Sampling Techniques - Continued.

02/02/2025

* ## Inverse Transform Sampling.

Method for generating random samples from a Probability distribution $P(X)$ using cumulative distribution function (CDF). It relies on probability integral transform & its inverse.

$X \rightarrow$ R.V. with a continuous CDF $F_X(x)$, then R.V. $Y = F_X(x) \rightarrow$ uniformly distributed on $[0,1]$.

$$Y = F_X(x) \sim U[0,1]$$

CDF $F_X(x)$ maps values of X to prob. in $[0,1]$.

$Y \sim U[0,1]$, R.V. $\hat{X} = F_X^{-1}(Y)$ has the same dist as X :-

$$\hat{X} = F_X^{-1}(Y) \sim X$$

Inverse CDF $F_X^{-1}(y)$ maps prob. $y \in [0,1]$ back to values of X.

Applying F_X^{-1} to uniformly distributed samples Y, we obtain $\hat{X}$ that follows the dist. of X.

__Algorithm__

- Compute CDF $F_X(x)$

  eg :- $P(x)$ - target Dist, then. CDF

$$\boxed{F_X(x) = \int_{-\infty}^{x} P(t)\,dt.}$$

- Compute the inverse CDF $F_X^{-1}(y)$
  - $F_X^{-1}(y)$ maps $y \in [0,1]$ to x.

- Generate uniform samples $Y \sim U[0,1]$
  using random no. generator from computer.

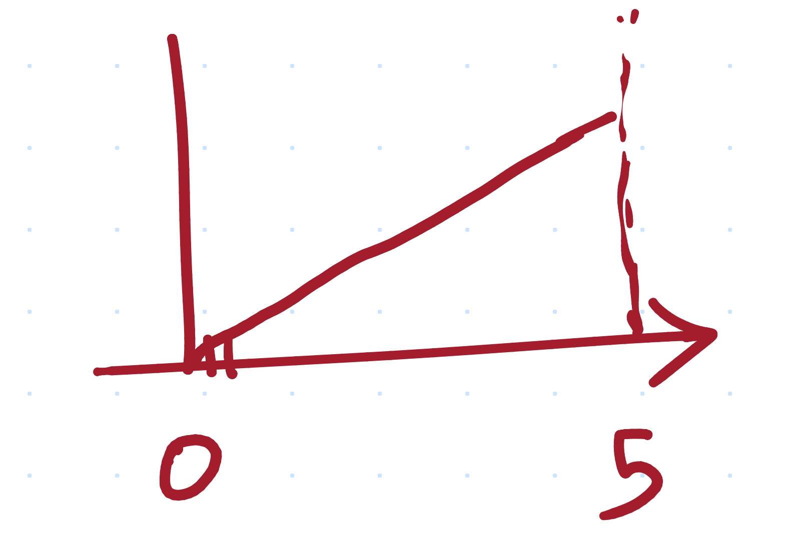
- Transform Y to $\hat{X}$.
  - Apply the inverse CDF to the uniform samples.

