International Workshop on Statistical-Mechanical Informatics 2009 (IW-SMI 2009) IOP Publishing
Journal of Physics: Conference Series **197** (2009) 012005 doi:10.1088/1742-6596/197/1/012005

Inference of a random potential from random walk realizations: Formalism and application to the one-dimensional Sinai model with a drift

S Cocco [1,3] **and R Monasson** [2,3]

[1] Laboratoire de Physique Statistique de l'ENS, CNRS, UMPC, 24 rue Lhomond, 75005 Paris, France
[2] Laboratoire de Physique Théorique de l'ENS, CNRS, UPMC, 24 rue Lhomond, 75005 Paris, France
[3] The Simons Center for Systems Biology, Institute for Advanced Study, Einstein Drive, Princeton NJ 08540, USA

E-mail: cocco@lps.ens.fr, monasson@lpt.ens.fr

Abstract. We consider the Sinai model, in which a random walker moves in a random quenched potential $\mathbf{V}$, and ask the following questions: 1. how can the quenched potential $\mathbf{V}$ be inferred from the observations of one or more realizations of the random motion? 2. how many observations (walks) are required to make a reliable inference, that is, to be able to distinguish between two similar but distinct potentials, $\mathbf{V}_1$ and $\mathbf{V}_2$? We show how question 1 can be easily solved within the Bayesian framework. In addition, we show that the answer to question 2 is, in general, intimately connected to the calculation of the survival probability of a fictitious walker in a potential $\mathbf{W}$ defined from $\mathbf{V}_1$ and $\mathbf{V}_2$, with partial absorption at sites where $\mathbf{V}_1$ and $\mathbf{V}_2$ do not coincide. For the one-dimensional Sinai model, this survival probability can be analytically calculated, in excellent agreement with numerical simulations.

1. Introduction

Random walks in random media have been intensively studied in the past decades in statistical physics, as well as in the theory of probability [1, 2]. Most efforts have been devoted to the characterization of the statistical properties of the walks, such as the velocity, the correlation functions, ... from the ones of the environment, that is, the energy potential. Fewer studies (see however [3]) have tackled the inverse problem: given one (or more) realizations of the random walks, can we guess the values of the potential?

The issue of the inference of a potential from the observation of a random, noisy motion naturally arises in the analysis of experiments on biophysics or biology. A nice illustration is provided by single-molecule experiments, in which biological objects, such as macro-molecules (DNA,RNA, enzymes) can be individually tracked. These objects, of the micrometer or nanometer size, are submitted to the thermal agitation of the solvent, and to intra- or inter-molecular interactions. Examples of interactions are ligand-receptor interactions in experiments dealing with the rupture of non-covalent bonds [4], base-pairing interactions in polynucleotides unzipping, *i.e.* the mechanical separation of DNA/RNA strands [5, 6]. The resulting potentials can be considered as disordered since they depend on the details of the molecular constituents,

and may vary a lot among ligand-receptor systems and DNA/RNA molecules. In addition, in order to modify the strength of the interactions between constituents, external forces can be exerted on the pico-Newton scale. The presence of such an external force tilts the potential, and affect the observed motion, which, in turn, should provide additional information on the interaction potential.

In a recent letter [8], motivated by the idea of infering the sequence of DNA molecule from the signal of the unzipping force in separation experiments [7], we have studied the case of a random walker moving in a quenched random potential $\mathbf{V}$, and asked the following two questions: **1.** how can the quenched potential $\mathbf{V}$ be inferred from the observations of one or more realizations of the random motion? **2.** how many observations (walks) are required to make a reliable inference, that is, to be able to distinguish between two similar but distinct potentials, $\mathbf{V}_1$ and $\mathbf{V}_2$? The present paper is intended as a pedagogical introduction to those questions and our results, as well as a numerical check of our main result, that is, the close and abtract relationship between question 2 and the survival probability of a fictitious random walker in a partially absorbing potential. Definitions are given in Section 2, questions 1 and 2 are studied in Sections 3 and 4 respectively, while the relationship mentioned above is exposed in Section 5.

2. Definitions of the random potentials, and random walks

The model we look at is the discrete space-continuous time Sinai model. A walker is moving on the set of non-negative integers $x = 0, 1, 2, 3, \ldots$, according to the set of values of the potential on those sites, $\mathbf{V} = \{V_x,\ x \geq 0\}$.

2.1. Dynamical rules for the random walk

Given the potential $\mathbf{V}$, the rate (probability per unit of time) for the walker to jump from x onto a neighbouring site $x' = x \pm 1$ reads

$$r_{\mathbf{V}}(x', x) = e^{\gamma(V_x - V_{x'})} \tag{1}$$

which ensures that detailed balance is satisfied at temperature $1/(2\gamma)$. We impose a reflecting boundary condition in $x = 0$, by setting $V_{-1} = +\infty$. The rate transition matrix $\hat{M}_{\mathbf{V}}$ is tridiagonal with non zero elements $M_{\mathbf{V}}(x \pm 1, x) = r_{\mathbf{V}}(x \pm 1, x)$, $M_{\mathbf{V}}(x, x) = -M_{\mathbf{V}}(x+1, x) - M_{\mathbf{V}}(x-1, x)$. An important property of $\hat{M}_{\mathbf{V}}$ is that the sum of entries along each column is equal to zero, which ensures the conservation of the probability for the walker to be somewhere on the postive half-line. A realization of the random walk is the set of positions of the walker at all times smaller than the duration T of the walk, $\mathbf{X} = \{x(t),\ 0 \leq t \leq T\}$. For definiteness, we assume that the walker is initially at the origin: $x(0) = 0$.

