# Mini Project: Modeling Stochastic Gradient Descent with SDEs

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 $\nabla f :\equiv \left(\frac{\partial f}{\partial x_i} \ \dots \ \frac{\partial f}{\partial x_p}\right)^\top : \mathbb{R}^p \to \mathbb{R}^p$ is the gradient of $f$.[1]

(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 step-size $h$, and $x(t)$ the solution of GF. Show that, for any fixed $T > 0$, $\sup_{k\leq \lfloor T/h \rfloor} \left\|x^k - x(hk)\right\| = O(h)$.

*Hint.* Show that

$$\left\|x^{k+1} - x(hk+h)\right\| \leq (1 + O(h)) \left\|x^k - x(hk)\right\| + \|[x(hk) - h\nabla f(x(hk))] - x(hk+h)\|$$

and that $\|[x(hk) - h\nabla f(x(hk))] - x(hk+h)\| = O(h^2)$.

Now suppose $f$ is of the form

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x).$$

Given $(f_i)_{i\leq n}$, $x^0$ and $h$, SGD constructs $(x^k)_{k\geq 0}$ by the stochastic update rule

$$x^{k+1} = x^k - h\nabla f_{i_k}(x^k) \qquad \text{where} \quad i_k \sim \text{Uniform}(\{1, ..., n\}).$$

In the next two questions, we derive a SDE that approximates SGD, in the spirit of [2] although the proposed technical pathway is a bit different.

(Q2) Write the SGD update rule as $x^{k+1} = x^k - h\nabla f(x^k) + \sqrt{h}V^k$. Show that $\mathbb{E}\left[V^k \middle| x^k\right] = 0$ and $\mathbb{E}\left[V^k V^{k\top} \middle| x^k\right] = h\Sigma(x^k)$ with

$$\Sigma(x) := \frac{1}{n}\sum_{i=1}^{n} \left[\nabla f(x) - \nabla f_i(x)\right]\left[\nabla f(x) - \nabla f_i(x)\right]^\top.$$

Can you anticipate why, intuitively, it makes sense to approximate SGD by (1)?

[1] We consider only constant step-sizes for simplicity. The notations used in this project are almost the standard ones used in the machine learning theory literature, except that the parameters to optimize are usually denoted by "$w$" or "$\theta$", with "$x$" being reserved for input data (a.k.a. covariates).

Consider the SDE

$$dX_t = -\nabla f(X_t)dt + \sqrt{h}\ \Sigma(X_t)^{1/2}dW_t \quad \text{and} \quad X_0 = x^0. \tag{1}$$

You will see, are seeing or saw in Chapter 6 of the lectures the following definitions:

$$\begin{aligned} C_{\mathrm{b}}^{\ell} &:= \left\{\phi \in C^{\ell}(\mathbb{R}^p, \mathbb{R}) \ \text{ s.t. } \quad \exists C > 0, \forall n \leq \ell, \forall x, \ \left\|\phi^{(n)}(x)\right\| \leq C\right\}, \\ C_{\mathrm{p}}^{\ell} &:= \left\{\phi \in C^{\ell}(\mathbb{R}^p, \mathbb{R}) \ \text{ s.t. } \ \exists m, C > 0, \forall n \leq \ell, \forall x, \ \left\|\phi^{(n)}(x)\right\| \leq C\left(1 + \|x\|^m\right)\right\} \end{aligned}$$

(the subscript "b" stands for "bounded" and "p" stands for "polynomial growth").

(Q3) Let $\tilde{x}^1 = x^0 - h\nabla f(x^0) + \sqrt{h}\widetilde{V}^0$ where $\widetilde{V}^0 \sim \mathcal{N}(0, h\Sigma(x^0))$. 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 $f_i \in C_{\mathrm{b}}^9$ for all $i$ and let any $\phi \in C_{\mathrm{p}}^4$. Show that there exists $C, M > 0$ such that

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

*Hint.* It is convenient to do Taylor expansions of $\phi$ around $x^0 - h\nabla f(x^0)$ with remainder in Lagrange or integral form.

