Learning Diffusion Priors from Observations by Expectation Maximization

François Rozet, Gérôme Andry, François Lanusse and Gilles Louppe

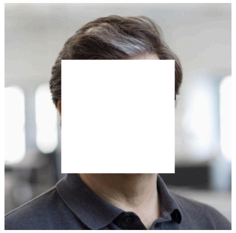
arXiv:2405.13712
francois-rozet/diffusion-priors

TL;DR We adapt the expectation-maximization algorithm to train diffusion models from (heavily) incomplete and noisy observations only. Additionally, we propose MMPS, a faster and more accurate posterior sampling scheme for unconditional diffusion models.

Introduction

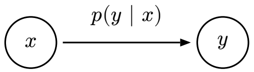
Many scientific applications are **inverse problems**, where the goal is to recover a latent x given an observation y.

$y = \text{mask}(x) + \text{noise}$



y is not sufficient to recover x unless we have **prior knowledge**

With a prior $p(x)$, the target becomes the posterior distribution $p(x \mid y)$.

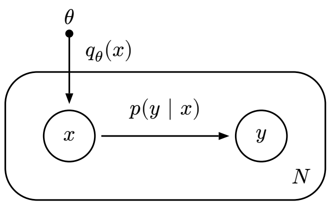


$$\overbrace{p(x \mid y)}^{\text{posterior}} \propto \overbrace{p(x)}^{\text{prior}} \overbrace{p(y \mid x)}^{\text{likelihood}} \quad (1)$$

Recently, **diffusion models** (DMs) proved to be remarkable priors for posterior inference. But can they be trained from **incomplete and noisy observations** only?

Empirical Bayes (EB)

EB formulates this problem as finding the parameters θ of a prior model $q_\theta(x)$ for which the evidence $q_\theta(y)$ is closest to the empirical distribution of observations $p(y)$.



$$q_\theta(y) = \int p(y \mid x)\, q_\theta(x)\, \mathrm{d}x \quad (2)$$

$$\begin{aligned} &\arg\min_\theta \mathrm{KL}(p(y) \parallel q_\theta(y)) \\ &= \arg\min_\theta \mathbb{E}_{p(y)}[-\log q_\theta(y)] \end{aligned} \quad (3)$$

Sadly, with a diffusion prior $q_\theta(x)$, the density $q_\theta(y)$ is not tractable.

Expectation-Maximization (EM) algorithm

For any two sets of parameters θ_a and θ_b,

$$\log q_{\theta_a}(y) - \log q_{\theta_b}(y) \geq \mathbb{E}_{q_{\theta_b}(x \mid y)}\left[\log q_{\theta_a}(x, y) - \log q_{\theta_b}(x, y)\right] \quad (4)$$

Therefore, starting from θ_0, the EM update

$$\theta_{k+1} = \arg\max_\theta \mathbb{E}_{p(y)} \mathbb{E}_{q_{\theta_k}(x \mid y)}\left[\log q_\theta(x, y) - \cancel{\log q_{\theta_k}(x, y)}\right] \quad (5)$$

leads to a **sequence of parameters** θ_k for which $\mathbb{E}_{p(y)}\left[\log q_{\theta_k}(y)\right]$ is monotonically increasing and converges to a local optimum.

Methods

In the context of EB, $q_\theta(x, y) = q_\theta(x)\, p(y \mid x)$ and the EM update becomes

$$\theta_{k+1} = \arg\max_\theta \mathbb{E}_{p(y)} \mathbb{E}_{q_{\theta_k}(x \mid y)}\left[\log q_\theta(x) + \cancel{\log p(y \mid x)}\right] \quad (6)$$

Intuitively, $q_{\theta_{k+1}}(x) \approx \int q_{\theta_k}(x \mid y)\, p(y)\, \mathrm{d}y$ is more consistent with the distribution of observations $p(y)$ than $q_{\theta_k}(x)$.

As long as we can
(i) generate samples from the posterior $q_{\theta_k}(x \mid y)$ and
(ii) train the prior $q_{\theta_{k+1}}(x)$ to fit these samples,
we can train any model $q_\theta(x)$ from observations, including DMs!

