Implicit Diffusion: Efficient Optimization through Stochastic Sampling

Recent Advances and Future Directions for Sampling, 16-18 October 2024

Pierre Marion ∗, **Anna Korba** †, Peter Bartlett, Mathieu Blondel, Valentin De Bortoli, Arnaud Doucet, Felipe Llinares-López, Courtney Paquette, Quentin Berthet – arXiv:2402.05468

* EPFL, † CREST, ENSAE IP Paris, every one else = Google.

General sampling problem

Suppose you are interested in some target probability distribution $\pi^\star \in \mathcal{P}(\mathbb{R}^d)$.

The goal of sampling is to generate samples from $\pi^\star$, having access only to some partial information, e.g.:

1. its unnormalized density

$$\pi^\star(x) = \frac{\exp(-V(x))}{Z}, \quad Z = \int \exp(-V(x))\,dx$$

   Example: $\pi^\star$ is a posterior distribution in Bayesian inference

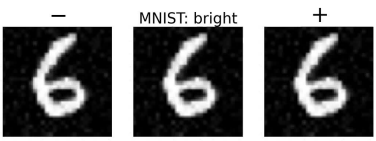
2. i.i.d. samples

$$x_1, \dots x_n \sim \pi^\star$$

   Example: $\pi^\star = p_{data}$ for some data of interest (e.g. images)

Motivation: sampling while optimizing (“Bilevel sampling”)

Example:



This work

- We will parametrize $\pi^\star(\theta) \in \mathcal{P}(\mathbb{R}^d)$ with $\theta \in \mathbb{R}^p$ and optimize $\mathcal{F}(\pi^\star(\theta))$ over θ.
- We study two sampling algorithms: Langevin and Diffusion models.

General framework

Optimization objective

$$\min_{\theta \in \mathbb{R}^p} \ell(\theta) := \min_{\theta \in \mathbb{R}^p} \mathcal{F}(\pi^\star(\theta)) \,,$$

where $\pi^\star(\theta) \in \mathcal{P}(\mathbb{R}^d)$ is the outcome (either in the limit when $s \to \infty$, or for some fixed $s = T$) of the iterates

$$p_{s+1} = \Sigma_s(p_s, \theta) \,.$$

Two questions

- How to evaluate gradients of ℓ w.r.t. θ?
- How to avoid a nested-loop approach?

Langevin diffusions

Continuous time

- Let $p_0 \in \mathcal{P}(\mathbb{R}^d)$. $\quad X_0 \sim p_0\,, \quad \mathrm{d}X_t = -\nabla_1 V(X_t, \theta)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,.$
- $\pi^\star(\theta)$ is the limiting distribution of X_t when $t \to \infty$, given by the Gibbs distributions

$$\pi^\star(\theta)[x] = \exp(-V(x,\theta))/Z_\theta\,, \text{ where } Z_\theta = \int \exp(-V(x,\theta))dx.$$

Example: $V(x,\theta) = 0.5(x-\theta)^2$, hence $\pi^\star : \theta \to \mathcal{N}(\theta, 1)$.

Discrete time

- $X_0 \sim p_0\,, \quad X_{s+1} = X_s - \eta\nabla_1 V(X_s, \theta) + \sqrt{2\eta}n_s \quad n_s \sim \mathcal{N}(0, I_d).$
- Defines implicitly an iterative sampling process Σ_s by letting p_s the distribution of X_s :

$$p_{s+1} = \Sigma_s(p_s, \theta)\,.$$

Denoising diffusions

Forward process

- $X_0 \sim p_{\text{data}}\,, \quad \mathrm{d}X_t = -X_t \mathrm{d}t + \sqrt{2}\mathrm{d}B_t,\ X_t \sim p_t, \quad p_\infty = \mathcal{N}(0, I_d).$

Backward process

- $Y_0 \sim \mathcal{N}(0, I_d)\,, \quad \mathrm{d}\,Y_t = \{\,Y_t + 2s_\theta(\,Y_t,\, T-t)\}\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,.$
- If $s_\theta(\cdot, t) \approx \nabla \log p_t$, then $Y_T \approx p_{\text{data}}$ for large T.
- Deterministic alternative ("probability flow") $\mathrm{d}\,Y_t = \{\,Y_t + s_\theta(\,Y_t,\, T-t)\}\mathrm{d}t\,.$

In practice

- time-discretized processes
- s_θ is learnt through a neural network with score matching.
- $\pi^\star(\theta)$ is the distribution of Y_T.

