Lectures 18–21 Langevin and diffusion models

Sampling by noisy gradient descent · the forward process and Tweedie's formula · reversal and score matching · guarantees

---

Reading guide. Section 1 is Lecture 18 (Nov 16), Section 2 is Lecture 19 (Nov 18), Section 3 is Lecture 20 (Nov 30), Section 4 is Lecture 21 (Dec 2). Everything is in discrete time with Gaussian steps; the continuous-time objects (the Langevin diffusion, the Fokker–Planck equation) appear only as limits of what is proved, and are marked as such. The tools are Gaussian integration by parts (Lecture 5), Gaussian conditioning (Lecture 2), the contraction argument of Lecture 16, and relative entropy (Lectures 1 and 3). Examples marked WORKED IN CLASS are done on the board. Exercises are typed A, B, C as before; the homework for a lecture is the exercise block closing its section.

---

1 Langevin dynamics

Lecture 16 found that constant-step SGD forgets its initial condition and then fluctuates forever at a noise floor proportional to the step. This lecture turns the floor into the purpose. Take $U : \mathbb{R}^d \to \mathbb{R}$ with $\int e^{-U} < \infty$, the probability density

$$\pi(x) = \frac{e^{-U(x)}}{Z}, \qquad Z = \int_{\mathbb{R}^d} e^{-U(x)}\, dx,$$

and run gradient descent on U with Gaussian noise injected at every step, calibrated to the step size:

$$\boxed{x_{k+1} = x_k - \eta\, \nabla U(x_k) + \sqrt{2\eta}\, g_{k+1}, \qquad g_1, g_2, \dots \text{ i.i.d. } \mathcal{N}(0, I_d).}$$

This is the *unadjusted Langevin algorithm*. The two theorems of the lecture say that it samples from π:

- π *is the equilibrium* (Theorem 1.2): if $x_k \sim \pi$ then $x_{k+1} \sim \pi$ up to an error $o(\eta)$ per step. The proof is integration by parts against e^{-U}, which for $U = \|x\|^2/2$ is Stein's identity of Lecture 5.
- *The chain forgets where it started* (Theorem 1.4): if U is strongly convex, two copies of the algorithm driven by the same noise approach each other at the rate $(1 - \mu\eta)^k$, the Lecture 16 rate. So after $O(1/\mu\eta)$ steps the law of x_k no longer depends on x_0, and the equilibrium is π up to a bias of order $\sqrt{\eta d}$.

Sampling from a density known up to its normalizing constant is the computational problem of Bayesian statistics (the posterior is prior $\times$ likelihood, and Z is the intractable marginal likelihood), of statistical physics (π is a Gibbs measure, Z the partition function), and of generative modelling (one wants fresh samples from the distribution of the data). The Langevin algorithm never looks at Z: it uses only $\nabla U = -\nabla \log \pi$, the *score* of π. The rest of the block is about the case where even U is unknown and only samples from π are available. Lecture 19 shows the score can then be learned, and Lectures 20–21 show that Langevin-type steps along a path of Gaussian-smoothed versions of π produce new samples. That is a diffusion model.

1.1 Setting

(U1) U is twice continuously differentiable and $\mu I \preceq \nabla^2 U(x) \preceq \beta I$ for all x, with $0 < \mu \le \beta$.

As in Lecture 16, (U1) gives $\langle \nabla U(x) - \nabla U(y), x - y\rangle \ge \mu\|x-y\|^2$, $\|\nabla U(x) - \nabla U(y)\| \le \beta\|x-y\|$, a unique minimizer x^*, and $U(x) \ge U(x^*) + \frac{\mu}{2}\|x - x^*\|^2$, so e^{-U} is integrable and π has all moments. The algorithm is a Markov chain: given $x_k = x$, $x_{k+1} \sim \mathcal{N}(x - \eta\nabla U(x), 2\eta I)$, and $x_k \in L^2(\mathcal{F}_k)$ for $\mathcal{F}_k = (g_1, \dots, g_k)$.

