UNIVERSITY OF LIVERPOOL

LEVERHULME TRUST

The Leverhulme Research Centre for Functional Materials Design

Maximally Dense Crystallographic Symmetry Group Packings through Entropic Trust Region

An Information Geometric Perspective

Miloslav Torda

CaLISTA Kick-off meeting

June 8, 2023

diamond · Hartree Centre · Science & Technology Facilities Council · Imperial College London · King Abdullah University of Science and Technology · University of Southampton

- M. TORDA, J. Y. GOULERMAS, R. PÚČEK AND V. KURLIN, *Entropic trust region for densest crystallographic symmetry group packings*, arXiv:2202.11959. To appear in SIAM Journal on Scientific Computing.

Crystal Structure Prediction (CSP)
motivation

[1] Chen W. et. al. (2008) Molecular orientation transition of organic thin films on graphite: the effect of intermolecular electrostatic and interfacial dispersion forces, Chemical Communications, pp. 4276–4278.

[2] Zhao Y. et. al. (2015) Molecular self-assembly on two-dimensional atomic crystals: insights from molecular dynamics simulations, The Journal of Physical Chemistry Letters, 6, pp. 4518–4524.

Crystal Structure Prediction (CSP)
motivation

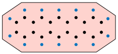
- An approach to accelerate Molecular CSP solvers:
  - Energy minimization → Geometric packing density maximization

[1] Chen W. et. al. (2008) Molecular orientation transition of organic thin films on graphite: the effect of intermolecular electrostatic and interfacial dispersion forces, Chemical Communications, pp. 4276–4278.

[2] Zhao Y. et. al. (2015) Molecular self-assembly on two-dimensional atomic crystals: insights from molecular dynamics simulations, The Journal of Physical Chemistry Letters, 6, pp. 4518–4524.

Crystal Structure Prediction (CSP)
motivation

- An approach to accelerate Molecular CSP solvers:
  - Energy minimization → Geometric packing density maximization

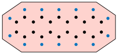


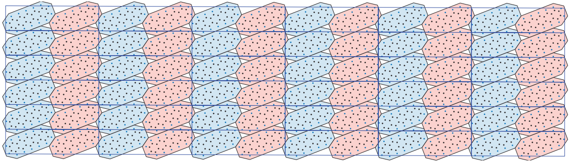
[1]Chen W. et. al. (2008) Molecular orientation transition of organic thin films on graphite: the effect of intermolecular electrostatic and interfacial dispersion forces, Chemical Communications, pp. 4276–4278.

[2]Zhao Y. et. al. (2015) Molecular self-assembly on two-dimensional atomic crystals: insights from molecular dynamics simulations, The Journal of Physical Chemistry Letters, 6, pp. 4518–4524.

Crystal Structure Prediction (CSP) motivation

- An approach to accelerate Molecular CSP solvers:
  - Energy minimization → Geometric packing density maximization





(**Left**) A geometric representation of pentacene as the convex hull of the atomic positions of the molecule with an offset given by hydrogen's van der Waals radius of 1.09Å. The dots symbolize atomic positions of (blue) hydrogen and (black) carbon. (**Right**) Visualization of the ETRPA output configuration of the densest $p2$-packing of the pentacene representation with density of 0.9533821 and resembles the configuration of single layer pentacene thin-film on graphite surface[1], and via MD simulation of self-assembly of a disordered system of pentacene molecules on a graphene surface driven by the minimization of molecule-molecule interactions using the Lennard–Jones potential[2].

[1] Chen W. et. al. (2008) Molecular orientation transition of organic thin films on graphite: the effect of intermolecular electrostatic and interfacial dispersion forces, Chemical Communications, pp. 4276–4278.

[2] Zhao Y. et. al. (2015) Molecular self-assembly on two-dimensional atomic crystals: insights from molecular dynamics simulations, The Journal of Physical Chemistry Letters, 6, pp. 4518–4524.

Crystallographic Symmetry Group (CSG) packing

- **Configuration space**: Crystallographic Symmetry Group (CSG)

Crystallographic Symmetry Group (CSG) packing

- **Configuration space**: Crystallographic Symmetry Group (CSG)
  - Discrete group of isometries of $\mathbb{R}^n$ containing a lattice subgroup

Crystallographic Symmetry Group (CSG) packing

- **Configuration space**: Crystallographic Symmetry Group (CSG)
  - Discrete group of isometries of $\mathbb{R}^n$ containing a lattice subgroup
- CSG packing

$$\mathcal{K}_G = \bigcup_{g \in G} gK,$$

$$\operatorname{int}(g_i K) \cap \operatorname{int}(g_j K) = \emptyset, \quad \forall\, g_i,\ g_j \in G,\ \ g_i \neq g_j$$

  - G - CSG
  - K - Compact subset of $\mathbb{R}^n$

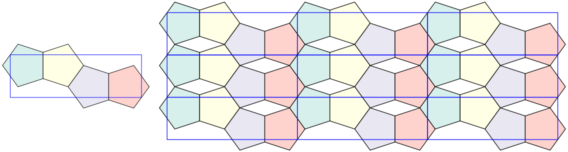
Crystallographic Symmetry Group (CSG) packing

- **Configuration space**: Crystallographic Symmetry Group (CSG)
  - Discrete group of isometries of $\mathbb{R}^n$ containing a lattice subgroup
- CSG packing

$$\mathcal{K}_G = \bigcup_{g \in G} gK,$$

$$\operatorname{int}(g_i K) \cap \operatorname{int}(g_j K) = \emptyset, \quad \forall\, g_i,\ g_j \in G,\ \ g_i \neq g_j$$

  - G - CSG
  - K - Compact subset of $\mathbb{R}^n$



The 2D periodic structure with the *p2mg* plane group symmetry where K is a regular convex pentagon with the packing density of approximately 0.8541019. (Left) A single primitive cell. (Right) 9 primitive cells. The blue parallelogram denotes the primitive cell of the respective configuration. Colors represent symmetry operations modulo lattice translations.

CSG packing problem

- CSG packing problem

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

  - $\rho\left(\mathcal{K}_G\right) = \frac{N \text{area}(K)}{\det(\mathbf{U})}$ - 2D packing density.
    - $\mathbf{U}$ - Unit cell.
    - N - number of symmetry operation modulo lattice translations.

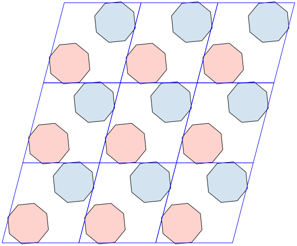
[1] Rogers, C. A. (1951). The closest packing of convex two-dimensional domains. Acta Mathematica, 86(1), 309-321.

CSG packing problem

- CSG packing problem

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

  - $\rho\left(\mathcal{K}_G\right) = \frac{N \text{area}(K)}{\det(\mathbf{U})}$ - 2D packing density.
    - $\mathbf{U}$ - Unit cell.
    - N - number of symmetry operation modulo lattice translations.



