Introduction to Generative AI

Clément Lalanne

Notation for this lecture

We denote measures by capital blackboard letters and densities in lowercase, so $\mathbb{P}$ has density p, $\mathbb{P}_\theta$ has p_θ, and $\mathbb{Q}$ has q. The data live in $\mathcal{X} \subseteq \mathbb{R}^d$, so d is the dimension of one data point, and there are no labels.

Sample. $S = (x_1, \dots, x_n) \sim \mathbb{P}^{\otimes n}$, and $\widehat{\mathbb{P}}_n = \frac{1}{n} \sum_{i \le n} \delta_{x_i}$ is the empirical measure. Anything computed from the sample carries it, as in $\hat{\theta}_S$ and $\widehat{\mathbb{P}}_n$.

Coordinates. $x = (x^{(1)}, \dots, x^{(d)})$, and $x^{(<j)} = (x^{(1)}, \dots, x^{(j-1)})$. Superscripts in parentheses are always coordinates, while subscripts are always sample indices, so that $x_i^{(j)}$ is coordinate j of the i-th point.

Latent variables. $z \in \mathcal{Z} = \mathbb{R}^k$ with prior $\mathbb{P}_z$, usually $\mathcal{N}(0, I_k)$, and k is the latent dimension. The variational density is q, amortised as $q_\phi(\cdot \mid x)$ with measure $\mathbb{Q}_{\phi,x}$. Note that θ always parametrises the model, while ϕ always parametrises the variational family.

Markov chains. The chain is $(X_t)_{t \ge 0}$, capital, indexed by time t. The sample is $x_1, \dots, x_n$, lowercase, indexed by i. They are different objects and the letters are deliberately different. The step size is γ, as in the optimization lecture, and $\epsilon \sim \mathcal{N}(0, I_d)$ is always a standard Gaussian noise.

Part I. The meaning of generation

The task

We observe $S = (x_1, \dots, x_n) \sim \mathbb{P}^{\otimes n}$ with $\mathbb{P}$ unknown, and we want an algorithm that outputs fresh points.

> **Definition.** A **generative model** is a family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ of distributions on $\mathcal{X}$ together with a procedure that, given θ, returns a random x with law $\mathbb{P}_\theta$, or an approximation of it.

The goal, informally. We choose $\hat{\theta}_S$ so that $\mathbb{P}_{\hat{\theta}_S}$ is close to $\mathbb{P}$, and we then draw from $\mathbb{P}_{\hat{\theta}_S}$.

Difference with prediction. There is no y, no loss $\ell(y, \hat{y})$, no risk to minimize, and no Bayes predictor to aim at. We want the law itself, which the introductory lecture described as usually more than we need, whereas here it is what we need.

Two abilities that do not come together

We can ask two different things of a model, and most families provide one of them cheaply and the other not at all.

Evaluation. Given x, the model returns $p_\theta(x)$, or $\log p_\theta(x)$, which is what a likelihood needs.

Sampling. The model returns a random x with law $\mathbb{P}_\theta$, which is what the user wants.

Two distinct abilities. A Gaussian mixture with a thousand components in dimension 10^4 is trivial to evaluate and to sample. In contrast, an Ising model on a 100×100 grid is trivial to evaluate up to a constant and, as Part II explains, extremely hard to sample. Conversely, a generative adversarial network is trivial to sample, but its density cannot be evaluated at all, since it does not have one.

Degeneracy of the empirical measure

The distribution closest to the data is the data itself. However, sampling from $\widehat{\mathbb{P}}_n$ means picking one of $x_1, \dots, x_n$ uniformly at random, which is trivial and useless.

Worse, if the model class is rich enough, maximum likelihood converges to the empirical measure. Indeed, let us consider, for $\sigma > 0$, the smoothed empirical measure

