Supplementary Information for cryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination

Supplementary Note 1: Stochastic Gradient Descent (SGD)

SGD iteratively optimizes an objective function by computing approximate gradients and taking steps in the parameter space according to those gradients. The two key parts of SGD are thus computation of the approximate gradient and determination of an appropriate parameter update step based on the gradient.

Objective function. In the case of cryo-EM structure determination for heterogeneous specimens, the optimization objective function is the log posterior probability distribution over K 3-D densities, $\mathbf{V} = \{V_i\}_{i=1\ldots K}$, given N particle images $\{X_j\}_{j=1\ldots N}$; i.e.,

$$\begin{aligned}
&\arg\max_{V_1,V_2,\ldots,V_K} \log p(V_1, V_2, \ldots, V_K | X_1, X_2, \ldots, X_N) && (1)\\
&= \arg\max_{V_1,\ldots,V_K} \log p(X_1, \ldots, X_N | V_1, \ldots, V_K) + \log p(V_1, \ldots, V_K) && (2)\\
&= \arg\max_{V_1,\ldots,V_K} \sum_{i=1}^{N} \log p(X_i|\mathbf{V}) + \sum_{j=1}^{K} \log p(V_j) && (3)\\
&\equiv \arg\max_{\mathbf{V}} f(\mathbf{V})
\end{aligned}$$

Here p denotes probability density. The objective function $f(\mathbf{V})$ is proportional to the log posterior probability of the heterogeneous structures given the observed images. This posterior probability is a marginal probability, with marginalization taken over the unknown variables of 3-D orientation, 2-D shift and class assignment for each image. Known parameters, including the CTF, are omitted for notational clarity.

In Equation (2) the second term is a joint prior over the 3-D structures. This prior can be set, for example, to restrict density to be strictly positive or to penalize high-frequency noise in structures. In this work, the prior is assumed to be independent over the structures, meaning that it factors into a separate prior for each structure in Equation (3).

Nature Methods: doi:10.1038/nmeth.4169

The objective function in Equation (3) is expanded below to explicitly indicate marginalization over parameters.

$$f(\mathbf{V}) = \sum_{i=1}^{N} \log p(X_i|\mathbf{V}) + \sum_{j=1}^{K} \log p(V_j)$$

where

$$p(X_i|\mathbf{V}) = \sum_{j=1}^{K} \pi_j p(X_i|V_j) \equiv U_i \tag{4}$$

$$p(X_i|V_j) = \int p(X_i|\phi, V_j) p(\phi) d\phi \ . \tag{5}$$

Therefore,

$$f(\mathbf{V}) = \sum_{i=1}^{N} \log \left(\sum_{j=1}^{K} \pi_j \int p(X_i|\phi, V_j) p(\phi) d\phi \right) + \sum_{j=1}^{K} \log p(V_j) \tag{6}$$

Equation (4) gives the likelihood of observing a single image, X_i, given all the 3-D structures $\mathbf{V}$, by marginalizing over the class assignment j. Mixing probabilities are given by π_j, and in this work they are assumed to be uniform over classes, i.e., $\pi_j = K^{-1}$. Equation (5) gives the likelihood of observing a single image from a single structure V_j, this time marginalizing over the unknown 3-D orientation and 2-D shift, denoted together by pose ϕ. A prior over poses $p(\phi)$ can be specified; in this work a uniform distribution is again used. The integrand $p(X_i|\phi, V_j)$ is the probability of observing an image X_i from a particular pose ϕ of a particular 3-D structure V_j, and is given by the microscope image formation model (including CTF) and sensor noise characteristics, as in existing likelihood-based methods [1, 2]. The specific form of $p(X_i|\phi, V_j)$ is given later in Equation (10). Equation (6) is the expanded form of the objective function.

Gradient. SGD optimizes the objective function in Equation (6) by iteratively updating the parameters $V_1, \ldots, V_K$. These structures are represented as voxels of density on 3-D grids.