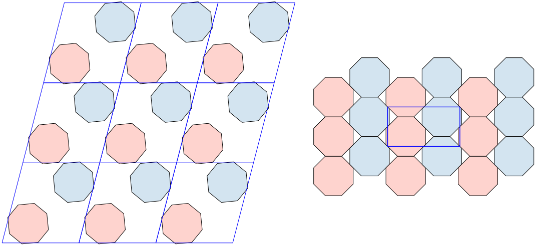
[1] Rogers, C. A. (1951). The closest packing of convex two-dimensional domains. Acta Mathematica, 86(1), 309-321.

CSG packing problem

- CSG packing problem

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

  - $\rho\left(\mathcal{K}_G\right) = \frac{N \text{area}(K)}{\det(\mathbf{U})}$ - 2D packing density.
    - **U** - Unit cell.
    - N - number of symmetry operation modulo lattice translations.



CSG packings where $\mathcal{G}$ of type $p2$ and K is a regular octagon. (Left) packing with density $\rho(\mathcal{K}_{p2}) \approxeq 0.413705$ and (right) optimal packing with density $\rho(\mathcal{K}_{p2}) = \frac{4+4\sqrt{2}}{5+4\sqrt{2}} \approxeq 0.90616$ [1]. The blue parallelogram denotes the primitive cell of the respective configuration. Colors represent symmetry operations modulo lattice translations.

[1] Rogers, C. A. (1951). The closest packing of convex two-dimensional domains. Acta Mathematica, 86(1), 309-321.

Entropic Trust Region Packing Algorithm

- Stochastic relaxation

$$\tilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} J(\boldsymbol{\theta}),$$

  - $J(\boldsymbol{\theta}) := E[\mathbf{F}|\boldsymbol{\theta}] = \int_{\mathbf{x} \in \mathcal{X}} \mathbf{F}(\mathbf{x}) dP(\boldsymbol{\theta})$
    - **F** - Fitness of the packing density
    - $dP(\boldsymbol{\theta})$ - Probability measure from a parametric family $S = \{dP(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^n\}$
    - $\mathcal{X}$ - Configuration space

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

Entropic Trust Region Packing Algorithm

- Stochastic relaxation

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \, J(\boldsymbol{\theta}),$$

  - $J(\boldsymbol{\theta}) := E[\mathbf{F}|\boldsymbol{\theta}] = \int_{\mathbf{x} \in \mathcal{X}} \mathbf{F}(\mathbf{x}) dP(\boldsymbol{\theta})$
    - $\mathbf{F}$ - Fitness of the packing density
    - $dP(\boldsymbol{\theta})$ - Probability measure from a parametric family $S = \{dP(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^n\}$
    - $\mathcal{X}$ - Configuration space
- Non-euclidean trust region method

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

Entropic Trust Region Packing Algorithm

- Stochastic relaxation
$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} J(\boldsymbol{\theta}),$$
  - $J(\boldsymbol{\theta}) := E[\mathbf{F}|\boldsymbol{\theta}] = \int_{\mathbf{x} \in \mathcal{X}} \mathbf{F}(\mathbf{x}) dP(\boldsymbol{\theta})$
    - $\mathbf{F}$ - Fitness of the packing density
    - $dP(\boldsymbol{\theta})$ - Probability measure from a parametric family $S = \{dP(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^n\}$
    - $\mathcal{X}$ - Configuration space
- Non-euclidean trust region method
  - Trust region defined by the Kullback-Leibler Divergence from $dP(\boldsymbol{\theta})$ to $dP(\boldsymbol{\theta} + \delta\boldsymbol{\theta})$: $D_{KL}(P_{\boldsymbol{\theta}} \,||\, P_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}) = \int_{\mathcal{X}} \ln\left(\frac{dP(\boldsymbol{\theta})}{dP(\boldsymbol{\theta}+\delta\boldsymbol{\theta})}\right) dP(\boldsymbol{\theta})$.

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

Entropic Trust Region Packing Algorithm

- Stochastic relaxation
$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}}\, J(\boldsymbol{\theta}),$$
  - $J(\boldsymbol{\theta}) := E[\mathbf{F}|\boldsymbol{\theta}] = \int_{\mathbf{x} \in \mathcal{X}} \mathbf{F}(\mathbf{x}) dP(\boldsymbol{\theta})$
    - $\mathbf{F}$ - Fitness of the packing density
    - $dP(\boldsymbol{\theta})$ - Probability measure from a parametric family $S = \{dP(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^n\}$
    - $\mathcal{X}$ - Configuration space
- Non-euclidean trust region method
  - Trust region defined by the Kullback-Leibler Divergence from $dP(\boldsymbol{\theta})$ to $dP(\boldsymbol{\theta} + \delta\boldsymbol{\theta})$: $D_{KL}\left(P_{\boldsymbol{\theta}} \,||\, P_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}\right) = \int_{\mathcal{X}} \ln\left(\frac{dP(\boldsymbol{\theta})}{dP(\boldsymbol{\theta}+\delta\boldsymbol{\theta})}\right) dP(\boldsymbol{\theta})$.
  - Update equations: $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \Delta^t \frac{\widetilde{\nabla} J(\boldsymbol{\theta}^t)}{\|\widetilde{\nabla} J(\boldsymbol{\theta}^t)\|_{\mathcal{I}_{\boldsymbol{\theta}^t}}}$

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

Entropic Trust Region Packing Algorithm

- Stochastic relaxation
$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}} J(\boldsymbol{\theta}),$$
  - $J(\boldsymbol{\theta}) := E[\mathbf{F}|\boldsymbol{\theta}] = \int_{\mathbf{x}\in\mathcal{X}} \mathbf{F}(\mathbf{x}) dP(\boldsymbol{\theta})$
    - $\mathbf{F}$ - Fitness of the packing density
    - $dP(\boldsymbol{\theta})$ - Probability measure from a parametric family $S = \{dP(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^n\}$
    - $\mathcal{X}$ - Configuration space
- Non-euclidean trust region method
  - Trust region defined by the Kullback-Leibler Divergence from $dP(\boldsymbol{\theta})$ to $dP(\boldsymbol{\theta}+\delta\boldsymbol{\theta})$: $D_{KL}\left(P_{\boldsymbol{\theta}} \mid\mid P_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}\right) = \int_{\mathcal{X}} \ln\left(\frac{dP(\boldsymbol{\theta})}{dP(\boldsymbol{\theta}+\delta\boldsymbol{\theta})}\right) dP(\boldsymbol{\theta})$.
  - Update equations: $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \Delta^t \frac{\widetilde{\nabla} J(\boldsymbol{\theta}^t)}{\|\widetilde{\nabla} J(\boldsymbol{\theta}^t)\|_{\mathcal{I}_{\boldsymbol{\theta}^t}}}$
    - $\widetilde{\nabla} J(\boldsymbol{\theta}) = \mathcal{I}_{\boldsymbol{\theta}}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ - Natural gradient[1] of the expected fitness $J(\boldsymbol{\theta})$

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

Entropic Trust Region Packing Algorithm

- Stochastic relaxation

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \, J(\boldsymbol{\theta}),$$

  - $J(\boldsymbol{\theta}) := E[\mathbf{F}|\boldsymbol{\theta}] = \int_{\mathbf{x} \in \mathcal{X}} \mathbf{F}(\mathbf{x}) dP(\boldsymbol{\theta})$
    - $\mathbf{F}$ - Fitness of the packing density
    - $dP(\boldsymbol{\theta})$ - Probability measure from a parametric family $S = \{dP(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^n\}$
    - $\mathcal{X}$ - Configuration space