(Q4) Assume that $f_i \in C_{\mathrm{b}}^9$ for all $i$. Show that for any $\phi \in C_{\mathrm{p}}^4$, there exists $C > 0$ such that

$$\forall k \leq \lfloor T/h \rfloor, \ \left|\mathbb{E}\phi(X_{hk}) - \mathbb{E}\phi(x^k)\right| \leq Ch.$$

We will say that (1) is an order-1 weak approximation of SGD (for $C_{\mathrm{b}}$ losses).
*Hint.* You may use the following multi-dimensional version of Theorem 2 from Chapter 6 of the lectures:

*Theorem.* Consider a SDE $\begin{cases} dX(t) = f(X(t))dt + g(X(t))dW(t) \\ X(0) = X_0 \end{cases}$ whose solution is denoted by $X(t)$ and a stochastic numerical method $\{x^k\}_{0\leq k\leq N}$, $h = T/N$, $N \in \mathbb{N}^*$ with the same initial value $X_0$ as the SDE.
For $r \geq 1$ assume that

(a) $f, g \in C_{\mathrm{b}}^{2r+2}$;

(b) The numerical method has weak local order $r$, i.e.,

$$\forall \varphi \in C_{\mathrm{p}}^{2r+2}, \exists C, M > 0 \text{ s.t. } \left|\mathbb{E}\left[\varphi(X(h))|X_0\right] - \mathbb{E}\left[\varphi(x^1)\middle|X_0\right]\right| \leq R$$
$$\text{where } \mathbb{E}R \leq C\left(1 + \mathbb{E}\left\|X_0\right\|^M\right)h^{r+1};$$

(c) $\mathbb{E}\left\|X_0\right\|^n < \infty$ for all $n \in \mathbb{N}^*$;

(d) We have

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

and

$$\left\|x^{k+1} - x^k\right\| \leq M_k\left(1 + \left\|x^k\right\|\right)\sqrt{h},$$

where $M_k$ is a r.v. independent of $x^k$ such that, for all $r \in \mathbb{N}$, $\mathbb{E}\left|M_k\right|^r \leq C_r$ for some constant $C_r$ independent of $k, h$.

$$\text{Then } \left|\mathbb{E}\varphi(x^k) - \mathbb{E}\varphi(X(t_n))\right| \leq Ch^r \text{ for any } t_k = k \cdot h \in [0,T] \text{ and any } \varphi \in C_{\mathrm{p}}^{2r+2}.$$

This will allow you to reduce the proof to a "local" or "one-step" weak approximation result. To show that local result, you may use that the (multi-dimensional) Euler-Maruyama method has weak local order of convergence 1.

Note that, since the randomness in SGD does not come from a Gaussian process, there is no obvious way to formulate a "strong approximation" property of SGD by a SDE. Also note that (informally) weak approximation is sufficient for the purpose of deriving convergence properties of the SGD algorithm.

**Application to quadratic minimization.** The next four questions consist in reproducing some of the experiments from [3].

(Q5) Let $M \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$, and $f_i(x) = \frac{1}{2}\left|M_{i\bullet}x - y_i\right|^2$. Check that $f(x) = \frac{1}{2n}\left\|Mx - y\right\|^2$ and

$$\Sigma(x) = \frac{1}{n}M^\top \left[\mathrm{Diag}(R)^2 - \frac{1}{n}RR^\top\right] M \quad \text{where} \quad R = Mx - y \quad \text{and} \quad \left[\mathrm{Diag}(R)^2\right]_{ij} = \delta_{ij}R_i^2.$$

(Q6) *(Overparametrized a.k.a. realizable regime.)* The SDE (1) for this particular $(f_i)_{i \leq n}$ does not have a simple closed form solution. (Try it!) Consider instead the SDE

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

Take $\widetilde{X}_0 = 0$ and assume that $y \in \mathrm{Im}(M)$ and $MM^\top$ is invertible – which is generically the case when $p \geq n$. Show that $\widetilde{X}_t$ converges to $x^* = \arg\min_{Mx=y}\left\|x\right\|^2$ in probability as $t \to \infty$ (for $h$ smaller than some constant).

*Hint.* Show that $d\left\|\widetilde{X}_t - x^*\right\|^2 = -bf(\widetilde{X}_t)dt + \sigma f(\widetilde{X}_t)dW_t$ for some constants $b, \sigma > 0$ to be specified, and that $\frac{1}{2n}\sigma_{\min}(MM^\top) \leq \frac{f(\widetilde{X}_t)}{\left\|\widetilde{X}_t - x^*\right\|^2} \leq \frac{1}{2n}\sigma_{\max}(MM^\top)$.