The above dynamics is defined in continuous time. It may be convenient in some cases to consider that time is discrete. Observations of the position are then made at times multiple of Δt. The number of observations is $T/\Delta t$. The transition matrix $\hat{Q}_{\mathbf{V},\Delta t}$ is simply obtained from the exponential of the rate transition matrix $M_{\mathbf{V}}$ through

$$\hat{Q}_{\mathbf{V},\Delta t} = \exp\left[\Delta t\ \hat{M}_{\mathbf{V}}\right] . \tag{2}$$

When Δt is very small, $\hat{Q}_{\mathbf{V},\Delta t} \simeq \mathrm{Id} + \Delta t\ \hat{M}_{\mathbf{V}}$, and is almost tridiagonal.

2.2. Random force model

The potential is itself a quenched random variable drawn as follows. On each bond $(x, x+1)$ with $x \geq 0$, we define a force f_x. The value of the potential at site x is defined through

$$V_x = -\sum_{y<x} f_y\ , \tag{3}$$

with $V_0 \equiv 0$. The forces are i.i.d. stochastic variables taking ± 1 values, and drawn from the law

$$P(f_x = f) = \frac{1+b}{2}\,\delta_{f,+1} + \frac{1-b}{2}\,\delta_{f,-1}\ , \tag{4}$$

where b is called tilt in the following. The tilt is the opposite of the average slope of the potential. Note that (3) defines a one-to-one mapping between the potential $\mathbf{V}$ and the set of forces $\{f_x\}$. Inferring the former is therefore equivalent to inferring the latter.

Hereafter, we focus on the $b > 0$ case, for which the random walk is transient. We show a sample of potential drawn with $b = 0.4$ in Figure 1A, as well as several realizations of random walks of duration $T = 5000$ in this potential, for $\gamma = 0.5$, in Figure 1B. The dynamics is affected by the presence of many deep local minima, in which the walked gets trapped for long times. The control parameter for the dynamics is $\alpha = \frac{T}{2}\,\log(\frac{1+b}{1-b})$[2]. When $\alpha > 1$, the velocity of the walk is finite, and self-averaging, *i.e.* it depends only on b and T, and not on the sample of the potential. When $\alpha < 1$, the velocity is vanishing. The average location of the walker increases as t^α only, due to the presence of deep traps with long sojourn times τ. As we will see later, while the identification of these two regimes is critical to the description of the random walk motion, it is less important as far as the inference problem is concerned.

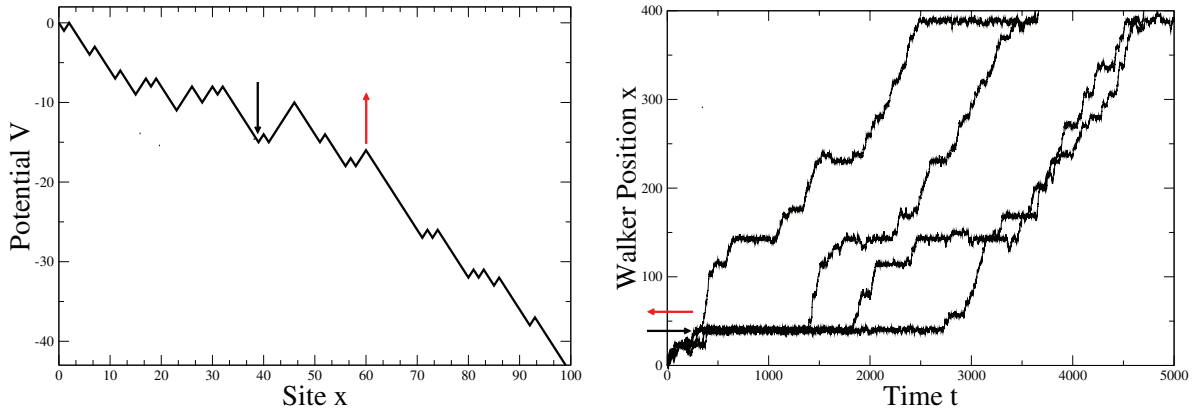

Figure 1. **A.** example of potential with $b = 0.4$. **B.** four realizations of random walks of duration $T = 5000$, and for $\gamma = 0.5$ in this potential. The up and right arrows point to a bond whose attached force is easy to infer, the down and left arrows point to a bond corresponding to a hardly predictable force, see Figure 2.

