Lecture 16 - Langevin Algorithm

赵尉辰

南开大学 统计与数据科学学院

目录

目录

Langevin Dynamics

定义 1 (Langevin Dynamics)

给定势能函数*(potential)* $V(x)$，*Langevin Dynamics*是如下形式的*SDE*

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t. \tag{1}$$

其解一般称之为*Langevin Diffusion.*

Langevin 扩散的生成元为

$$\mathcal{L}_{LD} f = -\nabla V \cdot \nabla f + \Delta f$$

生成元的伴随为

$$\mathcal{L}^*_{LD} g = \nabla \cdot (g \nabla V) + \Delta g$$

Langevin Dynamics

Kolmogorov backward方程：

$$\frac{\partial}{\partial t} P_t f(x) = \mathcal{L}_{LD} P_t f(x) = -\nabla V(x) \cdot \nabla P_t f(x) + \Delta P_t f(x)$$

Fokker-Planck方程：

$$\partial_t \mu(x,t) = \mathcal{L}^*_{LD} \mu(x,t) = \nabla \cdot (\mu(x,t) \nabla V(x)) + \Delta \mu(x,t)$$

命题 1

Langevin 扩散 $\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$ 的不变测度为

$$d\pi(x) \propto e^{-V(x)} dx$$

Langevin Algorithm

定义 2 (Langevin Algorithm)

对*Langevin*扩散*Euler–Maruyama*离散化：

$$X_{(k+1)h} := X_{kh} - h\nabla V(X_{kh}) + \sqrt{2}(B_{(k+1)h} - B_{kh}).$$

我们得到了一种*Langevin*扩散的实现方式，称为*(Unadjusted) Langevin Algorithm, ULA*/*Langevin Monte Carlo, LMC*. 其中h是迭代步长，k是迭代轮数。

Metropolis-adjusted Langevin Algorithm (MALA)

由于时间离散化，Langevin Monte Carlo与Langevin Dynamics不再一致，Langevin Monte Carlo的平稳分布也不再是目标分布。一般可以通过Metropolis调整保证采样分布的准确性。

Metropolis-Hastings 算法

1. 选择初始状态 x_0。
2. 对于每一步 $n = 1, 2, \ldots, N$:
   - 从提议分布 $q(x'|x_{n-1})$ 中生成候选状态 x'。
   - 计算接受概率：
   $$\alpha = \min\left(1, \frac{\pi(x')q(x_{n-1}|x')}{\pi(x_{n-1})q(x'|x_{n-1})}\right)$$
   - 以概率 α 接受候选状态 x'，否则保持状态 x_{n-1}。

Metropolis-adjusted Langevin Algorithm (MALA)

MALA

- Proposal step: same as in ULA

$$Z_{k+1} = X_k - h\nabla V(X_k) + \sqrt{2h}\xi_k$$

- Accept-reject step: go to

$$X_{k+1} = \begin{cases} Z_{k+1} & \text{with probability } \min\left\{1, \dfrac{\pi(Z_{k+1})\mathcal{P}_{Z_{k+1}}(X_k)}{\pi(X_k)\mathcal{P}_{X_k}(Z_{k+1})}\right\} \\ X_k & \text{with the remaining probability.} \end{cases}$$

注意到给定X_k，提议分布是一个均值为 $X_k - h\nabla V(X_k)$，方差为 $2h\mathbb{I}_n$ 的高斯分布，即提议分布显式表达为

$$\mathcal{P}_z(x) = \frac{1}{(2\pi \cdot 2h)^{\frac{n}{2}}} \exp\left(-\frac{\|x - (z - h\nabla V(z))\|_2^2}{4h}\right).$$

接受概率也具有显式表达：

$$\min\left\{1, \exp\left(-V(z) - \frac{1}{4h}\|x - (z - h\nabla V(z))\|_2^2 + V(x) + \frac{1}{4h}\|z - (x - h\nabla V(x))\|_2^2\right)\right\}$$

目录

耦合方法

定义 3 (Wasserstein distance)

