1 Principle of maximum entropy parameter estimation for banded stiffness matrices

Denote by $[[K]]_{\mathcal{N}}$ all the entries $(i,j) \in \mathcal{N}$ of K where $\mathcal{N}$ is a set of indices. For this exercise we will fix $\mathcal{N}$ to be the set of all indices associated to the cgDNA 18×18 block diagonal pattern with 6×6 overlaps. For sake of notation we will omit the $\mathcal{N}$, i.e, $[[K]]_{\mathcal{N}} = [[K]]$, for all $K \in \mathbb{R}^{12n-6 \times 12n-6}$, with $n \in \mathbb{N}$. Moreover with $[[\cdot]]^c$ we denote all the entries $(l,k) \in \mathcal{N}^c$, where $\mathcal{N}^c$ is the complement of $\mathcal{N}$.
Given $\mu \in \mathbb{R}^{12n-6}$ and $C \in \mathbb{R}^{12n-6 \times 12n-6}$ (the observed statistics, mean and covariance, of a n base–pair long molecule of DNA) define the following constraint set:

$$C = \left\{ \rho : \int_\Omega \rho dx = 1,\ \int_\Omega x_k \rho(x) dx = \mu_k,\ k = 1, \dots, 12n-6,\ \int_\Omega x_i x_j \rho(x) dx = c_{ij},\ (i,j) \in \mathcal{N} \right\}. \quad (1)$$

where $\Omega = \mathbb{R}^{12n-6}$. Using the principle of maximum entropy (Lecture, week 11), prove that the maximum entropy distribution is a Gaussian, i.e, it can be written as

$$\rho_{ME}(x) = \frac{1}{Z(\mu, K)} \exp\left\{ -\frac{1}{2}(x-\mu) \cdot K_{ME}(x-\mu) \right\}, \quad (2)$$

where μ is the observed mean and K_{ME} is such that $[[K^{-1}]] = [[C]]$, and $[[K]]^c = 0$.
[Remark: Thanks to the exercise 2 session 10, we know how to compute the matrix K_{ME} directly from the data $[[C]]$.]

2 Estimate of mean and stiffness from MD simulation data

Download from `http://lcvmwww.epfl.ch/teaching/modelling_dna/public_files/plotMatrix2D.m` a visualization tool for cgDNA matrices : `plot2DMatrix`. Load the file`muABC_S3.mat` used previously in Session 10.

1. Compute the inverse of the raw covariance `c1b` to obtain the raw stiffness `s1b`. Visualize both matrices using `plot2DMatrix`. What can you say about the two plots (differences, etc..) ?

2. The cgDNA inter and intra rotations are rescaled by $\frac{1}{5}$ during the parameter extraction procedure, thus

$$\frac{1}{5} u^{cgDNA} = u^{ob},$$
$$\frac{1}{5} \eta^{cgDNA} = \eta^{ob}.$$

   where u_i^{ob}, η_i^{ob} are the rotations observed from MD data. This scaling has been adopted in order to have all the entries of the stiffness matrix and of the mean within a similar range. Rescale correctly only the diagonal entries of the stiffness matrix `s1b`. Extract and then plot the diagonal entries of `s1b` and the diagonal entries of the rescaled matrix. Do the same for the entries of the ground state `shape`. What can you say?

[Note: The invariance of the Kullback-Leibler divergence means that the divergence between two Gaussians p and q does not change for different scaling, see Session 10 , exercise 1.1]

3 Palindromic symmetry of a shape vector and stiffness matrix

As seen in Exercise 2 of Session 5, the mean vector $\mu(S)$ an the stiffness matrix $K(S)$ of the cgDNA Gaussian distribution $\rho(\mathbf{w}, S)$, satisfy the following symmetric properties: $\mu(S) = E_n\mu(\overline{S})$, and $K(S) = E_nK(\overline{S})E_n$, where $\overline{S}$ is the complementary sequence to S and the matrix E_n is defined in Session 5. For palindromic sequences, i.e. when $S = \overline{S}$, these properties become $\mu(S) = E_n\mu(S)$ and $K(S) = E_nK(S)E_n$.

1. Load the file `muABC_S3.mat` with the DNA 18-mer data used previously in Session 10 . Construct the matrix E_n for $n = 18$, and check if the palindromic properties of the shape vector and the covariance matrix are satisfied.

2. Before using the MD data to fit a cgDNA parameter set, in order to minimize modelling errors, it is useful to first get shape and stiffness estimates satisfying palindromic symmetry conditions, if a sequence is palindromic. Symmetrize the shape estimate by computing $\bar{\mathbf{w}}_{sym} = \frac{1}{2}(\bar{\mathbf{w}} + E_n\bar{\mathbf{w}})$. Plot the difference between $\bar{\mathbf{w}}$ and $\bar{\mathbf{w}}_{sym}$.

3. The covariance matrix was obtained as $C = \frac{1}{M}\sum_{j=1}^{M}\mathbf{w}^{[j]}\otimes\mathbf{w}^{[j]} - \bar{\mathbf{w}}\otimes\bar{\mathbf{w}}$. Symmetrize $D := \frac{1}{M}\sum_{j=1}^{M}\mathbf{w}^{[j]}\otimes\mathbf{w}^{[j]}$ and then compute the symmetrized covariance matrix using D_{sym} and $\bar{\mathbf{w}}_{sym}$. Compute the symmetrized stiffness K_{sym} as the inverse of C_{sym}. Using `plot2DMatrix.m`, plot the differences between C and C_{sym} and between K and K_{sym}.

4. Download this scripts: `http://lcvmwww.epfl.ch/teaching/modelling_dna/public_files/MaxEntropy.m`. Use the code `MaxEntropy.m` to get the maximum entropy fit to the symmetrized covariance computed in the previous point.

5. Compute the following Kullback-Leibler divergences:

i) $D(\rho^{sym}_{obs}(S), \rho_{obs}(S))$,
ii) $D(\rho^{sym}_{band}(S), \rho^{sym}_{obs}(S))$,
iii) $D(\rho_{cgDNA}(S), \rho^{sym}_{band}(S))$.

Compare this error with the modelling errors (Kullback-Leibler divergences computed in Exercise 3 of Session 10). Is the convergence error big or small compared to the modelling errors?