# Stochastic Gradient Descent

Bulgarelli Luca and Ferrera Alessandro

## Introduction

The Stochastic Gradient Descent (SGD) is a method used to minimize a differentiable objective function. It is an optimization algorithm that updates the parameters of a model in order to minimize a loss function. The algorithm is called stochastic because it uses a randomly selected subset of the training data to compute the gradient at each iteration. This is in contrast to the standard gradient descent algorithm, which uses the entire training set to compute the gradient at each iteration. The stochastic gradient descent algorithm is often used in machine learning and deep learning applications, where the training data sets are very large and it is computationally expensive to compute the gradient using the entire training set.

## Q1.1

The sarting point of our analysis is the Gradient Descent algorithm, which can be seen as a discretization of the Gradient Flow differential equation, in particular applying the forward Euler method to the ODE. A generic forward Euler method for an ODE of the form $x'(t) = g(x(t))$ is given by the formula:

$$x^{k+1} = x^k + hg(x^k)$$

where $h$ is the step size and $x^0 = x(0)$. If we consider $g = -\nabla f$, we have that:

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

which is the formula for the gradient descent method.

## Q1.2

The Forward Euler method is a first order method, which means that the error of the method is $O(h)$. We want to prove this claim by induction. Let's consider the base case $k = 0$. We have that:

$$\|x^0 - x(0)\| = 0 = O(h)$$

Now let's consider the inductive step. The induction hypothesis is that $\|x^k - x(hk)\| = O(h)$. We want to prove that $\|x^{k+1} - x(hk + h)\| = O(h)$. We have that:

$$\begin{aligned}\|x^{k+1} - x(hk+h)\| &= \|x^k - h\nabla f(x^k) - x(hk) + x(hk) - x(hk+h) + h\nabla f(x(hk)) - h\nabla f(x(hk))\| \leq \\ &\leq \|x^k - x(hk) - h\nabla f(x^k) + h\nabla f(x(hk))\| + \|x(hk) - x(hk+h) - h\nabla f(x(hk))\| = \\ &= \|x^k - x(hk) - h\nabla f(x^k) + h\nabla f(x(hk))\| + \|x(hk+h) - x(hk) + h\nabla f(x(hk))\|\end{aligned}$$

Considering that by hypothesis $f \in C^2$, $\nabla f$ is Lipschitz continuous on $[0, T]$ with constant $L$, we have that $\|h\nabla f(x^k) - h\nabla f(x(hk))\| \leq hL\|x^k - x(hk)\|$. This implies that:

$$\begin{aligned}\|x^k - x(hk) - h\nabla f(x^k) + h\nabla f(x(hk))\| &\leq \|x^k - x(hk)\| + hL\|x^k - x(hk)\| = \\ &= (1 + hL)\|x^k - x(hk)\| = (1 + O(h))\|x^k - x(hk)\|\end{aligned}$$

Moreover, using Taylor expansion and the previous point, we have that:

$$\|x(hk+h) - x(hk) + h\nabla f(x(hk))\| = \|x(hk) + hx'(hk) + \frac{h^2}{2}x''(\xi) - x(hk) + h\nabla f(x(hk))\| =$$
$$= \|x(hk) - h\nabla f(x(hk)) + \frac{h^2}{2}\nabla^2 f(x(\xi))\nabla f(x(\xi)) - x(hk) + h\nabla f(x(hk))\| \leq$$
$$\leq \frac{h^2}{2}\|\nabla^2 f(x(\xi))\|\|\nabla f(x(\xi))\| \leq h^2\frac{LB}{2} = O(h^2)$$

Using the induction hypothesis, we have that:

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

If we consider a function which is the mean of $n$ functions, we can construct the Stochastic Gradient Descent by applying the GD with a randomly selected function at each iteration, which can be written as $x^{k+1} = x^k - h\nabla f(x^k) + \sqrt{h}V^k$ and it can be approximated with a SDE as the next steps will show.

# Q2.1

First of all, we want to investigate the distribution of $V^k$. From $x^{k+1} = x^k - h\nabla f(x^k) + \sqrt{h}V^k$, we have that:

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

which gives:

$$V^k = \sqrt{h}(\nabla f(x^k) - \nabla f_{i_k}(x^k))$$

Let's consider the conditional expectation of $V^k$ given $x^k$:

$$\mathbb{E}[V^k|x^k] = \mathbb{E}[\sqrt{h}(\nabla f(x^k) - \nabla f_{i_k}(x^k))|x^k] = \sqrt{h}\nabla f(x^k) - \sqrt{h}\mathbb{E}[\nabla f_{i_k}(x^k)|x^k] =$$
$$= \sqrt{h}\nabla f(x^k) - \sqrt{h}\frac{1}{n}\sum_{i=0}^{n}\nabla f_i(x^k) = 0$$

As for the variance, we have that:

$$\mathbb{E}[V^k V^{k\top}|x^k] = \mathbb{E}[h(\nabla f(x^k) - \nabla f_{i_k}(x^k))(\nabla f(x^k) - \nabla f_{i_k}(x^k))^\top|x^k] =$$
$$= \frac{h}{n}\sum_{i=1}^{n}(\nabla f(x^k) - \nabla f_i(x^k))(\nabla f(x^k) - \nabla f_i(x^k))^\top$$

since, when conditioning over $x^k$, the only random component is $i_k$, distributed as a uniform random variable over $\{1, \dots, n\}$.

# Q2.2

The approximation of the method with an SDE makes sense with the following reasoning. Our scheme is:

$$x^{k+1} - x^k = -h\nabla f(x^k) + \sqrt{h}V^k = -h\nabla f(x^k) + \sqrt{h}\Sigma^{1/2}(x^k)U^k$$

where $U^k$ is a random vector with $\mathbb{E}[U^k|x^k] = 0$ and $\mathbb{E}[U^k U^{k\top}|x^k] = hI$. For $h \to 0$, we have that:

$$x^{k+1} - x^k \approx dX_t, \quad h \approx dt, \quad U^k \approx dW_t$$

supposing that the distribution of $U^k$ converges to a Gaussian for $h \to 0$. This way we get the SDE:

$$dX_t = -\nabla f(X_t)dt + \sqrt{h}\Sigma(X_t)^{1/2}dW_t$$

The SDE we are considering does not have a simple closed form solution, but we can approximate it with the Euler-Maruyama scheme. Thus we need to assess the robustness of this approximation.