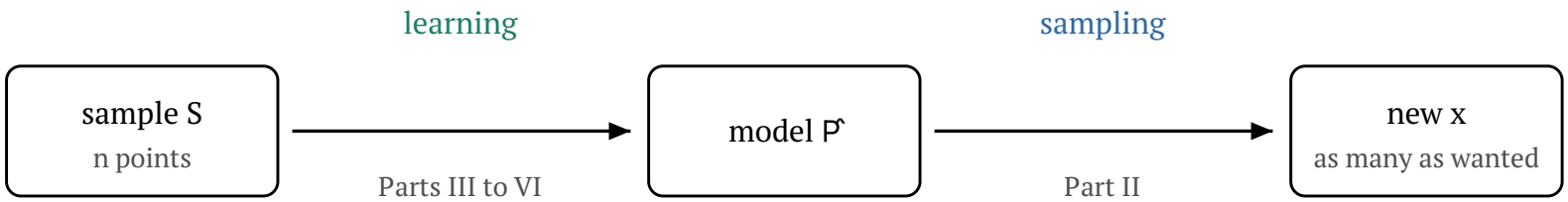
$$p_\sigma(x) = \frac{1}{n} \sum_{i \leq n} \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(- \frac{\|x - x_i\|_2^2}{2\sigma^2} \right) .$$

Keeping only the term $i = j$ in each summand, we get

$$\frac{1}{n} \sum_{j \leq n} \log p_\sigma(x_j) \geq \log \frac{1}{n} - \frac{d}{2} \log(2\pi\sigma^2)$$
$$\xrightarrow[\sigma \to 0]{} +\infty .$$

The likelihood is thus **unbounded** over this class, and the maximiser reproduces the training set. In other words, a perfect score on the training objective is obtained by a model that has learned nothing.

The two subproblems



the second arrow is hard even when the box in the middle is given to you

We take them in the opposite order to the one the arrows suggest. Part II assumes that $\mathbb{P}$ is given and only asks for the second arrow, since the algorithms that solve it reappear inside the training loop of several models of Parts V and VI.

Part II. Sampling from a known distribution

Densities known up to a constant

For the whole of Part II, we assume that $\mathbb{P}$ is given. In practice, it is given in the following form, and only in this form.

Definition. A density is known **up to a normalising constant** when we can evaluate a function $\tilde{p} \geq 0$ such that, with $Z = \int_{\mathcal{X}} \tilde{p}(x)dx \in (0, \infty)$,

$$p(x) = \frac{\tilde{p}(x)}{Z} \ ,$$

and Z is not available.

The usual situation. In practice, $\tilde{p}$ is a product of local terms (one per pixel or per pair of neighbours), and evaluating it costs $O(d)$. In contrast, the constant Z is one integral over $\mathbb{R}^d$, or one sum over 2^d configurations if $\mathcal{X}$ is a hypercube, and no method of numerical analysis computes it for $d = 10^4$.

Two consequences. An algorithm that only uses ratios $\tilde{p}(x')/\tilde{p}(x)$ never sees Z, and neither does an algorithm that only uses $\nabla_x \log p = \nabla_x \log \tilde{p}$. The methods below rely on these two facts.

The inverse CDF in dimension one

Proposition. Let $F(x) = \mathbb{P}((-\infty, x])$ and $F^{-1}(u) = \inf\{x : F(x) \geq u\}$. If $U \sim \mathcal{U}([0,1])$ then $F^{-1}(U) \sim \mathbb{P}$.

Proof. Since F is non-decreasing and right-continuous, $F^{-1}(u) \leq x$ if and only if $u \leq F(x)$. Indeed, the direct implication uses $F(F^{-1}(u)) \geq u$, which follows from right-continuity, and the converse follows from the definition of an infimum. Hence, we have

$$\begin{aligned} \mathbb{P}(F^{-1}(U) \leq x) &= \mathbb{P}(U \leq F(x)) \\ &= F(x)\,, \qquad\qquad \text{since } U \sim \mathcal{U}([0,1]) \end{aligned}$$

which concludes the proof. ∎

Higher dimensions. In dimension d, the analogue would be the inverse of the map $x \mapsto (F_1(x^{(1)}), F_2(x^{(2)} \mid x^{(1)}), \dots)$, which requires every conditional cumulative distribution function of $\mathbb{P}$. Obtaining them is harder than the problem we started from, except when the model is *built* out of them, which is the case of Part IV.

Rejection sampling

Let us pick a proposal $\mathbb{Q}$ that we can sample, and a constant M such that $\tilde{p}(x) \leq Mq(x)$ for all x.

