Project for the course "Numerical Integration of Stochastic Differential Equations"

Modeling Stochastic Gradient Descent with SDEs

Teacher: *Fabio Nobile*

Academic Year 2024/2025

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 can be considered in some sense an opposite problem to numerical integration: in the latter, the goal is to derive discrete algorithms to approximate SDEs, whereas here the goal is to derive a SDE to explain a discrete algorithm. This allows to get some intuition concerning SGD, at least in the small step-size regime.

Given a lost 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 $\nabla f := \left(\frac{\partial f}{\partial x_1} \ \dots \ \frac{\partial f}{\partial x_p}\right)^\top : \mathbb{R}^p \to \mathbb{R}^p$ is the gradient of f.[1]

- **(Q1)**
  1. Show that GD can be interpreted as a forward Euler method applied to the *gradient flow* (GF) differential equation: $\frac{dx(t)}{dt} = -\nabla f(x(t)),\ t > 0$, and $x(0) = x^0$.
  2. Suppose $f \in C^2(\mathbb{R}^p)$ and $L := \sup_{\mathbb{R}^p} \left\|\nabla^2 f\right\| < \infty$ ($\nabla^2 f$ denotes the Hessian of f and $\|\cdot\|$ the Euclidean norm for vectors and Frobenius norm for matrices). Let $(x^k)_k$ denote the GD iterates using step-size h, and $x(t)$ the solution of GF. Show that, for any fixed $T > 0$, $\sup_{k\leq\lfloor T/h\rfloor} \left\|x^k - x(hk)\right\| = O(h)$, where the hidden constant depends on T, L, and $B := \sup_{\mathbb{R}^p} \|\nabla f\|$.

     Hint. Show that

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

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

- **(Q2)** Now, suppose f is of the form

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x). \tag{2}$$

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

$$x^{k+1} = x^k - h\nabla f_{i_k}(x^k) \qquad \text{where } \ i_k \sim \text{Uniform}(\{1, ..., n\}). \tag{3}$$

In the next two questions, we derive a SDE that approximates SGD, in the spirit of [1] although the proposed technical pathway is a bit different.

1. 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

$$\Sigma(x) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \left[\nabla f(x) - \nabla f_i(x)\right]\left[\nabla f(x) - \nabla f_i(x)\right]^\top. \tag{4}$$

2. Why, intuitively, does it make sense to approximate SGD by the SDE (5) below?

[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 "θ", with "x" being reserved for input data (a.k.a. covariates).

- **(Q3)** Consider the SDE

$$dX_t = -\nabla f(X_t)dt + \sqrt{h}\,\Sigma(X_t)^{1/2}dW_t \quad \text{and} \quad X_0 = x^0. \tag{5}$$

Moreover, consider the following function spaces:

$$C_{\mathrm{b}}^{\ell} \coloneqq \left\{\phi \in C^{\ell}(\mathbb{R}^p;\mathbb{R}) \ \text{ s.t. } \ \exists C>0, \forall \alpha \in \mathbb{N}^p, |\alpha| \le \ell, \ \left\|\mathrm{D}^{\alpha}\phi\right\|_{L^\infty} \le C\right\}, \tag{6}$$

$$C_{\mathrm{p}}^{\ell} \coloneqq \left\{\phi \in C^{\ell}(\mathbb{R}^p;\mathbb{R}) \ \text{ s.t. } \ \exists m, C>0, \forall \alpha \in \mathbb{N}^p, |\alpha| \le \ell, \ \sup_{x\in\mathbb{R}^p} \frac{\|\mathrm{D}^{\alpha}\phi(x)\|}{(1+\|x\|^m)} \le C,\right\} \tag{7}$$

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

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 applied to (5) with step size $\Delta t = h$.

1. What is the distribution of $\tilde{x}^1$? And of $\tilde{x}^2$?
2. Assume that $f_i \in C_{\mathrm{b}}^9$ for all i and take any $\phi \in C_{\mathrm{p}}^4$. Show that there exist $C, M > 0$ such that

$$\left|\mathbb{E}\phi(x^1) - \mathbb{E}\phi(\tilde{x}^1)\right| \le C\left(1 + \left|x^0\right|^M\right) h^p. \tag{8}$$

for $p \in \mathbb{N}$ and determine p.

Hint. It is convenient to do Taylor expansion of ϕ 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

$$\forall k \le \lfloor T/h \rfloor, \ \left|\mathbb{E}\phi(X_{hk}) - \mathbb{E}\phi(x^k)\right| \le Ch. \tag{9}$$

We will say that (5) is an order-1 weak approximation of SGD (for C_{b} losses).

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

- **(Q5)**

1. 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

$$\Sigma(x) = \frac{1}{n}M^{\top}\left[\mathrm{Diag}(R)^2(x) - \frac{1}{n}R(x)R^{\top}(x)\right]M \ \text{ where } \ R(x) = Mx - y \ \text{ and } \ \left[\mathrm{Diag}(R)^2\right]_{ij} = \delta_{ij}R_i^2. \tag{10}$$

2. *(Overparametrized a.k.a. realizable regime.)* The SDE (5) for this particular set $(f_i)_{i\le n}$ of loss functions does not have a simple closed form solution. Consider instead the SDE

$$d\widetilde{X}_t = -\nabla f(\widetilde{X}_t)dt + \sqrt{h}\,\widetilde{\Sigma}(\widetilde{X}_t)^{1/2}dW_t \ \text{ where } \ \widetilde{\Sigma}(x) = \frac{1}{n}M^{\top}\left[\frac{1}{n}\left\|Mx-y\right\|^2 I\right]M. \tag{11}$$

Take $\widetilde{X}_0 = 0$ and assume that $y \in \mathrm{Im}(M)$ and $MM^{\top}$ is invertible – which is generically the case when $p \ge n$.