概率测度μ和ν之间的*2-Wasserstein Distance*定义为：

$$W_2(\mu,\nu) := \inf_{\gamma\in\mathcal{C}(\mu,\nu)} \left(\int \|x-y\|^2 \gamma(\mathrm{d}x,\mathrm{d}y)\right)^{\frac{1}{2}}. \tag{2}$$

其中$\mathcal{C}(\mu,\nu)$是μ和ν的耦合*(Couplings)*构成的空间，$\|\cdot\|$是欧式范数。

定理 1

设$\{X_t\}$为初值为$X_0 \sim \mu_0$，平稳分布为$\mu \propto e^{-V}$的*Langevin*扩散，假设μ是α-强*log-concave*的，那么

$$W_2^2(\mu_t,\mu) \le \exp(-2\alpha t) W_2^2(\mu_0,\mu).$$

耦合方法

定理 2

对于 $k \in \mathbb{N}$, 记μ_{kh}为*LMC*的的第*k*轮迭代的分布， $h > 0$为迭代步长。设目标分布为 $\mu \propto e^{-V}$，满足 $\alpha I_d \leq \nabla^2 V \leq \beta I_d$. 如果 $h \lesssim \frac{1}{\beta\kappa}$，那么对于所有 $N \in \mathbb{N}$,

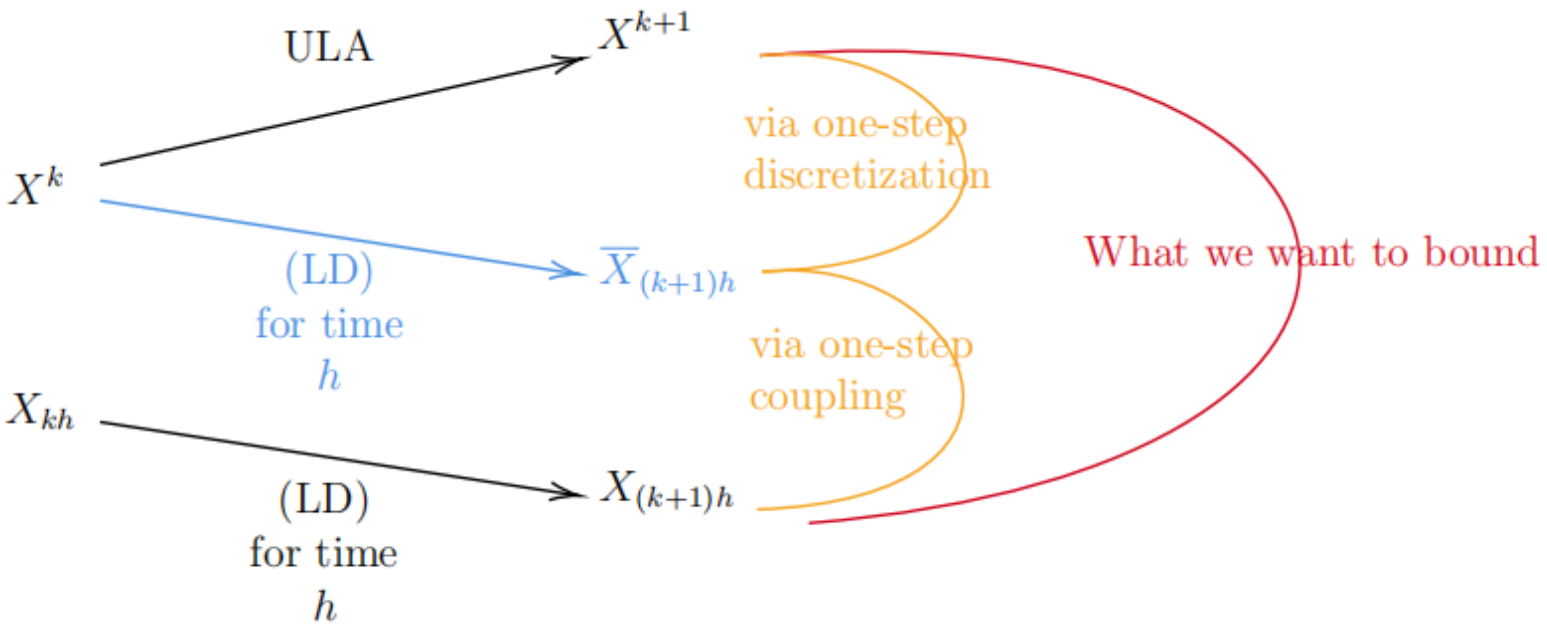
$$W_2(\mu_{Nh}, \mu) \leq \exp\left(-\frac{\alpha N h}{2}\right) W_2(\mu_0, \mu) + O\left(\frac{\beta d^{1/2} h^{1/2}}{\alpha}\right).$$

