École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland

Mathematics Section

School of Basic Sciences

Modeling Stochastic Gradient Descent with SDEs

— MATH-450 | Numerical Integration of Stochastic Differential Equations (Spring 2023)

By:
Eliott Van Dieren - 352934

Teacher and TA:
Dr. Adrian Blumenthal
Guillaume Wang

July 2023

1 Introduction

Stochastic Gradient Descent (SGD) is an optimization algorithm widely used in machine learning. It can be seen as a more computationally efficient variant of Gradient Descent (GD) that exploits composite structure of the function to optimize. In this project, we derive a SDE that approximates SGD. (This is in some sense an inverse problem to numerical integration, where the goal is to derive discrete algorithms to approximate SDEs.) This allows to get some intuition concerning SGD, at least in the small step-size regime.

Given a cost function $f : \mathbb{R}^p \to \mathbb{R}$, an initial value $x^0 \in \mathbb{R}^p$ and a step-size $h > 0$, GD constructs a sequence of iterates $(x^k)_{k\geq 0}$ by the update rule : $x^{k+1} = x^k - h\nabla f(x^k)$, where $\nabla f := (\frac{\partial f}{\partial x_1}, ..., \frac{\partial f}{\partial x_p}) : \mathbb{R}^p \to \mathbb{R}^p$ is the gradient of f.

2 From (S)GD to SDEs

Q1 Explain that GD is equivalent to the forward Euler method applied to the gradient flow (GF) differential equation: $\frac{dx}{dt} = -\nabla f(x(t))$ and $x(0) = x^0$. Suppose $f \in C^2(\mathbb{R}^p)$ and $\sup_{\mathbb{R}^p} \|\nabla^2 f\| < \infty$, let $(x^k)_k$ denote the GD iterates using stepsize h, and $x(t)$ the solution of GF. Show that, for any fixed $T > 0$, $\sup_{k\leq \lfloor T/h \rfloor} \|x^k - x(hk)\| = \mathcal{O}(h)$.

First, one intuitively finds that $dx = -\nabla f(x(t))dt$. Then, by definition of the forward Euler method, we get :

$$x^{k+1} - x^k = -h\nabla f(x^k)$$

with time step h, which is exactly the GD method.

For the second part of the question, we start by showing that

$$\|x^{k+1} - x(hk+h)\| \leq (1 + \mathcal{O}(h))\|x^k - x(hk)\| + \|x(hk) - h\nabla f(x(hk)) - x(hk+h)\|,$$

and that $\|x(hk) - h\nabla f(x(hk)) - x(hk+k)\| = \mathcal{O}(h^2)$.

To prove the first inequality, one starts with the definition of GD: by replacing x^{k+1}, we get

$$\|x^{k+1} - x(hk+h)\| = \|x^k - h\nabla f(x^k) - x(hk+h)\|,$$

then by sequentially adding and substrating $h\nabla f(x(hk)) + x(hk)$ inside the norm, and applying the triangular inequality, one gets:

$$\|x^k - h\nabla f(x^k) - x(hk+h)\| \leq \|x^k - x(hk)\| + \|x(hk) - h\nabla f(x(hk)) - x(hk+k)\| + h\|\nabla f(x(hk) - \nabla f(x^k)\|$$

Then, by applying the mean value theorem, we can bound $\|\nabla f(x(hk) - \nabla f(x^k)\|$ by $C\|x(hk) - x^k\|$, where $C = \max_{\mathbb{R}^p} \|\nabla^2 f\| < \infty$. Finally, by grouping terms, we get the desired result, as $Ch \in \mathcal{O}(h)$.

The second equality directly results from doing Taylor expansion on $x(hk+h)$ around hk, which cancels every term in the norm except its remainder which is in $\mathcal{O}(h^2)$, hence the norm in $\mathcal{O}(h^2)$.

Now to the main proof: thanks to those two results, one can see that

$$\|x^{k+1} - x(h(k+1))| \leq (1 + \mathcal{O}(h))\|x^k - x(hk)\| + \mathcal{O}(h^2)$$

then by recursion, we get

$$\|x^{k+1} - x(h(k+1))| \leq (1 + \mathcal{O}(h))^{k+1}\|x^0 - x(0)\| + \mathcal{O}(h) = \mathcal{O}(h) \text{ as } x^0 = x(0)$$

As this is true for every k, then it is true for its supremum below $\lfloor T/h \rfloor$, which yields the result.

Q2 Write the SGD update rule as $x^{k+1} = x^k - h\nabla f(x^k) + \sqrt{h}V^k$. Show that $\mathbb{E}[V^k|x^k] = 0$ and $\mathbb{E}[V^k V^{kT}|x^k] = h\Sigma(x^k)$ with $\Sigma(x) := \frac{1}{n}\sum_{i=1}^n [\nabla f(x) - \nabla f_i(x)][\nabla f(x) - \nabla f_i(x)]^\top$. Can you anticipate why, intuitively, it makes sense to approximate SGD by (1)?