1.2 The equilibrium

Lemma 1.1 (integration by parts against e^{-U})**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be continuously differentiable and of polynomial growth, together with its derivative. Then*

$$\mathbb{E}_\pi\big[\nabla\cdot f(X)\big] = \mathbb{E}_\pi\big[\langle f(X), \nabla U(X)\rangle\big].$$

Proof. $\nabla\cdot(fe^{-U}) = (\nabla\cdot f)e^{-U} - \langle f, \nabla U\rangle e^{-U}$; integrate coordinate by coordinate. The boundary terms vanish because U grows quadratically by (U1). □

For $U = \|x\|^2/2$ this is $\mathbb{E}[\nabla\cdot f(Z)] = \mathbb{E}\langle f(Z), Z\rangle$ for $Z \sim \mathcal{N}(0, I_d)$: Stein's identity. With $f = \nabla\varphi$ it reads $\mathbb{E}_\pi\Delta\varphi = \mathbb{E}_\pi\langle\nabla\varphi, \nabla U\rangle$, which is the form used next.

Theorem 1.2 (one step of Langevin)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be three times continuously differentiable with $\|\nabla^2 f\|_{\mathrm{op}} \le c_2$ and $\|D^3 f\| \le c_3$, and define the* generator

$$\mathcal{L}f(x) := -\langle\nabla U(x), \nabla f(x)\rangle + \Delta f(x).$$

Then $\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] = f(x_k) + \eta\,\mathcal{L}f(x_k) + R_k$ with $|R_k| \le \frac{1}{2}c_2\eta^2\|\nabla U(x_k)\|^2 + \frac{2}{3}c_3\big(\eta^3\|\nabla U(x_k)\|^3 + (2\eta)^{3/2}(d+1)^{3/2}\big)$. In particular, if $x_k \sim \pi$ then

$$\mathbb{E}f(x_{k+1}) = \mathbb{E}f(x_k) + O(\eta^{3/2}) :$$

π is stationary to first order in η.

Proof. Write $x = x_k$ and $\delta = -\eta\nabla U(x) + \sqrt{2\eta}\,g$ with $g = g_{k+1}$ independent of $\mathcal{F}_k$. By Taylor's formula, $f(x+\delta) = f(x) + \langle\nabla f(x), \delta\rangle + \frac{1}{2}\delta^\top\nabla^2 f(x)\delta + \rho$ with $|\rho| \le \frac{1}{6}c_3\|\delta\|^3$. Condition on $\mathcal{F}_k$ (freezing lemma). $\mathbb{E}\delta = -\eta\nabla U$ gives $-\eta\langle\nabla U, \nabla f\rangle$. Since $\mathbb{E}\delta\delta^\top = \eta^2\nabla U\nabla U^\top + 2\eta I$, the quadratic term has conditional mean $\frac{1}{2}\operatorname{tr}(\nabla^2 f\,\mathbb{E}\delta\delta^\top) = \eta\Delta f + \frac{1}{2}\eta^2\nabla U^\top\nabla^2 f\nabla U$. Finally $\|\delta\|^3 \le 4(\eta^3\|\nabla U\|^3 + (2\eta)^{3/2}\|g\|^3)$ and $\mathbb{E}\|g\|^3 \le (\mathbb{E}\|g\|^4)^{3/4} \le (d+1)^{3/2}$. For the last claim take expectations with $x_k \sim \pi$: $\mathbb{E}_\pi\mathcal{L}f = 0$ by Lemma 1.1 with $f \mapsto \nabla f$, and $\mathbb{E}|R_k| = O(\eta^{3/2})$ since π has all moments. □

The identity $\mathbb{E}_\pi\mathcal{L}f = 0$ is the whole reason e^{-U} is the target: the drift $-\nabla U$ pushes mass towards the minimizer, the noise spreads it, and π is the density at which the two cancel. In the small-step limit the cancellation is exact.

QUOTED · the Langevin diffusion and the Fokker–Planck equation

As $\eta \to 0$ with $t = k\eta$ fixed, the chain converges to the *Langevin diffusion* $dX_t = -\nabla U(X_t)\,dt + \sqrt{2}\,dB_t$, for which Theorem 1.2 becomes $\frac{d}{dt}\mathbb{E}f(X_t) = \mathbb{E}\mathcal{L}f(X_t)$; integrating by parts, the density

p_t of X_t solves the *Fokker–Planck equation* $\partial_t p = \nabla\cdot(p\nabla U + \nabla p)$, and $\pi \propto e^{-U}$ is stationary because the flux $p\nabla U + \nabla p$ vanishes for $p = \pi$. For $U \equiv 0$ this is the heat equation $\partial_t p = \Delta p$, solved from a point mass by $\mathcal{N}(0, 2tI)$: the heat kernel, which returns in Lecture 21. The construction of X_t needs stochastic calculus; nothing below depends on it.

