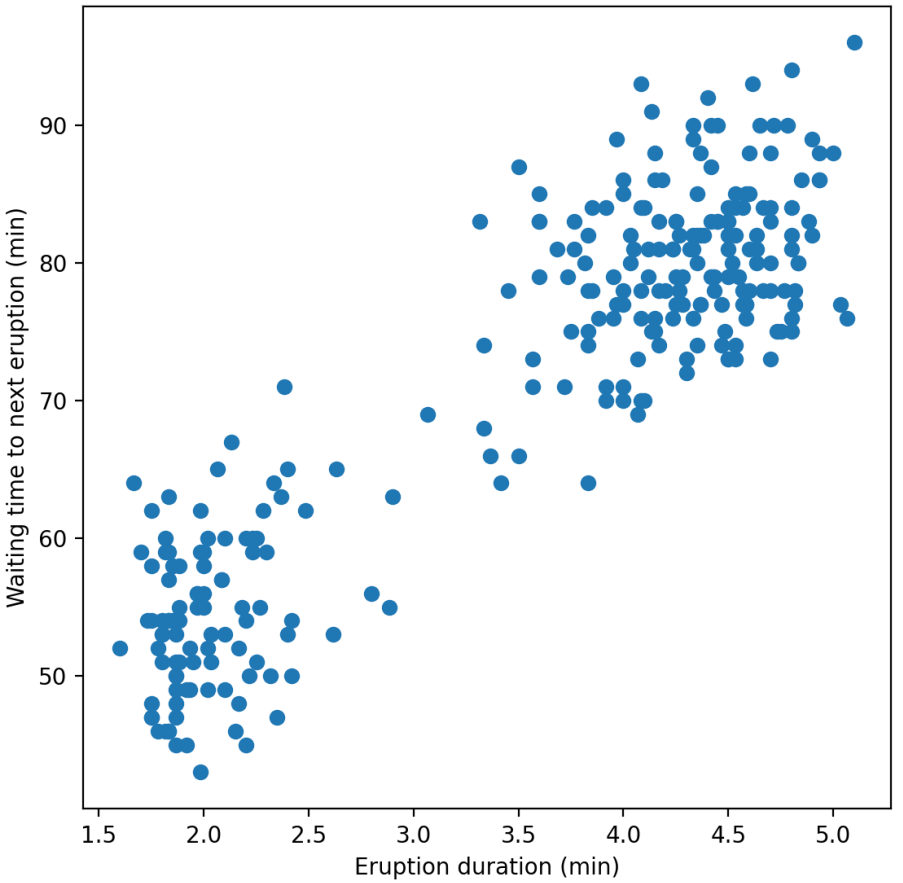
Foundations of Data Science

Lecture 7: Expectation-Maximization

Prof. Gilles Louppe
g.louppe@uliege.be

LIÈGE université

The Old Faithful geyser in Yellowstone National Park (USA) is famous for its frequent and predictable eruptions of hot water and steam.

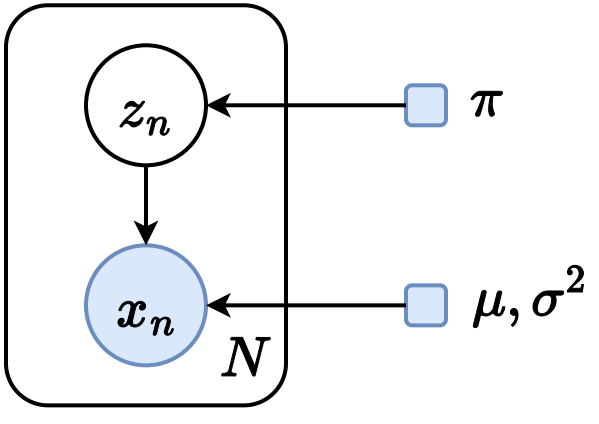

Waiting time to next eruption (min)
90
80
70
60
50
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Eruption duration (min)

2-component GMM

The observed data $\{x_n \in \mathbb{R}^2\}_{n=1}^N$ can be modeled as being generated from a mixture of two Gaussian distributions, with latents $z_n \in \{1, 2\}$ labeling observation membership and hyper-parameters $\pi, \mu = (\mu_1, \mu_2), \sigma^2 = (\sigma_1^2, \sigma_2^2)$ defining the mixture proportions, means and variances of the Gaussian components.

$$z_n \sim \text{Categorical}(\pi)$$

$$x_n \sim \mathcal{N}\left(x_{z_n}, \sigma_{z_n}^2\right)$$



How to fit the model parameters?

The marginal log-likelihood of the observed data is

$$\begin{aligned}\log p(\{x_n\}_{n=1}^N|\pi,\mu,\sigma^2) &= \log\prod_{n=1}^N p(x_n|\pi,\mu,\sigma^2)\\ &= \log\prod_{n=1}^N\sum_{z_n=1}^2 p(x_n,z_n|\pi,\mu,\sigma^2)\\ &= \sum_{n=1}^N\log\sum_{z_n=1}^2 p(x_n|z_n,\mu,\sigma^2)p(z_n|\pi),\end{aligned}$$

where

- $p(x_n|z_n=k,\mu,\sigma^2)=\mathcal{N}(x_n|\mu_k,\sigma_k^2 I)$
- $p(z_n=k|\pi)=\pi_k$ for $k=1,2$.

Direct maximization of the marginal log-likelihood w.r.t. parameters $\theta = (\pi, \mu, \sigma^2)$ is difficult.

- The objective is non-convex.
- No closed-form solution.
- Numerically unstable.

More generally, latent variable models lead to log-likelihoods involving sums or integrals inside the logarithm. As their domain grows, these sums/integrals **become intractable to even just evaluate**, let alone optimize.

If we knew the latent variables...

If the latent $\{z_n\}_{n=1}^N$ were known, the complete-data log-likelihood would be

$$\begin{aligned}\log p(\{x_n, z_n\}_{n=1}^N | \pi, \mu, \sigma^2) &= \log \prod_{n=1}^N p(x_n, z_n | \pi, \mu, \sigma^2) \\ &= \log \prod_{n=1}^N p(x_n | \mu_{z_n}, \sigma_{z_n}^2) p(z_n | \pi) \\ &= \sum_{n=1}^N \log p(x_n | \mu_{z_n}, \sigma_{z_n}^2) + \sum_{n=1}^N \log p(z_n | \pi).\end{aligned}$$

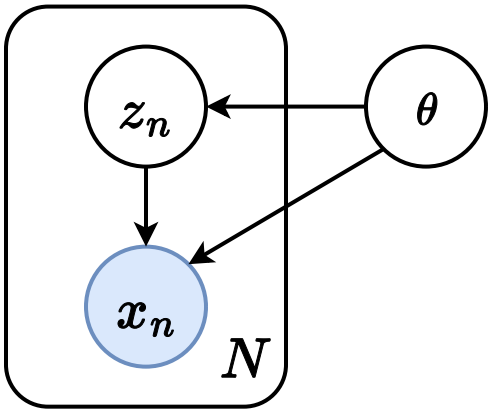
The evaluation of the log-likelihood becomes tractable, and so does its optimization.