By defining $V^k = \sqrt{h}\big(\nabla f(x^k) - \nabla f_{i_k}(x^k)\big)$, we get the equivalence. Now, one checks the first two moments of V^k. First, $\mathbb{E}[V^k|x_k] = \sqrt{h}\big(\nabla f(x^k) - \frac{1}{n}\sum_{i=1}^n f_i(x^k)\big) = 0$ by definition of f and the uniform distribution of i_k's. Secondly, one gets

$$\begin{aligned}
\mathbb{E}[V^k V^{k\top}|x_k] &= h\mathbb{E}\big[\big(\nabla f(x^k) - \nabla f_{i_k}(x^k)\big)\big(\nabla f(x^k) - \nabla f_{i_k}(x^k)\big)^\top |x^k\big] \\
&= h\left(\nabla f(x^k)\nabla f(x^k)^\top - \nabla f(x^k)\mathbb{E}[\nabla f_{i_k}(x^k)^\top|x^k] - \mathbb{E}[\nabla f_{i_k}(x^k)|x^k]\nabla f(x^k)^\top + \mathbb{E}[\nabla f_{i_k}(x^k)\nabla f_{i_k}(x^k)^\top|x^k]\right) \\
&= h\left(-\nabla f(x^k)\nabla f(x^k)^\top + \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\nabla f_i(x^k)\nabla f_i(x^k)^\top|x^k]\right) \\
&= h\Sigma(x^k) \text{ by definition of } \nabla f \text{ and } \Sigma(x).
\end{aligned}$$

Intuitively, it makes sense to approximate the SGD update rule by the SDE (1) as, if we define a Brownian motion with the same mean and variance as V^k, then we would, by using the EM method, obtain the same result by solving the SGD and the SDE. Furthermore, it is interesting to separate the deterministic (GD) part and the random part (containing V^k).

Q3 Let $\tilde{x}^1 = x^0 - h\nabla f(x^0) + \sqrt{h}\tilde{V}^0$ where $\tilde{V}^0 \sim \mathcal{N}(0, h\Sigma(x^0))$. That is $\tilde{x}^1$ is the iterate after one step of the Euler-Maruyama method with time-discretization $\Delta t = h$ applied to (1). Assume that $f_i \in C_b^9$ for all i and let any $\phi \in C_p^4$. Show that there exists $C, M > 0$ such that $|\mathbb{E}\phi(x^1) - \mathbb{E}\phi(\tilde{x}^1)| \le C\big(1 + |x^0|^M\big)h^2$.

One first rewrites $x^1 = x^0 - h\nabla f(x^0) + \sqrt{h}V^0 = x_t^0 + \sqrt{h}V^0$ and $\tilde{x}^1 = x_t^0 + \sqrt{h}\tilde{V}^0$. Then, by Taylor expansion on both terms around x_t^0, one gets (with abuse of notation on higher order vector multiplications):

$$\begin{aligned}
\phi(x^1) &= \phi(x_t^0) + \sqrt{h}\nabla\phi^\top(x_t^0)V^0 + \frac{h}{2}V^{0\top}\nabla^2\phi(x_t^0)V^0 + \frac{h^{\frac{3}{2}}}{6}D^3(\phi(x_t^0))(V^0)^3 + \frac{h^2}{24}D^4(\phi(\xi))(V^0)^4 \\
\phi(\tilde{x}^1) &= \phi(x_t^0) + \sqrt{h}\nabla\phi^\top(x_t^0)\tilde{V}^0 + \frac{h}{2}\tilde{V}^{0\top}\nabla^2\phi(x_t^0)\tilde{V}^0 + \frac{h^{\frac{3}{2}}}{6}D^3(\phi(x_t^0))(\tilde{V}^0)^3 + \frac{h^2}{24}D^4(\phi(\eta))(\tilde{V}^0)^4
\end{aligned}$$

Then, by taking the expectation on both equations and subtracting their values one gets

$$\begin{aligned}
\mathbb{E}[\phi(x^1) - \phi(\tilde{x}^1)] &= \frac{h}{2}\mathbb{E}[V^{0\top}\nabla^2\phi(x_t^0)V^0 - \tilde{V}^{0\top}\nabla^2\phi(x_t^0)\tilde{V}^0](= 0 \text{ as they have same variance}) \\
&+ \frac{h^{\frac{3}{2}}}{6}D^3(\phi(x_t^0))[\mathbb{E}[(V^0)^3 - (\tilde{V}^0)^3]](= \frac{h^{\frac{3}{2}}}{6}D^3(\phi(x_t^0))\mathbb{E}[(V^0)^3] \text{ by skewness of standard normal dist.} \\
&+ \frac{h^2}{24}D^4(\phi(\xi))\mathbb{E}[(V^0)^4] - \frac{h^2}{24}D^4(\phi(\eta))\mathbb{E}[(\tilde{V}^0)^4]
\end{aligned}$$

Now, using the hypothesis that $\phi \in C_p^4$, we will upper bound the term in $h^{\frac{3}{2}}$ and h^2. By rewriting $V^0 = \sqrt{h}(\nabla f(x^0) - \nabla f_{i_0}(x^0)) = \sqrt{h}M^0$, one first notes that M^0 is not dependent on h, nor its moments, and we get