Example 1.3 (the Gaussian target)**.** WORKED IN CLASS Let $U(x) = \frac{1}{2}(x-m)^\top H(x-m)$, $\mu I \preceq H \preceq \beta I$, so $\pi = \mathcal{N}(m, H^{-1})$. The algorithm is linear, $x_{k+1} - m = (I - \eta H)(x_k - m) + \sqrt{2\eta}\, g_{k+1}$, so x_k is Gaussian for every k (Lecture 1), with mean $m + (I-\eta H)^k(x_0 - m)$ and covariance $C_{k+1} = (I-\eta H)C_k(I-\eta H) + 2\eta I$. In the eigenbasis of H the coordinates decouple: $c_{k+1,i} = (1-\eta h_i)^2 c_{k,i} + 2\eta$, which for $\eta < 2/\beta$ converges geometrically to

$$c_{\eta,i} = \frac{2\eta}{1-(1-\eta h_i)^2} = \frac{1}{h_i}\cdot\frac{1}{1-\eta h_i/2}.$$

So the law of x_k converges to $\mathcal{N}(m, C_\eta)$: the right mean, and each variance too large by the factor $(1-\eta h_i/2)^{-1} = 1 + \frac{\eta h_i}{2} + \dots$, the *discretization bias*, of order η. This is Remark 1.3 of Lecture 16 read in the other direction: there the floor $\eta\sigma^2/(h(2-\eta h))$ was an error; here, with $\sigma^2 = 2/\eta$, it is $1/(h(1-\eta h/2))$, the target variance up to the bias. One step from equilibrium in $d = 1$: if $x_0 \sim \pi$ then $x_1 \sim \mathcal{N}(m, \frac{1}{h} + \eta^2 h)$; the variance moves by $\eta^2 h$, so for a Gaussian the stationarity of Theorem 1.2 is in fact second-order. (Constant-step SGD on a loss L with noise of covariance $\sigma^2 I$ has this form with $U = 2L/(\eta\sigma^2)$: it samples, approximately, from a Gibbs measure for the loss at temperature $\eta\sigma^2/2$; Exercise 1.8.)

1.3 Forgetting the initial condition

Theorem 1.4 (contraction)**.** *Assume (U1) and $\eta \le 1/\beta$. Let (x_k) and (y_k) be two runs of the Langevin algorithm* driven by the same noise $g_1, g_2, \dots$ *from initial points x_0, y_0. Then for every k,*

$$\|x_k - y_k\| \le (1-\mu\eta)^k \|x_0 - y_0\|.$$

Consequently, for random initial points with finite second moments and any ℓ-Lipschitz $f : \mathbb{R}^d \to \mathbb{R}$,

$$\big|\mathbb{E}f(x_k) - \mathbb{E}f(y_k)\big| \le \ell\,(1-\mu\eta)^k \big(\mathbb{E}\|x_0 - y_0\|^2\big)^{1/2}.$$

Proof. Subtracting the two recursions, the noise cancels: $x_{k+1} - y_{k+1} = T(x_k) - T(y_k)$ with $T(x) = x - \eta\nabla U(x)$, the gradient descent map. Along the segment $x_t = y + t(x-y)$, $T(x) - T(y) = \int_0^1 (I - \eta\nabla^2 U(x_t))(x-y)\,dt$, and by (U1) and $\eta \le 1/\beta$ the symmetric matrix $I - \eta\nabla^2 U(x_t)$ has eigenvalues in $[1-\eta\beta, 1-\eta\mu] \subset [0, 1-\mu\eta]$. So $\|T(x) - T(y)\| \le (1-\mu\eta)\|x-y\|$; iterate. For the second claim, $|\mathbb{E}f(x_k) - \mathbb{E}f(y_k)| \le \ell\,\mathbb{E}\|x_k - y_k\| \le \ell(1-\mu\eta)^k(\mathbb{E}\|x_0 - y_0\|^2)^{1/2}$ by Cauchy–Schwarz. □

This is Lemma 1.1 of Lecture 16 with the noise removed by the coupling instead of averaged out: for two chains driven by the same Gaussians, the difference evolves by *deterministic* gradient descent, and strong convexity contracts it. The noise is along for the ride; it is what makes the common limit interesting. In $k \approx \frac{1}{\mu\eta}\log\frac{1}{\epsilon}$ steps the law of x_k depends on x_0 only at the level ϵ, for every Lipschitz test function. Together with Theorem 1.2, which says that a chain started at π stays at π up to $O(\eta^{3/2})$ per step, this is the whole story: the errors $O(\eta^{3/2})$ accumulate over the $1/(\mu\eta)$ steps the chain remembers, for a total bias of order $\eta^{3/2}/(\mu\eta) = \sqrt{\eta}/\mu$. Making this precise requires one more ingredient, the size of the one-step error in the coupling sense rather than for test functions, and that is where the continuous-time process is needed; the result is the following.