- Non-euclidean trust region method
  - Trust region defined by the Kullback-Leibler Divergence from $dP(\boldsymbol{\theta})$ to $dP(\boldsymbol{\theta} + \delta\boldsymbol{\theta})$: $D_{KL}\left(P_{\boldsymbol{\theta}} \,||\, P_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}\right) = \int_{\mathcal{X}} \ln\left(\frac{dP(\boldsymbol{\theta})}{dP(\boldsymbol{\theta}+\delta\boldsymbol{\theta})}\right) dP(\boldsymbol{\theta})$.
  - Update equations: $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \Delta^t \frac{\widetilde{\nabla} J(\boldsymbol{\theta}^t)}{\|\widetilde{\nabla} J(\boldsymbol{\theta}^t)\|_{\mathcal{I}_{\boldsymbol{\theta}^t}}}$
    - $\widetilde{\nabla} J(\boldsymbol{\theta}) = \mathcal{I}_{\boldsymbol{\theta}}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ - Natural gradient[1] of the expected fitness $J(\boldsymbol{\theta})$
    - $\mathcal{I}_{\boldsymbol{\theta}}$ - Fisher metric tensor: $\mathcal{I}_{\boldsymbol{\theta}\,ij} = \int_{\mathcal{X}} \frac{\partial \ln(p(\boldsymbol{\theta}))}{\partial \theta_i} \frac{\partial \ln(p(\boldsymbol{\theta}))}{\partial \theta_j} dP(\boldsymbol{\theta})$;

      $p(\boldsymbol{\theta}) = \frac{dP(\boldsymbol{\theta})}{d\boldsymbol{\nu}}$ is the Radon-Nikodym derivative of $P(\boldsymbol{\theta})$ with respect to some reference measure $\boldsymbol{\nu}$ defined on $\mathcal{X}$.

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

Entropic Trust Region Packing Algorithm

- Stochastic relaxation

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \, J(\boldsymbol{\theta}),$$

  - $J(\boldsymbol{\theta}) := E[\mathbf{F}|\boldsymbol{\theta}] = \int_{\mathbf{x} \in \mathcal{X}} \mathbf{F}(\mathbf{x}) dP(\boldsymbol{\theta})$
    - $\mathbf{F}$ - Fitness of the packing density
    - $dP(\boldsymbol{\theta})$ - Probability measure from a parametric family $S = \{dP(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^n\}$
    - $\mathcal{X}$ - Configuration space
- Non-euclidean trust region method
  - Trust region defined by the Kullback-Leibler Divergence from $dP(\boldsymbol{\theta})$ to $dP(\boldsymbol{\theta} + \delta\boldsymbol{\theta})$: $D_{KL}\left(P_{\boldsymbol{\theta}} \,||\, P_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}\right) = \int_{\mathcal{X}} \ln\left(\frac{dP(\boldsymbol{\theta})}{dP(\boldsymbol{\theta}+\delta\boldsymbol{\theta})}\right) dP(\boldsymbol{\theta})$.
  - Update equations: $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \Delta^t \frac{\widetilde{\nabla} J(\boldsymbol{\theta}^t)}{\|\widetilde{\nabla} J(\boldsymbol{\theta}^t)\|_{\mathcal{I}_{\boldsymbol{\theta}^t}}}$
    - $\widetilde{\nabla} J(\boldsymbol{\theta}) = \mathcal{I}_{\boldsymbol{\theta}}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ - Natural gradient[1] of the expected fitness $J(\boldsymbol{\theta})$
    - $\mathcal{I}_{\boldsymbol{\theta}}$ - Fisher metric tensor: $\mathcal{I}_{\boldsymbol{\theta}\,ij} = \int_{\mathcal{X}} \frac{\partial \ln(p(\boldsymbol{\theta}))}{\partial \theta_i} \frac{\partial \ln(p(\boldsymbol{\theta}))}{\partial \theta_j} dP(\boldsymbol{\theta})$;

      $p(\boldsymbol{\theta}) = \frac{dP(\boldsymbol{\theta})}{d\boldsymbol{\nu}}$ is the Radon-Nikodym derivative of $P(\boldsymbol{\theta})$ with respect to some reference measure $\boldsymbol{\nu}$ defined on $\mathcal{X}$.
    - Δ^t - Step size

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

Extended Multivariate von Mises (EMvM) probability distribution

- The lattice subgroup L of a Crystallographic Symmetry Group induces quotient space $\mathbb{R}^n/\mathrm{L} \approx T^n$

[1] Mardia K. V. et. al. (2008). A multivariate von mises distribution with applications to bioinformatics. Canadian Journal of Statistics, 36, pp. 99–109.

Extended Multivariate von Mises (EMvM) probability distribution

- The lattice subgroup L of a Crystallographic Symmetry Group induces quotient space $\mathbb{R}^n/\mathrm{L} \approx T^n$
- Extended Multivariate von Mises[1] (EMvM) distribution:

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

$$c(\boldsymbol{\theta}-\boldsymbol{\mu}) = [cos(\theta_1-\mu_1),\ldots,cos(\theta_n-\mu_n)]^{\mathsf{T}} \quad 0 \le \theta_i \le 2\pi, \quad 0 \le \mu_i \le 2\pi, \quad \mathbf{D} = \begin{bmatrix} \mathbf{D}^{cc} & \mathbf{D}^{cs} \\ \mathbf{D}^{cs\,\mathsf{T}} & \mathbf{D}^{ss}\end{bmatrix}$$

$$s(\boldsymbol{\theta}-\boldsymbol{\mu}) = [sin(\theta_1-\mu_1),\ldots,sin(\theta_n-\mu_n)]^{\mathsf{T}} \quad 0 \le \kappa_i$$

  - $\mathbf{D}$ - $2n \times 2n$ real-valued symmetric matrix with diagonal elements of $\mathbf{D}^{cc}$, $\mathbf{D}^{ss}$ and $\mathbf{D}^{cs}$ set to zero.
    - Controls cosine-cosine, sine-sine and cosine-sine interactions.

[1] Mardia K. V. et. al. (2008). A multivariate von mises distribution with applications to bioinformatics. Canadian Journal of Statistics, 36, pp. 99–109.

Extended Multivariate von Mises (EMvM) probability distribution

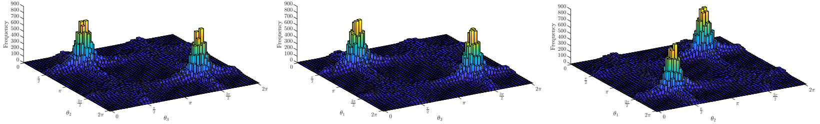
- The lattice subgroup L of a Crystallographic Symmetry Group induces quotient space $\mathbb{R}^n/\mathrm{L} \approx T^n$
- Extended Multivariate von Mises[1] (EMvM) distribution:

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

$$c(\boldsymbol{\theta}-\boldsymbol{\mu}) = [cos(\theta_1-\mu_1),\ldots,cos(\theta_n-\mu_n)]^{\mathsf{T}} \quad 0 \le \theta_i \le 2\pi, \quad 0 \le \mu_i \le 2\pi, \quad \mathbf{D} = \begin{bmatrix} \mathbf{D}^{cc} & \mathbf{D}^{cs} \\ \mathbf{D}^{cs\ \mathsf{T}} & \mathbf{D}^{ss}\end{bmatrix}$$

$$s(\boldsymbol{\theta}-\boldsymbol{\mu}) = [sin(\theta_1-\mu_1),\ldots,sin(\theta_n-\mu_n)]^{\mathsf{T}} \quad 0 \le \kappa_i$$

  - $\mathbf{D}$ - $2n \times 2n$ real-valued symmetric matrix with diagonal elements of $\mathbf{D}^{cc}$, $\mathbf{D}^{ss}$ and $\mathbf{D}^{cs}$ set to zero.
    - Controls cosine-cosine. sine-sine and cosine-sine interactions.