1. $\frac{h^{\frac{3}{2}}}{6}D^3(\phi(x_t^0))\mathbb{E}[(V^0)^3] \le \frac{h^3}{6}C(1 + |x_t^0|^M)\mathbb{E}[(M^0)^3] \le \frac{h^3}{6}C(1 + |x^0|^M + \mathcal{O}(h))\mathbb{E}[(M^0)^3] \le \tilde{C}_1(1 + |x^0|^M)h^3 + \mathcal{O}(h^4) \in \mathcal{O}(h^3)$
2. $\frac{h^2}{24}D^4(\phi(\xi))\mathbb{E}[(V^0)^4] \le \frac{h^4}{24}C(1 + |\xi|^M)\mathbb{E}[(M^0)^4] \in \mathcal{O}(h^4)$
3. $\frac{h^2}{24}D^4(\phi(\eta))\mathbb{E}[(\tilde{V}^0)^4] \le \frac{h^2}{24}C(1 + |\eta|^M)\mathbb{E}[(\tilde{V}^0)^4] \le \tilde{C}_3(1 + |x^0|^M)h^2 + \mathcal{O}(h^3) \in \mathcal{O}(h^2)$

by assuming finite moments for M^0 and $\tilde{V}^0$. Therefore, we get the final result for h small

$$|\mathbb{E}[\phi(x^1)] - \mathbb{E}[\phi(\tilde{x}^1)]| \le \tilde{C}_3\left(1 + |x^0|^M\right)h^2.$$

Q4 Assume that $f_i \in C_b^9$ for all i. Show that for any $\phi \in C_p^4$, there exists $C > 0$ such that $\forall k \leq \lfloor T/h \rfloor$, $|\mathbb{E}[\phi(X_{hk})] - \mathbb{E}[\phi(x^k)]| \leq Ch$.

To answer this question, we use Theorem 2 from Chapter 6 of the lectures which will yield the result. We now have to prove the four assumptions to use the theorem. Regarding notations, the $f(x)$ in Th. 2 here is $-\nabla f(x)$, and $g(x) = \sqrt{h}\Sigma(x)^{\frac{1}{2}}$

1. We first prove that for all $k, l \in \{1, ..., p\}$, $(\nabla f(x))_k$ and $(\sqrt{h}\Sigma(x)^{\frac{1}{2}})_{k,l}$ are in C_b^4. First, $(\nabla f(x))_k \in C_b^4$ for all $0 \leq i \leq p$ as $f_i \in C_b^9$ and ∇f is a linear combination of ∇f_i's which are in C_b^8. Second, we find that $(\Sigma(x))_{k,l} \in C_b^8$ for the same reasons. The square root of the matrix $\Sigma(x)$ will also have element-wise bounded derivatives, by contradiction with $(\Sigma(x))_{k,l} \in C_b^8$.

2. Now, we show that the SGD has weak local order 1. From Lemma 2 in Lecture 12, we have for $\phi \in C_p^4$, $\nabla f, \sqrt{h}\Sigma(x)^{\frac{1}{2}} \in C_b^2$ and $X(t)$ being the solution of the SDE such that

   $$\mathbb{E}[\phi(X_h | X_0 = x^0)] = \phi(x^0) + h\mathcal{L}\phi(x) + R_1 \text{ where } \mathbb{E}[R_1] \leq C_1(1 + \mathbb{E}[|x^0|^M])h^2.$$

   Furthermore, we have to prove for the SGD that $\mathbb{E}[\phi(x^1)|x^0] = \phi(x^0) + h\mathcal{L}\phi(x^0) + R_2$, with the same differential operator $\mathcal{L}$ and the same bound structure for the remainder R_2. With that, it will yield the result with, e.g. $R = R_1 + R_2$ such that $\mathbb{E}[R] \leq \tilde{C}(1 + \mathbb{E}[|X_0|^k])h^2$.

   Knowing that the multi-dimensional Euler-Maruyama method has weak local order of convergence 1, we know that

   $$\mathbb{E}[\phi(\tilde{x}^1 | X(0) = x^0] = \phi(x^0) + h\mathcal{L}\phi(x^0) + R_3 \text{ with same operator } \mathcal{L} \text{ and bounds for } R_3,$$

   and that from **Q3**, one has by replacing $\mathbb{E}[\phi(x^1)|X_0 = x^0]$ with the desired form

   $$\begin{aligned} &|\mathbb{E}\phi(x^1) - \mathbb{E}\phi(\tilde{x}^1)| \leq C\left(1 + |x^0|^M\right)h^2 \\ \Leftrightarrow &\left|\mathbb{E}\left[\mathbb{E}[\phi(x^1)|X_0 = x^0] - (\phi(x^0) + h\mathcal{L}\phi(x^0) + R_3)\right]\right| \leq C\left(1 + |x^0|^M\right)h^2 \\ \Leftrightarrow &|\mathbb{E}[R_2 - R_3]| \leq C\left(1 + |x^0|^M\right)h^2 \end{aligned}$$

   This confirms the structure of $\phi(x^1)|X_0 = x^0] = \phi(x^0) + h\mathcal{L}\phi(x^0) + R_2$, with the desired bound on R_2, and the result follows.

3. Now, we prove that $\mathbb{E}[\|X_0\|^n] < \infty, \forall n \in \mathbb{N}^*$. This is immediate as $X_0 = x^0 \in \mathbb{R}^p$, by definition of the SDE.

4. Lastly, we want to prove $\mathbb{E}[x^{k+1} - x^k | x^k = x] \leq C(1 + \|x\|)h$ and $\|x^{k+1} - x^k\| \leq M_k(1 + \|x^k\|)\sqrt{h}$.

   The first statement can be proved as follows. Rewriting the difference of the two iterations, we have $x_{k+1} - x_k = -h\nabla f(x_k) + \sqrt{h}V^k$, which yields

   $$\mathbb{E}[x^{k+1} - x^k | x^k = x] = -\nabla f(x)h \leq C(1 + \|x\|)h$$

   by using the growth assumption on ∇f.

