Practice 2 — The Random Walk, Forward and Backward

Mathematical Foundations of Diffusion Models
Monsoon 2026 · IIT Jammu

Worked derivations and exercises. Not graded, not submitted

The spine of this sheet

Toss a coin. Heads is $+1$, tails is -1. Add the result to a running total, and repeat n times:

$$X_1, X_2, \dots \text{ i.i.d.}, \quad \mathbb{P}(X_i = +1) = \mathbb{P}(X_i = -1) = \tfrac{1}{2}, \qquad S_n = \sum_{i=1}^{n} X_i.$$

That is the **simple symmetric random walk**, and it is the forward process of a diffusion model with the noise made as simple as noise can possibly be. Start somewhere, add independent noise repeatedly, end somewhere structureless. Compare the two update rules side by side:

$$\underbrace{S_n = S_{n-1} + X_n}_{\text{this sheet}} \qquad \underbrace{x_t = \sqrt{\alpha_t}\, x_{t-1} + \sqrt{1-\alpha_t}\, \varepsilon_t}_{\text{Practice 1, §6.3}}$$

Everything that makes the real process tractable is already present here, and nothing that makes it hard is. So we can do the whole thing by hand.

> **The concept**
>
> **What this sheet gets, and in what order.** Part A runs the walk forward: its mean and variance, its *exact* law at every step, the equation its density obeys, and what it becomes in the continuum. Part B runs it backward — and the reverse step turns out to have a closed form, so we end with a complete diffusion model in which nothing is learned and nothing is approximated.
>
> The last section is the honest one: what breaks when the data stops being a coin.

Part A — the walk forward

1. What the object is

S_n is a **discrete-time Markov chain** on the integers: given S_{n-1}, the next state is $S_{n-1} \pm 1$ with probability $\frac{1}{2}$ each, and nothing earlier matters. Its transition kernel is

$$K(x, y) = \begin{cases} \frac{1}{2} & y = x + 1 \\ \frac{1}{2} & y = x - 1 \\ 0 & \text{otherwise.} \end{cases}$$

Three features are worth naming immediately, because each is a property the real forward process shares.

- **The noise accumulates.** Each step adds a fresh independent increment and nothing removes it, so uncertainty grows with n. §2 makes "grows" precise.
- **No training is required.** The forward direction is fully specified in advance. This is the easy direction, exactly as in a diffusion model.
- **The destination is structureless.** §6 shows it is Gaussian, whatever the coin.

> **The concept**
>
> **One difference from Practice 1, and it is deliberate.** The DDPM update $x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\varepsilon_t$ shrinks the old state before adding noise, which keeps the total variance at 1 — *variance preserving*. The walk has no shrinkage at all, so its variance grows without bound — *variance exploding*.
>
> Both are used in practice (DDPM is the first, score-based SDE models often the second). Keep the pair in view: the difference is one multiplicative factor, and it is a design choice rather than a mathematical necessity.

2. Mean, variance, and the $\sqrt{n}$ envelope

Each increment has $\mathbb{E}[X_i] = \frac{1}{2}(+1) + \frac{1}{2}(-1) = 0$ and, since $X_i^2 = 1$ always,

$$\mathrm{Var}(X_i) = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 = 1 - 0 = 1.$$

The increments are independent, so **variances add** — this is Practice 1 §6.2, used for the first time on something other than a Gaussian:

$$\boxed{\mathbb{E}[S_n] = 0, \qquad \mathrm{Var}(S_n) = n, \qquad \mathrm{sd}(S_n) = \sqrt{n}.}$$

> **The concept**
>
> **This is the $\sqrt{t}$ envelope.** The W01 demo "Progressively Adding Noise" shows a cloud of walkers spreading, with a $\sqrt{t}$ curve drawn over it. That curve is the line above. A walk does not travel a distance proportional to the number of steps — it travels proportional to the *square root* of it, because the steps cancel in a way that lets variances rather than distances accumulate.
>
> Everything in the course that spreads like $\sqrt{t}$ spreads for this reason.