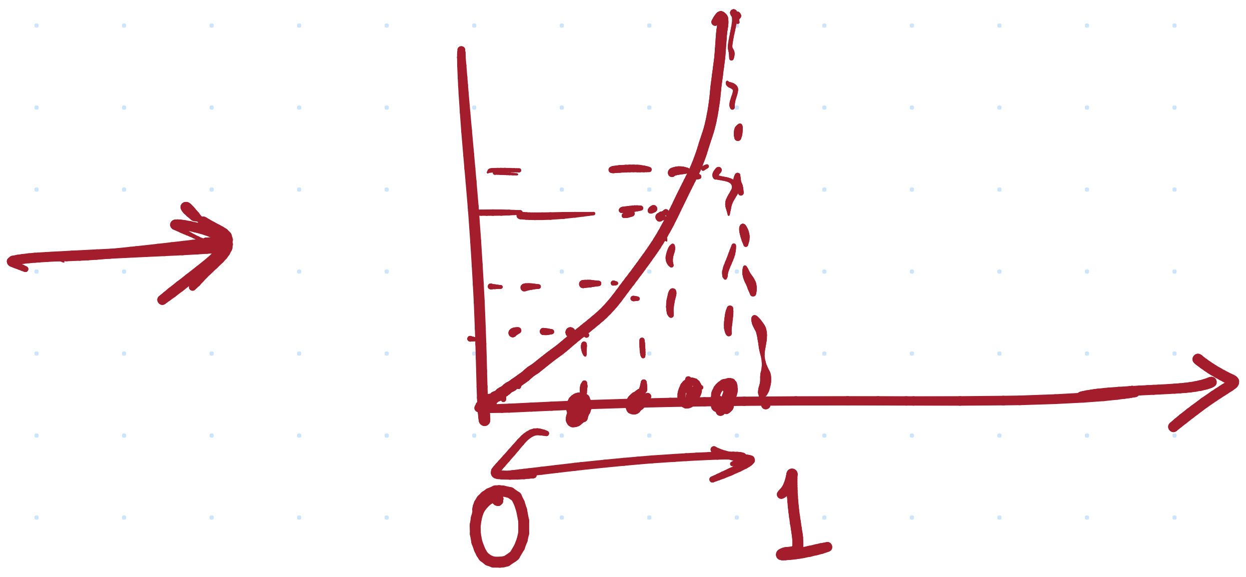
$$\hat{X} = F_X^{-1}(y).$$

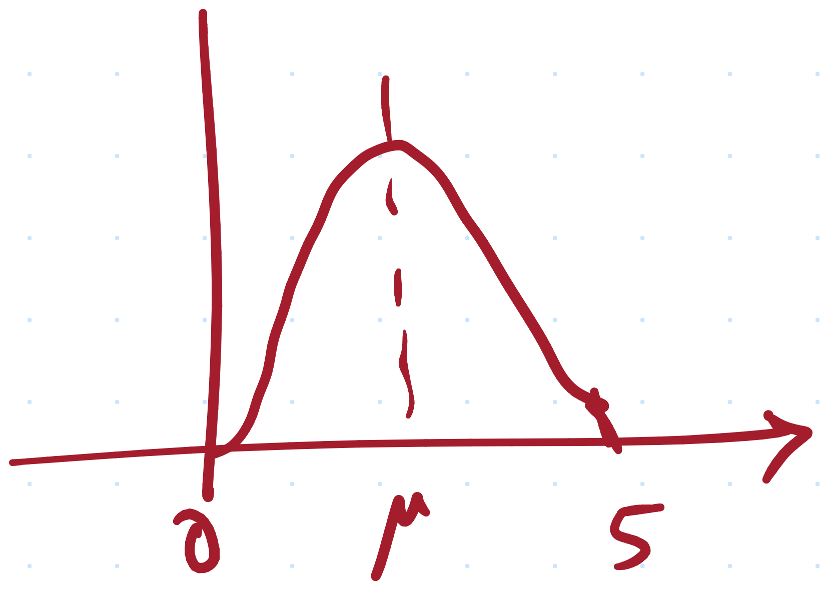
  - $\hat{x}$ will produce the target dist. $P(x)$.

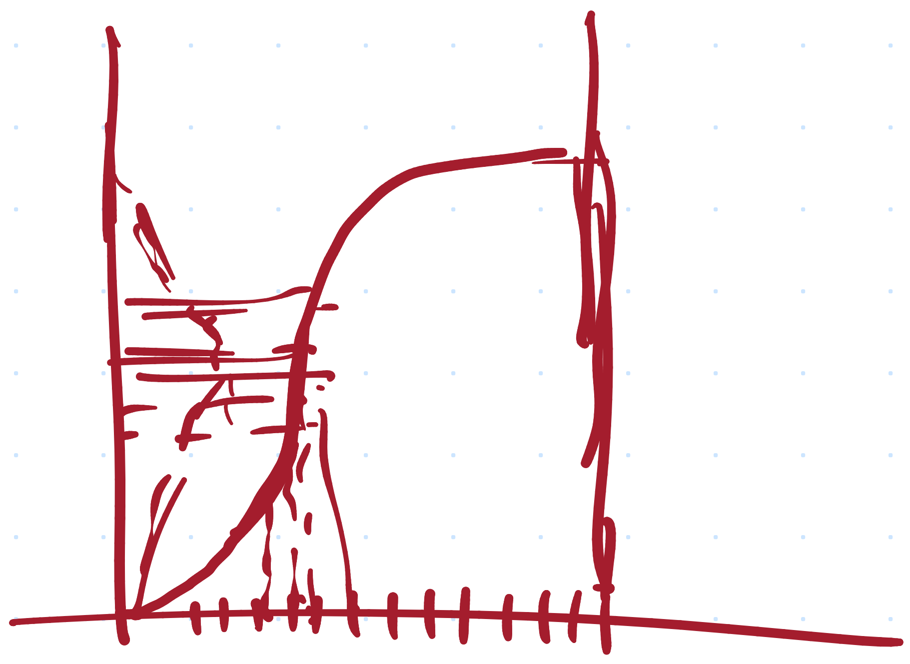
* Works with any dist. with a computable inverse CDF. $F_X^{-1}(y)$
* Computationally efficient, $F_X^{-1}(y)$ - easy to compute
* Dist's without closed form inverse CDF, we use numerical methods

