Reconstructing a random potential from its random walks

S. Cocco\(^1\) and R. Monasson\(^2\)

\(^1\) CNRS-Laboratoire de Physique Statistique de l’ENS - 24 rue Lhomond, 75005 Paris, France
\(^2\) CNRS-Laboratoire de Physique Théorique de l’ENS - 24 rue Lhomond, 75005 Paris, France

received 11 June 2007; accepted in final form 13 November 2007
published online 10 December 2007

PACS 02.30.Zz – Inverse problems
PACS 05.40.Fb – Random walks and Levy flights
PACS 02.50.Tt – Inference methods

Abstract – The problem of how many trajectories of a random walker in a potential are needed to reconstruct the values of this potential is studied. We show that this problem can be solved by calculating the probability of survival of an abstract random walker in a partially absorbing potential. The approach is illustrated on the discrete Sinai (random-force) model with a drift. We determine the parameter (temperature, duration of each trajectory, …) values making reconstruction as fast as possible.

Copyright © EPLA, 2008

Introduction. – Random walks (RWs) in random media have been intensively studied in the past decades as a paradigm for out-of-equilibrium dynamics, and have led to the discovery and understanding of important dynamical effects as anomalous diffusion, ageing ... [1,2]. Briefly speaking the issue is to determine the statistical properties of the walker from the ones of the energy potential. Much less attention has been devoted to the inverse problem [3]; given one (or more) observed RW(s) can we guess the potential values? This question naturally arises in biophysics where the use of AFM, optical and magnetic tweezers makes possible the mechanical separation of single protein-protein complexes [4], or the unfolding and refolding of single biomolecules [5-7]. The rupture of chemical bonds, the dynamics of folding/unfolding of nucleic acids, or proteins can be modeled as a RW motion affected by thermal noise, moving in a quenched potential determined by the composition of the chemical bonds, or the sequence of amino- or nucleic-acids. Reconstructing the free energy landscape of those processes is the object of current and intense efforts [4,7-10].

In this letter we show how the inverse RW problem can be practically solved within the Bayesian inference framework and address the crucial question of the accuracy of reconstruction. In practice information can be accumulated either by increasing the duration of one RW, or observing more than one RW, or combining the two. We discuss the optimal procedure minimizing the total number of data to be acquired, and show how this minimal amount of data can be calculated from the probability of survival of an abstract walker in a partially absorbing potential. The approach is illustrated in detail on the celebrated discrete random-force (RF) model (Sinai model with nonzero drift) [1,2].

Inference is a key issue in information theory and statistics [11], with applications in biology [12], social science [13], finance, ... A central question is the so-called hypothesis testing problem: which one of two candidate distributions is likely to have generated a set of measured data? This question was solved in the case of independent variables by Chernoff [14], and is the core issue of the asymptotic theory of inference [11]. Chernoff showed that the probability of guessing the wrong distribution decreases exponentially with the size of the data set [14]. Later on the case of correlated variables extracted from one recurrent realization of a finite Markov chain, an idealization of frequently encountered situations, was solved [15,16]. The present work can be seen as an extension of those studies to many transient realizations of an “infinite” chain.

Random-force model. – For an illustration of the problem consider the discrete, one-dimensional RF model defined on the set of sites \(x = 0, 1, 2, \ldots, N\) [1]. We start by choosing randomly a set of dimensionless forces \(f_x = \pm 1\) on each link \((x, x + 1)\) with a priori probability \(P_0 = \prod_{x}^{1+b/2} f_x\), where \(-1 < b < 1\) is called tilt. This defines the values of the potential \(V\) on each site, \(V_x = -\sum_{y<x} f_y\) (by definition \(V_0 = 0\)). An example of potential for \(b = 0.4\) is shown in fig. 1.
After the quenched potential has been drawn, a random walker starts in \( x = 0 \) at time \( t = 0 \). The walker then jumps from one site \( x \) to one of its neighbors \( x' = x \pm 1 \) with rate \( r_{x,x'}(x \rightarrow x') = r_0 \times e^{(V_x - V_{x'})/(2T)} \) to satisfy detailed balance at temperature \( T \); the attempt rate \( r_0 \) will be set to unity in the following. Reflecting boundary conditions are imposed by setting \( V_{N+1} = V_{-1} = +\infty \). We register the sequence of positions up to some time \( t_f \): \( X = \{x(t), 0 \leq t \leq t_f \} \). Figure 1 shows five RWs \( X_p, p = 1, \ldots, 5 \), each starting in the origin \( x(0) = 0 \) and of equal duration \( t_f \) for a temperature \( T = 1 \). The value of the temperature strongly affects the dynamics [2], and its relevance for the inverse problem will be discussed later.

