:=\; \frac{\partial^\ell K(\theta_0, \theta_1)}{\partial \theta_1^\ell} \;=\; \int_c^\infty (x-c)^\ell e^{\theta_0 + \theta_1 (x-c)} p_o(x)\, dx$$

are given by

$$\begin{aligned}
K_1(\theta_0, \theta_1) \;&=\; \frac{e^{\theta_0 - \theta_1 c}}{(1-\theta_1)^{\alpha+1}} \big(\alpha G_{\alpha+1}(\tilde{c}, \infty) - \tilde{c} G_\alpha(\tilde{c}, \infty) \big), \\
K_2(\theta_0, \theta_1) \;&=\; \frac{e^{\theta_0 - \theta_1 c}}{(1-\theta_1)^{\alpha+2}} \big(\alpha(\alpha+1) G_{\alpha+2}(\tilde{c}, \infty) - 2\alpha\tilde{a} G_{\alpha+1}(\tilde{c}, \infty) + \tilde{c}^2 G_\alpha(\tilde{c}, \infty) \big).
\end{aligned}$$

On the other hand, for fixed numbers $0 \leq a < b < \infty$ let

$$J(\theta_0, \theta_1) = J(\theta_0, \theta_1; a, b) \;=\; \int_a^b \exp\Big(\frac{b-x}{b-a}\theta_0 + \frac{x-a}{b-a}\theta_1 \Big) p_o(x)\, dx \;=\; \frac{e^{\tilde{\theta}_0}}{\Gamma(\alpha)} \int_a^b e^{(\tilde{\theta}_1 - 1)x} x^{\alpha-1}\, dx,$$

where

$$\tilde{\theta}_0 \;:=\; \frac{b\theta_0 - a\theta_1}{b-a} \quad \text{and} \quad \tilde{\theta}_1 \;:=\; \frac{\theta_1 - \theta_0}{b-a}.$$

With $\tilde{a} := (1 - \tilde{\theta}_1)a$ and $\tilde{b} := (1 - \tilde{\theta}_1)b$ we may write

$$J(\theta_0, \theta_1) \;=\; \begin{cases} \dfrac{e^{\tilde{\theta}_0} G_\alpha(\tilde{a}, \tilde{b})}{(1-\tilde{\theta}_1)^\alpha} & \text{if } \tilde{\theta} < 1, \\ \dfrac{e^{\tilde{\theta}_0}(b^\alpha - a^\alpha)}{\Gamma(\alpha+1)} & \text{if } \tilde{\theta} = 1. \end{cases}$$

Note that in our specific applications the slope parameter $\tilde\theta_1$ corresponds to the difference ratio $\big(\theta(b)-\theta(a)\big)/(b-a)$ of a function $\theta \in \mathbb{V}$. Thus it will be strictly smaller than 1 as soon as $\theta \in \Theta$ and $L(\theta) > -\infty$. During a Newton step the latter conditions may be violated temporarily, so in case of $\tilde\theta_1 > 1$ we use the simple bound

$$J(\theta_0,\theta_1) \;\le\; \frac{e^{\tilde\theta_0+(\tilde\theta_1-1)b}(b^\alpha - a^\alpha)}{\Gamma(\alpha+1)}.$$

In case of $\tilde\theta_1 < 1$, explicit expressions for

$$J_{\ell m}(\theta_0,\theta_1) \;:=\; \frac{\partial^{\ell+m} J(\theta_0,\theta_1)}{\partial\theta_0^\ell \partial\theta_1^m} \;=\; \int_a^b \frac{(b-x)^\ell (x-a)^m}{(b-a)^{\ell+m}} \exp\Big(\frac{b-x}{b-a}\theta_0 + \frac{x-a}{b-a}\theta_1\Big) p_o(x)\,dx$$