3. Question 1: How to infer the potential from one or more realizations in practice?

3.1. On realizations and signals

In this subsection, we focus on the discrete-time dynamics, and eventually send the time step $\Delta t \to 0$ at the end of the calculation. The probability of a time-trace (realization) $\mathbf{X}$ in the potential $\mathbf{V}$ reads

$$P[\mathbf{X}|\mathbf{V}] = \prod_{m=0}^{T/\Delta t-1} \hat{Q}_{\mathbf{V},\Delta t}(x_{m+1}, x_m)\ . \tag{5}$$

where the transition matrix $\hat{Q}_{\mathbf{V},\Delta t}$ is defined in (2). As the product over times in the above expression can be reordered, a sufficient statistics for $\mathbf{X}$ is the set $\mathbf{S} = \{N_x, u_x^+, u_x^-\}$, hereafter

called signal, where N_x is the number of time-steps m such that $x_{m+1} = x_m = x$, *i.e.* during which the walker remains at site x, and $u_x^{\pm}$ are the numbers of jumps from $x_m = x$ to $x_{m+1} = x \pm 1$. In term of the signal $\mathbf{S}$, the probability of $\mathbf{X}$ can be rewritten as

$$P[\mathbf{X}|\mathbf{V}] = \prod_{x\geq 0} \hat{Q}_{\mathbf{V},\Delta t}(x,x)^{N_x}\ \hat{Q}_{\mathbf{V},\Delta t}(x-1,x)^{u_x^-}\ \hat{Q}_{\mathbf{V},\Delta t}(x+1,x)^{u_x^+}\ . \tag{6}$$

with the obvious definition $u_0^- = 0$. Note that the above expression could be extended to the case of a walker not constrained to nearest-neighbour jumps. The signal S would then be a matrix, whose elements $S(x',x)$ would count the number of jumps (in the realization $\mathbf{X}$) from site x to site x' during a time-step.

Assume now that after time T (the duration of one walk), the walker is put again at the origin, and the random process is resumed. This way, more realizations of the random walk, $\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^R$, can be obtained in the same potential. As the random walks are independent of each other, their probabilities factorize, and the corresponding signals simply add up.

3.2. Bayesian expression for the most likely forces

Let us now take the $\Delta t \to 0$ limit. We expect $u_x^{\pm}$ to remain finite, and N_x to scale as $t_x/\Delta t$, where t_x is the local time at site x. The signal attached to a time-trace is now the set $\mathbf{S} = \{t_x, u_x^+, u_x^-\}$, and, from (6), we have

$$P[\mathbf{X}|\mathbf{V}] = \prod_{x\geq 0} \exp\left(-t_x\ (e^{\gamma f_x} + e^{-\gamma f_{x-1}}) - \gamma f_{x-1}\, u_x^- + \gamma f_x\, u_x^+\right) = C\ \exp\left(\gamma \sum_{x\geq 0} f_x\ h_x\right) \tag{7}$$

where C is a signal-independent factor, and the local field acting on force f_x is

$$h_x = \frac{\sinh\gamma}{\gamma}\ (t_{x+1} - t_x) + u_x^+ - u_x^-\ . \tag{8}$$

Finding the set of forces f_x^* with maximal *a priori* probabilities is straightforward: for each x, f_x^* is given by the sign of h_x. In other words, the forces are uncoupled in (7). Hence, the random force model is the simplest model one can think of from the inference point of view. More complicated models, with short-range couplings between neighboring forces, can be solved using transfer matrix techniques [7].

From a strictly Bayesian point of view, the *a posteriori* probability of the potential should include the prior over the potentials. Here, the prior is given by the distribution of forces (4). It is easy to see that this extra and multiplicative contribution to the probability amounts to an additive contribution, equal to α, to the local fields h_x acting on the forces f_x.

We show in Figure 2 the probabilities of inferring the wrong force f_x on bond $(x, x+1)$, $P_{error}(x)$, as a function of x for the potential $\mathbf{V}$ of Figure 1A. Those probabilities were calculated from 10,000 samples of $R = 1$ or 4 realizations of the random walk of equal duration T. For each sample, we extract the local times, t_x, and numbers of jumps, $u_x^{\pm}$, and compare the forces $f_x^* = \text{sign}(h_x)$ (8) to their true values. We see that $P_{error}(x)$ widely fluctuates from bond to bond: it is small close to local minima of $\mathbf{V}$, where the walker spends much time, and large in the vicinity of maxima or on the descending flanks of the potential.

As the random walk is transient, the walker visits each site only a finite number of times. The information about the values of the potential in the vicinity of a site can therefore only be increased by collecting more random walk realizations. Figure 2 shows that the probability of error strongly decreases when the number of realizations on which the inference is based grows from $R = 1$ to $R = 4$. A systematic numerical study shows that, for a given bond $(x, x+1)$, $P_{error}(x)$ asymptotically decays as $e^{-R/R_c(x)}$, where $R_c(x)$ depends on x and on the whole potential $\mathbf{V}$. The following section is devoted to the calculation of $R_c(x)$, the number of realizations necessary for an accurate inference of the force on the $(x, x+1)$ bond .