* $F_X(x) \rightarrow$ uniformly dist, $F_X^{-1}(y)$ transforms uniform samples into samples from $P(x)$.

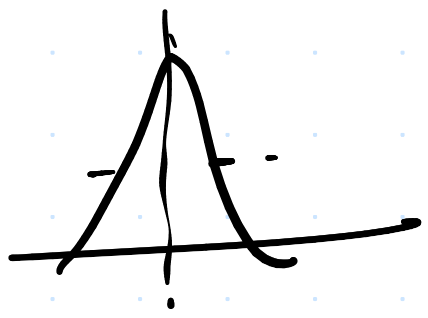








Cauchy Distribution



PDF: Cauchy $(x_0, \gamma) = \dfrac{1}{\pi\gamma\left[1 + \left(\frac{x - x_0}{2}\right)^2\right]}$

peak of the dist $\leftarrow x_0 \rightarrow$ location parameter

half width at half max $\rightarrow \gamma > 0 \rightarrow$ scale parameter

CDF:

$$P(X \le x) = \frac{1}{2} + \frac{1}{\pi} \tan^{-1}\left(\frac{x - x_0}{\gamma}\right)$$

Cauchy (0,1)

$$F_X^{-1}(y) = \tan\left[\pi\left(y - \frac{1}{2}\right)\right]$$

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\pi\gamma\left[1+\left(\frac{t-x_0}{\gamma}\right)^2\right]} \cdot dt$$

$$u = \frac{t-x_0}{\gamma} \qquad du = \frac{1}{\gamma}\,dt \;, \quad dt = \gamma\, du.$$

$$t \to -\infty \;, \quad u \to -\infty$$

$$t \to x \;, \quad u \to \frac{x-x_0}{\gamma}$$

$$\int_{-\infty}^{\frac{x-x_0}{\gamma}} \frac{1}{\pi\cancel{\gamma}\left[1+u^2\right]}\,\cancel{\gamma}\,du$$

$$\Rightarrow \frac{1}{\pi}\int_{-\infty}^{\frac{x-x_0}{\gamma}} \frac{1}{1+u^2}\,du = \frac{1}{\pi}\left[\tan^{-1}u\right]_{-\infty}^{\frac{x-x_0}{\gamma}}$$

$$= \frac{1}{\pi}\left[\tan^{-1}\left(\frac{x-x_0}{\gamma}\right) - \left(-\frac{\pi}{2}\right)\right]$$

as $u \to -\infty$

$\tan^{-1}u \to -\frac{\pi}{2}$

$u \to \frac{x-x_0}{\gamma}$

$\tan^{-1}u \to \tan^{-1}\left(\frac{x-x_0}{\gamma}\right)$

$$= \frac{1}{\pi}\left[\frac{\pi}{2} + \tan^{-1}\left(\frac{x-x_0}{\gamma}\right)\right]$$

$$F_X(x) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}\left(\frac{x-x_0}{\gamma}\right)$$

Cauchy $(0,1)$ $\rightarrow$

$\downarrow$ $\downarrow$

x y

$$F_X(x) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(x)$$

$$y = F_X(x) \qquad F_X^{-1}(y) = ?$$

$$\Rightarrow y = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(x)$$

$$\Rightarrow y - \frac{1}{2} = \frac{1}{\pi}\tan^{-1}(x)$$

$$\Rightarrow \tan^{-1}(x) = \left(y - \frac{1}{2}\right)\pi$$

$$\Rightarrow x = \tan\left(\pi\left(y - \frac{1}{2}\right)\right)$$

$$\boxed{x = F_X^{-1}(y) = \tan\left(\pi\left(y - \frac{1}{2}\right)\right)}$$

Exponential Dist :-

The PDF is given by :-

(a) $p(t) = Exp(\lambda) = \lambda e^{-\lambda x}$, $\forall\, x \in [0, \infty)$

