Complex Preferences for Different Convergent Priors in Discrete Graph Diffusion

Anonymous Author(s)
Affiliation
Address
email

Abstract

Diffusion models have achieved state-of-the-art performance in generating many different kinds of data, including images, text, and videos. Despite their success, there has been limited research on how the underlying diffusion process and the final convergent prior can affect generative performance; this research has also been limited to continuous data types and a score-based diffusion framework. To fill this gap, we explore how different *discrete* diffusion kernels (which converge to different prior distributions) affect the performance of diffusion models for graphs. To this end, we developed a novel formulation of a *family* of discrete diffusion kernels which are easily adjustable to converge to different Bernoulli priors, and we study the effect of these different kernels on generative performance. We show that the quality of generated graphs is sensitive to the prior used, and that the optimal choice cannot be explained by obvious statistics or metrics, which challenges the intuitions which previous works have suggested.

Submitted to ICML 2023 Workshop: Sampling and Optimization in Discrete Space. Do not distribute.

1 Introduction

In recent years, diffusion models have been applied successfully to many different problems and data types, achieving state-of-the-art generation quality Sohl-Dickstein et al. (2015); Ho et al. (2020); Song et al. (2021); Dhariwal & Nichol (2021); Rombach et al. (2021). Despite how central the underlying diffusion process is to a diffusion model, however, there has been very limited research that explores how different diffusion processes affect generative performance. A few works have found that performance *can* be affected by the diffusion process, but these findings have largely been limited to diffusion on continuous objects, and where diffusion time is also continuous (i.e. using a stochastic-differential-equation framework) Song et al. (2021); Dockhorn et al. (2021); Karras et al. (2022). In contrast, *discrete diffusion models* Austin et al. (2021) have recently emerged as a more effective way to model intrinsically discrete objects, such as graphs Vignac et al. (2022); Tseng et al. (2023), but the impact of different design choices in this setting has received little to no attention.

In this work, we explicitly explore how the underlying diffusion process may affect generative performance in a *discrete-time* and *discrete-object* setting. In particular, we will focus on generating *undirected graphs*, as they are simply represented, yet arguably one of the most versatile and expressive discrete data types (i.e. many problems can be phrased as graph problems).

The space of possible discrete diffusion kernels is large. To simplify our analysis, we formulate a family of diffusion kernels based on the Bernoulli distribution, where only the noise schedule is a free parameter. We will show that adjusting the noise schedule induces a convergent prior distribution which—on graphs—is an Erdös–Renyi graph with any arbitrary edge probability p.

A few recent works have suggested intuitions for selecting the best convergent prior in a diffusion model. On continuous data types, Lee et al. (2021) achieved better performance on generating audio tracks with a diffusion prior which is a Gaussian with covariance equal to that of the original data distribution. For discrete diffusion on graphs, Vignac et al. (2022) proposed that the optimal prior should have the probability of each edge state (e.g. present or absent) match the empirical distribution in the original data. Together, these works have strongly suggested that different generation tasks merit the use of different diffusion priors, and they have intimated that the optimal prior is one whose core statistic (e.g. Gaussian covariance, multinomial probabilities, etc.) matches that of the original distribution Lee et al. (2021); Vignac et al. (2022). We call this the *empirical prior*. Importantly, although these works propose that the empirical prior is optimal, their results merely suggest that the empirical prior outperforms a uniform prior (e.g. isotropic Gaussian or uniform probabilities).

To our knowledge, this is the first work which *systematically* explores how modifying the convergent prior directly affects generative performance in *discrete* diffusion. Our results will challenge previous intuitions of what the optimal prior is. In particular, we highlight the following contributions:

- We derive a novel family of discrete diffusion kernels based on asymmetric Bernoulli processes, that is easily adjustable so it converges to an arbitrary Erdös–Renyi prior.
- We demonstrate that different graph-generation tasks achieve optimal generative performance on diffusion kernels which converge to different priors.
- We show that the optimal prior for a given task is *not* simply given by the empirical prior (i.e. based on statistics of the original data distribution) as previous works have suggested.

