Modifications to the EMC Algorithm for Orientation Recovery in Single Particle Imaging Experiments on X-ray Free Electron Lasers

Sergei Zolotarev

National Research Centre “Kurchatov Institute,”, pl. Akademika Kurchatova 1,
Moscow, 123182, Russia

Abstract. The emergence of super-bright light sources - X-ray free electron lasers(XFELs) combined with Single Particle Imaging(SPI) method, makes it possible to obtain nanometer resolution 3D structure of biological particles such as proteins or viruses without needing to freeze them. SPI relies on the “diffraction before destruction” principle, meaning that each sample only produces a single diffraction image before being destroyed by an X-ray pulse. The orientation of the particle in the beam is random for each shot. This gives rise to the problem of orientation recovery, in which an array of 2D diffraction images has to be combined into a single 3D image, necessary for the reconstruction of 3D structure of the studied particle. The orientation recovery problem is most commonly solved by the EMC algorithm, which is the most computationally expensive part of data analysis for SPI experiments. In this work we introduce several modifications to the EMC algorithm aimed at improving the quality of reconstruction and/or increasing the algorithm’s speed of convergence. We analyse the effectiveness of these modifications using simulated diffraction data.

1 Introduction

The emergence of fourth generation light sources - X-ray Free Electron Lasers (XFEL), opens up new possibilities in many fields of science. Compared to third generation synchrotron light sources, XFELs produce extremely bright (more than 10^{12} photons) and extremely short (tens of femtoseconds) x-ray pulses [6]. These properties provide a new way to study the 3D structure of bioparticles such as proteins and viruses - Single Particle Imaging (SPI) [8]. Compared to other methods such as X-Ray crystallography or cryogenic electron microscopy, SPI has the advantage of being able to study non-crystalline bioparticles in their natural state (suspended in water).

SPI relies on ”diffraction before destruction”[2] principle, which abuses the unique properties of XFEL pulses. The pulses are bright enough to produce an informative diffraction pattern, and they are at the same time are short enough, so that all the scattering happens before the radiation damage takes place. Since the scattering particle is destroyed in the process, multiple identical particles are injected into the XFEL beam during the experiment.

This gives rise to orientation recovery problem: SPI produces a number of flat diffraction images of the particle in an unknown orientation. In order to reconstruct the 3D structure of said particle these images need to be combined to produce a single 3D diffraction density. The established way to solve this problem is by using the EMC algorithm [5].

In this work we introduce three modifications of orientation recovery algorithm EMC, which aim to improve quality of reconstruction and to increase the speed of convergence. We examine their effectiveness using simulated diffraction data.

2 Methods

2.1 Orientation recovery

In orientation recovery problem we have an array of M_{data} diffraction patterns scattered by identical particles in unknown orientations. Each such pattern is a spherical slice of a single 3D density in reciprocal space, and the goal of an orientation recovery algorithm is to reconstruct this 3D density W by finding to which orientation each diffraction pattern belongs to. This problem is most commonly solved by the EMC algorithm.

EMC algorithm starts with input data (diffraction images K) and an initial approximation of 3D diffraction density W^0. Then this initial approximation is iteratively improved until it converges to the final value of W^T. Each iteration of the algorithm consists of three steps which give the name to the algorithm: Expand, Maximize and Compress.

During Expansion step the current approximation W, which is typically represented by its values on regular 3D grid, is converted to tomographic representation W_{ij} - values of W in points corresponding to i-th pixel of the diffraction image with scattering particle in j-th orientation. In order to do this we sample the 3D rotation group SO(3) by a finite number of "evenly spaced" rotations R_j ($j = 1 \dots M_{rot}$). And for each pixel of the detector we calculate a corresponding point in reciprocal space q_i ($i = 1 \dots M_{pix}$). For typical flat detectors all these points will be lying on Ewald's sphere. After calculating rotations R_j and points q_i we can define W_{ij} as $W(R_j \cdot q_i)$, which is calculated via linear interpolation.

