Statistics and Computing (2024) 34:168
https://doi.org/10.1007/s11222-024-10480-y

ORIGINAL PAPER

Taming numerical imprecision by adapting the KL divergence to negative probabilities

Simon Pfahler[1] · Peter Georg[1] · Rudolf Schill[2] · Maren Klever[3] · Lars Grasedyck[3] · Rainer Spang[4] · Tilo Wettig[1]

Received: 1 February 2024 / Accepted: 23 July 2024

Abstract
The Kullback–Leibler (KL) divergence is frequently used in data science. For discrete distributions on large state spaces, approximations of probability vectors may result in a few small negative entries, rendering the KL divergence undefined. We address this problem by introducing a parameterized family of substitute divergence measures, the shifted KL (sKL) divergence measures. Our approach is generic and does not increase the computational overhead. We show that the sKL divergence shares important theoretical properties with the KL divergence and discuss how its shift parameters should be chosen. If Gaussian noise is added to a probability vector, we prove that the average sKL divergence converges to the KL divergence for small enough noise. We also show that our method solves the problem of negative entries in an application from computational oncology, the optimization of Mutual Hazard Networks for cancer progression using tensor-train approximations.

Keywords Kullback–Leibler divergence · Approximate Bayesian computation · Statistical optimization · Mutual Hazard Networks · Tensor trains

1 Introduction

1.1 Motivation

A notion of distance between probability distributions is required in many fields, e.g., information and probability theory, approximate Bayesian computation (Csilléry et al. 2010; Schill et al. 2019), spectral (re-) construction (Hoch 2014; Rothkopf 2017; Ha et al. 2019), or compression using variational autoencoders (Kingma and Welling 2022). In mathematical terms, this notion can be phrased in terms of divergence measures.

One of the most commonly used divergence measures is the Kullback–Leibler (KL) divergence (Kullback and Leibler 1951), which for two discrete probability vectors $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{q} = (q_1, \ldots, q_n)$ has the form

$$D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{q}) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \,. \tag{1}$$

The KL divergence includes the logarithm, which is well-defined in the usual case of probability vectors with nonnegative entries. However, in practical applications, situations can arise that lead to negative entries, for which the logarithm is not defined. Let us give a few examples. In high-dimensional spaces, approximations are often needed to keep calculations tractable and to avoid runtime explosion due to the curse of dimensionality (Thomas and Grima 2015; Georg et al. 2022). Such approximations can lead to negative entries when the exact entry is close to zero, as is frequently the case for probability vectors. Another source of negative entries are rounding errors that occur, e.g., when the probability vector is obtained as the solution of a linear system of equations (Philippe et al. 1992; Schill et al. 2019). Also, assumptions of a theory can lead to negative entries in the approximate probability distribution (Alkofer and Smekal 2001; Burnier et al. 2011; Rothkopf 2017).

If negative entries arise, one is faced with the choice of (a) giving up, (b) reformulating the theory, model, or approxima-

✉ Simon Pfahler
simon.pfahler@ur.de

1 Department of Physics, University of Regensburg, 93040 Regensburg, Germany

2 Department of Biosystems and Engineering, ETH Zürich, 4056 Basel, Switzerland

3 Institute for Geometry and Applied Mathematics, RWTH Aachen, 52062 Aachen, Germany

4 Department of Informatics and Data Science, University of Regensburg, 93040 Regensburg, Germany

Published online: 13 August 2024

tion such that negative entries cannot occur, or (c) devising a workaround that leads to correct results in a suitable limit. In Sect. 1.2 we review approaches from category (b) and (c) to address the problem of negative entries (Paatero and Tapper 1994; Hobson and Lasenby 1998; Welling and Weber 2001; Chi and Kolda 2012; Haas et al. 2014; Hansen et al. 2015; Lee et al. 2016; Rothkopf 2017). The methods we are aware of are typically tailored to the problem at hand or lead to reduced run-time performance.

Here, we suggest a method in category (c) that is generically applicable and does not suffer from decreased performance. Specifically, we propose the *shifted Kullback–Leibler (sKL) divergence* as a substitute divergence measure for the standard KL divergence in the case of approximate probability vectors with negative entries. It is parameterized by a vector of shift parameters. The sKL divergence is a modification of the KL divergence and retains many of the latter's useful properties while handling negative entries when they arise. To give theoretical support to our method, we consider a simple example in which i.i.d. Gaussian noise is added to the entries q_i in Eq. (1) so that negative entries can occur. For this example we prove that the difference between the KL divergence and the expected sKL divergence under the noise is quadratic in the standard deviation of the noise, provided that the shift parameters are suitably chosen. Therefore the sKL divergence converges to the KL divergence in the limit of small i.i.d. Gaussian noise. We also show this convergence numerically.

In a concrete application, we show that the sKL divergence enables efficient learning of a Mutual Hazard Network (MHN), a model of cancer progression as an accumulation of certain genetic events (Schill et al. 2019; Chen 2023; Luo et al. 2023). In an MHN, cancer progression is modeled as a continuous-time Markov chain whose transition rates are parameters of the model that are learned such that the model explains a given patient-data distribution. When considering $\gtrsim 25$ possible genetic events, exact model construction becomes impractical and requires efficient approximation techniques (Georg et al. 2022). We show that even though negative entries occur in the approximate probability vectors, meaningful models can still be constructed when the sKL divergence is employed.

1.2 Related work

Various approaches to handling or avoiding negative entries in approximate probability distributions exist in the literature.

In earlier work (Georg 2022), we considered introducing a threshold $\varepsilon > 0$ and replacing entries of the probability vector that are smaller than this threshold by ε. This method fails to preserve probability mass and leads to a loss of gradient information when used in an optimization task. Additionally, the comparison function obtained in this way no longer meets the requirements of a statistical divergence (Amari 2016), i.e., none of the required properties in Theorem 1 below are satisfied in this case.

For high-dimensional probability distributions, several nonnegative tensor decompositions have been proposed in order to break the curse of dimensionality while avoiding negative entries (Paatero and Tapper 1994; Welling and Weber 2001; Chi and Kolda 2012; Hansen et al. 2015; Lee et al. 2016). Typically, this change of format leads to a significant increase in the run time of the algorithms involved, as other important properties for a time-efficient approximation have to be omitted.

In situations where negative entries occur due to assumptions of the theory, a common approach is to modify the divergence measure (Hobson and Lasenby 1998; Rothkopf 2017; Haas et al. 2014). Such modifications are typically problem-specific and therefore not applicable in general.

Our approach is in a similar spirit, but our modification of an established divergence measure is not problem-specific, as it does not depend on special properties of a particular application. Furthermore, since we do not need specific formats to preserve nonnegativity, optimization tasks in high dimensions can be completed efficiently.

This paper is structured as follows. In Sect. 2 we define the sKL divergence, discuss its properties and give suggestions for how to choose the shift parameters. In Sect. 3 we show how the sKL divergence can be applied in optimization tasks on approximate probability vectors. In Sect. 4 we summarize our findings and give an outlook on future research topics. In Sect. A and B we provide proofs of three theorems on the sKL divergence and mathematical details of the approximation we employ.