## Q3.1

Let's begin by analysing the distribution of the first iterates of the scheme. Considering that $\tilde{x}^1$ is a linear transformation of $\tilde{V}^0$, which is normally distributed with mean 0 and variance $h\Sigma(x^0)$, we have that

$$\tilde{x}^1 \sim \mathcal{N}(x^0 - h\nabla f(x^0), h^2\Sigma(x^0))$$

In the same way we can obtain the distribution of $\tilde{x}^2$, conditioning on $\tilde{x}^1$. In this case we have that:

$$\tilde{x}^2 = \tilde{x}^1 - h\nabla f(\tilde{x}^1) + \sqrt{h}\tilde{V}^1 \quad \text{where} \quad \tilde{V}^1 \sim \mathcal{N}(0, h\Sigma(\tilde{x}^1))$$

so that:

$$\tilde{x}^2|\tilde{x}^1 \sim \mathcal{N}(\tilde{x}^1 - h\nabla f(\tilde{x}^1), h^2\Sigma(\tilde{x}^1))$$

## Q3.2

We then want to derive a bound for the error due to the approximation of SGD with the Euler-Maruyama scheme. We have:

$$x^1 = x^0 - h\nabla f(x^0) + \sqrt{h}V_0$$
$$\tilde{x}^1 = x^0 - h\nabla f(x^0) + \sqrt{h}\tilde{V}_0$$

We pose $\overline{x} = x^0 - h\nabla f(x^0)$, and $V_0 = \sqrt{h}W_0$, $\tilde{V}_0 = \sqrt{h}\tilde{W}_0$. The above equations become:

$$x^1 = \overline{x} + hW_0 \quad \text{with} \quad \mathbb{E}[W_0] = 0 \quad \text{and} \quad \mathbb{E}[W_0 W_0^\top] = \Sigma(x^0)$$
$$\tilde{x}^1 = \overline{x} + h\tilde{W}_0 \quad \text{with} \quad \tilde{W}_0 \sim \mathcal{N}(0, \Sigma(x^0))$$

We can now write the Taylor expansions of $\phi(x^1)$ and $\phi(\tilde{x}^1)$ around $\overline{x}$, with Lagrange remainder:

$$\phi(x^1) = \phi(\overline{x}) + h\nabla\phi(\overline{x})^\top W_0 + \frac{h^2}{2}W_0^\top\nabla^2\phi(\overline{x})W_0 + \frac{h^3}{6}\mathrm{D}^3\phi(\xi_1)\colon W_0^{\otimes 3}$$
$$\text{with} \quad \xi_1 = \overline{x} + \theta_1 hW_0 \quad \text{for some} \quad \theta_1 \in [0,1]$$
$$\phi(\tilde{x}^1) = \phi(\overline{x}) + h\nabla\phi(\overline{x})^\top \tilde{W}_0 + \frac{h^2}{2}\tilde{W}_0^\top\nabla^2\phi(\overline{x})\tilde{W}_0 + \frac{h^3}{6}\mathrm{D}^3\phi(\xi_2)\colon \tilde{W}_0^{\otimes 3}$$
$$\text{with} \quad \xi_2 = \overline{x} + \theta_2 h\tilde{W}_0 \quad \text{for some} \quad \theta_2 \in [0,1]$$

We can now compute:

$$|\mathbb{E}[\phi(x^1)] - \mathbb{E}[\phi(\tilde{x}^1)]| = \Big| h\nabla\phi(\overline{x})^\top\mathbb{E}[W_0 - \tilde{W}_0] + \frac{h^2}{2}\left(\mathbb{E}[W_0^\top\nabla^2\phi(\overline{x})W_0] + \right.$$
$$\left. - \mathbb{E}[\tilde{W}_0^\top\nabla^2\phi(\overline{x})\tilde{W}_0]\right) + \frac{h^3}{6}\left(\mathbb{E}[\mathrm{D}^3\phi(\xi_1)\colon W_0^{\otimes 3}] - \mathbb{E}[\mathrm{D}^3\phi(\xi_2)\colon \tilde{W}_0^{\otimes 3}]\right)\Big| =$$
$$= \frac{h^3}{6}\left|\mathbb{E}[\mathrm{D}^3\phi(\xi_1)\colon W_0^{\otimes 3}] - \mathbb{E}[\mathrm{D}^3\phi(\xi_2)\colon \tilde{W}_0^{\otimes 3}]\right| \leq$$
$$\leq \frac{h^3}{6}\left(\left|\mathbb{E}[\mathrm{D}^3\phi(\xi_1)\colon W_0^{\otimes 3}]\right| + \left|\mathbb{E}[\mathrm{D}^3\phi(\xi_2)\colon \tilde{W}_0^{\otimes 3}]\right|\right)$$

since $W_0$ and $\tilde{W}_0$ have 0 mean and the same covariance matrix, in particular:

$$\mathbb{E}[W_0^\top\nabla^2\phi(\overline{x})W_0] - \mathbb{E}[\tilde{W}_0^\top\nabla^2\phi(\overline{x})\tilde{W}_0] = \mathbb{E}\left[\mathrm{Tr}\left(\nabla^2\phi(\overline{x})W_0W_0^\top\right)\right] +$$
$$-\mathbb{E}\left[\mathrm{Tr}\left(\nabla^2\phi(\overline{x})\tilde{W}_0\tilde{W}_0^\top\right)\right] =$$
$$= \mathrm{Tr}\left(\nabla^2\phi(\overline{x})\mathbb{E}[W_0W_0^\top]\right) - \mathrm{Tr}\left(\nabla^2\phi(\overline{x})\mathbb{E}[\tilde{W}_0\tilde{W}_0^\top]\right) =$$
$$= \mathrm{Tr}\left(\nabla^2\phi(\overline{x})\Sigma(x^0)\right) - \mathrm{Tr}\left(\nabla^2\phi(\overline{x})\Sigma(x^0)\right) = 0$$

for the linearity of the trace operator.

