4.10 - SGLD (Pre 1)

Two goals: 1) in one step, $\theta_t \cdots$ 2) for the sequence $\{\theta_t\}_t \cdots$

1) When t is very large: $\nabla\theta_t = \frac{\varepsilon_t}{2}\left(\nabla\log p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n}\log p(x_{ti}|\theta_t)\right) + \eta_t$

could be regarded as Langevin equation

i.e. its equilibrium solution is posterior distribution over θ_t

$\Leftarrow$ in a specific step, injected noise η_t "$\gg$" (dominates) stochastic gradient

i.e. Var("noise") > Var(stochastic gradient)

Proof:

$g(\theta) := \nabla\log p(\theta) + \sum_{i=1}^{N}\nabla\log p(x_i|\theta)$ "Δ gradient"

$h_t(\theta) := \nabla\log p(\theta) + \frac{N}{n}\sum_{i=1}^{n}\log p(x_i|\theta) - g(\theta)$ ("actual" - "expected")

SGLD: $\Delta\theta_t = \underbrace{\frac{\varepsilon_t}{2}(g(\theta_t) + h_t(\theta_t))}_{①} + \underbrace{\eta_t}_{②}$ $\quad \eta_t \sim N(0, \varepsilon_t)$

$E(h_t(\theta)) = 0 \quad Var(h_t(\theta)) < \infty := V(\theta)$

$\therefore$ ①: $Var = \frac{\varepsilon_t^2}{4}V(\theta_t)$ ②: $Var = \varepsilon_t$

when $\varepsilon_t \to 0$ $Var(②) \gg Var(①)$ #.

2). $\{\theta_t\}_t$? non-stationary (ε_t changes) ; $\varepsilon_t \to 0$ ~~actual~~ expected "Δ"

$\Leftarrow$ subsequence $\theta_{t_1}, \theta_{t_2} \cdots \to$ posterior ② gradient: $g(\theta_t)$ "$\gg$" $h_t(\theta_t)$

$\Leftarrow$ For this subsequence: ① total injected noise "$\gg$" (dominated) total stochastic gradient

i.e. $\{\theta_t\}$ can be regarded as being sampling from normal LD

Proof: Find such $\{t_i\}$ $t_1 < t_2 < \cdots$ s.t. $\sum_{t_s+1}^{t_{s+1}}\varepsilon_t \to \varepsilon_0$, $s \to \infty$, $0 < \varepsilon_0 \ll 1$

找到这样一种分割 使得… "between sum" being restricted around ε_0

Total injected noise: $\|\sum_{t=t_s+1}^{t_{s+1}}\eta_t\|_2 = O(\sqrt{\varepsilon_0})$ $s \to \infty$ (very large)

Total gradient: $\sum_{t=t_s+1}^{t_{s+1}}\frac{\varepsilon_t}{2}(g(\theta_t) + h_t(\theta_t))$ $\quad \varepsilon_0 \ll 1 \therefore s \to \infty \ \|\theta_t - \theta_{t_s}\|_2 \ll 1, \forall t \in [t_{s+1}, t_{s+1}]$

$= \frac{\varepsilon_0}{2}g(\theta_{t_s}) + O(\varepsilon_0) + \sum_{t_s+1}^{t_{s+1}}\frac{\varepsilon_t}{2}h_t(\theta_t)$

smoothness

$\hookrightarrow$ dominated by mini-batch choice's random

if iid $Var(\sum\frac{\varepsilon_t}{2}h_t(\theta_t)) = \sum_{t_s+1}^{t_{s+1}}\frac{\varepsilon_t^2}{4}$

$= \frac{\varepsilon_0}{2}g(\theta_{t_s}) + O(\varepsilon_0) + O(\sqrt{\sum\frac{1}{4}\varepsilon_t^2})$

$= \frac{\varepsilon_0}{2}g(\theta_{t_s}) + \boxed{O(\varepsilon_0)}$ v.s. $O(\sqrt{\varepsilon_0})$

normal constant ε's LD — influence from total gradient — influence from total noise

#.