QUOTED · Dalalyan 2017; Durmus–Moulines 2017; Dalalyan–Karagulyan 2019

Under (U1) and $\eta \le 2/(\mu+\beta)$, for every k there is a coupling of x_k with a random vector $X \sim \pi$ such that

$$\big(\mathbb{E}\|x_k - X\|^2\big)^{1/2} \;\le\; (1-\mu\eta)^k \big(\mathbb{E}\|x_0 - X_0\|^2\big)^{1/2} + 1.65\,\frac{\beta}{\mu}\,\sqrt{\eta d}\,,$$

where (x_0, X_0) is any coupling of the initial point with π. The second term is the bias; it comes from the chain using the gradient at the start of each step, while the diffusion it approximates moves by $\sqrt{2\eta}\,g$ during the step and sees the gradient change by about $\beta\sqrt{2\eta d}$, a displacement of order $\beta\eta^{3/2}\sqrt{d}$ per step, divided by the contraction $\mu\eta$.

Two readings. To sample to accuracy ϵ one takes $\eta \approx \epsilon^2\mu^2/(3\beta^2 d)$ and then $k = O\big(\frac{\beta^2 d}{\mu^3\epsilon^2}\log\frac{1}{\epsilon}\big)$ steps: the dependence d/ϵ^2 is the sample complexity of SGD in Lecture 16, for the same reason, since each step injects noise of size $\sqrt{\eta d}$ and only many small steps average it out. And for the Gaussian target the bias is $O(\eta)$, not $O(\sqrt{\eta})$ (Example 1.3): the change of the gradient over a step averages out to second order for a quadratic. The $\sqrt{\eta}$ rate is sharp for general smooth potentials; a Metropolis accept/reject step on each proposal removes the bias altogether, at the cost of rejected moves.

1.4 Without strong convexity

Theorem 1.4 is the only place $\mu > 0$ was used, and it is a strong requirement: a strongly convex U has one well. For the mixture $\pi = \frac{1}{2}\mathcal{N}(-a,1) + \frac{1}{2}\mathcal{N}(a,1)$ in $d = 1$ with a large, $U = -\log\pi$ has two wells separated by a barrier of height $\Delta U \approx a^2/2$, and two chains started in different wells do not approach each other: each stays in its well for a time of order $e^{\Delta U}/\eta$ before the noise carries it across (Exercise 1.9). Theorem 1.2 holds verbatim, π is still the equilibrium; what fails is the rate, which is exponential in the barrier. For any feasible number of steps a chain started in one well samples the right local shape with the wrong weights.

This is where the Gaussians come back. Convolving π with $\mathcal{N}(0,\sigma^2)$ fills in the barrier: for $\sigma \gg a$ the smoothed density $\pi_\sigma = \pi * \mathcal{N}(0,\sigma^2)$ is nearly Gaussian and strongly log-concave, and Langevin samples it quickly. Lowering σ gradually, with a few Langevin steps at each level started from the samples of the previous one, carries samples from an easy distribution to the hard one. Two things make this a method: the score $\nabla \log \pi_\sigma$ at every noise level, which Lecture 19 shows is the same thing as a *denoiser* and can be learned from samples of π alone; and an accounting of the error made at each level, which is Lectures 20–21. The path from π to a Gaussian, $\sigma \uparrow \infty$, is the *forward process* of a diffusion model, and it is itself the Langevin algorithm for $U = \|x\|^2/2$: Example 1.3 run from the data.

Exercises for Lecture 18

Exercise 1.5 (the Gaussian chain). TYPE A In Example 1.3 with $d = 1$, $m = 0$: (a) verify $c_\eta = 1/(h(1-\eta h/2))$ and the one-step variance $\frac{1}{h} + \eta^2 h$ from equilibrium; (b) couple the chain with its own stationary version by the same noise and show $\mathbb{E}(x_k - \tilde{x}_k)^2 = (1-\eta h)^{2k}\mathbb{E}(x_0 - \tilde{x}_0)^2$: Theorem 1.4 is exact here; (c) show the bias $\sqrt{c_\eta} - h^{-1/2} = \frac{\eta\sqrt{h}}{4} + O(\eta^2)$ and compare with the quoted $1.65(\beta/\mu)\sqrt{\eta d}$. Why is the Gaussian case better by a factor $\sqrt{\eta}$?

