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
$$dX_\varepsilon(t) = -e^{-\gamma\left(U(X_\varepsilon(t))-\min_{s\le t} U(X_\varepsilon(s))\right)}\nabla U(X_\varepsilon(t))dt + \varepsilon dW(t),$$

(1) $\qquad X_\varepsilon(0) = x_0$

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 $x \in \mathbb{R}^d$:

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

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

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

on the interval $\tau_i < t < \tau_{i+1}$, where the times τ_i are the innovation times defined via
$$\tau_i = \inf\left\{ t \ge \tau_{i-1} : U(\hat{X}_\varepsilon(t)) < U(\hat{X}_\varepsilon(\tau_{i-1})) \right\},$$

with $\tau_0 = 0$ 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 $y \in \mathcal{B}(x_0)$, $\nabla U(y) \ne 0$. 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 $V(y) = 2 - 2e^{-\gamma(U(y)-U(x_0))}$. 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,
$$\gamma\Delta_{x_0} = \log\left(1 - \gamma\varepsilon^2 c\right)$$
$$= \log\left(1 + \sum_{i\ge 1}\gamma^i\varepsilon^{2i}c^i\right)$$
$$\approx \gamma\varepsilon^2 c + \gamma^2\varepsilon^4 c^2.$$

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

$$Y_\varepsilon(t_{i+1}) = Y_\varepsilon(t_i) - \int_{t_i}^{t_{i+1}} e^{-\gamma(U(Y_\varepsilon(s)) - U(x_0))} \nabla U(Y_\varepsilon(s))ds + \varepsilon \left(W(t_{i+1}) - W(t_i) \right)$$

$$\approx Y_\varepsilon(t_i) - \eta \left(e^{-\gamma(U(Y_\varepsilon(t_i)) - U(x_0))} \nabla U(Y_\varepsilon(s))ds + \frac{\varepsilon}{\sqrt{\eta}} \xi_i \right).$$

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,*

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

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

Now, as explained below, exit from the basin of attraction occurs in time