

# Instantons for rare events in heavy-tailed distributions

Mnerh Alqahtani\*  and Tobias Grafke 

Mathematics Institute, University of Warwick, Coventry CV4 7AL, United Kingdom

E-mail: [M.Alqahtani@warwick.ac.uk](mailto:M.Alqahtani@warwick.ac.uk)

Received 14 December 2020, revised 9 February 2021

Accepted for publication 15 February 2021

Published 6 April 2021



CrossMark

## Abstract

Large deviation theory and instanton calculus for stochastic systems are widely used to gain insight into the evolution and probability of rare events. At its core lies the fact that rare events are, under the right circumstances, dominated by their least unlikely realization. Their computation through a saddle-point approximation of the path integral for the corresponding stochastic field theory then reduces an inefficient stochastic sampling problem into a deterministic optimization problem: finding the path of smallest action, the instanton. In the presence of heavy tails, though, standard algorithms to compute the instanton critically fail to converge. The reason for this failure is the divergence of the scaled cumulant generating function (CGF) due to a non-convex large deviation rate function. We propose a solution to this problem by ‘convexifying’ the rate function through a nonlinear reparametrization of the observable, which allows us to compute instantons even in the presence of super-exponential or algebraic tail decay. The approach is generalizable to other situations where the existence of the CGF is required, such as exponential tilting in importance sampling for Monte-Carlo algorithms. We demonstrate the proposed formalism by applying it to rare events in several stochastic systems with heavy tails, including extreme power spikes in fiber optics induced by soliton formation.

Keywords: large deviation principle, exponentially tilted measures, nonconvex rate functions, nonlinear reparametrizations, instanton equations

(Some figures may appear in colour only in the online journal)

\*Author to whom any correspondence should be addressed.



Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](https://creativecommons.org/licenses/by/4.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

## 1. Introduction

In many situations of physical relevance, rare events are tremendously important despite their infrequent occurrence: heat waves, stock market crashes, or earth quakes all occur with small probability but devastating consequences. Unfortunately, due to their rareness, these events are hard to observe in experiment or numerical simulation, and require special treatment. Rare event algorithms [1] are typically based on one of the two following ideas: either to increase the rate of occurrence of the rare event by biasing the underlying system (importance sampling), or to substitute all possible ways of observing a rare event by its most common realization (large deviations/instanton theory). Under the hood both are connected to the *exponentially tilted* measure and the *cumulant generating function* (CGF). As we will see, when naively implementing standard schemes, both become ill-defined when the underlying probability densities become heavy-tailed. Examples of fat tailed distributions are ubiquitous in physical systems of relevance, such as the energy dissipation in fluid turbulence and the phenomenon of intermittency [2], the momentum of atoms in optical lattices [3], wealth distributions in economies [4, 5] or stock price changes in finance [6, 7].

Here, we focus on numerical algorithms connected to instanton theory and its rigorous counterpart, large deviation theory (LDT), to recover the tails of probability distributions in a stochastic system. A large deviation principle (LDP) states that the probability of rare events decays exponentially, and its exponential scaling is given by the minima of the corresponding rate function  $I$  [8, 9]. Roughly speaking, let  $P^\varepsilon$  be a family of probability measures on a suitable measurable space  $(\mathcal{X}, \Sigma)$ . We say  $P^\varepsilon$  satisfies an LDP with rate function  $I : \mathcal{X} \rightarrow \mathbb{R}$  if for all subsets  $\Omega \subset \Sigma$ , we have [10],

$$P^\varepsilon(\Omega) \asymp \exp\left(-\varepsilon^{-1} \inf_{x \in \Omega} I(x)\right), \quad (1)$$

where  $\asymp$  denotes log-asymptotic equivalence in the limit  $\varepsilon \rightarrow 0$ .

In a physical sense, the probability  $P^\varepsilon(\Omega)$  can formally be written as a path integral, and the estimate (1) becomes a saddle point approximation or Laplace method. In LDT, the *Gärtner–Ellis theorem* [11] provides a direct formula for the rate function  $I$ . Roughly speaking, if the limiting behavior of a scaled CGF is well-defined, then its *Legendre–Fenchel* (LF) transform is the rate function of the LDP of the process under consideration. The LF transform of a real-valued function  $f(x)$  defined on  $\mathbb{R}^n$  is defined as

$$f^*(y) = \sup_{x \in \mathbb{R}^n} (\langle x, y \rangle - f(x)), \quad (2)$$

where  $\langle x, y \rangle = x^T y$  is the inner product on  $\mathbb{R}^n$ . Let  $z^\varepsilon$  be a sequence of random variables in  $\mathbb{R}^n$ , with probability measures  $P^\varepsilon$ , and assume that its scaled CGF, defined as the limit,

$$G(\lambda) \equiv \lim_{\varepsilon \rightarrow 0} \varepsilon \log \mathbb{E} \left[ e^{\varepsilon^{-1} \langle \lambda, z^\varepsilon \rangle} \right], \quad (3)$$

exists for each  $\lambda \in \mathbb{R}^n$  and is differentiable in  $\lambda$ . Then, the Gärtner–Ellis theorem states that the family of probability measures  $\{P^\varepsilon\}$  satisfies an LDP, where the rate function  $I$  is the LF transform of  $G$ , i.e.  $I = G^*$ .

Intuitively, one can interpret  $\lambda$  as a Lagrange multiplier to condition on an outcome  $z = \lim_{\varepsilon \rightarrow 0} z^\varepsilon$ . Crucially, though, if the probability measure  $P^\varepsilon$  is super-exponential, or has even heavier tails, the expectation in equation (3) diverges and the CGF is no longer defined. Notably this does not mean that the corresponding rare events are special in any way, but merely that

the duality between the parameter  $\lambda$  and rare event observable  $z$  is broken. As a consequence, no tilt exists to realize an outcome  $z$ , and standard rare events algorithms fail.

In what follows, we will show how a *nonlinear tilt* allows to modify the connection between tilt  $\lambda$  and outcome  $z$ , so that heavy tails can be probed regardless of the non-convexity of their rate function. In particular, we will suggest a numerical algorithm to estimate the tail scaling for heavy-tailed probability densities by computing the corresponding instanton. To our knowledge, this procedure has not been described in the literature before. We note, though, that the duality between the CGF and the rate function is of central importance in statistical mechanics to describe the connection between free energy and entropy. Concretely, in the thermodynamic limit, the canonical ensemble and the microcanonical ensemble are not equivalent at nonconcave regions of a corresponding microcanonical entropy function. To overcome this issue, a generalized canonical ensemble that is universally equivalent to the microcanonical ensemble, regardless of the properties of the entropy, can be established by a penalty function of the exponent of Boltzmann distribution of the microstates of the system. Examples are the so-called *Gaussian ensemble* [12, 13], or the *Betrag ensemble* [14], which in spirit are similar to a convexification of the observable as proposed in this paper.

Here, we will concentrate specifically on the case of small noise sample-path large deviations for stochastic differential equations (SDEs), which will be introduced in section 2. We focus on the numerical computation of the instanton in section 2.1, and highlight the connection to a change of measure in path-space in section 2.2. We demonstrate the problem in the heavy-tailed case by reviewing the convex analysis for the Gärtner–Ellis theorem in section 3, and propose a solution modifying the instanton computation to yield finite outcomes for heavy tails and non-convex rate functions in section 3.1. To demonstrate the applicability of our approach, we show several examples of instantons for heavy-tailed distributions in section 4: toy models with super-exponential (section 4.1) and powerlaw tails (section 4.2), and a banana-shaped potential (section 4.3), and finally high-amplitude events in fiber-optics described by the focusing nonlinear Schrödinger (NLS) equation (section 4.4). We summarize our findings in section 5.

## 2. Instantons and Freidlin–Wentzell theory

Consider a stochastic system,

$$dX_t^\varepsilon = b(X_t^\varepsilon) dt + \sqrt{\varepsilon} \sigma dW_t, \quad X_{t_0}^\varepsilon = x_0, \quad (4)$$

where  $X_t^\varepsilon \in \mathbb{R}^n$  is a family of random processes indexed by the noise strength  $\varepsilon$ . The deterministic drift  $b: \mathbb{R}^n \rightarrow \mathbb{R}^n$  satisfies the Lipschitz condition,  $dW_t$  is an  $n$ -dimensional Brownian increment, and the noise covariance  $\chi = \sigma \sigma^T$  is assumed to be invertible for  $\sigma \in \mathbb{R}^{n \times n}$ . Intuitively, equation (4) describes the temporal evolution of a system perturbed by stochasticity, where we later assume the fluctuations to be small,  $\varepsilon \ll 1$ . This situation is ubiquitous in many application areas, where for example  $\varepsilon$  plays the role of the temperature in a chemical reaction, or the inverse number of particles in a thermodynamic system.

With vanishing noise,  $\varepsilon = 0$ , the solution  $x \in \mathbb{R}^n$  of the unperturbed (deterministic) system

$$\dot{x}(t) = b(x(t)), \quad x(t_0) = x_0, \quad (5)$$

converges to one of its attractors for long times. For example, consider a point attractor, or *asymptotically stable fixed points*,  $\bar{x} \in \mathbb{R}^n$ , with basin of attraction  $B$ , such that  $x(t) \rightarrow \bar{x}$  for  $t \rightarrow \infty$  for all initial conditions  $x_0$  in  $B$ . Solutions of the stochastic system (4) converge to

solutions of the deterministic system (5) in probability,  $P(\lim_{\varepsilon \rightarrow 0} \max_{t_0 \leq t \leq t_1} |X_t^\varepsilon - x(t)| = 0) = 1$  [10]. This is an instance of the law of large numbers, stating that for small noise and large times we expect solutions of the stochastic system to end up near the attractors of the deterministic one.

