# MixingOptimization

April 9, 2023

## 1 Introduction

### 1.1 Separable Hamiltonian and continuum dynamics

Consider the following Hamiltonian

$$\hat{H} = Vp^2 + V \tag{1}$$

This describes a particle moving in the potential $V$, with mass$^2 \propto V^{-1}$. This identification of the mass with the potential is responsible for making the dynamics converge to $V \to 0$. Indeed, the particle is moving without friction, along $V$ but it becomes very massive in regions where $V$ is very small, thus slowing down there.

Taking a logarithm we obtain an equivalent Hamiltonian $H = \log \hat{H}$, with the advantage of being separable:

$$H = \log(p^2 + 1) + \log(V). \tag{2}$$

The continuum dynamics of this Hamiltonian is given by the HJ equations

$$\dot{q}^i = \frac{2p_i}{1+p^2} \qquad \dot{p}^i = -\frac{\partial_i V}{V}\,. \tag{3}$$

Before considering explicit optimization algorithms constructed out of them, let us see some properties of this dynamics. Energy is conserved, meaning that the quantity

$$E = H = \log(p^2 + 1) + \log(V) \tag{4}$$

is constant during the evolution. If the system is mixing, after the mixing time the probability of finding the particle in a region of phase space is given by the volume of such region in phase space. Integrating over the momenta using the fact that the energy is conserved we get a probability in configuration space

$$\text{prob}(\Omega) \propto \int_\Omega \mathrm{d}q^n V^{-n/2} \tag{5}$$

where $n$ is the dimension of the configuration space.

We conclude by noticing a rewrite of the equations of motion that might require less computations. Since $e^E = V(p^2 + 1)$, we can rewrite the equations of motion as

$$\dot{q}^i = \frac{V}{e^E} 2p_i \qquad \dot{p}^i = -\frac{\partial_i V}{V}\,. \tag{6}$$

### 1.2 Optimizers

Suppose we are given a function $F(q)$ to minimize. Let us assume the global minimum is $F_0$. We can then relate $F$ to $V$ such that the Hamiltonian evolution will spend most of the time in the region where $F \sim F_0$.

A possible way to achieve this is to identify

$$V = (F - F_0)^\eta \,. \tag{7}$$

Doing so, the concentration of the measure near the minimum is enhanced by the power $\eta$ as

$$\text{prob}(\Omega) \propto \int_\Omega \mathrm{d}q^n (F - F_0)^{-\eta n/2} \tag{8}$$

In this section we are going to explore this choice in detail and construct optimizers based on this dynamics, but clearly more general choices are possible.

First of all, we can rewrite the continuum equations in terms of $F$, obtaining

$$\dot{q}^i = \frac{2p_i}{1+p^2}\,, \qquad \dot{p}^i = -\eta \frac{\partial_i F}{F - F_0}\,. \tag{9}$$

To integrate the equations we need an initial condition. The one for the $q$ is given by the initialization. This in turn implies an initial value for $V$, which we call $V(0)$. For the $p$, we choose to initialize along (minus) the initial gradient. Their norm is then related to the total energy $E$ through the energy equation (4) evaluated at initialization. More precise, define the constant $\delta E$ such that the total energy $E$ is given by

$$e^E \equiv V(0)(1 + \delta E)\,. \tag{10}$$

By construction $p^2(0) = \delta E$.

To recap, the momenta are initialized as

$$p^i(0) = \sqrt{\delta E} \frac{\partial_i V(0)}{|\partial V(0)|} \tag{11}$$

and the total energy of the system $E$ is obtained from (10).

### 1.3 Discrete dynamics

We are going to consider symplectic integration schemes for the dynamics (9). The reason to do so is that those preserve the symplectic form *exactly*, and thus volumes on phase space are exactly preserved by this discretization. An advantage of the separable Hamiltonian is that efficient second order symplectic integrators are readily available.