Gradient estimation through sampling

Implicit gradient estimation

We assume that there exists $\Gamma : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^p \to \mathbb{R}^p$ such that

$$\nabla \ell(\theta) = \Gamma(\pi^\star(\theta), \theta)\,.$$

Gradient estimation through sampling

Implicit gradient estimation

We assume that there exists $\Gamma : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^p \to \mathbb{R}^p$ such that

$$\nabla \ell(\theta) = \Gamma(\pi^\star(\theta), \theta)\,.$$

Gradient estimation through sampling

Implicit gradient estimation

We assume that there exists $\Gamma : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^p \to \mathbb{R}^p$ such that

$$\nabla \ell(\theta) = \Gamma(\pi^\star(\theta), \theta)\,.$$

In practice

- We don't have access to $\pi^\star(\theta)$ but to $\hat{\pi} \approx \pi^\star(\theta)$.
- We approximate $\nabla \ell(\theta)$ by $\Gamma(\hat{\pi}, \theta)$.

Example: Langevin diffusions

$\mathcal{F}(p) = -\mathbb{E}_{x\sim p}[R(x)]$ (Reward minimization)

- $\nabla \ell_{\text{reward}}(\theta) = \mathrm{Cov}_{X\sim\pi^\star(\theta)}[R(X), \nabla_2 V(X,\theta)]$.
- Thus $\Gamma_{\text{reward}}(p,\theta) := \mathrm{Cov}_{X\sim p}[R(X), \nabla_2 V(X,\theta)]$.
- Handles non-differentiable rewards!

$\mathcal{F}(p) = \mathrm{KL}(p_{\text{ref}} \,||\, p)$ (training of Energy-based models)

- $\nabla \ell_{\text{ref}}(\theta) = \mathbb{E}_{X\sim p_{\text{ref}}}[\nabla_2 V(X,\theta)] - \mathbb{E}_{X\sim\pi^\star(\theta)}[\nabla_2 V(X,\theta)]$ (contrastive learning).
- Thus $\Gamma_{\text{ref}}(p,\theta) := \mathbb{E}_{X\sim p_{\text{ref}}}[\nabla_2 V(X,\theta)] - \mathbb{E}_{X\sim p}[\nabla_2 V(X,\theta)]$.

- Extends naturally to linear combinations of Γ_{reward} and Γ_{ref}.

Example: Langevin diffusions

$\mathcal{F}(p) = -\mathbb{E}_{x\sim p}[R(x)]$ (Reward minimization)

- $\nabla \ell_{\text{reward}}(\theta) = \mathrm{Cov}_{X\sim\pi^\star(\theta)}[R(X), \nabla_2 V(X,\theta)]$.
- Thus $\Gamma_{\text{reward}}(p,\theta) := \mathrm{Cov}_{X\sim p}[R(X), \nabla_2 V(X,\theta)]$.
- Handles non-differentiable rewards!

$\mathcal{F}(p) = \mathrm{KL}(p_{\text{ref}} \,||\, p)$ (training of Energy-based models)

- $\nabla \ell_{\text{ref}}(\theta) = \mathbb{E}_{X\sim p_{\text{ref}}}[\nabla_2 V(X,\theta)] - \mathbb{E}_{X\sim\pi^\star(\theta)}[\nabla_2 V(X,\theta)]$ (contrastive learning).
- Thus $\Gamma_{\text{ref}}(p,\theta) := \mathbb{E}_{X\sim p_{\text{ref}}}[\nabla_2 V(X,\theta)] - \mathbb{E}_{X\sim p}[\nabla_2 V(X,\theta)]$.

- Extends naturally to linear combinations of Γ_{reward} and Γ_{ref}.

[1] Linear case studied by: Atchadé et al., 2017; De Bortoli et al., 2021; Xiao & Zhang, 2014; Rosasco et al., 2020; Nitanda, 2014; Tadic & Doucet, 2017.

Example: Denoising diffusions

- Gradients through differential equations solvers can be computed through the adjoint method
- Consider the ODE $\mathrm{d}\,Y_t = \mu(t, Y_t, \theta)\mathrm{d}t$ integrated between 0 and some $T > 0$ (e.g. denoising diffusion ODE with the appropriate choice of μ).
- Let $R : \mathbb{R}^d \to \mathbb{R}$. Let $Z_0 = Y_T, A_0 = \nabla R(Z_0), G_0 = 0$, and consider the ODE system

$$\mathrm{d}Z_t = -\mu(t, Z_t, \theta)\mathrm{d}t\,, \quad \mathrm{d}A_t = A_t^\top \nabla_2 \mu(T - t, Z_t, \theta)\mathrm{d}t\,, \quad \mathrm{d}G_t = A_t^\top \nabla_3 \mu(T - t, Z_t, \theta)\mathrm{d}t\,.$$