are given by

$$\begin{aligned}
J_{10}(\theta_0,\theta_1) &= \frac{e^{\tilde\theta_0}}{(1-\tilde\theta_1)^{\alpha+1}} \frac{\tilde b G_\alpha(\tilde a,\tilde b) - \alpha G_{\alpha+1}(\tilde a,\tilde b)}{b-a},\\
J_{01}(\theta_0,\theta_1) &= \frac{e^{\tilde\theta_0}}{(1-\tilde\theta_1)^{\alpha+1}} \frac{-\tilde a G_\alpha(\tilde a,\tilde b) + \alpha G_{\alpha+1}(\tilde a,\tilde b)}{b-a},\\
J_{20}(\theta_0,\theta_1) &= \frac{e^{\tilde\theta_0}}{(1-\tilde\theta_1)^{\alpha+2}} \frac{\tilde b^2 G_\alpha(\tilde a,\tilde b) - 2\alpha\tilde b G_{\alpha+1}(\tilde a,\tilde b) + \alpha(\alpha+1) G_{\alpha+2}(\tilde a,\tilde b)}{(b-a)^2},\\
J_{11}(\theta_0,\theta_1) &= \frac{e^{\tilde\theta_0}}{(1-\tilde\theta_1)^{\alpha+2}} \frac{-\tilde a\tilde b G_\alpha(\tilde a,\tilde b) + \alpha(\tilde a+\tilde b) G_{\alpha+1}(\tilde a,\tilde b) - \alpha(\alpha+1) G_{\alpha+2}(\tilde a,\tilde b)}{(b-a)^2},\\
J_{02}(\theta_0,\theta_1) &= \frac{e^{\tilde\theta_0}}{(1-\tilde\theta_1)^{\alpha+2}} \frac{\tilde a^2 G_\alpha(\tilde a,\tilde b) - 2\alpha\tilde a G_{\alpha+1}(\tilde a,\tilde b) + \alpha(\alpha+1) G_{\alpha+2}(\tilde a,\tilde b)}{(b-a)^2}.
\end{aligned}$$

Local parametrizations. Let us fix an arbitrary vector $\boldsymbol\tau$ with $m \ge 1$ components $0 \le \tau_1 < \cdots < \tau_m < x_n$. Any function $\theta : [0,\infty) \to \mathbb{R}$ which is constant on $[0,\tau_1]$ and linear on the intervals $\mathcal{X}_1, \dots, \mathcal{X}_m$ specified in Lemma 2.7 is uniquely determined by the vector $\boldsymbol\theta = (\theta_j)_{j=1}^{m+1} := \big(\theta(\tau_1), \dots, \theta(\tau_m), \theta'(\tau_m+)\big)^\top \in \mathbb{R}^{m+1}$. Then $L(\theta)$ is given by

$$\begin{aligned}
L(\boldsymbol\tau,\boldsymbol\theta) \;&:=\; \sum_{i=1}^n w_i\theta(x_i) - e^{\theta_1}F_0(\tau_1) - \sum_{j=1}^m \int_{\mathcal{X}_j} e^{\theta_j+\theta_j'(x-\tau_j)}\,P_o(dx) + 1\\
&=\; \sum_{j=1}^{m+1} \tilde w_j\theta_j - e^{\theta_1}G_\alpha(\tau_1) - \sum_{1\le j<m} J(\theta_j,\theta_{j+1};\tau_j,\tau_{j+1}) - K(\theta_m,\theta_{m+1};\tau_m) + 1
\end{aligned} \tag{28}$$

with the auxiliary functions $G_\alpha(\cdot)$, $J(\cdot,\cdot;\cdot,\cdot)$ and $K(\cdot,\cdot;\cdot)$ introduced before and the weights

$$\begin{aligned}
\tilde w_1 \;&:=\; \sum_{i=1}^n \min\Big(1, \frac{(\tau_2-x_i)^+}{\tau_2-\tau_1}\Big)\, w_i,\\
\tilde w_j \;&:=\; \sum_{i=1}^n \Big(1_{[x_i\le\tau_j]}\frac{(x_i-\tau_{j-1})^+}{\tau_j-\tau_{j-1}} + 1_{[x_i>\tau_j]}\frac{(\tau_{j+1}-x_i)^+}{\tau_{j+1}-\tau_j}\Big) w_i \quad \text{for } 1<j<m,\\
\tilde w_m \;&:=\; \sum_{i=1}^n \min\Big(1, \frac{(x_i-\tau_{m-1})^+}{\tau_m-\tau_{m-1}}\Big) w_i,\\
\tilde w_{m+1} \;&:=\; \sum_{i=1}^n (x_i-\tau_m)^+\, w_i.
\end{aligned}$$

In case of $m = 1$, the weight $\tilde w_1$ is just given by $\tilde w_1 = 1$.