Histograms of projections along (**left**) θ_1, (**middle**) θ_2 and (**right**) θ_3 coordinates of 100000 samples from the trivariate EMvM.

[1] Mardia K. V. et. al. (2008). A multivariate von mises distribution with applications to bioinformatics. Canadian Journal of Statistics, 36, pp. 99–109.

Exponential reformulation of the EMvM

- Exponential family Extended Multivariate von Mises distribution

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

$$\begin{matrix} c(\boldsymbol{\theta}) = \left[cos(\theta_1), \ldots, cos(\theta_n)\right]^{\mathsf{T}} \\ s(\boldsymbol{\theta}) = \left[sin(\theta_1), \ldots, sin(\theta_n)\right]^{\mathsf{T}} \end{matrix} \quad 0 \le \theta_i \le 2\pi, \; \boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\kappa} \odot c(\boldsymbol{\mu}) \\ \boldsymbol{\kappa} \odot s(\boldsymbol{\mu}) \end{bmatrix}, \; \mathbf{E} = \begin{bmatrix} \mathbf{E}^{cc} & \mathbf{E}^{cs} \\ \mathbf{E}^{sc} & \mathbf{E}^{ss} \end{bmatrix}, \quad \begin{matrix} 0 \le \mu_i \le 2\pi \\ 0 < \kappa_i \end{matrix}$$

  - **E** - $2n \times 2n$ real-valued symmetric matrix with diagonal elements of $\mathbf{E}^{cc}$, $\mathbf{E}^{ss}$ and $\mathbf{E}^{cs}$ set to zero.

Exponential reformulation of the EMvM

- Exponential family Extended Multivariate von Mises distribution

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

$$\begin{matrix} c(\boldsymbol{\theta}) = \left[cos(\theta_1), \ldots, cos(\theta_n)\right]^{\mathsf{T}} \\ s(\boldsymbol{\theta}) = \left[sin(\theta_1), \ldots, sin(\theta_n)\right]^{\mathsf{T}} \end{matrix} \quad 0 \le \theta_i \le 2\pi, \ \boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\kappa} \odot c(\boldsymbol{\mu}) \\ \boldsymbol{\kappa} \odot s(\boldsymbol{\mu}) \end{bmatrix}, \ \mathbf{E} = \begin{bmatrix} \mathbf{E}^{cc} & \mathbf{E}^{cs} \\ \mathbf{E}^{sc} & \mathbf{E}^{ss} \end{bmatrix}, \quad \begin{matrix} 0 \le \mu_i \le 2\pi \\ 0 < \kappa_i \end{matrix}$$

  - $\mathbf{E}$ - $2n \times 2n$ real-valued symmetric matrix with diagonal elements of $\mathbf{E}^{cc}$, $\mathbf{E}^{ss}$ and $\mathbf{E}^{cs}$ set to zero.
  - $\dim(S_{\mathrm{EMvM}}) = 2n^2 > \frac{n(n+3)}{2} = \dim\left(S_{\mathrm{MvM}}\right)$

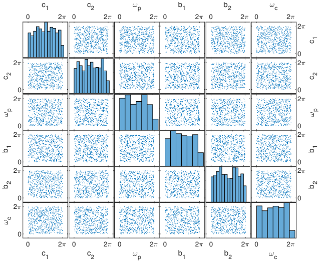
Exponential reformulation of the EMvM

- Exponential family Extended Multivariate von Mises distribution

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

$$c(\boldsymbol{\theta}) = [cos(\theta_1),\ldots,cos(\theta_n)]^{\mathbf{T}} \quad s(\boldsymbol{\theta}) = [sin(\theta_1),\ldots,sin(\theta_n)]^{\mathbf{T}} \quad 0\le\theta_i\le 2\pi,\ \boldsymbol{\eta} = \begin{bmatrix}\boldsymbol{\kappa}\odot c(\boldsymbol{\mu})\\ \boldsymbol{\kappa}\odot s(\boldsymbol{\mu})\end{bmatrix},\ \mathbf{E} = \begin{bmatrix}\mathbf{E}^{cc} & \mathbf{E}^{cs}\\ \mathbf{E}^{sc} & \mathbf{E}^{ss}\end{bmatrix},\quad 0\le\mu_i\le 2\pi,\ 0<\kappa_i$$

  - $\mathbf{E}$ - $2n\times 2n$ real-valued symmetric matrix with diagonal elements of $\mathbf{E}^{cc}$, $\mathbf{E}^{ss}$ and $\mathbf{E}^{cs}$ set to zero.
  - $\dim(S_{\mathrm{EMvM}}) = 2n^2 > \frac{n(n+3)}{2} = \dim(S_{\mathrm{MvM}})$



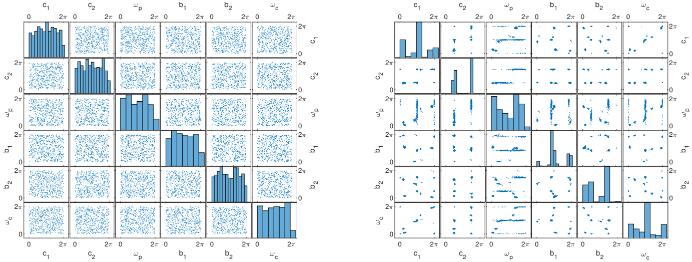
Exponential reformulation of the EMvM

- Exponential family Extended Multivariate von Mises distribution

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

$$\begin{aligned} c(\boldsymbol{\theta}) &= [cos(\theta_1), \ldots, cos(\theta_n)]^{\mathbf{T}} \\ s(\boldsymbol{\theta}) &= [sin(\theta_1), \ldots, sin(\theta_n)]^{\mathbf{T}} \end{aligned} \quad 0 \le \theta_i \le 2\pi, \ \boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\kappa} \odot c(\boldsymbol{\mu}) \\ \boldsymbol{\kappa} \odot s(\boldsymbol{\mu}) \end{bmatrix}, \ \mathbf{E} = \begin{bmatrix} \mathbf{E}^{cc} & \mathbf{E}^{cs} \\ \mathbf{E}^{sc} & \mathbf{E}^{ss} \end{bmatrix}, \quad \begin{matrix} 0 \le \mu_i \le 2\pi \\ 0 < \kappa_i \end{matrix}$$

  - $\mathbf{E}$ - $2n \times 2n$ real-valued symmetric matrix with diagonal elements of $\mathbf{E}^{cc}$, $\mathbf{E}^{ss}$ and $\mathbf{E}^{cs}$ set to zero.
  - $\dim(S_{\text{EMvM}}) = 2n^2 > \frac{n(n+3)}{2} = \dim(S_{\text{MvM}})$



600 realizations of the Extended Multivariate von Mises distribution defined on an T^6 from a single run of the ETRPA. (**Left**) Initial distribution and (**Right**) output distribution.