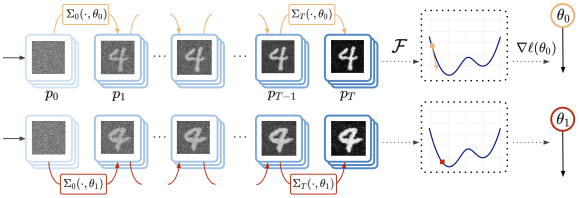
  G_T is the derivative of $R(Y_T)$ with respect to θ.
- Replacing $Z_0 \sim \pi^*(\theta)$ yields an estimator of $\nabla l(\theta)$, and $Z_0 \sim p$ of $\Gamma(p, \theta)$.
- Works both for ODEs and SDEs.
- We can also define Γ as above in the SDE case for the KL objective $\mathrm{KL}(p \,||\, \pi^\star(\theta_0))$ with Girsanov's theorem

$$\mathrm{KL}((Y_t^1)_{t\geq 0} \,||\, (Y_t^2)_{t\geq 0}) = \int_0^T \mathbb{E}_{y\sim q_t^1} \|\mu(t, y, \theta_1) - \mu(t, y, \theta_2)\|^2 \mathrm{d}t\,,$$

Baseline algorithm (with Langevin or Diffusion models)

Algorithm Vanilla nested-loop approach (Baseline)

input $\theta_0 \in \mathbb{R}^p$, $p_0 \in \mathcal{P}$
  for $k \in \{0, \dots, K-1\}$ (outer optimization loop) **do**
    $p_k^{(0)} \leftarrow p_0$
    for $s \in \{0, \dots, T-1\}$ (inner sampling loop) **do**
      $p_k^{(s+1)} \leftarrow \Sigma_s(p_k^{(s)}, \theta_k)$
    $\hat{\pi}_k \leftarrow p_k^{(T)}$
    $\theta_{k+1} \leftarrow \theta_k - \eta\Gamma(\hat{\pi}_k, \theta_k)$ (or another optimizer)
output θ_K



Single-loop (Langevin)

This
point of view is well-suited for stationary processes with infinite-time horizon.

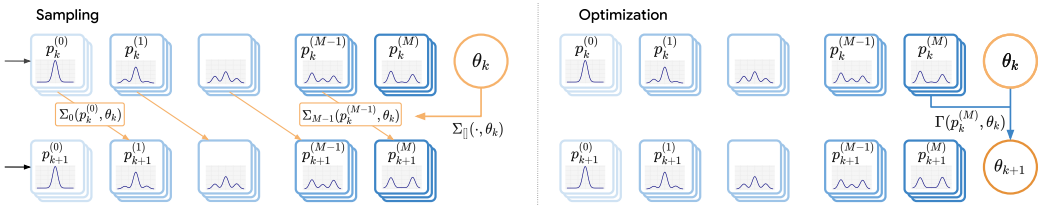
We show next
how to adapt our approach to sampling with diffusions with a finite-time horizon (and no stationary property).

Algorithm Implicit Diff. optimization, infinite time

input $\theta_0 \in \mathbb{R}^p$, $p_0 \in \mathcal{P}$
  for $k \in \{0, \dots, K-1\}$ (joint single loop) **do**
    $p_{k+1} \leftarrow \Sigma_k(p_k, \theta_k)$
    $\theta_{k+1} \leftarrow \theta_k - \eta \Gamma(p_k, \theta_k)$
output θ_K

Implicit diffusion

We evaluate in parallel several, say M, steps of the dynamics of the distribution p_k, through a queue of length M. Below $M = T$ for simplicity.



Left: Sampling - one step of the parameterized sampling is applied in parallel to all distributions in the queue.
Right: Optimization - the last element of the queue is used to compute a gradient for the parameter.

Implicit diffusion

runtime= $\mathcal{O}(K)$, gaining a factor of T compared to the nested-loop approach, but at a higher memory cost.

Algorithm Implicit Diff. optimization, finite time

input $\theta_0 \in \mathbb{R}^p$, $p_0 \in \mathcal{P}$
input $P_M = [p_0^{(0)}, \ldots, p_0^{(M)}]$
  for $k \in \{0, \ldots, K-1\}$ (joint single loop) **do**
    $p_{k+1}^{(0)} \leftarrow p_0$
    parallel $p_{k+1}^{(m+1)} \leftarrow \Sigma_m(p_k^{(m)}, \theta_k)$ for $m \in [M-1]$
    $\theta_{k+1} \leftarrow \theta_k - \eta\Gamma(p_k^{(M)}, \theta_k)$ (or another optimizer)
output θ_K

Experiments: (1) Reward training of Langevin diffusions

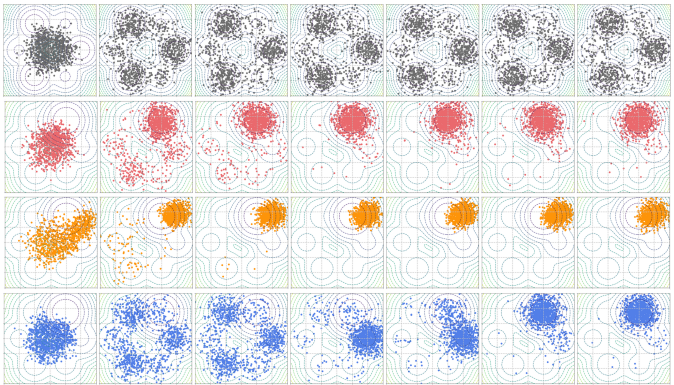
$$V(x,\theta) = -\log\left(\sum_{i=1}^{6} \sigma(\theta)_i \exp(-\|x-\mu_i\|^2)\right). \quad R(x) = \mathbf{1}_{x_1>0}\exp(-\|x-\mu\|^2).$$

Langevin diffusion with potential $V(\cdot, \theta_0)$ for some fixed $\theta_0 \in \mathbb{R}^p$, no reward.