The function $L(\boldsymbol\tau,\cdot) : \mathbb{R}^{m+1} \to [-\infty,\infty)$ is continuous and concave. On the open set $\big\{\boldsymbol\theta \in \mathbb{R}^{m+1} : L(\boldsymbol\tau,\boldsymbol\theta) > -\infty\big\} = \big\{\boldsymbol\theta \in \mathbb{R}^{m+1} : \theta_{m+1} < 1\big\}$ it is twice continuously differentiable with negative definite Hessian matrix, see the next paragraph.

Gradient vector and Hessian matrix for $L(\boldsymbol{\tau},\cdot)$ in (28). Let $\theta_{m+1} < 1$. In case of $m \geq 2$, the gradient $\boldsymbol{g}(\boldsymbol{\tau},\boldsymbol{\theta}) = \big(g_j(\boldsymbol{\tau},\boldsymbol{\theta})\big)_{j=1}^{m+1}$ of $L(\boldsymbol{\tau},\cdot)$ equals

$$g_j(\boldsymbol{\tau},\boldsymbol{\theta}) \;=\; \tilde{w}_j - \begin{cases} e^{\theta_1} G_\alpha(\tau_1) + J_{10}(\theta_1,\theta_2;\tau_1,\tau_2) & \text{if } j = 1, \\ J_{01}(\theta_{j-1},\theta_j;\tau_{j-1},\tau_j) + J_{10}(\theta_j,\theta_{j+1};\tau_j,\tau_{j+1}) & \text{if } 1 < j < m, \\ J_{01}(\theta_{m-1},\theta_m;\tau_{m-1},\tau_m) + K(\theta_m,\theta_{m+1};\tau_m) & \text{if } j = m, \\ K_1(\theta_m,\theta_{m+1};\tau_m) & \text{if } j = m+1, \end{cases}$$

while its negative Hessian matrix $\boldsymbol{H}(\boldsymbol{\tau},\boldsymbol{\theta}) = \big(H_{jk}(\boldsymbol{\tau},\boldsymbol{\theta})\big)_{j,k=1}^{m+1}$ is given by

$$\begin{aligned} H_{11}(\boldsymbol{\tau},\boldsymbol{\theta}) &= e^{\theta_1} G_\alpha(\tau_1) + J_{20}(\theta_1,\theta_2;\tau_1,\tau_2), \\ H_{j,j+1}(\boldsymbol{\tau},\boldsymbol{\theta}) = H_{j+1,j}(\boldsymbol{\tau},\boldsymbol{\theta}) &= J_{11}(\theta_j,\theta_{j+1};\tau_j,\tau_{j+1}) \;\text{ for } 1 \leq j < m, \\ H_{jj}(\boldsymbol{\tau},\boldsymbol{\theta}) &= J_{02}(\theta_{j-1},\theta_j;\tau_{j-1},\tau_j) + J_{20}(\theta_j,\theta_{j+1};\tau_j,\tau_{j+1}) \;\text{ for } 1 < j < m, \\ H_{mm}(\boldsymbol{\tau},\boldsymbol{\theta}) &= J_{02}(\theta_{m-1},\theta_m;\tau_{m-1},\tau_m) + K(\theta_m,\theta_{m+1};\tau_m), \\ H_{m,m+1}(\boldsymbol{\tau},\boldsymbol{\theta}) = H_{m+1,m}(\boldsymbol{\tau},\boldsymbol{\theta}) &= K_1(\theta_m,\theta_{m+1};\tau_m), \\ H_{m+1,m+1}(\boldsymbol{\tau},\boldsymbol{\theta}) &= K_2(\theta_m,\theta_{m+1};\tau_m), \\ H_{jk}(\boldsymbol{\tau},\boldsymbol{\theta}) &= 0 \;\text{ if } |j-k| > 1. \end{aligned}$$

In case of $m = 1$ we get the simplified formulae

$$L(\boldsymbol{\tau},\boldsymbol{\theta}) \;=\; \sum_{j=1}^{2} \tilde{w}_j \theta_j - e^{\theta_1} G_\alpha(\tau_1) - K(\theta_1,\theta_2;\tau_1) + 1,$$

$$g_j(\boldsymbol{\tau},\boldsymbol{\theta}) \;=\; \tilde{w}_j - \begin{cases} e^{\theta_1} G_\alpha(\tau_1) + K(\theta_1,\theta_2;\tau_1) & \text{if } j = 1, \\ K_1(\theta_1,\theta_2;\tau_1) & \text{if } j = 2, \end{cases}$$

and

$$\begin{aligned} H_{11}(\boldsymbol{\tau},\boldsymbol{\theta}) &= e^{\theta_1} G_\alpha(\tau_1) + K(\theta_1,\theta_2;\tau_1), \\ H_{12}(\boldsymbol{\tau},\boldsymbol{\theta}) = H_{21}(\boldsymbol{\tau},\boldsymbol{\theta}) &= K_1(\theta_1,\theta_2;\tau_1), \\ H_{22}(\boldsymbol{\tau},\boldsymbol{\theta}) &= K_2(\theta_1,\theta_2;\tau_1). \end{aligned}$$

