Stochastic Gradient Langevin Dynamics (SGLD)

-Summary-

Introduction

- Suggested Problem :
  - typical MCMC requires computations over the whole dataset $X = \{x_i\}_{i=1}^{N}$, which is intractable as data set size gets bigger.

- Suggested Solution:
  - Combine Robbins-Monro algorithm (~SGD R.V version) with Langevin dynamics (~noise injection).
  - Resulting algorithm smoothly transitions from stochastic optimization to sampling from the posterior using Langevin dynamics.

Preliminary – Robbin-Monro algorithm

- Goal : Solve equation $g(x^*) = 0$ numerically.
  - If the function ***g* is known** -> apply Newton's method: $x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)}$
    and it guarantees quadratic convergence (i.e : $|x_{n+1} - x_n| \le M \cdot |x_n - x_{n-1}|^2$)
  - When the function ***g* is unknown**, and but we observe some R.V whose mean is $g(x)$
    - It is equivalent to observe $y_n = g(x_n) + \zeta_n$, where $\mathbb{E}[\zeta_n] = 0$.
    - In this case, Newton's method does not guarantee the convergence.

(Additional) sufficient conditions :

1. **y_n is uniformly bounded**
2. **$g(x_n)$ is non-decreasing**
3. **$g'(x^*)$ exists and positive**

<Robbin-Monro algorithm> : $x_{n+1} = x_n - \epsilon_n y_n$

where $\sum_{n=1}^{\infty} \epsilon_n = \infty$ and $\sum_{n=1}^{\infty} \epsilon_n^2 < \infty$ (sufficient condition for convergence w.p 1.)

Ref : https://appliedprobability.blog/2019/01/26/robbins-munro-2/

Preliminary – Langevin dynamics

- SDE (Stochastic differential equation) : stochastically perturbed ODE

PDE

$$\frac{dX(t)}{dt} = f(X(t)), \qquad t > 0$$

$$X(0) = X_0$$

SDE

$$\frac{dX(t)}{dt} = f\big(t, X(t)\big) + \sigma\big(t, X(t)\big)\zeta(t), \qquad t > 0$$

$$X(0) = X_0$$

where $\zeta(t)$ is a white noise satisfying $\mathbb{E}[\zeta(t)] = 0,\ Cov\big(\zeta(t), \zeta(s)\big) = \delta(t - s)$.

- We formally set $\zeta(t) = \frac{dW(t)}{dt}$, then $dW(t) = \zeta(t)dt$ (differential form of the Brownian motion) and we obtain the following:

$$dX(t) = \underbrace{f\big(t, X(t)\big)dt}_{\text{Drift term}} + \underbrace{\sigma\big(t, X(t)\big)dW(t)}_{\text{Diffusion term}}$$

Ref : https://ethz.ch/content/dam/ethz/special-interest/mavt/dynamic-systems-n-control/idsc-dam/Lectures/Stochastic-Systems/SDE.pdf

Preliminary – Langevin dynamics

- Note that Drift term accounts for deterministic behavior as time goes while Diffusion term explain the unexpected stochastic behavior.

- [Langevin equation for Brownian motion]:

$$\frac{dX(t)}{dt} = -\nabla U\big(X(t)\big) + \sigma\zeta(t) \leftrightarrow dX(t) = -\nabla U\big(X(t)\big)dt + \sigma dW(t)$$

  where U is a potential function of $X(t)$

- By applying discrete approximation on above equation and taking $\zeta(t) \sim N(0,1)$

$$X_{t+1} - X_t = -\Delta t \cdot \nabla U(X_t) + \sigma \cdot \sqrt{\Delta t}\, N(0,1)$$

Algorithm (Langevin Dynamics sampling)

- Our intuition for Langevin Dynamics samping is to :
  - Maximize posterior distribution $p(\theta|X)$ (attain MAP) (By maximizing $\log p(\theta|X)$)
  - Avoid local mode by injecting random gaussian noise (adopting Brownian motion)