2 Theory

2.1 The sKL divergence

For two probability vectors $\mathbf{p}$ and $\mathbf{q}$ on a finite set of n elements, the Kullback–Leibler (KL) divergence of $\mathbf{p}$ from $\mathbf{q}$ is defined as (Kullback and Leibler 1951)

$$D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{q}) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}$$

with the convention $0 \log(0/x) = 0$. In the context of statistical modeling, the vector $\mathbf{p}$ (with $p_i \geq 0$) is typically given by the data, while the vector $\mathbf{q}$ is a theoretical model. This model can be fitted to the data by minimizing the KL divergence of $\mathbf{p}$ from $\mathbf{q}$. Approximations in the calculation of $\mathbf{q}$ can lead

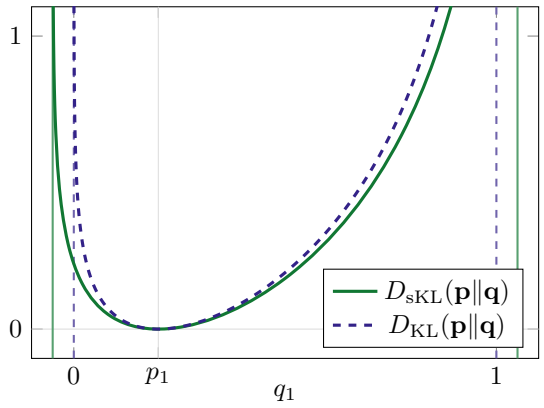

Fig. 1 Plot of the KL and sKL divergence of probability vector $\mathbf{p} = (0.2, 0.8)$ from $\mathbf{q} = (q_1, 1 - q_1)$. For the sKL divergence, $\varepsilon_i = 0.05$ was chosen for $i = 1, 2$

to negative entries $q_i < 0$. [1] In this case, the KL divergence is no longer defined and the optimization cannot be done. For this scenario, we propose the shifted Kullback–Leibler (sKL) divergence, a modification of the Kullback–Leibler divergence that allows for negative entries in $\mathbf{q}$. For a parameter vector $\boldsymbol{\varepsilon} \in \mathbb{R}^n_{\geq 0}$, we define it as

$$D_{\mathrm{sKL}}(\mathbf{p}\|\mathbf{q}) = \sum_i (p_i + \varepsilon_i) \log \frac{p_i + \varepsilon_i}{q_i + \varepsilon_i} \tag{2}$$

with the same convention as above. The sKL divergence is well defined for $q_i > -\varepsilon_i$. It is a modification of the KL divergence that reduces to the KL divergence for $\boldsymbol{\varepsilon} = 0$. Figure 1 shows the behavior of both the KL and the sKL divergence for probability vectors on a set with two possible outcomes. We can see that the domain on which the sKL divergence is defined now also includes approximate probability vectors with small negative entries. In the following two subsections we show that the sKL divergence retains many important properties of the KL divergence and discuss how to best choose the parameters ε_i.

2.2 Properties

The KL divergence belongs to the class of f-divergences (Basseville 2013). While the sKL divergence does not, it shares many important properties with the KL divergence: it is positive semidefinite, only zero for $\mathbf{p} = \mathbf{q}$, and locally a metric. Thus, it still satisfies the definition of a statistical divergence, see Theorem 1.

[1] For simplicity, we do not consider the possibility that the entries q_i are zero up to machine precision because this is extremely unlikely to happen in a numerical simulation. However, it is straightforward to extend our analysis to this case.

Theorem 1 *Let* $\mathbf{p}$, $\mathbf{q}$ *and* $\boldsymbol{\varepsilon}$ *be three vectors in* $\mathbb{R}^n$ *whose entries satisfy* $p_i, q_i > -\varepsilon_i$ *and* $\sum_i p_i = \sum_i q_i$. *The sKL divergence of* $\mathbf{p}$ *from* $\mathbf{q}$, *with parameter vector* $\boldsymbol{\varepsilon}$, *is a statistical divergence, i.e., it satisfies the following properties* (Amari 2016),

(1) $D_{\mathrm{sKL}}(\mathbf{p}\|\mathbf{q}) \geq 0$,
(2) $D_{\mathrm{sKL}}(\mathbf{p}\|\mathbf{q}) = 0$, *if and only if* $\mathbf{p} = \mathbf{q}$
(3) $\frac{\mathrm{d}^2 D_{\mathrm{sKL}}(\mathbf{p}\|\mathbf{q})}{\mathrm{d}q_i \mathrm{d}q_j}\big|_{\mathbf{q}=\mathbf{p}}$ *is a positive definite matrix.*

The first assumption of the theorem can always be satisfied by a suitable choice of the shift parameters ε_i, while the second assumption can always be satisfied by a rescaling of the q_i. Additionally, we show in Theorem 2 that the sKL divergence is convex in the pair of its arguments, like the KL divergence (van Erven and Harremos 2014).

Theorem 2 *For fixed parameter vector* $\boldsymbol{\varepsilon}$, *the sKL divergence is convex in the pair of its arguments. That is, if* $(\mathbf{p}^{(1)}, \mathbf{q}^{(1)})$ *and* $(\mathbf{p}^{(2)}, \mathbf{q}^{(2)})$ *are two pairs of vectors in* $\mathbb{R}^n$ *for which* $D_{\mathrm{sKL}}(\mathbf{p}^{(1)}\|\mathbf{q}^{(1)})$ *and* $D_{\mathrm{sKL}}(\mathbf{p}^{(2)}\|\mathbf{q}^{(2)})$ *are well-defined, and if* $\lambda \in [0, 1]$, *it satisfies*

$$\begin{aligned} &D_{\mathrm{sKL}}(\lambda\mathbf{p}^{(1)} + (1-\lambda)\mathbf{p}^{(2)}\|\lambda\mathbf{q}^{(1)} + (1-\lambda)\mathbf{q}^{(2)}) \\ &\leq \lambda D_{\mathrm{sKL}}(\mathbf{p}^{(1)}\|\mathbf{q}^{(1)}) + (1-\lambda) D_{\mathrm{sKL}}(\mathbf{p}^{(2)}\|\mathbf{q}^{(2)})\,. \end{aligned}$$

This property is of particular importance as it is often needed for well-behaved optimization (Fletcher 2000). The proofs of the two theorems are given in Sect. A.

2.3 Parameter choice

In the sKL divergence, we introduced a parameter vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$. The properties of the sKL divergence discussed in the previous subsection (Theorem 1 and 2) hold for any choice of the shift parameters ε_i. However, in applications, their values can have a large influence on the quality of the results, such as speed of convergence or accuracy of the learned model. In this section we therefore aim to provide a guide for choosing suitable values of the parameters ε_i for the case that $\mathbf{p}$ is an exact probability vector and $\mathbf{q}$ is an approximate one. In the KL divergence, the terms in the sum only need to be evaluated for the indices i for which $p_i \neq 0$. This property reduces the computational cost greatly if many entries of $\mathbf{p}$ are zero. This situation is encountered, for example, when optimizing an MHN (Schill et al. 2019). In order to preserve this advantage when working with the sKL divergence, one can simply choose $\varepsilon_i = 0$ if $p_i = 0$. To constrain possible values for ε_i, we first note that if $q_i < 0$, the sKL divergence is only well-defined if one chooses $\varepsilon_i > -q_i$.

Second, we compute the gradient of the sKL divergence,

$$\frac{\mathrm{d}D_{\mathrm{sKL}}(\mathbf{p}\|\mathbf{q})}{\mathrm{d}q_i} = -\frac{p_i+\varepsilon_i}{q_i+\varepsilon_i},$$

and notice that it approaches -1 for large values of ε_i. Most optimization algorithms use gradient information to minimize a loss function (Fletcher 2000). Therefore it is important to retain the gradient information, which implies that ε_i should not be too large.

Using these guidelines, we provide two particular choices of $\boldsymbol{\varepsilon}$. First, in Sect. 2.3.1, we discuss a natural, static choice of $\boldsymbol{\varepsilon}$ and examine its shortcomings. In Sect. 2.3.2, we then suggest a dynamic choice of $\boldsymbol{\varepsilon}$ and prove in Theorem 3 that the resulting sKL divergence is a good substitute for the KL divergence in the presence of small i.i.d. Gaussian noise. When possible, one should therefore prefer the dynamic choice of $\boldsymbol{\varepsilon}$, as we will also demonstrate in Sect. 3.3.

