The Leverhulme Research Centre for Functional Materials Design

Geometry of the n-torus entropic trust region packing algorithm

Miloslav Torda

Department of Computer Science, *University of Liverpool*

The Leverhulme Research Centre for Functional Materials Design, *University of Liverpool*

miloslav.torda@liverpool.ac.uk

www.milotorda.net

Abstract

Stochastic relaxation is a well-known approach to solve problems in machine learning and artificial intelligence in cases of complicated optimization landscapes. To solve the problem of densest packings of closed subsets of the n-dimensional Euclidean space restricted to a Crystallographic Symmetry Group (CSG), we construct, by the means of stochastic relaxation, a non-Euclidean trust region algorithm as a variant of the natural gradient learning with adaptive selection quantile fitness rewriting, the Entropic Trust Region Packing Algorithm (ETRPA). Since CSGs induce a toroidal topology on the configuration space, the ETRPA search is performed on a statistical manifold of extended multivariate von Mises probability distributions, a parametric family of probability measures defined on an n-dimensional torus. Using the connection with a generalized proximal method we examine the geometry of ETRPA. We provide a characterization of the algorithm via local dual geodesic flows which in fact maximize stochastic dependence among elements of the extended multivariate von Mises distributed random vector. Thus, the ETRPA's evolutionary computing, simulated annealing and recurrent neural computing theoretical background can be interpreted in terms of more general graphical interaction models.

Problem statement

Given K a compact subset of $\mathbb{R}^n$ and G a Crystallographic Symmetry Group (CSG), that is a discrete group of isometries of $\mathbb{R}^n$, find the configuration $\mathcal{K}_G$ consisting only of non-overlapping orbits of K with respect to G-action on $\mathbb{R}^n$, such that the ratio of the filled to the whole space is maximized over the whole isomorphism class $\mathcal{G}$ of G. Formally,

$$\mathcal{K}_{\max} = \underset{\mathcal{K}_G : G \in \mathcal{G}}{\operatorname{argmax}} \, \rho\left(\mathcal{K}_G\right), \; \mathcal{G} = \{H | H \cong G\}. \tag{1}$$

where $0 \leq \rho(\cdot) \leq 1$ is the packing density.

Our work is motivated by the problem of Crystal Structure Prediction, in which given some molecular shape, the goal is to predict a synthesizable periodic structure.

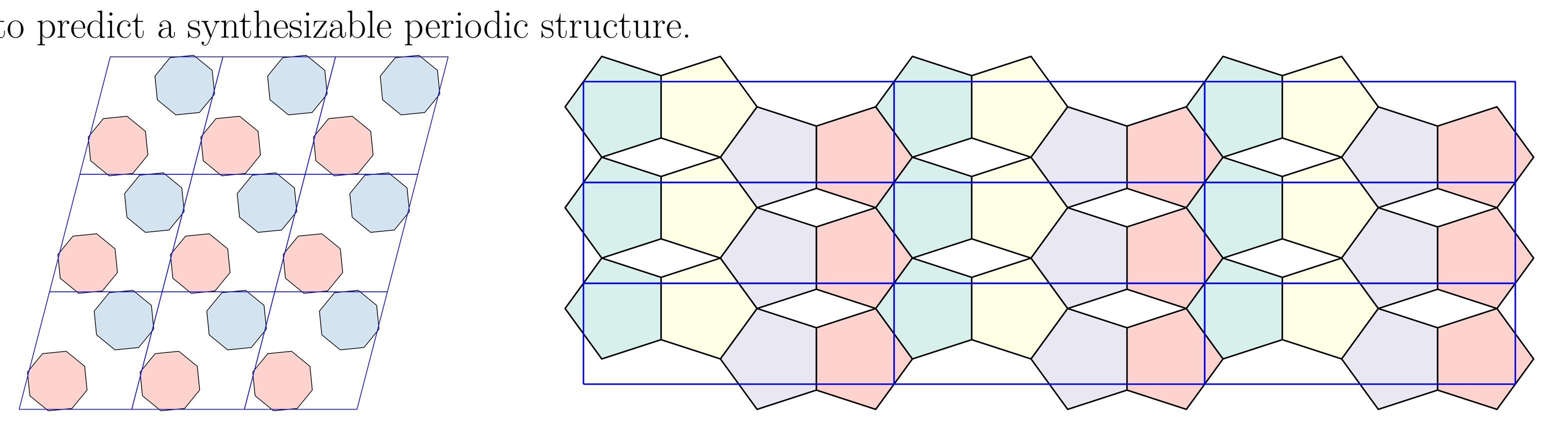

Figure 1: 2-dimensional CSG packings: (Left) A packing configuration where $\mathcal{G}$ is the $p2$ plane group and K is a regular octagon from the initial sampling of configurations visualized in Fig.3. (Right) Densest packing configuration found by ETRPA where $\mathcal{G}$ is the $p2mg$ plane group and K is a regular pentagon. Blue lines denote the crystal lattice. Colors denote symmetry operations represented by cosets gL where $g \in G$ is an element of 2-dimensional CSG and L is a lattice group associated with G.