Truncation selection transformation of the fitness function $\mathbf{F}$

- $\tilde{\mathbf{F}} := q \, \mathbf{1}_{\mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q}}}(\mathbf{x}) = \begin{cases} q & \text{if } \mathbf{F}(\mathbf{x}) \geq \mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q}} \\ 0 & \text{otherwise} \end{cases}$
    - $\mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q}}$ is the $P_{\tilde{\boldsymbol{\theta}}}$'s $(q-1)$th q-quantile of the fitness $\mathbf{F}$.

Truncation selection transformation of the fitness function $\mathbf{F}$

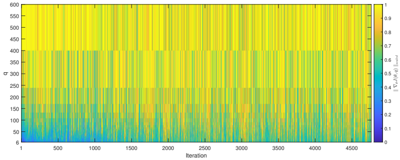
- $\tilde{\mathbf{F}} := q\, \mathbf{1}_{\mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q}}}(\mathbf{x}) = \begin{cases} q & \text{if } \mathbf{F}(\mathbf{x}) \geq \mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q}} \\ 0 & \text{otherwise} \end{cases}$
    - $\mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q}}$ is the $P_{\tilde{\boldsymbol{\theta}}}$'s $(q-1)$th q-quantile of the fitness $\mathbf{F}$.
- $S^e := \{dP^e(\boldsymbol{\theta}) = \exp\{\boldsymbol{\theta}^\intercal \mathbf{t}(\mathbf{x}) - \psi(\boldsymbol{\theta})\} dP_0(\mathbf{x})\}$,
    - Reference measure $dP_0(\mathbf{x})$.
    - The log-partition function $\psi(\boldsymbol{\theta}) = \ln \int \exp\{\boldsymbol{\theta}^\intercal \mathbf{t}(\mathbf{x})\} dP_0(\mathbf{x})$.

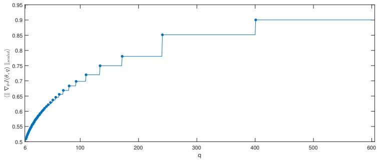
Truncation selection transformation of the fitness function $\mathbf{F}$

- $\tilde{\mathbf{F}} := q\ \mathbf{1}_{\mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q}}}(\mathbf{x}) = \begin{cases} q & \text{if } \mathbf{F}(\mathbf{x}) \geq \mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q}} \\ 0 & \text{otherwise} \end{cases}$
  - $\mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q}}$ is the $P_{\tilde{\boldsymbol{\theta}}}$'s $(q-1)$th q-quantile of the fitness $\mathbf{F}$.
- $S^e := \{dP^e(\boldsymbol{\theta}) = \exp\{\boldsymbol{\theta}^\intercal \mathbf{t}(\mathbf{x}) - \psi(\boldsymbol{\theta})\} dP_0\left(\mathbf{x}\right)\}$,
  - Reference measure $dP_0\left(\mathbf{x}\right)$.
  - The log-partition function $\psi\left(\boldsymbol{\theta}\right) = \ln \int \exp\{\boldsymbol{\theta}^\intercal \mathbf{t}\left(\mathbf{x}\right)\} dP_0\left(\mathbf{x}\right)$.
- $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \mid_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = \boldsymbol{\mu}_{\mathbf{F}_{1-\frac{1}{q^*}}} - \boldsymbol{\mu}$,
  - $\boldsymbol{\mu}_{\mathbf{F}_{1-\frac{1}{q_t}}} = \int\limits_{\mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q_t}}}^{\infty} \mathbf{F}^{-1}(y) \exp\left\{\tilde{\boldsymbol{\theta}}^\intercal \mathbf{t}\left(\mathbf{F}^{-1}(y)\right) - \psi(\tilde{\boldsymbol{\theta}}) + \ln\left(q_t\right)\right\} dP_0 \circ \mathbf{F}^{-1}(y)$.
  - $\boldsymbol{\mu} = \int \mathbf{x} \exp\{\tilde{\boldsymbol{\theta}}^\intercal \mathbf{t}\left(\mathbf{x}\right) - \psi(\tilde{\boldsymbol{\theta}})\} dP_0\left(\mathbf{x}\right)$.

Truncation selection transformation of the fitness function $\mathbf{F}$

- $\tilde{\mathbf{F}} := q\, \mathbf{1}_{\mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q}}}(\mathbf{x}) = \begin{cases} q & \text{if } \mathbf{F}(\mathbf{x}) \geq \mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q}} \\ 0 & \text{otherwise} \end{cases}$
  - $\mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q}}$ is the $P_{\tilde{\boldsymbol{\theta}}}$'s $(q-1)$th q-quantile of the fitness $\mathbf{F}$.
- $S^e := \{dP^e(\boldsymbol{\theta}) = \exp\{\boldsymbol{\theta}^\intercal \mathbf{t}(\mathbf{x}) - \psi(\boldsymbol{\theta})\} dP_0(\mathbf{x})\}$,
  - Reference measure $dP_0(\mathbf{x})$.
  - The log-partition function $\psi(\boldsymbol{\theta}) = \ln \int \exp\{\boldsymbol{\theta}^\intercal \mathbf{t}(\mathbf{x})\} dP_0(\mathbf{x})$.
- $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \mid_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = \boldsymbol{\mu}_{\mathbf{F}_{1-\frac{1}{q^*}}} - \boldsymbol{\mu}$,
  - $\boldsymbol{\mu}_{\mathbf{F}_{1-\frac{1}{q_t}}} = \int_{\mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q_t}}}^{\infty} \mathbf{F}^{-1}(y) \exp\left\{\tilde{\boldsymbol{\theta}}^\intercal \mathbf{t}\left(\mathbf{F}^{-1}(y)\right) - \psi(\tilde{\boldsymbol{\theta}}) + \ln(q_t)\right\} dP_0 \circ \mathbf{F}^{-1}(y)$.
  - $\boldsymbol{\mu} = \int \mathbf{x} \exp\{\tilde{\boldsymbol{\theta}}^\intercal \mathbf{t}(\mathbf{x}) - \psi(\tilde{\boldsymbol{\theta}})\} dP_0(\mathbf{x})$.





Influence of q on (Left) the scaled expected fitness gradient $\| \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t, q) \|_{\text{scaled}}$ and (Right) scaled expected fitness gradient averaged through all iterations $(<\| \nabla_\theta J(\theta, q) \|_{scaled}>)$.