2.3.1 Static choice

We first suggest a choice of $\boldsymbol{\varepsilon}$ that can be used with any optimization algorithm. This includes higher-order algorithms such as BFGS or conjugate-gradient methods, which make use of approximations of the Hessian matrix of the loss function (Fletcher 2000). For higher-order methods, the objective function must not change during optimization. This requires us to choose a suitable $\boldsymbol{\varepsilon}$ a priori, which can be a difficult task. A simple approach is to evaluate $\mathbf{q}$ once before starting the optimization and to set

$$\varepsilon_i = \begin{cases} 0\,, & p_i = 0\,, \\ \varepsilon\,, & \text{else}\,, \end{cases} \tag{3}$$

where ε is fixed to a number slightly larger than $\max_{p_j>0}(-q_j)$. If during optimization ε turns out to be too small, i.e., negative entries of $q_i+\varepsilon$ are encountered, the optimization has to be stopped early. The best possible result can then be found by tuning ε.

This parameter choice allows us to choose a wide variety of optimizers. However, in addition to the requirement of tuning ε, the static choice leads to an unnecessarily large loss of gradient information. This is because gradient information is reduced for all i, even though $q_i < 0$ is rarely encountered in practice.

2.3.2 Dynamic choice

We now suggest a second choice that can be used only if the objective function is allowed to change in every iteration of the optimizer. This does not pose a problem when using first-order optimization algorithms such as gradient descent (Fletcher 2000) or Adam (Kingma and Ba 2017), as these do not use information about the Hessian matrix of the loss function. In this scenario, we can choose a new $\boldsymbol{\varepsilon}$ for every new $\mathbf{q}$ during the optimization process. Our proposed choice is

$$\varepsilon_i = \begin{cases} 0\,, & p_i = 0 \text{ or } q_i > 0\,, \\ |q_i| + f(|q_i|)\,, & \text{else}\,, \end{cases} \tag{4}$$

where $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a nonnegative function. This choice of $\boldsymbol{\varepsilon}$ avoids a large loss in gradient information while extending the domain on which the sKL divergence is defined where necessary. A concrete choice of f, which we use in Sect. 3, is given by $f(x) = \delta \cdot x$ with a constant $\delta > 0$, but many other choices are possible.

2.4 Negative entries due to Gaussian noise

To further motivate Eq. (4), let us consider a simple example in which negative entries of q_i are caused by small Gaussian noise. In this case, we show that the difference between the KL divergence and the average of the sKL divergence over the Gaussian noise is quadratic in the standard deviation of the noise, see Theorem 3, the proof of which is provided in Sect. A. To keep the presentation simple, we restrict ourselves to nonnegative functions f that are finite sums of power laws, i.e., $f(x) = \sum_k a_k x^{b_k}$ with at least one $a_k \neq 0$.

Theorem 3 *Let* $\mathbf{p}$ *and* $\mathbf{q}$ *be two probability vectors on a finite set of* n *elements, with* $q_i > 0$ *whenever* $p_i > 0$. *Let further* $\mathbf{x}$ *be a vector of i.i.d. Gaussian random variables with mean* 0 *and standard deviation* σ. *If the* ε_i *are chosen according to Eq.* (4) *with the restriction on* f *stated above, the average of the sKL divergence of* $\mathbf{p}$ *from* $\mathbf{q}+\mathbf{x}$ *is given by*

$$\left\langle D_{\mathrm{sKL}}(\mathbf{p}\|\mathbf{q}+\mathbf{x})\right\rangle_{\mathbf{x}} = D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{q}) + \sigma^2 \sum_i \frac{p_i}{2q_i^2} + \mathcal{O}(\sigma^4)\,.$$

In fact, the restriction on f can be relaxed significantly so that a much larger class of functions is admissible, see Sect. A.3 for details.

In this example, we therefore see explicitly that the sKL divergence can be used as a substitute divergence measure for the KL divergence, provided that the condition $\sigma^2 \sum_i p_i/2q_i^2 \ll D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{q})$ is satisfied.

To validate Theorem 3 numerically, we first consider a generic toy model where $\mathbf{p}, \mathbf{q} \in \mathbb{R}^{10}$ are probability vectors whose entries are i.i.d. uniform random variables. As per the assumptions of Theorem 3, we consider i.i.d. Gaussian random variables added to $\mathbf{q}$, and we choose the ε_i according to Eq. (4) with the simple choice $f(x) = \delta \cdot x$. In order for the σ^2 correction to be small, we restrict ourselves to probability

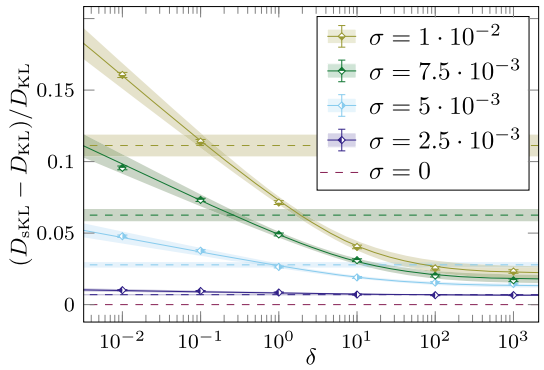

Fig. 2 Relative difference between the KL divergence and the sKL divergence with the dynamic choice of ε given by Eq. (4) with $f(x) = \delta \cdot x$, for different noise levels σ. For each σ, the dashed line shows the prediction from Theorem 3 truncated at order σ^2, the solid line shows the result obtained from numerical integration of Eq. (A1), and the points show the result obtained from simulation

vectors $\mathbf{p}$ and $\mathbf{q}$ that satisfy

$$\sigma^2 \sum_i \frac{p_i}{2q_i^2} \leq \frac{1}{2} D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{q})$$

for all σ considered. Figure 2 shows the convergence of the sKL divergence to the KL divergence in the limit of small σ of the noise. The plot shows averages of 200 pairs of $\mathbf{p}$ and $\mathbf{q}$. We can see that in this case Theorem 3 gives a good prediction for the relative difference between the sKL divergence and the KL divergence. Interestingly, for large values of δ, the relative difference is significantly lower than predicted. Therefore, large values $\delta \geq 10^1$ should be preferred in this toy model, but this may be different for other applications, in particular when the negative entries are not due to Gaussian noise. Finally, we observe that the parameter δ in the function f has only little effect on the result when σ is sufficiently small.

In many applications, the strength of the noise can be controlled by parameters that determine the numerical accuracy of the approximation. If the assumption of i.i.d. Gaussian noise is satisfied in a particular application, Theorem 3 is directly applicable. In other applications, the noise might follow a different distribution, for which one would have to do a similar analysis.

In the concrete application we consider in Sect. 3, we have seen numerically that the assumption of i.i.d. Gaussian noise is not satisfied. Nevertheless, as we increase the numerical accuracy of our approximation, we observe that the sKL divergence with the choice of Eq. (4) converges to the KL divergence computed without approximations. Also, the model results obtained from optimizing the sKL divergence with the choice of Eq. (4) converge to the model result obtained from the KL divergence without approximations.

3 Application

3.1 Mutual Hazard Networks

As a real-world application, we consider the modeling of cancer progression using Mutual Hazard Networks (MHNs) (Schill et al. 2019). Cancer progresses by accumulating genetic events such as mutations or copy-number aberrations (Michor et al. 2004). As each event can be either absent or present, the state of an event is represented by 0 (absent) or 1 (present). We consider d such events, thus representing a tumor as a d-dimensional vector $x \in \{0, 1\}^d$. An MHN aims to infer promoting and inhibiting influences between single events. The data used for the learning process are patient data of observed tumors. The data distribution $\mathbf{p}$ is thus a discrete probability distribution on the 2^d-dimensional state space $\mathcal{S} = \left\{x \in \{0, 1\}^d\right\}$ of possible tumors, i.e., $n = 2^d$ in the notation of Sect. 2.