2 An Adjustable, Aasymmetric Bernoulli Kernel

Consider a bit x_t. At each time t, the diffusion process will flip the bit with probability according to some noise schedule. Tseng et al. (2023) proposed three such diffusion kernels, in which the final prior was a Bernoulli distribution of $\pi(x = 1) = 0, \pi(x = 1) = 1$, or $\pi(x = 1) = 0.5$. We extend from Tseng et al. (2023) by defining *two* (potentially asymmetric) noise schedules: $\{\beta_t^0, \beta_t^1\}$ for $t \in \{1, \cdots, T\}$. At time t, the bit x_{t-1} is flipped to a 0 with probability β_t^0 (if $x_{t-1} = 1$), and is flipped to a 1 with probability β_t^1 (if $x_t = 0$). By defining these two distinct noise schedules, the final prior probability can be anything between 0 and 1. We will generally assume that $\beta_t^b \in [0, \frac{1}{2}]$.

We can derive the following forward-diffusion probability:

$$q(x_t = 1|x_0) = \frac{1 + (-1)^{t-1}}{2} + \sum_{i=1}^{t} \left[\frac{(-1)^i}{2} \epsilon_i^{\frac{1+(-1)^i}{2}} \prod_{j=i+1}^{t} \frac{\bar{\epsilon_j}}{2} \right] + x_0 \prod_{j=1}^{t} \frac{\bar{\epsilon_j}}{2} \tag{1}$$

where $\epsilon_t^b = 2(1 - \beta_t^b)$ and $\bar{\epsilon}_t = \epsilon_t^0 + \epsilon_t^1 - 2 = 2(1 - \beta_t^0 - \beta_t^1)$.

If $\lim_{t \to T} \beta_t^0 = p_0$ and $\lim_{t \to T} \beta_t^1 = p_1$ asymptotically, then the prior distribution is $q(x_T = 1) = \pi(x = 1) = \frac{p_0}{p_0 + p_1}$. Thus, simply modifying the asymmetric noise schedules causes the diffusion process to converge to a Bernoulli distribution of any probability in the range $[0, 1]$ (Supplementary Figure S1). A full derivation of the kernel family is in Appendix B.

In our work, we diffuse on graphs by treating the edges as binary states—either an edge exists or it does not. That is, for a graph of n nodes, we diffuse over $\binom{n}{2}$ binary variables. We consider unlabeled nodes. Our adjustable Bernoulli kernel induces an Erdös–Renyi prior with probability $p = \frac{p_0}{p_0 + p_1}$.

3 Generative Performance Depends on the Prior

We consider two well-known benchmark graph datasets: community (small) and stochastic block models. For each dataset, we trained discrete diffusion models using the adjustable Bernoulli kernel introduced in Section 2, exploring an extended range of prior probabilities corresponding to Erdös–Renyi graphs with p in $\{0, 0.05, 0.10, ..., 0.95, 1\}$.

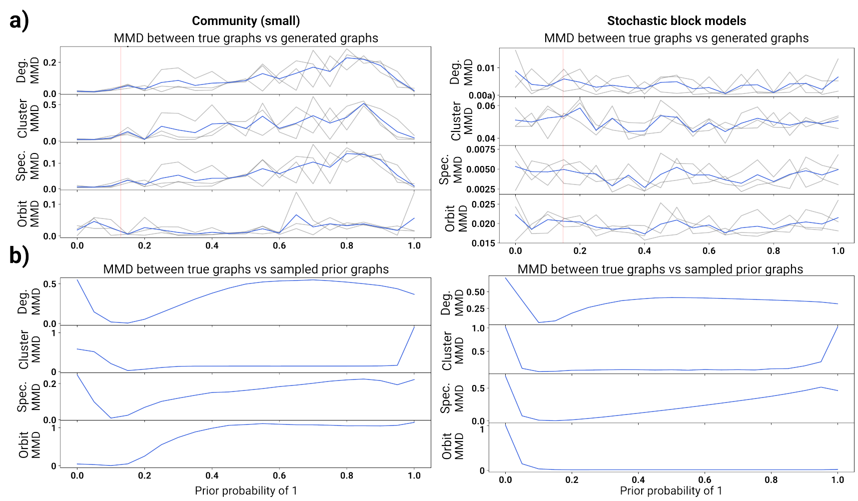

Figure 1: **a)** MMD of several graph distributions for our datasets, as a function of the prior probability in the diffusion kernel (the prior probability ranges from 0 to 1). A lower MMD is better. Three different random initializations are plotted in gray, and the average is in blue. The vertical red line marks the empirical probability of an edge in the original dataset. **b)** MMD between randomly sampled graphs from the prior distribution and the true data distribution, as a function of the prior probability.