For the Gaussian mixture model, differentiating the complete-data log-likelihood with respect to the parameters and setting to zero would even yield closed-form expressions for the maximum likelihood estimates:

- $\mu_k = \frac{1}{N_k} \sum_{\{n:z_n=k\}} x_n$
- $\sigma_k^2 = \frac{1}{N_k} \sum_{\{n:z_n=k\}} |x_n - \mu_k|^2$
- $\pi_k = \frac{N_k}{N}$ where $N_k = \sum_{n=1}^{N} 1(z_n = k)$.

What if we alternate between guessing the latent variables and optimizing the hyperparameters?

Expectation-Maximization



Assume a generic latent variable model $p(x, z|\theta)$ where θ are parameters mediating the joint distribution.

Evidence lower bound

The marginal log-likelihood for a single observation x can be rewritten as

$$\begin{aligned}\log p(x|\theta) &= \log \int p(x, z|\theta)\, dz \\ &= \log \int q(z) \frac{p(x, z|\theta)}{q(z)}\, dz \\ &= \log \mathbb{E}_{q(z)} \left[\frac{p(x, z|\theta)}{q(z)} \right],\end{aligned}$$

where $q(z)$ is any valid probability distribution over the latent variable z.

By Jensen's inequality, the log-likelihood can be lower-bounded as

$$\log p(x|\theta) \geq \mathbb{E}_{q(z)}\left[\log \frac{p(x,z|\theta)}{q(z)}\right] = \mathcal{L}(q,\theta),$$

where $\mathcal{L}(q,\theta)$ is known as the **evidence lower bound objective** (ELBO).

The ELBO can first be rewritten as

$$\begin{aligned}\mathcal{L}(q,\theta) &= \mathbb{E}_{q(z)}\left[\log \frac{p(x,z|\theta)}{q(z)}\right] \\ &= \mathbb{E}_{q(z)}\left[\log \frac{p(x|z,\theta)p(z|\theta)}{q(z)}\right] \\ &= \mathbb{E}_{q(z)}\left[\log p(x|z,\theta)\right] - \mathbf{KL}(q(z)||p(z|\theta)),\end{aligned}$$

where $\mathbf{KL}(q(z)||p(z|\theta)) = \mathbb{E}_{q(z)}\left[\log \frac{q(z)}{p(z|\theta)}\right]$ is the Kullback-Leibler divergence between distributions $q(z)$ and $p(z|\theta)$.

This expression highlights the trade-off between fitting the data well (first term) and keeping the variational distribution $q(z)$ close to the prior $p(z|\theta)$ (second term).

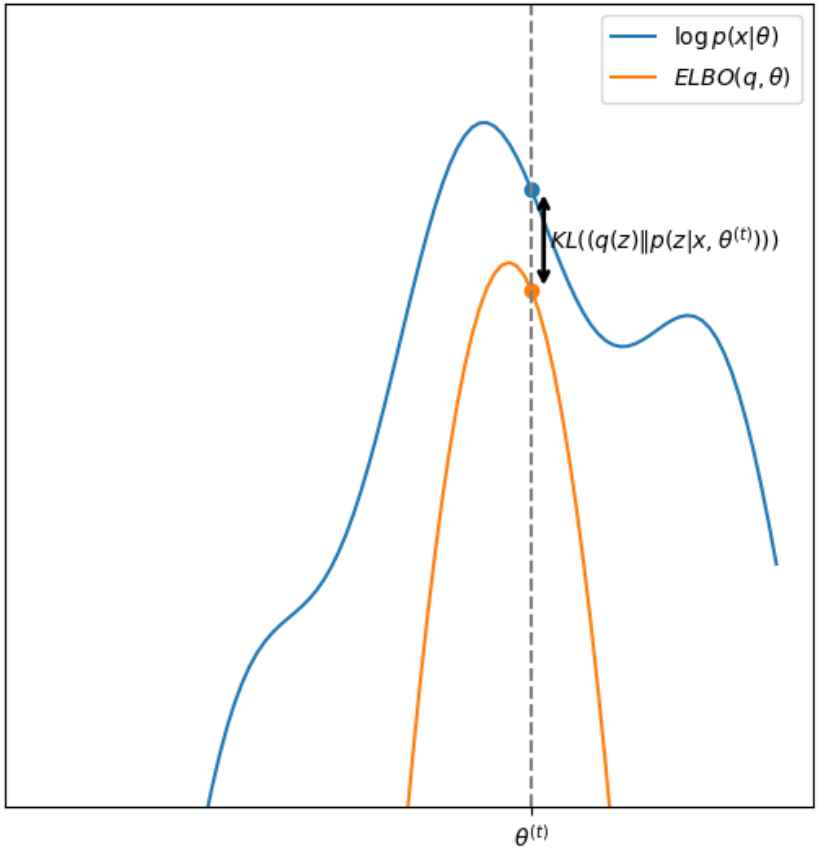
By factorizing the joint in the other way, the ELBO can also be rewritten as

$$\begin{aligned}\mathcal{L}(q,\theta) &= \mathbb{E}_{q(z)}\left[\log \frac{p(x,z|\theta)}{q(z)}\right] \\ &= \mathbb{E}_{q(z)}\left[\log \frac{p(z|x,\theta)p(x|\theta)}{q(z)}\right] \\ &= \log p(x|\theta) - \mathrm{KL}(q(z)||p(z|x,\theta)).\end{aligned}$$