An MHN models the progression as a continuous-time Markov chain under the assumptions that at time $t = 0$ no tumor has active events, that events occur only one at a time, that events are not reversible, and that transition rates follow the Proportional Hazard Assumption (Cox 1972). If we have two states $x, x^{+i} \in \mathcal{S}$ that differ only by $x_i = 0$ and $x_i^{+i} = 1$, the transition rate from state x to state x^{+i} is modeled as [2]

$$\mathbf{Q}_{x^{+i},x} = e^{\theta_{ii}} \prod_{x_j=1} e^{\theta_{ij}} ,$$

where $e^{\theta_{ii}}$ is the base rate of event i and $e^{\theta_{ij}}$ is the multiplicative influence event j has on event i. An MHN with d events can thus be described by a parameter matrix $\theta \in \mathbb{R}^{d\times d}$. Figure 3 shows all allowed transitions and their rates for $d = 3$. The transition-rate matrix can efficiently be written as a sum of d Kronecker products,

$$\mathbf{Q} = \sum_{i=1}^{d} \bigotimes_{j=1}^{d} Q_{ij} \tag{5}$$

with

$$Q_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & e^{\theta_{ij}} \end{pmatrix} \text{ for } i \neq j \text{ and } Q_{ii} = \begin{pmatrix} -e^{\theta_{ii}} & 0 \\ e^{\theta_{ii}} & 0 \end{pmatrix}.$$

Starting from the initial distribution $\mathbf{q}_\varnothing = (1, 0, 0, \ldots) \in \mathbb{R}^{2^d}$, the probability distribution at time $t \geq 0$ is given by Markov-chain theory as

$$\mathbf{q}(t) = e^{t\mathbf{Q}}\mathbf{q}_\varnothing ,$$

[2] All other off-diagonal transition rates are zero by assumption of the MHN.

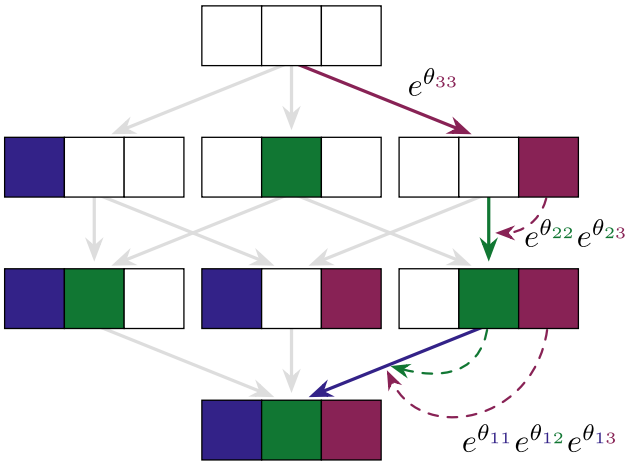

Fig. 3 Allowed transitions, along with transition rates, for an MHN with three possible events and parameter matrix $\theta \in \mathbb{R}^{3\times 3}$

where $e^{t\mathbf{Q}}$ denotes the matrix exponential. Since tumor age is not known in the data, MHNs assume that the age of tumors in $\mathbf{p}$ is an exponentially distributed random variable with mean 1. If we marginalize $\mathbf{q}(t)$ over t accordingly, we obtain

$$\mathbf{q}_\theta := \int_0^\infty \mathrm{d}t\ e^{-t} e^{t\mathbf{Q}} \mathbf{q}_\varnothing = (\mathbf{I} - \mathbf{Q})^{-1}\, \mathbf{q}_\varnothing\,, \tag{6}$$

where $\mathbf{I}$ denotes the identity.

A parameter matrix θ that best fits the data distribution $\mathbf{p}$ can now be obtained by minimizing a divergence measure of $\mathbf{p}$ from $\mathbf{q}_\theta$. For small d (i.e., $d \lesssim 25$), the KL divergence can be used since the calculation of $\mathbf{q}_\theta$ can be done without approximation. For larger d, this is no longer possible because of the exponential complexity (recall that the dimension of the state space is 2^d), and the approach has to be modified (Georg et al. 2022). One possible modification is the use of low-rank tensor methods (Grasedyck et al. 2013) in order to keep calculations tractable. In particular, we use the tensor-train (TT) format (Oseledets 2011), which usually minimizes the approximation error through the Euclidean distance. Specifically, when approximating a tensor $\mathbf{a}$ by $\tilde{\mathbf{a}}$, a maximum Euclidean distance $\Delta = ||\mathbf{a} - \tilde{\mathbf{a}}||_2$ can be specified. Thus, small negative entries can occur in $\tilde{\mathbf{a}}$ if the corresponding entry in $\mathbf{a}$ is smaller than Δ. In this case, the KL divergence is no longer defined. To be able to perform the optimization, we switch to the sKL divergence as our objective function. In Sect. 3.2, we explain the basics of the TT format. Sect. 3.3 shows how we can find suitable θ matrices by use of the sKL divergence even when negative entries are encountered.

3.2 Tensor Trains

For a large number d of events, storing $\mathbf{q}$ (and $\mathbf{p}$) as a 2^d-dimensional vector is computationally infeasible. This storage requirement can be alleviated through use of the TT format (Oseledets 2011). A d-dimensional tensor $\mathbf{a} \in \mathbb{R}^{n_1\times\cdots\times n_d}$ with mode sizes $n_k \in \mathbb{N}$ can be approximated in the TT format as

$$\mathbf{a}(i_1, \ldots, i_d) \approx \sum_{\alpha_0,\ldots,\alpha_d} \prod_{k=1}^{d} \mathbf{a}^{(k)}(\alpha_{k-1}, i_k, \alpha_k)$$

for all $i_1, \ldots, i_d$ with tensor-train cores $\mathbf{a}^{(k)} \in \mathbb{R}^{r_{k-1}\times n_k \times r_k}$, where the r_k are tensor-train ranks.[3] In order to represent a scalar by the right-hand side, the condition $r_0 = r_d = 1$ is required. The quality of approximation can be controlled through the TT ranks r_k. In particular, it can be shown that choosing the TT ranks large enough gives an exact representation of $\mathbf{a}$ (Oseledets 2011).

The TT format not only allows for efficient storage of high-dimensional tensors, but also supports many basic operations, e.g., addition, inner products, or operator-by-tensor products (Oseledets 2011). Furthermore, there are efficient algorithms for solving linear equations (Holtz et al. 2012; Dolgov and Savostyanov 2014).

By modifying the shape of the objects in Eq. (5) and changing the Kronecker products to tensor products, the transition-rate matrix $\mathbf{Q} \in \mathbb{R}^{\mathcal{S}\times\mathcal{S}}$ can be written in the tensor-train format (Hackbusch 2019) (for details see Sect. B). This leads to a tensor train with all TT ranks $r_1, \ldots, r_{d-1}$ equal to d. Thus, we can approximately solve Eq. (6) in the TT format using Holtz et al. (2012). Similar techniques can be used for the gradient calculation (Georg 2022).

Details of the algorithms involved (Holtz et al. 2012) lead to the conditions $r_{k+1} \le n_{k+1} r_k$ and $r_{k-1} \le n_k r_k$ on the TT ranks of $\mathbf{q}_\theta$. As a result, the first and last TT ranks are $r_1 \le n_1$ and $r_{d-1} \le n_d$. Towards the middle of the tensor train, ranks increase until they level off at the specified maximum rank. In our simulations, we specify a maximum TT rank $r_\mathbf{q}$ and choose the TT ranks $r_k \le r_\mathbf{q}$ to be as large as possible given the constraints on the r_k.

3.3 Simulations

We test how well an MHN can learn a probability distribution $\mathbf{p}$ when optimizing the sKL divergence instead of the classical KL divergence. We use simulated data for $d = 20$ events. This relatively small value of d allows for exact calculations so that we can compare results obtained with and without the

[3] Strictly speaking, the r_k are only called TT ranks if they are chosen minimally. To simplify the presentation we use the term "TT rank" regardless of this convention.

use of tensor trains. In the following, we describe how the data were generated, how the MHNs were learned, and how their quality was assessed.