如果 $h = O(\frac{\varepsilon^2}{\beta\kappa d})$, 那么对于任意 $\varepsilon \in [0, \sqrt{d}]$, 在

$$N = O\Big(\frac{\kappa^2 d}{\varepsilon^2} \log \frac{\sqrt{\alpha} W_2(\mu_0, \mu)}{\varepsilon}\Big)$$

轮迭代之后，有 $\sqrt{\alpha} W_2(\mu_{Nh}, \mu) \leq \varepsilon$.

Proof sketch[1]



- 计算ULA和LD之间的一步时间离散化误差；
- 通过LD在W_2距离下的指数压缩性分析多步迭代误差。

[1] https://chewisinho.github.io/main.pdf Sec.4.1

泛函不等式方法[2]

耦合方法的分析要求目标分布的log-concavity，泛函不等式方法可以减弱这一假设。

定义 4 (Log-Sobolev inequality)

称ν满足α *Log-Sobolev inequality*，如果对于$\mathbb{E}_\nu[g^2] < \infty$的光滑函数$g : \mathbb{R}^n \to \mathbb{R}$，

$$\mathrm{Ent}_\nu(g) \triangleq \mathbb{E}_\nu[g^2 \log g^2] - \mathbb{E}_\nu[g^2] \log \mathbb{E}_\nu[g^2] \le \frac{2}{\alpha}\mathbb{E}_\nu[\|\nabla g\|^2]. \tag{3}$$

定义 5 (Poincaré inequality)

称ν满足α *Poincaré inequality*，如果对于光滑函数$g : \mathbb{R}^n \to \mathbb{R}$, 有

$$\mathrm{Var}_\nu(g) \triangleq \mathbb{E}_\nu[g^2] - \mathbb{E}_\nu[g]^2 \le \frac{1}{\alpha}\mathbb{E}_\nu[\|\nabla g\|^2]. \tag{4}$$

[2]Bakry D, Gentil I, Ledoux M. Analysis and geometry of Markov diffusion operators[M]. Cham: Springer, 2014.

KL散度

定义 6 (KL散度)

μ对于ν的*KL*散度定义为

$$\mathrm{KL}(\mu\|\nu) = H_\nu(\mu) = \int_{\mathbb{R}^n} \mu(x) \log \frac{\mu(x)}{\nu(x)} dx. \tag{5}$$

命题 2 (Pinsker's inequality)

$$\mathrm{d_{TV}}(\mu,\nu)^2 \le \tfrac{1}{2} H_\nu(\mu).$$

命题 3 (Talagrand inequality)

若ν满足α *Log-Sobolev inequality*

$$\tfrac{\alpha}{2} W_2(\mu,\nu)^2 \le H_\nu(\mu).$$

Pinsker's inequality和Talagrand inequality说明KL散度是一个相对更强的距离度量，我们bound KL散度自然能够给出TV和W_2距离的界。

KL散度+LSI下Langevin Dynamics的指数收敛性

定理 3

若$\nu := e^{-f}$满足α *LSI*，那么*Langevin Dynamics*

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dW_t$$

的分布μ_t满足：

$$H_\nu(\mu_t) \leq e^{-2\alpha t} H_\nu(\mu_0).$$

进一步地，$W_2(\mu_t, \nu) \leq \sqrt{\frac{2}{\alpha} H_\nu(\mu_0)} e^{-\alpha t}$.

