1 Derivation of Consistent Majority Optimization

In this document we provide a full derivation of the EM algorithm presented in Section 3 of the main paper.

We represent uncertainty and linearize using conventions from Lie group theory; many researchers in the fields of computer vision, physics, and manifold optimization agree that this is the most natural and efficient method. We provide a very brief introduction to necessary Lie group concepts as needed in this section. Many complete introductions may be found in the literature, such as in the book by Murray, Li, and Sastry [1].

1.1 Model

In this section, let $\mathbb{P}$ be a Lie group of camera poses. In the case of inferring only rotations, $\mathbb{P} = \mathbb{SO}(3)$, and for the case of inferring full 6-DOF poses, $\mathbb{P} = \mathbb{SE}(3)$. $\dim \mathbb{P}$ is the tangent-space dimensionality (the same as the number of degrees-of-freedom) of the pose, i.e. $\dim \mathbb{SO}(3) = 3$ and $\dim \mathbb{SE}(3) = 6$. In each case, $\mathfrak{p}$ is the Lie algebra of $\mathbb{P}$, either $\mathfrak{so}(3)$ or $\mathfrak{se}(3)$. The Lie algebra is the tangent space about the identity element of a Lie group, and thus is the set of partial derivatives of the identity element with respect to the degrees of freedom. For rotations in $\mathbb{SO}(3)$, for instance, $\mathfrak{se}(3)$ is the 3D vector space of skew-symmetric 3×3 matrices.

The generative model of the i^{th} relative pose measurement z_i between a pair of cameras with poses x_j and x_k is

$$z_i = x_j^{-1} x_k \exp \hat{\epsilon}_i, \tag{1}$$

where $\epsilon_i \in \mathbb{R}^{\dim \mathbb{P}}$ is Gaussian noise drawn either from the inlier or outlier distribution:

$$\epsilon_i \sim \begin{cases} \mathcal{N}(\mathbf{0}, C_1), & y_i = 1 \\ \mathcal{N}(\mathbf{0}, C_0), & y_i = 0, \end{cases} \tag{2}$$

in which y_i is a latent indicator variable denoting whether the measurement z_i is correct (y_i=1) or erroneous (y_i=0). C_1 is the inlier covariance, which comes directly from the pairwise reconstructions. C_0 is the outlier covariance, which is chosen to be large.

Together, the hat operator and the exponential map in (1) convert a *vector* increment on the rotation and translation axes (in this case the random variable ϵ_{y_i}) into a relative *pose* increment that may be composed with another pose, i.e. $\exp \hat{\cdot} : \mathbb{R}^{\dim \mathbb{P}} \to \mathbb{P}$. Individually, the "hat" operator $\hat{\cdot} : \mathbb{R}^{\dim \mathbb{P}} \to \mathfrak{p}$ transforms coordinates of degrees-of-freedom (such as rotation axis-angles and translation directions) to a direction in the Lie algebra (e.g. for rotations the skew-symmetric matrices). $\exp : \mathfrak{p} \to \mathbb{P}$ is the matrix exponential map, which maps from an increment along a direction and distance in the Lie algebra back to the Lie group.

Later, we will make use of the logarithm map $\log : \mathbb{P} \to \mathfrak{p}$, which is the inverse of the exponential map. $\log^\vee : \mathbb{P} \to \mathbb{R}^{\dim \mathbb{P}}$ is the inverse of the combined exponential map and hat operator $\exp \hat{\cdot}$, mapping a Lie group element to a set of coordinates in the Lie algebra representing the corresponding increment from the identity element.

1.2 Inference

We perform inference on this model using EM. The M step amounts to finding the maximum likelihood solution for the poses x given an estimate of the expected values of the indicator variables y. The E step then estimates the expected value of each y_i, i.e., the probabilities of each of the edges being an inlier.

The key to efficient inference is that each latent variable y_i is independent of any other y_i when conditioned on the camera poses. This avoids a combinatorial search over all of y.

Thus, in EM, we optimize the expectation, with respect to the indicator variables, of the log-likelihood of the parameters. Inside each iteration, we evaluate this expectation with respect to the parameters *from the previous iteration*:

$$\begin{aligned} x^t &= \arg\max_x \left\langle \log p\left(x \,|\, z, y\right)\right\rangle_{y \,|\, x^{t-1}, z} \\ &= \arg\max_x \sum_y \left(\left(\log p\left(x \,|\, z, y\right)\right) \prod_i p\left(y_i \,|\, x_j^{t-1}, x_k^{t-1}, z_i\right) \right), \end{aligned} \tag{3}$$

where $\langle\cdot\rangle$ is the expectation with respect to the subscripted variables, $\sum_y$ is a sum over all possible combinations of y for each measurement i (note, however, that this soon simplifies into a much more tractable sum), and $\mathcal{L}\left(\cdot\right)$ is a log-likelihood. To write the parameter likelihood in terms of known densities we factor it using Bayes' law:

$$\begin{aligned} p\left(x|z, y\right) &\propto p\left(z \,|\, y, x\right) p\left(x\right) \\ &\propto \left(\prod_i p\left(z_i \,|\, y_i, x_j, x_k\right) \right) \prod_j p\left(x_j\right), \end{aligned} \tag{4}$$

where we have dropped the denominator since it is not a function of the optimization variable x, and assumed independent priors over the camera poses $p\left(x_j\right)$ that do not depend on any other variables. In our experiments we assumed uninformative pose priors, in which case the prior terms may simply be omitted.