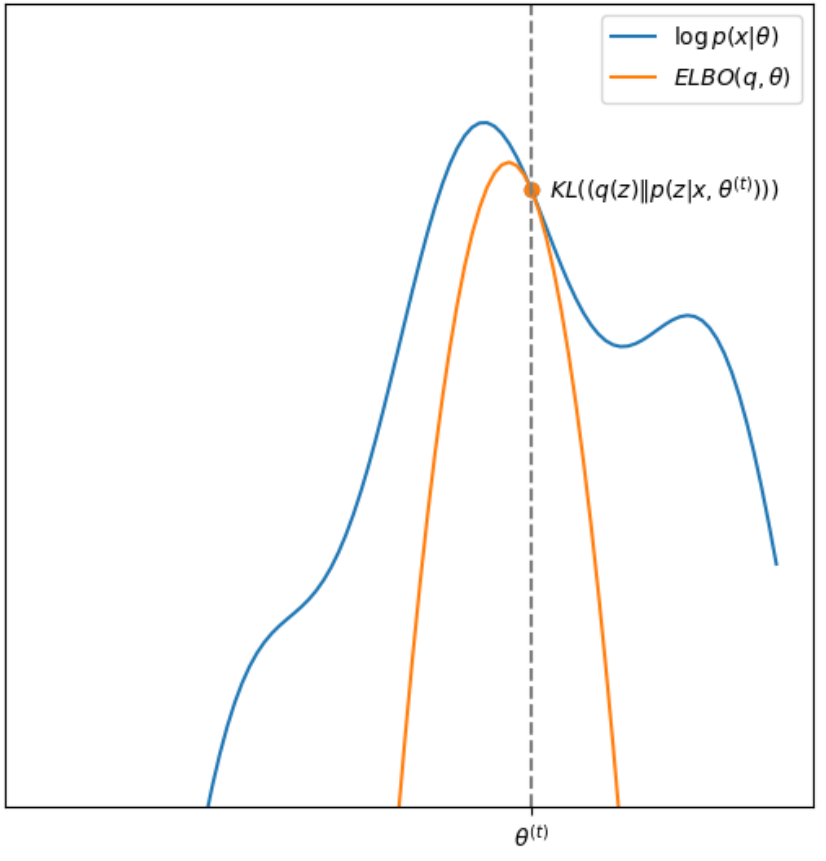
Therefore,

$$\log p(x|\theta) = \mathcal{L}(q,\theta) + \mathrm{KL}(q(z)||p(z|x,\theta)).$$

This decomposition reveals that the ELBO is a lower bound on the log-likelihood, with a gap measured by the KL divergence between the variational distribution $q(z)$ and the posterior distribution $p(z|x,\theta)$.



The KL gap $\mathbf{KL}(q(z)||p(z|x,\theta))$ between the log-likelihood $\log p(x|\theta)$ and the ELBO $\mathcal{L}(q,\theta)$ measures the tightness of the bound. (Sketch.)



The ELBO is tight when $q(z) = p(z|x, \theta)$. Maximizing the ELBO is then equivalent to maximizing the log-likelihood $\log p(x|\theta)$.

Expectation-Maximization algorithm

The EM algorithm maintains parameter estimates $\theta^{(t)}$ at iteration t and a variational distribution $q^{(t)}(z)$ over the latent variables. It iteratively maximizes the ELBO by alternating between two steps:

- E-step: maximize the ELBO $\mathcal{L}(q, \theta^{(t)})$ w.r.t. q while keeping $\theta^{(t)}$ fixed.
- M-step: maximize the ELBO $\mathcal{L}(q^{(t+1)}, \theta)$ w.r.t. θ while keeping $q^{(t+1)}(z)$ fixed.

The E-step consists in solving

$$q^{(t+1)} = \arg\max_q \mathcal{L}(q, \theta^{(t)})$$

which is achieved by setting q to the posterior distribution

$$q^{(t+1)}(z) = p(z|x, \theta^{(t)})$$

.

Proof. From the decomposition of the log-likelihood, we have

$$\mathcal{L}(q, \theta^{(t)}) = \log p(x|\theta^{(t)}) - \mathrm{KL}(q(z)||p(z|x, \theta^{(t)})).$$

Since the log-likelihood term $\log p(x|\theta^{(t)})$ does not depend on q, maximizing the ELBO w.r.t. q is equivalent to minimizing the KL divergence $\mathrm{KL}(q(z)||p(z|x, \theta^{(t)}))$. The KL divergence is minimized when $q(z) = p(z|x, \theta^{(t)})$. □