Proof sketch[3]

定义 7 (Fisher information)

μ对于ν的*Fisher information*定义为

$$J_\nu(\mu) = \int_{\mathbb{R}^n} \mu(x) \left\| \nabla \log \frac{\mu(x)}{\nu(x)} \right\|^2 dx. \tag{6}$$

令$g^2 = \frac{\mu}{\nu}$，Log-Sobolev inequality可以得到KL散度和Fisher information的如下关系：

$$H_\nu(\mu) \leq \frac{1}{2\alpha} J_\nu(\mu).$$

[3]Vempala S, Wibisono A. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices[J]. Advances in neural information processing systems, 2019, 32.

Proof sketch

引理 1

分布μ_t满足：

$$\frac{d}{dt}H_\nu(\mu_t) = -J_\nu(\mu_t). \tag{7}$$

利用*Langevin Dynamics*的*Fokker-Planck*方程：$\partial_t \mu_t = \nabla \cdot (\mu_t \nabla V(x)) + \Delta \mu_t$ 计算可得。

由Log-Sobolev inequality，

$$H_\nu(\mu) \leq \frac{1}{2\alpha} J_\nu(\mu).$$

结合(7)式，有

$$\frac{d}{dt}H_\nu(\mu_t) \leq -2\alpha H_\nu(\mu_t)$$

两边积分有

$$H_\nu(\mu_t) \leq e^{-2\alpha t} H_\nu(\mu_0).$$

KL散度+LSI下LMC的指数收敛性[4]

定理 4

若$\nu := e^{-V}$满足α *LSI* 并且是 *L-smooth*的$(-LI \preceq \nabla^2 V(x) \preceq LI \text{ for all } x \in \mathbb{R}^n)$，那么对于任意$x_0 \sim \mu_0$满足$H_\nu(\mu_0) < \infty$，步长$0 < \eta \leq \frac{\alpha}{4L^2}$的*ULA*

$$x_{k+1} = x_k - \eta \nabla V(x_k) + \sqrt{2\eta} z_k$$

的分布$x_k \sim \mu_k$满足：

$$H_\nu(\mu_k) \leq e^{-\alpha \eta k} H_\nu(\mu_0) + \frac{8\eta n L^2}{\alpha}.$$

因此，对任意精度$\delta > 0$，为了$H_\nu(\mu_k) < \delta$，*LMC*需要满足步长$\eta \leq \frac{\alpha}{4L^2} \min\{1, \frac{\delta}{4n}\}$，并且经过$k \geq \frac{1}{\alpha\eta} \log \frac{2H_\nu(\mu_0)}{\delta}$次迭代。

[4]Vempala S, Wibisono A. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices[J]. Advances in neural information processing systems, 2019, 32.

Proof sketch

- 给出一步LMC迭代的界

引理 2

若$\nu := e^{-V}$满足α *Log-Sobolev inequality* 并且 *L-smooth*，步长$0 < \eta \leq \frac{\alpha}{4L^2}$，那么*LMC*满足：

$$H_\nu(\mu_{k+1}) \leq e^{-\alpha\eta} H_\nu(\mu_k) + 6\eta^2 n L^2.$$

$$\frac{d}{dt} H_\nu(\mu_t) \leq -\frac{3}{4} J_\nu(\mu_t) + \frac{4t^2 L^4}{\alpha} H_\nu(\mu_0) + 2t^2 n L^3 + 2tnL^2.$$

由Log-Sobolev inequality

$$\frac{d}{dt} H_\nu(\mu_t) \leq -\frac{3\alpha}{2} H_\nu(\mu_t) + \frac{4t^2 L^4}{\alpha} H_\nu(\mu_0) + 2t^2 n L^3 + 2tnL^2.$$

$t = 0$到$t = \eta$积分，整理可得引理2。

- 给出多步迭代的界

Rényi散度+PI下Langevin Dynamics的指数收敛性

定义 8 (Rényi散度)

对于$q > 0,\ q \neq 1$，概率分布μ对于ν的q阶*Rényi*散度定义为：

$$R_{q,\nu}(\mu) := \frac{1}{q-1} \log F_{q,\nu}(\mu), \tag{8}$$

其中

$$F_{q,\nu}(\mu) := \mathbb{E}_\nu \left[\left(\frac{\mu}{\nu} \right)^q \right] = \int_{\mathbb{R}^n} \nu(x) \frac{\mu(x)^q}{\nu(x)^q} dx = \int_{\mathbb{R}^n} \frac{\mu(x)^q}{\nu(x)^{q-1}} dx.$$