   The second statement is solved as follows:

   $$\begin{aligned} \|x^{k+1} - x^k\| &\leq h\|\nabla f(x^k)\| + \sqrt{h}\|V^k\| \\ &\leq C(1 + \|x^k\|)h + h\|\nabla f(x^k)\| + h\|\nabla f_{i_k}(x^k)\| \\ &\leq \tilde{C}(1 + \|x^k\|)h + h\|\nabla f_{i_k}(x^k)\| \\ &\leq \tilde{C}(1 + \|x^k\|)h + C_{i_k}(1 + \|x^k\|)h \text{ from } \nabla f_i \in C_b^8 \\ &\leq (1 + \|x^k\|^m)\sqrt{h}M_k \end{aligned}$$

   where $M_k = \sqrt{h}[\tilde{C} + C_{i_k}] \sim \text{Uniform}(\{\sqrt{h}[\tilde{C} + C_1], ..., \sqrt{h}[\tilde{C} + C_n]\})$, where C_{i_k} is the bound for the gradient of ∇f_{i_k} (comes from uniform distribution of i_k and $\nabla f_i \in C_b^8$).

As those assumptions stand, we get the desired result from the theorem.

3 Application to quadratic minimization

Q5 Let $M \in \mathbb{R}^{n \times p}, y \in \mathbb{R}^n$, and $f_i(x) = \frac{1}{2}|M_{i,\cdot}x - y_i|^2$. Check that $f(x) = \frac{1}{2n}\|Mx - y\|^2$ and $\Sigma(x) = \frac{1}{n}M^\top \left[\mathrm{Diag}(R)^2 - \frac{1}{n}RR^\top\right] M$ where $R = Mx - y$ and $\mathrm{Diag}(R)^2_{i,j} = \delta_{i,j}R_i^2$.

Starting with f, we get $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x) = \frac{1}{2n}\sum_{i=1}^n |M_{i,\cdot}x - y_i|^2 = \frac{1}{2n}\|Mx - y\|^2$. Now, we can rewrite $\Sigma(x)$ as

$$\begin{aligned}\Sigma(x) &= \frac{1}{n}\sum_{i=1}^n \left[\frac{1}{n}M^\top(Mx-y) - M_{i,\cdot}^\top(M_{i,\cdot}x - y_i)\right]\left[\frac{1}{n}M^\top(Mx-y) - M_{i,\cdot}^\top(M_{i,\cdot}x - y_i)\right]^\top \\ &= \frac{1}{n}\sum_{i=1}^n \frac{1}{n^2}M^\top(Mx-y)(Mx-y)^\top M - \frac{2}{n^2}\sum_{i=1}^n M^\top(Mx-y)M_{i,\cdot}|M_{i,\cdot}x - y_i| + \frac{1}{n}\sum_{i=1}^n |M_{i,\cdot}x - y_i|^2 M_{i,\cdot}^\top M_{i,\cdot} \\ &= \frac{1}{n^2}M^\top RR^\top M - \frac{2}{n^2}M^\top RR^\top M + \frac{1}{n}M^\top \mathrm{Diag}(R)^2 M \\ &= \frac{1}{n}M^\top \left[\mathrm{Diag}(R)^2 - \frac{1}{n}RR^\top\right] M \text{ which yields the result.}\end{aligned}$$

Q6 Consider the SDE

$$d\widetilde{X}_t = -\nabla f(\widetilde{X}_t)dt + \sqrt{h}\widetilde{\Sigma}(\widetilde{X}_t)^{1/2}dW_t \text{ where } \widetilde{\Sigma}(x) = \frac{1}{n}M^\top[\frac{1}{n}\|Mx - y\|^2 I]M,$$

with $\widetilde{X}_0 = 0$. Show that $\widetilde{X}_t$ converges to $x^* = \arg\min_{Mx=y}\|x^2\|$ in probability as $t \to \infty$ (for h small enough).

First, we show that $d\|\widetilde{X}_t - x^*\|^2 = -bf(\widetilde{X}_t)dt + \sigma f(\widetilde{X}_t)dW_t$ for some constant $b, \sigma > 0$. This answer is inspired from "Rethinking SGD's noise" (Pillaud-Vivien, L. & Bach, F., 2022). Let $\widetilde{\Sigma}^{-1/2} = \frac{1}{n}\|M\widetilde{X}_t - y\|M^\top$, then by applying Ito's lemma on $g(x) = \|x - x^*\|^2$, we get

$$d\|\widetilde{X}_t - x^*\| = 2(\widetilde{X}_t - x^*)^\top d\widetilde{X}_t + \sum_{k=1}^p d\langle \widetilde{X}^k\rangle_t \text{ by definition of } g.$$

The first term can be expanded by the definition of $d\widetilde{X}_t$, and $Mx^* = y$, such that

$$\begin{aligned}2(\widetilde{X}_t - x^*)^\top d\widetilde{X}_t &= 2(\widetilde{X}_t - x^*)^\top[-\nabla f(\widetilde{X}_t)dt + \sqrt{h}\widetilde{\Sigma}(\widetilde{X}_t)^{1/2}dW_t] \\ &= -\frac{2}{n}(M(\widetilde{X}_t - x^*))^\top(M\widetilde{X}_t - y)dt + \frac{2\sqrt{h}}{n}\|M\widetilde{X}_t - y\|(M(\widetilde{X}_t - x^*))^\top dW_t \\ &= -\frac{2}{n}\|M\widetilde{X}_t - y\|^2 dt + \frac{2\sqrt{h}}{n}\|M\widetilde{X}_t - y\|^2 dW_t, \text{ by definition of BM (see reference)} \\ &= -4f(\widetilde{X}_t)dt + 4\sqrt{h}f(\widetilde{X}_t)dW_t.\end{aligned}$$