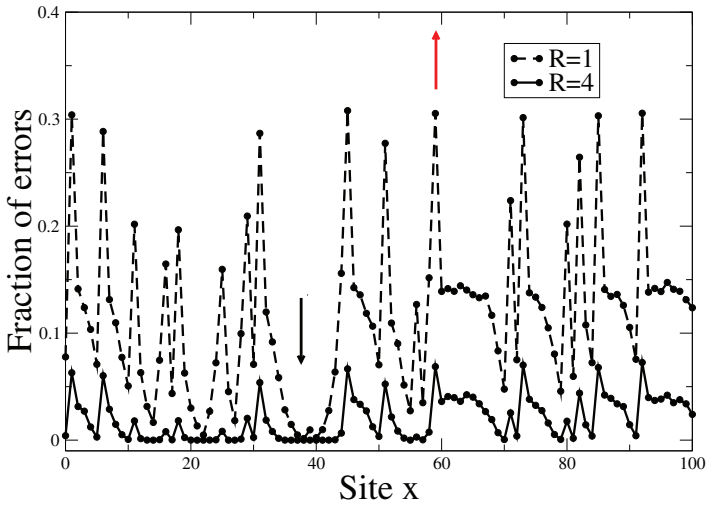

Figure 2. Fraction of errors, $P_{error}(x)$, for the forces over the first 100 bonds inferred from $R = 1$ and $R = 4$ realizations. The potential is the one shown in Figure 1. Parameters are $b = 0.4$, $\gamma = 0.5$, $T = 5000$. The fractions are calculated from the average over 10,000 samples of R random walks. The arrows locate the same bonds as in Figure 1, respectively, $(x, x+1) = (38, 39)$ and $(59, 60)$.

4. Question 2: how many realizations are necessary for a reliable inference?

4.1. A (much) simpler problem: 2-coin testing

To start with, let us consider the following and simpler problem. Suppose one has two biased coins $i = 1, 2$; a draw of coin i will give head with probability p_i, and tail with probability $1 - p_i$. The values p_1 and p_2 are known to the player. Then one picks up one of the two coins (with equal probability one half), and launch it R times. Can we infer which coin was picked up from the observation of the sequence of the R draws? How big should be R to ensure that the probability of guessing the wrong coin is smaller than, say, ϵ? These questions, called hypothesis testing, are very fundamental ones in statistics, see [9] for instance.

Let H be the number of heads in the sequence of R draws. As the draws are independent from each other, H is binomially distributed, see Figure 3A. When R is large, the probability of H can be approximated through

$$P(H = R\,h|i) = \exp\left(- R\, D(h|p_i) + o(R)\right) , \tag{9}$$

where i is the label of the 'true' coin, and

$$D(h|p_i) = h\, \ln(h/p_i) + (1-h)\, \ln\left((1-h)/(1-p_i)\right) \tag{10}$$

is the Kullback-Leibler divergence between h and p_i. $D(h|p_i)$ is a convex and positive function of h over the $[0; 1]$ range, vanishing in $h = p_i$ only (Figure 3B). In the Bayesian inference framework [9], the *a posteriori* probability for the coin label i given the value of H is

$$P(i|H) = \frac{P(H|i)}{P(H|1) + P(H|2)} . \tag{11}$$

Consequently, the most likely coin i is the one maximizing $P(H|i)$. For large R, one infers $i = 1$ if $H < h^* R$, $i = 2$ if $H > h^* R$, where h^* is the fraction of heads at which the two Kullback-Leibler divergences cross: $D(h^*|p_1) = D(h^*|p_2)$ (Figure 3B).

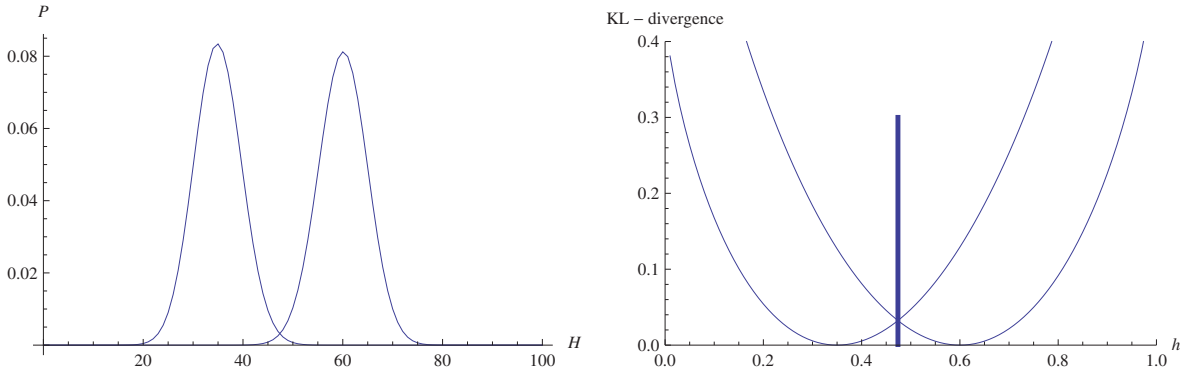

Figure 3. The two-coin testing problem. **A.** Probability of the number of head draws, H, after $R = 100$ draws for coins $i = 1, 2$, with $p_1 = 0.35 < p_2 = 0.6$. **B.** Kullback-Leibler divergences associated to the two coins, $D(h = H/R|p_i)$, see(10). The vertical line shows the location of the crossing point, $h^* \simeq 0.4739$. The inverse of the height of the crossing point is R_c (13).