Our objective is to reconstruct the potential over a region of the lattice, e.g., the value of the forces on some specific links from the observation of RWs. Within Baye’s inference framework this can be done by maximizing the joint probability of the potential \( V \) and of the observed RWs \( X_1, \ldots, X_R \) over \( V \) [11]. \( P \) is the product of the a priori probability of the potential, \( P_0 \), times the likelihood of the RWs given the potential, \( L \). Since the RW is Markovian, \( L \) depends only on the sets of total times \( t_x \) spent on every site \( x \), and of the numbers of jumps \( u(x \rightarrow x') \) from \( x \) to \( x' \) over the set of RWs:

\[
L = \prod_{x,x'} e^{-t_x} r_V(x \rightarrow x') r_V(x \rightarrow x')^{u(x \rightarrow x')},
\]

where the product runs over all sites \( x \) and their neighbors \( x' = x \pm 1 \). Expressing the rates in terms of the forces and maximizing the joint probability \( P \) we obtain the most likely values for the forces: \( f_x = \text{sign}(h_x + \alpha) \), where \( \alpha = T \ln[(1+b)/(1-b)] \) is a global “field” coming from the a priori distribution \( P_0 \) and \( h_x \) a local contribution due to the likelihood \( L \):

\[
h_x = 2T \sinh \left( \frac{1}{2T} \right) \left( t_{x+1} - t_x \right) + u(x \rightarrow x + 1) - u(x + 1 \rightarrow x).
\]

Figure 1 (left, bottom) shows predictions made from the collection of the first \( R \) RWs in the right panel, with \( R \) ranging from 1 to 5. The duration \( t_f \) of the RW is chosen to be much larger than the mean first-passage time in \( x = 200 \), and much smaller than the equilibration time \( t_{eq} \sim e^{bN/T} \). In this range the quality of prediction is essentially independent of \( t_f \) as will be discussed in detail below. As expected, the number of erroneous forces decreases with increasing \( R \) though atypical events may produce flaws in the prediction. The analysis of these atypical RWs, and how they lead to errors, is the keystone of what follows.

**Number of RWs necessary for a good reconstruction.** — Expression (1) for the likelihood of the RWs is true for any potential \( V \) and can be geometrically interpreted as follows. Given a set of RWs we extract a signal vector \( S \) whose components are the times \( t_x \) spent on site \( x \), the numbers \( u(x \rightarrow x') \) of transitions from site \( x \) to site \( x' \). When \( R \) is large we expect \( S \) to be extensive with \( R \) and define the intensive signal \( s = S/R \). Similarly, to each potential \( V \) we associate a vector \( v \) with components: minus the outgoing rate, i.e., \( \sum_{x'} v(x \rightarrow x') \) for each site \( x \), the logarithm of the rate \( r_V(x \rightarrow x') \) for

---

20002-p2
Reconstructing a random potential from its random walks

Fig. 2: Space of signals $s$ partitioned into Voronoi cells $C_V$, one for each potential $V$ (see text). The signal extracted from a large number $R$ of RWs in the potential $V$ is strongly concentrated around the typical signal $s^*_V$. Small deviations from $s^*_V$ (dark shading, small value of $\omega_V$) are more likely than large fluctuations (light shading, large values of $\omega_V$). The most dangerous potential $V$ corresponds to one of the contiguous cells, $C_V$; $\mu$ parametrizes the path between $s^*_V$ ($\mu = 0$) and $s^*_V$ ($\mu = 1$). The value of $\omega_V$ at the crossing point with the border between the two cells is the inverse of $R$, the number of RWs required to predict potential $V$.

of the RW, it turns out that its Legendre transform is easy to calculate, and is sufficient for our purpose.

