Supplementary Material for: Building Proteins in a Day: Efficient 3D Molecular Reconstruction

Marcus A. Brubaker, Ali Punjani and David J. Fleet
University of Toronto
{mbrubake,alipunjani,fleet}@cs.toronto.edu

1 Stochastic Optimization

This section provides algorithmic details of the Stochastic Averaged Gradient Descent (SAGD) optimization method used for MAP estimation. See the original SAGD paper [1] for details. Consider the objective function specified in Equation (6), rewritten as a sum of functions over subsets of the data:

$$\begin{aligned} f(\mathcal{V}) &= -\log p(\mathcal{V}) - \sum_{i=1}^{K} \log p(\tilde{\mathcal{I}}_i|\theta_i, \tilde{\mathcal{V}}) \\ &= \sum_{i=1}^{K} \left[-\frac{1}{K} \log p(\mathcal{V}) - \log p(\tilde{\mathcal{I}}_i|\theta_i, \tilde{\mathcal{V}}) \right] \\ &= \sum_{i=1}^{K} f_i(\mathcal{V}) \end{aligned}$$

At each iteration τ, SAGD computes the update given by

$$\mathcal{V}_{\tau+1} = \mathcal{V}_{\tau} - \frac{\epsilon}{L} \sum_{j=1}^{K} \left[d\mathcal{V}_j^{\tau} - \frac{1}{K} \frac{\partial}{\partial \mathcal{V}} \log p(\mathcal{V}) \right]$$

where $d\mathcal{V}_k^{\tau}$ is defined according to Equation (8). In practice, the sum in the above update equation is not computed at each iteration, but rather a running total is maintained and updated as follows:

$$\begin{aligned} \hat{\mathbf{g}}_{\tau} &= \sum_{k=1}^{K} d\mathcal{V}_k^{\tau} \\ \hat{\mathbf{g}}_{\tau+1} &= \hat{\mathbf{g}}_{\tau} - d\mathcal{V}_{k_\tau}^{\tau} + \mathbf{g}_{k_\tau}(\mathcal{V}_\tau) \end{aligned}$$

The SAGD algorithm requires a Lipschitz constant L which is not generally know. Instead it is estimated using a line search algorithm where an initial value of L is increased until the instantiated Lipschitz condition $f(\mathcal{V}) - f(\mathcal{V} - L^{-1} d\mathcal{V}) < \frac{\|d\mathcal{V}\|^2}{2L}$ is met. The line search for the Lipschitz constant L is only performed once every 20 iterations. Note that a more sophisticated line search could be performed if desired. A good initial value of L is found using a bisection search where the upper bound is the smallest L found so far to satisfy the condition and the lower bound is the largest L found so far which fails the condition. In between line searches, L is gradually decreased to try to take larger steps. The entire SAGD algorithm is provided in Algorithm (1).

Algorithm 1 SAGD

```
Initialize $\mathcal{V}$ and $L$
Initialize $\hat{\mathbf{g}} \leftarrow 0$
Initialize $d\mathcal{V}_k \leftarrow 0$ for all $k = 1..K$
for $\tau = 1..\tau_{\text{max}}$ do
    Select data subset $k_\tau$
    Compute objective gradient $\mathbf{g}_{k_\tau}(\mathcal{V})$
    $\hat{\mathbf{g}} \leftarrow \hat{\mathbf{g}} - d\mathcal{V}_{k_\tau} + \mathbf{g}_{k_\tau}(\mathcal{V})$
    $d\mathcal{V}_{k_\tau} \leftarrow \mathbf{g}_{k_\tau}(\mathcal{V})$
    $\mathcal{V} \leftarrow \mathcal{V} - \frac{\epsilon}{L}\left[\hat{\mathbf{g}} - \frac{\partial}{\partial \mathcal{V}} \log p(\mathcal{V})\right]$
    if mod($\tau$,20) == 0 then
        Perform line search
        while $f_{k_\tau}(\mathcal{V}) - f_{k_\tau}(\mathcal{V} - L^{-1} d\mathcal{V}_{k_\tau}) < \frac{\|d\mathcal{V}_{k_\tau}\|^2}{2L}$ do
            $L \leftarrow 2L$
        end while
    else
        $L \leftarrow \frac{K}{2^{\frac{1}{150}}}$
    end if
end for
```

2 Importance Sampling

Importance Sampling is a key part of the proposed reconstruction method for Cryo-EM and provides large speedups during optimization. We use importance sampling to efficiently compute the discrete sum in Equation (4). Note that importance sampling is applied independently for each image in the dataset, since the orientations and shifts which correspond to important terms in the discrete sum can be different for each image.