The gradient of Equation (6) with respect to each structure V_k is computed in order to take steps. This gradient is

$$\begin{aligned} \frac{\partial f}{\partial V_k} &= \sum_{i=1}^{N} \frac{1}{U_i} \frac{\partial U_i}{\partial V_k} + \frac{\partial}{\partial V_k} \log p(V_k) \\ &= \sum_{i=1}^{N} \frac{1}{U_i} \pi_k \int \frac{\partial}{\partial V_k} p(X_i|\phi, V_k)\, p(\phi) d\phi + \frac{\partial}{\partial V_k} \log p(V_k) \end{aligned} \tag{7}$$

Here U_i is the likelihood of observing image X_i, defined in Equation (4). The integrand in Equation (7) is the gradient of the cryo-EM image formation model with respect to structure V_k [2].

Nature Methods: doi:10.1038/nmeth.4169

Approximate gradient. The sum giving the gradient in Equation (7) is over all N single-particle images in the dataset. In SGD, the sum is approximated using subsampling. At each iteration, SGD selects a subset of images, called a minibatch, at random from the entire dataset, and uses only those images to approximate Equation (7). The size M of a minibatch $\mathbf{M}$ can vary over iterations. In this work the minibatch size is set automatically based on the current resolution and the number of classes K. The approximate gradients are given by

$$\frac{\partial f}{\partial V_k} \approx G_k \equiv \frac{N}{M} \sum_{i \in \mathbf{M}} \frac{1}{U_i} \pi_k \int \frac{\partial}{\partial V_k} p(X_i|\phi, V_k)\, p(\phi) d\phi + \frac{\partial}{\partial V_k} \log p(V_k) \qquad (8)$$

SGD update rule with momentum. The approximate gradient in Equation (8) points in a direction within the space of 3-D structures that will, in expectation over random selections of minibatches, improve the objective function in Equation (6). Over many iterations, following these noisy directions allows SGD to explore the space of 3-D structures. It is well known that in the general case, optimization of non-convex objective functions like Equation (6) is difficult and SGD only provides guarantees of local convergence [3]. Nevertheless, in practice we find that SGD performs well and finds the correct 3-D structures.

SGD maintains a current estimate of each structure V_k, denoted by $V_k^{(n)}$ at the n-th iteration. The update applied at the previous iteration, $dV_k^{(n-1)}$ is also recorded. SGD computes the update at the current iteration, $dV_k^{(n)}$, by scaling the current gradient $G_k^{(n)}$ with a step-size η_k and combining this linearly with the previous update in a ratio given by μ. This linear averaging is known as momentum [4] and serves to smooth the noisy approximate gradient directions in SGD.

$$dV_k^{(n)} = (\mu) dV_k^{(n-1)} + (1-\mu)(\eta_k) G_k^{(n)}$$
$$V_k^{(n+1)} = V_k^{(n)} + dV_k^{(n)}$$

Each structure V_k is updated using a different step-size η_k to allow classes with different numbers of particles and with different geometries to change at different rates over iterations. The momentum parameter μ is fixed at 0.9.

Step-sizes. In most gradient-descent algorithms, setting the step-size can often require tuning. Step-sizes that are too small yield slow convergence, while step-sizes that are too large can cause divergence of the algorithm. SGD generally has a similar property, but analysis of a particular optimization problem can yield methods for automatically setting the step-sizes η_k to appropriate values.

In this work the step-sizes are set using an approximation of the second-order curvature of the objective function $f(\mathbf{V})$. Generally $f(\mathbf{V})$ is non-convex, multi-modal, but smooth and differentiable. It is possible in principle to compute the direct Hessian (matrix of second derivatives), but this matrix would be

excessively large and slow to compute. Instead, two approximations are used. First, the Hessian in Fourier space is assumed to be diagonal. This approximation is reasonable due to the Fourier-slice theorem, and is used extensively in maximum-likelihood approaches for cryo-EM refinement [1], although not explicitly stated. Second, rather than computing the true second derivatives, a surrogate objective function $\tilde{f}(\mathbf{V})$ is constructed that is a convex quadratic lower bound on the true objective function, and the second derivatives of this surrogate are computed instead. This construction is identical to the one used to derive the maximization step of the Expectation-Maximization algorithm [5]. The two approximations allow the rapid computation of an approximate Hessian for $f(\mathbf{V})$ that gives diagonal Fourier space curvature information about the objective function. The maximum curvature over all dimensions in Fourier space is used directly as the inverse step-size for each structure. Computation of the approximate Hessian is carried out over the selected minibatch in each iteration, and re-uses most of the computation required for computing the gradient, leading to further efficiencies.