Now let's focus on one of the third order term:

$$\left|\mathbb{E}[\mathrm{D}^3\phi(\xi_2)\colon \tilde{W}_0^{\otimes 3}]\right| \leq \mathbb{E}\left[\left|\mathrm{D}^3\phi(\xi_2)\colon \tilde{W}_0^{\otimes 3}\right|\right] \leq \mathbb{E}\left[\|\mathrm{D}^3\phi(\xi_2)\|_F \|\tilde{W}_0^{\otimes 3}\|_F\right] \leq$$
$$\leq \mathbb{E}\left[C_1(1+\|x^0 - h\nabla f(x^0) + \theta_2 h\tilde{W}_0\|^m)\|\tilde{W}_0\|^3\right]$$

where we used the polinomial growth $\mathrm{D}^3\phi(x) < C_1(1+\|x\|^m)$ and the equality $\|\tilde{W}_0^{\otimes 3}\|_F = \|\tilde{W}_0\|^3$. Now we will exploit the fact that:

$$\|x^0 - h\nabla f(x^0) + \theta_2 h\tilde{W}_0\|^m \leq C_2(\|x^0\|^m + \|h\nabla f(x^0)\|^m + \|\theta_2 h\tilde{W}_0\|^m) \quad \text{for some} \quad C_2 > 0$$

So, using also the bound $\|\nabla f(x)\| < C_3$, we have:

$$\mathbb{E}\left[C_1(1+\|x^0 - h\nabla f(x^0) + \theta_2 h\tilde{W}_0\|^m)\|\tilde{W}_0\|^3\right] \leq$$
$$\leq \mathbb{E}\left[C_1(1+C_2(\|x^0\|^m + \|h\nabla f(x^0)\|^m + \|\theta_2 h\tilde{W}_0\|^m))\|\tilde{W}_0\|^3\right] \leq$$
$$\leq \mathbb{E}\left[C_1(1+C_2(\|x^0\|^m + (C_3 h)^m + (\theta_2 h)^m\|\tilde{W}_0\|^m))\|\tilde{W}_0\|^3\right] =$$
$$= C_1\left(1+C_2(\|x^0\|^m + (C_3h)^m)\right)\mathbb{E}\left[\|\tilde{W}_0\|^3\right] + C_1C_2(\theta_2 h)^m\mathbb{E}\left[\|\tilde{W}_0\|^{m+3}\right]$$

Since $\tilde{W}_0$ is a Gaussian random variable, we have that $\mathbb{E}\left[\|\tilde{W}_0\|^3\right]$ and $\mathbb{E}\left[\|\tilde{W}_0\|^{m+3}\right]$ are finite. Moreover exploiting that $h < 1$ and $\theta_2 \in [0,1]$, we have than:

$$C_1\left(1+C_2(\|x^0\|^m + (C_3h)^m)\right)\mathbb{E}\left[\|\tilde{W}_0\|^3\right] + C_1C_2(\theta_2 h)^m\mathbb{E}\left[\|\tilde{W}_0\|^{m+3}\right] \leq$$
$$\leq K_1(1+\|x^0\|^m) \quad \text{for some} \quad K_1 > 0$$

We can prove the same bound for the other term, since $W$ is a discrete random variable with a finite support. Finally we have:

$$|\mathbb{E}[\phi(x^1)] - \mathbb{E}[\phi(\tilde{x}^1)]| \leq \frac{h^3}{6}\left(\left|\mathbb{E}[\mathrm{D}^3\phi(\xi_1)\colon W_0^{\otimes 3}]\right| + \left|\mathbb{E}[\mathrm{D}^3\phi(\xi_2)\colon \tilde{W}_0^{\otimes 3}]\right|\right) \leq$$
$$\leq \frac{h^3}{6}\left(K_2(1+\|x^0\|^m) + K_1(1+\|x^0\|^m)\right) = C\ (1+\|x^0\|^m)\ h^3$$

We showed that the bound holds for $p = 3$.

This result yields the following bound for the weak error of the approximation of the SGD with the SDE.

# Q4

We have that

$$|\mathbb{E}\phi(X_{hk}) - \mathbb{E}\phi(x^k)| \leq |\mathbb{E}\phi(X_{hk}) - \mathbb{E}\phi(\tilde{x}^k)| + |\mathbb{E}\phi(\tilde{x}^k) - \mathbb{E}\phi(x^k)| \leq$$
$$\leq C_1 h + C_2(1+|x^0|^M)h^3 \leq Ch$$

using the result from the previous point generalized and the weak convergence order of the Euler-Maruyama scheme.

Let's focus on a particular choice of $f$. Given $f_i(x) = \frac{1}{2}\|M_i x - y_i\|^2$, we can now derive $f(x)$ and $\Sigma(x)$.

# Q5.1

We have that:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left|M_{i\bullet}x - y_i\right|^2 = \frac{1}{2n}\sum_{i=1}^{n}\left|\sum_{j=1}^{p}(M_{ij}x_j) - y_i\right|^2 = \frac{1}{2n}\sum_{i=1}^{n}|(Mx)_i - y_i|^2 =$$

$$= \frac{1}{2n}\sum_{i=1}^{n}|(Mx-y)_i| = \frac{1}{2n}\|Mx-y\|^2$$

Calculating the gradient of $f_i$ with respect to $x$, we have that:

$$\nabla f_i(x) = (Mx-y)_i M_{i\bullet}^\top = R_i M_{i\bullet}^\top$$

and rewriting $f$ as:

$$f(x) = \frac{1}{2n}(Mx-y)^\top(Mx-y) = \frac{1}{2n}(x^\top M^\top Mx - 2y^\top Mx + y^\top y)$$

we obtain that:

$$\nabla f(x) = \frac{1}{2n}(2M^\top Mx - 2M^\top y) = \frac{1}{n}M^\top R$$

We can now compute:

$$\Sigma(x) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n}M^\top R - R_i M_{i\bullet}^\top\right)\left(\frac{1}{n}M^\top R - R_i M_{i\bullet}^\top\right)^\top =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n^2}M^\top RR^\top M - \frac{1}{n}M^\top RR_i M_{i\bullet} - \frac{1}{n}R_i M_{i\bullet}RM^\top + M_{i\bullet}^\top R_i^2 M_{i\bullet} =$$

$$= \frac{1}{n}\left(\frac{1}{n}M^\top RR^\top M - \frac{2}{n}M^\top R\sum_{i=1}^{n}R_i M_{i\bullet} + \sum_{i=1}^{n}M_{i\bullet}^\top R_i^2 M_{i\bullet}\right) =$$

$$= \frac{1}{n}\left(\frac{1}{n}M^\top RR^\top M - \frac{2}{n}M^\top RR^\top M + M^\top \mathrm{Diag}(R)^2 M\right) =$$

$$= \frac{1}{n}M^\top\left[\mathrm{Diag}(R)^2(x) - \frac{1}{n}R(x)R^\top(x)\right]M$$