Adaptive selection quantile

- q as an equivalent to the annealing temperature control parameter

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

$$\cos(\alpha^t) = \frac{< \Delta\boldsymbol{\theta}^t, \Delta\boldsymbol{\theta}^{t-1} >_{\mathcal{I}_{\boldsymbol{\theta}^{t-1}}}}{||\Delta\boldsymbol{\theta}^t||_{\mathcal{I}_{\boldsymbol{\theta}^{t-1}}} ||\Delta\boldsymbol{\theta}^{t-1}||_{\mathcal{I}_{\boldsymbol{\theta}^{t-1}}}}, \qquad \begin{array}{c} \Delta\boldsymbol{\theta}^t = \boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}, \\ \Delta\boldsymbol{\theta}^{t-1} = \boldsymbol{\theta}^{t-1} - \boldsymbol{\theta}^{t-2}, \end{array}$$

  - $< \cdot, \cdot >_{\mathcal{I}_{\boldsymbol{\theta}^{t-1}}}$ - scalar product with respect to $\mathcal{I}_{\boldsymbol{\theta}^{t-1}}$.
  - $\Delta\boldsymbol{\theta}^t, \Delta\boldsymbol{\theta}^{t-1} \in T_{\boldsymbol{\theta}^{t-1}} S$.

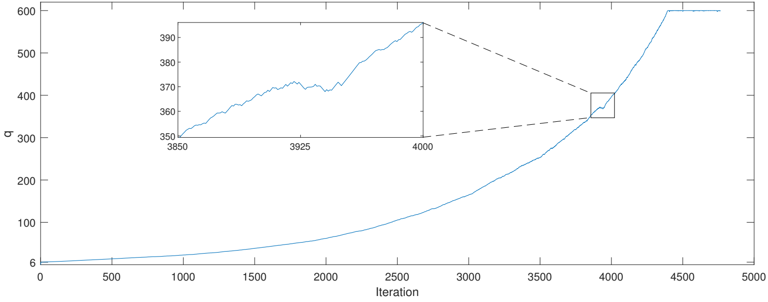
Adaptive selection quantile

- q as an equivalent to the annealing temperature control parameter

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

$$\cos(\alpha^t) = \frac{< \Delta\boldsymbol{\theta}^t, \Delta\boldsymbol{\theta}^{t-1} >_{\mathcal{I}_{\boldsymbol{\theta}^{t-1}}}}{||\Delta\boldsymbol{\theta}^t||_{\mathcal{I}_{\boldsymbol{\theta}^{t-1}}} ||\Delta\boldsymbol{\theta}^{t-1}||_{\mathcal{I}_{\boldsymbol{\theta}^{t-1}}}}, \qquad \begin{aligned} \Delta\boldsymbol{\theta}^t &= \boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}, \\ \Delta\boldsymbol{\theta}^{t-1} &= \boldsymbol{\theta}^{t-1} - \boldsymbol{\theta}^{t-2}, \end{aligned}$$

  - $< \cdot, \cdot >_{\mathcal{I}_{\boldsymbol{\theta}^{t-1}}}$ - scalar product with respect to $\mathcal{I}_{\boldsymbol{\theta}^{t-1}}$.
  - $\Delta\boldsymbol{\theta}^t, \Delta\boldsymbol{\theta}^{t-1} \in T_{\boldsymbol{\theta}^{t-1}} S$.



Evolution of the adaptive selection quantile q.

Entropic proximal maximization for exponential families

- Euclidean fitness gradient $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \,|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = \boldsymbol{\mu}_{\mathsf{F}_{1-\frac{1}{q^*}}} - \boldsymbol{\mu}$, for a fixed q^*, determines proximal map maximization[1]

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

[1] Beck A. and Teboulle M. (2003) Mirror descent and nonlinear projected subgradient methods for convex optimization, Operations Research Letters, 31, pp. 167–175.

[2] Raskutti G. and Mukherjee S. (2015). The information geometry of mirror descent, IEEE Transactions on Information Theory, 61, pp. 1451–1457.

Entropic proximal maximization for exponential families

- Euclidean fitness gradient $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \mid_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = \boldsymbol{\mu}_{\mathsf{F}_{1-\frac{1}{q^*}}} - \boldsymbol{\mu}$, for a fixed q^*, determines proximal map maximization[1]

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

- Natural gradient updates in dual coordinate systems via the Legendre transforms $\boldsymbol{\mu} = \nabla_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta})$ and $\boldsymbol{\theta} = \nabla_{\boldsymbol{\mu}} \phi(\boldsymbol{\mu})$ [2]:

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

[1] Beck A. and Teboulle M. (2003) Mirror descent and nonlinear projected subgradient methods for convex optimization, Operations Research Letters, 31, pp. 167–175.

[2] Raskutti G. and Mukherjee S. (2015). The information geometry of mirror descent, IEEE Transactions on Information Theory, 61, pp. 1451–1457.

Entropic proximal maximization for exponential families

- Euclidean fitness gradient $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \mid_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = \boldsymbol{\mu}_{\mathsf{F}_{1-\frac{1}{q^*}}} - \boldsymbol{\mu}$, for a fixed q^*, determines proximal map maximization[1]

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

- Natural gradient updates in dual coordinate systems via the Legendre transforms $\boldsymbol{\mu} = \nabla_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta})$ and $\boldsymbol{\theta} = \nabla_{\boldsymbol{\mu}} \phi(\boldsymbol{\mu})$ [2]:

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

- ETRPA: $\epsilon^t = \frac{\Delta^t}{\sqrt{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t)^\intercal (\mathcal{I}_{\boldsymbol{\theta}^t})^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t)}}$,

[1] Beck A. and Teboulle M. (2003) Mirror descent and nonlinear projected subgradient methods for convex optimization, Operations Research Letters, 31, pp. 167–175.

[2] Raskutti G. and Mukherjee S. (2015). The information geometry of mirror descent, IEEE Transactions on Information Theory, 61, pp. 1451–1457.

Information-geometry of ETRPA

- ETRPA as a solution of a maximin problem

$$\max_{P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}} \min_{P_{\boldsymbol{\theta}^1} \in S^1} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \mid\mid P_{\boldsymbol{\theta}^1}); S^{q^*}, S^1 \subset S.$$

Information-geometry of ETRPA

- ETRPA as a solution of a maximin problem

$$\max_{P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}} \min_{P_{\boldsymbol{\theta}^1} \in S^1} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \mid\mid P_{\boldsymbol{\theta}^1}); S^{q^*}, S^1 \subset S.$$

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

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

Information-geometry of ETRPA

- ETRPA as a solution of a maximin problem

$$\max_{P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}} \min_{P_{\boldsymbol{\theta}^1} \in S^1} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \mid\mid P_{\boldsymbol{\theta}^1}); S^{q^*}, S^1 \subset S.$$

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

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

- Fixing $P_{\boldsymbol{\theta}^{q^*}}$ minimizing with respect to $\boldsymbol{\theta}^1$, and fixing $P_{\boldsymbol{\theta}^1}$ maximizing with respect to $\boldsymbol{\theta}^q$

$$\boldsymbol{\theta}^{1^{t+1}} = \boldsymbol{\theta}^{1^t} + \epsilon \mathcal{I}^{-1}_{\boldsymbol{\theta}^{1^t}} (\boldsymbol{\mu}^{{q^*}^t} - \boldsymbol{\mu}^{1^t}), \quad \boldsymbol{\mu}^{{q^*}^{t+1}} = \boldsymbol{\mu}^{{q^*}^t} + \epsilon \mathcal{I}^{-1}_{\boldsymbol{\mu}^{{q^*}^t}} (\boldsymbol{\theta}^{{q^*}^t} - \boldsymbol{\theta}^{1^t}).$$