Etropic Trust Region Packing Algorithm (ETRPA)

Stochastic relaxation and the entropic trust region

We transform Eq. (1) via stochastic relaxation [3] to the problem of $\tilde{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} J(\boldsymbol{\theta})$, with $J(\boldsymbol{\theta}) := E[\mathbf{F}|\boldsymbol{\theta}] = \int_{\mathcal{X}} \mathbf{F}(\mathbf{x}) dP(\boldsymbol{\theta})$ being the expected fitness $\mathbf{F}$ of the packing density ρ under some probability measure $dP(\boldsymbol{\theta})$ from family of probability measures $S = \{dP(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^n\}$ defined on some configuration space $\mathcal{X}$ and solve it using first order trust region method where the trust region radius is given by the Kullback-Leibler divergence from $P_{\boldsymbol{\theta}}$ to $P_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}$, defined by $D_{KL}\left(P_{\boldsymbol{\theta}} \,||\, P_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}\right) = \int_{\mathcal{X}} \ln\left(\frac{dP(\boldsymbol{\theta})}{dP(\boldsymbol{\theta}+\delta\boldsymbol{\theta})}\right) dP(\boldsymbol{\theta})$. The update equation reads

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \Delta^t \frac{\widetilde{\nabla} J(\boldsymbol{\theta}^t)}{\| \widetilde{\nabla} J(\boldsymbol{\theta}^t) \|_{\mathcal{I}_{\boldsymbol{\theta}^t}}}, \tag{2}$$

where $\widetilde{\nabla} J(\boldsymbol{\theta}) = \mathcal{I}_{\boldsymbol{\theta}}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ is the natural gradient [1], $\| \cdot \|_{\mathcal{I}_{\boldsymbol{\theta}}}$ is the norm associated with the inner product induced by the Fisher metric tensor $\mathcal{I}_{\boldsymbol{\theta}}$ with elements $\mathcal{I}_{\boldsymbol{\theta}ij} = \int_{\mathcal{X}} \frac{\partial \ln(p(\boldsymbol{\theta}))}{\partial \theta_i} \frac{\partial \ln(p(\boldsymbol{\theta}))}{\partial \theta_j} dP(\boldsymbol{\theta})$, $p(\boldsymbol{\theta}) = \frac{dP(\boldsymbol{\theta})}{d\boldsymbol{\nu}}$ is the Radon-Nikodym derivative with respect to a reference measure $\boldsymbol{\nu}$ on $\mathcal{X}$ and $\nabla_{\boldsymbol{\theta}}$ is the Euclidean gradient in $\boldsymbol{\theta}$-coordinates.

Adaptive selection quantile

We implement a truncation selection transformation for $q \geq 1$ of the fitness function $\mathbf{F}$

$$q \, \mathbf{1}_{\mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q}}}(\mathbf{x}) = \begin{cases} q & \text{if } \mathbf{F}(\mathbf{x}) \geq \mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q}}, \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q}}$ is the $P_{\tilde{\boldsymbol{\theta}}}$'s $(q-1)$th q-quantile of the fitness $\mathbf{F}$ and $\mathbf{1}_{\cdot}(\cdot)$ is the indicator function. Assuming $P_{\tilde{\boldsymbol{\theta}}}$ is from the exponential family statistical model then, for a fixed q^*

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \,|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = \boldsymbol{\mu}_{\mathbf{F}_{1-\frac{1}{q^*}}} - \boldsymbol{\mu}, \tag{3}$$

where $\boldsymbol{\mu}_{\mathbf{F}_{1-\frac{1}{q^*}}} = \int_{\mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q^*}}}^{\infty} \mathbf{F}^{-1}(y) \exp\left\{\tilde{\boldsymbol{\theta}}^\intercal \mathbf{t}\left(\mathbf{F}^{-1}(y)\right) - \psi(\tilde{\boldsymbol{\theta}}) + \ln(q^*)\right\} dP_0 \circ \mathbf{F}^{-1}(y)$ is the expectation parametrisation of the truncated exponential probability distribution derived from $P_{\tilde{\boldsymbol{\theta}}}$'s (q^*-1)th q-quantile of the fitness $\mathbf{F}$ and $\boldsymbol{\mu} = \int \mathbf{x} \exp\{\tilde{\boldsymbol{\theta}}^\intercal \mathbf{t}(\mathbf{x}) - \psi(\tilde{\boldsymbol{\theta}})\} dP_0$ is the expectation parametrisation of $P_{\tilde{\boldsymbol{\theta}}}$.