3. The exact law, at every step

This section is the longest in the sheet, deliberately. The formula at the end is two lines to state, but getting there properly means being careful about what the underlying sample space is — and that carefulness is the point. We build the answer by hand for $n = 0, 1, 2, 3$, read the pattern, and then prove it by induction.

3.1 First, what is an outcome?

HW 1 Problem 1 insisted on three objects: a sample space Ω, an outcome $\omega \in \Omega$, and a random variable $X : \Omega \to \mathbb{R}$ that attaches a number to each outcome. On a die that distinction felt like pedantry, because the outcome was already a number and $X(\omega) = \omega$ was the identity map.

Here it is not pedantry, because **an outcome is a whole path**:

$$\Omega_n = \{-1, +1\}^n, \qquad \omega = (\omega_1, \omega_2, \ldots, \omega_n), \qquad |\Omega_n| = 2^n.$$

One ω is one complete record of every toss, in order — not a position, and not a number at all. The tosses are fair and independent, so every path is equally likely:

$$\mathbb{P}(\{\omega\}) = 2^{-n} \quad \text{for each of the } 2^n \text{ paths.}$$

Two random variables live on this space, and both are honest functions:

$$X_i(\omega) = \omega_i \qquad \text{(read off the } i\text{-th coordinate)}, \qquad\qquad S_n(\omega) = \sum_{i=1}^{n} \omega_i.$$

X_i is one-to-one on each coordinate and uninteresting. S_n is the interesting one, and the reason is that it is **massively many-to-one**: it crushes 2^n paths down onto at most $n + 1$ positions.

The concept

The chain from HW 1, running again.

$$\text{path } \omega \xrightarrow{S_n} \text{position} \longrightarrow \text{condition “}= k\text{”}$$
$$\longrightarrow \text{event } \{S_n = k\} = S_n^{-1}(\{k\}) \subseteq \Omega_n \longrightarrow \text{probability}$$

The event $\{S_n = k\}$ is a *set of paths* — the preimage of k. S_n is never evaluated at that set; the set is what the question “where does S_n return k?” produces.

And now the step that turns probability into combinatorics. Because the measure on Ω_n is **uniform**,

$$\boxed{\mathbb{P}(S_n = k) \;=\; \frac{\left|\, S_n^{-1}(\{k\}) \,\right|}{2^n} \;=\; \frac{\text{number of paths ending at } k}{\text{total number of paths}}.}$$

So the whole problem is: **count the paths**. Everything below is counting.

3.2 Small cases, written out in full

Write a path compactly as a string of $+$ and $-$. Let

$$N_n(k) = \left|\{\omega \in \Omega_n : S_n(\omega) = k\}\right|$$

be the number of paths of length n ending at k.

$n = 0$**.** There is exactly one path: the empty one. You have tossed nothing, so you are at the origin.

$$\Omega_0 = \{\,()\,\}, \qquad N_0(0) = 1, \qquad N_0(k) = 0 \;\text{ for } k \neq 0.$$

$n = 1$**.** Two paths.

k	paths with $S_1 = k$	$N_1(k)$
-1	`-`	1
$+1$	`+`	1

Total $1 + 1 = 2 = 2^1$. ✓

$n = 2$**.** Four paths, and the first genuine collision: two *different* paths land on the same position.

k	paths with $S_2 = k$	$N_2(k)$
-2	`--`	1
0	`+-`, `-+`	**2**
$+2$	`++`	1

Total $1 + 2 + 1 = 4 = 2^2$. ✓ Note also that $N_2(\pm 1) = 0$: after two steps you cannot be at an odd position.

$n = 3$**.** Eight paths.

k	paths with $S_3 = k$	$N_3(k)$
-3	`---`	1
-1	`+--`, `-+-`, `--+`	3
$+1$	`++-`, `+-+`, `-++`	3
$+3$	`+++`	1

Total $1 + 3 + 3 + 1 = 8 = 2^3$. ✓

Look at the three paths ending at $k = -1$: each has exactly one $+$ and two $-$, and they differ only in *which* of the three steps was the $+$. That observation is the whole theorem, and §3.4 makes it a proof.