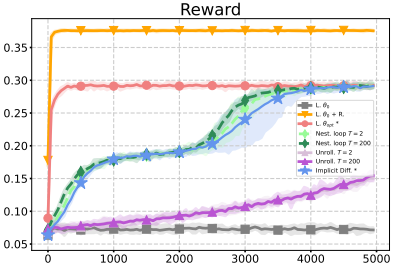
Langevin diffusion with potential $V(\cdot, \theta_{\text{opt}})$.

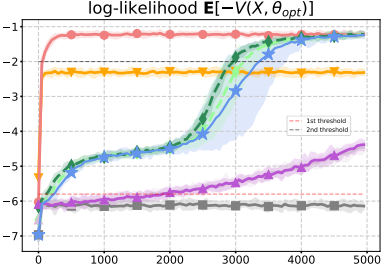
Langevin diffusion with potential $V(\cdot, \theta_0) - \lambda R_{\text{smooth}}$, where R_{smooth} is a smoothed version of R.

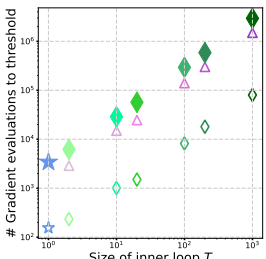
Implicit Diffusion: with $\mathcal{F}(p) = -\mathbb{E}_{X\sim p}[R(X)]$,



Quantitative metrics





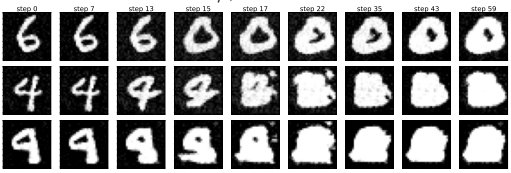


Experiments: (2) Reward training of denoising diffusions

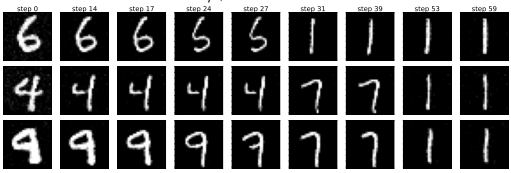
- Pretraining gives some parameters $\theta_{0,\text{MNIST}}$ and $\theta_{0,\text{CIFAR}}$.
- Objective for MNIST:

$$\mathcal{F}(p) = -\lambda \mathbb{E}_{X\sim p}(R_{\text{brightness}}(X)) + \beta\, \mathrm{KL}(p||\pi^\star(\theta_{0,\text{MNIST}}))\,,$$

$\lambda/\beta = 30$



$\lambda/\beta = -30$

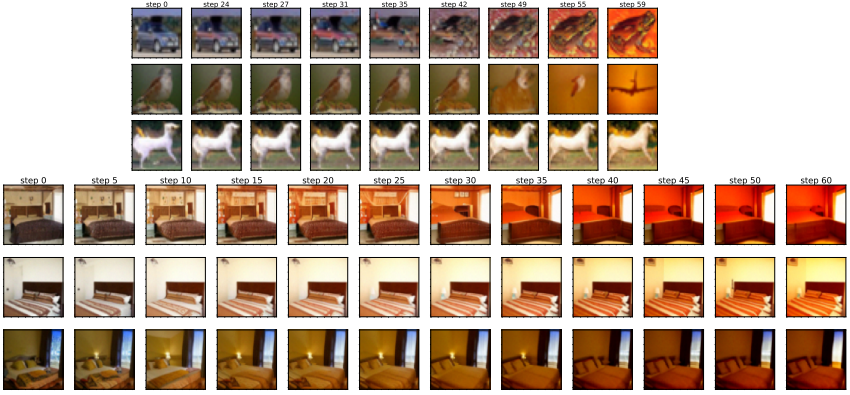


CIFAR-10 and LSUN

Objective for CIFAR-10:

$$\mathcal{F}(p) = \lambda \mathbb{E}_{X \sim p}(R_{\text{red}}(X)) + \beta \, \mathrm{KL}(p || \pi^{\star}(\theta_{0,\text{CIFAR}})) \, ,$$

where R_{red} is the average of the red channels minus the average of the other channels.