(b) $F_X(x) = P(X \le x) = 1 - e^{-\lambda x}$

(c) $F_X^{-1}(y) = -\log_e(1-y) \rightarrow$ for Exp(1)

$$F_X(x) = \int_{-\infty}^{x} p(t)\,dt = \int_{-\infty}^{x} \lambda e^{-\lambda t}\,dt$$

$$= \lambda \int_{0}^{x} e^{-\lambda t}\,dt = \lambda \left[\frac{e^{-\lambda t}}{-\lambda}\right]_0^x$$

$$= -1\left[e^{-x\lambda} - e^{0}\right]$$

$$= 1 - e^{-\lambda x} + e$$

$$F_X(x) = 1 - e^{-\lambda x} = y$$

$$y = 1 - e^{-\lambda x}$$

$$(y-1) = -e^{-\lambda x}$$

$$(1-y) = e^{-\lambda x}$$

$$\log_e(1-y) = -\lambda x$$

$\log_e e = ?\ 1$

$\ln$

$\boxed{Exp(1)} = ?$

λ

$$\boxed{x = -\frac{1}{\lambda}\log_e(1-y).}$$

$$\boxed{x = -\log_e(1-y)}$$

Gumbel Distribution

PDF $\quad Gumbel(\mu, \beta) = \dfrac{e^{-(z+e^{-z})}}{\beta}$

$z = \dfrac{x-\mu}{\beta}$

CDF $\quad P(X \leq x) = e^{-e^{-\frac{(x-\mu)}{\beta}}}$

$\mu \to$ location parameter

$\beta > 0$, scale param.

(For Gumbel $(0,1) \Rightarrow F_X^{-1}(y) = -\log_e(-\log_e(y))$)

wikipedia.

$\mu \quad \beta$

$$F_X(x) = \int_{-\infty}^{x} p(t)\, dt$$

X

$$= \int_{-\infty}^{x} \frac{e^{-(z+e^{-z})}}{\beta}\, dt$$

$$z = \frac{t-\mu}{\beta}$$

$$= \int_{-\infty}^{\frac{t-\mu}{\beta}} \frac{e^{-(z+e^{-z})}}{\cancel{\beta}} \cancel{\beta} dz \quad \left| \begin{array}{l} t \to \infty,\ z \to \infty \\ t \to x,\ z \to \frac{x-\mu}{\beta} \end{array} \right|$$

$$dz = \frac{dt}{\beta}$$

$$dt = \beta dz.$$

$$F_X(x) = \int_{-\infty}^{\frac{x-\mu}{\beta}} e^{-(z+e^{-z})}\, dz \quad ?$$

$$\ln u = -z,$$

$$u = e^{-z} \qquad z = -\ln u$$

$$du = e^{-z}(-1)\, dz$$

$$\boxed{dz = -\frac{1}{u}\, du}$$

$$\left\{ \begin{array}{l} z = -\infty,\ u = \frac{1}{e^z} = \frac{1}{e^{-\infty}} = e^{\infty} = \infty \\ u \to \infty \end{array} \right|$$

$$z \to \frac{x-\mu}{\beta}$$

$$u \to e^{-\frac{x-\mu}{\beta}}.$$

$$F_X(x) = \int_{\infty}^{e^{\frac{x-\mu}{\beta}}} e^{-\left(-\ln(u) + u\right)} \left(-\frac{1}{u}\right) du$$

messed up here!!

$$\int_{\infty}^{e^{-\frac{x-\mu}{\beta}}} \left(u + e^{-u}\right)\left(-\frac{1}{u}\right) du$$

Homework

$$F_X(x) = \int_{e^{-\frac{x-\mu}{\beta}}}^{\infty} e^{-u}\, du = \left[-e^{-u}\right]_{e^{-\frac{x-\mu}{\beta}}}^{\infty}$$