Since varying q influences the magnitude of Eq. (3), as visualized in Fig. 2, we introduce an adaptive quantile into the optimization schedule by

$$q_{t+1} = q_t \exp\{\beta \cos(\alpha^t)\},$$

where α^t is the angle between three consecutive updates of Eq. (2) considered as points on the statistical manifold S. Parameter q_t can be regarded as an equivalent of the temperature parameter in the simulated annealing schedule.

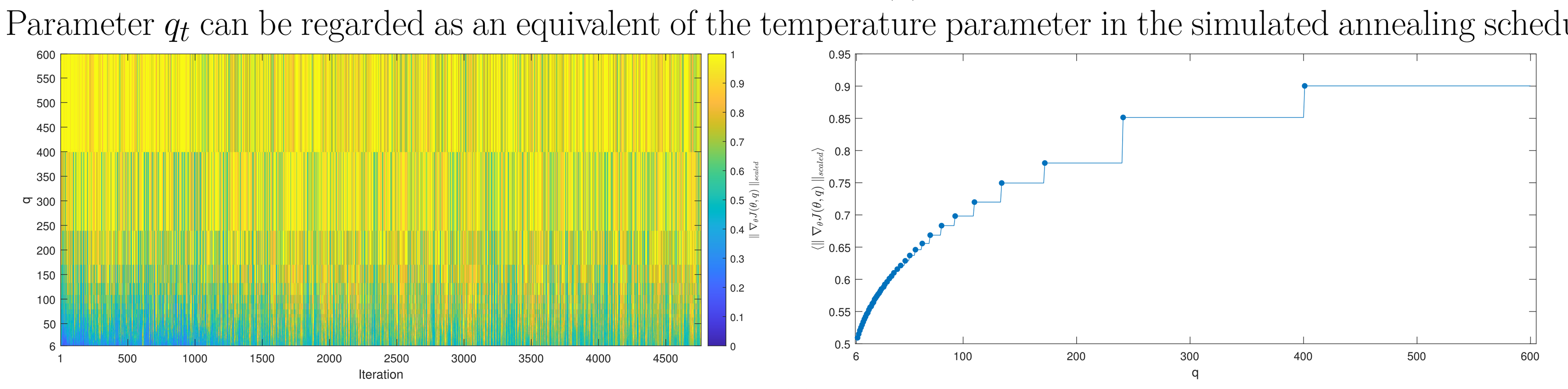

Figure 2: Influence of the q-quantile on the fitness gradient (Eq. (3)) in one run of the ETRPA on the problem of densest $p2$-packing of a regular octagon. (Left) Relationship between selection quantile q and the scaled expected fitness gradient $\| \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t, q) \|_{\text{scaled}} = \frac{\|\boldsymbol{\mu}^t_{\mathbf{F}_{1-\frac{1}{q}}} - \boldsymbol{\mu}^t\|_1}{\|\boldsymbol{\mu}^t_{\mathbf{F}_{1-\frac{1}{q^{\max}}}} - \boldsymbol{\mu}^t\|_1}$, where $q^{\max} = \operatorname{argmax}_q \| \boldsymbol{\mu}^t_{\mathbf{F}_{1-\frac{1}{q}}} - \boldsymbol{\mu}^t \|_1, \; q = 6, \ldots, 600$. (Right) Scaled expected fitness averaged through all iterations $<\| \nabla_\theta J(\theta, q) \|_{scaled}>$ where markers denote changes in the value.