Results: MNIST and CIFAR-10

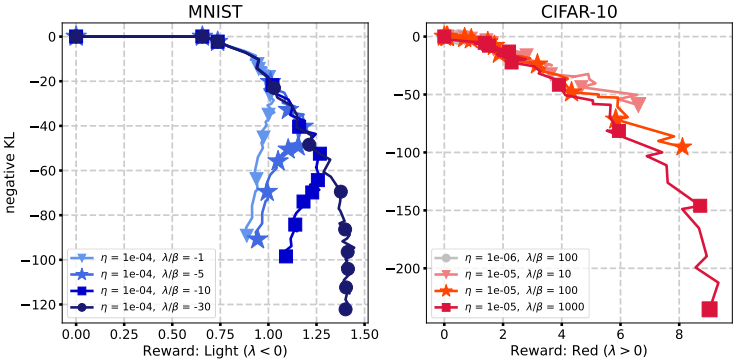


Figure: Reward training with **Implicit Diffusion** for various λ, η. For each dataset, we plot together the reward and the negative KL divergence w.r.t. $\pi^\star(\theta_0)$.

Theory: Implicit Diffusion for Langevin diffusions

Coupled differential system

$$\mathrm{d}X_t = -\nabla_1 V(X_t, \theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad X_t \sim p_t, \tag{1}$$

$$\mathrm{d}\theta_t = -\varepsilon_t \Gamma(p_t, \theta_t)\mathrm{d}t\,. \tag{2}$$

Theory: Implicit Diffusion for Langevin diffusions

Coupled differential system

$$\mathrm{d}X_t = -\nabla_1 V(X_t, \theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad X_t \sim p_t, \tag{1}$$

$$\mathrm{d}\theta_t = -\varepsilon_t \Gamma(p_t, \theta_t)\mathrm{d}t\,. \tag{2}$$

- p_t denotes the distribution of X_t.
- Simplification of the actual dynamics since the equation in θ is deterministic.

Theory: Implicit Diffusion for Langevin diffusions

Coupled differential system

$$\mathrm{d}X_t = -\nabla_1 V(X_t, \theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad X_t \sim p_t, \tag{1}$$

$$\mathrm{d}\theta_t = -\varepsilon_t \Gamma(p_t, \theta_t)\mathrm{d}t\,. \tag{2}$$

Simplification of the actual dynamics since the equation in θ is deterministic.

Link with bilevel optimization

- Let $\mathcal{G}_\theta(p) = \mathrm{KL}(p \,||\, \pi^\star(\theta))$. Our optimization problem can be recast as

$$\min_{\theta \in \mathbb{R}^p} \mathcal{F}(\pi^\star(\theta)) \quad \text{s.t.} \quad \pi^\star(\theta) \in \operatorname*{argmin}_{p \in \mathcal{P}} \mathcal{G}_\theta(p)\,.$$

Theory: Implicit Diffusion for Langevin diffusions

Coupled differential system

$$\mathrm{d}X_t = -\nabla_1 V(X_t, \theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad X_t \sim p_t, \tag{1}$$
$$\mathrm{d}\theta_t = -\varepsilon_t \Gamma(p_t, \theta_t)\mathrm{d}t\,. \tag{2}$$

Simplification of the actual dynamics since the equation in θ is deterministic.

Link with bilevel optimization

- Let $\mathcal{G}_\theta(p) = \mathrm{KL}(p \,||\, \pi^\star(\theta))$. Our optimization problem can be recast as

$$\min_{\theta \in \mathbb{R}^p} \mathcal{F}(\pi^\star(\theta)) \quad \text{s.t.} \quad \pi^\star(\theta) \in \operatorname*{argmin}_{p \in \mathcal{P}} \mathcal{G}_\theta(p)\,.$$

- Equation (1) is equivalent to $\mathrm{d}p_t = -\nabla_{W_2} \mathcal{G}_{\theta_t}(p_t)\mathrm{d}t$ (Jordan et al., 1998).

Theory: Implicit Diffusion for Langevin diffusions

Coupled differential system

$$\mathrm{d}X_t = -\nabla_1 V(X_t, \theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad X_t \sim p_t, \tag{1}$$

$$\mathrm{d}\theta_t = -\varepsilon_t \Gamma(p_t, \theta_t)\mathrm{d}t\,. \tag{2}$$

Assumptions

- $\pi^\star(\theta_t)$ verifies the Log-Sobolev inequality with constant $\mu > 0$ for all $t \geq 0$.
- The potential V is continuously differentiable and for $\theta \in \mathbb{R}^p, x \in \mathbb{R}^d$, $\|\nabla_2 V(x, \theta)\| \leq C$.
- For all $p \in \mathcal{P}$, $q \in \mathcal{P}$, $\theta \in \mathbb{R}^p$,

$$\|\Gamma(p, \theta)\| \leq C \text{ and } \|\Gamma(p, \theta) - \Gamma(q, \theta)\| \leq K_\Gamma \sqrt{\mathrm{KL}(p||q)}.$$

Example: $\|\Gamma_{\text{ref}}(p, \theta) - \Gamma_{\text{ref}}(q, \theta)\| = \|\mathbb{E}_{X\sim q}[\nabla_2 V(X, \theta)] - \mathbb{E}_{X\sim p}[\nabla_2 V(X, \theta)]\| \leq C\,\mathrm{TV}(p, q) \leq \frac{C}{\sqrt{2}}\sqrt{\mathrm{KL}(p||q)}$.