Evaluating $h_\theta(\tau) := DL(\theta, V_{\tau,\theta})$ and $h'_\theta(\tau+)$. Suppose first that $\theta \equiv 0$, so $D(\theta) = \emptyset$. Then one can show that

$$\begin{aligned} h'_\theta(\tau+) &= (\hat{F} - G_\alpha)(\tau), \\ h_\theta(\tau) &= \tau h'_\theta(\tau+) + \hat{\mu} - \int_{[0,\tau]} x\, \hat{P}(dx) - \alpha + \alpha G_{\alpha+1}(\tau). \end{aligned}$$

Now suppose that θ is given by a vector $\boldsymbol{\tau}$ of $m \geq 1$ points $\tau_1 < \cdots < \tau_m$ and a vector $\boldsymbol{\theta} = (\theta_j)_{j=1}^{m+1}$ as in (28). Then

$$h_\theta(0) \;=\; -\tau_1\big(\hat{F}(\tau_1) - e^{\theta_1} G_\alpha(\tau_1)\big) + \int 1_{[x \leq \tau_1]} x\, \hat{P}(dx) - e^{\theta_1} \alpha G_{\alpha+1}(\tau_1),$$

while for $0 \leq \tau < \tau_1$

$$\begin{aligned} h'_\theta(\tau+) &= \hat{F}(\tau) - e^{\theta_1} G_\alpha(\tau), \\ h_\theta(\tau) &= h_\theta(0) + \tau h'_\theta(\tau+) - \int 1_{[x \leq \tau]} x\, \hat{P}(dx) + e^{\theta_1} \alpha G_{\alpha+1}(\tau). \end{aligned}$$

For $1 \le j < m$ and $\tau \in [\tau_j, \tau_{j+1})$,

$$\begin{aligned} h'_\theta(\tau+) &= \hat{P}((\tau_j, \tau]) - J(\theta_j, \theta_*; \tau_j, \tau) - \int j_{10}(x; \tau_j, \tau_{j+1})\, \hat{P}(dx) + J_{10}(\theta_j, \theta_{j+1}; \tau_j, \tau_{j+1}), \\ h_\theta(\tau) &= (\tau - \tau_j)\big(h'_\theta(\tau+) + J_{01}(\theta_j, \theta_*; \tau_j, \tau)\big) - \int 1_{[\tau_j < x \le \tau]}(x - \tau_j)\, \hat{P}(dx), \end{aligned}$$

where $\theta_* := \theta(\tau) = (\tau_{j+1} - \tau_j)^{-1}\big((\tau_{j+1} - \tau)\theta_j + (\tau - \tau_j)\theta_{j+1}\big) = \theta_j + (\tau - \tau_j)\theta'_j$. Finally, for $\tau > \tau_m$,

$$\begin{aligned} h'_\theta(\tau+) &= K(\theta_*, \theta_{m+1}; \tau) - \hat{P}((\tau, \infty)), \\ h_\theta(\tau) &= (\tau - \tau_m)\big(h'_\theta(\tau+) + J_{01}(\theta_m, \theta_*; \tau_m, \tau)\big) - \int 1_{[\tau_m < x \le \tau]}(x - \tau_m)\, \hat{P}(dx), \end{aligned}$$

where $\theta_* := \theta_m + (\tau - \tau_m)\theta_{m+1}$.

