Single temperature for Monte Carlo optimization on complex landscapes

Denis Tolkunov\textsuperscript{1,2} and Alexandre V. Morozov\textsuperscript{1,2} \footnote{Department of Physics and Astronomy, Rutgers University, Piscataway, NJ 08854-8019} \footnote{BioMaPS Institute for Quantitative Biology, Rutgers University, Piscataway, NJ 08854-8019} \footnote{(Dated: February 3, 2012)}

We propose a new strategy for Monte Carlo (MC) optimization on rugged multi-dimensional landscapes. The strategy is based on querying the statistical properties of the landscape in order to find the temperature at which the mean first passage time across the current region of the landscape is minimized. Thus, in contrast to other algorithms such as simulated annealing (SA), we explicitly match the temperature schedule to the statistics of landscape irregularities. In cases where this statistics is approximately the same over the entire landscape, or where non-local moves couple distant parts of the landscape, single-temperature MC will outperform any other MC algorithm with the same move set. We also find that in strongly anisotropic Coulomb spin glass and traveling salesman problems, the only relevant statistics (which we use to assign a single MC temperature) is that of irregularities in low-energy funnels. Our results may explain why protein folding in nature is efficient at room temperatures.

PACS numbers: 05.10.Ln, 02.70.Tt, 02.60.Pn, 02.50.Ey

Numerous problems in science and technology such as protein structure prediction, evolution on fitness landscapes, stochastic dynamics of complex systems and machine learning require efficient global optimization of multivariate objective functions or “energies”. The objective function $U(\vec{x})$ can be viewed as a multi-dimensional (multi-$D$) landscape in which a certain quantity (potential energy, free energy, cost function, likelihood of a model) is assigned to every configuration $\vec{x}$ of an arbitrary number $D$ of discrete or continuous state variables. The optimization task is then to find a global minimum (or maximum) on arbitrary landscapes as efficiently as possible. Since exact global optimization methods are not available, various empirical approaches have been devised. A popular class of algorithms is based on the Metropolis MC scheme \cite{1}. This class includes the SA algorithm \cite{2}, as well as simulated tempering \cite{3}, parallel tempering \cite{4}, replica exchange \cite{5}, ensemble MC \cite{6,7} and multicanonical MC \cite{8}. Non-Metropolis schemes for global optimization such as random-cost \cite{9} and genetic algorithms \cite{10} have also been developed. Another class of algorithms enables more efficient exploration of the novel regions of the configuration space by making adaptive changes to the landscape \cite{11,14}.

Unfortunately, the empirical nature of these algorithms makes it impossible to predict which approach would perform best on a given problem. In addition, most algorithms depend on problem-dependent adjustable parameters such as the SA cooling schedule. Here we address these concerns by proposing a universal guiding principle for analyzing global optimization problems. Our interest is not only in developing efficient, parameter-free global optimization schemes, but also in understanding how stochastic simulations run by nature (such as protein folding at constant temperature driven by thermal fluctuations) appear to be so much simpler than corresponding human-designed algorithms.

Our intuition is based on the notion of the \textit{global gradient} that leads towards good solutions (Fig. 1A, upper panel). Landscapes without such a gradient are of the golf-course type or even the “misleading” type in which one has to go up before suddenly finding the global minimum (Fig. 1A, middle and lower panels). In both of the latter scenarios it is necessary to sample $O(N^D)$ possible states, where $N$ is the number of distinct values adopted by a (discretized) state variable. In contrast, global gradients define funnels on the landscape that can in principle be traversed in $O(N)$ steps, making efficient optimization possible. A famous problem of this kind is protein folding \cite{15}, but any landscape in which gradual improvements lead toward a good solution will have the funnel structure. However, in realistic problems the global gradient will be weak and obscured by the local “noise” or irregularities in the objective function (after all, in the absence of such noise any local optimizer would be successful). As a result, the global gradient will be invisible at the smallest scale of a single MC step and can only be detected from the average over a macroscopic local region. If this region is still relatively small, the global gradient will be approximately constant over it. Furthermore, in strongly anisotropic problems the gradient may not be present everywhere but only in the low-energy valleys, whereas the dominant high plateaus surrounding the valleys will be of the golf-course or misleading type.