Nevertheless, for any non-zero  $\varepsilon \ll 1$  there is a small but non-vanishing probability of finding the system far from the attractor. This can only happen if the noise conspires in just the right way to overcome the deterministic dynamics, and is consequently a rare event. Concretely, consider any domain  $D \subset \mathbb{R}^n$  attracted to  $\bar{x}$ , i.e.  $D \subset B$ . We are interested in the chance of trajectories  $X_t^\varepsilon$  departing from  $\bar{x}$  and eventually leaving  $D$ . These trajectories belong to the set

$$A_z := \{ \varphi \in \mathbf{C}_{t_0 t_1}(\mathbb{R}^n) \mid \varphi(t_0) = \bar{x}, \varphi(t_1) = z \notin D \}, \tag{6}$$

and we want to quantify the probability

$$p(z) = P[X^\varepsilon \in A_z] \quad \text{as } \varepsilon \rightarrow 0, \tag{7}$$

which is a question about the probability of *large deviations*. Under the stated conditions, there is a trajectory  $\varphi^* \in A_z$  such that the probability measure over  $A_z$  accumulates near  $\varphi^*$  for  $\varepsilon \rightarrow 0$ , namely if  $N(\varphi^*)$  is any neighborhood of  $\varphi^*$ ,

$$\lim_{\varepsilon \rightarrow 0} \frac{P[X^\varepsilon \in A_z \setminus N(\varphi^*)]}{P[X^\varepsilon \in N(\varphi^*)]} = 0. \tag{8}$$

In other words, in the small noise limit we will almost surely find our sample trajectory close to  $\varphi^*$ , such that  $\max_{t_0 \leq t \leq t_1} |X_t^\varepsilon - \varphi_t^*| \leq \delta$ , for an arbitrary small  $\delta$ .

In order to find this most likely trajectory  $\varphi^*$ , Freidlin–Wentzell theory [10] states that  $\varphi^*$  is actually the minimizer of large deviation *rate function*  $S(\varphi)$  associated with the stochastic system (4), given by

$$S(\varphi) = \frac{1}{2} \int_{t_0}^{t_1} \|\dot{\varphi}_t - b(\varphi(t))\|_\chi^2 dt, \tag{9}$$

where the integral exists, and  $S(\varphi) = \infty$  otherwise. The norm  $\|f\|_\chi^2 = \langle f, \chi^{-1} f \rangle$  is induced by the noise covariance  $\chi$ . With this rate function we can quantify the probability (7) of departing the domain  $D$  as

$$\lim_{\varepsilon \rightarrow 0} \varepsilon \log p(z) = -I(z) = -S(\varphi^*), \tag{10}$$

where

$$\varphi^* = \arg \min_{\varphi \in A_z} S(\varphi). \tag{11}$$

The problem of finding the rare event probability is now reduced to finding the minimizer  $\varphi^*$ .

In analogy to the principle of least action in classical mechanics or quantum mechanics, the rate function is often termed *action*, and the corresponding minimizer  $\varphi^*$  is called *instanton*. The integrand of  $S$  can be understood as a *Lagrangian*,

$$L(\varphi, \dot{\varphi}) = \frac{1}{2} \|\dot{\varphi}_t - b(\varphi)\|_\chi^2, \tag{12}$$

so that the maximum likelihood pathways leaving the attractors of (4) correspond to semi-classical trajectories of the field theory defined by  $L$ .

2.1. Instanton equations and large deviation Hamiltonian

It is helpful, both to increase understanding, and to simplify the numerical implementation, to rephrase the optimization problem (11) into the corresponding Hamiltonian formulation. To this end, we introduce the large deviation *Hamiltonian*  $H(\varphi, \vartheta)$  as the Legendre transform of the Lagrangian  $L(\varphi, \dot{\varphi})$ ,

$$H(\varphi, \vartheta) = \sup_{\dot{\varphi}} (\langle \vartheta, \dot{\varphi} \rangle - L(\varphi, \dot{\varphi})), \tag{13}$$

which, for the Lagrangian (12), corresponds to

$$H(\varphi, \vartheta) = \langle b(\varphi), \vartheta \rangle + \frac{1}{2} \langle \vartheta, \chi \vartheta \rangle. \tag{14}$$

Here  $\vartheta = \partial L / \partial \dot{\varphi}$  is the conjugate momentum of  $\varphi$  [15]. Now, the minimizer  $\varphi^*$  can also be expressed as the solution of Hamilton’s equations,

$$\begin{aligned} \dot{\varphi} &= \partial_{\vartheta} H(\varphi, \vartheta) = b(\varphi) + \chi \vartheta, \\ \dot{\vartheta} &= -\partial_{\varphi} H(\varphi, \vartheta) = -(\nabla_{\varphi} b(\varphi))^T \vartheta. \end{aligned} \tag{15}$$

with boundary conditions,

$$\varphi(t_0) = \bar{x}, \quad \varphi(t_1) = z. \tag{16}$$

Equations (15) and (16) are often termed *instanton equations*.

The fact that we are looking only for trajectories that will eventually leave the attractor,  $\varphi^* \in A_z$ , implies that the optimization problem (11) is a constrained one, i.e. we are looking only for solutions of the instanton equations conditioned on the endpoint  $z$ . Practically, this constrained optimization problem can be transformed into an unconstrained one,

$$\varphi^* = \arg \min_{\varphi \in C_{t_0 t_1}(\mathbb{R}^n)} (S(\varphi) - \langle \lambda, \varphi(t_1) - z \rangle), \tag{17}$$

by using a Lagrange multiplier  $\lambda \in \mathbb{R}^n$  to enforce the final constraint [16]. Note that the variation of this unconstrained action,

$$[\delta S(\varphi) - \langle \lambda, \delta \varphi(t_1) \rangle]_{\varphi = \varphi^*} = 0, \tag{18}$$

results in the same instanton equations, (15), but with different boundary conditions

$$\varphi(t_0) = 0, \quad \vartheta(t_1) = \lambda. \tag{19}$$

A more detailed derivation of the boundary conditions (16) and (19) is given in appendix A.1. We can solve the instanton equation (15) iteratively with these latter conditions, by solving the  $\varphi$ -equation forward in time, and using the result to solve the  $\vartheta$ -equation backward in time, until convergence [17, 18]. Note that this choice of temporal direction of integration is not only the one suggested by the boundary conditions, but is further the numerically stable choice of direction for the drifts  $b$  and  $\nabla b^T$ .

As we will see below, the mapping between Lagrange multipliers  $\lambda$  and final points  $z$  is nontrivial, and it is not clear *a priori* how to choose the correct  $\lambda$  to obtain a final configuration  $z = \varphi^*(t_1)$ . If we are interested in  $p^{\varepsilon}(z)$  for a whole range of  $z$ , we can instead choose to simply solve the instanton equations for a range of  $\lambda$  to cover a range of  $z$  without specifically needing to know the duality mapping  $\lambda(z)$ . Exactly this procedure is often used in the

literature to work out probability distributions of stochastic systems, from Burgers [17, 19], plasma turbulence [20], or Ginzburg–Landau [21] equations to the Kardar–Parisi–Zhang [22] and Kipnis–Marchioro–Presutti model [23].

Crucially, however, the existence of a corresponding dual  $\lambda$  for a given final point  $z = \varphi^*(t_1)$  is not necessarily guaranteed, as the next section clarifies. As a consequence, the above methodology might fail, in particular in situations with heavy tails.

## 2.2. Exponentially tilted measures

Interestingly, there is a probabilistic interpretation of the introduction of the Lagrange multiplier  $\lambda$  to the optimization problem (18) in the form of the *exponentially tilted measure* [24, 25], as for example employed in importance sampling for Monte-Carlo (MC) estimators [1]. Intuitively, by the procedure of *tilting*, one replaces the original random process (4) by a modified one, under which the rare events under consideration become more likely, while correcting for this modification *a posteriori* when computing their probability. As a consequence, with tilting, the rare event probability, or expectations over its realizations, can be estimated more efficiently, and with possibly smaller variance.

To be more precise, we denote by  $p_\lambda$  the measure exponentially tilted towards the outcome  $z$ , defined for our purposes as

$$p_\lambda(z) = \frac{\exp(\varepsilon^{-1} \langle \lambda, z \rangle)}{\mathbb{E}_p[\exp(\varepsilon^{-1} \langle \lambda, z \rangle)]} p(z), \quad (20)$$

where  $\mathbb{E}_p[\cdot]$  is the expectation under the original measure  $p$ . In equation (20), the probability measure  $p_\lambda$  of events resulting in  $z$ , i.e. trajectories in  $A_z$  (6), have been awarded extra weight by the *Radon–Nikodym derivative* of  $p_\lambda$  with respect to  $p$ , which is:

$$\frac{p_\lambda(z)}{p(z)} = \frac{dp_\lambda(z)}{dp(z)} = \frac{\exp(\varepsilon^{-1} \langle \lambda, z \rangle)}{\mathbb{E}_p[\exp(\varepsilon^{-1} \langle \lambda, z \rangle)]}. \quad (21)$$

Rearranging equation (20) gives:

$$p_\lambda(z) = \exp(\varepsilon^{-1} \{\langle \lambda, z \rangle - G(\lambda)\}) p(z), \quad (22)$$

where

$$G(\lambda) = \varepsilon \log \mathbb{E}_p[\exp(\varepsilon^{-1} \langle \lambda, z \rangle)], \quad (23)$$

is the scaled CGF,  $G: \mathbb{R}^n \rightarrow \mathbb{R}$ . This change of measure is optimal, in the sense of maximizing the tilted probability  $p_\lambda$ , at the choice  $\lambda = \lambda_z$  with

$$\lambda_z = \arg \max_{\lambda \in \mathbb{R}^n} \{\langle \lambda, z \rangle - G(\lambda)\}. \quad (24)$$

This can be seen, given the Gärtner–Ellis theorem,  $I(z) = G(\lambda)^* = \sup_{\lambda \in \mathbb{R}^n} (\langle \lambda, z \rangle - G(\lambda))$ , by realizing that

$$\begin{aligned} \log p_\lambda(z) &= \varepsilon^{-1} (\langle \lambda_z, z \rangle - G(\lambda_z)) + \log p(z) \\ &= \varepsilon^{-1} \sup_{\lambda \in \mathbb{R}^n} \{\langle \lambda, z \rangle - G(\lambda)\} + \log p(z) \\ &= \varepsilon^{-1} I(z) + \log p(z) \xrightarrow{\varepsilon \rightarrow 0} 0, \end{aligned} \quad (25)$$

where the last line is just the definition of the LDP,  $\varepsilon \log p(z) = -I(z)$  for  $\varepsilon \rightarrow 0$ .