What confidence can one have about the above inference? The probability of wrongly inferring, say, coin $i = 2$, while the true coin is $i = 1$ is given by

$$P_{error} = \sum_{H>h^*R} P(H|1) = \exp\left(- R/R_c(p_1, p_2) + o(R)\right) . \tag{12}$$

with

$$R_c(p_1, p_2) = \frac{1}{D(h^*|p_1)} = \frac{1}{D(h^*|p_2)} . \tag{13}$$

The above expression clearly shows that the inference error is a symmetric function of the coin probabilities when R is large. The height of the crossing point of the two Kullback-Leibler divergences is called Chernoff information in information theory [9]. Its inverse, R_c, can be interpreted as the order of magnitude of the number of draws needed to have a reliable inference. More precisely, the probability of error will be smaller than ϵ if the number of draws exceeds

$$R > \log(1/\epsilon) \times R_c(p_1, p_2) . \tag{14}$$

4.2. Back to the inverse Sinai problem: tiling of the signal space

Despite its simplicity, the 2-coin testing problem introduces the main idea needed to treat the more complicated case of the inverse Sinai problem. This key idea is the following: errors in the inference are due to very atypical observations. When the number of observations, R, is large, the main contribution to the probability of error comes from observations equally likely to be generated by any of the two candidate distributions *i.e.* the two coins in the previous section, or the two potentials $\mathbf{V}_1$, $\mathbf{V}_2$ for the Sinai problem. Those critical observations (and their probability) can be estimated based on extremal arguments, called large-deviation theory in probability literature [10].

Table 1 gives the correspondence between the quantities defined and studied in the 2-coin testing problem, and their counterparts in the problem of the inference of a potential from some realizations of a random walk.

As explained in Section 3.1, the information on a potential which can be extracted from the observation of the motion of the walker is contained in the signal $\mathbf{S}$. Signals corresponding to independent random walks in the same potential sum up. We thus expect that $\mathbf{S}$ will be an

Table 1. Correspondence between 2-coin testing and the potential inference problem.

	2-coin testing problem	potential inference problem
quantity to infer	coin i	potential $\mathbf{V}$
observation	sequence of draws	random walk $\mathbf{X}$
sufficient statistics	number H of heads	signal $\mathbf{S} = \{N_x, u_x^{\pm}; x \geq 0\}$
asymptotic probability for the sufficient statistics	$\exp\,(-R\,D(\frac{H}{R};p) + o(R))$	$\exp\,(-R\,\omega_{\mathbf{V}}(\frac{\mathbf{S}}{R}) + o(R))$
boundary between candidates	fraction h^*	critical signal $\mathbf{s}^*$

extensive function of the number R of realizations. In the following, we will denote the intensive ratio $\mathbf{S}/R$ by $\mathbf{s}$.

We now think of a signal $\mathbf{S}$ as a vector of components $\{\mathbf{S}_{x,0} = N_x, \mathbf{S}_{x,+} = u_x^+, \mathbf{S}_{x,-} = u_x^-; x \geq 0\}$ in a (infinite-dimensional) space. In the same way, we define the vector $\mathbf{v}$ attached to each potential $\mathbf{V}$, with components $\{\mathbf{v}_{x,0} = \ln \hat{Q}_{\mathbf{V}_1,\Delta t}(x,x), \mathbf{v}_{x,+} = \ln \hat{Q}_{\mathbf{V}_1,\Delta t}(x+1,x), \mathbf{v}_{x,-} = \ln \hat{Q}_{\mathbf{V}_1,\Delta t}(x-1,x); x \geq 0\}$ attached to each potential $\mathbf{V}$. From (6), the likelihood of the signal given the potential is simply the exponential of the dot product between the signal and potential vectors,

$$P[\mathbf{X}|\mathbf{V}] = \exp\left(\sum_{x\geq 0} \mathbf{S}_x\ \mathbf{v}_x\right)\ . \tag{15}$$

It is therefore natural to partition the space of signals into Voronoi cells $C_{\mathbf{v}}$. $C_{\mathbf{v}}$ is the set of signals having a larger dot product with $\mathbf{v}$ than with any other $\mathbf{v}'(\neq \mathbf{v})$, see Figure 4. It is also the set of signals from which the potential $\mathbf{V}$ will be inferred.

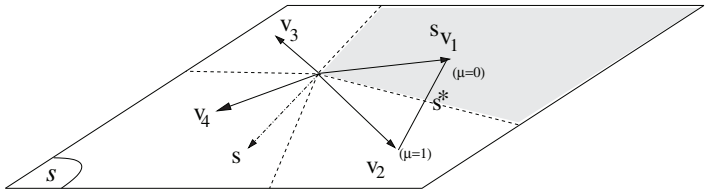

Figure 4. The space of signals, $\mathbf{s}$, tiled with Voronoi cells attached to every possible potentials. The cell corresponding to potential $\mathbf{V}_i$ is the set of signals $\mathbf{s}$ having a larger dot product with $\mathbf{v}_i$ than with any other potential vectors. The line joining the potential vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, parametrized by μ, is shown. Its intersection with the boundary between the cells is the critical signal $\mathbf{s}^*$.