The M-step consists in solving

$$\theta^{(t+1)} = \arg\max_{\theta} \mathcal{L}(q^{(t+1)}, \theta)$$

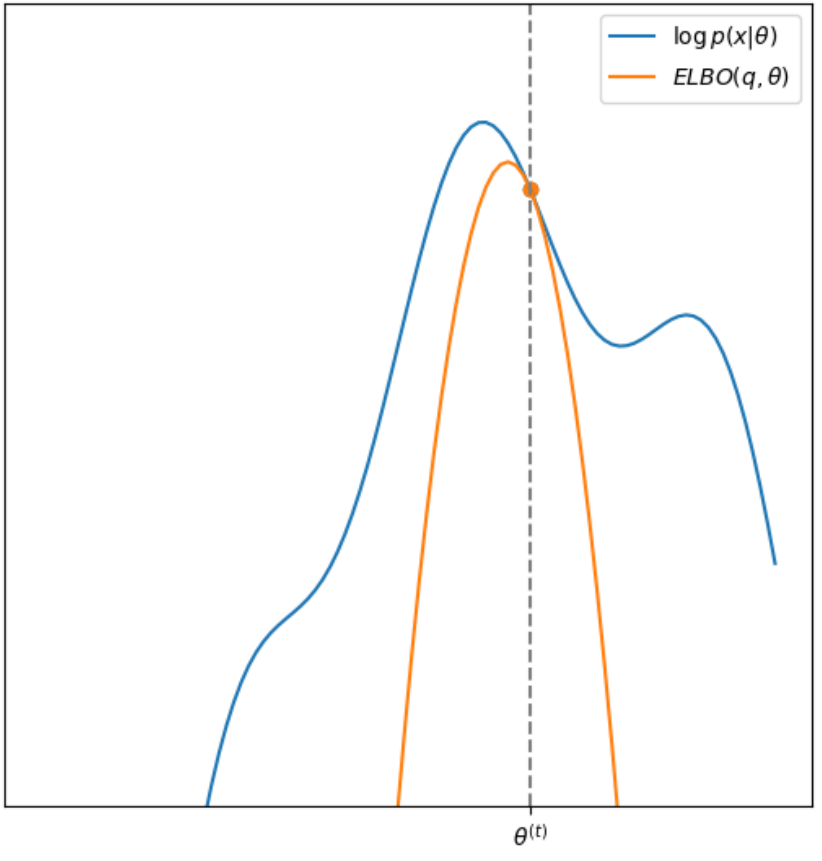
which can be rewritten as

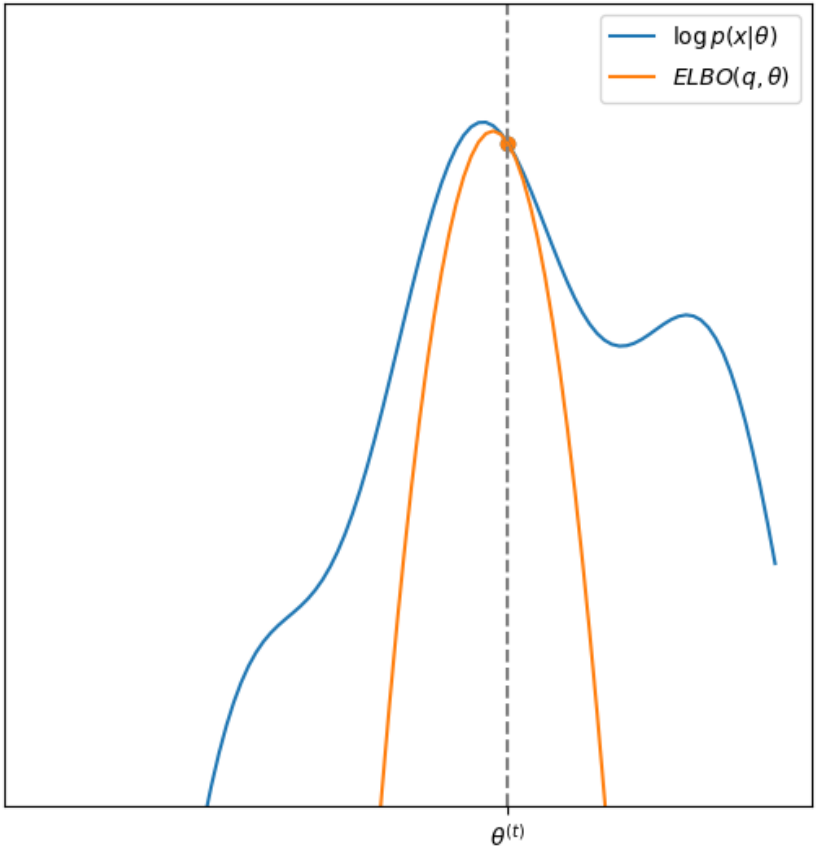
$$\begin{aligned}
\theta^{(t+1)} &= \arg\max_{\theta} \mathbb{E}_{q^{(t+1)}(z)} \left[\log \frac{p(x, z|\theta)}{q^{(t+1)}(z)} \right] \\
&= \arg\max_{\theta} \mathbb{E}_{q^{(t+1)}(z)} \left[\log p(x, z|\theta) \right] - \mathbb{E}_{q^{(t+1)}(z)} \left[\log q^{(t+1)}(z) \right] \\
&= \arg\max_{\theta} \mathbb{E}_{q^{(t+1)}(z)} \left[\log p(x, z|\theta) \right]
\end{aligned}$$

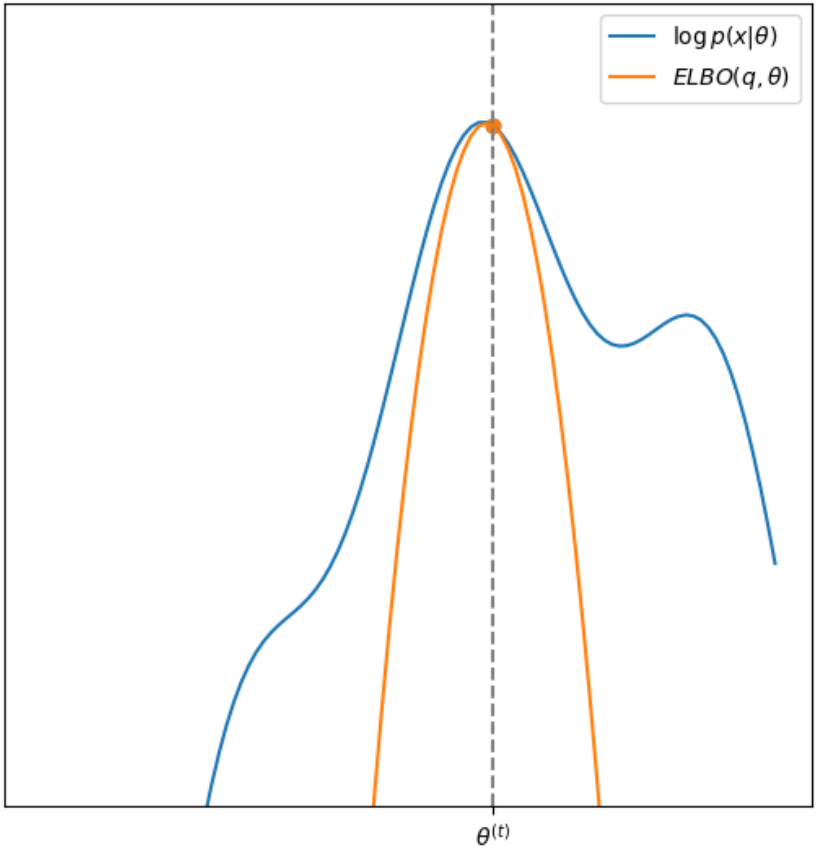
since the entropy term $H(q^{(t+1)}(z)) = \mathbb{E}_{q^{(t+1)}(z)} \left[-\log q^{(t+1)}(z) \right]$ does not depend on θ.

Depending on the model, this maximization can sometimes be done in closed-form. Otherwise, numerical optimization algorithms can be used.

Finally, to initialize the algorithm, we need to set initial parameters $\boldsymbol{\theta}^{(0)}$, which can be done randomly or based on prior knowledge.


$\log p(x|\theta)$
$ELBO(q, \theta)$
$\theta^{(t)}$


$\log p(x|\theta)$
$ELBO(q, \theta)$
$\theta^{(t)}$


$\log p(x|\theta)$
$ELBO(q, \theta)$
$\theta^{(t)}$

Proposition. The EM algorithm monotonically increases the marginal log-likelihood at each iteration, i.e.,

$$\log p(x|\theta^{(t)}) \leq \log p(x|\theta^{(t+1)})$$

for all $t \geq 0$.