3.3 The pattern: Pascal's triangle, lying on its side

Collect the counts:

n	-5	-4	-3	-2	-1	0	1	2	3	4	5	total
0	.	.	.	.	.	1	.	.	.	.	.	1
1	.	.	.	.	1	.	1	.	.	.	.	2
2	.	.	.	1	.	2	.	1	.	.	.	4
3	.	.	1	.	3	.	3	.	1	.	.	8
4	.	1	.	4	.	6	.	4	.	1	.	16
5	1	.	5	.	10	.	10	.	5	.	1	32

Three things are visible at once.

- **It is Pascal's triangle**, spread out so that consecutive entries sit two columns apart. The dots are the parity constraint: half the lattice is unreachable at any given time.
- **Each row sums to** 2^n, as it must — the rows partition Ω_n, and the paths have to go somewhere.
- **Each entry is the sum of the two above it**, one column left and one column right. That is §3.4's recursion, and it is Pascal's rule.

3.4 The recursion, and the induction

The recursion. Take any path of length $n+1$ ending at k and delete its last step. What remains is a path of length n, ending at $k-1$ if the deleted step was $+1$, and at $k+1$ if it was -1. This is a *bijection*: every length-n path ending at $k \mp 1$ extends in exactly one way to a length-$(n+1)$ path ending at k. Hence

$$\boxed{N_{n+1}(k) = N_n(k-1) + N_n(k+1).}$$

(Verified against direct enumeration for $n \le 11$ and all k.)

The claim. For $|k| \le n$ with $n+k$ even, and writing $j = (n+k)/2$,

$$N_n(k) = \binom{n}{j} = \binom{n}{\frac{n+k}{2}},$$

and $N_n(k) = 0$ otherwise.

Base case, $n = 0$. $N_0(0) = 1 = \binom{0}{0}$, and $N_0(k) = 0$ for $k \neq 0$. ✓

Inductive step. Assume the claim at n; prove it at $n+1$.

Parity. If $n+1+k$ is odd then both $n+(k-1)$ and $n+(k+1)$ are odd, so by the inductive hypothesis both terms of the recursion vanish and $N_{n+1}(k) = 0$, as claimed.

The main case. Let $n+1+k$ be even and put $j = (n+1+k)/2$. Then

$$\frac{n+(k-1)}{2} = j-1, \qquad \frac{n+(k+1)}{2} = j,$$

so the inductive hypothesis turns the recursion into

$$N_{n+1}(k) = \binom{n}{j-1} + \binom{n}{j} = \binom{n+1}{j},$$

the last equality being **Pascal's rule**. And $j = (n+1+k)/2$ is exactly the index the claim asks for at $n+1$. ■

> **The concept**
>
> **Where the binomial coefficient actually comes from.** The induction proves the formula; this one-line argument explains it.
>
> A path with j up-steps and $n-j$ down-steps ends at $j-(n-j) = 2j-n$. So "ending at k" is the same condition as "having exactly $j = (n+k)/2$ up-steps", and the paths ending at k are in bijection with the ways of choosing *which* j of the n steps went up — that is, with the j-element subsets of $\{1, 2, \ldots, n\}$. There are $\binom{n}{j}$ of those.
>
> The position does not record *when* the up-steps happened, only *how many* there were. That is precisely the information S_n destroys.