Thus the global optimization problem can be formulated as diffusion (i.e., Metropolis MC sampling) in a potential which consists of random fluctuations with arbitrary magnitude and correlation length superimposed onto a weak constant gradient. Note that global optimization is different from computing thermodynamic properties, which requires at least approximate equilibration and detailed balance. In contrast, global optimiza-
tion is a strongly non-equilibrium process, with diffusion at any point in the simulation affected only by the landscape features in its immediate neighborhood.

In the absence of random fluctuations, diffusion in a local region \( L \) subject to the constant force \( \mathbf{F} = -\partial U / \partial \mathbf{x} \) is described by a multi-D Fokker-Planck (FP) equation:

\[
\frac{\partial \rho}{\partial t} = D \frac{\partial^2 \rho}{\partial \mathbf{x}^2} - \mathbf{v} \cdot \nabla \rho,
\]

where \( \rho(\mathbf{x}, t) \) is the probability distribution, \( D \) is the diffusion coefficient, and \( \mathbf{v} = D \beta \mathbf{F} \) is the drift velocity \((\beta = 1/T \) is the inverse temperature). We choose a coordinate system in which one of the axes \( (x) \) is parallel to \( \mathbf{F} \).

In this system, Eq. (1) factorizes into a one-dimensional (1D) FP equation with a linear potential and \( D - 1 \) FP equations describing free diffusion. We impose an absorbing boundary perpendicular to \( \mathbf{F} \) and focus on the 1D FP equation with a linear potential.

The speed of propagation along \( \mathbf{F} \) is characterized by the mean first passage time (mfpt) \( \bar{\tau}(x, \beta) \), defined as the mean time required for a particle starting out at \( x \) to reach the absorbing boundary \( b \) for the first time. With a reflecting boundary at \( a < x < b \), the mfpt is given by [16, 17]:

\[
\bar{\tau}(x, \beta) = \frac{1}{D} \int_x^b dx' e^{\beta U(x')} \int_a^{x'} dx'' e^{-\beta U(x'')},
\]

where \( U(x) \) is the 1D potential (objective function) in the direction of \( \mathbf{F} \). For a linear potential in the absence of noise, \( U_0(x) = F(b - x) \) \((F > 0)\), we obtain:

\[
\bar{\tau}_0(x, \beta) = \frac{1}{D F b} \left[ (b - x) + \frac{1}{\beta F} \left( e^{-\beta F(a-b)} - e^{-\beta F(x-a)} \right) \right].
\]

In the \( \beta \to 0 \) (or \( F \to 0 \)) limit the free-diffusion expression is recovered [16]: \( \bar{\tau}_0(x, 0) = (1/2D)((b - a)^2 + (x - a)^2) \). However, for finite \( \beta \) and \( F \) we can always set \( a \) so that the exponential terms on the right-hand side of Eq. (6) are vanishingly small, eliminating the effect of the reflecting boundary on the diffusion process (formally, we take the \( a \to -\infty \) limit):

\[
\bar{\tau}_0(x, \beta) = \frac{b - x}{v},
\]

where \( v = D \beta F \). Now, assuming that the potential consists of the regular part and the irregular part, \( U(x) = U_0(x) + U_1(x) \), and that the characteristic length scale of \( U_1 \), \( l_c \), is much smaller than the size \( L \) of the region over which \( U_0 \) is approximately linear, one can take a spatial average over the irregular part [18]:

\[
\langle \bar{\tau}(x, \beta) \rangle = \frac{1}{D} \int_x^b dx' \int_{-\infty}^{x'} dx'' e^{\beta F(x'' - x')} \times
\]

\[
\int_{x-L/2}^{x+L/2} dx_1 e^{\beta U_1(x_1)} \int_{x''-L/2}^{x''+L/2} dx_2 e^{-\beta U_1(x_2)},
\]

where \( l_c \ll L \ll L \). Under these conditions and the additional assumption that \( U_1 \) statistics does not change over \( L \), the two spatial averages are independent of each other and of \( x' \), \( x'' \), yielding:

\[
\langle \bar{\tau}(x, \beta) \rangle = H(\beta) \bar{\tau}_0(x, \beta),
\]

where \( H(\beta) = \int_{-\infty}^{\infty} dU_1 \rho(U_1) e^{\beta U_1} \int_{-\infty}^{\infty} dU_2 \rho(U_2) e^{-\beta U_2} \) (we have switched from \( x_1 \) and \( x_2 \) to \( U_1 \equiv U_1(x_1) \) and \( U_2 \equiv U_1(x_2) \) in the spatial averages). Furthermore, \( H(\beta) = \int_{-\infty}^{\infty} d\Delta P(\Delta) e^{\beta \Delta} \), where \( P(\Delta) \) is the distribution of \( \Delta = U_1 - U_2 \) for \( x_1, x_2 \) constrained by \( |x_1 - x_2| > l_c \). The last condition guarantees that \( P(\Delta) \) is independent of \( |x_1 - x_2| \).