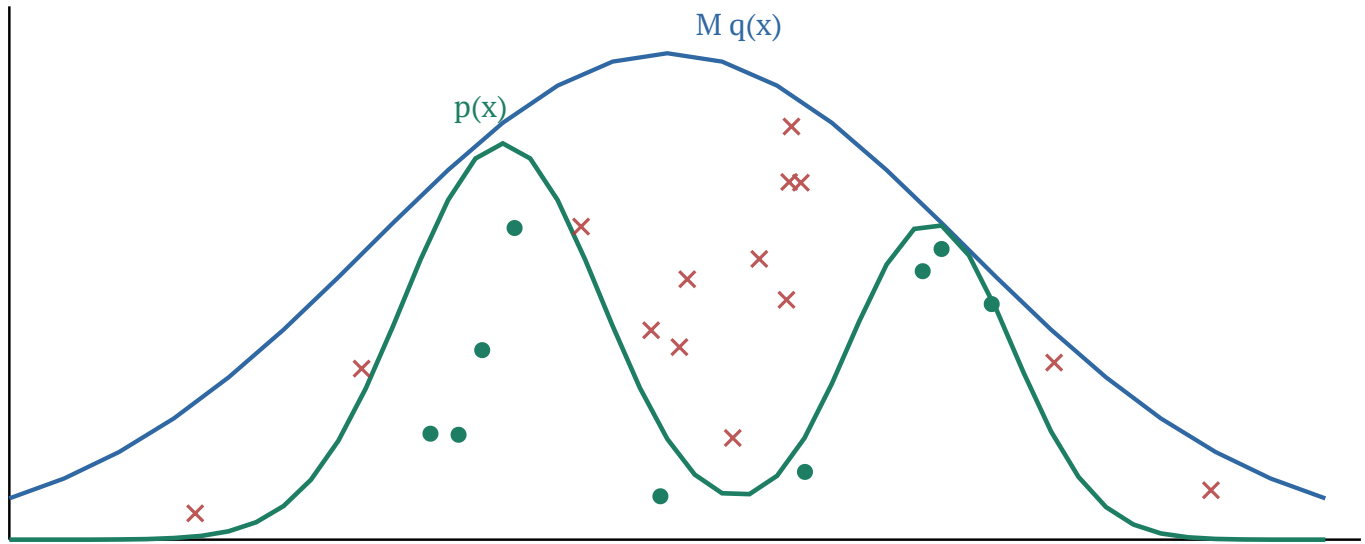
> **Proposition.** If we draw $x \sim \mathbb{Q}$ and $u \sim \mathcal{U}([0,1])$ independently, and accept x if $u \leq \tilde{p}(x)/(Mq(x))$, then conditionally on acceptance, $x \sim \mathbb{P}$, and the acceptance probability is Z/M.

Proof. For any measurable $B \subseteq \mathcal{X}$, we have

$$\begin{aligned}\mathbb{P}(x \in B, \text{accept}) &= \int_B q(x) \frac{\tilde{p}(x)}{Mq(x)} dx \quad \text{since } u \text{ is independent of } x \\ &= \frac{1}{M} \int_B \tilde{p}(x) dx \ .\end{aligned}$$

Taking $B = \mathcal{X}$ yields $\mathbb{P}(\text{accept}) = Z/M$, and dividing one by the other, we obtain $\mathbb{P}(x \in B \mid \text{accept}) = \int_B \tilde{p}/Z = \mathbb{P}(B)$, which concludes the proof. ■

Rejection sampling, drawn



accepted points fall under the green curve, rejected ones between the two curves

The accepted points are uniform on the region under p, which is what having law $\mathbb{P}$ means. The wasted work corresponds to the area between the two curves, so that the efficiency depends entirely on how tightly Mq hugs p.

Rejection sampling in high dimension

Let us consider the easiest possible case, $\mathbb{P} = \mathcal{N}(0, I_d)$ and $\mathbb{Q} = \mathcal{N}(0, \sigma^2 I_d)$ with $\sigma > 1$. Then

$$\frac{p(x)}{q(x)} = \sigma^d \exp\left(-\frac{\|x\|_2^2}{2}\left(1 - \frac{1}{\sigma^2}\right)\right),$$

which for $\sigma > 1$ is maximal at $x = 0$, so the smallest valid constant is $M = \sigma^d$ and the acceptance rate is σ^{-d}.

Numerically. With $\sigma = 1.1$ (i.e. a proposal wrong by ten percent in every coordinate), we get:

d	1	10	50	100
acceptance rate	0.91	0.39	8.5×10^{-3}	7.3×10^{-5}

Even with the friendliest target and a proposal as good as one can realistically hope for, in dimension 100 we throw away fourteen thousand draws per sample. More generally, every method that proposes a point *globally* and scores it degrades exponentially in d. A natural repair is to reweight rather than reject.