Proof. Assume we have parameter estimates $\theta^{(t)}$ at iteration t.

$$\begin{aligned}
\log p(x|\theta^{(t)}) &= \mathcal{L}(q^{(t+1)}, \theta^{(t)}) + \mathrm{KL}(q^{(t+1)}(z)||p(z|x, \theta^{(t)})) \\
&= \mathcal{L}(q^{(t+1)}, \theta^{(t)}) \\
&\leq \mathcal{L}(q^{(t+1)}, \theta^{(t+1)}) \\
&\leq \log p(x|\theta^{(t+1)}),
\end{aligned}$$

where the second equality holds since the E-step sets $q^{(t+1)} = p(z|x, \theta^{(t)})$, making the KL gap zero; the first inequality follows from the M-step optimization; and the last inequality holds because the ELBO is always a lower bound on the log-likelihood. □

EM for the Old Faithful Geyser model

For the Gaussian mixture model introduced earlier, the E-step consists in computing the posterior distribution over the latent variables:

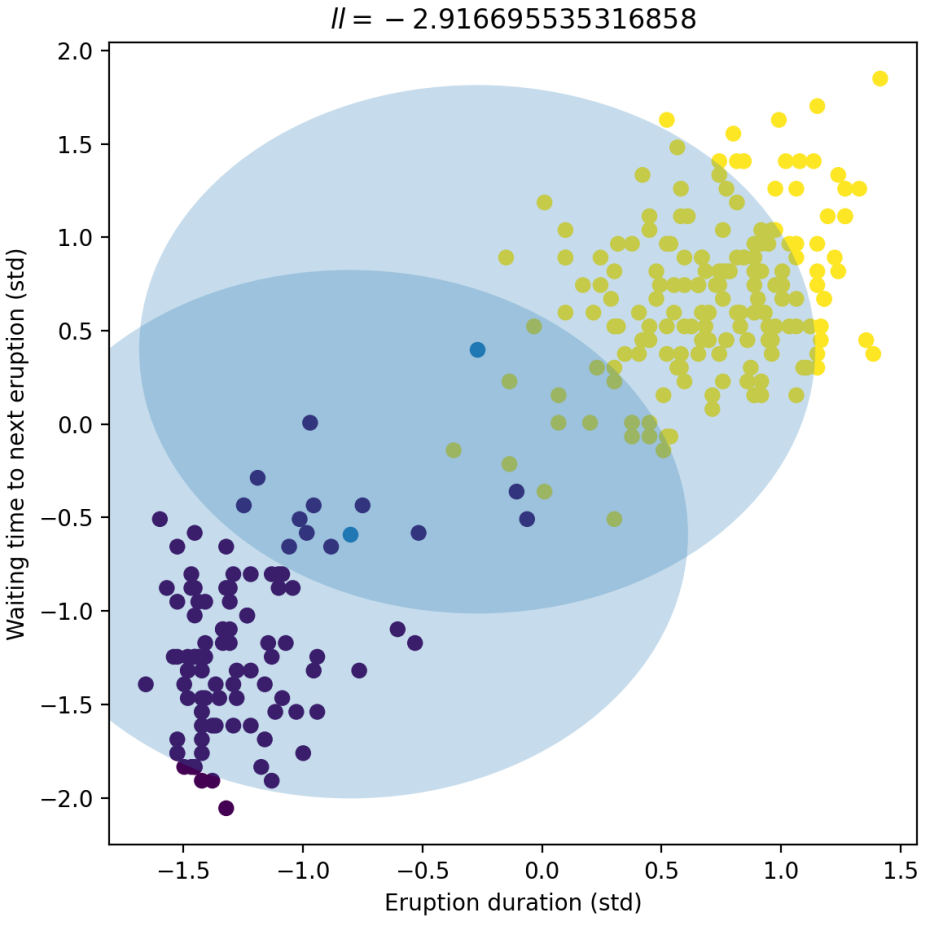
$$
\begin{aligned}
q^{(t+1)}(z_n = k) &= p(z_n = k | x_n, \pi^{(t)}, \mu^{(t)}, \sigma^{2(t)}) \\
&= \frac{p(x_n | z_n = k, \mu^{(t)}, \sigma^{2(t)}) p(z_n = k | \pi^{(t)})}{\sum_{j=1}^{2} p(x_n | z_n = j, \mu^{(t)}, \sigma^{2(t)}) p(z_n = j | \pi^{(t)})} \\
&= \frac{\pi_k^{(t)} \mathcal{N}(x_n | \mu_k^{(t)}, \sigma_k^{2(t)} I)}{\sum_{j=1}^{2} \pi_j^{(t)} \mathcal{N}(x_n | \mu_j^{(t)}, \sigma_j^{2(t)} I)}
\end{aligned}
$$

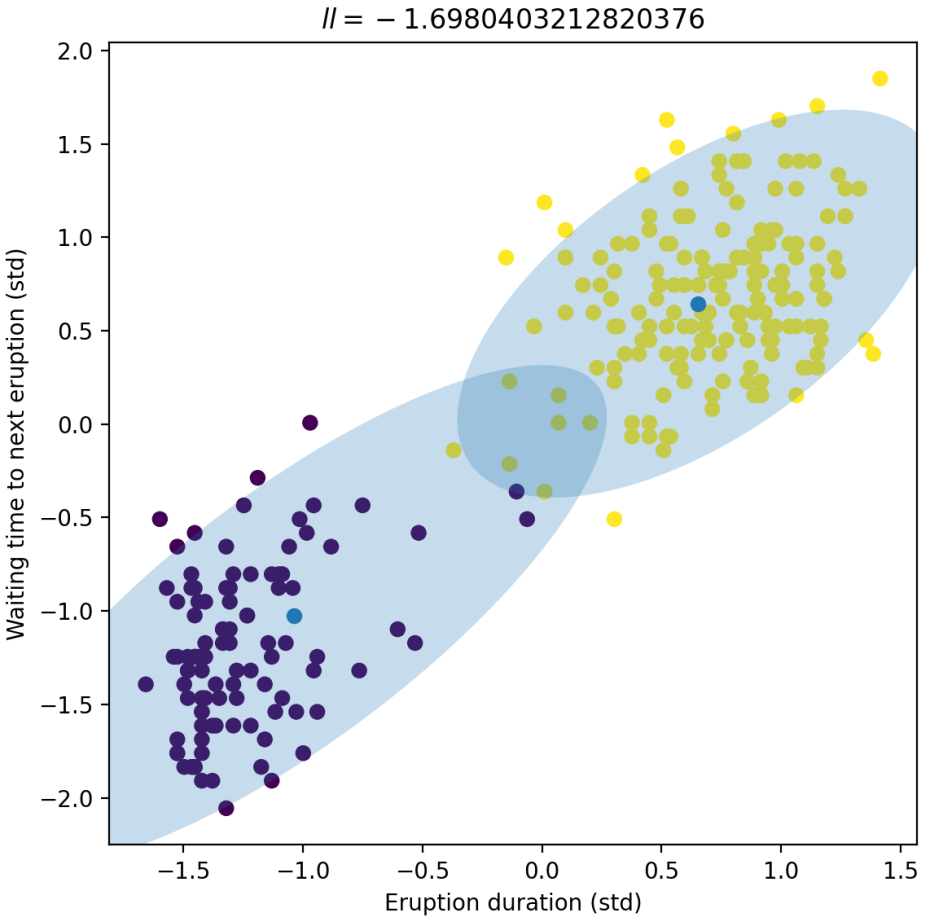
for $k = 1, 2$ and $n = 1, \ldots, N$.