The Maximization step updates current approximation of 3D diffraction density $W_{ij} \to W'_{ij}$ based on maximizing log-likelihood function $Q(W')$. This step is equivalent to one iteration of EM algorithm [3] and in itself consists of two steps. First we calculate P_{jk} - the probabilities of each image D_k ($k = 1 \dots M_{data}$) to be produced by the particle in j-th orientation, conditional on current model values W_{ij} as the product of Poisson probabilities at each detector pixel. Then we calculate new values of 3D density W'_{ij}, maximizing the expected log-likelihood:

$$W'_{ij} = \frac{\sum\limits_{k=1}^{M_{data}} P_{jk} D_{ik}}{\sum\limits_{k=1}^{M_{data}} P_{jk}}.$$

Finally, Compression step converts W'_{ij} back onto regular 3D grid by reversing the interpolation procedure used in the expansion step. In essence, EMC algorithm is equivalent to EM where after each iteration we perform combination of compression and expansion steps, which enforce extra constraints on current model W_{ij}. Whereas for EM all W_{ij} are treated as independent variables, in reality they are derived from values on 3D intensity grid, and when the distance between points $R_j \cdot q_i$ and $R_{j'} \cdot q_{i'}$ is small, values W_{ij} and $W_{i'j'}$ are not independent.

In this work we propose three modifications to the EMC algorithm:

Incremental EMC modifies the maximization procedure, instead of performing a single update of 3D intensity W using data from all images, only one image is used on each iteration. For randomly selected image D_{k^*} probabilities P_{jk^*} are calculated and then the current values of W_{ij} are immediately updated:

$$W'_{ij} = \frac{(\sum\limits_{k=1}^{M_{data}} P^{old}_{jk} D_{ik}) - P^{old}_{jk^*} D_{ik^*} + P_{jk^*} D_{ik^*}}{(\sum\limits_{k=1}^{M_{data}} P^{old}_{jk}) - P^{old}_{jk^*} + P_{jk^*}}.$$

Such an update can be performed in O(1) time if the numerator and denominator of W_{ij} are saved separately, as well as all the values of P^{old}_{jk} [7]. This way M_{data} of such updates can be performed in the same time as the single maximization step of EMC algorithm. This incremental M-step is followed by compression and expansion steps same as in regular EMC algorithm. However, expansion steps outputs only values of W_{ij}, and in order to use out incremental maximization step again we first need to perform a normal M-step to obtain separate values of numerators $\sum_{k=1}^{M_{data}} P_{jk} D_{ik}$, denominators $\sum_{k=1}^{M_{data}} P_{jk}$ and probabilities P^{old}_{jk}. Effectively, this modification of EMC alternates its M-steps between normal and incremental, and only performs compression and expansion steps after every other iteration.

Stepwise EMC uses batches of images in its maximization step. First a new 3D diffraction density W^*_{ij} is calculated using only a subset of images $D^* \in D$. Then we take weighted average between current desity W_{ij} and W^*_{ij} as the result of maximization:

$$W'_{ij} = (1 - (2 + t)^{-\gamma}) W_{ij} + (2 + t)^{-\gamma} W^*_{ij},$$

where t is the number of iteration and $0.5 < \gamma \leq 1$ is a coefficient that ensures that the weight before newly calculated density W^* exponentially decreases, thus guaranteeing convergence of the algorithm [10]. This maximization step is then performed several more times with different subsets of images, until all M_{data} images have been used. Then follow normal compression and expansion steps.

Adaptive EMC is the final proposed modification which does not modify any of the three steps of EMC, and instead it only takes effect after the M-step, when the new values of W'_{ij} are calculated. Instead of taking the point W' itself, which maximizes the expected log-likelihood function $Q(W')$, we try to go further in the same direction:

$$\hat{W}_{ij} = W_{ij} + \alpha(W'_{ij} - W_{ij}),$$

where coefficient $\alpha \geq 1$ is adaptively changed based on log-likelihood $Q(\hat{W})$. If $Q(\hat{W}) \geq Q(W)$ then α is increased and the $\hat{W}$ is accepted as the result of current iteration, otherwise α is reset to 1 and W' is taken as a result instead. This approach guarantees that log-likelihood never decreases and the algorithm thus converges [9].