Importance sampling

Since rejection sampling throws away most of its work, we now keep every draw and correct for having drawn it from the wrong law.

Proposition. Let $\mathbb{Q}$ have density $q > 0$ wherever $p > 0$, and let f be $\mathbb{P}$-integrable. With the **importance weight** $w = p/q$,

$$\mathbb{E}_{x\sim\mathbb{P}}[f(x)] = \mathbb{E}_{x\sim\mathbb{Q}}[f(x)w(x)]\ ,$$

so for $x_1, \dots, x_N \sim \mathbb{Q}$ i.i.d. the estimator $\hat{I}_N = \frac{1}{N}\sum_{j\le N} f(x_j)w(x_j)$ is unbiased.

Proof. Since $w = p/q$ and $q > 0$ wherever $p > 0$, we have $\mathbb{E}_{x\sim\mathbb{Q}}[f(x)w(x)] = \int_{\mathcal{X}} f(x)p(x)dx = \mathbb{E}_{x\sim\mathbb{P}}[f(x)]$, and the unbiasedness of $\hat{I}_N$ follows by linearity. ■

A letter. We denote the number of draws by N rather than M, since M is the envelope constant of the rejection slide.

Expectations rather than samples. The estimator $\hat{I}_N$ estimates an expectation under $\mathbb{P}$, but it does not return a point distributed as $\mathbb{P}$, since the points come from $\mathbb{Q}$ and carry weights. Resampling them with probabilities proportional to the weights only yields approximate samples.

Importance sampling without the constant

The weight $w = p/q$ requires Z, which we do not have. Instead, we divide it out.

Given $x_1, \dots, x_N \sim \mathbb{Q}$ i.i.d. and $\tilde{w} = \tilde{p}/q$, the **self-normalised** estimator is

$$\hat{I}_N^{\mathrm{sn}} = \frac{\sum_{j \le N} f(x_j)\tilde{w}(x_j)}{\sum_{j \le N} \tilde{w}(x_j)} \, .$$

Consistency. By the law of large numbers, applied to the numerator and to the denominator separately, we have

$$\frac{1}{N}\sum_{j \le N} f(x_j)\tilde{w}(x_j) \xrightarrow[N\to\infty]{\text{a.s.}} Z\mathbb{E}_{x\sim\mathbb{P}}[f(x)] \, , \quad \text{since } \int_{\mathcal{X}} f\tilde{p} = Z\mathbb{E}_{x\sim\mathbb{P}}[f(x)]$$

$$\frac{1}{N}\sum_{j \le N} \tilde{w}(x_j) \xrightarrow[N\to\infty]{\text{a.s.}} Z \, , \qquad \text{since } \int_{\mathcal{X}} \tilde{p} = Z$$

and since $Z > 0$, the ratio converges to $\mathbb{E}_{x\sim\mathbb{P}}[f(x)]$, which concludes the proof. ■

Properties of the self-normalised estimator

An estimate of the normalising constant. The denominator estimates Z on its own, which no other sampler of this part can do, since Metropolis and Hastings, Gibbs and Langevin all arrange for Z to cancel and cannot report it afterwards. Hence, it is the tool to use when Z is itself the quantity of interest, as it will be in Part V.

Loss of unbiasedness. A ratio of two estimators is not unbiased. However, the bias is of order N^{-1}, against a standard deviation of order $N^{-1/2}$, so it is not usually the binding problem.

Sensitivity to the tails. For a generic f, the estimator has finite asymptotic variance only if $\mathbb{E}_{x\sim\mathbb{Q}}[w(x)^2] < \infty$, which can fail as soon as q has lighter tails than p. Hence, a proposal that is slightly too narrow need not degrade gracefully. It can rather produce an estimator with infinite asymptotic variance and a perfectly innocent looking output.

Degeneracy of the importance weights

Proposition. $\mathrm{Var}_{x\sim\mathbb{Q}}(w(x)) = \chi^2(\mathbb{P}\|\mathbb{Q})$, where $\chi^2(\mathbb{P}\|\mathbb{Q}) = \int_{\mathcal{X}} (p(x)-q(x))^2/q(x)\,dx$.