In practice, we split the outer sum in Equation (4) into a double summation, one over orientations on the sphere and one over in-plane rotations of images and projections. We then compute each of the three sums (over shift, in-plane rotation, and orientation) with and independent importance sampler. This is equivalent to computing the full sum in Equation (4) using a single importance sampler with an importance distribution that is factored into three parts, one for each of shift, in-plane rotation, and orientation. This factoring is necessary, as the memory requirements of storing a fully joint importance distribution for each image in the dataset would become infeasible for high-resolution reconstructions.

For each of the three importance samplers, the importance distribution at each iteration is constructed according to Equation (14). At the first iteration during which a particular image is seen, the importance distribution is simply uniform, and in fact we explicitly sample every point once. The ϕ values resulting from this computation are stored. At the next iteration during which the same image is seen, these ϕ values are used in Equation (14) to construct a non-uniform importance distribution which is then sampled from. We use a number of samples proportional to the effective sample size of the importance distribution, so the number of samples used naturally decreases as the importance distribution becomes more peaked, leading to large speedups at late iterations during optimization.

Algorithm 2 Importance Sampling

```
Given $\phi_i$ for $i \in \mathfrak{I}$ from previous iteration
for $j \in 1..J$ do
    for $i \in \mathfrak{I}$ do
        Compute $\mathbf{K}_{i,j}$
    end for
end for
$\hat{\phi}_j \leftarrow \sum_{i \in \mathfrak{I}} \phi_i^{1/T} \mathbf{K}_{i,j} \quad \forall j \in 1..J$
$Z \leftarrow \sum_j \hat{\phi}_j$
$q_j \leftarrow (1-\alpha) Z^{-1} \hat{\phi}_j + \alpha \psi_j \quad \forall j \in 1..J$
$s \leftarrow \left( \sum_j q_j^2 \right)^{-1}$
$N \leftarrow s_0 s$
$\mathfrak{I} \leftarrow \varnothing$
for $k \in 1..N$ do
    $i \leftarrow$ sample from $q$
    insert $i$ into $\mathfrak{I}$
end for
Use $\mathfrak{I}$ to compute $\phi_i$ for next iteration
```

In Equation (14), the previous ϕ values are not directly used, but rather they are annealed by a temperature parameter and then smoothed by a kernel matrix. Both of these steps serve to guard against importance distributions which are too peaked around large ϕ values, which would inhibit the importance sampler from exploring the domain. The kernel matrix also serves the purpose of allowing use of ϕ values from a previous iteration even if the resolution of quadrature points being used has increased at the current iteration. The Von Mises-Fisher kernel is used for orientations and in-plane rotations, while a Gaussian kernel is used for shift:

$$\begin{aligned}\mathbf{K}_V(d_i, d_j; \kappa_V) &\propto \exp(\kappa_V d_i^T d_j) \\ \mathbf{K}_G(t_i, t_j; \kappa_G) &\propto \exp(-\kappa_G \|\mathbf{t}_i - \mathbf{t}_j\|^2)\end{aligned}$$

where κ_V and κ_G are precision parameters for each kernel which are set based on the resolution of the quadrature scheme used at the previous ϕ values, d_i and d_j are the quadrature directions (in $\mathcal{S}^2$ for particle orientation and $\mathcal{S}^1$ for in-plane rotation, and $\mathbf{t}_i$ and $\mathbf{t}_j$ are the quadrature shift values (in $\mathbb{R}^2$).

The algorithm for constructing an importance distribution and sampling from it are given in Algorithm (2). The sampled values are then used to compute (12). Note that some quadrature points can end up being sampled multiple times, this is detected and the value reused to reduce computation.

3 Error Measures

Because ground-truth is rarely available for Cryo-EM, measuring accuracy is often difficult. Traditionally, the field has used the *Fourier Shell Correlation* (FSC) to measure the resolution of a solved structure. The so-called gold-standard FSC works by splitting the dataset in half, estimating two densities separately and the computing the normalized correlation in Fourier space as a function of frequency. This curve would then be thresholded to provide an estimate of accuracy. However, we note that this measure is actually estimating the variance of the estimator, not the accuracy of the density it has produced. Further it is only theoretically justifiable when the estimator is unbiased, which is not true of the method proposed here or with other likelihood-based Bayesian methods such as RELION.