The probability to predict an erroneous potential is the probability that the stochastic signal $s$ does not belong to cell $C_V$. This probability of error thus decays exponentially with $R$ over a typical number of RWs,

$$R_e(V) = \left[ \min_{s \notin C_V} \omega_V(s) \right]^{-1},$$

where the minimum is taken over signals outside the “true” cell. It depends on the temperature, the duration of the RW, . . .

As the RWs are independently drawn, $\omega_V$ is a convex function of $s$, see appendix a) and [15]. The minimum in (3) is thus reached on the boundary between the true cell and another, bad cell, say, $C_b$. The attached potential, $V_b$, is the most “dangerous” one from the inference point of view. RWs generated from $V_b$ are hardly told from each other unless more than $R_e(V_b)$ of them are observed.

Assume $V$ is known. Then the boundary between $C_V$ and $C_V$ is the set of signals $s \perp v - \bar{v}$. We deduce

$$R_e(V) = \left[ \max_{\mu} \min_s (\omega_V(s) + \mu s \cdot (\bar{v} - v)) \right]^{-1},$$

where the Lagrange multiplier $\mu \in [0;1]$ ensures that $s$ is confined to the boundary. The Legendre transform of $\omega_V$ appearing in (4) is intimately related to the evolution operator of an abstract random-walk process, denoted by $\text{RW}(\mu)$ to distinguish from the original RW, see appendix b) and [17]. This $\text{RW}(\mu)$-er moves with the rates $r_{(1-\mu)\text{V} + \mu \text{V}}(x \rightarrow x')$ and may die on every site $x$ with positive rate,

$$d_{\text{V}, \mu}(x) = \sum_{x' \neq x} [(1-\mu)r_V(x \rightarrow x') + \mu r_V(x \rightarrow x')$$

$$- r_{(1-\mu)\text{V} + \mu \text{V}}(x \rightarrow x'),$$

see fig. 3. Consider now the probability $\pi(\mu)$ that the RW$(\mu)$-er, initially at the origin, has survived up to time $t_f$ (the duration of the original RW). Then $R_e(V) = \min_{\mu \in [0;1]} 1/|\ln \pi(\mu)|$, see appendix b).

Optimal working point for the RF model. – We apply the above theory to the discrete RF model, and want to predict the value of the force $f_y$ on the link $(y,y+1)$ for some specific $y$. The dangerous potential is $V$ obtained from $V$ upon reversal of the force $f_y \rightarrow -f_y$. We aim at calculating the probability $\pi(\mu)$ of survival of the RW$(\mu)$-er moving with rate $r(x \rightarrow x') = r_V(x \rightarrow x')$ and dying on site $x$ with rate $d(x) = 0$ except: $r(y \rightarrow y+1) = 1/r(y+1 \rightarrow y) = e^{(1-2\mu)f_y/(2T)}$, $d(y) = D(f_y)$, $d(y+1) = D(-f_y)$, where $D(f) \equiv (1-\mu)e^{f/(2T)} + \mu e^{-f/(2T)} - e^{(1-2\mu)f_y/(2T)}$ from (5). From the previous section the number of RWs required for a reliable prediction of $f_y$ is $R_e(y; V) = \min_{\mu \in [0;1]} 1/|\ln \pi(\mu)|$. 

---

1 The irrelevance of the a priori distribution in the asymptotic case of large data set is well known [11] and can be checked for the RF model: the local field (2) is extensive in $R$, while the global field $\alpha$ remains finite.

2 Let $v' \neq v$: $s^*_v \cdot (v - v') = \sum_{x \neq x'} u^*(x \rightarrow x')G(r_V(x \rightarrow x')/r_V(x \rightarrow x'))$, where $G(z) = z - \ln z - 1 > 0$ for $z \neq 1$. 

