1. Pedro Talk

The standard gradient descent dynamics is given by

$$dx(t) = -\nabla U(x(t))dt + \sigma dW(t), \quad x(0) = x_0.$$

In this note we will try to study the modified dynamics

$$\begin{aligned} &dX_\varepsilon(t) = -e^{-\gamma\left(U(X_\varepsilon(t)) - \min_{s\le t} U(X_\varepsilon(s))\right)}\nabla U(X_\varepsilon(t))dt + \varepsilon dW(t), \\ &X_\varepsilon(0) = x_0 \end{aligned} \tag{1}$$

2. Idea of Analysis

In the analysis of (1), let us assume that x_0 is a local minimum and a stationary point; that is, $\nabla U(x_0) = 0$. Define the deterministic flow induced by the vector field ∇U starting at an arbitrary point $x \in \mathbb{R}^d$:

$$\frac{d}{dt}S^t x = -\nabla U(S^t x), \quad S^0 x = x. \tag{2}$$

Then, consider the proxy process to (1) given by

$$d\hat{X}_\varepsilon(t) = -e^{-\gamma\left(U(\hat{X}_\varepsilon(t)) - \min_{s<t-\tau_i} U(S^s\hat{X}_\varepsilon(\tau_i))\right)}\nabla U(\hat{X}_\varepsilon(t))dt + \varepsilon dW(t), \tag{3}$$

on the interval $\tau_i < t < \tau_{i+1}$, where the times τ_i are the innovation times defined via

$$\tau_i = \inf\left\{t \ge \tau_{i-1} : U(\hat{X}_\varepsilon(t)) < U(\hat{X}_\varepsilon(\tau_{i-1})\right\}, i \ge 1,$$

and $\tau_0 = 0$ and the same initial condition as X_ε, $\hat{X}_\varepsilon(0) = x_0$. Let us now make some observations.

The time τ_1 is bounded below by the exit time of $\hat{X}_\varepsilon$ from the basin of attraction, $\mathcal{B}(x_0)$, of x_0; that is, $\tau_1 > \sigma(x_0) = \inf\left\{t : \hat{X}_\varepsilon(t) \in \partial\mathcal{B}(x_0)\right\}$, since for every $y \in \mathcal{B}(x_0)$, $\nabla U(y) \neq 0$. Now, from FW theory, it is well known that $\sigma(x_0)$ is exponentially distributed with mean $\varepsilon^{-2}\left(\min_{y\in\partial\mathcal{B}(x_0)} V(y) - V(x_0)\right)$, where V is the quasi-potential of (3). Since, $\hat{X}_\varepsilon$ does not find a new minimum of the function U before time $\sigma(x_0)$, and since $\min_{t>0} U(S^t x_0) = U(x_0)$ we observe that the drift in equation (3) is given by

$$b(x) = -\gamma^{-1}\nabla\left(1 - e^{-\gamma(U(y)-U(x_0))}\right),$$

so the quasi-potential is given by $V(y) = \frac{2}{\gamma}\left(1 - e^{-\gamma(U(y)-U(x_0))}\right)$. As a consequence, the exit time $\sigma(x_0)$ is exponentially distributed with mean λ given by

$$\log\lambda = \frac{2}{\gamma\varepsilon^2}\left(1 - e^{-\gamma\left(\min_{y\in\partial\mathcal{B}(x_0)} U(y) - U(x_0)\right)}\right).$$

So that if this expression is of constant order, then for some constant $c > 0$ and with $\Delta_{x_0} = \min_{y\in\partial\mathcal{B}(x_0)} U(y) - U(x_0)$ it follows that, for γ and ε small enough,

$$\begin{aligned} \gamma\Delta_{x_0} &= \log\left(1 - \gamma\varepsilon^2 c\right) \\ &= \log\left(1 + \sum_{i\ge1}\gamma^i\varepsilon^{2i}c^i\right) \\ &\approx \gamma\varepsilon^2 c + \gamma^2\varepsilon^4 c^2. \end{aligned}$$

This quadratic equation has a unique strictly positive solution given by $\varepsilon^4 c^2\gamma = \Delta_{x_0} - \varepsilon^2 c$, which leads to the following:

CLAIM 1. *By choosing $\gamma = \mathcal{O}(\varepsilon^{-4})$, as $\varepsilon \to 0$, the exit from the basin of attraction $\mathcal{B}(x_0)$ happens in almost constant time.*

3. Some Potential Discretizations

Let us consider SDE (1) up to time τ_1, so that the law of X_ε is the same as the law of

$$dY_\varepsilon(t) = -e^{-\gamma(U(Y_\varepsilon(t))-U(x_0))}\nabla U(Y_\varepsilon(t))dt + \varepsilon dW(t)$$

Then observe that, if we create a uniform equidistant discretization, $\mathcal{P} = \{0 = t_0 < t_1 < ....\}$ of $[0,\infty)$ with $t_1 = \eta$, then it follows that

$$\begin{aligned} Y_\varepsilon(t_{i+1}) &= Y_\varepsilon(t_i) - \int_{t_i}^{t_{i+1}} e^{-\gamma(U(Y_\varepsilon(s))-U(x_0))}\nabla U(Y_\varepsilon(s))ds + \varepsilon\left(W(t_{i+1}) - W(t_i)\right) \\ &\approx Y_\varepsilon(t_i) - \eta\left(e^{-\gamma(U(Y_\varepsilon(t_i))-U(x_0))}\nabla U(Y_\varepsilon(s))ds + \frac{\varepsilon}{\sqrt{\eta}}\xi_i\right). \end{aligned} \tag{4}$$

By comparing the last expression with the characterization of the algorithm, it follows that $\varepsilon = \sqrt{\eta}$, so that $\gamma \approx \eta^{-2}$ giving rise to the claim.