The second term is given as

$$\begin{aligned}\sum_{k=1}^p d\langle \widetilde{X}^k\rangle_t &= \sum_{k=1}^p \frac{h}{n^2}\|M\widetilde{X}_t - y\|^2(M_{i,\cdot})^\top M_{i,\cdot}dt \\ &= \frac{h}{n^2}\|M\widetilde{X}_t - y\|^2 \operatorname{tr}(M^\top M)dt = \frac{2h}{n}\operatorname{tr}(M^\top M)f(\widetilde{X}_t)dt\end{aligned}$$

This yields the final result with $b = 4 - \frac{2h}{n}\operatorname{tr}(M^\top M)$ and $\sigma = 4\sqrt{h}$.

Now, we prove the second statement, inspired by Course notes from MIT 6-241j. We want to show

$$\begin{aligned}&\sigma_{min}(MM^\top) \leq \frac{\|M\widetilde{X}_t - y\|^2}{\|\widetilde{X}_t - x^*\|^2} \leq \sigma_{max}(MM^\top) \\ \Leftrightarrow &\sigma_{min}(MM^\top) \leq \left(\frac{\|M(\widetilde{X}_t - x^*)\|}{\|\widetilde{X}_t - x^*\|}\right)^2 \leq \sigma_{max}(MM^\top)\end{aligned}$$

By defining $v_t = \widetilde{X}_t - x^*$, and using the SVD decomposition on $M = UDV^T$, with $D = \text{Diag}([\sigma_1, ..., \sigma_p])$, the diagonal matrix containing the *decreasingly ordered* singular values of M, we have

$$\frac{\|Mv_t\|^2}{\|v_t\|^2} = \frac{\|UDV^T v_t\|^2}{\|v_t\|^2} = \frac{\|Dv_t\|^2}{\|v_t\|^2} = \frac{\sum_{i=1}^p \sigma_i^2 (v_t)_i^2}{\sum_{i=1}^p (v_t)_i^2} = \sum_{i=1}^p \sigma_i^2 (\tilde{v}_t)_i^2.$$

where we used the invariance of the l2-norm wrt to multiplication with unitary matrices, and the definition of the l2-norm for the last equality. Furthermore, we note that

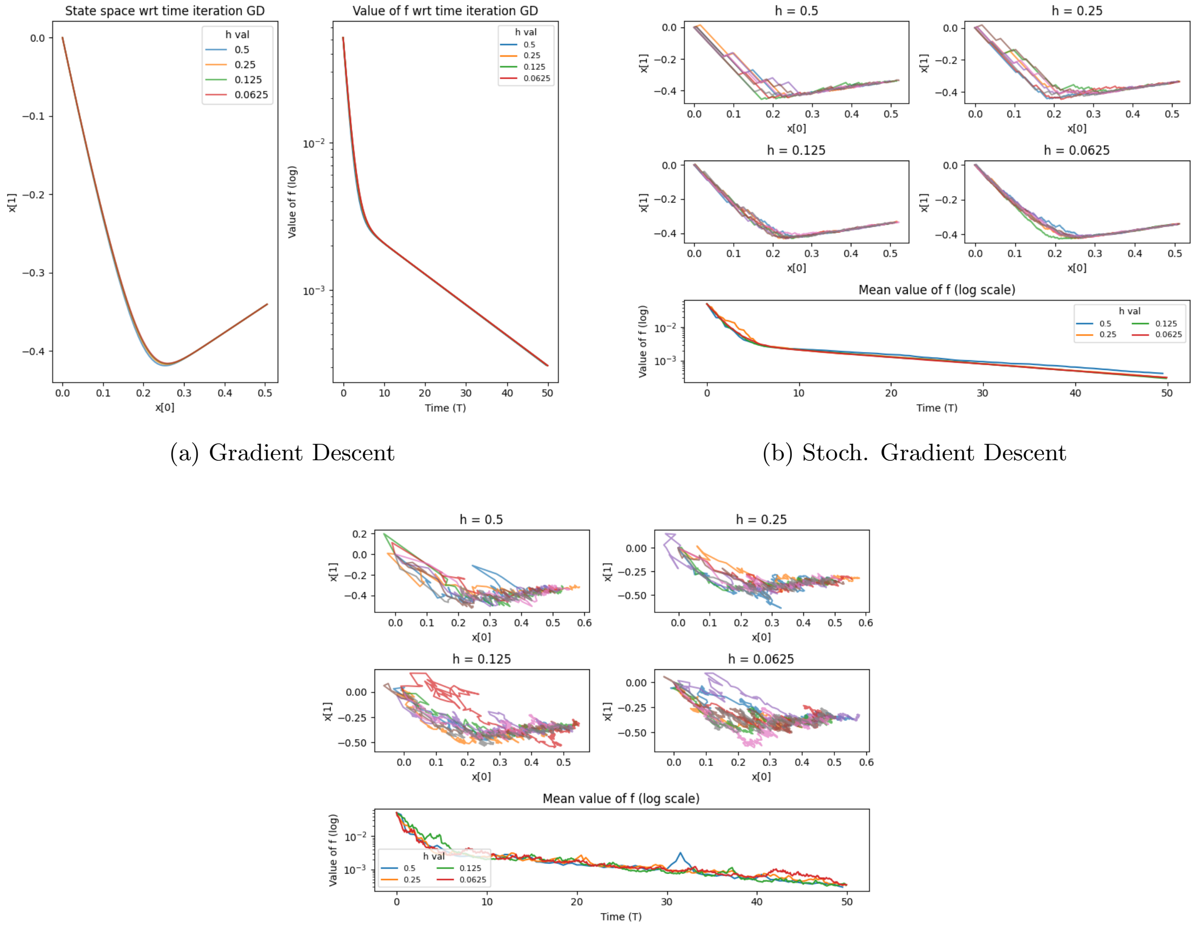
$$MM^T = UDV^\top (UDV^\top)^\top = UD^2V^\top,$$