Rényi散度来源于Rényi熵：$H_q(\mu) := \frac{1}{q-1} \log \int \mu(x)^q dx.$

定理 5

若$\nu := e^{-f}$满足α *Poincaré inequality*，$q \geq 2$，那么*Langevin Dynamics*的分布μ_t满足：

$$R_{q,\nu}(\mu_t) \leq \begin{cases} R_{q,\nu}(\mu_0) - \frac{2\alpha t}{q} & if R_{q,\nu}(\mu_0) \geq 1 \text{ and as long as } R_{q,\nu}(\mu_t) \geq 1, \\ e^{-\frac{2\alpha t}{q}} R_{q,\nu}(\mu_0) & if R_{q,\nu}(\mu_0) \leq 1. \end{cases}$$

Rényi散度+PI下LMC的指数收敛性

定理 6

若ν_η满足β *Poincaré inequality*，$q \geq 1$，$\nu := e^{-V}$是 *L-smooth*的，且$1 \leq R_{2q,\nu_\eta}(\mu_0) < \infty$，令$0 < \eta \leq \min\left\{\frac{1}{3L}, \frac{1}{9\beta}\right\}$，$q > 1$，那么对于$k \geq k_0 := \frac{2q}{\beta\eta}(R_{2q,\nu_\eta}(\mu_0) - 1)$, *LMC*满足：

$$R_{q,\nu}(\mu_k) \leq \left(\frac{q-\frac{1}{2}}{q-1}\right) e^{-\frac{\beta\eta(k-k_0)}{2q}} + R_{2q-1,\nu}(\nu_\eta).$$

对任意精度$\delta > 0$，为了$R_{q,\nu}(\mu_k) \leq \delta$, *LMC*需要满足步长$\eta = \Theta\left(\min\left\{\frac{1}{L}, \gamma_{2q-1}\left(\frac{\delta}{2}\right)\right\}\right)$，其中$\gamma_q(\delta) = \sup\{\eta > 0 \colon R_{q,\nu}(\nu_\eta) \leq \delta\}$, 并且经过$k = \Theta\left(\frac{1}{\beta\eta}\left(R_{2q,\nu_\eta}(\mu_0) + \log\frac{1}{\delta}\right)\right)$次迭代。

目录

SMLD

Recall that Langevin dynamics

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

具有不变测度$\pi \propto e^{-V}$, 若我们需要采样p_{data}，可以通过

$$\mathrm{d}X_t = \nabla \log p_{data}(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t.$$

其中$\nabla \log p$称为概率分布p的Score function.

如果p_{data}已知，那么可以显式计算Score，然而生成任务中，我们需要从数据中学习Score $\nabla \log p_{data}$，通过神经网络近似

$$\min_{\theta} \mathbb{E}[\|\nabla \log p_{data}(X) - s_\theta(X)\|_2^2]$$

其中s_θ为参数为θ的神经网络。

Score Matching

- Score matching[5]

$$\mathbb{E}_{X\sim p_{data}} \left\|\nabla \log p_{data}(X) - s_\theta(X)\right\|_2^2$$
$$= \underbrace{\mathbb{E}\left\|\nabla \log p_{data}(X)\right\|_2^2}_{\text{does not depend on } \theta} - 2\mathbb{E}\left\langle s_\theta(X), \nabla \log p_{data}(X)\right\rangle + \mathbb{E}\left\|s_\theta(X)\right\|_2^2 .$$

计算第二项

$$-\mathbb{E}\left\langle s_\theta(X), \nabla \log p_{data}(X)\right\rangle = -\int \left\langle s_\theta(x), \nabla \log p_{data}(x)\right\rangle p_{data}(x)dx$$
$$= \int \nabla \cdot s_\theta(x) p_{data}(x)dx = \mathbb{E}\nabla \cdot s_\theta(X),$$

那么我们的优化问题即为：

$$\min_\theta \mathbb{E}_{X\sim p_{data}} \left[\left\|s_\theta(X)\right\|_2^2 + 2\nabla \cdot s_\theta(X)\right] .$$

[5]Hyvärinen A, Dayan P. Estimation of non-normalized statistical models by score matching[J]. Journal of Machine Learning Research, 2005, 6(4).