Concretely, the inverse step-size for each structure is given by

$$\frac{1}{\eta_k} = \left\| \sum_{i \in \mathbf{M}} \pi_k \int p(\phi | X_i, V_k) \mathbf{P}_\phi^\top \frac{C_i^2}{\sigma^2} d\phi \right\|_\infty \tag{9}$$

Here $\mathbf{P}$ is the projection operator for pose ϕ, C_i are the CTF values for image i, and σ^2 are the noise variances. The term inside the infinity norm is a vector containing the diagonal values of the approximate Hessian, and the infinity norm is equivalent to selecting the maximum element.

Minibatch sizes. The SGD algorithm is generally quite robust to the setting of minibatch size, and in fact, the proofs of convergence of the SGD algorithm [3] are valid for a minibatch size as small as a single data point. In practice, however, it becomes computationally expensive to use excessively small minibatch sizes, as the full cost of updating the model with a new step is incurred for every minibatch, regardless of the minibatch size used. In this work, the minibatch size is set to $30 \times K$ initially, and changed to $100 \times K$ once the resolution being considered exceeds 20 Å.

Noise model. In the SGD algorithm, the objective function involves marginalizing over poses and class assignments. This marginalization is sensitive to the choice of noise model used in the image formation model. In most cryo-EM refinement algorithms, the noise model is understood to represent the shot noise and the readout noise of the microscope camera. In *ab initio* reconstruction, it is critical that the noise model also include model error. Model error refers to the distance between the current estimate of the 3-D structure and the true 3-D structure. Upon random initialization of SGD, this distance can be large. The noise model must therefore be initialized with a large variance, to account for image noise plus model error.

Nature Methods: doi:10.1038/nmeth.4169

In this work, the variance of the image noise is gradually estimated from the data during reconstruction. Specifically, the noise at frequency ℓ at iteration k is set to

$$\sigma^2_{\ell,k} = \frac{w_{\ell,k}\bar{\sigma}^2_{\ell,k} + \tilde{w}\tilde{\sigma}^2 + \hat{w}_k\hat{\sigma}^2}{w_{\ell,k} + \tilde{w} + \hat{w}_k}$$

where $w_{\ell,k} = \sum_{i=1}^{k} \gamma^{k-i} M_i C^2_{\ell,i}$, M_i is the batch size at iteration i, and $\bar{\sigma}^2_{\ell,k} = \sum_{i=1}^{k} \gamma^{k-i} M_i e^2_{\ell,i}$, with $e^2_{\ell,i}$ the average reconstruction error of frequency ℓ during iteration i and $C^2_{\ell,i}$ the average squared CTF of frequency ℓ at iteration i. This corresponds to an augmented approximate maximum a posterior (MAP) estimate of the variance based on decaying running averages with decay rate $\gamma = 0.9999$ and prior weight $\tilde{w} = 50$ and $\tilde{\sigma}^2$ is the initial white noise variance estimated from the corners of the particle images. To ensure that the objective function is relatively smooth initially an inflated initial noise prior is included with weight $\hat{w}_k = 2500\gamma^k$ and variance $\hat{\sigma}^2 = 8\tilde{\sigma}^2$. This inflated noise model accounts for the initial error in the 3-D model and decays over time until it effectively has no influence. This process causes the noise model to gradually decrease in variance from a high initial value to a final value that is roughly equal to a MAP estimate of the image noise once the 3-D map has converged.

Random initialization. SGD is able to converge to correct structures from arbitrary randomly generated initializations containing no prior structural knowledge or user expertise. In this work the initializations are generated by selecting a small random subset of images from the dataset (typically 10 images), assigning them randomly generated pose angles, and using them to reconstruct a 3-D volume. This process creates a 3-D map with random structure, but approximately correct overall scale and spatial extent. This structure is fed directly into SGD.