# Q5.2

Since the SDE with such parameters is quite difficult to analyse, we will focus on a different SDE:

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

Our goal now is to have a SDE for the mean squared error $\|e_t\|^2 = \|\tilde{X}_t - x^*\|^2$, with $x^* = \mathrm{argmin}_{Mx=y}\|x\|^2$. To achieve that, we will show the following results:

(a) We need to prove that $\tilde{W}_t$ verifies the three defining properties of a Brownian motion:

- $\tilde{W}_0 = 0$ is trivially verified.
- $\forall s \in [0,t]$ $\tilde{W}_t - \tilde{W}_s \sim \mathcal{N}(0, t-s)$ is verified since it is a linear tranformation of $W_t - W_s$ which is normally distributed. Moreover:

$$\mathbb{E}[\tilde{W}_t - \tilde{W}_s] = M(M^\top M)^{-\frac{1}{2}}\mathbb{E}[W_t - W_s] = 0$$

and the variance remains unchanged since $M(M^\top M)^{-\frac{1}{2}}$ is an orthogonal matrix:

$$\begin{aligned}\mathrm{Cov}(\tilde{W}_t - \tilde{W}_s) = \mathrm{Cov}(M(M^\top M)^{-\frac{1}{2}}(W_t - W_s)) = \\ = M(M^\top M)^{-\frac{1}{2}}\mathrm{Cov}(W_t - W_s)(M(M^\top M)^{-\frac{1}{2}})^\top = M(M^\top M)^{-\frac{1}{2}} I(t-s)(M(M^\top M)^{-\frac{1}{2}})^\top = \\ = (t-s)M(M^\top M)^{-\frac{1}{2}}(M^\top M)^{-\frac{1}{2}}M^\top = (t-s)M(M^\top M)^{-1}M^\top = \\ = (t-s)MM^{-1}M^{-\top}M^\top = (t-s)I\end{aligned}$$

- $\forall s \in [0,t]$ $\tilde{W}_t - \tilde{W}_s$ is independent of $\tilde{W}_u$ $\forall u \leq s$:

$$\mathbb{E}[(\tilde{W}_t - \tilde{W}_s)\tilde{W}_u^\top] = M(M^\top M)^{-\frac{1}{2}}\mathbb{E}[(W_t - W_s)W_u](M^\top M)^{-\frac{1}{2}}M^\top = 0$$

by the indipendence of the increments of $W_t$, and since uncorellation for Gaussian variables implies independence.

(b) We can now derive the SDE exploiting Ito formula on $\|e_t\|^2$:

$$\nabla\|e_t\|^2 = 2(\tilde{X}_t - x^*) \quad \nabla^2\|e_t\|^2 = 2I$$
$$d<\tilde{X}>_t = h\tilde{\Sigma}(\tilde{X}_t)^{\frac{1}{2}}\tilde{\Sigma}(\tilde{X}_t)^{\frac{1}{2}\top}dt = h\tilde{\Sigma}(\tilde{X}_t)dt = \frac{h}{n^2}\|M\tilde{X}_t - y\|^2(M^\top M)dt$$

$$\begin{aligned}d\|e_t\|^2 = \nabla\|e_t\|^{2\top}d\tilde{X}_t + \frac{1}{2}\nabla^2\|e_t\|^2 : d<\tilde{X}>_t = 2(\tilde{X}_t - x^*)^\top\left(-\nabla f(\tilde{X}_t)dt + \sqrt{h}\tilde{\Sigma}(\tilde{X}_t)^{\frac{1}{2}}dW_t\right) + \\ + \frac{1}{2}2I : \left(\frac{h}{n^2}\|M\tilde{X}_t - y\|^2 M^\top M dt\right) = -\frac{2}{n}(\tilde{X}_t - x^*)^\top M^\top R(\tilde{X}_t)dt + \\ + \frac{2\sqrt{h}}{n}(\tilde{X}_t - x^*)^\top\|M\tilde{X}_t - y\|(M^\top M)^{\frac{1}{2}}dW_t + \frac{h}{n}\mathrm{Tr}(M^\top M)\frac{1}{n}\|M\tilde{X}_t - y\|^2 dt = \\ = -\frac{2}{n}R(\tilde{X}_t)^\top R(\tilde{X}_t)dt + \frac{2\sqrt{h}}{n}\|M\tilde{X}_t - y\|(\tilde{X}_t - x^*)^\top(M^\top M)^{\frac{1}{2}}(M^\top M)^{\frac{1}{2}}(M^\top M)^{-\frac{1}{2}}dW_t + \\ + \frac{2h}{n}\mathrm{Tr}(M^\top M)f(\tilde{X}_t)dt = -2\left(2 - \frac{h}{n}\mathrm{Tr}(M\top M)\right)f(\tilde{X}_t)dt + \\ + \frac{2\sqrt{h}}{n}\|M\tilde{X}_t - y\|^2\frac{1}{\|R(\tilde{X}_t)\|}R(\tilde{X}_t)^\top d\tilde{W}_t = -2(2 - \frac{h}{n}\mathrm{Tr}(M\top M))f(\tilde{X}_t)dt + 4\sqrt{h}f(\tilde{X}_t)d\overline{W}_t\end{aligned}$$

where $\overline{W}_t = \frac{\langle R(\tilde{X}_t), \tilde{W}_t\rangle}{\|R(\tilde{X}_t)\|}$ is a Brownian motion, as [1] states.

(c) Finally, we want to state the convergence in probability of $X_t$ to the true solution, under some conditions on $h$. We start by noticing that:

$$\begin{aligned}\frac{f(\tilde{X}_t)}{\|\tilde{X}_t - x^*\|^2} = \frac{1}{2n}\frac{\|M\tilde{X}_t - y\|^2}{\|\tilde{X}_t - x^*\|^2} = \frac{1}{2n}\frac{\|M\tilde{X}_t - Mx^*\|^2}{\|\tilde{X}_t - x^*\|^2} = \frac{1}{2n}\frac{(M(\tilde{X}_t - x^*))^\top(M(\tilde{X}_t - x^*))}{(\tilde{X}_t - x^*)^\top(\tilde{X}_t - x^*)} = \\ = \frac{1}{2n}\frac{(\tilde{X}_t - x^*)^\top M^\top M(\tilde{X}_t - x^*)}{(\tilde{X}_t - x^*)^\top(\tilde{X}_t - x^*)}\end{aligned}$$

is actually the Rayleigh quotient of the matrix $M^\top M$ applied to the vector $\tilde{X}_t - x^*$ multiplied by $\frac{1}{2n}$, we can apply the following inequalities:

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

where $\sigma_{\min}(M^\top M)$ and $\sigma_{\max}(M^\top M)$ are the minimum and maximum eigenvalues of $M^\top M$ respectively, which coincide with the minimum and maximum eigenvalues of $MM^\top$.