Let $\mathbf{V}_1$ be the 'true' potential in which the walker moves. We know that, in the large R limit, the signals $\mathbf{S}$ most likely to lead to the erroneous inference of $\mathbf{V}_2$ are located on the boundary between cells $C_{\mathbf{v}_1}$ and $C_{\mathbf{v}_2}$. The probability of erroneously predicting $\mathbf{V}_2$ will therefore be given, when R is large, by the probability of signals extracted from potential $\mathbf{V}_1$, and constrained to

have equal dot products with both potential vectors,

$$P_{error} = \sum_{\mathbf{S}} P_R(\mathbf{S}|\mathbf{V}_1)\ \delta_{\mathbf{S}\cdot(\mathbf{v}_1-\mathbf{v}_2),0}\ . \tag{16}$$

It is important to distinguish the probability $P(\mathbf{X}|\mathbf{V}_1)$ of a realization $\mathbf{X}$, given by (15), from the probability $P_R(\mathbf{S}|\mathbf{V}_1)$ of the signal $\mathbf{S}$ itself. The latter is equal to the former multiplied by the number of random walks sharing the same signal. We denote this multiplicity by $\mathcal{M}_R(\mathbf{S})$, where we make explicit the dependence on the number R of realizations.

When R gets large, the multiplicity $\mathcal{M}_R(\mathbf{S})$ is expected to scale as $\exp(R\,\sigma(\mathbf{s})+o(R))$, where the growth function σ depends on the duration T of the walk. Hence the probability of a signal given the 'true' potential is

$$P_R(\mathbf{S}|\mathbf{V}_1) = \exp\left\{R\left[\sigma(\mathbf{s}) + \mathbf{s}\cdot\mathbf{v}_1\right] + o(R)\right\}\ . \tag{17}$$

Usinf a Lagrange multiplier to enforce the constraint in the sum on the right-hand side of (16), we conclude that

$$P_{error} = \min_{\mu} \sum_{\mathbf{s}} \exp\left\{R\left[\sigma(\mathbf{s}) + \mathbf{s}\cdot\left((1-\mu)\,\mathbf{v}_1 + \mu\,\mathbf{v}_2\right)\right] + o(R)\right\}\ . \tag{18}$$

The role of the optimization over μ is precisely to locate the critical signal $\mathbf{s}^*$ on the boundary between the two cells attached to the two potentials, see Figure 4.

Interestingly, there is no need to calculate the growth function σ to exploit the above formula. An immediate comparison of (18) and (17) shows that P_{error} is, apart the minimization over μ, the sum over all possible signals of the probability of signals attached to a fictitious 'potential', whose vector is $(1-\mu)\,\mathbf{v}_1 + \mu\,\mathbf{v}_2$. This potential is fictitious in that it is generally not (for $\mu \neq 0, 1$) one of the potentials attached to a sequence of binary forces. The nature of the fictitious potential and of the random walk taking place therein, is unveiled in the next section.

5. Relationship with a fictitious and partially absorbing random walk

5.1. Interpretation of P_{error} as a probability of survival

The second member of (18) defines an effective transition matrix, whose elements are given by, to the first order in Δt,

$$\begin{aligned}
\hat{Q}_{\mathbf{V}_1,\mathbf{V}_2,\mu,\Delta t}(x',x) &= \hat{Q}_{\mathbf{V_1},\Delta t}(x',x)^{1-\mu}\ \hat{Q}_{\mathbf{V_2},\Delta t}(x',x)^{\mu} \qquad (19)\\
&= \begin{cases} \Delta t\ r_{\mathbf{V}_1}(x',x)^{1-\mu}\, r_{\mathbf{V}_2}(x',x)^{\mu} & \text{if} \quad x' = x \pm 1 \\ 1 - \Delta t \sum_{y=\pm 1} \left((1-\mu)\, r_{\mathbf{V}_1}(x+y,x) + \mu\, r_{\mathbf{V}_2}(x+y,x)\right) & \text{if} \quad x' = x \end{cases}
\end{aligned}$$

Hence, in the $\Delta t \to 0$ limit, we may interpret $\hat{Q}_{\mathbf{V}_1,\mathbf{V}_2,\mu,\Delta t}$ as a transition matrix describing a one-dimensional random walk on the set of non-negative integers. The associated potential, $\mathbf{W}$, can be read from the off-diagonal transition rates. For instance

$$r_{\mathbf{W}}(x+1,x) = (e^{\gamma f_{1,x}})^{1-\mu}\ (e^{\gamma f_{2,x}})^{\mu} = e^{\gamma\left((1-\mu)f_{1,x}+\mu f_{2,x}\right)}\ , \tag{20}$$

where $f_{i,x}$ is the force acting on bond $(x, x+1)$ in the potential $\mathbf{V}_i$. A similar result holds for the transition from x to $x-1$. Hence $\mathbf{W} = (1-\mu)\mathbf{V}_1 + \mu\mathbf{V}_2$, as could be guessed from the comparison of (17) and (18).