It is in this sense that the optimal tilting parameter of the end-point distribution corresponds to the Lagrange multiplier in the instanton equations constraining the endpoint to  $z$ .

### 3. Convex analysis and the Gärtner–Ellis theorem

In order to use the described methodology to find the instanton for a rare outcome  $z$ , or equivalently make sense of the corresponding exponentially tilted measure  $p_\lambda(z)$ , we must demand that the mapping  $z \rightarrow \lambda(z)$  is a bijection: for every outcome  $z$  there must be a unique tilt  $\lambda$  such that the instanton  $\varphi$ , solution of (15) with boundary conditions (19), is a unique solution with  $\varphi(t_1) = z$ . If that is the case then we can estimate the probability

$$p(z) \asymp \exp(-\varepsilon^{-1}S(\varphi^*)) = \exp(-\varepsilon^{-1}I(z)). \tag{26}$$

The precise properties of the duality mapping between tilting parameter  $\lambda$  and outcome  $z$  can be understood by the interplay between the Gärtner–Ellis theorem and convex analysis. We have,

$$\begin{aligned} G(\lambda) &= \sup_{z \in \mathbb{R}^n} (\langle \lambda, z \rangle - I(z)) \\ &= \langle \lambda, z(\lambda) \rangle - I(z(\lambda)), \end{aligned} \tag{27}$$

where the solution  $z(\lambda)$  of the form

$$\nabla I(z) = \lambda, \tag{28}$$

does only hold when the rate function is strictly convex. If instead the rate function is not strictly convex (i.e. has concave, and/or affine linear regions or is even just asymptotically linear), the LF transform is applied only to the region at which  $I(z)$  admits *supporting hyperplanes*. If there exists  $\lambda \in \mathbb{R}^n$  such that [26],

$$I(y) \geq I(z) + \lambda(y - z), \quad \forall y \in \mathbb{R}^n, \tag{29}$$

then we say  $I$  admits a supporting hyperplane at  $z$ , where the slope of the supporting hyperplane is  $\lambda$ . In this sense, we can define non-convex regions to be the ones that do not admit any supporting hyperplane, so do not have any corresponding  $\lambda$ . Note that the absence of these hyperplanes can affect the LF transform  $I^*(z) = G(\lambda)$  in two different ways,

**Case I:** having an asymptotically linear part of  $I(z)$  leads to a divergent LF transform  $G(\lambda)$ .

**Case II:** having a concave or affine linear part of  $I(z)$  leads to an existent but nondifferentiable

LF transform  $G(\lambda)$ .

Both cases will be discussed specifically in the applications in section 4.

Assuming for now there are supporting hyperplanes (i.e. existent  $\lambda$ ) for all  $z \in \mathbb{R}^n$ , then equation (28) leads to,

$$z(\lambda) = (\nabla I)^{-1}(\lambda), \tag{30}$$

i.e.  $\nabla I$  must be invertible for  $z(\lambda)$  to be so. Up to a choice of sign, this implies that  $\nabla I$  is *strictly monotonically increasing* (SMI), which is equivalent to  $I$  being a strictly convex function [24]. Also note that if  $z(\lambda)$  is invertible, this implies that  $G(\lambda)$  is a differentiable function, since equation (27) gives,

$$\begin{aligned} \nabla G(\lambda) &= z(\lambda) + (\nabla z(\lambda))^T \lambda - (\nabla z(\lambda))^T \nabla I(z(\lambda)) \\ &= z(\lambda), \end{aligned} \tag{31}$$

where equation (28) is used. What we have demonstrated above is nothing but the well-known fact that the LF transform of a convex, differentiable function  $G(\lambda)$  is strictly convex. This perspective, though, makes it clear that the existence of a tilting parameter (Lagrange multiplier)  $\lambda$  to enforce an outcome  $z$  depends on the finiteness and differentiability of the scaled CGF. In other words, both exponential tilting, and finding a Lagrange parameter to constrain the endpoint to  $z$ , depends on the rate function being strictly convex. Only if this is the case, it is possible to compute a unique instanton in order to estimate  $I(z)$  and therefore  $p(z)$ , (26).

Since in the low noise limit we have  $p(z) \sim \exp(-\varepsilon^{-1}I(z))$ , it is easy to construct cases where the rate function  $I(z)$  is not strictly convex. In fact, every situation where the tails of  $p(z)$  are *fat*, i.e. exponential ( $I(z) \sim z$ ), stretched exponential ( $I(z) \sim z^\alpha, \alpha < 1$ ), or even algebraic ( $I(z) \sim \alpha \log(z), \alpha < 0$ ) tails will break the above assumption.

The main contribution of this paper is the realization that the introduction of a nonlinear map  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  allows us to loosen the restriction of the convexity of  $I(z)$ . The idea is to define a *nonlinear tilt* through  $F$  via  $\exp(\varepsilon^{-1}\langle \lambda, F(z) \rangle)$ , such that the map  $\lambda \rightarrow z(\lambda)$  is replaced by a new map  $\lambda \rightarrow F \circ z(\lambda)$ . We are free to choose any appropriate  $F$ . As we will see next, this allows us to suitably reparametrize the space of outcomes, so that the *effective* rate function  $I \circ F^{-1}$  is strictly convex.

### 3.1. Nonlinear tilt

In analogy to equation (22) and the description in sections 2.2 and 3, we can now define the *nonlinearly tilted measure*

$$\begin{aligned} p_\lambda^F(z) &= \frac{\exp(\varepsilon^{-1}\langle \lambda, F(z) \rangle)}{\mathbb{E}_p \exp(\varepsilon^{-1}\langle \lambda, F(z) \rangle)} p(z) \\ &= \exp(\varepsilon^{-1}(\langle \lambda, F(z) \rangle - G_F(\lambda))) p(z), \end{aligned} \tag{32}$$

where the nonlinearly tilted CGF is given by

$$\begin{aligned} G_F(\lambda) &= \sup_{z \in \mathbb{R}^n} (\langle \lambda, F(z) \rangle - I(z)) \\ &= \langle \lambda, F(z(\lambda)) \rangle - I(z(\lambda)), \end{aligned} \tag{33}$$

(compare equation (27)) which is assumed to be finite and differentiable. Its gradient fulfills

$$\begin{aligned} \nabla G_F(\lambda) &= F(z(\lambda)) + (\nabla z(\lambda))^T (\nabla F(z(\lambda)))^T \lambda - (\nabla z(\lambda))^T \nabla I(z(\lambda)) \\ &= F(z(\lambda)), \end{aligned} \tag{34}$$

where the last equality is due to  $z(\lambda)$  being the solution of  $\nabla I(z) = \lambda^T \nabla F(z)$  in equation (33).

This proposed remapping can be chosen to overcome the above problem by creating a new function  $G_F(\lambda)$ , which plays the role of the CGF, while simultaneously being a bounded and differentiable function. At the same time,  $I \circ F^{-1}(y)$  can be understood as the effective rate function, since equation (33) can be written as,

$$G_F(\lambda) = \sup_{F^{-1}(y) \in \mathbb{R}^n} (\langle \lambda, y \rangle - I \circ F^{-1}(y)). \tag{35}$$



Obviously, the right choice of  $F$  depends on the nature of the tail scaling at hand. We will derive the necessary properties of  $F(\cdot)$  next.

### 3.2. Properties of the reparametrization and the nonlinearly tilted instanton

In the following, we denote by  $y \in \mathbb{R}^n$  the reparametrized outcome,  $y = F(z)$ . Our goal is to choose  $F$  such that  $F \circ z(\lambda) = y(\lambda)$  is a bijection. From above, it is clear that

$$\lambda^T = \nabla I(z)(\nabla F(z))^{-1}. \quad (36)$$

Following the same argument as in section 3,  $y(\lambda)$  is bijective if

- $F$  is a diffeomorphism, and
- $I \circ F^{-1}(y)$  is strictly convex, i.e.

$$\langle v, \text{Hess}(I \circ F^{-1})(y) v \rangle > 0 \quad \forall v \in \mathbb{R}^n. \quad (37)$$

Assuming these conditions on  $F$  implies that the gradient of the reparametrized rate function  $I \circ F^{-1}(y)$ , given by  $\lambda \circ F^{-1}(y)$ , is an SMI function, implying that it is invertible. The desired bijective mapping then becomes  $\lambda \rightarrow F(z(\lambda)) = \nabla G_F(\lambda)$  (compare equation (34)). For a non-convex  $I$ , intuitively,  $F$  must be chosen to suitably reparametrize the space of outcomes for the effective rate function to become strictly convex. For example for the  $n = 1$  case with heavy tails, one might imagine a strong enough *compression* of the observable  $z$  such that a fat tail becomes non-fat.