which yields that the singular values of $MM^\top$ are $(\sigma_i^2)_{i=1,...p}$. To get the bonds, one takes respectively the maximum and minimum over $\tilde{v}_t$ for the upper and lower bound. For the minimum, $\tilde{v}_t^{*,min} = [0, ..., 0, 1]^\top \in \mathbb{R}^p$ as the singular values are ordered, and for for maximum, $\tilde{v}_t^{*,max} = [1, 0, ..., 0]^\top \in \mathbb{R}^p$. This yields the desired result

$$\frac{1}{2n}\sigma_{min}(MM^\top) \leq \frac{f(\widetilde{X}_t)}{\|\widetilde{X}_t - x^*\|^2} \leq \frac{1}{2n}\sigma_{max}(MM^\top).$$

Thanks to this result, we showed that the ratio between the objective function $f(\widetilde{X_t})$ is uniformly bounded, then as $b > 0$ for h small enough, we will get $\|\widetilde{X}_t - x^*\| \xrightarrow{p} 0$ as $t \to \infty$ (close to a GBM with negative drift).

Q7 Plot numerical results for GD, SGD and EM, varying the step size.



(a) Gradient Descent

(b) Stoch. Gradient Descent

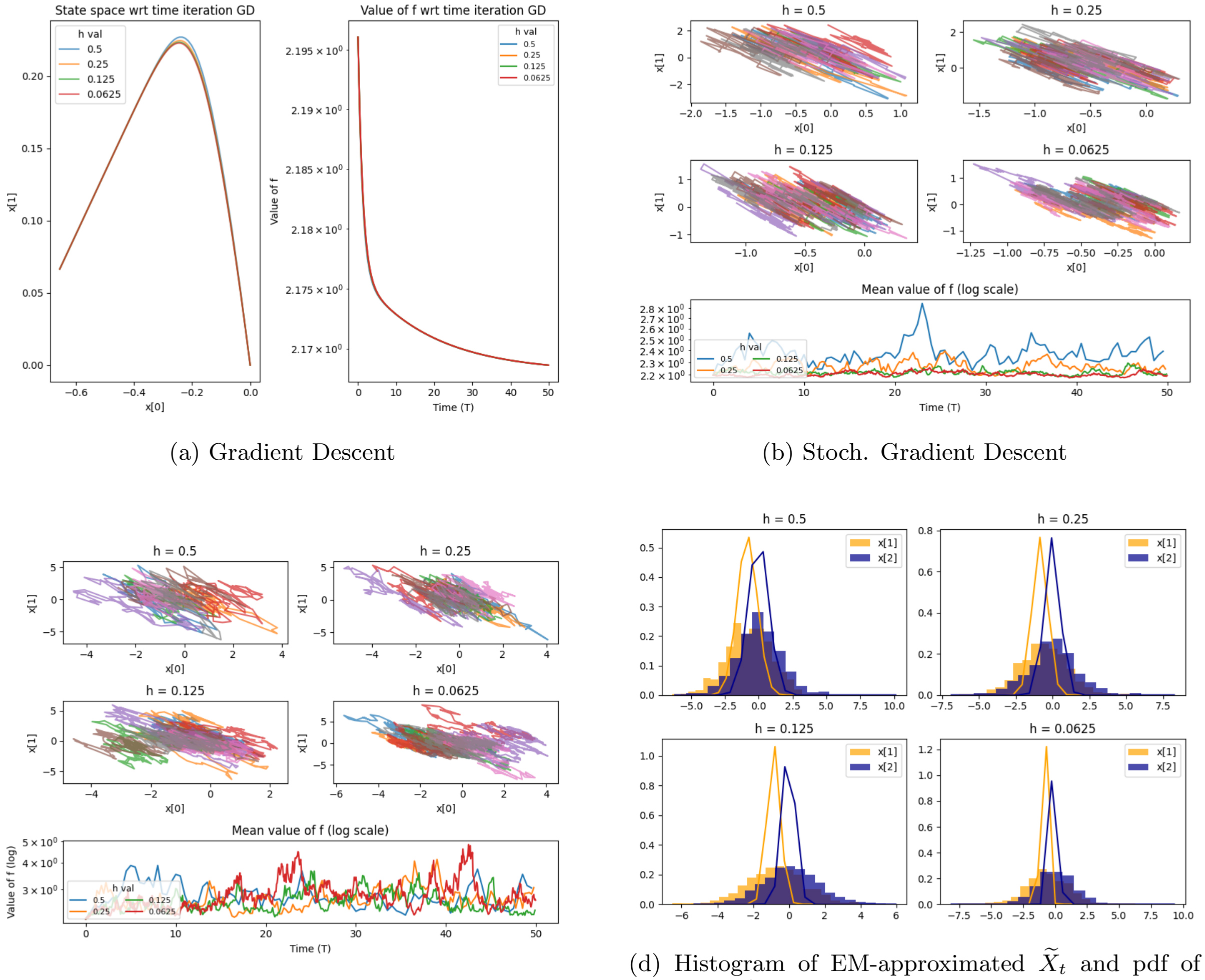
(c) Euler Maruyama

Figure 1: Comparison of state spaces and objective function for GD, SGD and EM.

Figure 1 shows a good convergence to the same state space, for each value of h, and a decrease of the objective value, as expected. We also see a smoother state space converge as h decreases. However, please note that for this question, we obtained a matrix with negative eigenvalues for $\Sigma(x^k)$, which produced either a complex squared root matrix or did not work with Cholesky (due to negative eigenvalues). Therefore, Figure 1c shows the computation over $\widetilde{X}_t$ instead of X_t, which did not cause any problem for the square root matrix.

Q8 Underparametrized problem and convergence of $\widetilde{X}_t$

With this new problem, Figure 2 shows a stable objective function value throughout the iterations and methods, with $f \approx 2$. Regarding the convergence of $\widetilde{X}_t$ in distribution to $\mathcal{N}(x^*, \sigma^2 I_p)$, Figure 2d shows the convergence wrt to the mean but the variance is bigger for the empirical results compared to the theoretical value of $\sigma^2 = \frac{h}{2} f(x^*)$.



(a) Gradient Descent

(b) Stoch. Gradient Descent

(c) Euler Maruyama

