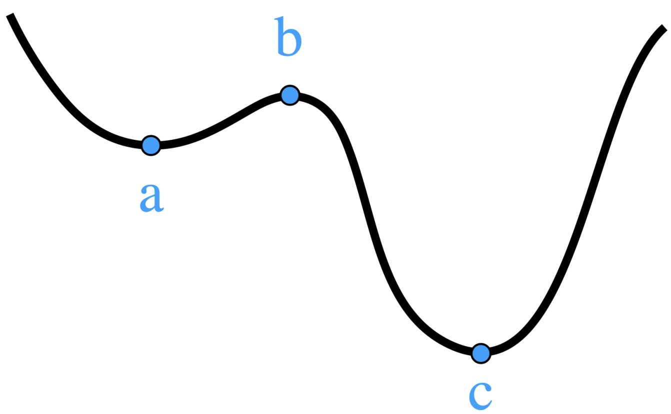
Kramer's escape problem with SGLD and SGD

Hikaru Ibayashi

This note is my own reproduction of the results in Xie et al. (2020) for my understanding.

We will formulate the escape time of SGLD and SGD in the framework of Kramer's escape problem, where we consider how much probability escapes from a to c.



SGLD analysis We model SGLD as follows,

$$d\theta = -dt\frac{\partial L(\theta)}{\partial \theta} + \sqrt{T}\mathcal{N}(0, dtI).$$

First, we discuss the case of SGLD. We assume probability distribution is almost in the stationary distribution and and most of the probability is located around a. Therefore the amount of probability around a can be calculated as follows,

$$\begin{aligned} P(\theta \in V_a) &= P(a)\int_{\theta \in V_a} \exp\left(-\frac{L(\theta') - L(a)}{T}\right) d\theta' \\ &\approx P(a)\int_{\theta \in V_a} \exp\left(-\frac{H_a(\theta' - a)^2}{2T}\right) d\theta' \\ &\approx P(a)\frac{(2\pi T)^{\frac{1}{2}}}{H_a^{\frac{1}{2}}} \end{aligned}.$$

As a next step, we consider how much flux escapes from a. We know that the probability density flow follows Fokker-Plank equation,

$$\frac{\partial P(\theta, t)}{\partial t} = \frac{\partial}{\partial \theta} \cdot \frac{\partial L(\theta)}{\partial \theta} P(\theta, t) + T\frac{\partial^2 P(\theta, t)}{\partial \theta^2}.$$

As we know $\frac{\partial P(\theta,t)}{\partial t} = -\frac{\partial J}{\partial \theta}$, the flux is as follows,

$$J = -\frac{\partial L(\theta)}{\partial \theta}P(\theta, t) - T\frac{\partial P(\theta, t)}{\partial \theta}$$

An a alternative equivalent representation,

$$J = -T \exp\left(-\frac{L(\theta)}{T}\right) \frac{\partial}{\partial \theta} \left(\exp\left(\frac{L(\theta)}{T}\right) P(\theta, t)\right).$$

Or,

$$\frac{\partial}{\partial \theta} \left(\exp\left(\frac{L(\theta)}{T}\right) P(\theta, t)\right) = -\frac{J}{T} \exp\left(\frac{L(\theta)}{T}\right). \tag{1}$$

As we assumed probability distribution is almost in the stationary distribution, we can assume J is independent of θ. In other words, a constant escaping flux is flowing from a to c. So integrating over $[a, c]$ on both sides, we can get

$$\left[\exp\left(\frac{L(\theta)}{T}\right) P(\theta, t)\right]_a^c = -\frac{J}{T} \int_a^c \exp\left(\frac{L(\theta')}{T}\right) d\theta'.$$

As we assumed most of the probability is around a, the left hand side equals to $-\exp\left(\frac{L(a)}{T}\right) P(a, t)$. Hence we can formulate the flux as follows,

$$J = \frac{T \exp\left(\frac{L(a)}{T}\right) P(a)}{\int_a^c \exp\left(\frac{L(\theta')}{T}\right) d\theta'}$$

Now, let's take a close look at $\int_a^c \exp\left(\frac{L(\theta')}{T}\right) d\theta'$. As we can see from the figure, $\exp\left(\frac{L(\theta')}{T}\right)$ has a peak at $\theta' = b$. So we can approximate as follows,

$$\begin{aligned}
\int_a^c \exp\left(\frac{L(\theta')}{T}\right) d\theta' &= \exp\left(\frac{L(b)}{T}\right) \int_a^c \exp\left(\frac{L(\theta') - L(b)}{T}\right) d\theta' \\
&\approx \exp\left(\frac{L(b)}{T}\right) \int_a^c \exp\left(\frac{H_b (\theta' - b)^2}{2T}\right) d\theta' \\
&\approx \exp\left(\frac{L(b)}{T}\right) \int_{\infty}^{-\infty} \exp\left(\frac{H_b (\theta' - b)^2}{2T}\right) d\theta' \\
&= \exp\left(\frac{L(b)}{T}\right) \frac{(2\pi T)^{\frac{1}{2}}}{|H_b|^{\frac{1}{2}}}
\end{aligned}$$