If τ is restricted to some interval I not containing any observations x_i or knots τ_j, the expressions for $h'_\theta(\tau+)$ are constant in τ except for one term $e^{\theta_1} G_\alpha(\tau)$, $J(\theta_j, \theta_*; \tau_j, \tau)$ or $K(\theta_*, \theta_{m+1}; \tau)$. Hence finding τ such that $h'_\theta(\tau+) = 0$ leads to equations of the following type: For given real numbers $\theta_0, \theta_1, \tau_0$ and c, find $\tau \in [0, \infty)$ such that

$$e^{\theta_0} G_\alpha(\tau) = c, \tag{29}$$

$$J(\theta_0, \theta_0 + \theta_1(\tau - \tau_0); \tau_0, \tau) = c, \tag{30}$$

$$K(\theta_0 + \theta_1(\tau - \tau_0), \theta_1; \tau) = c, \tag{31}$$

and check whether $\tau \in I$. The unique solution of (29) is given by

$$\tau = G_\alpha^{-1}(c e^{-\theta_0})$$

with the quantile function $G_\alpha^{-1} : [0, 1) \to [0, \infty)$ of Gamma$(\alpha, 1)$, provided that $0 \le c e^{-\theta_0} < 1$; otherwise no solution exists. It follows from $J(\theta_0, \theta_0 + \theta_1(\tau - \tau_0); \tau_0, \tau) = (1 - \theta_1)^{-\alpha} e^{\theta_0 - \theta_1 \tau_0}\big(G_\alpha((1 - \theta_1)\tau) - G_\alpha((1 - \theta_1)\tau_0)\big)$ that the unique solution of (30) is given by

$$\tau = (1 - \theta_1)^{-1} G_\alpha^{-1}\big(c(1 - \theta_1)^\alpha e^{\theta_1 \tau_0 - \theta_0} + G_\alpha((1 - \theta_1)\tau_0)\big),$$

provided that $0 \le \theta_1 < 1$ and $0 \le c(1 - \theta_1)^\alpha e^{\theta_1 \tau_0 - \theta_0} + G_\alpha\big((1 - \theta_1)\tau_0\big) < 1$; otherwise no solution exists. Likewise it follows from $K(\theta_0 + \theta_1(\tau - \tau_0), \theta_1; \tau) = (1 - \theta_1)^{-\alpha} e^{\theta_0 - \theta_1 \tau_0}\big(1 - G_\alpha((1 - \theta_1)\tau)\big)$ that the unique solution of (31) is given by

$$\tau = (1 - \theta_1)^{-1} G_\alpha^{-1}\big(1 - c(1 - \theta_1)^\alpha e^{\theta_1 \tau_0 - \theta_0}\big),$$

provided that $0 \le \theta_1 < 1$ and $0 < c(1 - \theta_1)^\alpha e^{\theta_1 \tau_0 - \theta_0} \le 1$; otherwise no solution exists.

Activating one constraint. The activation of one constraint is identical to Setting 2A, except that here is no weight $\tilde{w}_0$.

Data Simulation. Let $P_o = \text{Gamma}(\alpha, \beta)$, and let $\theta \in \Theta$ such that $\gamma = \gamma(\theta) := \lim_{x \to \infty} \theta'(x+) < \beta$ and $\int f_\theta \, dP_o = 1$ with $f_\theta := e^\theta$. To simulate data from the density $f_\theta := e^\theta$ with respect to P_o, we use the acceptance rejection method of von Neumann (1951). We simulate independent random variables $Y \sim \text{Gamma}(\alpha, \beta - \gamma)$ and $U \sim \text{Unif}[0, 1]$. Note that Y has density $h(x) := (1 - \gamma/\beta)^{-\alpha} e^{\gamma x}$ with respect to P_o and that

$$(f_\theta / h)(x) = (f_\theta / h)(0) \exp\big(\theta(x) - \theta(0) - \gamma x\big)$$

is monotone decreasing in $x \ge 0$. Hence the conditional distribution of Y, given that $U \le \exp\big(\theta(Y) - \theta(0) - \gamma Y\big)$ is equal to the desired distribution P_θ. This leads to the following pseudocode for generating an independent sample $\boldsymbol{X}$

of size n from f_θ:

```
i ← 0
while i < n do
   simulate Y ~ Gamma(α, β − γ)
   simulate U ~ Unif([0, 1])
   if U ≤ exp(θ(Y) − θ(0) − γY) then
      i ← i + 1
      X_i ← Y
   end if
end while
```

A.5. Further proofs

Continuity of L on $(\mathbb{V}, \|\cdot\|)$ (Section 3.2). In Setting 1, the assertion is obvious, so we prove it for Settings 2A-B. Recall that a sequence $(\theta_k)_k$ in $\mathbb{V}$ converges to a function $\theta \in \mathbb{V}$ with respect to $\|\cdot\|$ if and only if it converges uniformly on any bounded subset of $\mathcal{X}$. Assuming this from now on, we want to show that $L(\theta_k) \to L(\theta)$ as $k \to \infty$. If $L(\theta) = -\infty$, then it follows from Fatou's lemma that

$$\limsup_{k\to\infty} L(\theta_k) \;=\; \int \theta \, d\hat{P} - \liminf_{k\to\infty} \int e^{\theta_k} \, dP_o + 1 \;\le\; L(\theta) = -\infty.$$