(Q7) Let $n = 2$, $p = 2$. Let $x^0 = 0$ and pick any $M$, $y \in \mathrm{Im}(M)$ fixed. For example, you may draw $M_{ij} \sim \mathcal{N}(0,1)$ for each $i, j$, and $y \sim \mathcal{N}(0, I_n)$.[2]

Let $T = 50$. For each $h \in \{2^{-1}, 2^{-2}, 2^{-3}, 2^{-4}\}$, letting $K = \lfloor T/h \rfloor$,

- Plot the GD iterates $(x_{\mathrm{GD}}^k)_{k \leq K}$, as well as $f(x_{\mathrm{GD}}^k)$ (on another figure).
- On the same figures, plot the SGD iterates $(x^{(1)k})_k$, ..., $(x^{(L)k})_k$ for $L = 8$ independent runs of SGD, as well as $\frac{1}{L}\sum_{\ell=1}^L f\left(x^{(\ell)k}\right)$.
- On the same figures, plot $L = 8$ sample paths $X_t^{(\ell)}$ of (1) over $[0,T]$ (for example using the Euler-Maruyama method with a small time discretization), as well as $\frac{1}{L}\sum_{\ell=1}^L f(X_t^{(\ell)})$.

Optionally, on the same figures, display some level sets of $f$.

[2] For readability of the figures, please choose (cherry-pick) an easy instance, i.e., $M$ such that $\sigma_1(M)/\sigma_2(M)$ is not too large.

(Q8) *(Underparametrized a.k.a. non-realizable regime.)* In the setting and notations of (Q6), when we instead assume $y \notin \mathrm{Im}(M)$ and $M^\top M$ is invertible – which is generically the case when $p < n$ –, one can show that $\widetilde{X}_t$ converges in distribution to $\mathcal{N}(x^*, \sigma^2 I_n)$ where $x^* = \arg\min f$ and $\sigma^2 = \frac{h}{2} f(x^*)$ [3].

Do the same experiments as in (Q7), but with $n = 5$, $p = 2$ and $y \notin \mathrm{Im}(M)$. Does the limiting behavior of $\widetilde{X}_t$ match the one of $X_t$?

**Comparison with Langevin dynamics.** Another variant of GD that involves randomness is *noisy Gradient Descent* (NGD), which given $f$, $x^0$ and $h, \tau > 0$ constructs iterates $(x^k)_k$ by the update rule

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

(Note that $f$ does not have to be of the form $\frac{1}{n}\sum_i f_i$.)

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

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

(Q10) Let $M \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$ and $f(x) = \frac{1}{2n}\|Mx - y\|^2$. Write the SDE (2) for this particular $f$. 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.[3]

(Q11) Same question as (Q7), with "(1)" replaced by "(2)" and "SGD" replaced by "NGD", and $\tau = 0.01$. The parameter $\tau$ is commonly referred to as "temperature"; can you explain why?

# References

[1] Li, Q., Tai, C., & Weinan, E. (2019). Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations. The Journal of Machine Learning Research, 20(1), 1474-1520.

[2] Li, Q., Tai, C., & Weinan, E. (2017, July). Stochastic modified equations and adaptive stochastic gradient algorithms. In International Conference on Machine Learning (pp. 2101-2110). PMLR.

[3] Pillaud-Vivien, L. & Bach, F. Posted on July 25, 2022. Rethinking SGD's noise. https://francisbach.com/rethinking-sgd-noise/.

[4] Lu, H. (2022). An $O(s^r)$-resolution ODE framework for understanding discrete-time algorithms and applications to the linear convergence of minimax problems. Mathematical Programming, 194(1-2), 1061-1112.

[3] More generally one can show that, provided that $\forall x, f(x) \geq \mu x^2 - A$ for some $\mu > 0$ and $A < \infty$, the solution $X_t$ of (2) converges in distribution and the limiting distribution has probability density function $p_\infty(x) \propto e^{-\frac{1}{\tau}f(x)}$. But that's another story...

## Rules

The rules for the submission of your project are the following:

(*1*) Your report should address all the previous points, with clear references to the correspondence between the questions (Qx) and your answers.

(*2*) Submit your solution via email to guillaume.wang@epfl.ch in an archive folder named `familyname.zip` (e.g., `wang.zip`) which should contain your report and a subfolder with your implementation. The deadline for submitting your solution is **12 June 2023 at 23:59**.

(*3*) Your report must not exceed the length of **10 pages** (minimum font size 10pt, figures and references included), and should be typeset in LaTeX. The setting and results of your numerical experiments have to be included, and all questions above have to be addressed in your report.

(*4*) Your implementation should be clear and a set of easy-to-run numerical tests should be provided. The programming language is of your choice, but a `Matlab`, `Julia` or `Python` implementation would be appreciated.

(*5*) Whenever you exploit results from existing literature, please cite your source accordingly in the bibliography.

(*6*) The project **is optional**. In case you submit a solution, your final grade $F$ for the course will be computed as

$$F = \max\{0.8W + 0.2P, W\},$$

where $W$ is the grade of the written exam and $P$ is the grade of the project.