20002-p3
Let $\pi_x(\mu, t)$ be the probability that $\text{RW}(\mu)$, initially on site $x$, is still alive at time $t$. The time evolution of $\pi_x$ is described by

$$
\frac{\partial \pi_x}{\partial t} = \sum_{x' \neq x} r(x \to x')(\pi_{x'} - \pi_x) - d(x) \pi_x,
$$

with initial condition $\pi_x(\mu, 0) = 1$ (by convention $\pi_{-1} = \pi_{N+1} = 0$). After Laplace transform over time, eqs. (6) are turned into recurrence equations for the ratios $\pi_x/\pi_{x+1}$ and solved with great numerical accuracy. We obtain this way the probability of survival, $\pi(\mu) = \pi_0(\mu, t_f)$, and optimize over $\mu$. Though $R_c$ depends on the potential $V$, its general behavior for tilt $b > 0$ as a function of the duration $t_f$ is sketched in fig. 4. Three regimes are observed:

- for $t_f \ll \tau_y$ (mean first-passage time in $y$) $\text{RW}(\mu)$ has a low probability to visit $y$ and is almost surely alive, hence $R_c$ is very large;
- for $\tau_y \ll t_f \ll t_{eq}$ $\text{RW}(\mu)$ has visited the region surrounding $y$ and escaped from this region (transient regime), hence its probability of survival remains constant, and so does $R_c$;
- for $t_f \gg t_{eq}$ $\text{RW}(\mu)$ visits again and again the region surrounding $y$, hence the probability of survival decreases exponentially with duration: $R_c \propto 1/t_f$.

The total time $R_c \times t_f$ for a good reconstruction is minimal when we choose $t_f \sim \tau_y$. This marginally transient regime corresponds to the plateau of fig. 4: $\text{RWs}$ are long enough to visit site $y$ but short enough not to wander much away from $y$. To calculate the corresponding value of $R_c$ we take the limits, in order, $N \to \infty$, $t_f \to \infty$, and look for the stationary solution of (6) with boundary condition $\pi_{x \to \infty} = 1$. The result for the probability of survival is

$$
\pi(\mu) = \frac{e^{-D}}{1 - \mu e^{-D} + \mu(1 - \mu) t_{y+1}^* (e^{-D} - e^{-D/2})^2},
$$

where the mean sojourn time on site $y+1$ in $V$ is [2]

$$
t_{y+1}^* = \sum_{z \geq 0} \exp \left[ \frac{1}{T} \left( \frac{V_{y+z+1} + V_{y+z+1}}{2} - V_{y+1} \right) \right].
$$

Distribution of $\{R_c\}$ over potentials. – The number $R_c(y; V)$ of $\text{RWs}$ necessary to predict the value of $f_y$ depends on the potential $V$ through the sojourn time $t_{y+1}^*$ (8). By randomly drawing potentials (or varying site $y$) we obtain the distribution of $R_c$ shown in fig. 5. The main features are:

- Small $R_c$ correspond to sites where the $\text{RW}$ spends long time $t^*$ (traps); $R_c \sim \frac{1}{\ln \pi} \sim \frac{1}{\ln \tau^*}$ from (7). The power law tail of the distribution of sojourn times, $P(t^*) \sim (t^*)^{-(\alpha+1)}$ [2], gives rise to an essential singularity at the origin in the cumulative distribution, $Q(R_c) \sim e^{-\alpha/R_c}$. The potential is easy to predict over trapping regions since the $\text{RW}$ spends a long time there, and accumulates information about the energy landscape.
- Conversely the largest value of $R_c$, denoted by $R_c^H$, corresponds to the homogeneous potential $V^H = -x$ in which the walker is never trapped and is quickly

---

3 $\text{RW}(\mu)$, due to conditioning to survival, is likely to stay for $\sim 1/d(y) \ll t^*$ in the trap only.
Reconstructing a random potential from its random walks

Fig. 5: Cumulative probability distribution $Q$ of $R_c(y; V)$ at temperature $T = 1$ and for three tilt values $b$. Full lines are numerical results from $10^5$ samples, and dashed lines are the outcomes from the $\ell$-pattern approximation. The singularities in $Q$ are in one-to-one correspondence with the local potential patterns, some of which are shown. Their abscissas (value of $R_0 \sim \nu(T)$) when $T \to 0$, where $\nu$ is the highest barrier to the right of $y$ in the potential defined by the pattern (fig. 5). When the temperature exceeds the temperature $T_b$ such that $\alpha = 1$ the velocity of the RWer is finite $\nu(T) \sim v(T) > 0$ [2]. The rate of reconstruction $\nu(T)$, i.e. the velocity at which bases can be correctly predicted, is at least $v(T)/R_c^H(T)$, that is,

$$
\nu(T) \geq 1 - \frac{1}{\sinh \frac{1}{\beta_0}} - \frac{1}{b} \sinh \frac{1}{\beta_0} \frac{1}{\cosh \frac{1}{\beta_0} - \frac{1}{b}} / R_c^H(T).
$$