Let now $z_t = log\|e_t\|^2$. Applying the Ito formula we have that:

$$dz_t = \frac{d\|e_t\|^2}{\|e_t\|^2} - \frac{1}{2\|e_t\|^4}16hf(\tilde{X}_t)^2dt =$$

$$= \frac{f(\tilde{X}_t)}{\|e_t\|^2}(2h\mathrm{Tr}(M^\top M) - 4)dt + \frac{f(\tilde{X}_t)}{\|e_t\|^2}4\sqrt{h}d\overline{W}_t - \left(\frac{f(\tilde{X}_t)}{\|e_t\|^2}\right)^2 8hdt =$$

$$= \left(\frac{f(\tilde{X}_t)}{\|e_t\|^2}(2h\mathrm{Tr}(M^\top M) - 4) - \left(\frac{f(\tilde{X}_t)}{\|e_t\|^2}\right)^2 8h\right)dt + \frac{f(\tilde{X}_t)}{\|e_t\|^2}4\sqrt{h}d\overline{W}_t$$

We can now write $z_t$ as follows:

$$z_t = z_0 + \int_0^t \left(\frac{f(\tilde{X}_s)}{\|e_s\|^2}(2h\mathrm{Tr}(M^\top M) - 4) - \left(\frac{f(\tilde{X}_s)}{\|e_s\|^2}\right)^2 8h\right)ds + \int_0^t \frac{f(\tilde{X}_s)}{\|e_s\|^2}4\sqrt{h}d\overline{W}_s$$

with $z_0 = log\|x^0 - x^*\|^2$.
Let's take the expectation of $z_t$, recalling that the expectation of a stochastic integral is 0:

$$\mathbb{E}[z_t] = z_0 + \mathbb{E}\left[\int_0^t \left(\frac{f(\tilde{X}_s)}{\|e_s\|^2}(2h\mathrm{Tr}(M^\top M) - 4) - \left(\frac{f(\tilde{X}_s)}{\|e_s\|^2}\right)^2 8h\right)ds\right] \leq$$

$$\leq z_0 + \left(\left(2h\mathrm{Tr}(M^\top M) - 4\right)\frac{\sigma_{max}}{2n} - 8h\frac{\sigma_{min}^2}{4n^2}\right)t \leq$$

$$\leq z_0 + \left(h\sigma_{max}^2 - \frac{2}{n}\sigma_{max} - \frac{2h}{n^2}\sigma_{min}^2\right)t$$

exploiting the results previously proved and the fact that $\mathrm{Tr}(M^\top M) \leq n\sigma_{max}$.
We want to show that $\mathbb{E}[z_t] \to -\infty$ for $t \to \infty$, so we need:

$$h\sigma_{max}^2 - \frac{2}{n}\sigma_{max} - \frac{2h}{n^2}\sigma_{min}^2 < 0$$

which is verified for:

$$h < \frac{2n\sigma_{max}}{n^2\sigma_{max}^2 - 2\sigma_{min}^2}$$

As for the variance, we can show that $\mathrm{Var}[z_t] \leq Cht$. As $\{z_t\}_t$ is an Ito process and thus a martingale, its variance is equal to the expectation of the quadratic variation. Therefore:

$$\mathrm{Var}(z_t) = \mathbb{E}[<z>_t] = \mathbb{E}\left[\int_0^t \left(\frac{f(\tilde{X}_s)}{\|e_s\|^2}4\sqrt{h}\right)^2 ds\right] \leq \frac{4h}{n^2}\sigma_{max}^2 t = Cht$$

Finally, we can show that $\tilde{X}_t \to x^*$ for $t \to \infty$, in probability:

$$\tilde{X}_t \xrightarrow{\mathbb{P}} x^* \iff \|e_t\|^2 \xrightarrow{\mathbb{P}} 0$$

since convergence in probability is stable under contnuous functions, such as the squared norm.
We can now prove $\forall \epsilon > 0$:

$$\mathbb{P}(\|e_t\|^2 \geq \epsilon) = \mathbb{P}(z_t \geq log(\epsilon)) = \mathbb{P}(z_t - \mathbb{E}[z_t] \geq log(\epsilon) - \mathbb{E}[z_t]) \leq$$

$$\leq \frac{\mathrm{Var}[z_t]}{(log(\epsilon) - \mathbb{E}[z_t])^2} \leq \frac{Cht}{(log(\epsilon) - \mathbb{E}[z_t])^2} \to 0$$

exploiting the Chebyshev inequality and the fact that $\mathbb{E}[z_t] = O(t)$.

Let's assess the performance of the GD, SGD and SDE of $X_t$ through numerical simulations.

## Q6.1

*Figure 1* shows the sequence of iterates generated by the three algorithms (Gradient Descent, Stochastic Gradient Descent and SDE). We can see that all the algorithms converge to the minimum of the function, but the Stochastic Gradient Descent and the SDE have a higher variance in the iterates, which decreases with the step $h$. The Gradient Descent algorithm converges faster than the other two, since it always moves in the direction of the gradient of the function, which is the direction of maximum decrease.