Proof. We have $\mathbb{E}_{x\sim\mathbb{Q}}[w(x)] = \int p = 1$, so $\mathrm{Var}_{x\sim\mathbb{Q}}(w) = \int p^2/q - 1 = \int (p^2 - 2pq + q^2)/q$, which is the stated integral. ∎

We call $N_{\text{eff}} = N / \mathbb{E}_{x\sim\mathbb{Q}}[w(x)^2]$ the **effective sample size** (i.e. the number of unweighted draws carrying the same information). On the Gaussian test of the rejection slide, with $\mathbb{P} = \mathcal{N}(0, I_d)$ and $\mathbb{Q} = \mathcal{N}(0, \sigma^2 I_d)$ and $\sigma^2 > 1/2$, one dimension and independence yield

$$\mathbb{E}_{x\sim\mathbb{Q}}[w(x)^2] = \left(\frac{\sigma^2}{\sqrt{2\sigma^2 - 1}} \right)^d .$$

Reading the comparison

At $\sigma = 1.1$, against the acceptance rate of the rejection slide:

d	1	10	50	100	1000
N_{eff}/N	0.98	0.86	0.47	0.22	2.3×10^{-7}
acceptance rate	0.91	0.39	8.5×10^{-3}	7.3×10^{-5}	4.0×10^{-42}

Two exponential rates. Reweighting is much better than rejecting, but both are exponential in d with different rates, and both are unusable at $d = 1000$.

A common cause. A point proposed from a global $\mathbb{Q}$ almost surely lands where p/q is tiny. Rejection reports this as a low acceptance rate, while importance sampling reports a weight vector in which one entry holds all the mass.

Local proposals. The way out is to propose *locally*, from wherever the last accepted point was, which is what a Markov chain does and what the rest of this part studies.

Invariant distributions of Markov chains

We now give up on independent samples, and we build instead a chain $(X_t)_{t \geq 0}$ whose law converges to $\mathbb{P}$, in order to read the sample off the trajectory.

Definition. A **transition kernel** K assigns to each $x \in \mathcal{X}$ a probability measure $K(x, \cdot)$ on $\mathcal{X}$, the law of X_{t+1} given $X_t = x$. A probability measure $\mathbb{P}$ is **invariant** for K when, for every measurable $B \subseteq \mathcal{X}$,

$$\int_{\mathcal{X}} K(x, B) d\mathbb{P}(x) = \mathbb{P}(B) \ .$$

In other words, if the chain starts at $\mathbb{P}$, it stays at $\mathbb{P}$. Note that invariance does not say that the chain reaches $\mathbb{P}$ from elsewhere, which is a separate statement.

The design problem. We are given $\mathbb{P}$ and we must produce a K that leaves it invariant and that we can simulate. However, checking invariance from the definition requires an integral against $\mathbb{P}$, which we cannot compute, and detailed balance removes that difficulty.

Detailed balance

Proposition. Suppose $K(x, \cdot)$ has density $k(x, \cdot)$ and that, for all $x, x' \in \mathcal{X}$,

$$p(x)k(x, x') = p(x')k(x', x) \ .$$

Then $\mathbb{P}$ is invariant for K.

Proof. Integrating detailed balance in the first variable, we obtain

$$\begin{aligned} \int_{\mathcal{X}} p(x)k(x, x')dx &= \int_{\mathcal{X}} p(x')k(x', x)dx && \text{by detailed balance} \\ &= p(x') \int_{\mathcal{X}} k(x', x)dx \\ &= p(x') \ , && \text{since } k(x', \cdot) \text{ is a probability density} \end{aligned}$$

which concludes the proof. ■

We have thus replaced a condition involving an integral against $\mathbb{P}$ by a condition on one pair of points at a time, which is what makes Markov chain Monte Carlo practical. Note that detailed balance is invariant under

Consequences of invariance

Theorem (ergodic theorem). If K is irreducible and aperiodic with invariant distribution $\mathbb{P}$, then for $\mathbb{P}$-almost every initial point X_0 and every f with $\mathbb{E}_{x\sim\mathbb{P}}[|f(x)|] < \infty$,

$$\frac{1}{T}\sum_{t=1}^{T} f(X_t) \xrightarrow[T\to\infty]{\text{a.s.}} \mathbb{E}_{x\sim\mathbb{P}}[f(x)] \;,$$

$$\|\mathcal{L}(X_T) - \mathbb{P}\|_{\mathrm{TV}} \xrightarrow[T\to\infty]{} 0 \;.$$

The proof is out of the scope of this lecture. The two limits are Fact 5 and Theorem 4 of Roberts and Rosenthal (2004), or Theorems 17.0.1 (i) and 13.3.4 (ii) of Meyn and Tweedie (1993).