Exercise 1.6 (Stein's identity as stationarity). TYPE A Let $U = \|x\|^2/2$. (a) Compute $\mathcal{L}f$ for $f(x) = \|x\|^2$ and for $f(x) = e^{\langle \lambda, x\rangle}$, and read off $\mathbb{E}_\pi\|X\|^2$ and $\mathbb{E}_\pi e^{\langle \lambda, X\rangle}$ from $\mathbb{E}_\pi \mathcal{L}f = 0$. (b) For $f(x) = \|x\|^2$ show that Theorem 1.2 holds with $R_k = \eta^2\|x_k\|^2$ exactly, and deduce the recursion for $\mathbb{E}\|x_k\|^2$ and its limit.

Exercise 1.7 (the best step for contraction). TYPE B Under (U1), show that $T(x) = x - \eta\nabla U(x)$ satisfies $\|T(x) - T(y)\| \le \max(|1-\eta\mu|, |1-\eta\beta|)\|x-y\|$ for every $\eta > 0$, that the constant is attained for a quadratic U, and that it is minimized at $\eta = 2/(\mu+\beta)$ with value $(\kappa-1)/(\kappa+1)$, $\kappa = \beta/\mu$. Compare the number of steps to contract by e^{-1}, about $\kappa/2$, with the κ steps of Theorem 1.4 at $\eta = 1/\beta$, and with the κ^2 steps given by Lemma 1.1 of Lecture 16 at its best step $\eta = \mu/\beta^2$. Where did Lecture 16 lose a factor κ?

Exercise 1.8 (SGD at a temperature). TYPE B Constant-step SGD with additive Gaussian noise, $\theta_{k+1} = \theta_k - \eta\nabla L(\theta_k) - \eta\xi_{k+1}$, $\xi \sim \mathcal{N}(0, \sigma^2 I)$. (a) Match it to the Langevin algorithm for a potential U with step η': $\eta' = \eta^2\sigma^2/2$, $U = 2L/(\eta\sigma^2)$. With mini-batches of size B the temperature $\eta\sigma^2/2$ becomes $\eta\sigma^2/(2B)$: the "linear scaling rule" $\eta \propto B$. (b) For $L = \frac{1}{2}\theta^\top H\theta$ compute the exact stationary covariance of SGD and compare it with the Gibbs covariance $\frac{\eta\sigma^2}{2}H^{-1}$. (c) By Laplace's method, the Gibbs measure gives a neighbourhood of a local minimizer θ_j^* with Hessian eigenvalues $h_i^{(j)}$ the weight $\propto e^{-2L(\theta_j^*)/(\eta\sigma^2)}\prod_i (h_i^{(j)})^{-1/2}$. Show that among minimizers of equal loss the flattest is preferred, independently of the temperature, and say how large a loss difference δ the temperature can trade against a factor 2 in $\prod_i h_i^{(j)}$.

Exercise 1.9 (the double well, and smoothing it). TYPE C Let $\pi = \frac{1}{2}\mathcal{N}(-a,1) + \frac{1}{2}\mathcal{N}(a,1)$ in $d = 1$ and $U = -\log\pi$. (a) Compute ∇U in closed form and show $U(0) - U(\pm a) = \frac{a^2}{2} - \log 2 + O(e^{-2a^2})$. (b) Run the Langevin algorithm with $\eta = 0.1$ from $x_0 = -a$ for $a = 2, 3, 4$ and 10^5 steps; record the fraction of time at $x > 0$ (target $\frac{1}{2}$) and the mean number of steps before the first visit to $x > 0$ over 100 runs; compare with $e^{U(0)-U(-a)}/\eta$. (c) For $\sigma > 0$ let $\pi_\sigma = \pi * \mathcal{N}(0,\sigma^2) = \frac{1}{2}\mathcal{N}(-a, 1+\sigma^2) + \frac{1}{2}\mathcal{N}(a, 1+\sigma^2)$; show $-\log\pi_\sigma$ is strongly convex when $1 + \sigma^2 > a^2$ and compute its score. (d) *Annealed Langevin.* For $a = 4$ run 200 Langevin steps at each level $\sigma = 8, 4, 2, 1, 0.5, 0.25, 0$ in turn with $\eta = \sigma^2/10$ ($\eta = 0.1$ at the last level), each level started from the end of the previous one, for 200 chains from $x_0 = -a$; report the fraction ending at $x > 0$. This is a diffusion model with the exact score.