- Define unnormalized log-posterior $U(\theta)$ by:

$$U(\theta) := -\sum_{i=1}^{N} \log p(x_i|\,\theta) - \log p(\theta)$$

- Now, we update our parameter θ using Langevin dynamics:

$$\theta_{t+1} - \theta_t = -\frac{\epsilon_t}{2} \cdot \nabla U(\theta_t) + \eta_t = \frac{\epsilon_t}{2}\left(\nabla \log p(\theta_t) + \sum_{i=1}^{N} \log p(x_i|\theta_t)\right) + \eta_t$$

where ϵ_t : learning rate and $\eta_t \sim N(0, \epsilon_t)$

Algorithm (Stochastic Langevin Dynamics sampling = SGLD)

- **Problem** : We must pass whole data X with size N, which becomes intractable as N increases
- **Suggested Solution** : Use stochastic version where convergence is guaranteed by Robbin-Monro algorithm.

- Set stochastic unnormalize log-posterior $\widetilde{U}(\theta) = -\frac{N}{n}\sum_{i=1}^{n} \log p(x_{ti}|\theta) - \log p(\theta)$ (where n is batch size)
- Now, update θ_t using SGD:

$$\theta_{t+1} - \theta_t = -\frac{\epsilon_t}{2} \cdot \nabla\widetilde{U}(\theta_t) + \eta_t = \frac{\epsilon_t}{2} \cdot \left(\nabla \log p(\theta_{ti}) + \frac{N}{n}\sum_{i=1}^{n} \log p(x_{ti}|\theta_t) \right) + \eta_t$$

- Note : $\sum_{t=1}^{\infty} \epsilon_t = \infty$ and $\sum_{t=1}^{\infty} \epsilon_t^2 < \infty$ for convergence (ex: $\epsilon_t = a(b+t)^{-\gamma}$ with $\gamma \in (0.5,1])$)

Features of SGLD

1. Transition from stochastic optimization to Langevin dynamics during training
   - The variance of $\Delta\theta_t \;= \theta_{t+1} - \theta_t$ due to stochastic GD:
     - Let $V_s = \frac{1}{n}\sum_{i=1}^{n}(s_{ti} - \bar{s_t})(s_{ti} - \bar{s}_t)^T$ where $s_{ti} = \nabla \log p(x_{ti}|\theta_t) + \frac{1}{N}\nabla \log p(\theta_t)$
     - Then, $Var(\Delta\theta_t|\eta_t) = \frac{\epsilon_t^2 N^2}{4n} V_s$, and $\|Var(\Delta\theta_t|\eta_t)\|_2 \leq \frac{\epsilon_t^2 N^2}{4n}\lambda_{max}(V_s) = \alpha$

   - When $\alpha \ll 1$, the SGD noise becomes negligible and transition into Langevin dynamics sampling happens. (guaranteed since $\epsilon_t \to 0$ as $t \to \infty$)

   - To adjust the transition point, we can pre-multiply preconditioned matrix M by:

$$\Delta\theta_t = \frac{\epsilon_t}{2} \cdot M\left(\nabla \log p(\theta_{ti}) + \frac{N}{n}\sum_{i=1}^{n} \log p(x_{ti}|\theta_t)\right) + \eta_t$$

where $\eta_t \sim N(0, \epsilon_t, M)$ ($\to \|Var(\Delta\theta_t|\eta_t)\|_2 = \frac{\epsilon_t^2 N^2}{4n}\lambda_{max}\left(M^{1/2} V_s M^{1/2}\right)$ which is controllable.)

Features of SGLD

2. Ignore of acceptance step
    - There is no need to set proposal distribution or computing $p(\theta_t)$

3. Use sub-data in each iteration (following from SGD)
    - One drawback : Possibility to stuck in a local mode (or local minimum) -> Solution by 4.

4. Adding random gaussian noise (Brownian motion):
    - Effectively escape local minimum and leads to successful estimation of posterior distribution $p(\theta|X)$

- Note : Why is the gaussian noise variance fixed to be ϵ_t ? → for guarantee of a correct sampler (by the Fokker-Planck Equation)