Theory: Implicit Diffusion for Langevin diffusions

Coupled differential system

$$\mathrm{d}X_t = -\nabla_1 V(X_t, \theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \qquad X_t \sim p_t, \tag{1}$$

$$\mathrm{d}\theta_t = -\varepsilon_t \Gamma(p_t, \theta_t)\mathrm{d}t\,. \tag{2}$$

Theorem

Take $\varepsilon_t = \min(1, \frac{1}{\sqrt{t}})$. Then, under the Assumptions,

$$\frac{1}{T}\int_0^T \|\nabla \ell(\theta_t)\|^2 \mathrm{d}t \leq \frac{c(\ln T)^2}{T^{1/2}}$$

for some $c > 0$ depending on the constants of the problem.

Theory: Implicit Diffusion for Langevin diffusions

Coupled differential system

$$\mathrm{d}X_t = -\nabla_1 V(X_t, \theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad X_t \sim p_t, \tag{1}$$

$$\mathrm{d}\theta_t = -\varepsilon_t \Gamma(p_t, \theta_t)\mathrm{d}t\,. \tag{2}$$

$$\frac{\mathrm{d}\ell}{\mathrm{d}t}(t) = \left\langle \nabla \ell(\theta_t), \frac{\mathrm{d}\theta_t}{\mathrm{d}t} \right\rangle$$

Theory: Implicit Diffusion for Langevin diffusions

Coupled differential system

$$\mathrm{d}X_t = -\nabla_1 V(X_t, \theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad X_t \sim p_t, \tag{1}$$

$$\mathrm{d}\theta_t = -\varepsilon_t \Gamma(p_t, \theta_t)\mathrm{d}t\,. \tag{2}$$

$$\begin{aligned}\frac{\mathrm{d}\ell}{\mathrm{d}t}(t) &= \left\langle \nabla\ell(\theta_t), \frac{\mathrm{d}\theta_t}{\mathrm{d}t} \right\rangle \\ &= -\varepsilon_t \langle \nabla\ell(\theta_t), \Gamma(p_t, \theta_t)\rangle\end{aligned}$$

Theory: Implicit Diffusion for Langevin diffusions

Coupled differential system

$$\mathrm{d}X_t = -\nabla_1 V(X_t, \theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad X_t \sim p_t, \tag{1}$$

$$\mathrm{d}\theta_t = -\varepsilon_t \Gamma(p_t, \theta_t)\mathrm{d}t\,. \tag{2}$$

$$\begin{aligned}
\frac{\mathrm{d}\ell}{\mathrm{d}t}(t) &= \left\langle \nabla\ell(\theta_t), \frac{\mathrm{d}\theta_t}{\mathrm{d}t} \right\rangle \\
&= -\varepsilon_t \langle \nabla\ell(\theta_t), \Gamma(p_t, \theta_t)\rangle \\
&= -\varepsilon_t \langle \nabla\ell(\theta_t), \Gamma(\pi^\star(\theta_t), \theta_t)\rangle + \varepsilon_t \langle \nabla\ell(\theta(t)), \Gamma(\pi^\star(\theta_t), \theta_t) - \Gamma(p_t, \theta_t)\rangle
\end{aligned}$$

Theory: Implicit Diffusion for Langevin diffusions

Coupled differential system

$$\mathrm{d}X_t = -\nabla_1 V(X_t, \theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad X_t \sim p_t, \tag{1}$$

$$\mathrm{d}\theta_t = -\varepsilon_t \Gamma(p_t, \theta_t)\mathrm{d}t\,. \tag{2}$$

$$\begin{aligned}
\frac{\mathrm{d}\ell}{\mathrm{d}t}(t) &= \left\langle \nabla\ell(\theta_t), \frac{\mathrm{d}\theta_t}{\mathrm{d}t} \right\rangle \\
&= -\varepsilon_t \langle \nabla\ell(\theta_t), \Gamma(p_t, \theta_t)\rangle \\
&= -\varepsilon_t \langle \nabla\ell(\theta_t), \Gamma(\pi^\star(\theta_t), \theta_t)\rangle + \varepsilon_t \langle \nabla\ell(\theta(t)), \Gamma(\pi^\star(\theta_t), \theta_t) - \Gamma(p_t, \theta_t)\rangle \\
&\leq -\varepsilon_t \|\nabla\ell(\theta_t)\|^2 + \varepsilon_t \|\nabla\ell(\theta_t)\| \|\Gamma(\pi^\star(\theta_t), \theta_t) - \Gamma(p_t, \theta_t)\|
\end{aligned}$$