A law of large numbers without independence. The introductory lecture used the classical law of large numbers to justify $\widehat{R}_S \approx R$, and this theorem is the same statement for a dependent sequence, which makes a trajectory usable as a sample.

No rate. The theorem is silent on how large T must be, and the answer can be e^{cd}, since a chain that has to cross a low density region to move between two modes waits an exponentially long time. In other words, MCMC fails through its mixing time, not through its correctness.

The Metropolis and Hastings algorithm

Let us pick any proposal density $q(x' \mid x)$ that is easy to sample, for instance $\mathcal{N}(x, \tau^2 I_d)$.

From $X_t = x$, the algorithm draws $x' \sim q(\cdot \mid x)$, computes

$$\alpha(x, x') = \min\left(1, \frac{\tilde{p}(x')q(x \mid x')}{\tilde{p}(x)q(x' \mid x)}\right),$$

and sets $X_{t+1} = x'$ with probability $\alpha(x, x')$, and $X_{t+1} = x$ otherwise.

No normalising constant. The density $\tilde{p}$ appears only through the ratio $\tilde{p}(x')/\tilde{p}(x)$, in which Z cancels, as we observed when introducing $\tilde{p}$.

Cost per step. Each step requires two evaluations of $\tilde{p}$, one draw from q and one uniform draw, and nothing depends on d beyond the cost of evaluating $\tilde{p}$ itself.

Correctness of Metropolis and Hastings

Proposition. The Metropolis and Hastings kernel satisfies detailed balance with respect to $\mathbb{P}$, hence $\mathbb{P}$ is invariant for it.

Proof. For $x' \neq x$, the transition density is $k(x, x') = q(x' \mid x)\alpha(x, x')$, and writing $p = \tilde{p}/Z$, we have

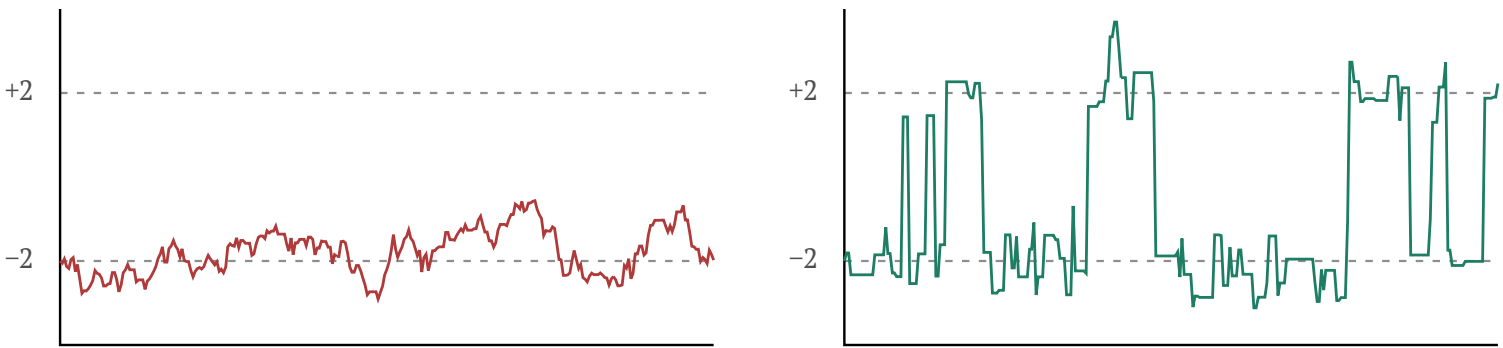
$$\begin{aligned} p(x)k(x,x') &= \frac{\tilde{p}(x)}{Z} q(x' \mid x) \min\left(1, \frac{\tilde{p}(x')q(x \mid x')}{\tilde{p}(x)q(x' \mid x)}\right) && \text{by definition of } \alpha \\ &= \frac{1}{Z} \min\left(\tilde{p}(x)q(x' \mid x), \tilde{p}(x')q(x \mid x')\right) . && \text{since } a\min(1, b/a) = \min(a,b) \text{ for } a > 0 \end{aligned}$$

Since the right-hand side is symmetric in x and x', it equals $p(x')k(x', x)$, and the case $x' = x$ is trivial, which concludes the proof. ■

The proof works for *any* q with the right support, which is both the strength of the method and its weakness. Indeed, correctness comes for free, and all the difficulty has been pushed into the mixing time, which depends entirely on q.