Suppressing the vector indices and using subscripts to denote the time-step, a second order symplectic integration is given by the following leapfrog scheme. Starting from the initial $(q_0, p_0)$ as described above, the first update is

$$p_{1/2} = p_0 - \frac{\Delta t}{2}\eta \frac{\partial F(q_0))}{F(q_0)} \tag{12}$$

$$q_1 = q_0 + 2\Delta t \frac{p_{1/2}}{1 + |p_{1/2}|^2} \tag{13}$$

Then, the other steps are

$$p_1 = p_{1/2} - \frac{\Delta t}{2}\eta \frac{\partial F(q_1))}{F(q_1)} \tag{14}$$

$$p_{3/2} = p_1 - \frac{\Delta t}{2}\eta \frac{\partial F(q_1))}{F(q_1)} \tag{15}$$

$$q_2 = q_1 + 2\Delta t \frac{p_{3/2}}{1 + |p_{3/2}|^2} \tag{16}$$

$\ldots$

**Remark 1.1.** Notice that each leapfrog step requires a single gradient evaluation, and we have decided to split them in this way since at a single step we only have $F(q_i)$ and $\partial(F(q_i)$.

**Remark 1.2.** If we do not need to access $p_i$ and $q_i$ at the same instant, we could combine the two momentum updates and obtain a simpler update. However, we do it in this way to be able to compute energy and monitor/restore it.

### 1.3.1 Enforcing energy conservation

After the first momentum update is performed, we have access to $p_i$ and $q_i$ at the same $i$, and we can compute whether the value of the energy at step $i$ agrees with the initial value by comparing the constant (10) to

$$E_i \equiv \log(1 + |p_i|^2) + \eta \log(F(q_i) - F_0). \tag{17}$$

At this stage, an option is to restore energy conservation by rescaling the momenta. More precisely, if energy were exactly conserved, after the first momentum update we should have

$$|p_i|^2 = \frac{e^E}{(F(q_i) - F_0)^\eta} - 1. \tag{18}$$

To enforce it, we can simply rescale all the $p_i$ homogeneously to achieve that. To recap, strict energy conservation can be enforced by the rescaling

$$p_i \to \frac{p_i}{|p_i|}\sqrt{\frac{e^E}{(F(q_i) - F_0)^\eta} - 1}, \tag{19}$$

performed after the first momentum update on each leapfrog step.

Enforcing strict energy conservation in this way might not really be necessary: if the system is well-behaved and $\Delta t$ not too big, the energy would fluctuate but not drift. This is another advantage of symplectic integration schemes. We leave this as an option in the optimizer.

## 1.4 Encouraging mixing

There are two different ways we can use to encourage mixing.

### 1.4.1 Bounces

This is a random rotation of the $p$, happening once in a while.

### 1.4.2 Generalized bounces

Slightly rotate $p_i$ by a random tiny amount at each step.

$$\Pi' = |\Pi| \frac{\frac{\Pi}{|\Pi|} + \nu z}{\left|\frac{\Pi}{|\Pi|} + \nu z\right|} \tag{20}$$

Slightly rotate $\Pi_i$ by a random tiny amount at each step.

$$p' = |p| \frac{\frac{p}{|p|} + \left(\frac{V(1+\delta E)}{e^E}\right)^\gamma \nu z}{\left|\frac{p}{|p|} + \left(\frac{V(1+\delta E)}{e^E}\right)^\gamma \nu z\right|} \tag{21}$$

Here $z$ is Gaussian random vector (all components are independently drawn from a standard Gaussian) and $\nu$ and $\gamma$ are hyperparameters. Notice that

$$1 \geqslant \frac{V(1+\delta E)}{e^E} \geqslant 0, \tag{22}$$

starting from 1 at initialization. Thus the role of the $\gamma$ parameter is to tune down the momentum decoherence when the minimum is being approached.

### 1.5 Weight decay

For a non-linear optimizer, weight decay and $L^2$ regularization act differently. The former is defined as exponential decay of the weights during training [1], while the latter is the addition of a $\theta^2$ term to the loss. While for SGD there is no difference between the two (with an appropriate rescaling) the difference appears already for Adam, as stressed in [2].

More precisely, the first way of implementing weight decay changes the update rule of the parameter as

$$\theta^i_{t+1} = \cdots - \Delta t w_{d0} \theta^i_t \tag{23}$$

where the dots denote the terms that would be there in the update rule if weight decay was zero. Since this deformation modifies the Hamiltonian dynamics it introduces an energy violation. This could be corrected when energy violation is enforced. The second one, being akin to an $L^2$ term, preserves energy. To implement it, we define a constant $w_d$ and modify the update rules by the shift

$$F(q) \to F(q) + \frac{w_d}{2} q^2 \qquad\qquad \partial_i F(q) \to \partial_i F(q) + w_d q_i \tag{24}$$

In the code, we have implemented this second, energy-conserving, choice.