Every dataset was generated from a ground-truth model described by $\theta_{\mathrm{GT}} \in \mathbb{R}^{d\times d}$. The diagonal entries of θ were drawn from a Gaussian distribution $\exp(-(x-\mu)^2/2\sigma^2)/\sigma\sqrt{2\pi}$ with $\mu = -1$ and $\sigma = 1$. A random set of 10% of the off-diagonal entries of θ was drawn from a Laplace distribution $\exp(-|x-\mu|/b)/2b$ with $\mu = 0$ and $b = 1$, and the remaining entries were set to 0. These choices were made to mimic MHNs obtained from real data we studied. The data distribution $\mathbf{p}$ was obtained from θ_{GT} by drawing 1000 samples from its time-marginalized probability distribution, as defined in Eq. (6).

Given a dataset, MHNs were learned by optimizing the sKL divergence. We indirectly control the magnitude of the largest negative entries of $\mathbf{q}_\theta$ through our choice of the maximum possible TT rank $r_{\mathbf{q}}$ of $\mathbf{q}_\theta$. 10 datasets were generated, and for all of them, MHNs were learned for specific choices of $r_{\mathbf{q}}$ and ε. The results shown below are arithmetic means from these 10 runs. To avoid overfitting, an L1 penalty term of the form $\lambda \sum_{i\neq j} |\theta_{ij}|$ was added to the objective function. The factor λ was not made part of the optimization but set to a constant value of 10^{-3} for all simulation runs to make comparison of different models easier.

A learned MHN's quality was assessed using the KL divergence of the ground truth model's time-marginalized probability distribution from the time-marginalized probability distribution of the learned MHN, see Eq. (6). This KL divergence was calculated without the use of the TT format to ensure that no negative entries can occur.

3.3.1 Static choice

First, we consider the static choice of ε given in Eq. (3). In this case, higher-order optimizers can be used for faster convergence to an optimum. However, for a fair comparison with the dynamic choice, we used gradient descent (a first-order optimizer) for all optimizations. Table 1 and Fig. 4 show the results for various combinations of $r_{\mathbf{q}}$ and ε. In the last column ("exact"), optimization using the sKL divergence was also done when the TT format was not used, even though the KL divergence is well-defined and the introduction of a positive ε is not necessary in this case. The additional entries are written in grey to indicate this.

For increasing value of $r_{\mathbf{q}}$, the TT approximation improves, and accordingly the approximate results tend towards the exact result, although the convergence is quite slow. If $q_{\theta,i} + \varepsilon < 0$ occurred in any iteration of the optimization procedure, the optimization was stopped and the last θ matrix was returned. The colors indicate the percentage of runs where this happened.

Table 1 Average KL divergence (without approximation) from the ground-truth model θ_{GT} for MHNs obtained by optimizing the sKL divergence with the static choice of ε given by Eq. (3)

	$r_{\mathbf{q}}$					
ε	4	8	16	24	32	exact
0	0.354	0.241	0.199	0.131	0.106	0.087
10^{-10}	0.355	0.237	0.192	0.135	0.105	0.088
10^{-9}	0.354	0.237	0.193	0.147	0.101	0.088
10^{-8}	0.354	0.237	0.190	0.132	0.103	0.090
10^{-7}	0.353	0.228	0.163	0.106	0.094	0.100
10^{-6}	0.338	0.202	0.145	0.126	0.122	0.150
10^{-5}	0.358	0.295	0.316	0.321	0.326	0.434
10^{-4}	0.609	0.833	0.910	0.953	0.935	1.155
10^{-3}	2.069	3.202	3.319	3.327	3.341	3.454

If $q_{\theta,i}+\varepsilon < 0$ occurred during optimization, the last available MHN was used to calculate the KL divergence from the ground truth. The cell color indicates the percentage of simulations where this occurred (a darker color corresponds to a higher percentage). In the column "exact", the TT format was not used during optimization, thus no negative entries occurred here

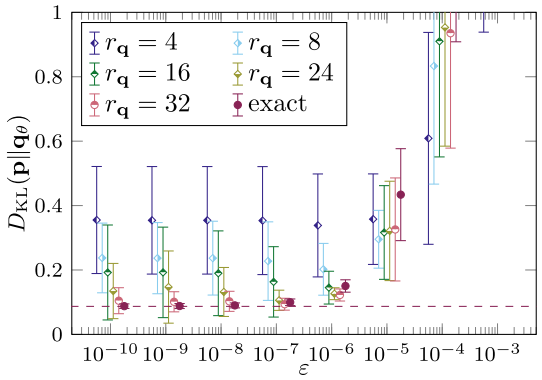

Fig. 4 Visualization of the data in Table 1. The dashed line indicates the result when optimizing the KL divergence without the use of the TT format

The first row of Table 1 shows the results of optimization using the KL divergence. It can clearly be seen that, when the TT format is used, optimization using the sKL divergence leads to MHNs that describe the data more closely. The optimal ε value is between 10^{-7} and 10^{-6} across all cases where the TT format is used. The results also clearly show that, as explained in Sect. 2.3, if ε is chosen too small or too large, we obtain poor optimization results. From the colors we can see that early stopping due to negative entries of $q_{\theta,i}+\varepsilon$ does not have a big influence on the optimization results. This is because early stopping often occurred late in the optimization procedure, i.e., after a good estimate for the optimum was already found.

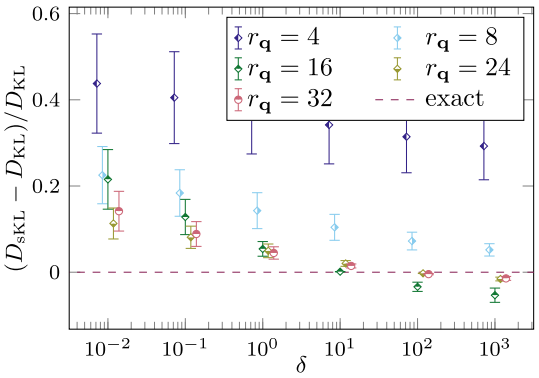

Fig. 5 Relative difference between the KL divergence and the sKL divergence with the dynamic choice of ε given by Eqs. (4) and (7), for different TT ranks $r_{\mathbf{q}}$

3.3.2 Dynamic choice

Next, we similarly investigate the dynamic choice of ε given in Eq. (4), choosing the function f as

$$f(x) = \delta \cdot x \tag{7}$$

with $\delta > 0$. In Theorem 3 we showed that the average of the sKL divergence converges to the KL divergence in the limit of small i.i.d. Gaussian noise. In our particular application, numerical data show that the assumption of i.i.d. Gaussian noise is violated. Nevertheless, as $r_{\mathbf{q}}$ increases and thus the quality of the approximation improves, we observe in Fig. 5 that the sKL divergence converges to the KL divergence, as already mentioned at the end of Sect. 2.3.2. This statement is true for all values of the parameter δ, although the details of the convergence may depend on δ.

We now discuss the results obtained from optimizing the sKL divergence, similar to Sect. 3.3.1. Since optimization using higher-order optimizers is not possible with the dynamic choice, simple gradient descent was used for optimization. The dynamic choice of ε ensures $q_{\theta,i} + \varepsilon_i > 0$, so optimization was always done until a stopping criterion was satisfied. In Table 2 and Fig. 6 we show numerical results for various combinations of δ and $r_{\mathbf{q}}$. As $r_{\mathbf{q}}$ is increased, we again observe a convergence towards the exact result, but now at a much faster rate than for the static choice. Table 2 also shows that the results are quite stable with respect to the parameter δ. For the exact calculation without the TT format, δ played no role, since it is only important when encountering negative entries in $\mathbf{q}_\theta$.

Comparing the static and dynamic choice of ε, it can be seen clearly that dynamically choosing ε generally leads to MHNs that are closer to the exact results at the same level of approximation. This is because usually only a few entries of $\mathbf{q}_\theta$ are negative. Therefore, dynamically choosing ε leads to an objective function that closely resembles the KL divergence, while the static choice introduces a shift for all entries even when the shift is not needed.