3.5 The law

Dividing the counts by 2^n, as §3.1 licensed:

$$\boxed{\mathbb{P}(S_n = k) = \binom{n}{\frac{n+k}{2}} 2^{-n}, \qquad k \in \{-n,\, -n+2,\, \ldots,\, n-2,\, n\}.}$$

Equivalently and more compactly: if K_n is the number of up-steps then $K_n \sim \text{Binomial}(n, \frac{1}{2})$ and

$$S_n = K_n - (n - K_n) = 2K_n - n, \qquad \text{so} \qquad K_n = \frac{n+S_n}{2}.$$

The walk is a relabelled binomial. Everything known about Binomial$(n, \frac{1}{2})$ transfers, including $\mathbb{E}[K_n] = n/2$ and $\mathrm{Var}(K_n) = n/4$, which reproduce $\mathbb{E}[S_n] = 0$ and $\mathrm{Var}(S_n) = 4 \cdot n/4 = n$ from §2. ✓

Written out for $n = 4$:

k	$j = (4+k)/2$	$N_4(k) = \binom{4}{j}$	$\mathbb{P}(S_4 = k)$	decimal
-4	0	1	$1/16$	0.0625
-2	1	4	$4/16 = 1/4$	0.2500
0	2	6	$6/16 = 3/8$	0.3750
2	3	4	$4/16 = 1/4$	0.2500
4	4	1	$1/16$	0.0625

Sums to 1; mean 0; variance 4, as §2 promised.

The **parity constraint** deserves one more sentence because it is easy to state carelessly. After n steps only positions of the same parity as n are reachable, since each step changes S by an odd amount. So $\mathbb{P}(S_4 = 1) = 0$ — not small, *exactly* zero, and no approximation that assigns it positive mass is telling the truth about the walk.

> **The concept**
>
> **This is the discrete $q(x_t \mid x_0)$.** Practice 1 §6.3 derived $q(x_t \mid x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)$ and made the point that a *closed-form marginal* is what lets you jump to any noise level in one step, instead of simulating the chain. The boxed formula above is that same property, with a binomial coefficient where the Gaussian would be.
>
> Without it, sampling S_{1000} would mean tossing a thousand coins. With it, one binomial draw.

> **The concept**
>
> **And here is why Part B will need Bayes.** S_n maps 2^n paths onto $n+1$ positions, so by the time you observe $S_n = k$ you have thrown away almost everything: which steps went up, and in what order. At $n = 20$ there are 1,048,576 paths and only 21 possible answers.
>
> That collapse is exactly what makes running the walk backwards a genuine inference problem rather than a lookup. It is the same collapse as noising an image — many clean images are consistent with one noisy observation — and §7 is where we invert it.

4. The equation the density obeys

Now watch the distribution evolve rather than the walker. Write $p_n(k) = \mathbb{P}(S_n = k)$.

There are exactly two ways to be at k at time $n+1$: be at $k-1$ and step up, or be at $k+1$ and step down. Each has probability $\frac{1}{2}$. So

$$\boxed{p_{n+1}(k) = \tfrac{1}{2}\,p_n(k-1) + \tfrac{1}{2}\,p_n(k+1)}$$

— the **master equation** of the chain. (Verified directly against the binomial formula of §3.)

Subtract $p_n(k)$ from both sides and regroup, adding and subtracting nothing:

$$p_{n+1}(k) - p_n(k) = \tfrac{1}{2}\,p_n(k-1) - p_n(k) + \tfrac{1}{2}\,p_n(k+1) = \tfrac{1}{2}\Big[\,p_n(k-1) - 2p_n(k) + p_n(k+1)\Big].$$

Now read the two sides:

$$\underbrace{p_{n+1}(k) - p_n(k)}_{\text{difference in } \textit{time}} \;=\; \tfrac{1}{2}\underbrace{\Big[p_n(k+1) - 2p_n(k) + p_n(k-1)\Big]}_{\text{second difference in } \textit{space}}$$

The left side is a discrete time derivative. The right side is the standard three-point second difference, the discrete version of ∂_x^2. Passing to the continuum (§5 does this properly),

$$\boxed{\frac{\partial p}{\partial t} = \frac{1}{2}\frac{\partial^2 p}{\partial x^2}}$$

> **The concept**
>
> **That is the heat equation**, and it is the Fokker–Planck equation for Brownian motion — Module 5's centrepiece, derived here from a coin toss with nothing but addition and subtraction.
>
> It says something the walk-level picture hides: *the density diffuses.* Where p is concave (a peak) the second difference is negative and the density falls; where it is convex (a trough) it rises. Peaks flatten, troughs fill, and the whole thing spreads — which is what "adding noise" does to a distribution.
>
> Practice 1 tracked a *sample* through the forward process. This tracks the *distribution*, and Modules 5 and 6 are conducted almost entirely in this second language.