If $L(\theta) > -\infty$, then $\int \exp\bigl(\theta(x) + \varepsilon(1 + |x|)\bigr) \, P_o(dx) < \infty$ for sufficiently small $\varepsilon > 0$, and for sufficiently large k, $\theta_k(x) \le \theta(x) + \varepsilon(1 + |x|)$ for all $x \in \mathbb{R}$. Hence, it follows from dominated convergence that $L(\theta_k) \to L(\theta)$ as $k \to \infty$. □

Proof of Remark 3.2. Let $(\theta_k)_k$ be a sequence in $\Theta \cap \mathbb{V}$ such that $L(\theta_k) \to L(\hat{\theta})$ but $\theta_k \not\to \hat{\theta}$ pointwise as $k \to \infty$. As in the proof of Lemmas 2.6 and 2.7, we may replace this sequence by a subsequence, if necessary, such that it converges to some function $\theta_* \in \Theta \setminus \{\hat{\theta}\}$ with respect to $\|\cdot\|$. Since L is continuous, this implies that $L(\theta_k) \to L(\theta_*)$ as $k \to \infty$, whence $L(\theta_*) = L(\hat{\theta})$. Now, uniqueness of the maximizer of L on Θ leads to the contradiction that $\theta_* = \hat{\theta}$. □

Proof of Lemma 3.1 for Setting 2B and Setting 1. We only indicate the main changes in the proof for Setting 2A.

In Setting 2B, the constant C_ℓ may be replaced with 0, and the set $\mathcal{V}$ of basis functions consists of $v_0 \equiv 1$ and V_τ, $\tau \in \mathcal{D}$. This leads to $v_{\max}(x) = \max(1, x)$, and $\theta_{\max}(x) = C_o + C_r(x - x_n)^+$. Moreover, $\hat{\theta} - \theta = \alpha_0 + \sum_{\tau\in\mathcal{D}} \beta_\tau V_\tau$ with $|\alpha_0| = \bigl|\hat{\theta}(0) - \theta(0)\bigr| \le 2C_o$, and

$$\sum_{\tau\in\mathcal{D}} \beta_\tau^+ \;\le\; \hat{\theta}'(x_n) \;\le\; C_r, \quad \sum_{\tau\in\mathcal{D}} \beta_\tau^- \;\le\; \theta'(x_n) \;\le\; C_r.$$

Here $|DL(\theta, \eta_{\tau,\theta})| \le (1 + x_n)\sqrt{C_{\mathrm{N}}\delta_{\mathrm{Newton}}(\theta)}$, and this leads to obvious changes in the upper bound for $DL(\theta, \hat{\theta} - \theta)$.

In Setting 1, the main changes are as follows. We do not need the constants C_ℓ, C_r, and integrals $\int \cdots \, P_o(dx)$ have to be replaced with integrals $\int_{x_1}^{x_n} \cdots \, dx$. Here $v_{\max}(x) = \max(1, x - x_1)$, and $\theta_{\max} \equiv C_o$. The difference $\hat{\theta} - \theta$ equals $\alpha_0 v_0 + \alpha_1 v_1 + \sum_{\tau\in\mathcal{D}} \beta_\tau V_\tau$ with

$$\begin{aligned}
|\alpha_0| \;&=\; \bigl|\hat{\theta}(x_1) - \theta(x_1)\bigr| \;\le\; 2C_o, \\
|\alpha_1| \;&=\; \bigl|\hat{\theta}'(x_1+) - \theta'(x_1+)\bigr| \;\le\; 4C_o/(x_2 - x_1), \\
\beta_\tau \;&=\; \hat{\theta}'(\tau-) - \hat{\theta}'(\tau+) - \bigl(\theta'(\tau-) - \theta'(\tau+)\bigr) \;\begin{cases} \le\; \hat{\theta}'(\tau-) - \hat{\theta}'(\tau+), \\ \ge\; -\bigl(\theta'(\tau-) - \theta'(\tau+)\bigr). \end{cases}
\end{aligned}$$

