Equivalence of noisy gradient descent and sampling from a Gibbs distribution

Jesse de Wringer

July 20, 2021

Abstract

The purpose of this report is to show equivalence between stochastic gradient descent with added standard normal noise to sampling from a Gibbs distribution defined over the solution space. We can then directly extend this to perform Simulated Annealing. SA is a technique that in theory finds global minima given certain annealing schedules, but in practice is mostly used for avoiding poor local minima by properly exploring the solution space.

1 Preface

The conclusion of this report is unlikely to be the first of its kind and similar conclusions are often implicitly mentioned in research. However, a straight forward derivation shows a clear interpretation of SGD with added noise.

2 The Gibbs distribution

Most learning/optimization problems can be defined in terms of the minimization of some risk function $R(X, \theta)$, where $(\mathbf{x}_i)_{1 \leq i \leq N} = X \in \mathcal{X}$ is a dataset where $\mathbf{x}_i \in \mathbb{R}^d$ and $\theta \in \mathbb{R}^p$ is a parameterization of a solution. We furthermore assume that $R(X, \theta) = \frac{1}{N} \sum_{i=1}^{N} R_i(\mathbf{x}_i, \theta)$, which simply tells us that we can evaluate the risk function at one data point, which is a sufficient condition for most kinds of batch optimization. We now wish to define a probability density function $p(\theta|X)$ over the solution space. Under the constraint that $\mathbb{E}_{p(\cdot|X)}[R(X, \theta)] = \mu$, the maximization of the entropy $H[p]$ gives us that $p(\theta|X) \propto e^{-\frac{1}{T} R(X,\theta)}$, where T is defined as the temperature and is a free hyperparameter. Given the decomposability condition on the risk, we can also define $p(\theta|\mathbf{x}_i) \propto e^{-\frac{1}{TN} R(\mathbf{x}_i,\theta)}$, which gives $\prod_{i=1}^{n} p(\theta|\mathbf{x}_i) = p(\theta|X)$. Technically, $p(\theta|X)$ defines a family of distributions, of which the members can be uniquely identified by T.

3 Stochastic gradient Langevin dynamics

The SGLD algorithm [1] is an MCMC algorithm that can produce samples from a non-normalized distribution, which in this case, is $p(\theta|X)$. Given a randomly initialized chain $\theta_0 \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_p)$ and a random subset (without replacement) of the data $\mathcal{I}$, we can describe the next sample in the chain as follows:

$$\theta_{k+1} = \theta_k + \tau \frac{N}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \nabla_\theta \log p(\theta|\mathbf{x}_i)|_{\theta=\theta_k} + \sqrt{2\tau}\mathbf{z}_k \quad (1)$$

Where $\mathbf{z}_k \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_p)$. Here τ is some hyperparameter. Only under some decaying schedules do samples of this chain actually converge to samples from the stationary distribution $p(\theta|X)$.

4 Showing equivalence

Starting from equation 1, and using the definition of the Gibbs distribution:

$$\begin{aligned}
\theta_{k+1} &= \theta_k - \tau' \frac{N}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \nabla_\theta \frac{1}{TN} R(\mathbf{x}_i, \theta)|_{\theta=\theta_k} + \sqrt{2\tau'}\mathbf{z}_k \\
&= \theta_k - \tau' \frac{1}{T|\mathcal{I}|} \sum_{i \in \mathcal{I}} \nabla_\theta R(\mathbf{x}_i, \theta)|_{\theta=\theta_k} + \sqrt{2\tau'}\mathbf{z}_k \\
&= \theta_k - \tau \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \nabla_\theta R(\mathbf{x}_i, \theta)|_{\theta=\theta_k} + \sqrt{2T\tau}\mathbf{z}_k
\end{aligned}$$

Where we defined $\tau := \frac{\tau'}{T}$. Which proves that performing SGD with added standard normal noise is equivalent to sampling from a Gibbs distribution over the solution space. The amount of noise that is added directly determines which Gibbs distribution we are sampling from.

5 Simulated annealing

SA is an algorithm to find global optima of some function. Although SA is quite a general algorithm, we will now look at a specific version of SA. In our problem, we are trying to find a minima of $R(X, \theta)$. We proceed by using the previously defined family of density functions. For every member of this family, its maxima correspond to the minima of the risk function. It also also interesting to look at the limiting forms of the distributions. When $T \to \infty$, the distributions converges to the uniform distribution over all solutions. As $T \to 0$, the distributions converge to the uniform distribution over the minimal risk solutions. The aforementioned SGLD algorithm now enables us to sample from every such distribution, when given a temperature. Theoretically it is now possible to use a low termpature and sample (close to) global maximima. However, the mixing time will become a practical burden. To ameliorate this, we can instead use a specific decaying temperature schedule to still produce global maxima, but even still, this procedure is computationally costly. In practice, we use a faster temperature decaying schedule, and can only hope for low risk solutions.