Theory: Implicit Diffusion for Langevin diffusions

Coupled differential system

$$\mathrm{d}X_t = -\nabla_1 V(X_t, \theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad X_t \sim p_t, \tag{1}$$

$$\mathrm{d}\theta_t = -\varepsilon_t \Gamma(p_t, \theta_t)\mathrm{d}t\,. \tag{2}$$

$$\begin{aligned}
\frac{\mathrm{d}\ell}{\mathrm{d}t}(t) &= \left\langle \nabla\ell(\theta_t), \frac{\mathrm{d}\theta_t}{\mathrm{d}t} \right\rangle \\
&= -\varepsilon_t \langle \nabla\ell(\theta_t), \Gamma(p_t, \theta_t)\rangle \\
&= -\varepsilon_t \langle \nabla\ell(\theta_t), \Gamma(\pi^\star(\theta_t), \theta_t)\rangle + \varepsilon_t \langle \nabla\ell(\theta(t)), \Gamma(\pi^\star(\theta_t), \theta_t) - \Gamma(p_t, \theta_t)\rangle \\
&\leq -\varepsilon_t \|\nabla\ell(\theta_t)\|^2 + \varepsilon_t \|\nabla\ell(\theta_t)\| \|\Gamma(\pi^\star(\theta_t), \theta_t) - \Gamma(p_t, \theta_t)\| \\
&\leq -\varepsilon_t \|\nabla\ell(\theta_t)\|^2 + \varepsilon_t K_\Gamma \|\nabla\ell(\theta_t)\| \sqrt{\mathrm{KL}(p_t \,||\, \pi^\star(\theta_t))}
\end{aligned}$$

Theory: Implicit Diffusion for Langevin diffusions

Coupled differential system

$$\mathrm{d}X_t = -\nabla_1 V(X_t, \theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad X_t \sim p_t, \tag{1}$$

$$\mathrm{d}\theta_t = -\varepsilon_t \Gamma(p_t, \theta_t)\mathrm{d}t\,. \tag{2}$$

$$\begin{aligned}
\frac{\mathrm{d}\ell}{\mathrm{d}t}(t) &= \left\langle \nabla\ell(\theta_t), \frac{\mathrm{d}\theta_t}{\mathrm{d}t} \right\rangle \\
&= -\varepsilon_t \langle \nabla\ell(\theta_t), \Gamma(p_t, \theta_t)\rangle \\
&= -\varepsilon_t \langle \nabla\ell(\theta_t), \Gamma(\pi^\star(\theta_t), \theta_t)\rangle + \varepsilon_t \langle \nabla\ell(\theta(t)), \Gamma(\pi^\star(\theta_t), \theta_t) - \Gamma(p_t, \theta_t)\rangle \\
&\leq -\varepsilon_t \|\nabla\ell(\theta_t)\|^2 + \varepsilon_t \|\nabla\ell(\theta_t)\| \|\Gamma(\pi^\star(\theta_t), \theta_t) - \Gamma(p_t, \theta_t)\| \\
&\leq -\varepsilon_t \|\nabla\ell(\theta_t)\|^2 + \varepsilon_t K_\Gamma \|\nabla\ell(\theta_t)\| \sqrt{\mathrm{KL}(p_t \,||\, \pi^\star(\theta_t))} \\
&\leq -\frac{1}{2}\varepsilon_t \|\nabla\ell(\theta_t)\|^2 + \frac{1}{2}\varepsilon_t K_\Gamma^2 \,\mathrm{KL}(p_t \,||\, \pi^\star(\theta_t))\,.
\end{aligned}$$

Theory: Implicit Diffusion for Langevin diffusions

Coupled differential system

$$\mathrm{d}X_t = -\nabla_1 V(X_t, \theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad X_t \sim p_t, \tag{1}$$

$$\mathrm{d}\theta_t = -\varepsilon_t \Gamma(p_t, \theta_t)\mathrm{d}t\,. \tag{2}$$

- Upper bound on the gradients

$$\|\nabla \ell(\theta_t)\|^2 \leq -\frac{2}{\varepsilon_t}\frac{\mathrm{d}\ell}{\mathrm{d}t}(t) + K_\Gamma^2 \,\mathrm{KL}(p_t \,||\, \pi^\star(\theta_t))\,.$$

Theory: Implicit Diffusion for Langevin diffusions

Coupled differential system

$$\mathrm{d}X_t = -\nabla_1 V(X_t, \theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad X_t \sim p_t, \tag{1}$$

$$\mathrm{d}\theta_t = -\varepsilon_t \Gamma(p_t, \theta_t)\mathrm{d}t\,. \tag{2}$$

- Upper bound on the gradients

$$\|\nabla \ell(\theta_t)\|^2 \leq -\frac{2}{\varepsilon_t}\frac{\mathrm{d}\ell}{\mathrm{d}t}(t) + K_\Gamma^2 \,\mathrm{KL}(p_t \,||\, \pi^\star(\theta_t))\,.$$