Next, substituting $p\left(x \,|\, z, y\right)$ from (4) into (3) and simplifying, we obtain

$$\begin{aligned} \left\langle \mathcal{L}\left(x \,|\, z, y\right)\right\rangle_{y \,|\, x^{t-1}, z} &= \sum_y \left(\sum_i \mathcal{L}\left(z_i \,|\, y_i, x_j, x_k\right) + \sum_j \mathcal{L}\left(x_j\right) \right) \prod_i p\left(y_i \,|\, x_j^{t-1}, x_k^{t-1}, z_i\right) \\ &= \sum_y \left(\sum_i \mathcal{L}\left(z_i \,|\, y_i, x_j, x_k\right) \right) \prod_i p\left(y_i \,|\, x_j^{t-1}, x_k^{t-1}, z_i\right) + \sum_j \mathcal{L}\left(x_j\right) \\ &= \sum_i \sum_{y_i} \mathcal{L}\left(z_i \,|\, y_i, x_j, x_k\right) p\left(y_i \,|\, x_j^{t-1}, x_k^{t-1}, z_i\right) + \sum_j \mathcal{L}\left(x_j\right), \end{aligned} \tag{5}$$

where $\mathcal{L}\left(\cdot\right) \triangleq \log p\left(\cdot\right)$ is the log-likelihood, and $\sum_{y_i}$ is a sum over the two possible values of y_i, 0 and 1. Note that in the second step, we are able to pull $\sum_j \mathcal{L}\left(x_j\right)$ outside of the sum over y because this quantity does not depend on y and $\sum_y p\left(y \,|\, x, z\right) = 1$.

The M Step In the M step, we maximize the expected log-likelihood in Eq. 5 with respect to the camera poses x. Note that this would be considerably more difficult if the expectation were not with respect to the poses from the previous iteration. Since this is constant in this maximization, we let the shorthand $\lambda_i \triangleq \left\langle y_i \right\rangle_{y_i \,|\, x^{t-1}, z_i} = p\left(y_i{=}1 \,|\, x_j^{t-1}, x_k^{t-1}, z_i\right)$. Then expanding the form for the Gaussian, we obtain

$$\begin{aligned} x^t &= \arg\max_x C + \mathcal{L}\left(x\right) + \sum_i \left(\lambda_i^t \left(\frac{-1}{2} \left\| \log^\vee\left(z_i^{-1} x_j^{-1} x_k\right) \right\|_{C_1}^2 \right) \right. \\ &\qquad \left. + \left(1 - \lambda_i^t\right) \left(\frac{-1}{2} \left\| \log^\vee\left(z_i^{-1} x_j^{-1} x_k\right) \right\|_{C_0}^2 \right) \right) \\ &= \arg\max_x \mathcal{L}\left(x\right) + \frac{-1}{2} \sum_i \left\| \log^\vee\left(z_i^{-1} x_j^{-1} x_k\right) \right\|_{(\lambda_i C_1 + (1-\lambda_i) C_0)}^2, \end{aligned} \tag{6}$$

where C is the constant from the Gaussian normalization terms. Note that the mixture of Gaussians simplifies to a weighted average of the inlier and outlier covariance matrices, with responsibility λ_i^t. In our implementation, we solve the nonlinear least-squares problem in (6) to compute the updated poses x^t using the Levenberg-Marquardt algorithm.

The E Step In the E step, we determine the sufficient statistics of $p(y|x,z)$. As seen above we need only the probability that y_i=1, i.e. the expectation

$$\begin{aligned}
\lambda_i^t &= \langle y_i \rangle_{y_i \,|\, x^{t-1}, z_i} \\
&= \sum_{y_i} p(y_i|x,z_i)\, y_i \\
&= p\left(y_i{=}1 \,|\, x_j^{t-1}, x_k^{t-1}, z_i\right) \\
&= \frac{p\left(z_i \,|\, x_j^{t-1}, x_k^{t-1}, y_i{=}1\right) p(y_i{=}1)}{\sum_{y_i} p\left(z_i \,|\, x_j^{t-1}, x_k^{t-1}, y_i\right) p(y_i)} \\
&= \frac{\mathcal{N}\left(\log^\vee\left(z_i^{-1} x_j^{-1} x_k\right); \mathbf{0}, C_1\right) p(y_i{=}1)}{\sum_{y_i} \mathcal{N}\left(\log^\vee\left(z_i^{-1} x_j^{-1} x_k\right); \mathbf{0}, C_{y_i}\right) p(y_i)},
\end{aligned} \tag{7}$$

where Σ_i is either the inlier or outlier covariance, according to y_i. The notation $\mathcal{N}(\xi; \mu, \Sigma)$ is the evaluation of a Gaussian PDF at ξ. The prior $p(y_i)$ may be estimated online, e.g. as a Gaussian mixture prior, although in our experiments we use an uninformative prior.

References

[1] R.M. Murray, Z. Li, S. Sastry, and S.S. Sastry. *A mathematical introduction to robotic manipulation*. CRC Press, 1994.