To harness this nonlinear tilt in the computation of instantons for distributions with heavy tails, we need to modify the approach outlined in section 2.1 as follows: the variation of the unconstrained action (18) now reads

$$[\delta S(\varphi) - \langle \lambda, \nabla F(\varphi(t_1)) \delta \varphi(t_1) \rangle]_{\varphi=\varphi^*} = 0. \quad (38)$$

Consequently, the boundary conditions of the instanton equations are modified to

$$\varphi(t_0) = 0, \quad \vartheta(t_1) = \lambda \nabla F(\varphi(t_1)), \quad (39)$$

which will yield an instanton trajectory  $\varphi^*$  that reaches  $z$ ,  $\varphi(t_1) = z$ , despite the fact that the rate function  $I(z) = S(\varphi^*)$  is not convex around  $z$ . Since  $F$  is continuous, the probability measure  $P^\varepsilon \circ F^{-1}(y)$  in the limit  $\varepsilon \rightarrow 0$  is the same as  $P^\varepsilon(z)$ , according to the contraction principle [9, 10].

Note that this reparametrization through  $F$  is introduced solely to adequately define the tilted measure, or equivalently numerically compute the instanton without encountering divergences. Afterwards, the reparametrization can be reverted to obtain the probability distribution in the original coordinates  $z$ .

As additional remark, methods that compute the instanton by solving the global optimization problem, for example by solving the associated Euler–Lagrange equations instead of integrating the instanton equations [27, 28], do not require the above treatment: the tilting parameter disappears in these cases as the boundary conditions are fixed in the field variable instead of the conjugate momentum. Therefore, in principle, these methods can be chosen in the non-convex case. The solution of the instanton equations, though, is generally preferred [29] due to numerical efficiency, and sometimes even required (such as when the noise covariance in (4) is not invertible).

### 4. Applications

We will now consider a number of examples that show how to compute tail probabilities in stochastic systems. To demonstrate the wide applicability of our approach, we consider several cases that highlight different complications. We start with two toy models that feature stretched exponential (section 4.1) or powerlaw tails (section 4.2). Then, in section 4.3, we consider a two-dimensional system with a bent (‘banana-shape’) potential, where the non-convexity is not due to heavy tails, but due to the shape of the unimodal invariant probability density. Lastly, we demonstrate the practical applicability of our method by considering an example motivated from fiber optics in section 4.4. Here we compute the probability of measuring extreme power spikes at the end of extended optical fibers, where the probability distribution of the input signal is known. Due to soliton formation, this distribution features heavy tails for long fiber lengths ( $L \gg 10$  m), so in order to compute probabilities via an instanton approach, our corrections are necessary.

#### 4.1. Stretched exponential

Consider the stochastic gradient flow,

$$dX_t^\varepsilon = -\nabla U(X_t^\varepsilon)dt + \sqrt{2\varepsilon} dW_t, \quad t \in [t_0, t_1]. \tag{40}$$

The potential  $U : \mathbb{R}^n \rightarrow \mathbb{R}$  determines completely the stationary probability distribution function (PDF)

$$\rho_\infty(z) = Z^{-1} \exp(-\varepsilon^{-1}U(z)), \tag{41}$$

with normalization constant  $Z$ . We further assume that  $U$  has a unique minimum, i.e. we are only considering unimodal distributions. For large times,  $t_1 - t_0 = T \rightarrow \infty$ , the distribution of endpoints of  $X_{t_1}^\varepsilon = z$  will converge to  $\rho_\infty(z)$ . From the perspective of LDT, comparing (26) to (41), the rate function for the final point distribution is equivalent to the potential,  $I(z) = U(z)$ , and

$$\lim_{\varepsilon \rightarrow 0} \varepsilon \log p(z) = -U(z). \tag{42}$$

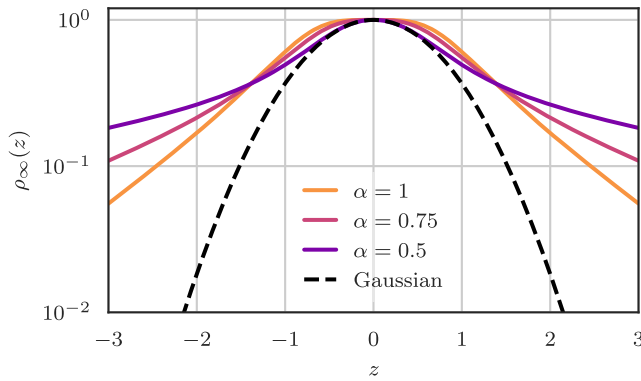
Therefore, in order to approximate the tails of the stationary distribution, we can compute the instanton  $\varphi^*$  ending at  $z$  and estimate  $\rho_\infty(z) \approx \exp(-\varepsilon^{-1}S(\varphi^*))$ .

We choose  $n = 1$  and consider the non-convex potential,

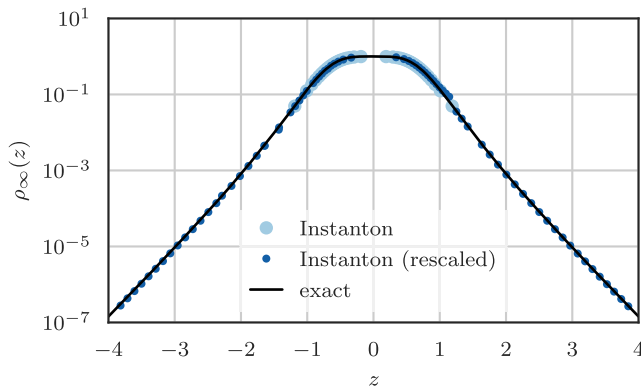
$$U(z) = \left( \frac{z^4}{1 + |z|^3} \right)^\alpha, \quad 0 < \alpha \leq 1, \tag{43}$$

which corresponds to a stretched exponential stationary distribution: at the tails, the dominant exponent is  $\alpha$ , and  $\rho_\infty(z) \approx \exp(-\varepsilon^{-1}|z|^\alpha)$  for large  $z$ , as shown in figure 1. For this distribution,  $\mathbb{E}[\exp(\varepsilon^{-1}\lambda z)]$  diverges for large  $\lambda$  as in case I, and hence numerical methods to find the instanton fail in the tail.

This can be seen in figure 2: we employ the numerical scheme by Chernykh–Stepanov (C–S) [17, 19] to compute the instanton starting at  $\varphi^*(t_0) = 0$  and ending at  $\varphi^*(t_1) = z$  (see appendix A.2 for details). The iterative algorithm converges towards the minimizer of the action, and once converged, we can estimate the probability of reaching  $z$  by  $p(z) \approx \exp(-\varepsilon^{-1}S(\varphi^*))$ . As expected, though, computing the instanton (light blue dots) fails beyond the inflection points at  $z \approx \pm 1.26$ , i.e. the points at which the second derivative of  $U(z)$  changes



**Figure 1.** The potential  $U(z) = \left(z^4/(1 + |z|^3)\right)^\alpha$ ,  $0 < \alpha \leq 1$ , for the gradient flow SDE (40) leads to heavy (stretched exponential) tails of the stationary density  $\rho_\infty(z)$  as  $\alpha$  decreases.



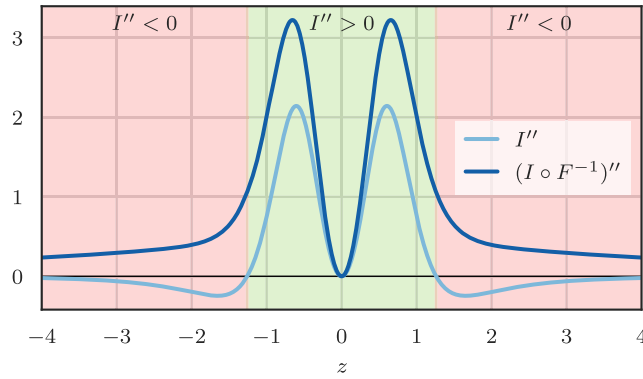
**Figure 2.** Stationary density with exponential tails (equations (41) and (43) for  $\varepsilon = 0.25, \alpha = 1$ ). Probing the tails with the traditional instanton method (light blue) leads to numerical divergence around the non-convex tail region. Reparametrizing the observable via  $F(z) = \text{sign}(z) \log |z|$  convexifies the tail, so that the instanton (dark blue) correctly predicts the exact tail probabilities (solid black).

sign. Here, the tails become stretched exponentials: no choice of  $\lambda$  leads to endpoints  $z$  of the instanton beyond these, as the linear tilt diverges and the CGF is undefined. Instead, we need to choose a non-linear tilt, such as

$$F(z) = \text{sign}(z) \log |z|, \quad z \in \mathbb{R} \setminus \{0\}, \tag{44}$$

for which even in the tails the reparametrized expectation  $\mathbb{E}[\exp(\varepsilon^{-1} \lambda F(X_{t_1}^\varepsilon))] < \infty$  remains bounded. Note that other choices of  $F(z)$  would be possible. For the choice (44), the derivative of the CGF  $G_F(\lambda)$  is a bijection, so that every value of  $\lambda$  has a corresponding  $z$ . This map is explicitly given by

$$\lambda(z) = \lambda \circ F^{-1}(y) = e^{4y} (4 + e^{3y}) / (1 + e^{3y})^2, \tag{45}$$



**Figure 3.** Convexity condition for the stretched exponential tails: the second derivative of the rate function becomes negative beyond the inflection points at  $z \approx \pm 1.26$ . The nonlinearly tilted rate function, instead, remains strictly convex in the whole domain,  $(I \circ F^{-1})'' > 0$ .