Table 2 Average KL divergence (without approximation) from the ground truth model θ_{GT} for MHNs obtained by optimizing the sKL divergence with the dynamic choice of ε given by Eqs. (4) and (7)

	$r_{\mathbf{q}}$					
δ	4	8	16	24	32	exact
10^{-2}	0.320	0.189	0.154	0.092	0.094	0.087
10^{-1}	0.302	0.177	0.112	0.093	0.090	0.087
1	0.258	0.164	0.114	0.093	0.087	0.087
10^{1}	0.263	0.159	0.112	0.091	0.087	0.087
10^{2}	0.256	0.163	0.113	0.093	0.090	0.087
10^{3}	0.280	0.156	0.113	0.096	0.090	0.087

The colors indicate the number of negative entries $q_{\theta,i}$ in the final results, for indices i with $p_i > 0$ (a darker color corresponds to a larger number of negative entries). In the column "exact", the TT format was not used during optimization, thus no negative entries occurred here

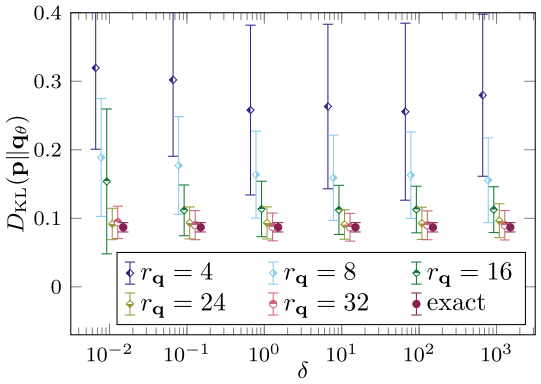

Fig. 6 Visualization of the data in Table 2

4 Summary and Outlook

We have introduced a new method to handle negative entries in approximate probability vectors. This method is very general and can be used in a wide variety of applications (see Sect. 1.2 for examples). Moreover, it does not come with significant computational overhead.

We showed that the sKL divergence shares many desirable properties with the KL divergence. We discussed two possible choices of the shift parameters that occur in the sKL divergence. The static choice allows for the use of higher-order optimizers, but it requires tuning of the parameters and leads to a large loss of gradient information. In contrast, the dynamic choice restricts us to first-order optimizers, but it offers more freedom in the choice of the parameters and preserves most of the gradient information. For the dynamic choice we showed that, when negative entries occur due to i.i.d. Gaussian noise, the difference between the KL diver-

gence and the average sKL divergence is quadratic in the strength of the noise and thus goes to zero for small noise.

In this work, we only considered the sKL divergence in the context of approximate discrete probability distributions. The investigation of possible use cases in other contexts is left for future work.

We applied our method to a real-world application, the modeling of cancer progression by Mutual Hazard Networks, where the use of tensor-train approximations can lead to negative entries. We showed that the sKL divergence and the corresponding model results converge to the KL divergence and the exact model results, respectively, when the parameter that controls the quality of the approximation is increased. We also showed that the dynamic choice of the shift parameters leads to a faster convergence to the exact results than the static choice, as expected from the theoretical considerations in Sect. 2.

When using the sKL divergence as an objective function, first-order optimizers are desirable because they allow for more freedom when choosing the parameters of the sKL divergence. So far, we only used a standard gradient-descent optimizer. In future work, we will investigate the effect of different first-order optimizers, including stochastic and momentum-based optimizers.

Regarding MHNs, it was shown in Georg (2022) that the computational complexity of model construction can be reduced from exponential to cubic in the number of events using low-rank tensor formats. The remaining problem of negative entries in the approximate probability vectors is solved by the current work, which thus enables the construction of large Mutual Hazard Networks with $\gg 30$ events. Expected applications will include as many as 800 events. Furthermore, modifications of the MHN have been conceived which allow for more realistic modeling of tumor progression, but which are currently limited to a very small number of events. We will apply the techniques described in this work to these large and extended MHNs in future work.

Appendix A Proofs

A.1 Proof of Theorem 1

We first prove property (3) by direct calculation,

$$\left.\frac{\mathrm{d}^2 D_{\mathrm{sKL}}(\mathbf{p}\|\mathbf{q})}{\mathrm{d}q_i \mathrm{d}q_j}\right|_{\mathbf{q}=\mathbf{p}} = \frac{\delta_{ij}}{p_i+\varepsilon_i}\,.$$

This means that $\mathrm{d}^2 D_{\mathrm{sKL}}(\mathbf{p}\|\mathbf{q})/\mathrm{d}q_i\mathrm{d}q_j|_{\mathbf{q}=\mathbf{p}}$ are the entries of a diagonal matrix. Since all diagonal entries are positive by assumption, the matrix is positive definite.

Next, we prove that $\mathbf{q}=\mathbf{p}$ is the only extremum of D_{sKL}. The vectors $\mathbf{p}$ and $\mathbf{q}$ obey $\sum_i p_i = \sum_i q_i$. We can use a Lagrange multiplier λ to find the extremal points of

$$\mathcal{L}_{D_{\mathrm{sKL}}}(\mathbf{q},\lambda) = \sum_i (p_i+\varepsilon_i)\log\frac{p_i+\varepsilon_i}{q_i+\varepsilon_i} + \lambda\sum_i (p_i-q_i)\,.$$

By taking derivatives w.r.t. q_i and λ, we find that the only local extremum of $\mathcal{L}_{D_{\mathrm{sKL}}}(\mathbf{q},\lambda)$ is at $\mathbf{q}=\mathbf{p}$ and $\lambda=-1$. Together with property (3) and $D_{\mathrm{sKL}}(\mathbf{p}\|\mathbf{p})=0$, this proves properties (2) and (1).

A.2 Proof of Theorem 2

With $\lambda_1=\lambda$ and $\lambda_2=1-\lambda$, we can write the left-hand side as

$$D_{\mathrm{sKL}}(\lambda_1\mathbf{p}^{(1)}+\lambda_2\mathbf{p}^{(2)}\|\lambda_1\mathbf{q}^{(1)}+\lambda_2\mathbf{q}^{(2)}) = \sum_i\left(\sum_{j=1,2}\lambda_j(p_i^{(j)}+\varepsilon_i)\right)\log\frac{\sum_{j=1,2}\lambda_j(p_i^{(j)}+\varepsilon_i)}{\sum_{j=1,2}\lambda_j(q_i^{(j)}+\varepsilon_i)}\,.$$

For fixed i, we now use the log sum inequality (Cover and Thomas 1991)

$$\left(\sum_j a_j\right)\log\frac{\sum_j a_j}{\sum_j b_j} \le \sum_j a_j\log\frac{a_j}{b_j}$$

for $a_j, b_j > 0$ to complete the proof.

A.3 Proof of Theorem 3

The probability distribution of i.i.d. Gaussian noise $\mathbf{x}$ with mean 0 and standard deviation σ is

$$P(\mathbf{x}) = \prod_i \rho(x_i,\sigma) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x_i^2}{2\sigma^2}\right).$$

The average of the sKL divergence over the Gaussian noise is given by

$$\left\langle D_{\mathrm{sKL}}(\mathbf{p}\|\mathbf{q}+\mathbf{x})\right\rangle_{\mathbf{x}} = \sum_i\int_{-\infty}^{\infty}\mathrm{d}x_i\,\rho(x_i,\sigma)(p_i+\varepsilon_i)\log\frac{p_i+\varepsilon_i}{q_i+x_i+\varepsilon_i}\,. \tag{A1}$$

We therefore consider the integral

$$I = c\int_{-\infty}^{\infty}\mathrm{d}x\,e^{-\frac{x^2}{2\sigma^2}}(p+\varepsilon)\log\frac{p+\varepsilon}{q+x+\varepsilon}\,,$$

where $c=1/\sqrt{2\pi\sigma^2}$ is the normalization factor of the Gaussian. To expand this integral in σ^2, it is convenient to perform a saddle-point analysis (also known as method of steepest

descent or stationary-phase approximation) (Mathews and Walker 1970). This is a useful method to deal with integrals of the form

$$\int \mathrm{d}x\, e^{-Ng(x)} h(x)\,,$$

where N is a large parameter and g and h are functions of x. The saddle-point method yields a systematic expansion of such integrals in powers of $1/N$. In our case, the large parameter is $N = 1/\sigma^2$, and $g(x) = x^2/2$. The remaining part of the integrand is identified with $h(x)$.