5. The continuum limit, and where $\sqrt{\Delta t}$ comes from

The walk lives on the integers and moves once per tick. To get a limit, shrink both: let the lattice spacing be h and the time step Δt, so that after time t the walker has taken $n = t/\Delta t$ steps of size h. Then, by §2,

$$\text{Var}\left(\text{position at time } t\right) = nh^2 = \frac{t\,h^2}{\Delta t}.$$

Now ask what choice of h keeps this finite and non-zero as $\Delta t \to 0$. There is only one.

choice	variance at time t	limit
$h = \Delta t$	$t\,\Delta t \to 0$	the walk **dies**: no randomness survives
$h = \sqrt{\Delta t}$	t	**Brownian motion**
$h = (\Delta t)^{1/4}$	$t/\sqrt{\Delta t} \to \infty$	the walk **explodes**

> **The concept**
>
> **This is why an SDE is written $dX = b\,dt + \sigma\,dW$ with dW of size $\sqrt{dt}$**, and not dt.
>
> It is the question every student asks when they first meet stochastic calculus, and the answer is the arithmetic above: *variances add, so standard deviations accumulate like $\sqrt{n}$.* Noise scaled by Δt is too weak to survive the limit; noise scaled by any smaller power blows up. Only the square root balances.
>
> The same fact explains a detail of the DDPM parametrisation: the noise coefficient is $\sqrt{1-\alpha_t}$, a square root, because it multiplies the *variable* while variances scale by its *square.*

Under this scaling the master equation of §4 becomes the heat equation exactly: the time difference divided by Δt approaches $\partial_t p$, the second difference divided by $h^2 = \Delta t$ approaches $\partial_x^2 p$, and the two sides match with the factor $\frac{1}{2}$ intact.

6. The Gaussian is a limit, not an assumption

6.1 The central limit theorem, on this example

S_n is a sum of n i.i.d. terms with mean 0 and variance 1, so

$$\frac{S_n}{\sqrt{n}} \implies \mathcal{N}(0,1).$$

Equivalently, and this is the older statement: the binomial distribution of §3, once centred and scaled, approaches a Gaussian density. That is the de Moivre–Laplace theorem, and it is the original central limit theorem.

> **The concept**
>
> **The Gaussian in a diffusion model is not an assumption. It is a limit.**
>
> The increments here are as un-Gaussian as a random variable can be — two atoms, no density at all. After enough steps the marginal is Gaussian anyway. So the choice of Gaussian noise in a diffusion model is not a modelling restriction being smuggled in; after many small steps you would arrive at a Gaussian whatever the increments looked like. Choosing Gaussians simply means the closed forms hold at *every* t instead of asymptotically.
>
> The Gaussian Lab makes the same point from the other side, with four non-Gaussian sources converging under averaging.

6.2 Which also explains the shape of the walk's law

Substituting the CLT into §3, $\mathbb{P}(S_n = k) \approx \sqrt{\frac{2}{\pi n}}\, e^{-k^2/2n}$ for k of the right parity and $|k| \ll n$ — the factor 2 because only half the integers are reachable. The binomial coefficients *are* a Gaussian, in disguise and on a lattice.

6.3 And in d dimensions: the walk lives on a shell

Run d independent walks, one per coordinate, and let $S_n \in \mathbb{Z}^d$ be the vector. Each coordinate has variance n, so

$$\mathbb{E}\big[\|S_n\|^2\big] = nd, \qquad \text{typical distance from the origin} \approx \sqrt{nd}.$$

The surprise is how *tight* that is. Simulation, with $n = 400$ steps per coordinate, reporting $\|S_n\|/\sqrt{n}$:

d	mean	$\sqrt{d}$	relative width
1	0.801	1.000	0.749
10	3.068	3.162	0.226
100	9.961	10.000	0.070
1000	31.609	31.623	0.022

The standard deviation of $\|S_n\|/\sqrt{n}$ stays near $1/\sqrt{2} \approx 0.707$ no matter how large d gets, while the mean grows like $\sqrt{d}$. So the relative width shrinks like $1/\sqrt{2d}$, and in high dimension essentially every sample lies in a thin shell at radius $\sqrt{nd}$.