In particular,

$$\left.\begin{aligned} \sum_{\tau\in\mathcal{D}} \beta_\tau^+ \;&\le\; \hat{\theta}'(x_1+) - \hat{\theta}'(x_n-) \\ \sum_{\tau\in\mathcal{D}} \beta_\tau^- \;&\le\; \theta'(x_1+) - \theta'(x_n-) \end{aligned}\right\} \;\le\; 2C_o/\min\{x_2 - x_1, x_n - x_{n-1}\}.$$

Setting 2A

n	$\hat{\kappa}_{n,0.10}$	$\hat{\kappa}_{n,0.05}$	$\hat{\kappa}_{n,0.01}$	time (ms)
100	2.923	3.763	5.653	3.087
400	3.298	4.179	6.133	4.282
1000	3.531	4.434	6.473	5.880
2000	3.682	4.613	6.678	8.355

Setting 2B

n	$\hat{\kappa}_{n,0.10}$	$\hat{\kappa}_{n,0.05}$	$\hat{\kappa}_{n,0.01}$	time (ms)
100	1.228	1.863	3.378	1.795
400	1.481	2.160	3.751	2.736
1000	1.622	2.317	3.879	4.228
2000	1.736	2.418	4.128	6.676

Table 2
Some estimated critical values for goodness-of-fit tests and mean running time per sample from P_o.

n	0	1	2	3	4	5	> 5
100	0.164	0.324	0.296	0.154	0.050	0.011	0.002
400	0.100	0.258	0.301	0.208	0.095	0.030	0.008
1000	0.075	0.217	0.290	0.231	0.123	0.047	0.017
2000	0.059	0.187	0.277	0.245	0.146	0.062	0.025

Table 3
Estimators of $P(M = m)$, $0 \le m \le 5$, and $P(M > 5)$ in Setting 2A.

n	0	1	2	3	4	> 4
100	0.360	0.445	0.165	0.028	0.002	0.000
	(0.000)	(0.069)	(0.029)	(0.006)	(0.000)	(0.000)
400	0.292	0.432	0.216	0.053	0.007	0.001
	(0.000)	(0.050)	(0.030)	(0.010)	(0.001)	(0.000)
1000	0.252	0.419	0.244	0.072	0.012	0.001
	(0.000)	(0.040)	(0.031)	(0.010)	(0.002)	(0.000)
2000	0.229	0.403	0.263	0.086	0.017	0.002
	(0.000)	(0.034)	(0.030)	(0.011)	(0.002)	(0.000)

Table 4
Estimators of $P(M = m)$, $0 \le m \le 4$, and $P(M > 4)$ in Setting 2B. In brackets are the estimators of $P(J = 1, M \ldots)$.

Here we utilized the fact that $v'(x_i+) = v'(x_{i+1}-) = \big(v(x_{i+1}) - v(x_i)\big)/(x_{i+1} - x_i)$ for $v \in \mathbb{V}$ and $1 \le i < n$. Finally, $|DL(\theta, \eta_{\tau,\theta})| \le \sqrt{C_{\mathrm{N}} \delta_{\mathrm{Newton}}(\theta)}$, because $\eta_{\tau,\theta}$ is always a convex combination of two basis functions in $\mathcal{V} \cap \mathbb{V}_{D(\theta)}$. □

A.6. On the distribution of T_{LR} under the null hypothesis

For the goodness-of-fit tests with a given sample size n, we simulated $10^5 - 1$ times a sample $X_1, \ldots, X_n$ from P_o and recorded the test statistic $T_{LR} = T_{LR}(X_1, \ldots, X_n)$ as well as the number $M = M(X_1, \ldots, X_n)$ of kinks of the estimator $\hat{\theta} = \hat{\theta}(\cdot \mid X_1, \ldots, X_n)$. The reference distribution P_o was $\mathcal{N}(0, 1)$ in Setting 2A and χ^2_1 in Setting 2B. In the latter setting, we also recorded the indicator $J = J(X_1, \ldots, X_n)$ that $\hat{\theta}$ has a kink at 0, i.e. $\hat{\theta}'(0+) > 0$.

Table 2 contains critical values $\hat{\kappa}_{n,\alpha}$ for different sample sizes n and different test levels α. Tables 3 and 4 contain the estimated distribution of the random number M in Settings 2A and 2B, respectively. In the latter setting, Monte Carlo estimators of probabilities $P(J = 1, M \cdots)$ are listed as well.