The dependence of $\nu$ upon temperature is sketched in the inset of fig. 4: it is maximal and equal to $\nu^M$ for some temperature $T^M$ realizing a trade-off between fast motion (large velocity) and accurate reading-out (small $R_c$). Even in the small $b$ limit the optimal reconstruction rate is finite, $\nu^M \sim b^2$, by working at high temperature $T^M \sim \frac{1}{b}$, while in the absence of optimization procedure the number of predicted forces scales only as the squared logarithm of the time [18].

**Conclusion.** – We have shown how the number of RWs required for a good reconstruction of the potential can be deduced from the probability of survival of an absorbing RW process. This connection is intimately related to the Feynman-Kac formula for evaluating Wiener functionals over Brownian paths [19]. Here the Brownian paths of interest are the RWs in the true potential, and the functional corresponds to the *a posteriori* probability of predicting the most dangerous potential from those observed RWs [17]. Our result is of practical interest since the survival probability can be estimated through numerical simulations, e.g. in dimension $\geq 2$. Furthermore we have determined, for the special case of the RF model, the optimal “experimental” protocol for reconstruction (number of RWs, duration, temperature).
Our formalism can be extended to models involving discrete variables $f_x$ taking more than 2 values, and with nearest-neighbour (or, more generally, local) interactions. An example is the DNA-inspired model of [9], where the base $f_x$ on each site may take 4 values. This case is interesting in that the rate matrix $r(x \rightarrow x+1; f_x, f_{x+1})$ is degenerate: base pair $(f_x, f_{x+1})$ and its complementary, reverse base pair, e.g. AG and CT, give rise to the same rate. Quasi-“zero mode”-like excitations appear which induce strong correlations between decay constants $R_e(y)$ at different sites $y$.

The application to continuously parametrized potentials, e.g. the RF model with forces taking continuous instead of binary values, is possible too. The aim is now to predict the true potential values up to some accuracy on each site; this accuracy bounds turns de facto the continuous problem into a discretized one, and determines an acceptable neighborhood around $s^*_x$ in the space of signals. The rate function $\omega_V$ is generically parabolic around $s^*_x$, with a curvature matrix called Fisher information matrix [11]. Our approach can be easily extended to the case of a finite delay between two measures of the positions, and Chernoff’s result is recovered in the finite $N$, infinite delay limits [9,14].

***

We are grateful to D. THIRUMALAI for his suggestion of illustrating our formalism on the RF model. This work was partially funded by ANR under contract 06-JCJC-051.

Appendix: rate function for signals. – In this appendix we call $P_R(S \mid \mathbf{V})$ the probability that $R$ RWs in potential $\mathbf{V}$ provide signal $S$. For large $R$ we expect $P_R(S \mid \mathbf{V}) \sim \exp(-R \omega_V(s = S/R))$.

a) Consider two sets of, respectively, $R_1$ and $R_2$ RWs with signals $S_1$ and $S_2$. Then the combination of those $R = R_1 + R_2$ RWs gives signal $S_1 + S_2$. We deduce the inequality $P_{R_1 + R_2}(S_1 + S_2 \mid \mathbf{V}) \geq P_{R_1}(S_1 \mid \mathbf{V}) \times P_{R_2}(S_2 \mid \mathbf{V})$. Taking the logarithm and sending $R \rightarrow \infty$ at fixed $\eta = R_i/R$, $s_i = S_i/R_i$ ($i = 1, 2$) we find that $\omega_V(\eta_1 s_1 + \eta_2 s_2) \geq \eta_1 \omega_V(s_1) + \eta_2 \omega_V(s_2)$.