Extended multivariate von Mises distribution

The lattice subgroup L of a CSG induces quotient space $\mathbb{R}^n/\mathrm{L}$. Thus, we restrict ETRPA's search to a statistical model consisting of probability distributions defined on an n-torus. This is done by extending the multivariate von Mises model [4] to the family of distributions with the probability density function

$$f(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{D}) = \frac{1}{Z(\boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{D})} \exp\left\{\boldsymbol{\kappa}^\intercal c(\boldsymbol{\theta}-\boldsymbol{\mu}) + \frac{1}{2}\begin{bmatrix} c(\boldsymbol{\theta}-\boldsymbol{\mu}) \\ s(\boldsymbol{\theta}-\boldsymbol{\mu}) \end{bmatrix}^\intercal \mathbf{D} \begin{bmatrix} c(\boldsymbol{\theta}-\boldsymbol{\mu}) \\ s(\boldsymbol{\theta}-\boldsymbol{\mu}) \end{bmatrix}\right\} \tag{4}$$

$$c(\boldsymbol{\theta}-\boldsymbol{\mu}) = [cos(\theta_1-\mu_1), \ldots, cos(\theta_n-\mu_n)]^\intercal \qquad \mathbf{D} = \begin{bmatrix} \mathbf{D}^{cc} & \mathbf{D}^{cs} \\ \mathbf{D}^{cs\intercal} & \mathbf{D}^{ss} \end{bmatrix}$$

$$s(\boldsymbol{\theta}-\boldsymbol{\mu}) = [sin(\theta_1-\mu_1), \ldots, sin(\theta_n-\mu_n)]^\intercal$$

$$0 \leq \theta_i \leq 2\pi \quad 0 \leq \mu_i \leq 2\pi \quad 0 \leq \kappa_i,$$

where the diagonal elements of $\mathbf{D}^{cc}$, $\mathbf{D}^{ss}$ and $\mathbf{D}^{cs}$ are zero and $\mathbf{D}$ is a $2n \times 2n$ real-valued symmetric matrix that controls cosine-cosine, sine-sine and cosine-sine interactions. Additionally, we remove non-identifiability of Eq. (4) by the restriction $\kappa_i > 0$ and introduce the exponential family re-parametrisation of the extended multivariate von Mises distribution Eq.(4) to

$$f(\boldsymbol{\theta}|\boldsymbol{\eta}, \mathbf{E}) = \exp\left\{\begin{bmatrix} c(\boldsymbol{\theta}) \\ s(\boldsymbol{\theta}) \end{bmatrix}^\intercal \boldsymbol{\eta} + \operatorname{vec}\left(\begin{bmatrix} c(\boldsymbol{\theta}) \\ s(\boldsymbol{\theta}) \end{bmatrix} \begin{bmatrix} c(\boldsymbol{\theta}) \\ s(\boldsymbol{\theta}) \end{bmatrix}^\intercal\right)^\intercal \operatorname{vec}(\mathbf{E}) - \psi(\boldsymbol{\eta}, \mathbf{E})\right\}$$

where $\operatorname{vec}(\cdot)$ denotes vectorization.

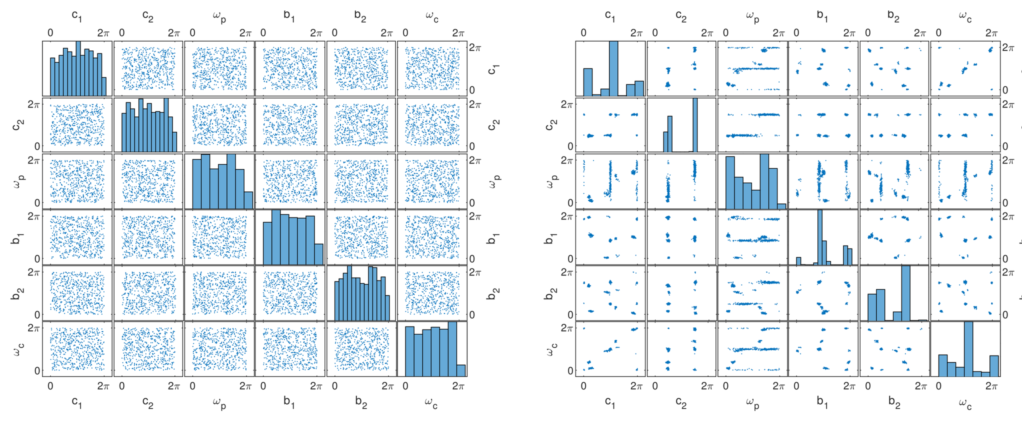

Figure 3: 2-dimensional projections along coordinate axes and histograms of univariate marginals corresponding to the respective optimization variables of 600 realizations of the exponential re-parametrisation of the extended multivariate von Mises distribution defined on an 6-torus. (Left) Initial distribution and (Right) output distribution of a single run of the ETRPA on the problem of densest $p2$-packing of a regular octagon with following optimization variables: octagon centroid fractional coordinates c_1 and c_2 in the $p2$ fundamental region, angle of rotation of the octagon ω_p, lengths of the lattice generators b_1 and b_2, and angle between lattice generators ω_c.