Clearly, if mfpt along \( \mathbf{F} \) is minimized for all local regions \( L \) with the constant gradient, the total time to reach a good solution will also be minimized. The inverse temperature \( \beta^* \) that minimizes mfpt is given by:

\[
\frac{dH(\beta)}{d\beta} \bigg|_{\beta^*} = \frac{H(\beta^*)}{\beta^*}.
\]

Note that \( \beta^* \) is independent of \( F \). Eq. (6) can be used to find \( \beta^* \) numerically for any \( P(\Delta) \). If \( P(U_1) = N(0, \sigma^2) \), \( P(\Delta) = N(0, 2\sigma^2) \) and \( H(\beta) = e^{\beta^2 \sigma^2} \), yielding \( \beta^* = 1/\sqrt{2}\sigma \). With \( T \ll T^* = \sqrt{2}\sigma \) the diffusing particle gets stuck in local minima \( (\bar{\tau}) \sim e^{\sigma^2/T^2} \), while for \( T \gg T^* \) diffusion is no longer optimally along the gradient of \( U_0 \) (\( \bar{\tau} \sim T^2 \)) (Fig. 1C). \( \langle \bar{\tau}(x, \beta^*) \rangle \sim \sqrt{F} / \beta \), indicating that diffusion is impeded by noise and aided by the gradient. If \( U_1 = 0 \) everywhere, \( P(\Delta) = \delta(\Delta) \) and \( T^* = 0 \). Thus, as expected, the optimal solution in the absence of noise is to roll down the potential at zero temperature. However, \( \bar{\tau}_0(x, \infty) = 0 \) since Eq. (2) does not accurately describe the ballistic regime or strong forces. Finally, if \( F \to 0 \), \( T^* = \infty \) and \( \langle \bar{\tau}(x, 0) \rangle \) reduces to the expression for free...
diffusion, although any \( T > \sqrt{2}\sigma \) will work almost as well (Fig. 1B).

Thus, if \( U_1 \) statistics is approximately constant and isotropic throughout the landscape, there is a unique MC temperature for the most efficient minimization of the objective function (the anisotropic case will be presented elsewhere). If not, different parts of the landscape are to be assigned different temperatures matched to the \( U_1 \) statistics. All other schemes such as SA will yield suboptimal performance. In fact, if the amplitude of \( U_1 \) increases with decreasing \( U_0 \), our prescription calls for increasing the temperature as the simulation progresses –– the exact opposite of the SA cooling schedule [2].

Thus, to find the best MC temperature \( T^* \), we need to estimate \( P(U_1) \) by sampling in the neighborhood of the current state (\( P(U_1) \) can be resampled periodically during the simulation and \( T^* \) recomputed). Unfortunately, it is difficult to extract \( P(U_1) \) from \( P(U) \) by detrending multi-D walks. Instead, we consider \( P(\Delta) \), where \( \Delta = U(\vec{x} + \Delta\vec{x}) - U(\vec{x}) \) and \( \Delta\vec{x} \) is a single MC step with constant length. \( |\Delta\vec{x}| \) can be made so small that the contribution from \( U_0 \) is negligible. With uncorrelated noise (\( |\Delta\vec{x}| > l_c \)), \( P(\Delta) = P(\Delta) \) and Eq. 10 can be applied immediately. However, if \( |\Delta\vec{x}| < l_c \), \( U_1 \) is smooth at the scale of a single MC step and \( P(\Delta) \) and \( T^* \) will depend on the move set. Indeed, \( T^* \sim |\Delta\vec{x}| \) if the MC steps are so fine that \( U_1(\vec{x} + \Delta\vec{x}) - U_1(\vec{x}) \) is approximately linear. Nonetheless, we find that for complex landscapes where \( U(\vec{x}) \) is a sum of many independent terms, \( P(\Delta) \) quickly adopts a Gaussian shape if the sampling is over a region \( \gg l_c \). Since the MC walk is memoryless, \( T^* = \sqrt{2}\sigma \) still holds but now \( \sigma \) depends on the move set. As \( |\Delta\vec{x}| \) increases beyond \( l_c \), \( \sigma \) converges to a universal value.