Denoising score matching

实际训练神经网络优化经验风险函数：

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \left[\|s_\theta(x_i)\|_2^2 + 2\nabla \cdot s_\theta(x_i) \right].$$

然而高维情形计算散度项$\nabla \cdot s_\theta(x_i)$比较困难，考虑通过Denoising score matching避免散度的计算。

- Denoising score matching[6]

考虑扰动$\tilde{x} = x + \sigma z$, 其中$z \sim N(0, I)$，Denoising score matching的目标为

$$\min_{\theta} \mathbb{E}_{q_\sigma(\tilde{x}|x)p_{data}(x)}[\|s_{\boldsymbol{\theta}}(\tilde{x}) - \nabla_{\tilde{x}} \log q_\sigma(\tilde{x} \mid x)\|_2^2].$$

可以证明$s_{\theta^*}(\tilde{x}) = \nabla_{\tilde{x}} \log q_\sigma(\tilde{x})$几乎处处成立[7]，其中$q_\sigma(\tilde{x}) \triangleq \int q_\sigma(\tilde{x} \mid x) p_{\text{data}}(x) \mathrm{d}x$.

[6] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In Advances in Neural Information Processing Systems, pp. 11895–11907, 2019.

[7] Vincent P. A connection between score matching and denoising autoencoders[J]. Neural computation, 2011, 23(7): 1661-1674.

Denoising score matching

虽然只有在σ比较小的时候，有

$$s_{\theta^*}(\tilde{x}) = \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}) \approx \nabla_x \log p_{\text{data}}(x)$$

但是$q_\sigma(\tilde{x} \mid x) \sim \mathcal{N}(x, \sigma^2 I)$是条件高斯的在计算上十分高效

$$\begin{aligned}
\nabla_{\tilde{x}} \log q_\sigma(\tilde{x} \mid x) &= \nabla_{\tilde{x}} \log \frac{1}{(\sqrt{2\pi\sigma^2})^d} \exp\left\{-\frac{\|\tilde{x} - x\|^2}{2\sigma^2}\right\} \\
&= \nabla_{\tilde{x}} \left\{-\frac{\|\tilde{x} - x\|^2}{2\sigma^2} - \log(\sqrt{2\pi\sigma^2})^d\right\} \\
&= -\frac{\tilde{x} - x}{\sigma^2}.
\end{aligned}$$

Denoising score matching的优化问题即为：

$$\min_\theta \mathbb{E}_{x \sim p_{data}, \tilde{x} \sim \mathcal{N}(x, \sigma^2 I)} \left\| s_\theta(\tilde{x}, \sigma) + \frac{\tilde{x} - x}{\sigma^2} \right\|_2^2. \tag{9}$$

Noise Conditional Score Networks

(9)中参数化的神经网络模型$s_\theta(x,\sigma)$称为 Noise Conditional Score Networks。由于只有在σ比较小的时候，有

$$s_{\theta^*}(x,\sigma) \approx \nabla_x \log p_{\text{data}}(x)$$

考虑设计一个time schedule, 使得$\sigma_t \to 0$.

Algorithm 1 Annealed Langevin dynamics.

Require: $\{\sigma_i\}_{i=1}^L, \epsilon, T$.

Initialize $\tilde{\mathbf{x}}_0$
for $i \leftarrow 1$ to L **do**
  $\alpha_i \leftarrow \epsilon \cdot \sigma_i^2/\sigma_L^2$ ▷ α_i is the step size.
  for $t \leftarrow 1$ to T **do**
    Draw $\mathbf{z}_t \sim \mathcal{N}(0, I)$
    $\tilde{\mathbf{x}}_t \leftarrow \tilde{\mathbf{x}}_{t-1} + \frac{\alpha_i}{2}\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_{t-1}, \sigma_i) + \sqrt{\alpha_i}\,\mathbf{z}_t$
  end for
  $\tilde{\mathbf{x}}_0 \leftarrow \tilde{\mathbf{x}}_T$
end for
return $\tilde{\mathbf{x}}_T$

Thanks & Questions