CLAIM 2. *Let $y_0 = x_0$, and consider the algorithm,*

$$y_{i+1} = y_i - \eta\left(e^{-\eta^{-2}(U_i - U_i^*)}\nabla U_i + \xi_i\right), \quad i \in \mathbb{N}, \tag{5}$$

where we have used the short hand notation $U_i = U(y_i)$ and $U_i^* = \min_{j\le i} U_j$. Then, with high probability, after $\mathcal{O}(\eta^{-1})$ time steps, y_i scapes the set $\mathcal{B}(x_0)$.

3.1. Quantitative Estimates for the Discrete Approximation. A necessary towards the proof of Claim 2 is to get an estimate of the approximation given in (4) comparing the continuous SDE with the discrete algorithm proposed in (5).

First observe that, if we define $\rho(y) = e^{-\eta^{-2}(U(y)-U(x_0))}$ with its respective short hand discrete reduction $\rho_i = \rho(y_i)$, it follows that

$$\begin{aligned} Y_\varepsilon(t_{i+1}) - y_{i+1} &= Y_\varepsilon(t_i) - y_i - \int_{t_i}^{t_{i+1}} \left(\rho(Y_\varepsilon(s))\nabla U(Y_\varepsilon(s)) - \rho_i \nabla U_i\right) ds \\ &= Y_\varepsilon(t_i) - y_i - \int_{t_i}^{t_{i+1}} \left(\rho(Y_\varepsilon(s)) - \rho_i\right)\nabla U(Y_\varepsilon(s))ds \\ &\quad - \int_{t_i}^{t_{i+1}} \rho_i\left(\nabla U(Y_\varepsilon(s)) - \nabla U_i\right) ds, \end{aligned}$$

so that if we define $\alpha_i = |Y_\varepsilon(t_{i+1}) - y_{i+1}|$, triangle inequality combined with jensen's inequality implies that $\alpha_{i+1} \le \alpha_i + A_i + B_i$, where we have defined

$$\begin{aligned} A_i &= \int_{t_i}^{t_{i+1}} |\rho(Y_\varepsilon(s)) - \rho_i|\,|\nabla U(Y_\varepsilon(s))|\,ds \text{ and} \\ B_i &= \int_{t_i}^{t_{i+1}} \rho_i\,|\nabla U(Y_\varepsilon(s)) - \nabla U_i|\,ds. \end{aligned}$$

Let us start with the analysis for A_i, by taking the supremum inside the integral,

$$\begin{aligned}
A_i &\leq \eta \left|\nabla U\right|_\infty \sup_{s\in[t_i,t_{i+1})} \left|\rho(Y_\varepsilon(s)) - \rho_i\right| \\
&\leq \eta\rho_i \left|\nabla U\right|_\infty \sup_{s\in[t_i,t_{i+1})} \left|e^{-\eta^{-2}(U(Y_\varepsilon(s))-U_i)} - 1\right| \\
&\leq \rho_i \left|\nabla U\right|_\infty \sum_{j\geq 1} \frac{\eta^{1-2j}}{j!} \sup_{s\in[t_i,t_{i+1})} \left|U(Y_\varepsilon(s)) - U_i\right|^j \\
&\leq \eta^{-1}\rho_i \left|\nabla U\right|_\infty^2 \sum_{j\geq 1} \frac{\eta^{-2(j-1)}}{(j-1)!} \left|\nabla U\right|_\infty^{j-1} \alpha_i^{j-1}\alpha_i \\
&\leq \eta^{-1}\rho_i\alpha_i C,
\end{aligned}$$

for some $C > 0$. Likewise, using the lipshitz constant, ℓ, off ∇U, we see that B_i satisfies $B_i \leq \eta\ell\alpha_i$, and as a consequence, we get that

$$\alpha_{i+1} \leq \alpha_i \left(1 + C\eta^{-1}\rho_i + \ell\eta\right),$$

which readily implies that

$$\log\alpha_{i+1} \leq \log\alpha_0 + \sum_{1\leq j\leq i} \log\left(1 + C\eta^{-1}\rho_j + \ell\eta\right).$$

From the definition of ρ_i, it follows that $\limsup_{\eta\to 0}\eta^{-k}\max_{j\in\mathbb{N}}\rho_j = 0$ for every $k \geq 1$, so in particular, for a constant $c_1 > 0$, we get that

$$\log\alpha_{i+1} \leq \log\alpha_0 + ic_1\eta.$$

So, using the modulus of continuity of the continuous process, the following claim follows:

Claim 3. *After $T > 0$ steps, for some constant $c > 0$, it follows that*

$$\max_{1\leq i\leq T} |Y_\varepsilon(t_i) - y_i| \leq e^{c\eta T}\sqrt{-\eta\log\eta}.$$

In particular, with high probability, after $\mathcal{O}(\eta^{-1})$ number of time steps, y_i is at a distance of at most $\mathcal{O}(\sqrt{-\eta})$ from escaping $\mathcal{B}_{x_0}$.