Langevin Dynamics for Inference in Sparse Coding

Rajit R [1] Daniel Abraham[1]

[1]UC Berkeley

Backgorund on Sparse Coding

Setup: We are given a set of measurements $y_i \sim Az_i + \epsilon_i$ where the z_i are k-sparse, the ϵ_i are iid noise, and the dictionary A is overcomplete.

Goal: Learn the dictionary A given a set of measurements.

Applications: Neuroscience, Image Processing [2]

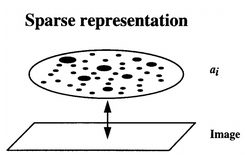

Figure 1. Sparse Image Model.

The vanilla sparse coding model assumes $\epsilon \sim N(0, \sigma^2 I)$. It imposes a Laplacian prior on z_i such that the optimization problem for inference is formulated as: $\min U(z_i) = |Az_i - y_i|^2 + \lambda|z_i|_1$.

$$\min_{x,D} ||Ds - Y||_F^2 \text{ subject to } ||s_i||_0 \leq T$$

The problem is solved using alternating minimization:

$$s^{(k+1)} = \min_X ||D^{(k)}s - Y||_F^2 \text{ subject to } ||s_i||_0 \leq T$$
$$D^{(k+1)} = \min_D ||Ds^{(k+1)} - Y||_F^2$$

Which alternates between updating the sparse representation x and the dictionary D.

Langevin Dynamics

We refer to *continuous-time Langevin dynamics*:

$$\theta_t = -\frac{1}{L}\nabla U(\theta_t)\, dt + \sqrt{2/L}\, dB_t. \tag{1}$$

The L parameter in this case refers to the Lipschitz constant of the energy function $U(\theta_t)$. It is well known that continuous-time Langevin dynamics converges to the target distribution $p^*(\theta) = exp(-U(\theta))$ exponentially quickly.

We discretize the process to give us Euler-Maruyama method. Notice that if we set the Noise term $B_t = 0$, we simply recover a gradient descent step with stepsize $= 1/L$.

$$\frac{dx}{dt} = -\nabla E(x) + \eta(t)$$

It is essentially gradient descent with additive Gaussian Noise that is uncorrelated in continous time as illustrated here: $\langle \eta(t), \eta(t') \rangle = 2\delta(t - t')$

Second-order Langevin Dynamics incorporates a momentum term to allow the trajectory to become more smooth and decrease discretization error.

$$\begin{cases} \theta_t &= r_t\, dt \\ r_t &= -\frac{1}{L}\nabla U(\theta_t)\, dt - \xi r_t\, dt + \sqrt{2\xi/L}\, dB_t^r, \end{cases} \tag{2}$$

The stationary distribution of this is $p^*(\theta) = exp(-U(\theta) + \frac{1}{2}||r||^2)$.

We can also define second-order Langevin Dynamics which includes a momentum based parameter that improves smoothness of trajectory and non-reversibility.

Learning and Inference

We instead use the spike and slab prior [3]. What is desired is for the activities of s to be L0-sparse. That is we require s_i to be zero with some probability π, so $p(s_i = 0; \pi) = \pi$ and $p_s(s_i) = \pi e^{-\lambda} + (1 - \pi)\delta(s)$

To achieve this, we can define a variable u that follows an exponential distribution. With π as the probability of being 'active', $1 - \pi$ quantifies the L0 sparsity, or how likely s is to be zero. When s is in the active state it is exponentially distributed with mean $1/\lambda$ as in Fig 2. We then take the latent variables s to be given by $s_i = f(u_i)$ where s is a ReLU function offset by u_0. We can show that s_i is then distributed according to the prior $p_0(s)$ by marginalizing the joint distribution $p(s, u)$ over u as follows:

$$p_S(s) = \int p(s|u)p(u)du = \int_0^{u_0} \delta(s)p_U(u)du + \int_{u_0}^{\infty} \delta(s - (u - u_0))p_U(u)du \tag{3}$$

$$= \delta(s)\int_0^{u_0} p_U(u)du + p_U(s + u_0) = \delta(s)[1 - e^{-\lambda u_0}] + \lambda e^{-\lambda s}e^{-\lambda u_0} \tag{4}$$

$$= [1 - \pi]\delta(s) + \pi\lambda e^{-\lambda s} = p_0(s) \tag{5}$$

We then re-write the energy function in terms of u.