However, $\hat{Q}_{\mathbf{V}_1,\mathbf{V}_2,\mu,\Delta t}$ is not equal to $\hat{Q}_{\mathbf{W},\Delta t}$, since the latter conserves the probability, while the former does not. Indeed, according to (19),

$$\sum_{x'} \hat{Q}_{\mathbf{V}_1,\mathbf{V}_2,\mu,\Delta t}(x',x) = 1 - \Delta t\ d_{\mathbf{V}_1,\mathbf{V}_2,\mu}(x) + O(\Delta t^2)\ , \tag{21}$$

where

$$d_{\mathbf{V}_1,\mathbf{V}_2,\mu}(x) = \sum_{y=\pm 1} \Big[(1-\mu)\, r_{\mathbf{V}_1}(x+y,x) + \mu\, r_{\mathbf{V}_2}(x+y,x) - r_{\mathbf{W}}(x+y,x) \Big] \ . \tag{22}$$

When $0 < \mu < 1$, by convexity of the exponential function, $d_{\mathbf{V}_1,\mathbf{V}_2,\mu}$ is a positive quantity, which vanishes only if $\mathbf{V}_1$ and $\mathbf{V}_2$ coincide around site x, that is, if the forces $f_{i,x-1}$ and $f_{i,x}$ on, respectively, bonds $(x-1,x)$ and $(x,x+1)$ take the same values in both potentials. Hence $d_{\mathbf{V}_1,\mathbf{V}_2,\mu}$ can be interpreted as the rate of absorption at site x for the random walk process associated to $\hat{Q}$. In other words, it is the probability per unit of time for the walker to die at site x. We conclude that the probability of error (18) is

$$P_{error} = \pi^{R+o(R)} \ , \tag{23}$$

where π is the probability that the walker, initially at site $x = 0$, survives till time T. We find back the exponential decay of the probability of error with R, with a decay constant $R_c = -1/\ln \pi$.

We emphasize that the more different are the potentials $\mathbf{V}_1$ and $\mathbf{V}_2$ in the neighbourhood of x, the higher is the absorption rate $d_{\mathbf{V}_1,\mathbf{V}_2,\mu}(x)$, and the more likely is the walker to die in x. Hence the probability of survival, $\pi(0,T)$, is large, *i.e.* close to 1, when the potentials $\mathbf{V}_1$ and $\mathbf{V}_2$ are difficult to tell from one another, and small when they are easily distinguishable based on the observation of one random walk realization.

5.2. Calculation of the survival probability for the one-dimensional Sinai model

The results derived in the above section are easily generalizable to the case of random walk in spaces of dimension larger than one. Formula (23) then gives a practical way to estimate R_c by means of numerical simulations. In the case of the one-dimensional Sinai model, the survival probability can be analytically calculated as we now show.

Let $\pi(x,t)$ be the probability that a random walker, initially in x, has survived until time t; we are eventually interested in calculating $\pi = \pi(0,T)$ (23). To obtain the equation fulfilled by $\pi(x,t)$, we write that, in order to be alive until time $t + \Delta t$ where Δt is an infinitesimal time interval, the walker can first move during the first Δt period, and then has to survive during a period of duration equal to t,

$$\pi(x, t+\Delta t) = \sum_{x'} \pi(x',t)\ \hat{Q}_{\mathbf{V}_1,\mathbf{V}_2,\mu,\Delta t}(x',x) \ , \tag{24}$$

In the $\Delta t \to 0$ limit, we obtain the following backward Fokker-Planck equation,

$$\frac{d\pi}{dt}(x,t) = \sum_{x'(\neq x)} r_{\mathbf{W}}(x',x)\ (\pi(x',t) - \pi(x,t)) - d_{\mathbf{V}_1,\mathbf{V}_2,\mu}(x)\ \pi(x,t) \ , \tag{25}$$

with the initial condition $\pi(x,0) = 1$.

Consider the case of where the differences between the two sequences of forces are localized on a subset of bonds arount site z, *e.g.* $f_{1,x} = f_{2,x}$ for all x but $x = z$, and $f_{1,z} = -f_{2,z}$. The absorption rates vanish everywhere, except in the vicinity of z. Let t_z denotes the mean first-passage time at site z in one of the two potentials. If T is much smaller than t_z, the random walker is alive with high probability since the neighborhood of z has not been visited. On the contrary, when T is much larger than t_z (or, more precisely, the average last-passage time in z), the random walk is very likely not to visit z anymore, and the walker will wander to infinity without risking of being absorbed.