We first split the integral I into two contributions I_1 and I_2 corresponding to the integration intervals $(-\infty, -q]$ and $[-q, \infty)$, respectively. For $q + x \geq 0$, Eq. (4) gives $\varepsilon = 0$ so that I_2 takes the form

$$I_2 = c \int_{-q}^{\infty} \mathrm{d}x\, e^{-\frac{x^2}{2\sigma^2}}\, p \log \frac{p}{q+x}\,.$$

The saddle point, i.e., the location of the minimum of $g(x)$, is at $x = 0$, which is within the integration interval. A standard saddle-point analysis up to next-to-leading order then yields

$$I_2 = p \log \frac{p}{q} + \frac{p}{2q^2}\sigma^2 + \mathcal{O}(\sigma^4)\,.$$

In the remaining region, where $q + x < 0$, Eq. (4) gives

$$\varepsilon = |q + x| + f(|q + x|)\,.$$

Substituting $x = -q - z$, we obtain

$$I_1 = c \int_0^{\infty} \mathrm{d}z\, e^{-\frac{(q+z)^2}{2\sigma^2}} (p + z + f(z)) \log \frac{p + z + f(z)}{f(z)}\,.$$

Before we can expand this integral in σ^2, we first need to make sure that it exists. This depends on the choice of the (nonnegative) function f. In particular, I_1 exists if f is a sum of power laws, as assumed in Theorem 3. In fact, I_1 only diverges for rather exotic choices of f that go to zero too rapidly. For example, it diverges at the lower end for $f(z) = \exp(-1/z)$ and at the upper end for $f(z) = \exp(-\exp(z^4))$. In the following we restrict ourselves to functions f for which I_1 exists.

To extract the small-σ behavior of I_1, we again perform a saddle-point analysis. We still have $N = 1/\sigma^2$, but now $g(z) = (q + z)^2/2$. In this case, we encounter two complications that require modifications of the saddle-point method. The first complication arises because the saddle point obtained from minimizing $g(z)$ is at $z = -q$, which is outside the integration interval. Hence the integral is not dominated by the region near a saddle point but by the region near the minimum of $g(z)$ within the integration interval, which is at the lower bound $z = 0$, where $g'(0) = q > 0$. The standard result for this case is obtained by expanding $g(z)$ about zero, which yields

$$\int_0^{\infty} \mathrm{d}z\, e^{-Ng(z)} h(z) = \frac{e^{-Ng(0)} h(0)}{Ng'(0)} \left(1 + \mathcal{O}(1/N)\right) \qquad \text{(A2)}$$

if $h(0)$ is well-defined. Applied to our case we find

$$I_1 = \frac{\sigma}{\sqrt{2\pi} q} e^{-\frac{q^2}{2\sigma^2}} (p + f(0)) \log \frac{p + f(0)}{f(0)} \left(1 + \mathcal{O}(\sigma^2)\right),$$

which is well-defined if $f(0) > 0$. By assumption, f is nonnegative, but if $f(0) = 0$ a second complication arises since $\log f(0)$ is not defined. In this case, let us assume that the leading behavior near $z = 0$ is $f(z) = az^b$ with $a, b > 0$. If we split the logarithm in the integrand of I_1, the integral involving $\log(p + z + f(z))$ is of the form of Eq. (A2) with $h(0) = p \log p$. The integral involving $\log f(z)$ is not covered by Eq. (A2), but it is still dominated by the region near $z = 0$. For this part, we need the integral (valid for $\mathrm{Re}(\alpha) > -1$)

$$\begin{aligned}
&\int_0^{\infty} \mathrm{d}z\, e^{-Ng(z)} z^{\alpha} \log z \\
&\quad \approx e^{-Ng(0)} \int_0^{\infty} \mathrm{d}z\, e^{-Ng'(0)z} z^{\alpha} \log z \\
&\quad = \frac{e^{-Ng(0)} \Gamma(\alpha + 1)}{[Ng'(0)]^{1+\alpha}} \left(\log \frac{1}{Ng'(0)} + \Psi(\alpha + 1) \right),
\end{aligned}$$

where Γ and Ψ denote the gamma and digamma function, respectively, and the $\approx$ sign indicates that we have only kept the leading behavior for large N (corresponding to small σ below). Applying this to our case (with $\alpha = 0$, 1, and b) and collecting terms, we now find the leading behavior for small σ,

$$I_1 \approx \frac{\sigma}{\sqrt{2\pi} q} e^{-\frac{q^2}{2\sigma^2}}\, p \left(\log \frac{p}{a} - b \log \frac{\sigma^2}{q} + b\gamma \right),$$

where $\gamma = -\Psi(1)$ is Euler's constant. Incidentally, our choice of f in Sect. 3.3.2 corresponds to $a = \delta$ and $b = 1$. We conclude that for the functions f we admit, the integral I_1 is suppressed for small σ by a factor of $\exp(-q^2/2\sigma^2)$ and thus goes to zero faster than any power of σ. (This statement remains true if the leading behavior of f near zero is $f(z) = \exp(-a/z^b)$ with $a > 0$ and $b < 1$.) Adding the two integrals we obtain

$$\begin{aligned}
&\left\langle D_{\mathrm{sKL}}(\mathbf{p}\|\mathbf{q}+\mathbf{x})\right\rangle_{\mathbf{x}} \\
&\quad = \sum_i \left(p_i \log \frac{p_i}{q_i} + \frac{p_i}{2q_i^2}\sigma^2 \right) + \mathcal{O}(\sigma^4) \\
&\quad = D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{q}) + \sigma^2 \sum_i \frac{p_i}{2q_i^2} + \mathcal{O}(\sigma^4)\,,
\end{aligned}$$

which completes the proof.

Appendix B Transition-rate matrix as a tensor train

Here we provide a more detailed description of how the transition-rate matrix is stored in the tensor-train format. We start from Eq. (5). We can define 4-dimensional tensors $\tilde{Q}_{\ell k} \in \mathbb{R}^{1\times 2\times 2\times 1}$ through $(\tilde{Q}_{\ell k})_{1ij1} = (Q_{\ell k})_{ij}$ and construct tensor-train operators $\tilde{\mathbf{Q}}_\ell$ with all TT ranks equal to 1,

$$\begin{aligned}
&\tilde{\mathbf{Q}}_\ell(i_1,\ldots,i_d,j_1,\ldots,j_d) \\
&\qquad = \sum_{\alpha_0,\ldots,\alpha_d} \prod_{k=1}^{d} \tilde{Q}_{\ell k}(\alpha_{k-1},i_k,j_k,\alpha_k)\,.
\end{aligned}$$

The full transition-rate matrix is now obtained in the TT format by summing up the tensor trains $\tilde{\mathbf{Q}}_\ell$, giving

$$\begin{aligned}
&\tilde{\mathbf{Q}}(i_1,\ldots,i_d,j_1,\ldots,j_d) \\
&\qquad = \sum_{\ell=1}^{d} \tilde{\mathbf{Q}}_\ell(i_1,\ldots,i_d,j_1,\ldots,j_d)\,.
\end{aligned}$$

This sum can be performed in the format without approximation, resulting in a tensor train with all TT ranks equal to d (except for $r_0 = r_d = 1$) (Hackbusch 2019; Oseledets 2011).

