SGLD

A Preprint

1 Abstract

By adding the right amount of noise to a standard stochastic gradient optimization algorithm we show that the iterates will converge to samples from the true posterior distribution as we anneal the stepsize. This seamless transition between optimization and Bayesian posterior sampling provides an inbuilt protection against overfitting

2 Legend

- θ : parameters
- $p(\theta)$ − prior.distribution
- $X = \{x_i\}_{i=1}^N$

3 Main

As we want to maximize the posterior $p(\theta|X)$ taking a log as basically coming to log-posterior(Maximizing this is equivalent to minimizing the negative of this function). So we approximate grad:

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

where ϵ_t is a sequence of step sizes.

The issue with ML or MAP estimation, as stochastic optimization aims to do, is that they do not capture parameter uncertainty and can potentially overfit data so Gaussian noise is used so they do not collapse:

$$\Delta\theta_t = \frac{\epsilon_t}{2}(\nabla \log p(\theta_t) + \frac{N}{n}\sum_{i=1}^{N} \nabla \log p(x_{ti}|\theta_t)) + \eta_t \quad \eta_t \sim N(0;\epsilon)$$

This simple modification causes the dynamics to transition from stochastic optimization to approximate posterior sampling as $\epsilon_t \to 0$.

Convergence Intuition

Let $g(\theta)$ be the true gradient:

$$g(\theta) = \nabla \log p(\theta) + \sum_{i=1}^{N} \nabla \log p(x_i \mid \theta),$$

and let the stochastic approximation be:

$$h_t(\theta) = \frac{N}{n}\sum_{i=1}^{n} \nabla \log p(x_{ti} \mid \theta) - \sum_{i=1}^{N} \nabla \log p(x_i \mid \theta).$$

Then the update becomes:

$$\Delta\theta_t = \frac{\epsilon_t}{2}(g(\theta_t) + h_t(\theta_t)) + \eta_t.$$

As $t \to \infty$, $\epsilon_t \to 0$, the noise η_t dominates h_t, and updates behave like Langevin dynamics, which has the posterior as its stationary distribution.

Why as $t \to \infty$, $\epsilon_t \to 0$, the noise η_t dominates h_t, and the updates resemble Langevin dynamics

1. Stochastic Gradient Noise h_t: Its contribution is scaled by:

$$\text{Magnitude} \sim \frac{\epsilon_t}{2} h_t(\theta_t) \quad \Rightarrow \quad \text{Variance} \propto \left(\frac{\epsilon_t}{2}\right)^2 = \mathcal{O}(\epsilon_t^2)$$

2. Injected Noise η_t: Since $\eta_t \sim \mathcal{N}(0, \epsilon_t I)$, we have:

$$\text{Variance} = \mathcal{O}(\epsilon_t)$$

3. Relative Comparison: Now compare the two noise sources:

$$\frac{\text{Var(stochastic gradient noise)}}{\text{Var(injected noise)}} \propto \epsilon_t \to 0 \quad \text{as } t \to \infty$$

Conclusion: As t increases and $\epsilon_t \to 0$, the injected Gaussian noise η_t dominates the stochastic gradient error h_t.

Convergence to Langevin Dynamics

When $h_t(\theta_t)$ becomes negligible, the update becomes:

$$\Delta\theta_t = \frac{\epsilon_t}{2} g(\theta_t) + \eta_t,$$

which is a discretization of the continuous-time Langevin diffusion:

$$d\theta = \frac{1}{2}\nabla \log p(\theta \mid X)\, dt + dW_t,$$

where W_t is a Wiener process (Brownian motion).

It is known that this stochastic differential equation has the posterior distribution $p(\theta \mid X)$ as its stationary distribution, under mild regularity conditions.

Summary

Therefore, as:

- $\epsilon_t \to 0$,
- the mini-batch noise h_t vanishes faster than η_t,

the SGLD updates effectively become Langevin dynamics, and the iterates θ_t asymptotically follow the posterior distribution.

Algorithm: Stochastic Gradient Langevin Dynamics

Algorithm 1 Stochastic Gradient Langevin Dynamics

Require: Initial parameters θ_0, step size schedule $\{\epsilon_t\}$, data $\{x_i\}_{i=1}^N$, minibatch size n

1: for $t = 1$ to T do

2: Sample minibatch $X_t = \{x_{t1}, \ldots, x_{tn}\}$

3: Compute stochastic gradient:

$$G_t = \nabla \log p(\theta_t) + \frac{N}{n} \sum_{i=1}^{n} \nabla \log p(x_{ti} \mid \theta_t)$$

4: Sample noise $\eta_t \sim \mathcal{N}(0, \epsilon_t I)$

5: Update parameters:

$$\theta_{t+1} = \theta_t + \frac{\epsilon_t}{2} G_t + \eta_t$$

6: end for

Notes

- Step size ϵ_t must decay such that:

$$\sum_t \epsilon_t = \infty, \qquad \sum_t \epsilon_t^2 < \infty$$

- This guarantees convergence to the posterior without requiring a Metropolis-Hastings correction step.
- SGLD can be seen as a natural extension of SGD for Bayesian inference.