Geometry of ETRPA

Recasting ETRPA as a solution of a maximin problem, which has an interpretation in terms of maximizing multi-information [2] and can be observed visually in Fig. 3, allows us to view ETRPA (information) geometrically as follows. The aforementioned maximin problem reads

$$\max_{P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}} \min_{P_{\boldsymbol{\theta}^1} \in S^1} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \,||\, P_{\boldsymbol{\theta}^1}) \tag{5}$$

and is induced by dual gradient flows between two codimension 1 submanifolds of the ambient statistical manifold

$$S = \bigcup_{q \in [1;\infty)} S^q, \; S^q = \{dP^q(\boldsymbol{\theta}) | \, \boldsymbol{\theta} \in \boldsymbol{\Theta}^q \subset \mathbb{R}^n\},$$

where $dP^q(\boldsymbol{\theta}) = \begin{cases} \exp\left\{\boldsymbol{\theta}^\intercal \mathbf{t}\left(\mathbf{F}^{-1}(y)\right) - \psi(\boldsymbol{\theta}) + \ln(q)\right\} dP_0 \circ \mathbf{F}^{-1}(y) & y \geq \mathbf{F}^{\boldsymbol{\theta}}_{1-\frac{1}{q}} \\ 0 & \text{otherwise} \end{cases}.$

Indeed, the Euclidean fitness gradient Eq. (3) for a fixed q^* determines proximal map maximization

$$\boldsymbol{\theta}^{t+1} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \left\{\boldsymbol{\theta}^\intercal \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t) - \frac{1}{\epsilon} D_{KL}\left(P_{\boldsymbol{\theta}^t} \,||\, P_{\boldsymbol{\theta}}\right)\right\},$$

and hence natural gradient updates in dual coordinate systems

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \epsilon^t \mathcal{I}_{\boldsymbol{\theta}^t}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t), \; \boldsymbol{\mu}^{t+1} = \boldsymbol{\mu}^t + \epsilon^t \mathcal{I}_{\boldsymbol{\mu}^t}^{-1} \nabla_{\boldsymbol{\mu}} J(\boldsymbol{\mu}^t), \tag{6}$$

which when lifted to S via $\boldsymbol{\mu}_{\mathbf{F}_{1-\frac{1}{q^*}}} = q^* \boldsymbol{\mu} - (q^*-1) \boldsymbol{\mu}_{\mathbf{F}_{\frac{1}{q^*}}}$ for the natural gradient in $\boldsymbol{\mu}$-coordinates, give the flows. Moreover, for $\epsilon^t = \frac{\Delta^t}{\sqrt{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t)^\intercal (\mathcal{I}_{\boldsymbol{\theta}^t})^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t)}}$ the $\boldsymbol{\theta}$-coordinate natural gradient update part of Eq. (6) coincides with trust region update Eq. (2). In addition, submanifolds S^q induce a dually flat structure on S and following the information projection theorems, solving Eq. (5) is equivalent to solving

$$\max_{P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \,||\, P_{\hat{\boldsymbol{\theta}}^1_s}) \tag{7}$$

where $P_{\hat{\boldsymbol{\theta}}^1_s} \in S^1$ is a model without any interaction whatsoever corresponding to the maximum entropy estimate of $P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}$. Consequently, $D_{KL}(\cdot \,||\, P_{\hat{\boldsymbol{\theta}}^1_s})$ in Eq.(7) is considered as a special case of multi-information, a measure of stochastic dependence in complex systems.

Further information

Torda, M., Goulermas, J. Y., Púček, R., & Kurlin, V., Entropic trust region for densest crystallographic symmetry group packings, arXiv:2202.11959, (2022).

References

[1] S.-I. Amari, *Natural gradient works efficiently in learning*, Neural computation, 10 (1998), pp. 251–276.

[2] N. Ay and A. Knauf, *Maximizing multi-information*, Kybernetika, 42 (2006), pp. 517–538.

[3] S. Geman and D. Geman, *Stochastic relaxation, gibbs distributions and the bayesian restoration of images*, Journal of Applied Statistics, 20 (1993), pp. 25–62.

[4] K. V. Mardia, G. Hughes, C. C. Taylor, and H. Singh, *A multivariate von mises distribution with applications to bioinformatics*, Canadian Journal of Statistics, 36 (2008), pp. 99–109.

The authors thank the Leverhulme Trust for funding this research via the Leverhulme Research Centre for Functional Materials Design