Instead, we introduce a novel metric based on reconstruction error of a held test set. To quantify the ability of marginal likelihood methods, such as ours, to model and explain the observed data we introduce the *Expected Mean Squared Error*

$$\mathcal{E}^2(\mathcal{I}|\theta, \mathcal{V}) \equiv E_{\mathbf{R},\mathbf{t}|\mathcal{I},\theta,\mathcal{V}} \left[\|\mathcal{I} - \mathbf{C}_\theta \mathbf{S}_\mathbf{t} \mathbf{P}_\mathbf{R} \mathcal{V}\|^2 \right] \tag{1}$$

to be the expectation of the squared error between the image and its reconstruction under the image formation model. Note that the expectation is conditioned on the current density and the CTF parameters and is taken over the unknown pose and translation, $\mathbf{R}$ and $\mathbf{t}$. After switching to Fourier space and with some manipulation this formula becomes

$$\mathcal{E}^2(\mathcal{I}|\theta, \mathcal{V}) = Z^{-1} \int_{\mathbb{R}^2} \int_{\mathcal{SO}(3)} \|\tilde{\mathcal{I}} - \tilde{\mathbf{C}}_\theta \tilde{\mathbf{S}}_\mathbf{t} \tilde{\mathbf{P}}_\mathbf{R} \tilde{\mathcal{V}}\|^2 p(\tilde{\mathcal{I}}|\theta, \mathbf{R}, \mathbf{t}, \tilde{\mathcal{V}}) p(\mathbf{R}) p(\mathbf{t}) d\mathbf{R} d\mathbf{t} \tag{2}$$

where the

$$\begin{aligned} Z &= p(\tilde{\mathcal{I}}|\theta, \tilde{\mathcal{V}}) && (3) \\ &= \int_{\mathbb{R}^2} \int_{\mathcal{SO}(3)} p(\tilde{\mathcal{I}}|\theta, \mathbf{R}, \mathbf{t}, \tilde{\mathcal{V}}) p(\mathbf{R}) p(\mathbf{t}) d\mathbf{R} d\mathbf{t} && (4) \end{aligned}$$

is a normalization constant. To compute this efficiently, we can use the same importance sampling technique described in the main paper to approximate it as

$$\hat{\mathcal{E}}^2(\mathcal{I}|\theta, \mathcal{V}) = \hat{Z}^{-1} \sum_{j \in \mathfrak{J}^{\mathbf{R}}} \frac{w_j^{\mathbf{R}}}{N_{\mathbf{R}} q_j^{\mathbf{R}}} \left(\sum_{\ell \in \mathfrak{J}^{\mathbf{t}}} \frac{w_\ell^{\mathbf{t}}}{N_{\mathbf{t}} q_\ell^{\mathbf{t}}} p_{j,\ell} \|\tilde{\mathcal{I}} - \tilde{\mathbf{C}}_\theta \tilde{\mathbf{S}}_{\mathbf{t}} \tilde{\mathbf{P}}_{\mathbf{R}} \tilde{\mathcal{V}}\|^2 \right) \tag{5}$$

where

$$\hat{Z} = \sum_{j \in \mathfrak{J}^{\mathbf{R}}} \frac{w_j^{\mathbf{R}}}{N_{\mathbf{R}} q_j^{\mathbf{R}}} \left(\sum_{\ell \in \mathfrak{J}^{\mathbf{t}}} \frac{w_\ell^{\mathbf{t}}}{N_{\mathbf{t}} q_\ell^{\mathbf{t}}} p_{j,\ell} \right) \tag{6}$$

is the approximation of the normalization constant. The above quantities can be readily computed along with the main likelihood computation using the same importance sampling scheme described above.

We compute the average value of $\hat{\mathcal{E}}^2(\mathcal{I}|\theta, \mathcal{V})$ on a held out set of test images whose gradients are never used. To normalize for different datasets we report the *Relative Root Expected Mean Squared Error* (RREMSE) as

$$\sqrt{\frac{1}{\sigma^2 N_{\text{test}}} \sum_{\mathcal{I}} \hat{\mathcal{E}}^2(\mathcal{I}|\theta, \mathcal{V})} \tag{7}$$

where the sum is taken over the test set which has N_{test} images and σ^2 is the noise variance of the dataset. Values near 1 indicate that the data is being well explained.

References

[1] N. Le Roux, M. Schmidt, and F. Bach, "A stochastic gradient method with an exponential convergence rate for strongly convex optimization with finite training sets," in *NIPS*, 2012. 1