To calculate the probability of survival in this large T limit, $\pi = \pi(0, \infty)$, we look for a stationary solution of (25), denoted by $\pi_s(x)$ for a given μ. Defining $q^{\pm}(x) = r_{\mathbf{W}}(x \pm 1, x)/(r_{\mathbf{W}}(x+1, x) + r_{\mathbf{W}}(x-1, x) + d_{\mathbf{V}_1, \mathbf{V}_2, \mu}(x))$, the ratio $\rho(x) = \pi_s(x)/\pi_s(x+1)$ satisfies the recurrence equation

$$\rho(x) = \frac{q^+(x)}{1 - q^-(x)\,\rho(x-1)}\ , \tag{26}$$

for $x \geq 1$. We iterate this recursive equation, with the boundary condition $\rho(0) = q^+(0)$, to get $\pi(0) = \prod_{x \geq 0} \rho(x)$. The probability of survival can then be minimized over μ to obtain π. Figure 5 shows the application of this procedure to the potential of Figure 1A.

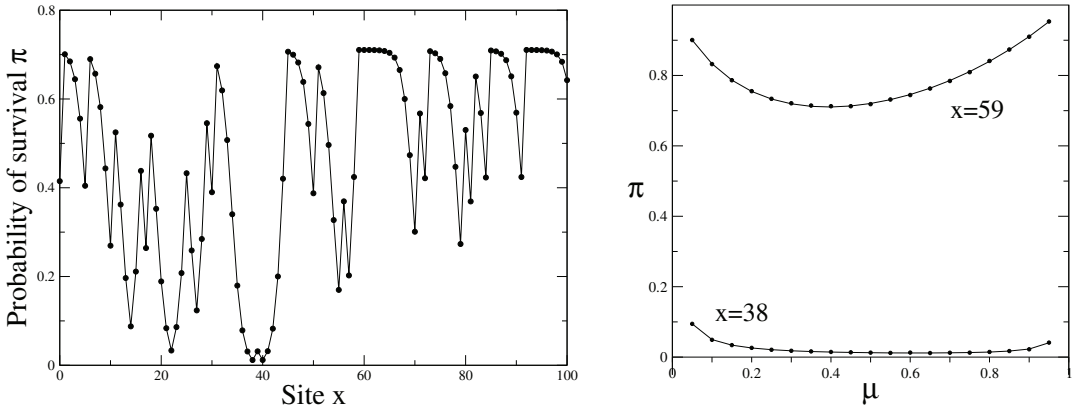

Figure 5. A. Probability of survival π corresponding to the comparison of the potential of Figure 1 and the same potential with the force on bond $(x, x+1)$ inverted. The probability of error from R realizations decays as π^R for large R, compare to Figure 2. **B.** Probability of survival π as a function of μ for sites $x = 38$ and $x = 59$. Lines are theoretical predictions. Dots show results from numerical simulations averaged over 100,000 realizations of the absorbing random walk; error bars are smaller than the dot size. The survival probability corresponds to the minimum of the curves over μ.

6. Conclusion

In this paper, we have shown how the observation of a random walker in a potential can be used to infer the value of this potential. We have also investigated the issue of the reliability of the inferred potential, and, more precisely, of how many observations are necessary to distinguish between two candidate potentials. From a technical point of view, R_c is obtained from the probability of survival of a random walker in a fictitious and partially absorbing potential. The interest is two-fold. First, it allowed us to avoid the calculation of the entropy σ of signals. Secondly, it gives a practical way to estimate R_c for any model, and not only for the one-dimensional Sinai model, where many quantities can be calculated exactly.

We have here focused on the case of transient walks. The issue of inferring a random potential from recurrent walks, *i.e.* when the tilt b vanishes, is a very interesting one from a mathematical point of view, and was recently studied by Andreoletti [3]. In this case, the motion is very slow: very few sites are visited (of the order of $(\log T)^2$ after a time T), but the forces on the corresponding bonds can be accurately inferred. From a practical point of view, it may be more profitable to work at a finite tilt, so that the walk acquires a finite velocity. The number of

visited bonds grows linearly with T and, though the information collected by a single random walk realization is incomplete, the strategy is on the overall faster as far as the inference is concerned [8].

Acknowledgement: We thank P. Andreoletti and R. Diel for useful discussions. This work was partially funded by the ANR 06-JCJC-051 contract.

References

[1] Hughes B D 1996 *Random Walks and Random Environments* (Oxford University Press)
[2] Bouchaud J P and Georges A 1990 *Phys. Rep.* **195** 127
[3] Andreoletti P 2006 *Preprint* arxiv:math.PR/0612208
[4] Evans E and Ritchie K 1997 *Biophys. J.* **72** 1541
[5] Essevaz-Roulet B, Bockelmann U and Heslot F 1997 *Proc. Natl. Acad. Sci. USA* **94** 11935
[6] Danilowitcz C *et al* 2003 *Proc. Natl. Acad. Sci. USA* **100** 1694
[7] Baldazzi V *et al* 2006 *Phys. Rev. Lett.* **96** 128102; 2007 *Phys. Rev.* E **75** 011904
[8] Cocco S and Monasson R 2008 *Europhys. Lett.* **81** 20002
[9] Cover T M and Thomas J A 1991 *Elements of Information Theory* (Wiley)
[10] Dembo A and Zeitouni O 1998 *Large Deviations Techniques and Applications* (Springer-Verlag)
[11] Chernoff H 1952 *Ann. Math. Stat.* **23** 493