(for  $z > 0$ , and negative for  $z < 0$ ).

With this choice, the instanton prediction for the stationary PDF is almost exact far into the heavy tails (dark blue dots vs black solid in figure 2). Here, we again employ the iterative instanton computation, but are solving the instanton equations with the boundary condition (39) instead. The underlying reason for convergence is that the reparametrization with  $F$  convexifies the rate function, i.e. creating supporting lines with slopes  $\lambda$  for all the domain of  $I \circ F^{-1}$ . As shown in figure 3, while the second derivative of the rate function becomes negative beyond the inflection points, the second derivative of the nonlinearly tilted rate function remains positive throughout. This corresponds precisely to the necessary condition (37), and therefore implies that bijectivity.

For the numerics in this example, we chose  $\alpha = 1$  and a noise parameter  $\varepsilon = 0.25$ , with  $N_t = 10^3$  timesteps, and a time interval of  $T = 6$ .

#### 4.2. Powerlaw distribution

Even heavier tails are given by power law distributions,

$$p(z) \sim |z|^{-\beta}, \quad \beta > 0, \tag{46}$$

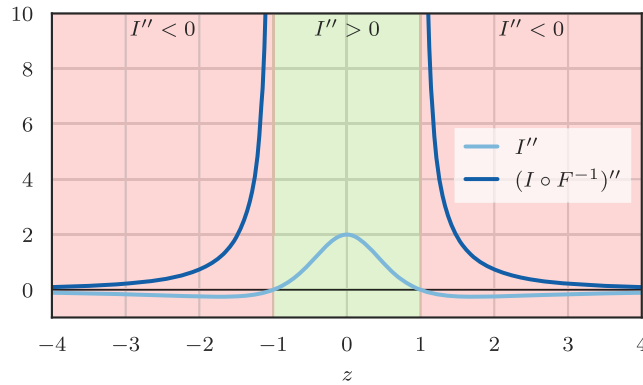
which are associated with a multitude of phenomena in wide areas of science, in part due to their connection to scale invariance, self-similarity, universality classes and criticality in phase transitions. Here, we construct a simple SDE in  $n = 1$  dimensions which has a powerlaw invariant density. Consider

$$dX_t^\varepsilon = -\frac{\beta X_t^\varepsilon}{1 + (X_t^\varepsilon)^2} dt + \sqrt{2\varepsilon} dW_t. \tag{47}$$

It can easily be shown that the invariant density for the process (47) is given by

$$\rho_\infty(z) = Z^{-1}(1 + z^2)^{-\beta/2\varepsilon}, \tag{48}$$

where  $Z$  is a normalization constant. For  $z \gg 1$  and  $\varepsilon = 1$ , this takes the limiting form (46), but is regularized for small  $z$ . Again, we are interested in computing tail probabilities for this toy



**Figure 4.** Convexity condition for the powerlaw test case: the second derivative of the rate function is negative in the tails beyond the inflection points at  $z = \pm 1$  (light blue). The nonlinearly tilted rate function is strictly convex in this region instead (dark blue).

model, by computing the instanton  $\varphi^*$  realizing large values of  $z$ , which yields the respective probability by evaluating the corresponding action.

As in section 4.1, the LDT rate function, given here by

$$I(z) = \frac{1}{2}\beta \log(1 + z^2), \tag{49}$$

does not admit supporting lines in the tails, and consequently its LF transform is undefined (case I as well). This is reflected in the fact that the moment generating function

$$\mathbb{E} \exp(\varepsilon^{-1} \lambda z) = Z^{-1} \int_{\mathbb{R}} \exp(\varepsilon^{-1} \lambda z) (1 + z^2)^{-\beta/2\varepsilon} dz, \tag{50}$$

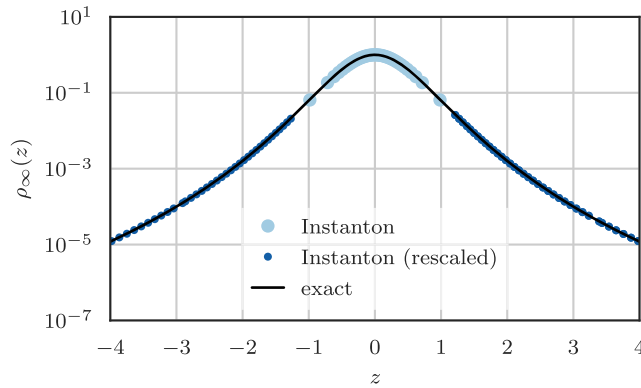
diverges. We can convexify the rate function (49) by reparametrizing via

$$F(z) = \text{sign}(z) \log \log |z|, \quad z \in \mathbb{R} \setminus [-1, 1], \tag{51}$$

which is an even more drastic tail compression than needed for the stretched exponential. Here, again the choice of  $F(z)$  is arbitrary, and only subject to the convexity condition (37). Note further that here we only convexify in the tails,  $|z| > 1$ , where the problem occurs, and do not attempt to find a global map. Indeed, for this choice of  $F$ , the reparametrized rate function  $I \circ F^{-1}(y)$  is convex in the tails, as shown in figure 4: its second derivative remains positive in the tails, which is not true for the original rate function  $I(z)$ . It therefore fulfills the convexity condition given in (37). Consequently, we can explicitly map  $\lambda$  to  $z$  via

$$\lambda(z) = \lambda \circ F^{-1}(y) = \beta e^{(2e^y + y)} / (1 + e^{2e^y}), \tag{52}$$

(for  $z > 1$  and negative for  $z < -1$ ). Numerically, as before, we solve the optimization problem posed by the instanton equations with tilted boundary conditions (39) to obtain instantons  $\varphi^*$  for events with large  $z$ . The corresponding action,  $S(\varphi^*)$ , yields the tail probability. This computation is shown in figure 5: the naive instanton computation (light blue) leads to numerically diverging results in the tail region,  $|z| > 1$ , which are captured accurately by the reparametrized instanton (dark blue). Parameters are  $\beta = 2$ ,  $\varepsilon = 0.25$ ,  $N_t = 10^3$ , and  $T = 10$ .



**Figure 5.** Stationary density with powerlaw tails ( $\beta = 2, \varepsilon = 0.25$ ). Naively computing the instanton for tail events fails beyond  $z = 1$  (light blue). The nonlinear tilt  $F(z) = \text{sign}(z) \log \log |z|$  yields probabilities of events with  $z \gg 1$  within the non-convex powerlaw tail (dark blue) in good agreement with the theoretical result (black solid).

### 4.3. Banana potential

In higher dimensions, non-convexity can manifest in more subtle ways than in 1D. Consider for example the 2D system,

$$dX_t^\varepsilon = b(X_t^\varepsilon) dt + \sqrt{2\varepsilon} dW_t, \tag{53}$$

where,

$$b(x) = -2 \begin{bmatrix} x_1 (1 - 2(x_2 - x_1^2)) \\ x_2 - x_1^2 \end{bmatrix}. \tag{54}$$

This system is a gradient flow for the potential  $U(x) = x_1^2 + (x_2 - x_1^2)^2$ , so that again we have that the rate function for the stationary distribution is equivalent to this potential,  $I(z) = U(z)$ , i.e.

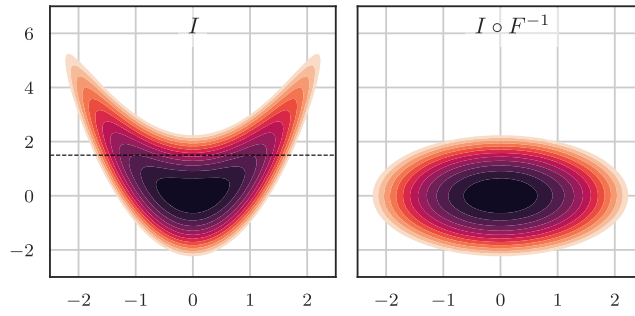
$$I(z) = z_1^2 + (z_2 - z_1^2)^2. \tag{55}$$

The system has a unique stable fixed point at the origin, which is the deepest point of a banana-shaped valley (the set of points  $\{(x_1, x_2) \mid x_2 = x_1^2\}$ ) of the potential, as can be seen in figure 6 (left). The rate function  $I(z)$  (55) does not admit supporting hyperplanes in the region  $\{z = (a, b) \mid b > a^2\}$ , as can be checked by solving (29). As a consequence, this leads to a situation where no tilt variable  $\lambda \in \mathbb{R}^2$  exists to reach an outcome  $z$  within that region.

Unlike the previous examples, the challenge in this case is therefore not the far tails of the stationary density, but actually probing the core of the distribution. The non-convexity of the rate function of the previous examples amounted to the divergence of its LF transform, the CGF (case I), while here the non-convexity leads to the non-differentiability of the CGF (case II).

To fix this non-differentiability of the CGF  $G(\lambda)$ , we propose a nonlinear reparametrization that satisfies the criteria of section 3.2. Consider

$$F(z) = \begin{bmatrix} z_1 \\ z_2 - z_1^2 \end{bmatrix}. \tag{56}$$



**Figure 6.** Left: contour plot of the rate function (55) shows a banana-like valley surrounding a non-convex plateau. The black dashed line represents the position of the marginal shown in figure 7. Right: the rate function composed with the inverse of the non-linear observable (56) deforms the landscape so that the rate function becomes strictly convex.