We have tested our approach on a set of standard functions often used to check performance of global optimization algorithms [19] (Fig. 2 Table I). To estimate \( P(\Delta) \), we use 50 trials with randomized starting positions and \( 10^5 \) random steps each (all steps are accepted). These parameters ensure that \( P(\Delta) \) is close to a Gaussian (with the rate of convergence dependent on the long-range order in the landscape and on the complexity of the potential function), and \( T^* \) is predicted as its \( \sigma \). This simple procedure allows us to guess the best temperature correctly (Table I), despite the fact that the landscapes are correlated and anisotropic and the gradient is not guaranteed to be weak. Note that \( U_1 \) is the 4D Griewank (G) function (Table I). For each \( T \), \( N_{\text{trials}} \) independent trajectories with \( N_{\text{iter}} \) MC steps each were created by Metropolis MC sampling. The lowest energy \( U_{\text{best}} \) from each trajectory is shown as a grey dot. Red dots are the average of \( U_{\text{best}} \) at a given \( T \). Inset: Histogram of energy differences from a random sample of the landscape, with a Gaussian fit (blue solid curve).

FIG. 2: (Color online) Distribution of best predicted energies as a function of temperature for the 4D Griewank function (Table I). For each \( T \), \( N_{\text{trials}} \) independent trajectories with \( N_{\text{iter}} \) MC steps each were created by Metropolis MC sampling. The lowest energy \( U_{\text{best}} \) from each trajectory is shown as a grey dot. Red dots are the average of \( U_{\text{best}} \) at a given \( T \). Inset: Histogram of energy differences from a random sample of the landscape, with a Gaussian fit (blue solid curve).

TABLE I: Predicted and computed optimal temperatures for standard test functions [19]: the 4D Griewank (G) function \( U(\vec{x}) = 1 + \frac{1}{4000} \sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n} \cos\left(\frac{\pi}{\sqrt{2}} x_i\right) \), \( x_i \in [-600, 600], \forall i \), the 4D Rastrigin (R) function \( U(\vec{x}) = 4 + \sum_{i=1}^{n} (x_i^2 - \cos(18x_i)) \), \( x_i \in [-5, 5], \forall i \) and the 4D Ackley (A) function \( U(\vec{x}) = 20 + e - \exp\left(-0.2 \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2}\right) - \exp\left(\frac{1}{n} \sum_{i=1}^{n} \cos(2\pi x_i)\right) \), \( x_i \in [-32.8, 32.8], \forall i \). All three functions have multiple local minima and a single global minimum located at \( \vec{x} = 0 \) (\( U(0) = 0 \)). Each MC step is taken in a random direction and has a constant length of 1.0 (G), 0.05 (R) and 0.3 (A); \( T_{\text{pred}} \) is based on a Gaussian fit to the histogram of \( \Delta \). \( T_{\text{comp}} \) is the temperature at which the average \( U_{\text{best}} \) is at minimum. \( U_{\text{best}} \) distribution at each \( T \) was estimated using \( N_{\text{trials}} \) Metropolis MC runs of \( N_{\text{iter}} \) steps each. The lowest \( U_{\text{best}} \) at \( T_{\text{pred}}, U_{\text{best}}(T_{\text{pred}}), \) was obtained using \( N_{\text{iter}} = 6 \times 10^4 \) (G), \( 2 \times 10^4 \) (R), and \( 10^4 \) iterations (A), for improved sampling of the tail of the \( U_{\text{best}}(T_{\text{pred}}) \) distribution. In all cases \( x_{\text{min}} \) was within a single step from the global minimum.
Euclidean distances to compute $U$ and employ non-local moves in which a segment of the trajectory is chosen at random and the direction in which all cities within that segment are traversed is inverted [2]. To reduce the degeneracy of low-scoring solutions, we start all trajectories from the same city. The TS landscape has a complex multi-funnel structure (Fig. 3C) with high plateaus that dominate the landscape, so that only $\Delta$ statistics within the funnels is relevant. As in CSG, we employ the funnel-sampling algorithm (with $N_{\text{iter}} = 2 \times 10^3$) to obtain $\sigma = 7.6 \times 10^{-4}$. This value is confirmed by scanning a range of temperatures with fixed-temperature Metropolis MC (Fig. 3D).