Supplementary Note 2: Branch and Bound Search

SGD is able to compute *ab initio* structures to medium resolution (approx. 10 Å). Once SGD has converged, the Expectation-Maximization algorithm (also known as iterative refinement) is used to refine the resulting structures to high resolution. The computationally expensive part of iterative refinement is the expectation step, in which each 2-D image is aligned over 3-D orientations and 2-D translations to the current estimate of each 3-D structure. In cryoSPARC this search problem is solved efficiently using a branch and bound search technique. In general, branch and bound search involves repeatedly computing a lower bound and an upper bound on the search criterion of interest. It then uses these bounds to exclude regions of search space from further search at each repetition.

Image alignment problem setup. The problem of image alignment is to find the optimal pose ϕ (3-D orientation and 2-D translation) that aligns a given

image X_i with a structure V_k. The fit between X_i and a projection of V_k from pose ϕ is given by the probability of having observed X_i at pose ϕ, so the task reduces to maximizing this probability. For mathematical and computational convenience, the task is equivalently rewritten as the minimization of the negative log probability. Negative log probability is a measure of image alignment error, and is a function of the pose of the particle in image X_i.

In this section, the pose variable ϕ is broken into two parts, with r denoting the 3-D orientation of the particle, and t denoting the 2-D translation of the particle within the particle image. The following describes image alignment for a single image X and a single structure V, so subscripts i and k are omitted for notational clarity.

As is common in the literature, the probability of observing an image from a particular pose is given in the Fourier domain as follows:

$$p(X|\phi, V) = p(X|r, t, V) = \frac{1}{Z} \exp\left(\sum_{\ell} \frac{-1}{2\sigma_\ell^2} |C_\ell Y_\ell(r) - S_\ell(t) X_\ell|^2\right) \tag{10}$$

where

$$Y_\ell(r) = \Theta_\ell(r) V \ .$$

Here, with the image X (2-D) and model V (3-D) represented in Fourier space, the log likelihood involves a sum over Fourier coefficients ℓ. Accordingly, $Y_\ell(r)$ denotes the projection of model V according to pose r, at frequency ℓ. Poses can be parameterized in any suitable fashion, but in this work the axis-angle formulation is used. The subscript ℓ denotes a two-component index of a particular Fourier coefficient, also known as a wavevector. The sum over ℓ is shorthand for summing over all wavevectors in 2-D (i.e., the Fourier domain of the image). C denotes the contrast transfer function (CTF) of the microscope, and $\Theta_\ell(r)$ is a linear projection operator, corresponding to the slice operator in Fourier space, with pose r, for wavevector ℓ. S denotes the 2-D phase shift corresponding to a 2-D translation of t pixels. The normalizing constant Z can be ignored because it does not depend on the unknown pose r, t. The noise parameter σ_ℓ represents the level of Gaussian noise expected at each frequency, with a possibly different variance for each Fourier coefficient (allowing for white or colored noise models). For notational clarity in what follows we assume a white noise model with $\sigma_\ell = \sigma = 1$ but the general case with colored noise is a simple extension. In practice and in the results presented, a colored noise model is used, with σ_ℓ estimated from the data.

Taking the negative log of Equation (10) gives the image alignment error (the negative log likelihood), which is the squared error in the Fourier coefficients:

$$E(r, t) = \sum_{\ell} \frac{1}{2} |C_\ell Y_\ell(r) - S_\ell(t) X_\ell|^2 \tag{11}$$

The aim of image alignment is to find r and t that minimize this function for the given image X and model V.

Intuition behind a lower bound. The core challenge in employing a branch and bound method for cryo-EM is to derive a lower bound that is inexpensive to evaluate but informative about the image alignment error function $E(r,t)$. Constructing a useful lower bound requires insight into the characteristics of $E(r,t)$.

The following derivation starts from a simple, well-known intuition: if an image aligns poorly to a structure at low resolution, it will not align well at high resolution. This property means that if we evaluate the likelihood of an image across poses using only low-resolution Fourier coefficients, the resulting values should give us some knowledge about which regions of pose space are worth pursuing at high resolution.

To make this intuition concrete and move towards a bound on E, note that the negative log-likelihood in Equation (11) is a sum of squared error terms. As a consequence, each Fourier coefficient contributes independently with equal weight (assuming white noise) to the total squared error E. Critically, the contribution of each coefficient is related to how much *power* there is in that Fourier coefficient. If a Fourier coefficient in the model V with wavevector ℓ has no power, that coefficient will only contribute a term equal to $\frac{1}{2}|X_\ell|^2$ to E, and that term does not depend on the pose (r,t) and thus does not need to be considered during search. The bound developed in this work exploits the fact that a Fourier coefficient in the model that has non-zero but small power also gives a small and limited possible pose-dependent contribution to E.