- The samples converge to $\pi^\star(\theta_t)$ up to an $\mathcal{O}(\varepsilon_t)$ error

$$\mathrm{KL}(p_t \,||\, \pi^\star(\theta_t)) \leq \mathrm{KL}(p_0 \,||\, \pi^\star(\theta_0))e^{-2\mu t} + 2C^2 \int_0^t \varepsilon_s e^{2\mu(s-t)}\mathrm{d}s\,.$$

Theory: Implicit Diffusion for Langevin Monte Carlo

Coupled updates

$$X_{k+1} = X_k - \eta_k \nabla_1 V(X_k, \theta_k) + \sqrt{2\gamma_k} B_{k+1} \,,$$
$$\theta_{k+1} = \theta_k - \eta_k \varepsilon_k \Gamma(p_k, \theta_k) \,.$$

Theory: Implicit Diffusion for Langevin Monte Carlo

Coupled updates

$$X_{k+1} = X_k - \eta_k \nabla_1 V(X_k, \theta_k) + \sqrt{2\gamma_k} B_{k+1}\,,$$
$$\theta_{k+1} = \theta_k - \eta_k \varepsilon_k \Gamma(p_k, \theta_k)\,.$$

Additional assumptions

- $\nabla_1 V(\cdot, \theta)$ is L_X-Lipschitz for all $\theta \in \mathbb{R}^p$.
- $\nabla_1 V(x, \cdot)$ is L_Θ-Lipschitz for all $x \in \mathbb{R}^d$.
- $\nabla \ell$ is L-Lipschitz.

Theory: Implicit Diffusion for Langevin Monte Carlo

Coupled updates

$$X_{k+1} = X_k - \eta_k \nabla_1 V(X_k, \theta_k) + \sqrt{2\gamma_k} B_{k+1} \,,$$
$$\theta_{k+1} = \theta_k - \eta_k \varepsilon_k \Gamma(p_k, \theta_k) \,.$$

Additional assumptions

- $\nabla_1 V(\cdot, \theta)$ is L_X-Lipschitz for all $\theta \in \mathbb{R}^p$.
- $\nabla_1 V(x, \cdot)$ is L_Θ-Lipschitz for all $x \in \mathbb{R}^d$.
- $\nabla \ell$ is L-Lipschitz.

Theorem

Take $\eta_k = \frac{c_1}{\sqrt{k}}$ and $\varepsilon_k = \frac{1}{\sqrt{k}}$. Then, under the Assumptions,

$$\frac{1}{K} \sum_{k=1}^{K} \|\nabla \ell(\theta_k)\|^2 \leq \frac{c_2 \ln K}{K^{1/3}} \,,$$

where $c_1, c_2 > 0$ depend on the constants of the problem.

Conclusion

- Implicit Diffusion is an algorithm to differentiate through iterative sampling.
- It has applications to Langevin diffusions and denoising diffusions.
- Next steps: find other applications, scale up experiments, refine the theoretical analysis.
- In particular we have guarantees in continuous time for diffusion models based inner loop in the 1D Gaussian case, but the more general case is missing.

Bibliography I

Jordan, R., Kinderlehrer, D., & Otto, F. (1998). The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, *29*(1), 1–17.

Guarantees for diffusion (1D Gaussian case)

1D Gaussian case . Considering $p_{\text{data}} = \mathcal{N}(\theta_{\text{data}}, 1)$ and the forward process , the score is $\nabla \log p_t(x) = -(x - \theta_{\text{data}} e^{-t})$. A natural score function is therefore $s_\theta(x, t) := -(x - \theta e^{-t})$.

With this score function, the output of the sampling process is $\pi^\star(\theta) = \mathcal{N}(\theta(1 - e^{-2T}), 1)$. Remarkably, $\pi^\star(\theta)$ is Gaussian for all $\theta \in \mathbb{R}$, making the analytical study tractable.

Assume that pretraining with samples of p_{data} yields $\theta = \theta_0$, and we want to finetune the model towards some other $\theta_{\text{target}} \in \mathbb{R}$ by optimizing the reward $R(x) = -(x - \theta_{\text{target}})^2$.

A short computation shows that $\nabla \ell(\theta) = -\mathbb{E}_{x \sim \pi^\star(\theta)} R'(x)(1 - e^{-2T})$, hence $\Gamma(p, \theta) = -\mathbb{E}_{x \sim p} R'(x)(1 - e^{-2T})$.

Proposition

Let $(\theta_t)_{t \geq 0}$ be given by the continuous-time equivalent of Implicit diffusion. Then $\|\theta_{2T} - \theta_{\text{target}}\| = \mathcal{O}(e^{-T})$, and $\pi^\star(\theta_{2T}) = \mathcal{N}(\mu_{2T}, 1)$ with $\mu_{2T} = \theta_{\text{target}} + \mathcal{O}(e^{-T})$.