The concept

Nobody is at the origin. In $d = 1000$ the walker is at distance $31.6\sqrt{n}$, give or take 2%, essentially always — even though the origin is the single most likely individual point.

This is the Gaussian annulus, and it is the geometric content of note 06's curse of dimensionality. It is also why $x_T \sim \mathcal{N}(0, I_d)$ in a diffusion model is not "a blob around zero" but a sphere of radius $\sqrt{d}$: the reverse process starts on a shell, not at a point.

Part B — the walk backward

7. The reverse step, exactly

Here is the diffusion question, asked of a coin: *you are told where the walk ended. What was the last step?*

$$\mathbb{P}\big(X_n = +1 \,\big|\, S_n = k\big) = \,?$$

7.1 By Bayes

To have $S_n = k$ with $X_n = +1$, the walk must have been at $k-1$ after $n-1$ steps and then stepped up. With $j = (n+k)/2$ the number of heads, being at $k-1$ after $n-1$ steps means $j-1$ heads out of $n-1$:

$$\mathbb{P}(X_n = +1 \mid S_n = k) = \frac{\mathbb{P}(S_{n-1} = k-1)\cdot \frac{1}{2}}{\mathbb{P}(S_n = k)} = \frac{\binom{n-1}{j-1}2^{-(n-1)}\cdot\frac{1}{2}}{\binom{n}{j}2^{-n}} = \frac{\binom{n-1}{j-1}}{\binom{n}{j}}.$$

The powers of two cancel completely. Now use $\binom{n}{j} = \frac{n}{j}\binom{n-1}{j-1}$:

$$\boxed{\mathbb{P}\big(X_n = +1 \mid S_n = k\big) = \frac{j}{n} = \frac{n+k}{2n}.}$$

Sanity checks, which students should do before believing it: at $k = n$ (all heads) this is 1; at $k = -n$ it is 0; at $k = 0$ it is $\frac{1}{2}$. Verified exhaustively over all 2^n paths for small n.

7.2 By symmetry, in one line

Now the argument that explains the answer rather than merely producing it.

Given $S_n = k$, every path with j heads has the same probability — the probability of a path depends only on *how many* heads it contains, not on where they fall. So conditionally the increments are **exchangeable**, and $\mathbb{E}[X_i \mid S_n = k]$ is the same number c for every i. But they must add up:

$$\sum_{i=1}^{n} \mathbb{E}[X_i \mid S_n = k] = \mathbb{E}[S_n \mid S_n = k] = k \qquad \Longrightarrow \qquad nc = k.$$

Therefore

$$\boxed{\mathbb{E}\big[X_i \mid S_n = k\big] = \frac{k}{n} \quad \text{for every } i.}$$

The expected step is the average step. No binomial coefficients, no algebra.

The concept

This is a denoiser. Read it against Practice 1 §7.1: a hidden quantity (X_n), a noisy observation (S_n), a posterior, and its mean — and §7.1 proved that the conditional mean is exactly what squared-error training converges to.

So k/n is what a network trained to predict the increment from the endpoint would learn. Here we have it in closed form, so we can check the answer a network would have to discover.

8. Reverse sampling: a diffusion model with nothing learned

The reverse step is all you need to run the process backwards. Starting from $S_n = k$:

(a) Set $j = (n + k)/2$.

(b) With probability j/n declare the last step to have been $+1$ and set $k \leftarrow k - 1$; otherwise declare -1 and set $k \leftarrow k + 1$.

(c) Set $n \leftarrow n - 1$ and repeat until $n = 0$.