$u \to \infty,\ -e^{-u} \to 0$

$u = e^{-\frac{(x-\mu)}{\beta}},\ -e^{-u}$

$\longrightarrow -e^{-e^{-\frac{(x-\mu)}{\beta}}}$

$$F_X(x) = e^{-e^{-\frac{(x-\mu)}{\beta}}}$$

$$\Rightarrow \quad e^{-e^{-\frac{(x-\mu)}{\beta}}} = y$$

$$\Rightarrow \quad -e^{-\frac{(x-\mu)}{\beta}} = \ln y$$

$$\Rightarrow \quad -\frac{x-\mu}{\beta} = \ln(-\ln(y))$$

$$\Rightarrow \quad \frac{x-\mu}{\beta} = -\ln(-\ln(y))$$

$$\Rightarrow \quad x = -\underset{\downarrow\\1}{\beta}\ln(-\ln(y)) + \underset{\downarrow\\0}{\mu}.$$

$\mu = 0, \ \beta = 1$

$$F_X^{-1}(y) = \boxed{x = -\ln(-\ln(y))}$$

Langevin Monte Carlo

Stochastic Gradient Langevin Dynamics (SGLD) is a Markov chain Monte Carlo (MCMC) method used to sample from a target dist.

It combines gradient based optimization and stochastic differential equations.

Goal: Sample from a target dist $P(X)$ with probability density $e^{-U(x)}$, $U(x) \rightarrow$ potential function.

Simulates a stochastic process, Langevin SDE that explores space in a way that converges to a target dist.

$$\left\{ \underline{\underline{dX_t}} = -\underline{\nabla U(X_t)\, dt} + \underline{\underline{\sqrt{2}\, dB_t}} \right\}$$

$X_t \rightarrow$ State of the system at time t

$U(X_t) \rightarrow$ potential function related to the target distribution.

$\nabla U(X_t) \rightarrow$ gradient of potential function.

$dB_t \rightarrow$ increment of a standard Brownian Motion (Wiener Process)

$\sqrt{2} \rightarrow$ scaling factor for the noise.

$-\nabla U(X_t)dt \rightarrow$ drift term, pushes X_t towards regions of lower potentials, higher probability under $P(X)$.

$\sqrt{2}\,dB_t \rightarrow$ diffusion term $\rightarrow$ add random noise, to mitigate stucking in local minima.

$P(X) \propto e^{-U(X)} \rightarrow$ Stationary dist is the target dist.

After running for a long time, the samples X_t will be distributed according to $P(X)$.

Discretized to generate samples :-

$$\left\{ X_{t+1} = X_t - \eta \nabla U(X_t) + \sqrt{2\eta}\, Z_t \right\}$$

$\eta \rightarrow$ step size / learning rate

$Z_t \rightarrow$ standard normal R.V. $Z_t \sim N(0, I)$

$-\eta \nabla U(X_t) \rightarrow$ performs a gradient step towards regions of lower potential.

$\sqrt{2\eta}\, Z_t \rightarrow$ adds Gaussian noise to the exploration.

Algorithm :-

Start with initial point X_0.

For each step:-

- Compute gradient $\nabla U(X_t)$
- update

$$X_{t+1} = X_t - \eta \nabla U(X_t) + \sqrt{2\eta}\, Z_t$$

- Store X_{t+1} as a sample.

Example :- — Sampling from a Gaussian dist.

target → Gaussian $N(\mu, \sigma^2)$

$$U(x) = \frac{(x-\mu)^2}{2\sigma^2} \qquad \propto e^{-U(x)}$$

$$\cancel{\frac{1}{\sqrt{2\pi\sigma^2}}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\nabla U(x) = \frac{x-\mu}{\sigma^2}$$

Langevin Update :-

$$x_{t+1} = x_t - \eta \frac{x-\mu}{\sigma^2} + \sqrt{2\eta}\, z_t$$

$$U(z) = \underbrace{\frac{1}{2}\left(\frac{\|z\| - 2}{0.4}\right)^2} - \log\left(e^{-0.5\left[\frac{z_1-2}{0.6}\right]} + e^{-0.5\left[\frac{z_1+2}{0.6}\right]}\right)$$

$$p(z) \propto e^{-U(z)}$$

$$\|z\| = \sqrt{z_1^2 + z_2^2}$$ → Euclidean norm of z

$\frac{1}{2}\left(\frac{\|z\|-2}{0.4}\right)^2$ → quadratic penalty that encourages $\|z\|$ to be close to 2

$-\log(\cdots)$ → log-sum-exp → creates two modes, $z_1 = 2$, $z_1 = -2$

Target dist $= p(z)$

$$p(z) \propto e^{-U(z)}$$

$$p(z) \propto \underbrace{\exp\left(-\frac{1}{2}\left(\frac{\|z\|-2}{0.4}\right)^2\right)}_{\substack{\text{ring shaped gaussian} \\ \text{centred at } \|z\|=2 \\ \text{\& radius}=2 \text{ (encourages)}}} \cdot \left(e^{-0.5\left[\frac{z_1-2}{0.6}\right]} + e^{-0.5\left[\frac{z_1+2}{0.6}\right]}\right)$$

$\downarrow$

two modes

$z_1 = 2$, $z_1 = -2$

(peaks)

symmetric abt. $z_1 = 0$.