For each model, we quantified the generative performance by computing the maximum mean discrepancy (MMD) for several graph distributions, following previous works in the space of graph generation You et al. (2018); Liao et al. (2019); Cao & Kipf (2018); Martinkus et al. (2022); Vignac et al. (2022). This performance metric compares several distributions of various statistics over the generated and true graphs (i.e. distribution of node degrees, clustering coefficients, spectrum of the normalized Laplacian, and node orbit counts). Averaging over several random initializations, the MMD values show a clear *preference* for which diffusion kernels—which vary in the convergent prior probability—yield the best performance overall (Figure 1a). This preference is consistent regardless of which graph statistic MMD is computed on. Furthermore, the best kernel is *different* between our datasets, and critically, the optimal kernel does not converge to the prior probability which matches the empirical probability in the original dataset. That is, the empirical prior is not necessarily optimal in our experiments. We also found that the generative performance of the optimal kernel yields better performance than previous graph-generative methods, including other discrete diffusion models (Supplementary Table S1).

It may seem intuitive to believe that the optimal prior should have a final edge probability that matches the empirical probability in the dataset (e.g. if the original dataset has a probability p of having an edge, it may seem that the optimal diffusion kernel should also converge to a probability of p). In this regard, Lee et al. (2021) showed that a prior which matches the empirical data in covariance (in continuous Gaussian diffusion) could be learned by a simpler neural network (thus leading to a more efficient training). These same intuitions were in Vignac et al. (2022), which showed some limited results for graphs suggesting that a diffusion kernel which converges to the empirical edge probability might have some moderate benefits over a uniform prior. Our experiments further extend these intuitions, and show that although the empirical prior may outperform the uniform prior, the optimal prior (at least for discrete graph diffusion) is not always the empirical prior.

3.1 The Optimal Kernel is not Explained by Empirical MMD

As the optimal diffusion kernel is not explained by the empirical distribution's edge probability, one may ask whether the kernel which yields the optimal MMD of generated graphs is the one whose prior distribution also has the optimal MMD (i.e. the one whose prior distribution matches the data distribution closest using MMD). In order to explore this, we sampled graphs from the prior distribution of each kernel, and computed the MMD between these randomly sampled graphs with the true data distribution.

Although we found that there was a trend in the convergent prior probability and the MMD between the prior distribution and the original data distribution (Figure 1b), this optimum was *not* the same as the optimal prior which maximizes generative performance (i.e. minimizes the MMD between the *generated graphs* and the original data distribution). This optimum, however, *does* match the empirical edge probability in the original data distribution.

4 Searching for the Optimal Kernel in Practice

Our results show that for discrete graph diffusion, the choice of diffusion prior can have large effects on the final generative performance. Additionally, the optimal prior is not simply the one which statistically matches the empirical data, or which maximizes similarity with the original data when measured by the MMD performance metric. Thus, we propose treating the diffusion kernel as a hyperparameter. In order to identify the optimal diffusion kernel, one may fix a family of diffusion kernels (e.g. Gaussian, or asymmetric Bernoulli as presented in Section 2, etc.) and search over it.

In order to aid in the efficient search for the optimal kernel, we found that by training only for a short time, the average training loss in the first few epochs is already somewhat predictive of the optimal kernel. That is, early training loss is correlated with final generative performance across different diffusion kernels in a family (Supplementary Figure S2). Furthermore, at least within the asymmetric-Bernoulli kernel family, we showed that the performance varies smoothly with the prior's probability of an edge (Figure 1). This property is expected in other families of diffusion kernels (e.g. Gaussian kernels in continuous diffusion) and enables search through efficient hyper-optimization techniques, such as Bayesian optimization.

5 Discussion

In this work, we developed a family of diffusion kernels based on the Bernoulli distribution which is easily modified to tune the final prior probability of an edge. We demonstrated that the generative performance of a graph-generation task depends on the specific diffusion prior, and that the optimal kernel is different for different tasks. Critically, we showed how the optimal kernel is not defined by a prior whose underlying probability distribution is the same as the empirical probability distribution of the original data, as prior works have intuited. Instead, we suggested that the optimal kernel may be treated as a hyperparameter and tuned for, which can be done relatively efficiently.