The relevance to our problem is that performing SGD with decaying noise is equivalent to performing SA when the chosen MCMC algorith is SGLD. Especially in low dimensional problems, this can greatly speed up or even enable learning.

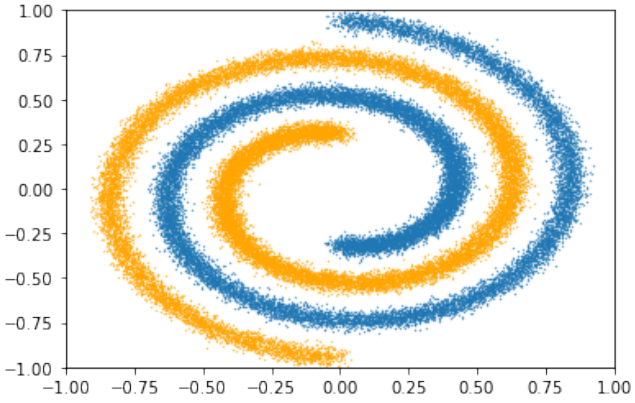

Figure 1: The dataset in question. Blue points correspond to label 0 and orange points correspond to label 1.

6 Experiments

As an easily visualizable classification problem, we look at a Swiss roll dataset, where the purpose is to perform binary classification of the points as shown in figure 1.

In figure 2, we can see the loss over time for vanilla SGD versus noise SGD. It is important to notice that the early solutions are worse than random initialization, which shows that the solution landscape is properly being explored.

The solutions after training can be seen in figures 3 and 4. All hyperparameters can easily be found in the corresponding notebook.

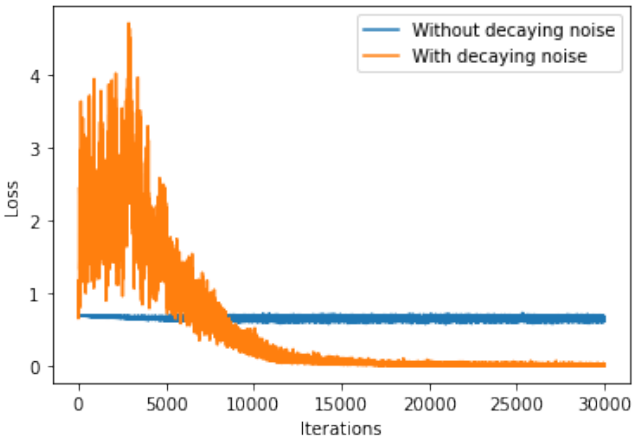

Figure 2: This figure shows that standard SGD does not manage to properly learn a good solution to a relatively simple classification problem, while noisy SGD finds a very strong solution

References

[1] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

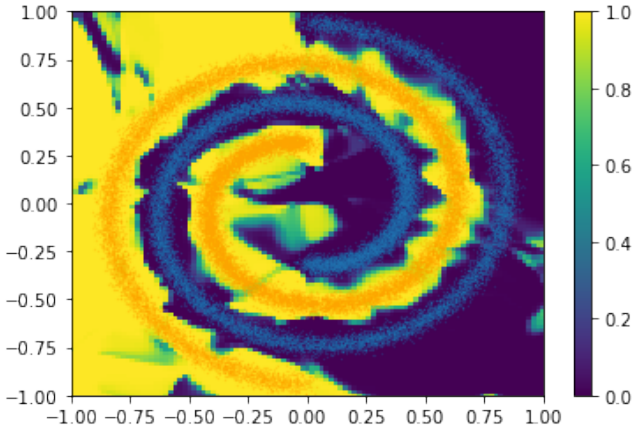

Figure 3: The final output after training using noisy SGD

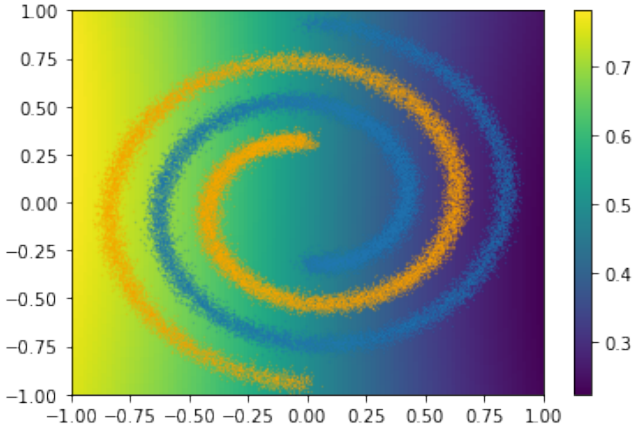

Figure 4: The final output after training using vanilla SGD