To sample from $p(z)$, we use langevin Monte Carlo or other MCMC methods.

$$\nabla U(z) = \begin{bmatrix} \frac{\partial U}{\partial z_1} \\ \frac{\partial U}{\partial z_2} \end{bmatrix}$$

$P(x) \propto e^{-U(x)}$ → Related to energy based models, states with lower energy are more likely (higher probability) while states with higher energy are less likely (lower probability).

$P(x) \propto e^{-0} \propto 1$. $\quad$ $U(x) \to 0$ lower energy

$P(x) \propto e^{-\infty} \propto \frac{1}{e^{\infty}} \propto 0$. $\quad$ $U(x) \to \infty$ → less likely higher energy

$P(x)$ = always non-ve, the dist is normalized.

$$P(x) = \frac{1}{Z} e^{-U(x)} \qquad Z = \int e^{-U(x)} dx$$

↓

This satisfies the requirements of Boltzmann distribution → lower energy states are more probable, consistent with laws of thermodynamics, statistical mechanics. Arises naturally from max. entropy.

Euler - Maruyama Discretization (ULA)

Unadjusted Langevin Algorithm

$$X_{k+1} = X_k - \gamma_{k+1} \nabla U(X_k) + \sqrt{2\gamma_{k+1}}\, Z_{k+1}$$

$X_k \rightarrow$ state at the k-th iteration

$\gamma_{k+1} \rightarrow$ step size at k+1 th iteration.

$Z_{k+1} \rightarrow$ standard normal R-V.

$Z_{k+1} \sim N(0, I)$

$-\gamma_{k+1} \nabla U(X_k)$ – performs a gradient descent step towards regions of lower potential

$\sqrt{2\gamma_{k+1}}\, Z_{k+1} \rightarrow$ Adds Gaussian noise to ensure exploration.

step size $\gamma_{k+1} \leftarrow$ controls trade off b/w exploitation (small $\underline{\underline{\gamma_{k+1}}}$ to move towards regions of lower potential) and exploration (to add noise to explore the space, i.e., large $\underline{\underline{\gamma_{k+1}}}$).

In practice, a constant step size is often used.

Decreasing step size $\gamma_{k+1} = \frac{c}{k+1}$ → can improve convergence but requires careful tuning.

X_0 → typically sampled from a simple dist.

initial state $X_0 \sim N(0, I)$ (standard $\underline{\text{normal}}$)

ULA assumes : $\nabla U(x)$ is $\underline{\text{L- Lipschitz}}$ $\underline{\underline{\text{continuous}}}$.

$$\underline{\|\nabla U(x) - \nabla U(y)\|} \leq \underline{\underline{L\|x-y\|}} \quad \underline{\underline{\forall x, y}}$$

this assume the gradient doesn't change too rapidly which is necessary for the convergence of Euler-Maruyama scheme

$U(x) \rightarrow$ smooth & differentiable.

<u>Sampling</u> :- ULA can be used to sample from target dist. of the form

$$p(x) = \frac{e^{-U(x)}}{Z}$$

$Z \rightarrow$ normalizing constant which may/maynot be known.

Algorithm :-

Initialize :-

- Start with an initial state X_0

Iterate :-

For each iteration k :-

- Compute the gradient $\nabla U(X_k)$
- Propose a new state :-

$$X_{k+1} = X_k - \gamma_{k+1} \nabla U(X_k) + \sqrt{2 \gamma_{k+1}}\, Z_{k+1}$$

- Store x_{k+1} as sample.

- Burn-in - discard the first few samples to allow the chain to converge

* ULA doesn't needs Metropolis-Hastings acceptance step, unlike-MALA. Hence the samples may have some bias due to discretization error.

* ULA- simpler & computationally cheaper than MALA.