(d) Histogram of EM-approximated $\widetilde{X}_t$ and pdf of $\mathcal{N}(x^*, \sigma(h)^2 I_p)$

Figure 2: Comparison of state spaces and objective function for GD, SGD and EM for underparameterized problem, and histograms of EM-approximated $\widetilde{X}_t$ with respect to h.

4 Comparison with Langevin dynamics.

Q9 Explain that NGD is equivalent to the Euler-Maruyama method applied to the (damped) Langevin equation

Writing down the EM formula for the damped Langevin equation, we get:

$$x^{k+1} - x^k = -h\nabla f(x^k) + \sqrt{2\tau h}\Delta W_k \text{ where } \Delta W_k \sim \mathcal{N}(0, I_p)$$

with time step h, which is exactly the NGD method.

Q10 Write out the Langevin equation for $f(x) = \frac{1}{2n}\|Mx - y\|^2$. Solve it. Show that, if $M^\top M$ is invertible, the solution X_t converges in distribution to a Gaussian random variable, and specify the mean and variance of the limiting distribution.

First, we get $\nabla f(x) = \frac{1}{n}M^\top(Mx - y)$, which yields

$$dX_t = -\frac{1}{n}M^\top M X_t dt + \frac{1}{n}M^\top y dt + \sqrt{2\tau}dW_t.$$

Secondly, we can solve this equation inspired of the notes from Prof. R. Dalang (Theory of Stochastic Calculus, EPFL 2019). Let $\Phi(t)$ be the unique solution to deterministic differential equation

$$d\Phi(t) = -\frac{1}{n}M^\top M\Phi(t)dt,$$

one gets $\Phi(t) = \exp\left(\frac{-1}{n}M^\top Mt\right)$, which is strictly positive, and therefore that $\Phi^{-1}(t)$ is well defined for all $t \geq 0$. With this function Φ, we now show that

$$\begin{aligned}
X_t &= \Phi(t)\left(x_0 + \frac{1}{n}\int_0^t \Phi^{-1}(s)M^\top y ds + \int_0^t \Phi^{-1}(s)\sqrt{2\tau}dW_s\right)\\
&= \exp\left(\frac{-1}{n}M^\top Mt\right)\left(x_0 + \frac{1}{n}\int_0^t \exp\left(\frac{1}{n}M^\top Ms\right)M^\top y ds + \sqrt{2\tau}\int_0^t \exp\left(\frac{1}{n}M^\top Ms\right)dW_s\right)\\
&= \exp\left(\frac{-1}{n}M^\top Mt\right)(x_0 + V_t + M_t)
\end{aligned}$$

where V_t is of bounded variation, and M_t is a martingale (it is a BM), is the solution of the Langevin equation for $f(x) = \frac{1}{2n}\|Mx - y\|^2$. In other words, it must satisfy

$$X_t - x_0 = \int_0^t -\frac{1}{n}M^\top(MX_t - y)dt + \int_0^t \sqrt{2\tau}dW_t.$$