This is exact. It produces a trajectory distributed exactly as a forward walk conditioned to end at k — I checked it by simulation against the uniform distribution over all 15 paths with $S_6 = 2$, and the deviations sit inside sampling noise.

The concept

A complete diffusion model, in nine lines of code and with nothing trained.

Forward: add noise, ending at a distribution you can sample. Reverse: a posterior over the last increment, applied repeatedly, walking the noise back into a trajectory. That is the whole architecture, and here every piece is a fraction.

Note 01's two-point model is the other exact toy in this course. That one is continuous and exact; this one is discrete and exact. Between them they show the skeleton twice, with no approximation anywhere.

9. Starting from data: the discrete denoiser

So far $S_0 = 0$ always, which is the honest weakness of the model: there is no data distribution, only the noise process. Fix that. Let the walk start at a data point,

$$x_0 \in \{-1, +1\} \ \text{ with probability } \tfrac{1}{2} \text{ each}, \qquad S_n = x_0 + \sum_{i=1}^{n} X_i,$$

which is exactly the two-point dataset of note 01 with a coin-toss forward process in place of a Gaussian one. Now ask the real question: *given the noised state, what was the data?*

By Bayes, with the prior cancelling,

$$\mathbb{P}(x_0 = +1 \mid S_n = k) = \frac{\mathbb{P}(\sum X_i = k - 1)}{\mathbb{P}(\sum X_i = k - 1) + \mathbb{P}(\sum X_i = k + 1)},$$

and so, writing $B(m) = \mathbb{P}(\sum_{i=1}^{n} X_i = m)$ for the binomial law of §3,

$$\boxed{\mathbb{E}\big[x_0 \mid S_n = k\big] = \frac{B(k-1) - B(k+1)}{B(k-1) + B(k+1)}.}$$

A ratio of binomial coefficients — exact, and computable. Now compare it with the denoiser Practice 1 §7.3 derived for Gaussian noise, $\tanh(x_\sigma/\sigma^2)$, at the matching noise level $\sigma^2 = n$:

k (with $n = 64$)	exact	$\tanh(k/n)$
1	0.015385	0.015624
9	0.138462	0.139705
17	0.261538	0.259549
33	0.507692	0.474316
65	1.000000	0.768079

The concept

Two lessons, and the second is the better one.

They agree in the bulk. The coin-toss denoiser and the Gaussian denoiser are the same function where the probability mass is — which is §6.1 again: after enough steps it does not matter that the noise was made of coins.

They disagree in the tail, and the exact answer is the right one. At $k = 65$ with $n = 64$ the walk is beyond anywhere it could have reached from $x_0 = -1$, so the data *must* have been $+1$ and the exact denoiser returns 1 with certainty. The Gaussian approximation, which puts mass everywhere, hedges at 0.768.

The approximation is excellent where you are and wrong where you are not. That is worth saying out loud in a course that will spend Modules 3–6 relying on Gaussian approximations.

10. Where learning has to enter

Everything above was exact. It is worth being precise about which step stops being exact when the data stops being a coin, because that step is the entire subject of the rest of the course.

The reverse process needs $\mathbb{P}(x_0 \mid S_n)$. In §9 we could write it down because x_0 had *two* possible values and we could enumerate them. Replace the two-point dataset with images and three things break at once:

- there is no formula for the prior $p(x_0)$ — note 06 is a catalogue of failed attempts to get one;
- the sum over possible x_0 becomes an integral over $\mathbb{R}^d$ that cannot be done;
- and by §6.3 the relevant region is a thin shell in a huge space, so no amount of sampling will cover it.

So the posterior mean $\mathbb{E}[x_0 \mid x_t]$ is not computable. But Practice 1 §7.1 says it is exactly what squared-error regression converges to — and a regression target *is* available, because the forward process is known and we can noise a real image ourselves.

The concept

That is the whole idea of the course, and this sheet has now assembled all of it: a forward process that is free to run and has a closed-form marginal (§3); a reverse process that is a posterior over what was added (§7); and a training objective that reaches that posterior without ever writing it down (Practice 1 §7.1).

