LANGEVIN DYNAMICS

$$d\vec{\theta} = -\nabla_\theta U(\theta)\,dt + \sqrt{2}\,dW_t \qquad \text{(LD)}$$

targets

$$p(\theta) = e^{-U(\theta)}/Z$$

Time-Evolution Obeys Fokker-Planck Eq:

$$\partial_t p(\theta) = \partial_{\theta_i}\left(\partial_{\theta_i} U\, p(\theta)\right) + \partial_{\theta_i}\partial_{\theta_i}\left(p(\theta)\right)$$

$$\boxed{\text{Conservation of Probability}}$$

1-D: $\partial_t p = \frac{\partial}{\partial\theta}\left(\frac{\partial U}{\partial\theta}\, p(\theta)\right) + \frac{\partial^2 p}{\partial\theta^2}$

If $\partial_t p = 0$ check $p = e^{-U}/Z$ is a solution

$$\frac{\partial}{\partial\theta}\left(+\frac{\partial U}{\partial\theta} e^{-U}/Z\right) + \frac{\partial}{\partial\theta}\overbrace{\left(-U e^{-U}/Z\right)}^{\partial p/\partial\theta} = 0$$

✓

SGLD

$$\Delta\theta_t = \frac{\varepsilon_t}{2}\Big(\underbrace{\nabla \log p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n} \nabla \log p(x_i|\theta_t)}_{\tilde{\nabla} U(\theta) \approx \nabla U(\theta_t) + \underbrace{g(\theta_t)}_{\text{subsampling noise}}}\Big) + N(0,\varepsilon_t)$$

Noisy Gradient

$\cancel{V(\log g(\theta_t))}$

$$V(\text{Gradient Noise}) = V\left(\frac{\varepsilon_t}{2} g(\theta_t)\right) = \frac{\varepsilon_t^2}{2} V(g(\theta_t))$$

$$\ll V(\text{Real Noise}) = V(N(0,\varepsilon_t)) = \varepsilon_t$$

if $\varepsilon_t \to 0$, SGLD $\approx$ LD

If $\sum \varepsilon_t^2 < \infty$, $\sum \varepsilon_t \to 0$

RM conditions

SGD converges to local mode.

As $\varepsilon_t \to 0$ MH Rejection Probability $\to 0$.

SGLD for BNN

$=$

SGD for NN

$+$

Noise

Prior on Weights : $p(W)$

Probabilistic Output :

i.e. softmax $\xleftrightarrow{\text{Bregman Divergence}}$ Cross-Entropy Loss

Gaussian $\longleftrightarrow$ L^2 loss

Do BACKPROP on $W + \mathcal{N}(0, \varepsilon_t)$

to sample

BDN

- Predictive Distribution

$$q(y|x) = \frac{1}{S} \sum_{s=1}^{S} p(y|x, \theta^s)$$

S is MC samples of "teachers"

$$|\theta^s| \approx 10^6 - 10^8 \text{ params}$$

$q(y|x) \rightsquigarrow$ "teachers" $\to$ Bayesian Predictive Ensemble

$S(y|x, w) \rightsquigarrow$ "Student" $\to$ Deep Net

SGLD trains $q(\theta)$ online while simultaneously minimizing

~~mini~~ $KL\left(q(y|x) \,||\, S(y|x, w)\right)$

using SGD

$$L(w|x) = KL(p(y|x, D_N) \,||\, S(y|x,w))$$
$$= -\mathbb{E}_{p(y|x,D_N)} \log S(y|x,w)$$
$$= -\int \left[\int p(y|x,\theta) p(\theta|D_N) \, d\theta \right] \log S(y|x,w) \, dy$$
$$\doteq -\int p(\theta|D_N) \int p(y|x,\theta) \log S(y|x,w) \, dy \, d\theta$$
$$= -\int p(\theta|D_N) \left[\mathbb{E}_{p(y|x,\theta)} \log S(y|x,w) \right] d\theta$$

? ?

x is the input data to

STUDENT

Monte-Carlo to integrate out x
is high-dimensions, near "training data"
$\boxed{D'}$

$$\hat{L}(w) \approx \frac{1}{|D'|} \sum_{x' \in D'} L(w|x').$$

$$\hat{L}(w) \approx -\frac{1}{|\Theta||D'|} \sum_{\theta^* \in \Theta} \sum_{x' \in D'} \mathbb{E}_{p(y|x,\theta^*)} \log S(y|x,w)$$

Distilled / Online SGLD $\begin{cases} \theta & \text{Teacher} \\ W & \text{Student} \end{cases}$

for $t = 1 : T$ do

Update θ : (SGLD Step)

$$\theta_{t+1} = \theta_t + \frac{\varepsilon_t}{2}\left(\nabla_\theta \log p(\theta) + \frac{N}{n}\sum_{i\in[n]} \nabla_\theta \log p(y_i \mid x_i, \theta)\right) + \mathcal{N}(0, \varepsilon_t)$$

Update W : (Student Step)

Sample D' from student generator

$$W_{t+1} = W_t - \rho_t \left(\frac{1}{|D'|} \sum_{x' \in D'} \nabla_W \hat{L}(\omega, \theta_{t+1} \mid x') \right.$$

$$\left. + \underbrace{\gamma W_t}_{L^2 \text{ Reg.}} \right)$$

Recall for softmax output

$$\hat{L}(W, \theta_{t+1} \mid x) = - \sum_{k=1}^{K} \underbrace{p(y = k \mid x, \theta^*)}_{\text{TEACHER}} \log \underbrace{S(y = k \mid x, W)}_{\text{STUDENT}}$$

USE SGD (BACKPROP) to TRAIN