Although the optimal kernel/prior was not obviously informed by the empirical data, our exploration paves the way for more research toward designing optimal priors for discrete diffusion models. Future work may explore potentially more inscrutable relationships which may explain the optimal kernel, as this remains an open problem in both discrete and continuous diffusion.

References

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and Berg, R. V. D. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 22:17981–17993, 7 2021. ISSN 10495258. doi: 10.48550/arxiv.2107.03006. URL `https://arxiv.org/abs/2107.03006v2`.

Cao, N. D. and Kipf, T. Molgan: An implicit generative model for small molecular graphs. 5 2018. doi: 10.48550/arxiv.1805.11973. URL `https://arxiv.org/abs/1805.11973v2`.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 11:8780–8794, 5 2021. ISSN 10495258. doi: 10.48550/arxiv.2105.05233. URL `https://arxiv.org/abs/2105.05233v4`.

Dockhorn, T., Vahdat, A., and Kreis, K. Score-based generative modeling with critically-damped langevin diffusion. 12 2021. doi: 10.48550/arxiv.2112.07068. URL `https://arxiv.org/abs/2112.07068v4`.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020-December, 6 2020. ISSN 10495258. doi: 10.48550/arxiv.2006.11239. URL `https://arxiv.org/abs/2006.11239v2`.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. 6 2022. doi: 10.48550/arxiv.2206.00364. URL `https://arxiv.org/abs/2206.00364v1`.

Lee, S.-G., Kim, H., Shin, C., Tan, X., Liu, C., Meng, Q., Qin, T., Chen, W., Yoon, S., and Liu, T.-Y. Priorgrad: Improving conditional denoising diffusion models with data-dependent adaptive prior. 6 2021. URL `https://arxiv.org/abs/2106.06406v2`.

Liao, R., Li, Y., Song, Y., Wang, S., Hamilton, W. L., Duvenaud, D., Urtasun, R., and Zemel, R. Efficient graph generation with graph recurrent attention networks. *Advances in Neural Information Processing Systems*, 32, 10 2019. ISSN 10495258. doi: 10.48550/arxiv.1910.00760. URL `https://arxiv.org/abs/1910.00760v3`.

Martinkus, K., Loukas, A., Perraudin, N., and Wattenhofer, R. Spectre : Spectral conditioning helps to overcome the expressivity limits of one-shot graph generators. 4 2022. doi: 10.48550/arxiv.2204.01613. URL `https://arxiv.org/abs/2204.01613v1`.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. 12 2021. doi: 10.48550/arxiv.2112.10752. URL `https://arxiv.org/abs/2112.10752v2`.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. *32nd International Conference on Machine Learning, ICML 2015*, 3:2246–2255, 3 2015. doi: 10.48550/arxiv.1503.03585. URL `https://arxiv.org/abs/1503.03585v8`.

Song, Y., Sohl-Dickstein, J., Brain, G., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. 2021.

Tseng, A. M., Diamant, N., Biancalani, T., and Scalia, G. Graphguide: interpretable and controllable conditional graph generation with discrete bernoulli diffusion. 2 2023. URL `https://arxiv.org/abs/2302.03790v1`.

Vignac, C., Krawczuk, I., Siraudin, A., Wang, B., Cevher, V., and Frossard, P. Digress: Discrete denoising diffusion for graph generation. 9 2022. doi: 10.48550/arxiv.2209.14734. URL `https://arxiv.org/abs/2209.14734v1`.

You, J., Ying, R., Ren, X., Hamilton, W. L., and Leskovec, J. Graphrnn: Generating realistic graphs with deep auto-regressive models. *35th International Conference on Machine Learning, ICML 2018*, 13:9072–9081, 2 2018. doi: 10.48550/arxiv.1802.08773. URL `https://arxiv.org/abs/1802.08773v3`.