This reparametrization ‘straightens the banana’, i.e. it deforms the space of outcomes such that the rate function becomes strictly convex, as figure 6 (right) shows. Using this reparametrization produces a continuous and differentiable CGF of the observable  $F$ , resulting from the LF transform of  $I \circ F^{-1}(y)$ ,

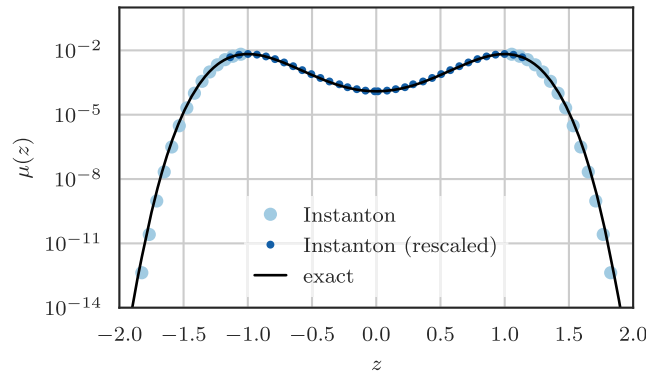
$$\begin{aligned}
 G_F(\lambda) &= \sup_{F^{-1}(y) \in \mathbb{R}^2} (\langle \lambda, y \rangle - I \circ F^{-1}(y)), \\
 &= \frac{1}{4} (\lambda_1^2 + \lambda_2^2), \tag{57}
 \end{aligned}$$

which allows Lagrange multipliers to reach any outcome  $z$ , in particular ones within the nonconvex region.

As numerical experiment, we choose to look at the marginal stationary distribution  $\mu$  in  $z_1$  direction for a fixed value of  $z_2 = \frac{3}{2}$ , i.e.  $\mu(z) = \rho^\infty(z, \frac{3}{2})$ . Since at any fixed value  $z_2 > 0$  we cut through the non-convex region  $z_2 > z_1^2$ , the marginal distribution  $\mu(z)$  looks like a double-well potential. We stress, though, that the whole system indeed has only a single fixed point. We then solve the optimization problem posed by the instanton equations with linear tilt, and compare to the minimization problem with nonlinear tilt. As shown in figure 7, the linearly tilted instanton computation produces acceptable results in the tails of the probability density (light blue dots), it fails to converge within the non-convex region  $-1 < z < 1$ : since in that region there are no supporting planes of the rate function (55), there is no  $\lambda \in \mathbb{R}^2$  corresponding to the slope of the supporting plane at that  $z$ , and consequently no tilt exists to produce the desired outcome  $z$ .

For the reparametrized observable (56), on the other hand, the effective rate function  $I \circ F^{-1}(y)$  is convexified and admits supporting planes at every  $z$ . Indeed, as demonstrated in figure 7, the reparametrized optimization problem leads instanton trajectories reaching outcomes (shown as dark blue dots) within the non-convex region  $-1 < z < 1$ .

As a final remark, convexifying the rate function by the above method, even though it guarantees the existence of a tilt for every outcome, might nevertheless lead to numerical convergence issues. For example, in regions where the original rate function was convex, the rescaled optimization problem might be harder to solve, or necessitate more iterations or smaller time-steps. Similarly, even in the convexified region, the problem might become ill-posed, for example for  $z_1 \ll 1$  and  $z_2 > 2$ , where the observable is approximately linear.



**Figure 7.** The marginal distribution  $\mu(z) = \rho^\infty(z, \frac{3}{2})$  [as denoted by the blacked dashed line in figure 6 (left)]. Instantons with linear tilt (light blue) fail to reach the region  $-1 < z < 1$  without supporting hyperplanes of the rate function  $I(z)$  (55). Performing a reparametrization using the observable (56) produces a strictly convex effective rate function  $I \circ F^{-1}(y)$  that admits supporting hyperplanes everywhere. The corresponding instanton actions successfully capture the non-convex region (dark blue).

#### 4.4. Nonlinear Schrödinger equation

As a practical example for our proposed method, we consider the formation of extreme events in nonlinear wave equations [30–33]. In the field of nonlinear optics and photonics it has been established that heavy tailed statistics frequently occur [34]. Physical mechanisms such as soliton formation [35, 36] and nonlinear amplification [37] are responsible for the emergence of extreme power spikes out of incoherent, Gaussian initial conditions, and have been subject to investigation by a multitude of rare event algorithms [38, 39].

Here, we consider the one-dimensional propagation of an optical pulse along a fiber, described by the NLS equation

$$-i\partial_x\psi = \frac{1}{2}\partial_t^2\psi + |\psi|^2\psi, \quad \psi(x=0, t) = \psi_0(t), \tag{58}$$

for a complex wave envelope  $\psi : [0, L] \times [0, T] \rightarrow \mathbb{C}$ . Boundary conditions are given at location  $x = 0$  at the beginning of the fiber for all times  $t \in [0, T]$ , and the output is measured at the end of the fiber at  $x = L$ . The input signal is considered random, with a Gaussian distribution of known energy spectrum. Specifically, we are mimicking an experimental setup such as [40] of a partially coherent light source, where the input signal is designed as a Gaussian shape in frequency space with covariance

$$\chi_n \sim \exp(-\frac{1}{2}\omega_n^2/\Delta\nu^2), \quad |n| < N, \tag{59}$$

with spectral bandwidth  $1/\Delta\nu$  and truncation frequency  $\omega_N$ , so that the input signal is given by

$$\psi_0(t) = \sum_{n=-N}^N e^{i\omega_n t} \sqrt{\chi_n} \xi_n, \tag{60}$$

where  $\xi_n$  are i.i.d. mean zero, unit variance complex Gaussian random variables.



For this setup, we are interested in the probability of measuring large spikes in the optical power  $A(x, t) = |\psi(x, t)|^2$  at the fiber end,  $x = L$ . Within the presented instanton formalism, this can be achieved by tilting the distribution of initial conditions towards a high-power outcome at the fiber end, and estimating the tail probability by its most likely (‘instantonic’) realization. The corresponding LDP is given by

$$p(z) = P[A(L, T/2) \geq z] \asymp \exp(-I(z)), \tag{61}$$

for a power spike of size  $z$  taken arbitrarily at the center of the temporal domain,  $t = T/2$ . Due to the Gaussianity of the initial conditions, the rate function  $I(z)$  simply is [39]

$$I(z) = \inf_{\xi \in \mathbb{C}^{2N+1}} \left( \frac{1}{2} |\xi|^2 - \lambda(z) |\psi(L, T/2)| \right). \tag{62}$$

Here,  $\xi$  determines the source signal  $\psi_0(t)$  through (60), while,  $\lambda(z)$  can be interpreted as a Lagrange multiplier enforcing the power constraint  $|\psi|^2 = z$  at the end of the fiber. Equation (62) is therefore simply saying that the rate function is given by the most likely random configuration  $\xi$  that determines a source signal with high power output. Note that, similar to the examples above, the tilt in (62) is linear, and we can therefore expect the expectation

$$\mathbb{E} \exp(-\lambda |\psi(L, T/2)|), \tag{63}$$

over light source signals to diverge for fiber lengths  $L$  long enough for solitons to emerge and for the tails of the power distribution to become fat.

Since the probability of high power output signals at the fiber end is not known analytically, the only option we have to get comparison data is to perform MC simulations to sample the power distribution. To this end, we simulate the evolution of a wave packet along the fiber with a random input signals with energy spectrum (59) by numerically integrating equation (58). This equation is non-dimensionalized, with  $x, t$  and  $\psi$  normalized by characteristic parameters  $\mathcal{L}_0, \mathcal{T}_0$  and  $\mathcal{P}_0$  respectively, such that

$$x = \tilde{x}/\mathcal{L}_0, \quad t = \tilde{T}/\mathcal{T}_0, \quad \psi = \tilde{\psi}/\sqrt{\mathcal{P}_0}, \tag{64}$$

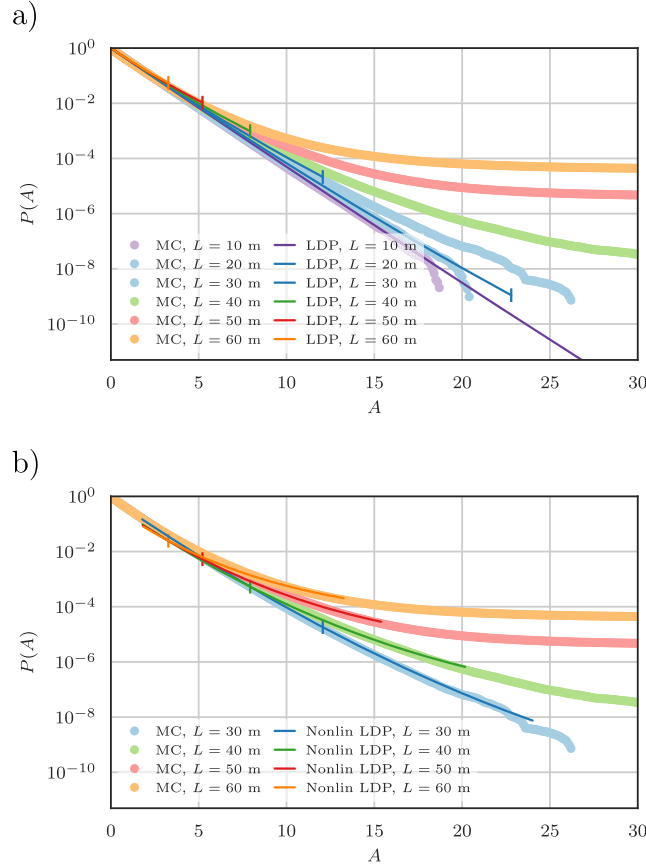
where  $\tilde{x}, \tilde{t}$  and  $\tilde{\psi}$  are the corresponding dimensional variables. These parameters  $\mathcal{L}_0, \mathcal{T}_0$  and  $\mathcal{P}_0$  also determine the dispersion  $\beta_2$  and nonlinearity  $\gamma$  properties of the optical fiber via

$$\beta_2 = \mathcal{T}_0^2/\mathcal{L}_0, \quad \gamma = 1/(\mathcal{L}_0\mathcal{P}_0). \tag{65}$$