b) Consider $\pi_R = \int ds P_R(s \mid \mathbf{V}) e^{u s \cdot (\mathbf{V} - \mathbf{V})}$. A saddle-point calculation leads to $-\ln \pi_R/R = \min_s [\omega_V(s) + \mu s \cdot (\mathbf{V} - \mathbf{V})]$ when $R \rightarrow \infty$. Assume time is discrete, with a time delay $\Delta t$ between two measures of the position $x_i$ of the RW, $i = 0, 1, 2, \ldots, \ell = t/\Delta t$. Going back from a sum over signals to a sum over the RWs $X_1, \ldots, X_R$ we have $\pi_R = \pi^R$, where $\pi = \sum_x \prod_{i=1}^{\ell} \omega(x_{i-1} \rightarrow x_i)^{1-\mu} \times q(x_{i-1} \rightarrow x_i)\mu = \sum_x [(q(x)\mu \times (q(x))\mu)]^t(x_i, 0)$, where $q(x \rightarrow y)$ is the probability to go from $x$ to $y$ in time $\Delta t$ in the potential $\mathbf{W}$. Here $\times, (\ldots)^{1-\mu}$ refer to Hadamard i.e. entry-wise product, and $[\ldots]^t$ to matrix power. Taking the limit $\Delta t \rightarrow 0$ we obtain $\pi = \sum_s (e^s M)_{s,0}$, where $M$ is the transition matrix with entries $M(y, x) = r(1-\mu)\mu v(x \rightarrow y)$, and $M(x, x) = \sum_s [(1-\mu)\mu v(x \rightarrow y) + \mu v(x \rightarrow y)]$. Note that $M(x, x) + \sum_y M(y, x) = -d_{\mathbf{V}^{\nu,\mu}}(x) = 0$. The RW process does not conserve probability, and $\pi$ can be interpreted as the probability of survival up to time $t$ of a RWer in potential $(1-\mu)\mathbf{V} + \mu \mathbf{V}$ in the presence of absorbing sites with rates $d_{\mathbf{V}^{\nu,\mu}}(x)$.

REFERENCES

[1] Hughes B. D., Random Walks and Random Environments (Oxford University Press) 1996.
[2] Bouchaud J.-P. and Georges A., Phys. Rep., 195 (1990) 127.
[3] ADELMAN O. and ENRIQUEZ N., Isr. J. Math., 142 (2004) 205.
[4] MERKEL R. et al., Nature, 397 (1999) 50.
[5] FERNANDEZ J. M. and LI H., Science, 303 (2004) 1674.
[6] ESSEZOU-ROUET B., BOCKELMANN U. and HESLOT F., Proc. Natl. Acad. Sci. U.S.A., 94 (1997) 11395.
[7] WOODSIDE M. T. et al., Science, 314 (2006) 1001.
[8] HYEON C. and THIRUMALAI D., Proc. Natl. Acad. Sci. U.S.A., 100 (2003) 10249.
[9] BALDAZZI V. et al., Phys. Rev. Lett., 96 (2006) 128102; Phys. Rev. E, 75 (2007) 011904.
[10] MANOSAS M., COLLIN D. and RITORT F., Phys. Rev. Lett., 96 (2006) 218301.
[11] COVER T. M. and THOMAS J. A., Elements of Information Theory (Wiley) 1991.
[12] EIN-DOR L., ZUK O. and DOMANY E., Proc. Natl. Acad. Sci. U.S.A., 103 (2006) 5923.
[13] ANDERSON T. W. and GOODMAN L., Ann. Math. Stat., 28 (1957) 89.
[14] CHERNOFF H., Ann. Math. Stat., 23 (1952) 493.
[15] DEMBO A. and ZEITOUNI O., Large Deviations Techniques and Applications (Springer-Verlag) 1998.
[16] BOZA L. B., Ann. Math. Stat., 42 (1971) 1992.
[17] COCCO S. and MONASSON R., in preparation.
[18] ANDREOLETTI P., Sinai Walk: A Statistical Aspect, preprint arxiv:math.PR/0612209 (2006).
[19] SCHULMAN L. S., Techniques and Applications of Path Integrals (Wiley, New York) 1981.