Metropolis and Hastings, drawn

We run three hundred steps on $p \propto \frac{1}{2}\mathcal{N}(-2, 0.5^2) + \frac{1}{2}\mathcal{N}(2, 0.5^2)$, with the same target and the same code, for two values of the proposal width τ.



τ = 0.2, acceptance 0.84

never leaves the mode it started in

τ = 2, acceptance 0.32

crosses thirteen times

horizontal axis: iteration. vertical axis: the state. dashed lines: the two modes.

Both chains are correct, since both have $\mathbb{P}$ as invariant law by the proposition just proved. However, the left one is useless, and no diagnostic run on it alone would say so.

Gibbs sampling

When the conditionals $p(x^{(j)} \mid x^{(-j)})$ are available in closed form, we cycle through the coordinates and resample each of them from its conditional, holding the others fixed.

Each update is a Metropolis and Hastings step with the proposal $q(x' \mid x) = p(x'^{(j)} \mid x^{(-j)})$, which only changes coordinate j. For such a move, we have

$$\begin{aligned} \frac{\tilde{p}(x')q(x \mid x')}{\tilde{p}(x)q(x' \mid x)} &= \frac{p(x'^{(j)}, x^{(-j)})p(x^{(j)} \mid x^{(-j)})}{p(x^{(j)}, x^{(-j)})p(x'^{(j)} \mid x^{(-j)})} \\ &= 1 \,, \end{aligned}$$

where we used $p(a, b) = p(a \mid b)p(b)$ twice and $x'^{(-j)} = x^{(-j)}$. As a result, **the acceptance probability is always one.**

Scope. The model must be specified by its conditionals, as for Ising and Potts models, latent Dirichlet allocation, or any graphical model with conjugate factors. Each conditional is then a law on one coordinate, whose normalisation is a sum over the alphabet and costs nothing.

Failure mode. When coordinates are strongly dependent, the chain follows a narrow ridge in small steps, one coordinate at a time, and the mixing time explodes.

Langevin dynamics

A different idea is to follow the gradient of $\log p$, and to add just enough noise to stop the trajectory from collapsing onto a mode.

> The **overdamped Langevin diffusion** is the solution of the stochastic differential equation
>
> $$dX_t = \nabla_x \log p(X_t) dt + \sqrt{2} dB_t \ ,$$
>
> where (B_t) is a standard Brownian motion on $\mathbb{R}^d$.

Dependence on $\tilde{p}$ only. We have $\nabla_x \log p = \nabla_x \log \tilde{p}$, since $\log Z$ is a constant in x. Hence, the normalising constant disappears a second time, and this time without any ratio.

A noisy gradient descent. The drift is minus the gradient of the potential $-\log p$, so that the deterministic part of the dynamics is the gradient flow of the optimization lecture, run on the objective $-\log p$. The Brownian term then turns a search for the mode into a sampler for the whole distribution.

Invariance of $\mathbb{P}$ for the Langevin diffusion

Proposition. $\mathbb{P}$ is invariant for the overdamped Langevin diffusion.

Proof. The density ρ_t of X_t solves the Fokker and Planck equation associated with the drift $b = \nabla_x \log p$ and the diffusion coefficient $\sqrt{2}$,

$$\partial_t \rho_t = -\nabla_x \cdot (\rho_t \nabla_x \log p) + \Delta_x \rho_t \ .$$

Substituting $\rho_t = p$, we obtain

$$\begin{aligned} -\nabla_x \cdot (p \nabla_x \log p) + \Delta_x p &= -\nabla_x \cdot (\nabla_x p) + \Delta_x p \quad \text{since } p \nabla_x \log p = \nabla_x p \\ &= 0 \ , \end{aligned}$$

which concludes the proof. ■

The computation shows more than invariance, since the drift was *chosen* to make the probability current $\rho \nabla_x \log p - \nabla_x \rho$ vanish identically at $\rho = p$. Any other drift would leave a nonzero current, which changes the invariant law unless it is divergence free.