We chose these parameters according to the experimental setup in [40], where they are given as

$$\mathcal{L}_0 = 160.3 \text{ m}, \quad \mathcal{T}_0 = 1.8778 \text{ ps}, \quad \mathcal{P}_0 = 2.6 \text{ W}. \tag{66}$$

Therefore, the optical fiber has dispersion parameter  $\beta_2 = 0.022 \text{ ps}^2 \text{ m}^{-1}$  and nonlinearity constant  $\gamma = 0.0024(\text{Wm})^{-1}$ . The spectral bandwidth  $1/\Delta\nu$  is taken to be ( $\Delta\nu = 0.5 \text{ THz}$ ). We pick fiber lengths between 10 m and 60 m, periodic boundary conditions in time treated pseudo-spectrally, and integrate with a second-order Runge–Kutta exponential time differencing method (ETDRK2) [41] in the spatial variable. The discretization is  $\Delta x = 6.24 \times 10^{-3}$ ,  $\Delta t = 1.3 \times 10^{-2}$ ,  $T = 106$ , and frequency cut-off  $N = 45$ . As expected, and shown in figure 8, the tails of the PDF of optical power become heavier with increasing fiber length  $L$ . Its samples number is  $10^6$ , where the property of  $A$  being statistically homogeneous in time is used to improve the statistics.



**Figure 8.** (a) PDF of optical power  $A(x, t) = |\psi(x, t)|^2$  at the end of the fiber. Compared are MC simulations (light color) with instanton prediction (dark color) for different fiber lengths,  $L \in \{10 \text{ m}, \dots, 60 \text{ m}\}$ . The tails become fatter with increasing  $L$ . A vertical marker is inserted at the maximal power achievable with the naive instanton method, highlighting how fat tails prevent useful instanton predictions. (b) The same, but for the instanton with nonlinear tilt, equation (71). The LDP computation now reaches far into the fat tails. The vertical markers of the naive instanton are copied over here for comparison, highlighting the increased tail reach (e.g. by more than a factor 3 for  $L = 60 \text{ m}$ ). Only lengths  $L > 30 \text{ m}$  are shown.

To compare these brute-force sampling estimates to the instanton prediction, we have to solve the optimization problem (62). This is done by defining the cost functional

$$E(\xi) = \frac{1}{2} |\xi|^2 - \lambda |\psi(L, T/2)|, \tag{67}$$

for a given  $\lambda$  and performing gradient descent, where the gradient is given by

$$dE/d\xi = \xi - \lambda J(L, T/2) d|\psi(L, T/2)|/d\psi, \tag{68}$$

with Jacobian  $J(x, t) = d\psi(x, t)/d\xi$ . This gradient can be evaluated by simultaneously integrating the NLS equation (58) and the evolution equation of the Jacobian,

$$\partial_x J = i \left( \frac{1}{2} \partial_t^2 J + \psi^2 \bar{J} + 2|\psi|^2 J \right), \tag{69}$$

(where  $\bar{a}$  is the complex conjugate of  $a \in \mathbb{C}$ ). The iterative gradient descent algorithm yields the optimal choice  $\xi^*$  that will lead to the desired outcome of the final power exceeding the power threshold  $z$ . As can be seen in figure 8(a), the corresponding prediction for the probability,  $\exp(-I(z))$  (from equation (61)) correctly describes the tail decay of high power events at the fiber end, but crucially *only as long as the rate function admits supporting lines, i.e. remains convex*. Therefore, the instanton prediction is basically useless for optical fibers longer than  $L = 30$  m. As a side note, the gradient computation could instead be performed in the adjoint formalism, leading to two coupled forward–backward equations similar in spirit to the instanton equation (15), but identically yielding meaningful results only in the convex region of the rate function.

Now, applying the idea from above, we can instead *nonlinearly tilt* the probability distribution of input signals towards high power outcomes. For this, we choose instead the nonlinearly tilted rate function

$$I(x) = \inf_{\xi \in \mathbb{C}^{2N+1}} \left( \frac{1}{2} |\xi|^2 - \lambda(y) F(\psi(L, T/2)) \right), \tag{70}$$

with

$$F(z) = \log \log |z|, \quad |z| > 1. \tag{71}$$

For this tilt, the cost functional becomes

$$E(\xi) = \frac{1}{2} |\xi|^2 - \lambda F(\psi(L, T/2)), \tag{72}$$

and the gradient is

$$dE/d\xi = \xi - \lambda J(L, T/2) dF(\psi(L, T/2))/d\psi, \tag{73}$$

instead. The results of this are shown in figure 8(b) for the four longest fiber lengths of 30–60 m in 10 m increments, where the tails are fattest. In the revised formalism (dark color), the nonlinearly tilted instanton prediction is able to reach far into the stretched tail and gives the right order of magnitude for the probability of power spikes obtained from sampling (light color). The end of the region of convergence for the naive instanton is shown for comparison (vertical markers).

Note that due to the choice of reparametrization  $F$  in (71), the nonlinearly tilted instanton prediction is restricted to the region of normalized power  $|\psi|^2 > 1$ , but of course this is exactly the tail region that we care about.

### 5. Conclusion

Estimating the probability of tail events can efficiently be done via LDT and instanton calculus, which transforms an inefficient sampling problem into a deterministic optimization problem. Unfortunately, for systems with heavy tails, or more generally non-convex rate functions, standard mechanisms of exponentially tilting the measure, or numerically solving the optimization problem, fail. The reason is the absence of a bijective map between Lagrange multiplier (tilting parameter) and desired outcome, caused by the breakdown of their Lagrange duality, or equivalently by the non-convexity of the rate function.

We put forward the idea of a nonlinear tilt that reparametrizes the output space, effectively convexifying the rate function of the observed probability distribution. We discuss the necessary conditions required for this reparametrization to yield a unique outcome variable and ensure a bijective mapping between tilt and outcome: it needs to be a diffeomorphism chosen such that its composition with the rate function is strictly convex. Note further that the reparametrization can be chosen locally, i.e. the conditions on the nonlinear observable need only apply in a subdomain of the events of interest.

Finding such nonlinear observable can be subtle, especially when the system is highly nonlinear, influencing the rate function landscape. However, drawing inspiration from toy problems with stretched exponential and algebraic tails, which can be treated analytically, yields candidate reparametrizations for physically relevant problems. We show the applicability to real-world problems by demonstrating how instantons determine the probability of extreme optical power events in a fiber optical cable, where solitons lead to a heavy-tailed power distribution at the fiber end.

## Acknowledgments

The authors thank Eric Vanden-Eijnden for helpful discussions and Giovanni Dematteis for help with the source code for the fiber optics example. MA acknowledges the PhD funding received from UKSACB. TG acknowledges the support received from the EPSRC projects EP/T011866/1 and EP/V013319/1.

## Appendix.

### A.1. Instanton equations with two different sets of boundary conditions

In this section, we will explain in more detail the difference between the derivation of instanton equations with boundary conditions (16) and (19). The goal is to find the minimizer of the rate function, equation (11), where  $\varphi$  must belong to  $A_z$ , (6), which is a set that enforces the boundary conditions (16). The formula for the rate function is given by (9), where the integrand is the Lagrangian  $L(\varphi, \dot{\varphi})$ , which can be written in terms of the Hamiltonian  $H(\varphi, \vartheta)$  as follows,

$$S(\varphi) = \int_{t_0}^{t_1} L(\varphi, \dot{\varphi}) dt = \int_{t_0}^{t_1} [\langle \vartheta, \dot{\varphi} \rangle - H(\varphi, \vartheta)] dt, \quad (\text{A.1})$$

with the Hamiltonian  $H(\varphi, \vartheta)$  given by equation (14). With this, the variation of the rate function yields

$$\begin{aligned} \delta S(\varphi, \vartheta) \Big|_{\varphi=\varphi^*, \vartheta=\vartheta^*} &= \int_{t_0}^{t_1} \left[ \left( \left\langle \vartheta, \frac{\partial \dot{\varphi}}{\partial \varphi} \right\rangle - \frac{\partial H(\varphi, \vartheta)}{\partial \varphi} \right) \delta \varphi \right. \\ &\quad \left. + \left( \dot{\varphi} - \frac{\partial H(\varphi, \vartheta)}{\partial \vartheta} \right) \delta \vartheta \right] dt = 0. \end{aligned} \quad (\text{A.2})$$

After performing integration by parts on the first term of equation (A.2), this results in

$$\dot{\vartheta} = -\partial_{\varphi} H(\varphi, \vartheta), \quad \dot{\varphi} = \partial_{\vartheta} H(\varphi, \vartheta), \quad (\text{A.3})$$

which are the instanton equation (15) written in terms of the Hamiltonian (14). These two coupled equations solved with the boundary conditions induced by the set  $A_z$  correspond to the boundary value problem given in equation (16).