The intuition above indicates that if Fourier coefficients of the model at higher resolutions have limited power, there is a limit to how much they can impact the squared error E. If the low-frequency coefficients already have a given error at a particular pose, the high-frequency coefficients cannot make this error much better or worse. In this work, inexpensive evaluations of the squared error at low resolutions are used to bound true values of E, allowing branch and bound to eliminate search regions without computing the sum in Equation (11) entirely.

Derivation of a lower bound. To derive a lower bound, which is always less than E, we first split E into two parts, denoted A and B, as follows:

$$E(r,t) = \underbrace{\sum_{\|\ell\| \leq L} \frac{1}{2}|C_\ell Y_\ell(r) - S_\ell(t) X_\ell|^2}_{\equiv A(r,t)} + \underbrace{\sum_{\|\ell\| > L} \frac{1}{2}|C_\ell Y_\ell(r) - S_\ell(t) X_\ell|^2}_{\equiv B(r,t)} \qquad (12)$$

Here, A is the squared error of Fourier coefficients at or below a certain radius L in Fourier space, and B is the squared error of coefficients above that radius. The derivation that follows is general and valid for any radius L. In the branch and bound algorithm, L is initialized as a small value, and then increased with each iteration until reaching the Nyquist frequency. The schedule for increasing L is discussed in a subsequent section.

In order to bound E, we compute A directly, which is inexpensive when L is small. We then bound B from below. To derive that bound, it is convenient

Nature Methods: doi:10.1038/nmeth.4169

to further split B into three parts:

$$\begin{aligned} B(r,t) &= \sum_{\|\ell\|>L} \frac{1}{2}|C_\ell Y_\ell(r) - S_\ell(t)X_\ell|^2 \\ &= \underbrace{\sum_{\|\ell\|>L} \frac{1}{2}|X_\ell|^2}_{\equiv B_1} + \underbrace{\sum_{\|\ell\|>L} \frac{1}{2}C_\ell^2|Y_\ell(r)|^2}_{\equiv B_2} - \underbrace{\sum_{\|\ell\|>L} C_\ell \Re(Y_\ell(r)^* S_\ell(t) X_\ell)}_{\equiv B_3} \ , \end{aligned} \quad (13)$$

where $\Re(z)$ denotes the real part of a complex-valued z. Here, the fact that $|S_\ell(t)| = 1$ and the fact that the CTF is real-valued are used, and * denotes complex conjugation. The first term, B_1, is the total power of the image at high frequencies, and does not depend on r, t. The second term B_2 is the total power at high frequencies of a slice of the model from pose r. B_2 does not depend on t. The third term B_3 is the correlation between the shifted image X and the slice of the 3-D model in the Fourier domain.

First consider B_3; an upper bound on B_3 contributes to a lower bound on B. The cryo-EM image formation model states that the observed image X is the true signal $\tilde{X}$ modulated by the CTF, plus independent identically-distributed noise in Fourier space:

$$\begin{aligned} X_\ell &= C_\ell \tilde{X}_\ell + \epsilon_\ell \\ \text{where } \epsilon_\ell &\sim \mathcal{CN}\left(0, \frac{1}{2}\right) \end{aligned} \quad (14)$$

Here, each ϵ_ℓ is a complex normal random variable. The variance is $\frac{1}{2}$ in the white noise case due to the real-valued white noise signal. In the case of colored noise it would be $\sigma_\ell^2/2$.