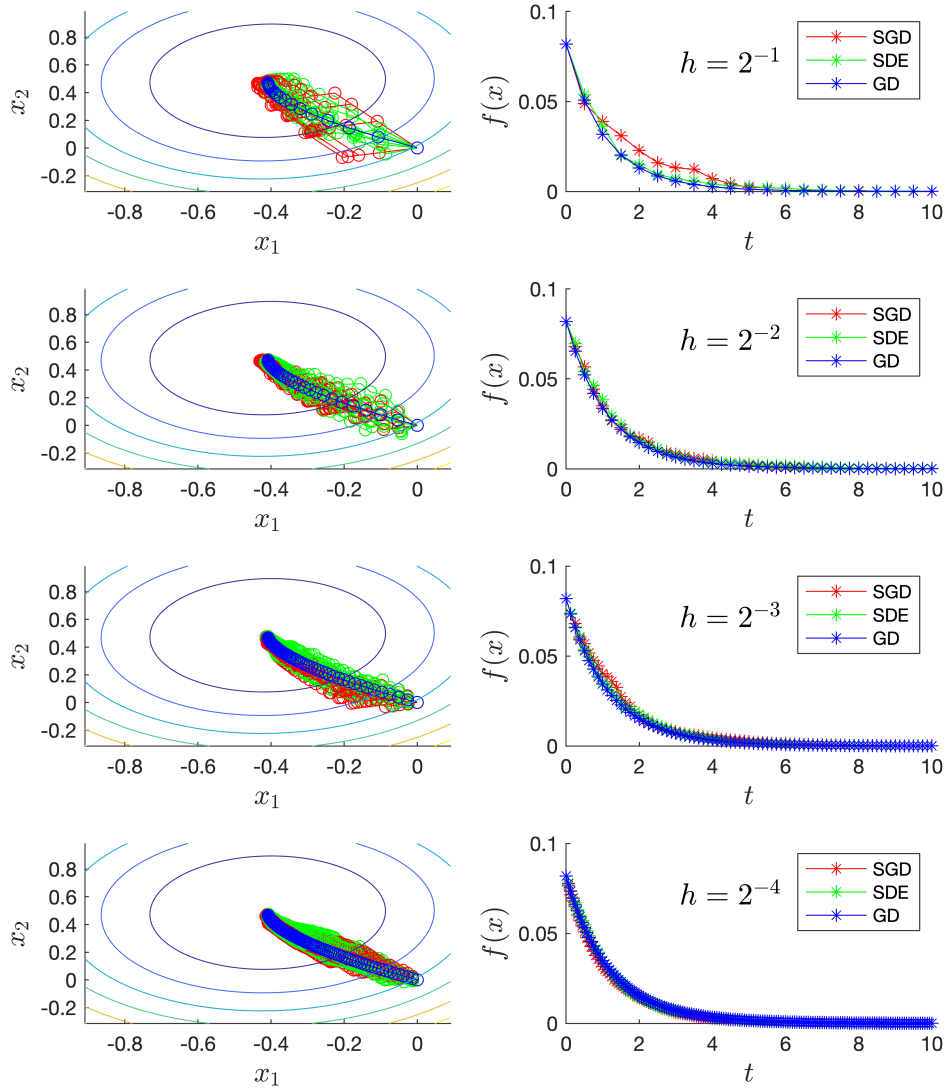


Figure 1: Iterates of GD, SGD and SDE in the realizable regime

## Q6.2

*Figure 2* shows the sequence of iterates generated by the three algorithms in the non-realizable regime. In this setting, starting from $x^0 = 0$, we have $\nabla f(x^0) = \frac{1}{n}(M^\top M x^0 - M^\top y) = -\frac{1}{n} M^\top y = 0$, since $y \notin \mathrm{Im}(M) \iff y \in \ker(M^\top)$. As a result, the Gradient Descent algorithm won't move from the initial point. The Stochastic Gradient Descent and the SDE, on the other hand, will keep moving around the minimum, due to their random nature.
In particular, from theory, we know that $\tilde{X}_t$ converges in distribution to a Gaussian with mean $x^* = \arg\min f = 0$ and variance $\sigma^2 = \frac{h}{2} f(x^*) = \frac{h}{2} f(0)$. We showed that the same behaviour holds for $X_t$, as displayed in the histograms in *Figure 2*.

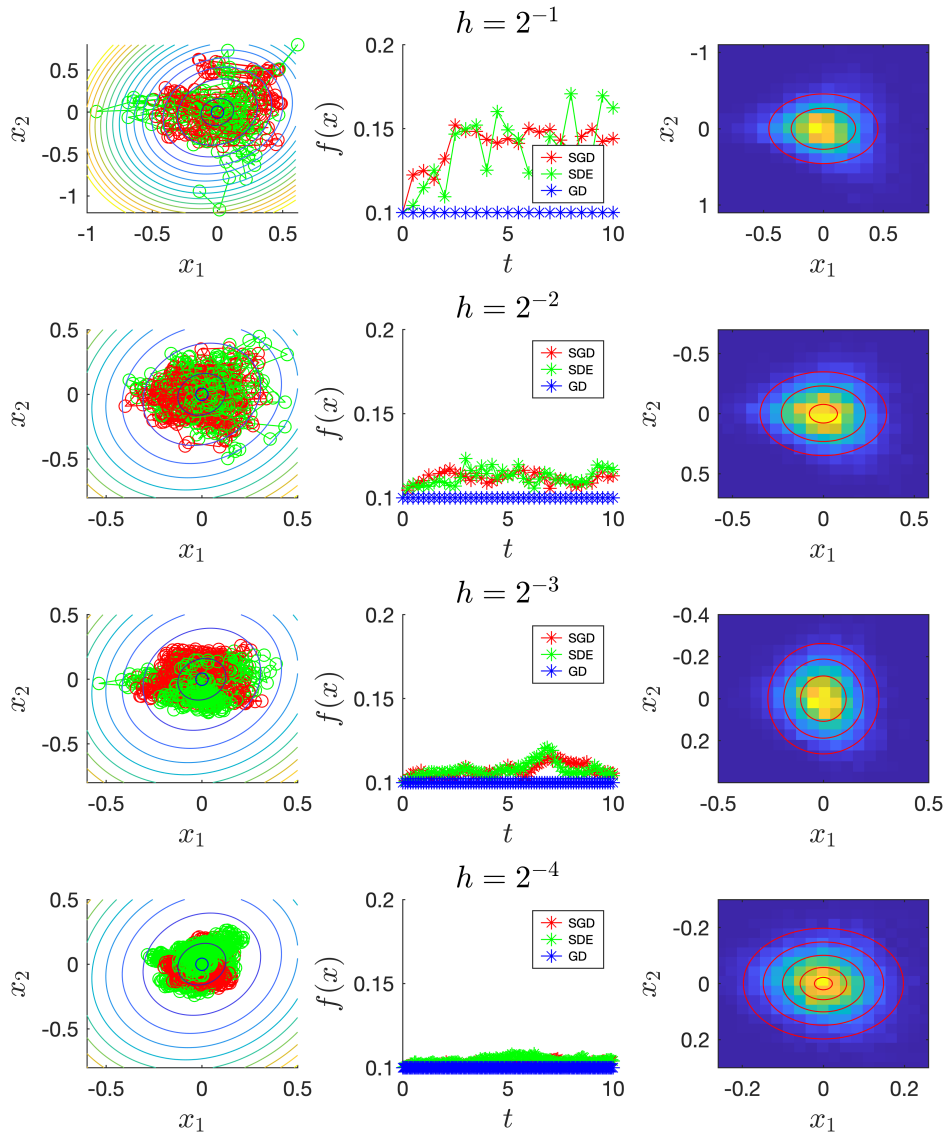


Figure 2: Iterates of GD, SGD and SDE in the non-realizable regime

We would like to find a higher-order approximation of the SGD, based on:

$$\frac{dx(t)}{dt} = -\nabla f(x(t)) + \frac{h}{2}\nabla^2 f(x(t))\nabla f(x(t))$$

## Q7.1

We now prove that the GD is a order-2 approximation of the previous ODE. We will follow the same method of Q1.2. By induction we will show that:

$$\|x^k - x(hk)\| = O(h^2) \quad \forall k$$

The base case is trivial. Now suppose that the result holds for $k$. Then, for $k+1$, exploiting Taylor expansion up to second order, we have:

$$\|x^{k+1} - x(h(k+1))\| = \|x^k - h\nabla f(x^k) - x(hk) + h\nabla f(x(hk)) + \frac{h^2}{2}\nabla^2 f(x(hk))\nabla f(x(hk)) + O(h^2)\| \leq$$

$$\leq \|x^k - x(hk)\| + h\|\nabla f(x^k) - \nabla f(x(hk))\| + \frac{h^2}{2}\|\nabla^2 f(x(hk))\nabla f(x(hk))\| + O(h^2) \leq$$
$$\leq \|x^k - x(hk)\| + hL\|x^k - x(hk)\| + \frac{h^2}{2}LB + O(h^2) = O(h^2)$$

where we used the induction hypothesis and the fact that $\nabla f$ is bounded by $B$, and Lipschitz continuous with constant $L$.

## Q7.2

With a similar reasoning, the following SDE is a weak order-2 approximation of the SGD scheme:

$$dX_t = \left[-\nabla f(X_t) - \frac{h}{2}\nabla^2 f(X_t)\nabla f(X_t)\right] dt + \sqrt{h}\Sigma(X_t)^{\frac{1}{2}} dW_t$$

We computed $\max |\mathbb{E}\phi(X_{hk} - \mathbb{E}\phi(x^k|)$ with $x^k$ the SGD iterate, and $\phi(x)$ a linear function.
As shown in *Figure 3*, the error decreases as $h^2$ as expected. The irregularity of the plot is due to the random nature of the Monte Carlo simulation.

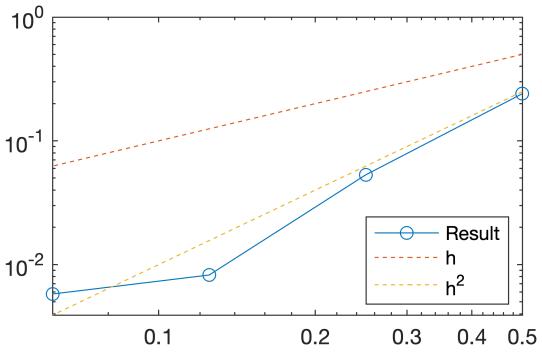


Figure 3: Error of the Euler-Maruyama scheme

In the end, we considered a variant of the SGD, the Noisy Gradient Descent (NGD), that has the folllowing update rule:

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

As before, this dynamics is connected with a SDE, that has the following form:

$$dX_t = -\nabla f(X_t)dt + \sqrt{2\tau}dW_t$$

We will now show that the Euler-Maruyama scheme applied to the SDE is equivalent to the NGD scheme.

## Q8.1

Applying the scheme to the SDE on a partition $0 = t_0 < t_1 < \cdots < t_n = T$, we have that:

$$x^{k+1} = x^k - h\nabla f(x^k) + \sqrt{2\tau}\tilde{W}^k \quad \text{where} \quad \tilde{W}^k \sim \mathcal{N}(0, hI)$$

This is equivalent to the following scheme:

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

## Q8.2

We could also consider the Milstein scheme for the SDE, but in this case it coincides with the Euler-Maruyama scheme, since the additive noise term is independent of the state $X_t$.

## Q8.3

Let's consider the same particular choice of $f$ as before, and solve the SDE.
With the previous results, we can write it as:

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

To solve the SDE, we consider an integrating factor $e^{\frac{1}{n}M^\top Mt}$, which gives:

$$d(e^{\frac{1}{n}M^\top Mt}X_t) = e^{\frac{1}{n}M^\top Mt}dX_t + \frac{1}{n}M^\top M e^{\frac{1}{n}M^\top Mt}X_t dt = e^{\frac{1}{n}M^\top Mt}dX_t + \frac{1}{n}e^{\frac{1}{n}M^\top Mt}M^\top M X_t dt$$

exploiting the commutativity of $e^{\frac{1}{n}M^\top Mt}$ and $M^\top M$ since they are both diagonalizable and function of the same matrix.

$$d(e^{\frac{1}{n}M^\top Mt}X_t) = e^{\frac{1}{n}M^\top Mt}\left(-\frac{1}{n}M^\top M X_t + \frac{1}{n}M^\top y\right)dt + \sqrt{2\tau}e^{\frac{1}{n}M^\top Mt}dW_t + \frac{1}{n}e^{\frac{1}{n}M^\top Mt}M^\top M X_t dt =$$
$$= \frac{1}{n}e^{\frac{1}{n}M^\top Mt}M^\top y dt + \sqrt{2\tau}e^{\frac{1}{n}M^\top Mt}dW_t$$

Integrating both sides, we have that:

$$e^{\frac{1}{n}M^\top Mt}X_t = X_0 + \frac{1}{n}\int_0^t e^{\frac{1}{n}M^\top Ms}M^\top y ds + \sqrt{2\tau}\int_0^t e^{\frac{1}{n}M^\top Ms}dW_s$$

$$e^{\frac{1}{n}M^\top Mt}X_t = X_0 + \left(e^{\frac{1}{n}M^\top Mt} - I\right)(M^\top M)^{-1}M^\top y + \sqrt{2\tau}\int_0^t e^{\frac{1}{n}M^\top Ms}dW_s$$

$$X_t = e^{-\frac{1}{n}M^\top Mt}X_0 + \left(I - e^{-\frac{1}{n}M^\top Mt}\right)(M^\top M)^{-1}M^\top y + \sqrt{2\tau}e^{-\frac{1}{n}M^\top Mt}\int_0^t e^{\frac{1}{n}M^\top Ms}dW_s \tag{1}$$

$X_t$ is a strong solution since it is continuous, it is $\mathcal{F}_t$-adapted, the drift is $\mathcal{M}^1([0,T])$, the diffusion is $\mathcal{M}^2([0,T])$ and (1) holds a.s.