**Algorithm 1.** C–S algorithm.

**Choose:**

$$\lambda, \eta \in [0, 1], [t_0, t_1], \Delta t, N = \frac{t_1 - t_0}{\Delta t}, \text{TOL}$$

**Set:**

$$k = 0, \varphi^k(t) = \bar{x}, \vartheta^k(t) = 0$$

**repeat**

Given  $\varphi^k$  and  $\tilde{\vartheta}(t=t_1) = \lambda \nabla F(\varphi^k(t=t_1))$ , integrate  $\tilde{\vartheta}(t)$  backward in time

Given  $\tilde{\vartheta}$  and  $\tilde{\varphi}(t=t_0) = \bar{x}$ , integrate  $\tilde{\varphi}(t)$  forward in time

**Update:**

$$\varphi^{k+1} = \eta \tilde{\varphi} + (1 - \eta) \varphi^k$$

$$\vartheta^{k+1} = \eta \tilde{\vartheta} + (1 - \eta) \vartheta^k$$

$$k \leftarrow k + 1$$

**until**  $\max |\varphi^{k+1} - \varphi^k| / |\varphi^k| < \text{TOL}$

**Terminate.**

The boundary value problem defined in equation (19) is different. Here, the constraint is instead encoded in the Lagrange multiplier, which ends up modifying the boundary conditions of the instanton equations, as we will show next. Starting from equation (17), the variation of the rate function (compare it with (A.2)) becomes

$$\begin{aligned} \delta S(\varphi, \vartheta) \Big|_{\varphi=\varphi^*, \vartheta=\vartheta^*} &= \int_{t_0}^{t_1} \left[ \left( \left\langle \vartheta, \frac{\partial \dot{\varphi}}{\partial \varphi} \right\rangle - \frac{\partial H(\varphi, \vartheta)}{\partial \varphi} - \lambda \frac{\partial \varphi(t)}{\partial \varphi} \delta(t - t_1) \right) \right. \\ &\quad \left. \times \delta \varphi + \left( \dot{\varphi} - \frac{\partial H(\varphi, \vartheta)}{\partial \vartheta} \right) \delta \vartheta \right] dt = 0. \end{aligned} \tag{A.4}$$

This now results in,

$$\dot{\vartheta} = -\partial_{\varphi} H(\varphi, \vartheta), \quad \dot{\varphi} = \partial_{\vartheta} H(\varphi, \vartheta), \quad \text{subject to } \vartheta(t_1) = \lambda, \tag{A.5}$$

while the derivation of the instanton equations remains unchanged. The main difference is the fact that the final time condition  $\varphi(t_1)$  has been encoded in the rate function as a Lagrange multiplier, resulting in a final time condition on the conjugate momentum  $\vartheta$ , equation (19).

The instanton equations with this second set of boundary conditions defined in (A.5) can be solved very efficiently by an adjoint formulation of gradient descent (‘C–S’ scheme), while using the original conditions, equation (16), would necessitate solving the instanton equations with comparably inefficient shooting methods.

*A.2. Chernykh–Stepanov numerical scheme*

In this subsection we lay out implementation details of the recursive C–S scheme. It can be used to solve the instanton equation (15) with boundary conditions given in (19), or, for our purposes, the tilted boundary conditions of equation (39).

The main idea is to solve for  $\vartheta(t)$  backward in time and, subsequently, using this solution  $\vartheta$ , to solve for  $\varphi(t)$  forward in time. This choice of integration direction is not only consistent with the given boundary values, but turns out to be the numerically stable choice for the right-hand-sides of the instanton equations.

The whole scheme is explained in pseudocode as algorithm 1. Note that it is analogous to the *adjoint formalism* of a gradient descent in numerical optimization, where  $\vartheta$  is considered as the adjoint variable [29]. The interpolated update with  $\eta$  is only needed in case of numerical


instabilities. In this case, the variable  $\eta$  can be interpreted as step size of the gradient descent, and could instead be chosen via e.g. Armijo line search. For the tolerance TOL of the exit criterion, we choose  $\text{TOL} = 10^{-8}$ . A practical limitation of this scheme is the fact that whenever we are interested in statements about the stationary densities of our process, we would be forced to let  $t_1 - t_0 = T \rightarrow \infty$ , which can only be approximated numerically (by taking a large enough  $T$ ). A more sophisticated solution to this problem is to replace the physical time parameter  $t$  by a geometric parameter [18]. However, such a treatment was not needed in this paper.

### Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

### ORCID iDs

Mnerh Alqahtani  <https://orcid.org/0000-0002-7082-0034>

Tobias Grafke  <https://orcid.org/0000-0003-0839-676X>

### References

- [1] Bucklew J 2013 *Introduction to Rare Event Simulation* (Berlin: Springer)
- [2] Frisch U 1995 *Turbulence* (Cambridge: Cambridge University Press)
- [3] Kessler D A and Barkai E 2010 *Phys. Rev. Lett.* **105** 120602
- [4] Drăgulescu A and Yakovenko V M 2001 *Physica A* **299** 213–21
- [5] Sinha S 2006 *Physica A* **359** 555–62
- [6] Gopikrishnan P, Plerou V, Gabaix X and Stanley H E 2000 *Phys. Rev. E* **62** R4493
- [7] Plerou V, Gopikrishnan P, Nunes Amaral L A, Gabaix X and Eugene Stanley H 2000 *Phys. Rev. E* **62** R3023
- [8] Varadhan S R S 1966 *Commun. Pure Appl. Math.* **19** 261–86
- [9] Dembo A and Zeitouni O 2010 *Large Deviations Techniques and Applications* (Berlin: Springer)
- [10] Freidlin M I and Wentzell A D 2012 *Random Perturbations of Dynamical Systems* vol 260 (Berlin: Springer)
- [11] Ellis R S 1984 *Ann. Probab.* **12** 1–12
- [12] Costeniuc M, Ellis R S, Touchette H and Turkington B 2005 *J. Stat. Phys.* **119** 1283–329
- [13] Costeniuc M, Ellis R, Touchette H and Turkington B 2006 *Phys. Rev. E* **73** 026105
- [14] Touchette H 2010 *J. Stat. Mech.* **P05008**
- [15] Deriglazov A 2017 *Classical Mechanics* (Berlin: Springer)
- [16] Rindler F 2018 *Calculus of Variations* (Berlin: Springer)
- [17] Chernykh A I and Stepanov M G 2001 *Phys. Rev. E* **64** 026306
- [18] Grafke T, Grauer R, Schäfer T and Vanden-Eijnden E 2014 *Multiscale Model. Simul.* **12** 566–80
- [19] Grafke T, Grauer R and Schäfer T 2013 *J. Phys. A: Math. Theor.* **46** 062002
- [20] Kim E-J and Anderson J 2008 *Phys. Plasmas* **15** 114506
- [21] Rolland J, Bouchet F and Simonnet E 2016 *J. Stat. Phys.* **162** 277–311
- [22] Meerson B, Katzav E and Vilenkin A 2016 *Phys. Rev. Lett.* **116** 070601
- [23] Zarfaty L and Meerson B 2016 *J. Stat. Mech.* **033304**
- [24] Touchette H 2009 *Phys. Rep.* **478** 1–69
- [25] Cohen S N and Elliott R J 2015 *Stochastic Calculus and Applications* vol 2 (Berlin: Springer)
- [26] Touchette H 2005 Legendre–Fenchel transforms in a nutshell published online: [www.maths.qmul.ac.uk/ignorespacesht/archive/lfth2.pdf](http://www.maths.qmul.ac.uk/ignorespacesht/archive/lfth2.pdf)
- [27] Weinan E, Ren W and Vanden-Eijnden E 2004 *Commun. Pure Appl. Math.* **57** 637–56

- [28] Grafke T, Schäfer T and Vanden-Eijnden E 2017 Long term effects of small random perturbations on dynamical systems: theoretical and computational tools *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science (Fields Institute Communications)* (New York: Springer) pp 17–55
- [29] Grafke T and Vanden-Eijnden E 2019 *Chaos* **29** 063118
- [30] Zakharov V E 1968 *J. Appl. Mech. Tech. Phys.* **9** 190–4
- [31] Osborne A R, Onorato M and Serio M 2000 *Phys. Lett. A* **275** 386–93
- [32] Mori N, Onorato M, Janssen P A E M, Osborne A R and Serio M 2007 *J. Geophys. Res.* **112** C09011
- [33] Onorato M, Proment D, El G, Randoux S and Suret P 2016 *Phys. Lett. A* **380** 3173–7
- [34] Akhmediev N, Dudley J M, Solli D R and Turitsyn S K 2013 *J. Opt.* **15** 060201
- [35] Kibler B, Fatome J, Finot C, Millot G, Dias F, Genty G, Akhmediev N and Dudley J M 2010 *Nat. Phys.* **6** 790
- [36] Tikan A *et al* 2017 *Phys. Rev. Lett.* **119** 033901
- [37] Onorato M, Osborne A R, Serio M and Cavaleri L 2005 *Phys. Fluids* **17** 078101
- [38] Farazmand M and Sapsis T P 2017 *Sci. Adv.* **3** e1701533
- [39] Dematteis G, Grafke T and Vanden-Eijnden E 2019 *SIAM/ASA J. Uncertain. Quantification* **7** 1029–59
- [40] Tikan A, Bielawski S, Sz waj C, Randoux S and Suret P 2018 *Nat. Photon.* **12** 228–34
- [41] Du Q and Zhu W 2005 *BIT Numer. Math.* **45** 307–28