2.2 Testing the algorithms

To evaluate the performance of our modifications and to compare them to regular EMC we tested them using simulated diffraction data.

In the first test we used a binary contrast torus in reciprocal space as our object. Using such a simple model allows us to successfully perform orientation recovery in a short time and without tuning of any reconstruction parameters. This model allows us to easily compare different algorithms, but it has a drawback of being very far removed from actual experimental data. For that reason we performed a second more life-like test.

In the second test we used diffraction patterns of keyhole limpet hemocyanin [4] simulated by Dragonfly [1] software package. In order to produce successful reconstructions from this data, we used deterministic annealing modification of EMC as described in [1].

In both tests for each algorithm we performed the reconstruction 5 times, using a random initial approximation of 3D diffraction density W^0. Then we compared the output of each iteration W^t with the initial density W_{true}, that was used to generate the diffraction images. We used root mean square difference (RMS difference) between these two 3D diffraction densities as our metric of quality of reconstruction.

3 Results

In the first test we used 1000 images with 4900 pixels in each image. Number of possible rotations M_{rot} was increasing every 10 iterations. Starting from 420

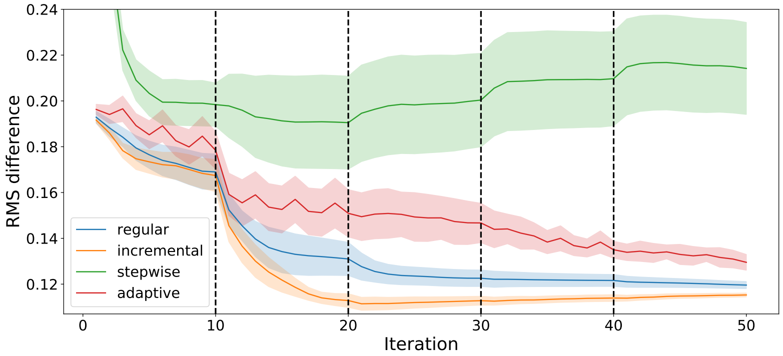

Fig. 1. Average and standard deviation of root mean square difference between the output of each algorithm and the initial 3D diffraction density. Dashed lines indicate the points where the number of rotations considered in reconstruction changed.

possible orientations on the first ten iterations and up to 10860 orientations for iterations 41 through 50. The results of this test are presented on Fig. 1.

One can see that out of three proposed modifications, only incremental EMC demonstrated better results than unmodified algorithm. Stepwise EMC proved to be too dependent on the random selection of the first batch of images used in reconstruction. Adaptive EMC didn't provide any boost to the speed of convergence, due to the fact that the coefficient α never became greater than 1 without decreasing the expected log-likelihood.

For these reasons stepwise and adaptive modifications were ruled out as ineffective, and for the second test only Incremental EMC was evaluated. In this test we used 5000 images with 10000 pixels and 10860 possible orientations. The results of this test are presented on Fig. 2.

Both algorithms demonstrated similar results, however our chosen metric of quality doesn't perform very well on this test. The difference between 0-th iteration which is just random noise and the final output of the algorithm is quite small, when looking only at the RMS difference between the reconstructed and initial 3D diffraction density. Due to this we cannot conclusively evaluate the performance of incremental modification of EMC algorithm.

4 Conclusion

In this work we proposed, developed, and tested three modifications to the EMC algorithm, used to solve orientation recovery problem in SPI experiments. After the first preliminary test, adaptive and stepwise modifications have shown worse results than unmodified algorithm and were thus ruled out as non-viable. For the more promising incremental modification we performed a second set of tests with more life-like input data. In these tests both incremental and regular EMC