Acknowledgements This work is supported in part by the German Research Foundation (DFG) through the grants "Tensorapproximationsmethoden zur Modellierung von Tumorprogression"(project 458051812), "PUNCH4NFDI—Teilchen, Universum, Kerne und Hadronen für die NFDI"(project 460248186), and "Striking a moving target: From mechanisms of metastatic organ colonization to novel systemic therapies"(TRR 305). We would like to thank Alexander Rothkopf and Phiala Shanahan for a stimulating discussion.

Author contributions S.P., P.G. and T.W. worked out the initial idea. S.P. and P.G. wrote the implementation. R.Sc., M.K., L.G., and R.Sp. contributed ideas to improve the method and to provide a theoretical basis. S.P. and T.W. wrote the paper. All authors reviewed and improved upon the draft.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availibility No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors have no Conflict of interest to declare that are relevant to the content of this article.

References

Amari, S.-I.: Information geometry and its applications. Springer, Tokyo (2016). https://doi.org/10.1007/978-4-431-55978-8

Alkofer, R., Smekal, L.: The infrared behaviour of QCD Green's functions: confinement, dynamical symmetry breaking, and hadrons as relativistic bound states. Phys. Rep. **353**(5), 281 (2001). https://doi.org/10.1016/S0370-1573(01)00010-2

Basseville, M.: Divergence measures for statistical data processing—an annotated bibliography. Signal Process. **93**(4), 621 (2013). https://doi.org/10.1016/j.sigpro.2012.09.003

Burnier, Y., Laine, M., Mether, L.: A test on ananlytic continuation of thermal imaginary-time data. Eur. Phys. J. C **71**(4), 1619 (2011). https://doi.org/10.1140/epjc/s10052-011-1619-0

Csilléry, K., Blum, M.G.B., Gaggiotti, O.E., François, O.: Approximate Bayesian computation (ABC) in practice. Trends Ecol. Evol. **25**(7), 410–418 (2010). https://doi.org/10.1016/j.tree.2010.04.001

Chen, J.: Time hazard networks: incorporating temporal difference for oncogenetic analysis. PLoS ONE **18**(3), 1 (2023). https://doi.org/10.1371/journal.pone.0283004

Chi, E.C., Kolda, T.G.: On tensors, sparsity, and nonnegative factorizations. SIAM J. Matrix Anal. Appl. **33**(4), 1272 (2012). https://doi.org/10.1137/110859063

Cox, D.R.: Regression models and life-tables. J. Roy. Stat. Soc.: Ser. B (Methodol.) **34**(2), 187 (1972). https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Cover, T.M., Thomas, J.A.: Elements of information theory. Wiley, Hoboken (New Jersey) (1991). https://doi.org/10.1002/0471200611

Dolgov, S.V., Savostyanov, D.V.: Alternating minimal energy methods for linear systems in higher dimensions. SIAM J. Sci. Comput. **36**(5), 2248 (2014). https://doi.org/10.1137/140953289

Fletcher, R.: Practical methods of optimization. Wiley, Chichester (2000). https://doi.org/10.1002/9781118723203

Georg, P.: Tensor Train Decomposition for solving high-dimensional Mutual Hazard Networks. PhD thesis, University of Regensburg (2022). https://epub.uni-regensburg.de/53004

Georg, P., Grasedyck, L., Klever, M., Schill, R., Spang, R., Wettig, T.: Low-rank tensor methods for Markov chains with applications to tumor progression models. J. Math. Biol. **86**(1), 7 (2022). https://doi.org/10.1007/s00285-022-01846-9

Grasedyck, L., Kressner, D., Tobler, C.: A literature survey of low-rank tensor approximation techinques. GAMM-Mitteilungen **36**(1), 53 (2013). https://doi.org/10.1002/gamm.201310004

Hackbusch, W.: Tensor spaces and numerical tensor calculus. Springer series in computational mathematics, vol. 57. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-35554-8

Haas, M., Fister, L., Pawlowski, J.M.: Gluon spectral functions and transport coefficients in Yang–Mills theory. Phys. Rev. D **90**(9), 091501 (2014). https://doi.org/10.1103/PhysRevD.90.091501

Hobson, M.P., Lasenby, A.N.: The entropic prior for distributions with positive and negative values. Mon. Not. R. Astron. Soc. **298**(3), 905–908 (1998). https://doi.org/10.1046/j.1365-8711.1998.01707.x

Hoch, J.C.: Nonuniform sampling and maximum entropy reconstruction in multidimensional NMR. Acc. Chem. Res. **47**(2), 708 (2014). https://doi.org/10.1021/ar400244v

Hansen, S., Plantenga, T., Kolda, T.G.: Newton-based optimization for Kullback–Leibler nonnegative tensor factorizations. Optim. Methods Softw. **30**(5), 1002 (2015). https://doi.org/10.1080/10556788.2015.1009977

Holtz, S., Rohwedder, T., Schneider, R.: The alternating linear scheme for tensor optimization in the tensor train format. SIAM J. Sci. Comput. **34**(2), 683 (2012). https://doi.org/10.1137/100818893

Ha, W., Sidky, E.Y., Barber, R.F., Schmidt, T.G., Pan, X.: Estimating the spectrum in computed tomography via Kullback–Leibler divergence constrained optimization. Med. Phys. **46**(1), 81 (2019). https://doi.org/10.1002/mp.13257

Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2017). arxiv.org/abs/1412.6980

Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**(1), 79 (1951). https://doi.org/10.1214/aoms/1177729694

Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2022). arxiv.org/abs/1312.6114

Luo, X.G., Kuipers, J., Beerenwinkel, N.: Joint inference of exclusivity patterns and recurrent trajectories from tumor mutation trees. Nat. Commun. **14**(1), 3676 (2023). https://doi.org/10.1038/s41467-023-39400-w

Lee, N., Phan, A.-H., Cong, F., Cichocki, A.: Nonnegative Tensor train decomposition for multi-domain feature extraction and clustering. In: Hirose, A., Ozawa, S., Doya, K., Ikeda, K., Lee, M., Liu, D. (eds.) Neural Information Processing, p. 87. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46675-0_10

Michor, F., Iwasa, Y., Nowak, M.A.: Dynamics of cancer progression. Nat. Rev. Cancer **4**(3), 197 (2004). https://doi.org/10.1038/nrc1295

Mathews, J., Walker, R.L.: Mathematical methods of physics. Addison-Wesley, New York (1970)

Oseledets, I.V.: Tensor-Train Decomposition. SIAM J. Sci. Comput. **33**(5), 2295 (2011). https://doi.org/10.1137/090752286

Philippe, B., Saad, Y., Stewart, W.J.: Numerical methods in markov chain modeling. Oper. Res. **40**(6), 1156 (1992). https://doi.org/10.1287/opre.40.6.1156

Paatero, P., Tapper, U.: Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics **5**(2), 111 (1994). https://doi.org/10.1002/env.3170050203

Rothkopf, A.: Bayesian inference of nonpositive spectral functions in quantum field theory. Phys. Rev. D **95**(5), 056016 (2017). https://doi.org/10.1103/physrevd.95.056016

Schill, R., Solbrig, S., Wettig, T., Spang, R.: Modelling cancer progression using Mutual Hazard Networks. Bioinformatics **36**(1), 241 (2019). https://doi.org/10.1093/bioinformatics/btz513

Thomas, P., Grima, R.: Approximate probability distributions of the master equation. Phys. Rev. E **92**, 012120 (2015). https://doi.org/10.1103/PhysRevE.92.012120

van Erven, T., Harremos, P.: Rényi divergence and Kullback–Leibler divergence. IEEE Trans. Inf. Theory **60**(7), 3797 (2014). https://doi.org/10.1109/TIT.2014.2320500

Welling, M., Weber, M.: Positive tensor factorization. Pattern Recognition Lett. **22**(12), 1255 (2001). https://doi.org/10.1016/S0167-8655(01)00070-8