As a result, J can be approximated as follows

$$J \approx \frac{T \exp\left(\frac{L(a)}{T}\right) P(a)}{\exp\left(\frac{L(b)}{T}\right) \frac{(2\pi T)^{\frac{1}{2}}}{|H_b|^{\frac{1}{2}}}}$$

Combine everything together, the escape time is

$$\frac{J}{P(\theta \in V_a)} \approx \frac{\sqrt{H_a |H_b|}}{2\pi} \exp\left(\frac{L(a) - L(b)}{T}\right)$$

SGD analysis We model SGD as follows,

$$d\theta = -dt \frac{\partial L(\theta)}{\partial \theta} + \sqrt{\frac{\eta}{B} H(\theta)}\, \mathcal{N}(0, dtI).$$

Getting $P(\theta \in V_a)$ follows the same process with SGLD except for T_a.

$$P(\theta \in V_a) \approx P(a) \frac{(2\pi T_a)^{\frac{1}{2}}}{H_a^{\frac{1}{2}}}.$$

For the second part, to get J, we use Fokker-Plank equation similarly, Eq. (1).

$$\frac{\partial}{\partial \theta}\left(\exp\left(\frac{L(\theta)}{T}\right)P(\theta,t)\right) = -\frac{J}{T}\exp\left(\frac{L(\theta)}{T}\right).$$

Let s be the middle point between a and b, where T_a is dominant in $[a,s]$ and T_b is dominant in $[s,b]$. Here, we introduce $L(s) = (1-s)L(a) + sL(b)$. [1]

$$\frac{\partial}{\partial \theta}\left(\exp\left(\frac{L(\theta)-L(s)}{T}\right)P(\theta,t)\right) = -\frac{J}{T}\exp\left(\frac{L(\theta)-L(s)}{T}\right)$$

As we did in the case of SGLD, we take integrate over [a,c].

$$\begin{aligned}
\text{LHS} &= \left[\exp\left(\frac{L(\theta)-L(s)}{T_a}\right)P(\theta,t)\right]_a^s + \left[\exp\left(\frac{L(\theta)-L(s)}{T_b}\right)P(\theta,t)\right]_s^c \\
&= P(s) - \exp\left(\frac{L(a)-L(s)}{T_a}\right)P(a) - P(s) \\
&= -\exp\left(\frac{L(a)-L(s)}{T_a}\right)P(a) \\
\text{RHS} &= -J\int_a^c T^{-1}\exp\left(\frac{L(\theta)-L(s)}{T}\right)d\theta.
\end{aligned}$$

As a result, we can get the flux J as follows,

$$J = \frac{\exp\left(\frac{L(a)-L(s)}{T_a}\right)P(a)}{\int_a^c T^{-1}\exp\left(\frac{L(\theta)-L(s)}{T}\right)d\theta}$$

Here, the denominator can be approximated in the same way with the case of SGLD. Note that T_b is dominant around b.

$$\int_a^c \exp\left(\frac{L(\theta')}{T}\right)d\theta' \approx \exp\left(\frac{L(b)}{T_b}\right)\frac{(2\pi T_b)^{\frac{1}{2}}}{|H_b|^{\frac{1}{2}}}$$

Combining everything together, the escape time is

$$\begin{aligned}
\frac{J}{P(\theta \in V_a)} &\approx \frac{\exp\left(\frac{L(a)-L(s)}{T_a}\right)P(a)}{\exp\left(\frac{L(b)}{T_b}\right)\frac{(2\pi T_b)^{\frac{1}{2}}}{|H_b|^{\frac{1}{2}}}P(a)\frac{(2\pi T_a)^{\frac{1}{2}}}{H_a^{\frac{1}{2}}}} \\
&= \frac{\sqrt{T_b H_a |H_b|}}{2\pi\sqrt{T_a}}\exp\left(-\frac{L(s)-L(a)}{T_a} - \frac{L(b)-L(s)}{T_b}\right)
\end{aligned}$$

SGD model by SGN, it is known that $T_a = \frac{\eta}{B}H_a$ and $T_b = -\frac{\eta}{B}H_b$. So the simplified result turns out to be as follows

$$\frac{|H_b|}{2\pi}\exp\left(\frac{B}{\eta}(L(a)-L(b))\left(\frac{s}{H_a} + \frac{(1-s)}{|H_b|}\right)\right)$$

References

Xie, Z., Sato, I., and Sugiyama, M. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. February 2020.

[1] This is introduced because Eq. (1) is true only around critical point but this looks artificial.