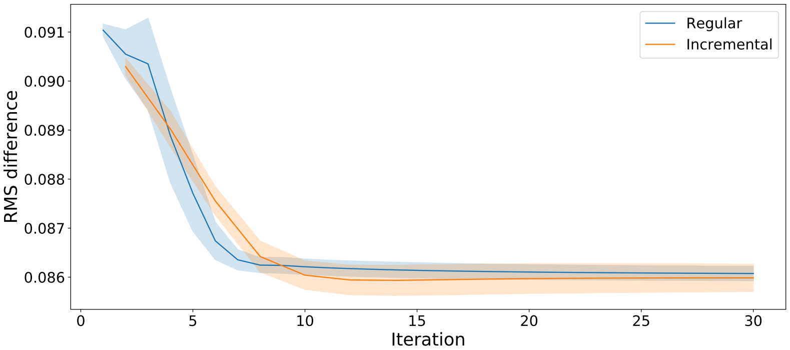

Fig. 2. Average and standard deviation of root mean square difference between the output of regular and incremental EMC algorithms and the initial 3D diffraction density.

demonstrated similar results, but due to instability of reconstruction process and difficulty of the evaluation of reconstruction quality, we can not definitively say that one algorithm is better than the other. Additional tests may be required to establish that.

5 Acknowledgements

This research was supported by Russian Science Foundation (project No. 18-41-06001) and Helmholtz Associations Initiative Networking Fund (HRSF-0002).

References

1. Ayyer, K., Lan, T.Y., Elser, V., Loh, N.D.: *Dragonfly*: an implementation of the expand–maximize–compress algorithm for single-particle imaging. Journal of Applied Crystallography **49**(4), 1320–1335 (Aug 2016). https://doi.org/10.1107/S1600576716008165
2. Chapman, H.N., Barty, A., Bogan, M.J., Boutet, S., Frank, M., Hau-Riege, S.P., Marchesini, S., Woods, B.W., Bajt, S., Benner, W.H., London, R.A., Plonjes, E., Kuhlmann, M., Treusch, R., Dusterer, S., Tschentscher, T., Schneider, J.R., Spiller, E., Moller, T., Bostedt, C., Hoener, M., Shapiro, D.A., Hodgson, K.O., van der Spoel, D., Burmeister, F., Bergh, M., Caleman, C., Huldt, G., Seibert, M.M., Maia, F.R.N.C., Lee, R.W., Szoke, A., Timneanu, N., Hajdu, J.: Femtosecond diffractive imaging with a soft-x-ray free-electron laser. Nature Physics **2**(12), 839–843 (Nov 2006). https://doi.org/10.1038/nphys461
3. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological) **39**(1), 1–22 (1977). https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

4. Gatsogiannis, C., Markl, J.: Keyhole limpet hemocyanin: 9-a CryoEM structure and molecular model of the KLH1 didecamer reveal the interfaces and intricate topology of the 160 functional units. Journal of Molecular Biology **385**(3), 963–983 (Jan 2009). https://doi.org/10.1016/j.jmb.2008.10.080
5. Loh, N.T.D., Elser, V.: Reconstruction algorithm for single-particle diffraction imaging experiments. Physical Review E **80**(2) (Aug 2009). https://doi.org/10.1103/physreve.80.026705
6. Mcneil, B., Thompson, N.: X-ray free-electron lasers. Nature Photonics **4** (12 2010). https://doi.org/10.1038/nphoton.2010.239
7. Neal, R.M., Hinton, G.E.: A View of the Em Algorithm that Justifies Incremental, Sparse, and other Variants, pp. 355–368. Springer Netherlands, Dordrecht (1998). https://doi.org/10.1007/978-94-011-5014-9-12
8. Neutze, R., Wouts, R., van der Spoel, D., Weckert, E., Hajdu, J.: Potential for biomolecular imaging with femtosecond x-ray pulses. Nature **406**(6797), 752–757 (Aug 2000). https://doi.org/10.1038/35021099
9. Salakhutdinov, R., Roweis, S.: Adaptive overrelaxed bound optimization methods **2** (10 2003)
10. Sato, M., Ishii, S.: On-line EM algorithm for the normalized gaussian network. Neural Computation **12**(2), 407–432 (Feb 2000). https://doi.org/10.1162/089976600300015853