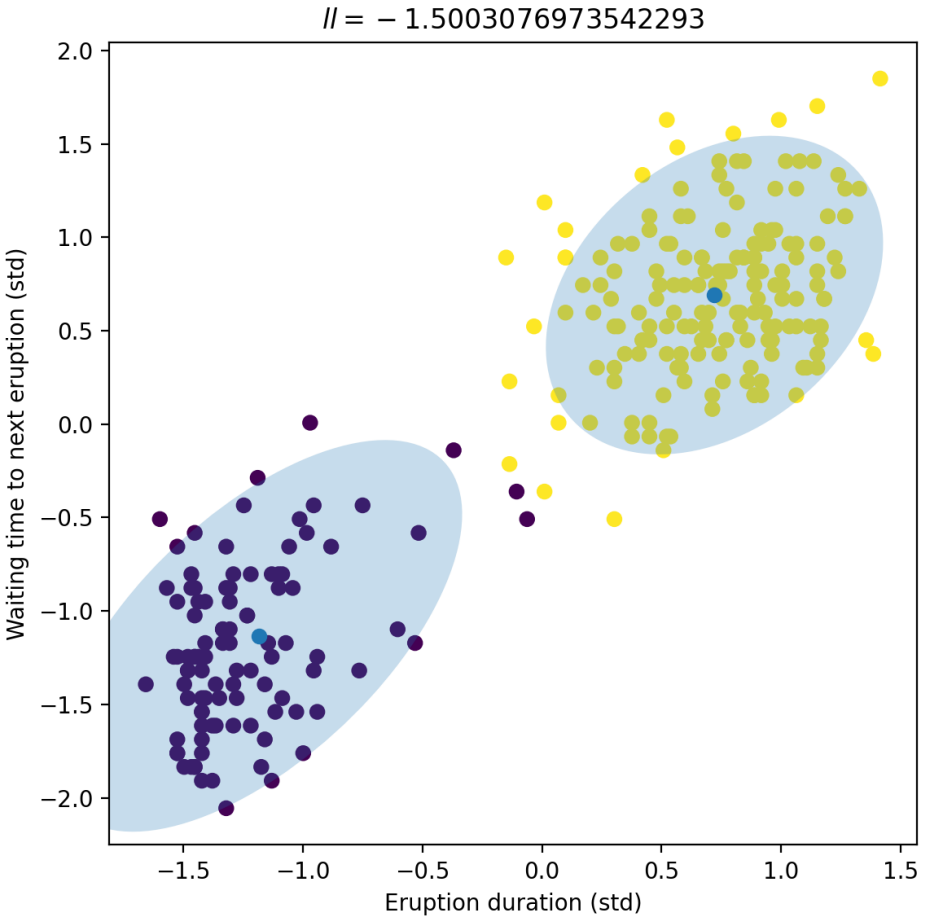
The M-step consists in updating the parameters as

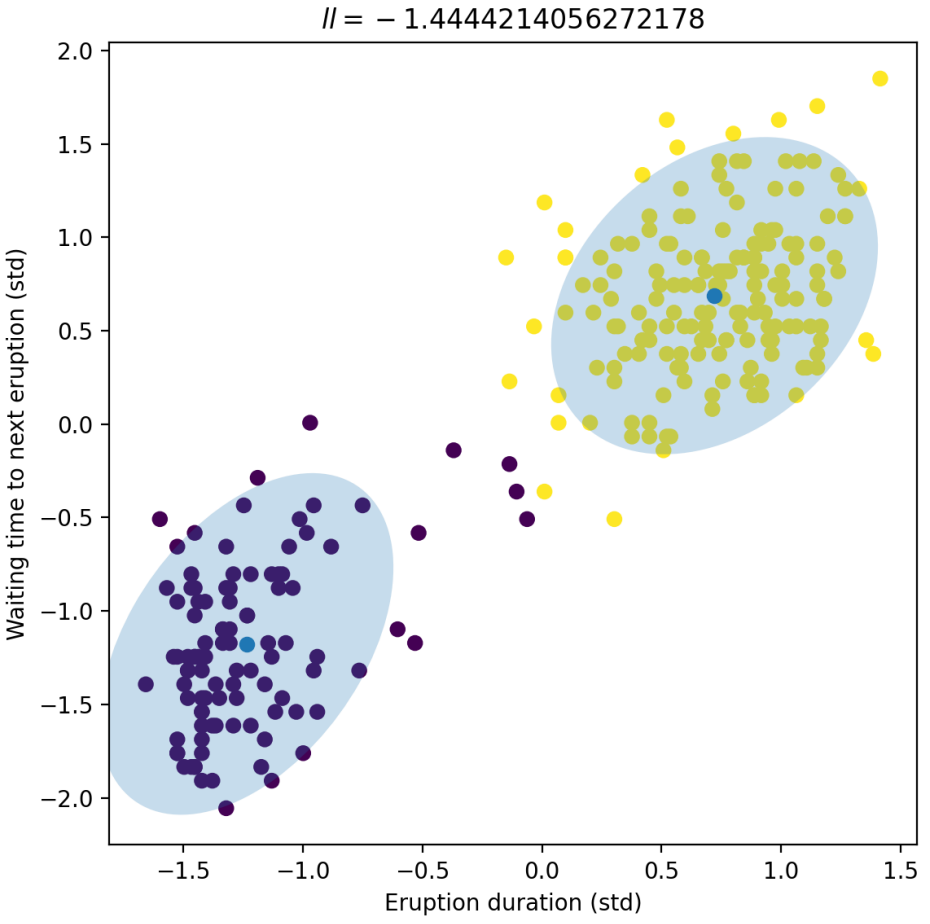
$$\mu_k^{(t+1)} = \frac{\sum_{n=1}^N q^{(t+1)}(z_n = k) x_n}{\sum_{n=1}^N q^{(t+1)}(z_n = k)}$$
$$\sigma_k^{2(t+1)} = \frac{\sum_{n=1}^N q^{(t+1)}(z_n = k) |x_n - \mu_k^{(t+1)}|^2}{\sum_{n=1}^N q^{(t+1)}(z_n = k)}$$
$$\pi_k^{(t+1)} = \frac{1}{N} \sum_{n=1}^N q^{(t+1)}(z_n = k)$$

