Stochastic Gradient Langevin Dynamics

Clayton W. Seitz

September 6, 2022

Langevin dynamics

Bayesian inference of parameters starts with Bayes rule

$$P(\theta|\mathcal{D},\mathcal{M}) = \frac{P(\mathcal{D}|\theta,\mathcal{M})P(\theta|\mathcal{M})}{\int P(\mathcal{D}|\theta,\mathcal{M})P(\theta|\mathcal{M})d\theta}$$

where $P(\mathcal{D}|\theta,\mathcal{M}) = \prod_{i=1}^{N} P(\mathcal{D}_i|\theta,\mathcal{M})$. In gradient based Bayesian learning we can define a loss function $\mathcal{L}$

$$\begin{aligned}\nabla_\theta \mathcal{L} &= \nabla_\theta \log P(\theta|\mathcal{D},\mathcal{M}) \\ &= \sum_{i=1}^{N} \nabla_\theta \log P(\mathcal{D}_i|\theta,\mathcal{M}) + \nabla_\theta \log P(\theta|\mathcal{M})\end{aligned}$$

Gradient-based Bayesian learning

In Stochastic gradient descent (SGD), we use minibatches to estimate the true gradient

$$\theta_{t+1} = \theta_t - \eta \tilde{\nabla}_\theta \mathcal{L}$$

This is an unbiased estimator and we can model it as an SDE

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L} + \eta \epsilon_1 \quad \epsilon_1 \sim \mathcal{N}(0, \sigma^2)$$