Moment Matching Posterior Sampling (MMPS)

To generate from $p(x)$, DMs approximate the score $\nabla_{x_t} \log p(x_t)$ of a series of increasingly noisy distributions $p(x_t) = \int \mathcal{N}(x_t \mid x, \Sigma_t)\, p(x)\, \mathrm{d}x$. To sample from the posterior $p(x \mid y)$, we need to approximate

$$\overbrace{\nabla_{x_t} \log p(x_t \mid y)}^{\text{posterior score}} = \overbrace{\nabla_{x_t} \log p(x_t)}^{\text{prior score}} + \overbrace{\nabla_{x_t} \log p(y \mid x_t)}^{\text{likelihood score}} \quad (7)$$

For a linear Gaussian observation process $p(y \mid x) = \mathcal{N}(y \mid Ax, \Sigma_y)$, the approximation $p(x \mid x_t) \approx \mathcal{N}(x \mid \mathbb{E}[x \mid x_t], \mathbb{V}[x \mid x_t])$ leads to

$$\begin{aligned} \nabla_{x_t} \log p(y \mid x_t) &\approx \nabla_{x_t} \log \mathcal{N}\left(y \mid A\mathbb{E}[x \mid x_t], \Sigma_y + A\mathbb{V}[x \mid x_t]A^\top\right) \\ &\approx \nabla_{x_t} \mathbb{E}[x \mid x_t]^\top A^\top \underbrace{\left(\Sigma_y + A\mathbb{V}[x \mid x_t]A^\top\right)^{-1}\left(y - A\mathbb{E}[x \mid x_t]\right)}_{\text{symmetric positive definite linear system}} \end{aligned} \quad (8)$$

$\mathbb{E}[x \mid x_t]$ and $\mathbb{V}[x \mid x_t]$ are linked to the score via Tweedie's formulae

$$\begin{aligned} \mathbb{E}[x \mid x_t] &= x_t + \Sigma_t \nabla_{x_t} \log p(x_t) \\ \mathbb{V}[x \mid x_t] &= \Sigma_t + \Sigma_t \nabla_{x_t}^2 \log p(x_t) \Sigma_t = \Sigma_t \nabla_{x_t}^\top \mathbb{E}[x \mid x_t] \end{aligned} \quad (9)$$

Instead of computing an expensive matrix inverse, we can solve the linear system in Eq. (8) with the **conjugate gradient** method.

Results

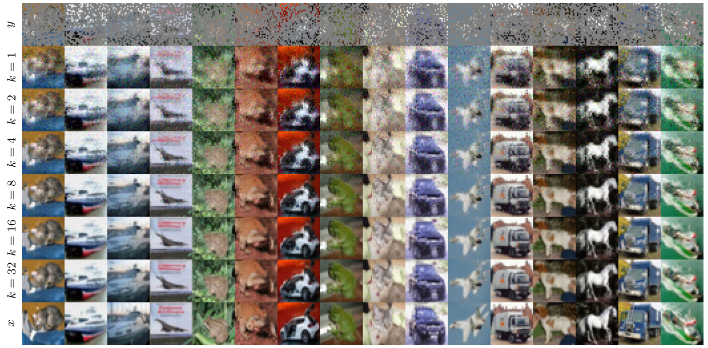


Figure 1. Samples from the posterior $q_{\theta_k}(x \mid y)$ along the EM iterations for the corrupted (75%) CIFAR-10 experiment. Samples become gradually more detailed and less noisy with iterations.

Daras (2023)

Corruption	FID ↓	IS ↑
0.20	11.70	7.97
0.40	18.85	7.45
0.60	28.88	6.88

Ours

Corruption	FID ↓	IS ↑
0.25	5.88	8.83
0.50	6.76	8.75
0.75	13.18	8.14

Table 1. Evaluation of final priors trained on corrupted CIFAR-10.

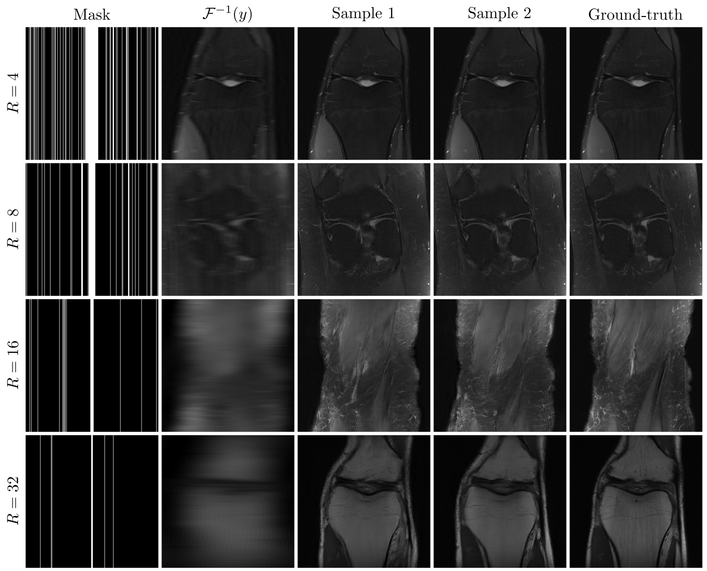


Figure 2. Accelerated MRI posterior samples using a diffusion prior trained from incomplete ($R = 8$) spectral observations only. Samples are detailed and varied, while remaining consistent with the observation.