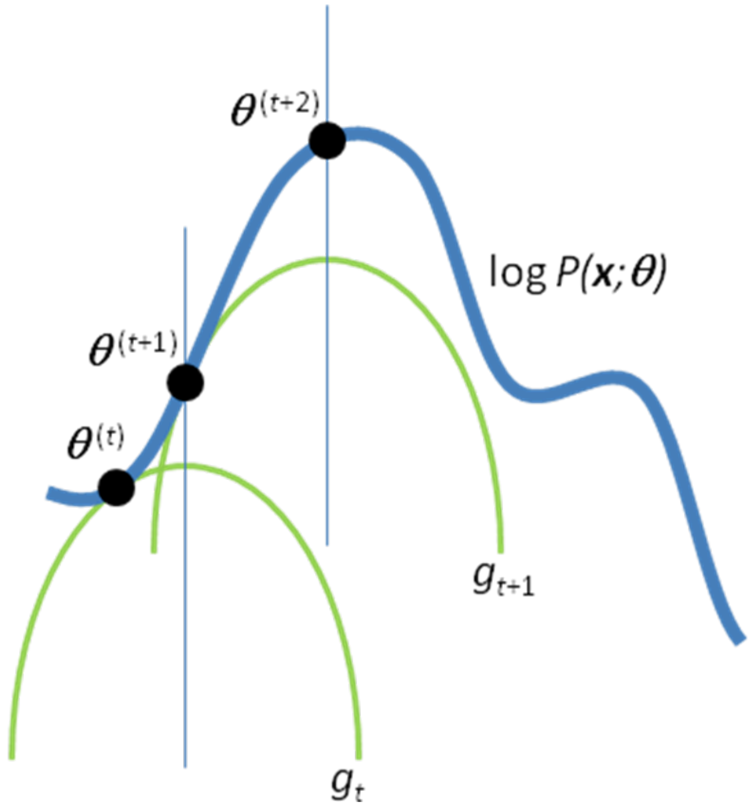

Supplementary Figure 1 Convergence of the EM algorithm. Starting from initial parameters $\boldsymbol{\theta}^{(t)}$, the E-step of the EM algorithm constructs a function g_t that lower-bounds the objective function $\log P(\boldsymbol{x};\boldsymbol{\theta})$. In the M-step, $\boldsymbol{\theta}^{(t+1)}$ is computed as the maximum of g_t. In the next E-step, a new lower-bound g_{t+1} is constructed; maximization of g_{t+1} in the next M-step gives $\boldsymbol{\theta}^{(t+2)}$, etc.

Supplementary Note 1

Here, we provide a short derivation of the EM algorithm based on the idea of bound maximization. A more general (though somewhat subtle) argument, which leads to a number of variants of the EM algorithm, was described in:

> Neal RM and Hinton GE. 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants. In MI Jordan, ed., *Learning in Graphical Models*, Kluwer Academic Publishers, 355-358.

Mathematically, the EM algorithm derives from the fact that for any probability distribution $Q(\mathbf{z})$,

$$\log\left(\sum_{\mathbf{z}} P(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\right) = \log\left(\sum_{\mathbf{z}} Q(\mathbf{z}) \cdot \frac{P(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{Q(\mathbf{z})}\right) \geq \sum_{\mathbf{z}} Q(\mathbf{z}) \log\left(\frac{P(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{Q(\mathbf{z})}\right). \tag{3}$$

where the inequality is tight (i.e. holds with equality) whenever $Q(\mathbf{z}) = P(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})$. The derivation of (3) relies on a tool from mathematical analysis known as **Jensen's inequality** for concave functions.[1]

Now, consider the update rule, $\hat{\boldsymbol{\theta}}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} g_t(\boldsymbol{\theta})$, where

$$g_t(\boldsymbol{\theta}) = \sum_{\mathbf{z}} P\left(\mathbf{z}\middle|\mathbf{x};\hat{\boldsymbol{\theta}}^{(t)}\right) \log\left(\frac{P(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{P\left(\mathbf{z}\middle|\mathbf{x};\hat{\boldsymbol{\theta}}^{(t)}\right)}\right). \tag{4}$$

Applying the tightness conditions of equation (3), $g_t\left(\hat{\boldsymbol{\theta}}^{(t)}\right) = \log P\left(\mathbf{x};\hat{\boldsymbol{\theta}}^{(t)}\right)$. However, observe that $g_t\left(\hat{\boldsymbol{\theta}}^{(t)}\right) \leq g_t\left(\hat{\boldsymbol{\theta}}^{(t+1)}\right)$ by definition of the update rule. Furthermore, $g_t\left(\hat{\boldsymbol{\theta}}^{(t+1)}\right) \leq \log P\left(\mathbf{x};\hat{\boldsymbol{\theta}}^{(t+1)}\right)$ since (3) guarantees that $g_t(\boldsymbol{\theta})$ is a lower bound on $\log P(\mathbf{x};\boldsymbol{\theta})$ for any parameter vector $\boldsymbol{\theta}$. Therefore, the update rule results in monotonic improvement of the maximum likelihood objective for incomplete data.

To see the connection between (4) and the description of the EM algorithm given in the text, consider the following equivalent update rule:

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \sum_{\mathbf{z}} P\left(\mathbf{z}\middle|\mathbf{x};\hat{\boldsymbol{\theta}}^{(t)}\right) \log P(\mathbf{x},\mathbf{z};\boldsymbol{\theta}). \tag{5}$$

Here, equivalence follows from the fact that the objective function of (5) differs from $g_t(\boldsymbol{\theta})$ by a constant offset which does not depend on $\boldsymbol{\theta}$. In this final form, we see that the EM update rule effectively maximizes the log-likelihood of a dataset expanded to contain all possible completions $\mathbf{z}$

---

[1] In brief, Jensen's inequality states that a concave function (e.g., $\log(\cdot)$) of the expectation of a random variable (e.g., y) is never less than expected value of that concave function applied to the random variable (i.e., $\log(E[y]) \geq E[\log(y)]$); furthermore, equality holds if and only if the random variable is constant with probability 1 (i.e., $y = E[y]$). Equation (3) follows from Jensen's inequality, using the random variable $y = \dfrac{P(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{Q(\mathbf{z})}$.

of the unobserved variables, where each completion is weighted by the posterior probability, $P(\mathbf{z}|\mathbf{x};\hat{\boldsymbol{\theta}}^{(t)})$.