$$E(A, u, x) = \frac{1}{2}\frac{||x - Af(|u|)||_2^2}{\sigma^2} + \lambda||u||_1 \tag{6}$$

We let u_i move freely between positive and negative values and then use only their absolute value in evaluating the energy to avoid the problems associated with having an infinite energy barrier at $u_i = 0$.

$$p(|u||x) \propto exp(-\frac{1}{\sigma^2}||x - Af(|u|)||_2^2 - \lambda||u||_1) \tag{7}$$

$$\propto p(x|f(|u|)p_U(|u|) = p(x|s)p_0(s) \tag{8}$$

Thus, by following the Langevin Dynamics governed by equation 14, we can obtain samples from the posterior $p(s|x)$ with $s = f(|u|)$.

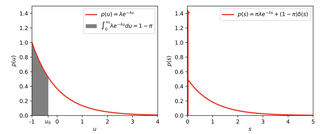

Figure 2. Left: Exponential distribution with $\lambda = 0.5$, Right, spike-and-slab prior $\pi = 0.5$

We use the following coupled system of stochastic differential equations to perform our inference and learning step.

$$\tau_u \dot{u} = -A^T(As - x)\Theta(|u| - u_0) - \lambda sign(u) + \sqrt{2}\xi(t) \tag{9}$$

$$\tau_A \dot{A} = -(As - x)s^T \tag{10}$$

where $\Theta(u)$ is the heaviside function and $\xi(t)$ is independent Gaussian white noise.

Numerical Experiments

We conduct inference each step using first order and second order Langevin dynamics. The inner loop is inference and the outer loop is learning. In the inner loop, we are able to quantify uncertainty as we sample the entire distribution. We then use the MAP estimate. Interestingly, we observe that Langevin Dynamics converges faster too [Fig 4]. The second order Langevin dynamics appears to have a smoother trajectory of exploration of the energy function [Fig 5,6]. We learn the dictionary and observe that the redundant dictionary elements vanish [Fig 3] just like in [1].

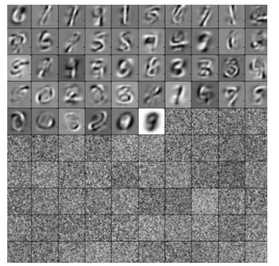

Figure 3. KLS-learned MNIST Dictionary

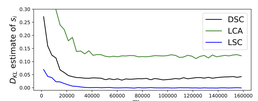

Figure 4. The KL-Divergence for coefficients s_i

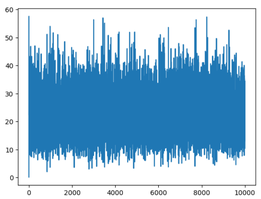

Figure 5. Evolution of energy function for first-order

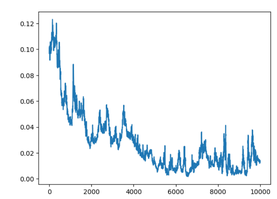

Figure 6. Evolution of energy function for second-order

References

[1] Michael Fang et al.
Langevin sparse coding.
Neural Computation, 34:1676–1700, 2022.

[2] Bruno Olshausen.
Sparse coding with an overcomplete basis set: A strategy employed by V1?
PhD thesis, Stanford, 1997.

[3] J.J. Beauchamp T.J. Mitchell.
Bayesian variable seleciton in linear regression.
Journal of American Statistical Association, 83:1023–1032, 1988.

[4] Martin J Wainwright Peter L Bartlett Michael I Jordan Wenlong Mou, Yi-An Ma.
High-order langevin diffusion yields an accelerated mcmc algorithm.
Journal of Machine Learning Research, 22:1–41, 2021.