Information-geometry of ETRPA

- ETRPA as a solution of a maximin problem

$$\max_{P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}} \min_{P_{\boldsymbol{\theta}^1} \in S^1} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \mid\mid P_{\boldsymbol{\theta}^1}); S^{q^*}, S^1 \subset S.$$

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

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

- Fixing $P_{\boldsymbol{\theta}^{q^*}}$ minimizing with respect to $\boldsymbol{\theta}^1$, and fixing $P_{\boldsymbol{\theta}^1}$ maximizing with respect to $\boldsymbol{\theta}^q$

$$\boldsymbol{\theta}^{1^{t+1}} = \boldsymbol{\theta}^{1^t} + \epsilon \mathcal{I}^{-1}_{\boldsymbol{\theta}^{1^t}} (\boldsymbol{\mu}^{q^{*^t}} - \boldsymbol{\mu}^{1^t}), \quad \boldsymbol{\mu}^{q^{*^{t+1}}} = \boldsymbol{\mu}^{q^{*^t}} + \epsilon \mathcal{I}^{-1}_{\boldsymbol{\mu}^{q^{*t}}} (\boldsymbol{\theta}^{q^{*^t}} - \boldsymbol{\theta}^{1^t}).$$

- Considering S^{q^*} and S^1 as embeddings of S^e in S we recover S^e from S

Information-geometry of ETRPA

- ETRPA as a solution of a maximin problem

$$\max_{P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}} \min_{P_{\boldsymbol{\theta}^1} \in S^1} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \mid\mid P_{\boldsymbol{\theta}^1}); S^{q^*}, S^1 \subset S.$$

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

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

- Fixing $P_{\boldsymbol{\theta}^{q^*}}$ minimizing with respect to $\boldsymbol{\theta}^1$, and fixing $P_{\boldsymbol{\theta}^1}$ maximizing with respect to $\boldsymbol{\theta}^q$

$$\boldsymbol{\theta}^{1^{t+1}} = \boldsymbol{\theta}^{1^t} + \epsilon\mathcal{I}^{-1}_{\boldsymbol{\theta}^{1^t}}(\boldsymbol{\mu}^{q^{*\,t}} - \boldsymbol{\mu}^{1^t}), \quad \boldsymbol{\mu}^{q^{*\,t+1}} = \boldsymbol{\mu}^{q^{*\,t}} + \epsilon\mathcal{I}^{-1}_{\boldsymbol{\mu}^{q^*t}}(\boldsymbol{\theta}^{q^{*\,t}} - \boldsymbol{\theta}^{1^t}).$$

- Considering S^{q^*} and S^1 as embeddings of S^e in S we recover S^e from S
  - For $\boldsymbol{\theta}^1 \in S^1$ the $\boldsymbol{\theta}$-coordinates of S^e: $[\theta^1_i]_{i=1,\ldots,n} \to \boldsymbol{\theta}$.

Information-geometry of ETRPA

- ETRPA as a solution of a maximin problem

$$\max_{P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}} \min_{P_{\boldsymbol{\theta}^1} \in S^1} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \mid\mid P_{\boldsymbol{\theta}^1}); S^{q^*}, S^1 \subset S.$$

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

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

- Fixing $P_{\boldsymbol{\theta}^{q^*}}$ minimizing with respect to $\boldsymbol{\theta}^1$, and fixing $P_{\boldsymbol{\theta}^1}$ maximizing with respect to $\boldsymbol{\theta}^q$

$$\boldsymbol{\theta}^{1^{t+1}} = \boldsymbol{\theta}^{1^t} + \epsilon \mathcal{I}^{-1}_{\boldsymbol{\theta}^{1^t}}(\boldsymbol{\mu}^{q^{*t}} - \boldsymbol{\mu}^{1^t}), \quad \boldsymbol{\mu}^{q^{*t+1}} = \boldsymbol{\mu}^{q^{*t}} + \epsilon \mathcal{I}^{-1}_{\boldsymbol{\mu}^{q^*t}}(\boldsymbol{\theta}^{q^{*t}} - \boldsymbol{\theta}^{1^t}).$$

- Considering S^{q^*} and S^1 as embeddings of S^e in S we recover S^e from S
  - For $\boldsymbol{\theta}^1 \in S^1$ the $\boldsymbol{\theta}$-coordinates of S^e: $[\theta^1_i]_{i=1,\ldots,n} \to \boldsymbol{\theta}$.
  - For $\boldsymbol{\mu}^{q^*} \in S^{q^*}$ the $\boldsymbol{\mu}$-coordinates of S^e:

$$\boldsymbol{\mu} = \frac{1}{q^*}[\mu^{q^*}_i]_{i=1,\ldots,n} + \frac{q^*-1}{q^*}\boldsymbol{\mu}_{\mathbf{F}^C_{1-\frac{1}{q^*}}},$$

$$\boldsymbol{\mu}_{\mathbf{F}^C_{1-\frac{1}{q^*}}} = \int_{-\infty}^{\mathbf{F}_{1-\frac{1}{q^*}}} \mathbf{F}^{-1}(y) \exp\left\{\boldsymbol{\theta}^\intercal \mathbf{t}\left(\mathbf{F}^{-1}(y)\right) - \psi(\boldsymbol{\theta}) + \ln\left(\frac{q^*}{q^*-1}\right)\right\} dP_0 \circ \mathbf{F}^{-1}(y)$$

Maximization of stochastic dependence

- $\max_{P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}} \min_{P_{\boldsymbol{\theta}^1} \in S^1} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \mid\mid P_{\boldsymbol{\theta}^1})$

[1] Studený M. and Vejnarová J. (1998). The multiinformation function as a tool for measuring stochastic dependence, in Learning in Graphical Models. pp. 261–297.

Maximization of stochastic dependence

- $\max_{P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}} \min_{P_{\boldsymbol{\theta}^1} \in S^1} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \mid\mid P_{\boldsymbol{\theta}^1})$
- For a fixed $P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}$

$$\hat{\boldsymbol{\theta}}^1 = \operatorname*{argmin}_{P_{\boldsymbol{\theta}^1} \in S^1} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \mid\mid P_{\boldsymbol{\theta}^1})$$

  - $P_{\hat{\boldsymbol{\theta}}^1} \in S^1$ - maximum entropy estimate of $P_{\boldsymbol{\theta}^{q^*}}$

[1] Studený M. and Vejnarová J. (1998). The multiinformation function as a tool for measuring stochastic dependence, in Learning in Graphical Models. pp. 261–297.

Maximization of stochastic dependence

- $\max_{P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}} \min_{P_{\boldsymbol{\theta}^1} \in S^1} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \parallel P_{\boldsymbol{\theta}^1})$
- For a fixed $P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}$

$$\hat{\boldsymbol{\theta}}^1 = \underset{P_{\boldsymbol{\theta}^1} \in S^1}{\operatorname{argmin}} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \parallel P_{\boldsymbol{\theta}^1})$$

  - $P_{\hat{\boldsymbol{\theta}}^1} \in S^1$ - maximum entropy estimate of $P_{\boldsymbol{\theta}^{q^*}}$

- $D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \parallel P_{\hat{\boldsymbol{\theta}}^1}) = D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \parallel P_{\hat{\boldsymbol{\theta}}^1_s}) - D_{KL}(P_{\hat{\boldsymbol{\theta}}^1} \parallel P_{\hat{\boldsymbol{\theta}}^1_s})$
  - $P_{\hat{\boldsymbol{\theta}}^1_s} \in S^1$ - split model corresponding to $P_{\hat{\boldsymbol{\theta}}^1}$

[1] Studený M. and Vejnarová J. (1998). The multiinformation function as a tool for measuring stochastic dependence, in Learning in Graphical Models. pp. 261–297.