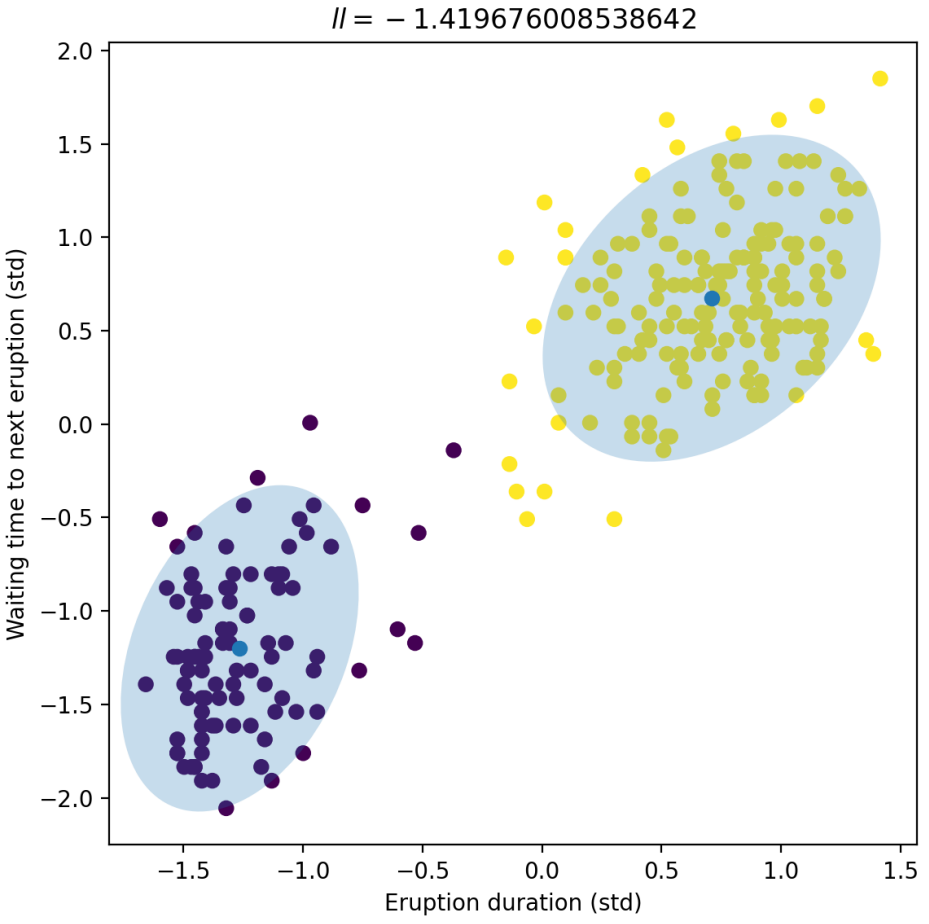
for $k = 1, 2$, which follows differentiating the expected complete-data log-likelihood and setting to zero.

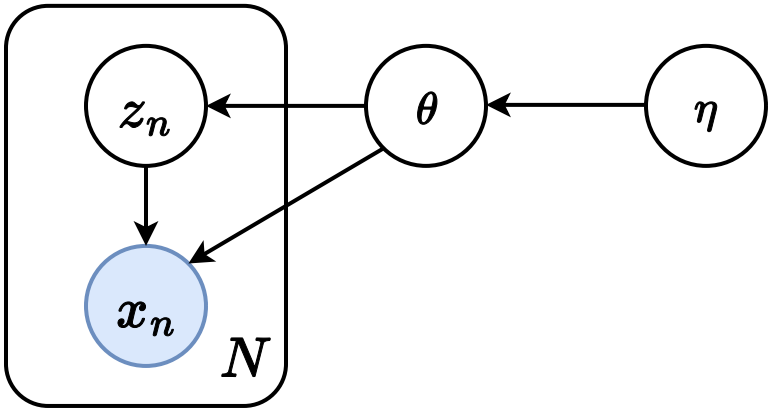

$ll = -2.916695535316858$
Waiting time to next eruption (std)
Eruption duration (std)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


ll = −1.6980403212820376
Waiting time to next eruption (std)
Eruption duration (std)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


$ll = -1.5003076973542293$
Waiting time to next eruption (std)
Eruption duration (std)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


ll = − 1.4444214056272178
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
Waiting time to next eruption (std)
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Eruption duration (std)


$ll = -1.419676008538642$
Waiting time to next eruption (std)
Eruption duration (std)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5



Empirical Bayes

In hierarchical Bayesian models, model parameters θ are treated as random variables with prior distribution $p(\theta|\eta)$, where η are hyper-parameters with their own hyper-prior $p(\eta)$.

A full Bayesian treatment to obtain $p(\theta|x)$ would require integrating out both the latent variables z and the hyper-parameters η, thus computing

$$p(\theta|x) = \iint p(\theta, z, \eta|x)\, dz\, d\eta.$$

This is often intractable.

An alternative is to use an **empirical Bayes** approach, which consists in approximating $p(\theta) = \int p(\theta|\eta)p(\eta)\,d\eta$ by a point estimate $p(\theta|\hat{\eta})$, where

$$\hat{\eta} = \arg\max_{\eta} p(x|\eta) = \arg\max_{\eta} \iint p(x, z|\theta)p(\theta|\eta)\,dz\,d\theta.$$

Put otherwise, empirical Bayes consists in estimating the prior $p(\theta)$ over model parameters from the data.

While empirical Bayes is unorthodox from a fully Bayesian perspective, it can lead to good practical results and is used in many applications.

The maximization can be performed using the EM algorithm by treating both the latent variables z and the model parameters θ as unobserved data.