Inserting Equation (14) into B_3 yields

$$\begin{aligned} B_3 &= \sum_{\|\ell\|>L} C_\ell^2 \Re(Y_\ell(r)^* S_\ell(t) \tilde{X}_\ell) + \underbrace{\sum_{\|\ell\|>L} C_\ell \Re(Y_\ell(r)^* S_\ell(t) \epsilon_\ell)}_{\equiv H} \\ &= \sum_{\|\ell\|>L} C_\ell^2 \Re(Y_\ell(r)^* S_\ell(t) \tilde{X}_\ell) + H \\ &\leq \sum_{\|\ell\|>L} C_\ell^2 |Y_\ell(r)| \, |\tilde{X}_\ell| + H \end{aligned} \quad (15)$$

Here the triangle inequality has been used, and H is a random variable corresponding to the correlation between the model slice $Y(r)$ and the random noise ϵ. H captures the expectation of the effect of noise on E. H can be simplified

as follows:

$$\begin{aligned}
H &= \sum_{\|\ell\|>L} C_\ell \Re(Y_\ell(r)^* S_\ell(t)\epsilon_\ell) \\
&= \sum_{\|\ell\|>L} C_\ell \Re(Y_\ell(r)^* \epsilon_\ell) \\
&= \sum_{\|\ell\|>L} C_\ell \Re\left(Y_\ell(r)^* \mathcal{CN}\left(0, \frac{1}{2}\right)\right) \\
&= \sum_{\|\ell\|>L} C_\ell \Re\left(\mathcal{CN}\left(0, \frac{1}{2}|Y_\ell(r)|^2\right)\right) \\
&= \sum_{\|\ell\|>L} \mathcal{N}\left(0, \frac{1}{2}C_\ell^2|Y_\ell(r)|^2\right) \\
&= \mathcal{N}\left(0, \sum_{\|\ell\|>L} \frac{1}{2}C_\ell^2|Y_\ell(r)|^2\right)
\end{aligned}$$

The first line above uses the fact that the noise variables ϵ_ℓ are uniform over phase, and so their distribution is invariant to the phase shift of $S_\ell(t)$. The final line shows that H, the contribution to B_3 from noise in the image, is normally distributed with variance

$$\sigma_H^2 = \sum_{\|\ell\|>L} \frac{1}{2}C_\ell^2|Y_\ell(r)|^2 \tag{16}$$

Returning to B_2 and B_3 and incorporating Equation (15):

$$\begin{aligned}
B_2 - B_3 &= \sum_{\|\ell\|>L} \frac{1}{2}C_\ell^2|Y_\ell(r)|^2 - \sum_{\|\ell\|>L} C_\ell \Re(Y_\ell(r)^* S_\ell(t) X_\ell) \\
&\geq \sum_{\|\ell\|>L} \frac{1}{2}C_\ell^2|Y_\ell(r)|^2 - \sum_{\|\ell\|>L} C_\ell^2|Y_\ell(r)||\tilde{X}_\ell| - H \\
&= \underbrace{\sum_{\|\ell\|>L} \frac{1}{2}\left(C_\ell^2|Y_\ell(r)|^2 - 2C_\ell^2|Y_\ell(r)||\tilde{X}_\ell|\right)}_{\equiv Q} - H
\end{aligned} \tag{17}$$

Each term of Q is a positive-definite quadratic function of $Y_\ell(r)$. Therefore

Q can be bounded from below:

$$\begin{aligned} Q &\geq \min_{Y_\ell(r)} Q \\ &= \min_{Y_\ell(r)} \sum_{\|\ell\|>L} \frac{1}{2} \left(C_\ell^2 |Y_\ell(r)|^2 - 2C_\ell^2 |Y_\ell(r)||\tilde{X}_\ell| \right) \\ &\quad \text{attained at } Y_\ell(r) = \tilde{X}_\ell \\ &= \sum_{\|\ell\|>L} -\frac{1}{2} C_\ell^2 |\tilde{X}_\ell|^2 \ . \end{aligned} \tag{18}$$

Here, the minimum of Q is found by taking derivatives of Q with respect to $Y_\ell(r)$ and setting them equal to zero. Unfortunately, Equation (18) cannot be computed directly because the true signal $\tilde{X}$ is unknown. The image formation model in Equation (14) is now employed again, along with an assumption that in the image, the signal $\tilde{X}$ is actually a projection (i.e., a slice in Fourier space) of the model V from the unknown true pose r^*:

$$\tilde{X}_\ell = Y_\ell(r^*)$$

and therefore

$$Q \geq \sum_{\|\ell\|>L} -\frac{1}{2} C_\ell^2 |Y_\ell(r^*)|^2 \ .$$

The true pose r^* is still unknown, but r^* must be one of the poses in the entire space of poses. Therefore, Q can again be bounded from below:

$$\begin{aligned} Q &\geq \sum_{\|\ell\|>L} -\frac{1}{2} C_\ell^2 |Y_\ell(r^*)|^2 \\ &\geq \min_r \sum_{\|\ell\|>L} -\frac{1}{2} C_\ell^2 |Y_\ell(r)|^2 \\ &= -\max_r \sum_{\|\ell\|>L} \frac{1}{2} C_\ell^2 |Y_\ell(r)|^2 \ , \end{aligned} \tag{19}$$

which is attained at

$$r = \hat{r}, \text{ with } \hat{Y}_\ell \equiv Y_\ell(\hat{r}) \ . \tag{20}$$

As a consequence, it follows that

$$Q \geq -\sum_{\|\ell\|>L} \frac{1}{2} C_\ell^2 |\hat{Y}_\ell|^2 \equiv \hat{Q} \tag{21}$$

Here, $\hat{Y}$ is the slice of model V that has the maximum CTF-modulated total power, as given by Equation (19). Finding this slice is simple, because it does not depend on the image X or shift t. Once $\hat{Y}$ (and the corresponding pose $\hat{r}$)

is identified, the bound $\hat{Q}$ on Q is fixed. Inserting Equation (21) into Equation (17) and inserting the result into Equation (13) finally yields a lower bound on $B(r,t)$:

$$B(r,t) \geq \sum_{\|\ell\|>L} \frac{1}{2}|X_\ell|^2 - \sum_{\|\ell\|>L} \frac{1}{2}C_\ell^2|\hat{Y}_\ell|^2 - H$$

Inserting the above bound on $B(r,t)$ into Equation (12) yields a lower bound on $E(r,t)$:

$$E(r,t) \geq \sum_{\|\ell\|\leq L} \frac{1}{2}|C_\ell Y_\ell(r) - S_\ell(t)X_\ell|^2 + \sum_{\|\ell\|>L} \frac{1}{2}|X_\ell|^2 - \sum_{\|\ell\|>L} \frac{1}{2}C_\ell^2|\hat{Y}_\ell|^2 - H$$

Due to the presence of H, the above expression is a *probabilistic* bound on E, giving the probability of E being greater than the value of the expression. In practice, a probability of 0.999936, corresponding to four standard deviations of H, provides a threshold that serves as an upper bound on H, and hence a deterministic lower bound for E above. That is,

$$\begin{aligned} H &\leq 4\sigma_H \qquad \text{(with probability 0.999936)} \\ &= 4\sqrt{\sum_{\|\ell\|>L} \frac{1}{2}C_\ell^2|Y_\ell(r)|^2} \\ &\leq 4\max_r \sqrt{\sum_{\|\ell\|>L} \frac{1}{2}C_\ell^2|Y_\ell(r)|^2} \\ &= 4\sqrt{\max_r \sum_{\|\ell\|>L} \frac{1}{2}C_\ell^2|Y_\ell(r)|^2} \end{aligned}$$

from which it follows that

$$H \leq 4\sqrt{\sum_{\|\ell\|>L} \frac{1}{2}C_\ell^2|\hat{Y}_\ell|^2}\ .$$

Here, the maximum power slice $\hat{Y}$ from Equation (19) is used. Incorporating the above expression finally yields a complete lower bound on $E(r,t)$:

$$\begin{aligned} E(r,t) \geq &\sum_{\|\ell\|\leq L} \frac{1}{2}|C_\ell Y_\ell(r) - S_\ell(t)X_\ell|^2 + \sum_{\|\ell\|>L} \frac{1}{2}|X_\ell|^2 \\ &- \sum_{\|\ell\|>L} \frac{1}{2}C_\ell^2|\hat{Y}_\ell|^2 - 4\sqrt{\sum_{\|\ell\|>L} \frac{1}{2}C_\ell^2|\hat{Y}_\ell|^2} \\ \equiv &\beta_L(r,t) \end{aligned} \tag{22}$$

Equation (22), with very high probability, bounds $E(r,t)$ from below. The bound is inexpensive to compute for a particular r,t (since only $A(r,t)$ depends on these). The remainder only needs to be computed once for all r,t, and also only once for all images that share the same CTF (i.e., that come from the same micrograph). To compute the bound, the slice of the model that has the most power ($\hat{Y}$) is first found. Then the expression for $\beta_L(r,t)$ is used to compute values of the lower bound.