Maximization of stochastic dependence

- $\max_{P_{\boldsymbol{\theta}^{q*}} \in S^{q*}} \min_{P_{\boldsymbol{\theta}^1} \in S^1} D_{KL}(P_{\boldsymbol{\theta}^{q*}} \mid\mid P_{\boldsymbol{\theta}^1})$
- For a fixed $P_{\boldsymbol{\theta}^{q*}} \in S^{q^*}$

$$\hat{\boldsymbol{\theta}}^1 = \underset{P_{\boldsymbol{\theta}^1} \in S^1}{\operatorname{argmin}} D_{KL}(P_{\boldsymbol{\theta}^{q*}} \mid\mid P_{\boldsymbol{\theta}^1})$$

  - $P_{\hat{\boldsymbol{\theta}}^1} \in S^1$ - maximum entropy estimate of $P_{\boldsymbol{\theta}^{q*}}$

- $D_{KL}(P_{\boldsymbol{\theta}^{q*}} \mid\mid P_{\hat{\boldsymbol{\theta}}^1}) = D_{KL}(P_{\boldsymbol{\theta}^{q*}} \mid\mid P_{\hat{\boldsymbol{\theta}}^1_s}) - D_{KL}(P_{\hat{\boldsymbol{\theta}}^1} \mid\mid P_{\hat{\boldsymbol{\theta}}^1_s})$
  - $P_{\hat{\boldsymbol{\theta}}^1_s} \in S^1$ - split model corresponding to $P_{\hat{\boldsymbol{\theta}}^1}$
- For a fixed $P_{\hat{\boldsymbol{\theta}}^1}$:

$$\max_{P_{\boldsymbol{\theta}^{q*}} \in S^{q^*}} D_{KL}(P_{\boldsymbol{\theta}^{q*}} \mid\mid P_{\hat{\boldsymbol{\theta}}^1}) \Leftrightarrow \max_{P_{\boldsymbol{\theta}^{q*}} \in S^{q^*}} D_{KL}(P_{\boldsymbol{\theta}^{q*}} \mid\mid P_{\hat{\boldsymbol{\theta}}^1_s})$$

[1] Studený M. and Vejnarová J. (1998). The multiinformation function as a tool for measuring stochastic dependence, in Learning in Graphical Models. pp. 261–297.

Maximization of stochastic dependence

- $\max_{P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}} \min_{P_{\boldsymbol{\theta}^1} \in S^1} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \parallel P_{\boldsymbol{\theta}^1})$
- For a fixed $P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}$

$$\hat{\boldsymbol{\theta}}^1 = \underset{P_{\boldsymbol{\theta}^1} \in S^1}{\operatorname{argmin}} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \parallel P_{\boldsymbol{\theta}^1})$$

  - $P_{\hat{\boldsymbol{\theta}}^1} \in S^1$ - maximum entropy estimate of $P_{\boldsymbol{\theta}^{q^*}}$

- $D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \parallel P_{\hat{\boldsymbol{\theta}}^1}) = D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \parallel P_{\hat{\boldsymbol{\theta}}^1_s}) - D_{KL}(P_{\hat{\boldsymbol{\theta}}^1} \parallel P_{\hat{\boldsymbol{\theta}}^1_s})$
  - $P_{\hat{\boldsymbol{\theta}}^1_s} \in S^1$ - split model corresponding to $P_{\hat{\boldsymbol{\theta}}^1}$
- For a fixed $P_{\hat{\boldsymbol{\theta}}^1}$:

$$\max_{P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \parallel P_{\hat{\boldsymbol{\theta}}^1}) \Leftrightarrow \max_{P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \parallel P_{\hat{\boldsymbol{\theta}}^1_s})$$

- Special case of multi-information/total correlation[1]:

$$D_{KL}(\cdot \parallel P_{\hat{\boldsymbol{\theta}}^1_s})$$

  - A measure of stochastic dependence in complex systems.

[1] Studený M. and Vejnarová J. (1998). The multiinformation function as a tool for measuring stochastic dependence, in Learning in Graphical Models. pp. 261–297.

Maximization of stochastic dependence

- Multi-information as generalization[1] of the Infomax principle[2]:

$$\max_{\{f \in \mathcal{F} \mid O = f(I)\}} MI(I; O)$$

  - A rule for training artificial neural networks.

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

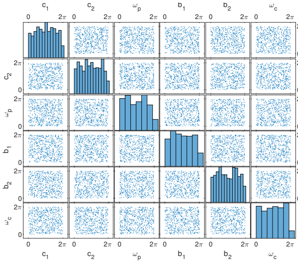
[2] Linsker R. (1997). A local learning rule that enables information maximization for arbitrary input distributions, Neural Computation, 9, pp. 1661–1665.

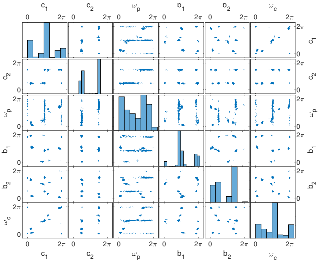
Maximization of stochastic dependence

- Multi-information as generalization[1] of the Infomax principle[2]:

$$\max_{\{f \in \mathcal{F} | O = f(I)\}} MI(I; O)$$

  - A rule for training artificial neural networks.
- ETRPA moves towards maximizing stochastic dependence among elements of the extended multivariate von Mises distributed random vector.





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

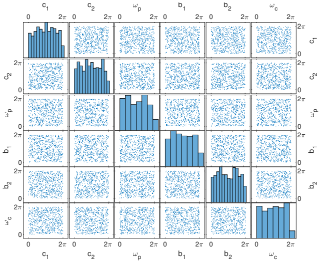
[2]Linsker R. (1997). A local learning rule that enables information maximization for arbitrary input distributions, Neural Computation, 9, pp. 1661–1665.

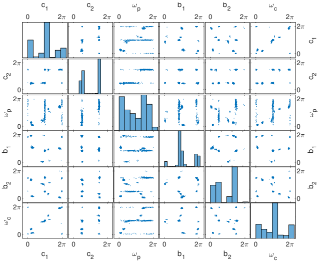
Maximization of stochastic dependence

- Multi-information as generalization[1] of the Infomax principle[2]:

$$\max_{\{f \in \mathcal{F} \mid O = f(I)\}} MI(I; O)$$

  - A rule for training artificial neural networks.
- ETRPA moves towards maximizing stochastic dependence among elements of the extended multivariate von Mises distributed random vector.
  - Learns the optimization landscape given by the CSG packing problem.





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

[2] Linsker R. (1997). A local learning rule that enables information maximization for arbitrary input distributions, Neural Computation, 9, pp. 1661–1665.

THANK YOU