Thanks to the regularity of the integrand, the stochastic integral is well defined and has a Gaussian distribution, implying that $X_t$ is Gaussian distributed as well, with:

$$\mathbb{E}[X_t] = e^{-\frac{1}{n}M^\top Mt}X_0 + \left(I - e^{-\frac{1}{n}M^\top Mt}\right)(M^\top M)^{-1}M^\top y$$

$$\mathrm{Var}(X_t) = \mathrm{Var}\left(\sqrt{2\tau}e^{-\frac{1}{n}M^\top Mt}\int_0^t e^{\frac{1}{n}M^\top Ms}dW_s\right) = 2\tau e^{-\frac{2}{n}M^\top Mt}\int_0^t e^{\frac{2}{n}M^\top Ms}ds =$$
$$= \tau n\left(I - e^{-\frac{2}{n}M^\top Mt}\right)(M^\top M)^{-1}$$

using Ito isometry.
Moreover, we have that:

$$\mathbb{E}[X_t] \to (M^\top M)^{-1}M^\top y \quad \text{for} \quad t \to +\infty$$
$$\mathrm{Var}(X_t) \to \tau n (M^\top M)^{-1} \quad \text{for} \quad t \to +\infty$$

Since $X_t$ is Gaussian distributed $\forall t$ and the mean and the covariance matrix converge to the values above, we can conclude that:

$$X_t \xrightarrow{\mathcal{L}} \mathcal{N}\left((M^\top M)^{-1}M^\top y,\ n\tau(M^\top M)^{-1}\right) \quad \text{for} \quad t \to +\infty$$

## Q8.4

We performed the same simulations with $\tau = 0.01$, considering NGD and its associated SDE. *Figure 4* shows the sequence of iterates generated by the three algorithms. Comparing these plots with the previous ones, we observe a similar pattern. Both NGD and its SDE are noisier than GD, but now the variance does not decrease with $h$. Nonetheless, they still converge to the minimum of the function. However, the behavior of NGD and its SDE is more erratic than that of SGD and its SDE, as noise is introduced at each iteration. Repeating the same simulations with lower $\tau$, we observed that the sparsity of the iterates decreases, which justifies referring to this parameter as "temperature," since it characterizes the entropy of the system. .

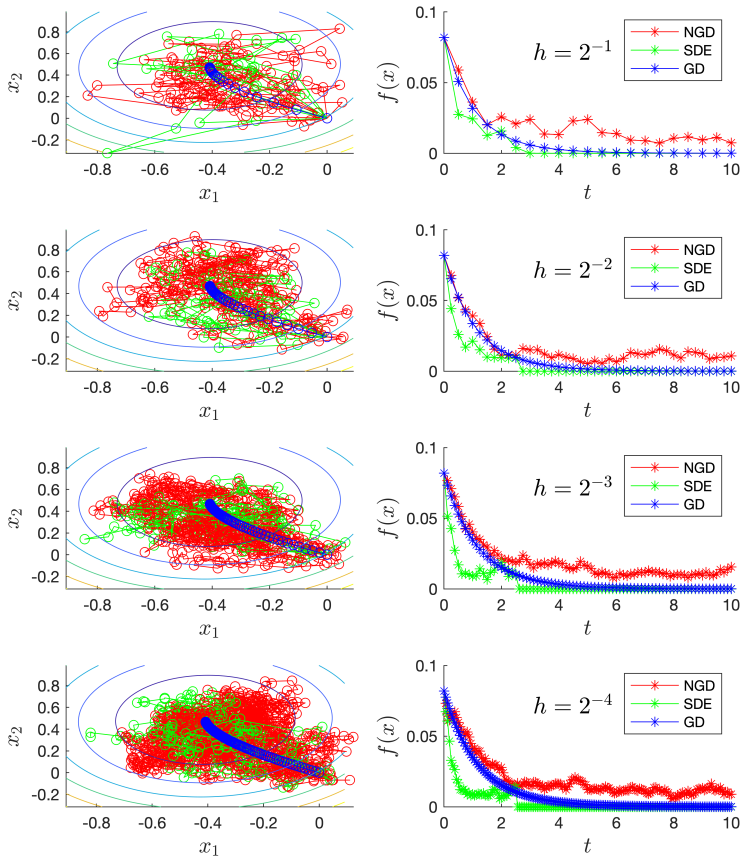


Figure 4: Iterates of GD, NGD and SDE

# Conclusion

In this project, we analyzed the Stochastic Gradient Descent algorithm, focusing on its interpretation as SDE. This connection offers a framework to understand the behavior of SGD, particularly when the full gradient is unavailable, as in large-scale optimization problems. The SDE perspective helps in analyzing the dynamics of SGD, especially in terms of its convergence properties and the role of noise. It also provides insight into how the step size $h$ influences the trade-off between exploration and convergence speed. We also studied Noisy Gradient Descent, a variant of SGD that introduces controlled noise through a "temperature" parameter. This noise allows NGD to escape local minima, improving exploration in complex, non-convex landscapes. The convergence of NGD was also analyzed through its interpretation as an SDE, which provided a rigorous understanding of its dynamics and guarantees for convergence. Both SGD and NGD demonstrate the power of stochastic methods as practical approximations of deterministic optimization algorithms, offering unique advantages in tackling high-dimensional and non-convex optimization tasks.

# References

[1] *Rethinking SGD's noise*, F. Pillaud-Vivien L. & Bach. Tech., 2022 rep. `https://francisbach.com/rethinking-sgd-noise/`.