Modules 3 to 6 replace the coin with a Gaussian and the enumeration with a network. Nothing else changes.

Closing: what this sheet installed

Result	Section	Where it returns
$\mathrm{Var}(S_n) = n$, $\mathrm{sd} = \sqrt{n}$	§2	the $\sqrt{t}$ envelope, W01 demo
$\mathbb{P}(S_n = k) = \binom{n}{(n+k)/2} 2^{-n}$	§3	closed-form $q(x_t \mid x_0)$
$p_{n+1} = \frac{1}{2}p_n(k-1) + \frac{1}{2}p_n(k+1)$	§4	the master equation of any chain
$\partial_t p = \frac{1}{2}\partial_x^2 p$	§4	Fokker–Planck, Module 5
$h = \sqrt{\Delta t}$	§5	why dW has size $\sqrt{dt}$
$S_n/\sqrt{n} \Rightarrow \mathcal{N}(0,1)$	§6.1	why the forward endpoint is Gaussian
$\|S_n\| \approx \sqrt{nd}$, thin shell	§6.3	why $\mathcal{N}(0, I_d)$ is a sphere
$\mathbb{E}[X_i \mid S_n = k] = k/n$	§7	the reverse process; the denoiser
Reverse sampling by peeling	§8	the sampling loop, Module 3
$\mathbb{E}[x_0 \mid S_n] \to \tanh(k/n)$	§9	note 01, and Practice 1 §7.3

The concept

The one sentence. A diffusion model is a random walk you know how to run forwards and have to learn to run backwards — and on a coin, the backwards step is k/n.

Exercises

(a) Compute $\mathbb{P}(S_6 = k)$ for every k and check the values sum to 1. Which k are unreachable, and why?

(b) Derive $\mathrm{Var}(S_n) = n$ twice: once from "variances add", and once from the binomial law of §3 using $S_n = 2K_n - n$. *(For the second you will need* $\mathrm{Var}(K_n) = n/4$*.)*

(c) Verify the master equation at $n = 3 \to 4$, $k = 0$ and $k = 2$, using the explicit numbers of §3.

(d) A walker starts at 0. Estimate how many steps until it is typically 100 away. Now do the same for 1000 away, and comment on the ratio.

(e) Show $\mathbb{E}[X_i \mid S_n = k] = k/n$ for *every* i, not just $i = n$, by the exchangeability argument. Why does this not contradict the increments being independent *unconditionally*?

(f) **The biased coin.** Let $\mathbb{P}(X_i = +1) = p$. Show $\mathbb{E}[S_n] = n(2p-1)$ and $\mathrm{Var}(S_n) = 4np(1-p)$. Then show that $\mathbb{P}(X_n = +1 \mid S_n = k)$ is *still* $(n+k)/2n$, independent of p. Explain in one sentence why the bias vanishes from the reverse step.

(g) In §5, take $h = \Delta t$ and show the limiting variance is 0. What does the trajectory look like? Then take $h = (\Delta t)^{1/4}$ and say what goes wrong.

(h) Implement the §8 reverse sampler. Starting from $S_{20} = 4$, generate 10,000 trajectories and check that the fraction of paths whose *first* step is $+1$ matches $\mathbb{E}[X_1 \mid S_{20} = 4]$ from §7.2.

(i) Compute $\mathbb{E}[x_0 \mid S_n = k]$ exactly for $n = 8$ and every reachable k, and plot it against $\tanh(k/8)$. Where is the agreement worst, and why is the *exact* curve the correct one there?

(j) For d independent walks, compute $\mathbb{E}\|S_n\|^2$ exactly. Then, using the CLT on each coordinate, explain why the relative width of $\|S_n\|$ shrinks like $1/\sqrt{d}$. What does this say about where the reverse process of a diffusion model begins?

(k) **Connecting the two sheets.** Practice 1 §6.3 gave $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon$. Which choice of α_t turns that update into the random walk of this sheet? What does that tell you about the difference between variance-preserving and variance-exploding processes?