Throughout this paper we have focused on minimizing mfpt. However, instead one may want to maximize the fraction of runs with $U_{\text{best}}$ (the lowest energy from each MC trajectory) below a certain cutoff. The tail of the $U_{\text{best}}$ distribution at a given $T$ is affected by both its mean and standard deviation $\sigma'$, making it possible that the temperature with the best mean is not the same as the temperature optimized for yielding extremely low-energy solutions. However, from Fig. 2 and Figs. 3B,D we see that $\sigma'$ varies with $T$ rather slowly. As a result, the mfpt-based $T^*$ remains valid, but in some cases the interplay between the mean and $\sigma'$ may make temperatures in a small range around $T^*$ equally acceptable.

If $U_1$ statistics is the same everywhere, global optimization is carried out most efficiently by MC runs with a fixed temperature $T^*$. However, if the nature of irregularities changes across the landscape, two scenarios are possible. First, if $P(U_1)$ stays approximately the same in $\gg \mathbf{L}$, the best temperature can be found for each region but needs to be updated as the landscape is traversed. Mixing statistics from multiple regions will yield a single $T^*$ that will not be the absolute best solution but may still be a good approximate one. Second, it is possible that different scales are mixed in a region $\ll \mathbf{L}$, e.g. due to anisotropy. In this case $P(\Delta)$ will be non-Gaussian but Eq. 10 still applies, yielding a single $T^*$. Using a single temperature works especially well with non-local steps such as those employed in the TS problem, which can traverse a sizable part of the landscape in a single leap. Indeed, we find that even a multi-scale TS problem, in which cities are clustered rather than randomly distributed [2], has a unique best temperature with non-local steps.

Our procedure can be viewed as an extension of the SA algorithm, which the priori assumes that all scales are present in the problem and, moreover, that they appear according to a specific cooling schedule. While SA may be the best way to proceed if the properties of the landscape are completely unknown, querying some of the landscape statistics allows us to improve on the “one size fits all” SA technique by matching a given landscape to the appropriate temperature(s). We look forward to applying our approach to protein structure prediction and
other global optimization challenges.

This research was supported by National Institutes of Health (HG 004708) and by an Alfred P. Sloan Research Fellowship to AVM.

* Corresponding author: morozov@physics.rutgers.edu

[1] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, J. Chem. Phys. 21, 1087 (1953).
[2] S. Kirkpatrick, C. Gelatt, Jr., and M. Vecchi, Science 220, 671 (1983).
[3] E. Marinari and G. Parisi, Europhys. Lett. 19, 451 (1992).
[4] U. Hansmann, Chem. Phys. Lett. 281, 140 (1997).
[5] K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn. 65, 1604 (1996).
[6] F.-M. Dittes, Phys. Rev. Lett. 76, 4651 (1996).
[7] B. Hesselbo and R. Stinchcombe, Phys. Rev. Lett. 74, 2151 (1995).
[8] B. Berg and T. Neuhaus, Phys. Lett. B 267, 249 (1991).
[9] B. Berg, Nature 361, 708 (1993).
[10] D. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning (Addison Wesley, Reading, MA, 1989).
[11] J. Barhen, V. Protopopescu, and D. Reister, Science 276, 1094 (1997).
[12] W. Wenzel and K. Hamacher, Phys. Rev. Lett. 82, 3003 (1999).
[13] K. Hamacher, Europhys. Lett. 74, 944 (2006).
[14] D. Cvijović and J. Klinowski, Science 267, 664 (1995).
[15] J. Bryngelson, J. Onuchic, N. Socci, and P. Wolynes, Proteins: Struc. Func. Genet. 21, 167 (1995).
[16] S. Lifson and J. Jackson, J. Chem. Phys. 36, 2410 (1962).
[17] G. Weiss, Adv. Chem. Phys. 13, 1 (1966).
[18] R. Zwanzig, Proc. Natl. Acad. Sci. USA 85, 2029 (1988).
[19] A. Törn and Å. Zilinskas, Global Optimization (Springer-Verlag, Berlin, Germany, 1989).