- E-step: compute the posterior distribution over both latents and parameters

$$q^{(t+1)}(z, \theta) = p(z, \theta | x, \eta^{(t)}).$$

- M-step: update the hyper-parameters as

$$\eta^{(t+1)} = \arg\max_{\eta} \mathbb{E}_{q^{(t+1)}(z,\theta)} \left[\log p(x, z | \theta) + \log p(\theta | \eta) \right].$$

Learning diffusion priors by EM (Rozet et al, 2024)

EM can also be used to learn complex prior distributions parameterized by deep generative models, such as diffusion models, from noisy and incomplete observations only.

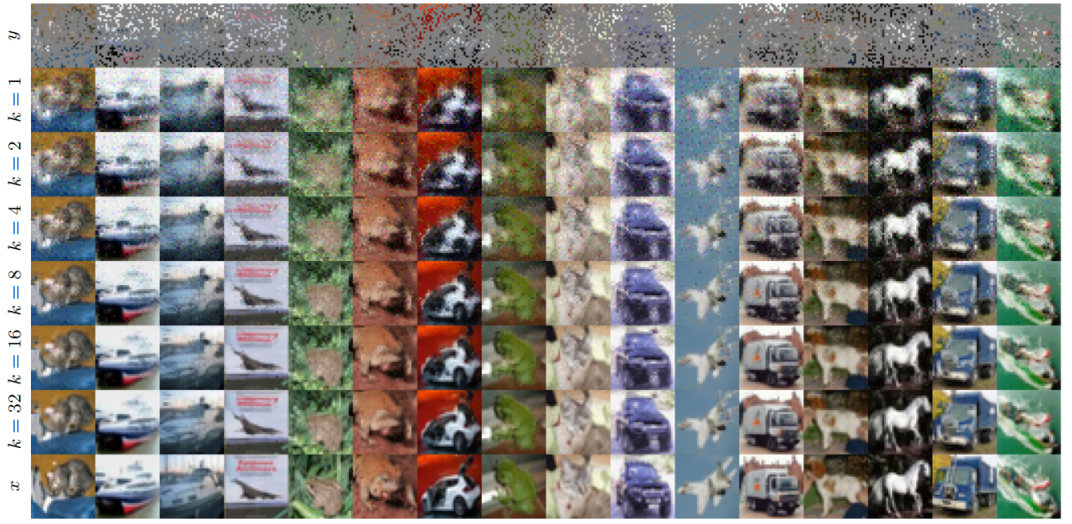


Figure 8. Example of samples from the posterior $q_{\theta_k}(x \mid y)$ along the EM iterations for the CIFAR-10 experiment. The generated images become gradually more detailed and less noisy.

—

Credits: Rozet et al (arXiv:2405.13712), 2024.

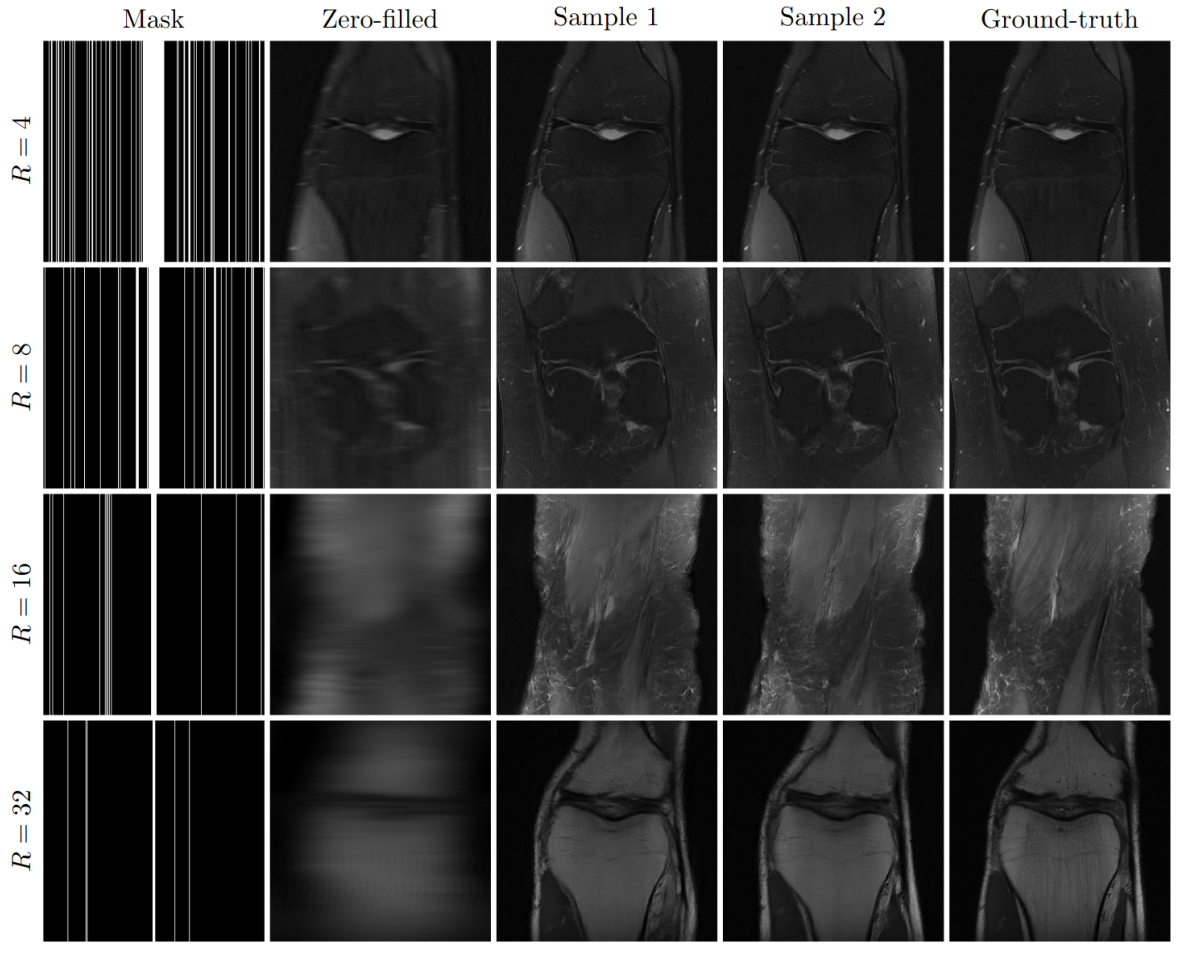


Figure 6. Examples of posterior samples for accelerated MRI using a diffusion prior trained from k-space observations only. Posterior samples are detailed and present plausible variations, while remaining consistent with the observation. We provide the zero-filled inverse, where missing frequencies are set to zero, as baseline.

—

The end.