The bound above is actually a family of bounds, one for each radius L. As L is increased and a greater number of Fourier coefficients are included in $A(r,t)$ rather than $B(r,t)$, the bound becomes more expensive but tighter. Finally, when L reaches the Nyquist rate then the bound becomes exact, but is as expensive as directly computing $E(r,t)$.

Subdivision scheme. The branch and bound algorithm relies on a method for representing regions in the space of 3-D poses and 2-D shifts so that when the bound above is computed, it can be used to discard regions and the remaining candidate poses are recorded in an organized fashion.

To accomplish this goal, this work uses a cartesian grid in the axis-angle representation of 3-D pose, and a second cartesian grid in the 2-D space of pixel shifts. These grids can be subdivided by a factor of two in each dimension, meaning that each subdivision increases the number of gridpoints in the pose space by a factor of eight, and a factor of four in the space of shifts.

In the first iteration of branch and bound image alignment, the pose and shift grids are initialized at a spacing of approximately 24 degrees and 5 pixels respectively. Seven iterations of branch and bound are used, each one subdividing the grids, yielding a final precision of 0.18 degrees and 0.04 pixels. The first iteration uses a radius of $L = 12$ Fourier coefficients. Each subsequent iteration uses double the radius in Fourier space, up to a maximum radius that is determined from the current resolution of the 3-D map.

Approximations. The lower bound derived above provides large speed improvements for the alignment of most images in a dataset. Some images, however, are pathological. Consider the case of an image that is an outlier, containing a non-particle or only noise. In this case the assumptions of the lower bound are violated as the image does not come from the standard cryo-EM image formation model. In this case the lower bound and branch-and-bound search process will still be valid, but will have nearly equally poor image alignment at many poses. Thus the bound will not be able to reject large portions of the pose space, as even the best pose will have high error.

In these cases, the branch and bound approach can be almost as slow as exhaustive search over all poses. To guard against this case, in this work an upper limit is set on the number of candidate poses (on the current discrete grid) that can remain after each iteration of branch and bound search. This limit is set to 12.5% of 3-D orientations, and 25% of 2-D image shifts. Empirically we find these limits do not have any negative effect on 3-D refinement resolution,

Nature Methods: doi:10.1038/nmeth.4169

as they only significantly affect non-particle images, which do not contribute useful signal.

The bound $\beta_L(r,t)$ depends on the CTF of the image that is being aligned. Therefore the components of the bound must be recomputed for each micrograph. Instead, in cryoSPARC we approximate the magnitude of the oscillating CTF at high resolutions (i.e., above L) using the root mean squared value of the CTF, which is a constant $\frac{1}{\sqrt{2}}$. This approximation does not generally make the bound tighter, but it removes the dependence on the CTF so that the last three terms of the bound only needs to be computed once for all images, given the structure V.

Finally, it is assumed that the lower bound is sufficiently smooth that it does not need to be sampled at full resolution, which would require a prohibitively large number of poses to be evaluated even with small values of L. Consequently, the subdivision scheme uses an empirically set initial sampling of poses. Formalizing the spacing of these poses based on the continuity of $\beta_L(r,t)$ remains a direction for future work.

References

[1] Sjors H W Scheres. A bayesian view on cryo-EM structure determination. *Journal of Molecular Biology*, 415(2):406–418, 2012.

[2] M.A. Brubaker, A. Punjani, and D.J. Fleet. Building proteins in a day: Efficient 3D molecular reconstruction. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June, 2015.

[3] León Bottou. Large-Scale Machine Learning with Stochastic Gradient Descent. *Proceedings of COMPSTAT'2010*, pages 177–186, 2010.

[4] Ilya Sutskever, James Martens, George E Dahl, and Geoffrey E Hinton. On the importance of initialization and momentum in deep learning. *Jmlr W&Cp*, 28(2010):1139–1147, 2013.

[5] A Dempster, N Laird, and D Rubin. *Maximum likelihood from incomplete data via the EM algorithm*, volume 39. 1977.