(a) Show that $\tilde{W}_t = M(M^{\top}M)^{-\frac{1}{2}}W_t$ is a Brownian motion.

(b) Let $x^* = \arg\min_{Mx=y}\left\|x\right\|^2$ and $e_t = \tilde{X}_t - x^*$. Exploiting the fact that $Me_t = R(\tilde{X}_t)$, show that

$$d\|e_t\|^2 = -b(h)f(\widetilde{X}_t)dt + \sigma(h)f(\widetilde{X}_t)d\tilde{W}_t,$$

for some $b, \sigma > 0$ independent of $\tilde{X}_t$ (but otherwise dependent on h, n, M).

(c) Show that $\frac{1}{2n}\sigma_{\min}(MM^{\top}) \le \frac{f(\tilde{X}_t)}{\left\|\tilde{X}_t - x^*\right\|^2} \le \frac{1}{2n}\sigma_{\max}(MM^{\top})$ and conclude that $\tilde{X}_t$ converges in probability to $x^* = \arg\min_{Mx=y}\left\|x\right\|^2$ as $t \to \infty$ (for h smaller than some constant).

Hint. Write the SDE for $z_t = \log\|e_t\|^2$ and show that $\mathbb{E}[z_t] \to -\infty$ as $t \to +\infty$, $\mathrm{Var}[z_t] \le Cht, ...$

- **(Q6)**

1. Let $n = 2$, $p = 2$, $x^0 = 0$ and pick any M and $y \in \mathrm{Im}(M)$. 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 $T = 10$. For each $h \in \{2^{-1}, 2^{-2}, 2^{-3}, 2^{-4}\}$, letting $K = \lfloor T/h \rfloor$,
   - Plot the GD iterates $(x^k_{\mathrm{GD}})_{k \le K}$, as well as $f(x^k_{\mathrm{GD}})$ (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 (5) over $[0, T]$ (for example using the Euler-Maruyama method with a small time discretization), as well as $\frac{1}{L}\sum_{\ell=1}^{L} f(X_t^{(\ell)})$.

   Moreover, on the same figures, display some level sets of f.
2. *(Underparametrized a.k.a. non-realizable regime.)* In the setting and notations of **(Q5)-2** (Over-parametrized a.k.a. realizable regime.), 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 of **(Q6)-1**, 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?

A higher-order approximation. As shown in [2], by backward error analysis, one may derive ODEs approximating GD with arbitrarily high order. As for SDEs modeling SGD, we are only aware of works showing how to get one order higher [1]. Here we go just one order higher than GF resp. (5).

- **(Q7)**

1. Suppose $f \in C^3(\mathbb{R}^p)$ and $\sup_{\mathbb{R}^p} \left\|\nabla^2 f\right\| < \infty$, let $(x^k)_k$ denote the GD iterates using step-size h, and let $x(t)$ be the solution of the *higher-resolution ODE of GD*

$$\frac{dx(t)}{dt} = -\nabla f(x(t)) - \frac{h}{2}\nabla^2 f(x(t))\nabla f(x(t)) \quad \text{and} \quad x(0) = x^0. \tag{12}$$

   Show that, for any fixed $T > 0$, $\sup_{k \le \lfloor T/h \rfloor} \left\|x^k - x(hk)\right\| = O(h^2)$, where hidden constants can have similar dependency as in **(Q1)** .
2. One can show that

$$dX_t = \left[-\nabla f(X_t) - \frac{h}{2}\nabla^2 f(X_t)\nabla f(X_t)\right] dt + \sqrt{h}\ \Sigma(X_t)^{1/2} dW_t \quad \text{and} \quad X_0 = x^0 \tag{13}$$

   is an order-2 weak approximation of SGD; i.e., for any $\phi \in C^6_{\mathrm{p}}$, there exists $C > 0$ such that

$$\forall k \le \lfloor T/h \rfloor,\ \left|\mathbb{E}\phi(X_{hk}) - \mathbb{E}\phi(x^k)\right| \le Ch^2. \tag{14}$$

   Verify this fact numerically for some linear ϕ, by plotting $\max_{k \le \lfloor T/h \rfloor} \left|\mathbb{E}\phi(X_{hk}) - \mathbb{E}\phi(x^k)\right|$ as a function of h in log-log scale, where X_t is the solution of (13).

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

$$x^{k+1} = x^k - h\nabla f(x^k) + \sqrt{2h\tau}\ W^k \quad \text{where} \quad W^k \sim \mathcal{N}(0, I_p). \tag{15}$$

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

- **(Q8)**

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

$$dX_t = -\nabla f(X_t)dt + \sqrt{2\tau}\ dW_t. \tag{16}$$

2. How does the Milstein scheme applied to (16) reads? Argument.

[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.

3. 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 (16) 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]
4. Same question as in **(Q6)-1**, with "(5)" replaced by "(16)" and "SGD" replaced by "NGD", and $\tau = 0.01$. The parameter τ is commonly referred to as "temperature"; can you explain why?

References

[1] Qianxiao Li, Cheng Tai, and E Weinan. "Stochastic modified equations and adaptive stochastic gradient algorithms". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 2101–2110.

[2] Haihao Lu. "An o (sr)-resolution ode framework for understanding discrete-time algorithms and applications to the linear convergence of minimax problems". In: *Mathematical Programming* 194.1 (2022), pp. 1061–1112.

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

[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 (16) 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...