If we further define $f(\phi, v, m) = \phi(x_0 + v + m)$, we can apply Ito's formula on f, which yields

$$\begin{aligned}
X_t &= f(\Phi(t), V_t, M_t)\\
&= f(1, 0, 0) + \int_0^t (x_0 + V_s + M_s)d\Phi(s) + \int_0^t \Phi(s)dV_s + \int_0^t \Phi(s)dM_s\\
&= x_0 + \int_0^t \left(\frac{-1}{n}M^\top M\right)X_s\Phi^{-1}(s)\Phi(s)ds + \int_0^t \Phi(s)\Phi^{-1}(s)\frac{1}{n}M^\top y ds + \int_0^t \Phi(s)\Phi^{-1}(s)\sqrt{2\tau}dW_s\\
&= x_0 + \int_0^t -\frac{1}{n}M^\top(MX_t - y)ds + \int_0^t \sqrt{2\tau}dW_s, \text{ which concludes the result.}
\end{aligned}$$

From the solution

$$X_t = \exp\left(\frac{-1}{n}M^\top Mt\right)\left(x_0 + \frac{1}{n}\int_0^t \exp\left(\frac{1}{n}M^\top Ms\right)M^\top y ds + \sqrt{2\tau}\int_0^t \exp\left(\frac{1}{n}M^\top Ms\right)dW_s\right),$$

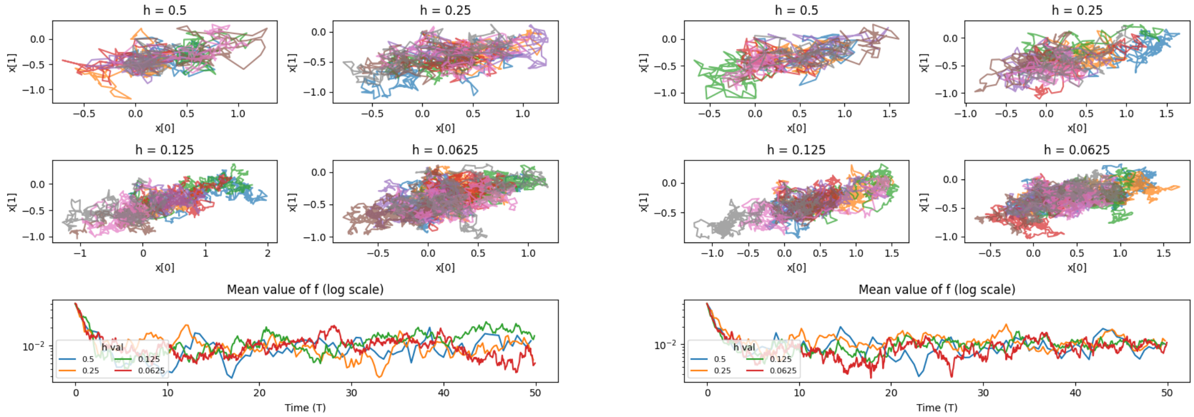
we first find that it follows a gaussian random variable by definition of the Ito integral M_t. With that, we can compute the mean of X_t as

$$\begin{aligned}
\mathbb{E}[X_t] &= \exp\left(\frac{-1}{n}M^\top Mt\right)\left(x_0 + \frac{1}{n}\int_0^t \exp(\frac{1}{n}M^\top Ms)M^\top y ds + \sqrt{2\tau}\mathbb{E}\left[\int_0^t \exp(\frac{1}{n}M^\top Ms)dW_s\right]\right) \\
&= \exp\left(\frac{-1}{n}M^\top Mt\right)\left(x_0 + \frac{1}{n}\int_0^t \exp(\frac{1}{n}M^\top Ms)ds M^\top y\right) \\
&= \exp\left(\frac{-1}{n}M^\top Mt\right)\left(x_0 + (M^\top M)^{-1}\left(\exp(\frac{1}{n}M^\top Mt) - I\right)M^\top y\right) \\
\Leftrightarrow \lim_{t\to\infty} \mathbb{E}[X_t] &= (M^\top M)^{-1}M^\top y = M^{-1}y,
\end{aligned}$$

and variance

$$\begin{aligned}
\mathrm{Var}[X_t] &= 2\tau \exp\left(\frac{-1}{n}M^\top Mt\right)\mathrm{Var}\left[\int_0^t \exp(\frac{1}{n}M^\top Ms)dW_s\right]\exp\left(\frac{-1}{n}M^\top Mt\right)^\top \\
&= 2\tau \exp\left(\frac{-1}{n}M^\top Mt\right)\mathbb{E}\left[\int_0^t \exp(\frac{2}{n}M^\top Ms)ds\right]\exp\left(\frac{-1}{n}M^\top Mt\right) \\
&= n\tau\left[\exp(\frac{1}{n}M^\top Mt) - \exp\left(\frac{-1}{n}M^\top Mt\right)\right](M^\top M)^{-1}\exp\left(\frac{-1}{n}M^\top Mt\right) \\
\Leftrightarrow \lim_{t\to\infty} \mathrm{Var}[X_t] &= n\tau(M^\top M)^{-1}
\end{aligned}$$

Q11 Reproduce **Q7** with noisy Gradient Descent on the new SDE



(a) Noisy Gradient Descent

(b) Euler-Maruyama

Figure 3: Comparison of state spaces and objective function for NGD and EM for Langevin equation with damping $\tau = 0.01$.

From Figure 3, we observe similar noisy state spaces, and the decrease up to a plateau at $f \approx 10^{-2}$ for the objective function. The GD did not change from previous plots, and is therefore not in this figure.

Regarding the interpretation of τ, one sees from the SDE that τ is the multiplicative coefficient of the noise (described by the Brownian motion increment dW_t). Physically, a particle is more excited when temperature increases, which matches the behaviour of X_t for the Langevin equation.