The unadjusted Langevin algorithm

We discretise the diffusion with the Euler and Maruyama scheme, with step $\gamma > 0$ and noises $\epsilon_1, \epsilon_2, \cdots \sim \mathcal{N}(0, I_d)$ i.i.d.:

$$X_{t+1} = X_t + \gamma \nabla_x \log \tilde{p}(X_t) + \sqrt{2\gamma}\epsilon_{t+1} \ .$$

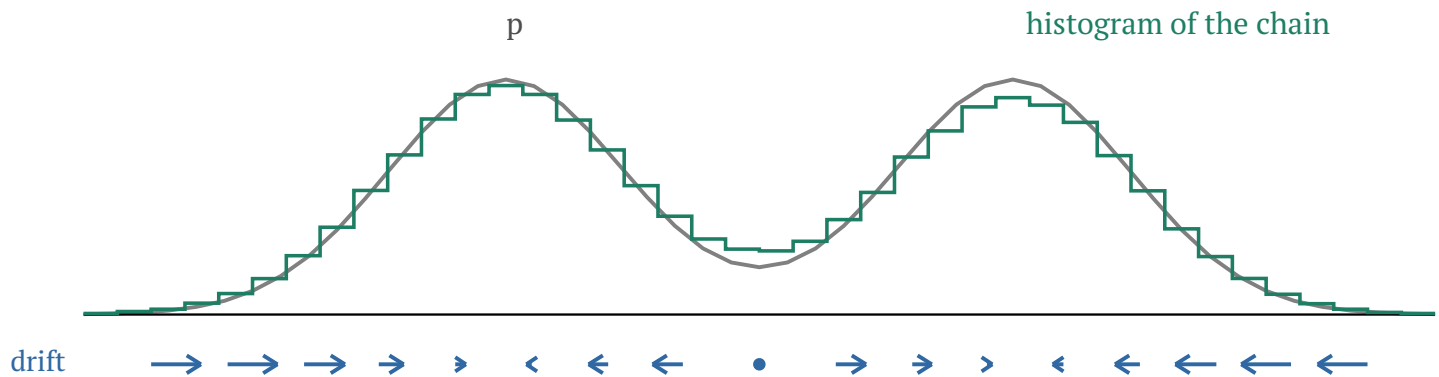
A gradient and a Gaussian per step. There is no accept or reject step, no evaluation of $\tilde{p}$ itself, and no proposal to tune beyond γ, which makes it the only sampler of this part whose cost per step is one backward pass through a network.

A bias. The invariant law of the discretised chain is a perturbation of $\mathbb{P}$, at distance $O(\gamma)$ in the usual metrics. Shrinking γ removes the bias and slows the mixing, which is the same trade-off as for a learning rate.

The correction. Adding a Metropolis and Hastings accept or reject step on top, with $q(\cdot \mid x) = \mathcal{N}(x + \gamma \nabla_x \log p(x), 2\gamma I_d)$, removes the bias entirely. We obtain the Metropolis adjusted Langevin algorithm, which is the practical default whenever $\tilde{p}$ itself can be evaluated.

Langevin, drawn

We show the drift $\nabla_x \log p$ and four hundred thousand steps of the unadjusted algorithm at $\gamma = 0.08$, on $p \propto \frac{1}{2}\mathcal{N}(-1.5, 0.7^2) + \frac{1}{2}\mathcal{N}(1.5, 0.7^2)$.



the drift points uphill and vanishes at the modes and at the saddle between them

The drift alone, which is gradient ascent on $\log p$, would stop at a mode. The noise turns it into a sampler, and the histogram sits on the density because the two are in the balance described by the Fokker and Planck computation.

Scope and limits of Part II

Scope. We can sample from any $\mathbb{P}$ that we can evaluate up to a constant, or whose $\nabla_x \log p$ we can evaluate, at the cost of correlated samples and an unknown mixing time.

Limits. We do not have $\mathbb{P}$ itself, since everything above assumed that $\tilde{p}$, or $\nabla_x \log p$, was handed to us, whereas in generative modelling we only have S.

The rest of the lecture. We learn a $\tilde{p}_\theta$, or a $\nabla_x \log p_\theta$, from S, and we then run Part II on it. Some families make the second step trivial at the cost of the first, and others do the reverse.

Part II inside the training loop. Energy based models need samples from the current $\mathbb{P}_\theta$ at every gradient step, and score based models learn the object that the Langevin dynamics consumes. Both are studied in Part VI.