A Supplementary Figures and Tables

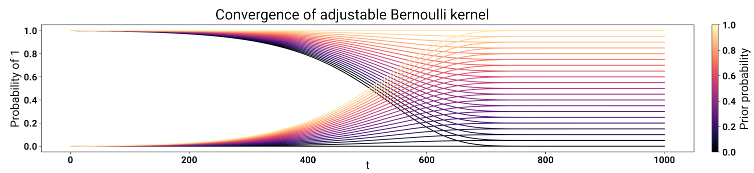

Figure S1: Visualization of the diffusion process of the adjustable Bernoulli kernel, for several different noise schedules. There are two lines of each color, showing the probability of a bit being 1 at each time t, if the original bit started at 0 or 1. Each color is a different asymmetric noise schedule, and the final probability converges to a prior defined by the asymptotic behavior of the noise schedules.

Table S1: MMD ratio

Model	Community (small)			Stochastic block models		
	Deg. ↓	Clus. ↓	Orbit ↓	Deg. ↓	Clus. ↓	Orbit ↓
GraphRNN	2.00	1.31	2.00	2.62	1.33	1.75
GRAN	1.73	1.25	1.00	3.76	1.29	1.46
MolGAN	1.73	1.36	1.00	5.42	1.87	1.67
SPECTRE	1.00	1.73	1.00	3.14	1.26	**0.54**
DiGress	1.00	0.95	1.00	1.26	1.22	1.30
Optimal prior	**0.99**	**0.57**	**0.79**	**0.56**	**1.18**	0.83

Comparison of generative performance of optimal prior to other works

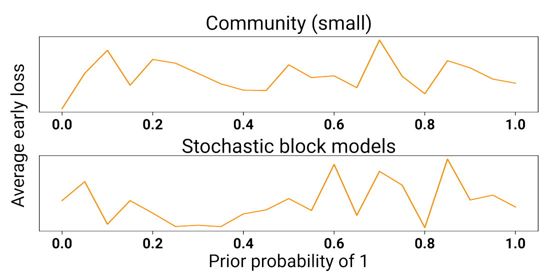

Figure S2: Average value of the loss for the first 10 epochs of training, for each diffusion kernel on each task.

B Derivation of asymmetric Bernoulli kernel

B.1 Forward diffusion distribution

Here, we derive the forward distribution $q_t(x_t|x_{t-1}, x_0)$. Note that every x is a single bit.

Let us define a noising process $\{\beta_t^0, \beta_t^1\}$ for $t \in \{1, \cdots, T\}$. In particular, we have $q(x_t = 1|x_{t-1} = 0) = \beta_t^0$ and $q(x_t = 0|x_{t-1} = 1) = \beta_t^1$.

We will generally assume that $\beta_t^b \in [0, \frac{1}{2}]$.

In our derivation, we will use the following changes of variables to assist in simplification:

$\beta_t^b = 1 - \frac{1}{2}\epsilon_t^b$ (or equivalently, $\epsilon_t^b = 2(1 - \beta_t^b)$)

$\bar{\epsilon}_t = \epsilon_t^0 + \epsilon_t^1 - 2 = 2(1 - \beta_t^0 - \beta_t^1)$

Below are the forward-distribution probabilities for the first four time steps:

$P(x_1 = 1|x_0) = \frac{1}{2}(2 - \epsilon_1^0 + x_0\bar{\epsilon}_1)$

$P(x_2 = 1|x_0) = \frac{1}{4}(2\epsilon_2^1 - \epsilon_1^0\bar{\epsilon}_2 + x_0\bar{\epsilon}_1\bar{\epsilon}_2)$

$P(x_3 = 1|x_0) = \frac{1}{8}(8 - 4\epsilon_3^0 + 2\epsilon_2^1\bar{\epsilon}_3 - \epsilon_1^0\bar{\epsilon}_2\bar{\epsilon}_3 + x_0\bar{\epsilon}_1\bar{\epsilon}_2\bar{\epsilon}_3)$

$P(x_4 = 1|x_0) = \frac{1}{16}(8\epsilon_4^1 - 4\epsilon_3^0\bar{\epsilon}_4 + 2\epsilon_2^1\bar{\epsilon}_3\bar{\epsilon}_4 - \epsilon_1^0\bar{\epsilon}_2\bar{\epsilon}_3\bar{\epsilon}_4 + x_0\bar{\epsilon}_1\bar{\epsilon}_2\bar{\epsilon}_3\bar{\epsilon}_4)$

Or in general:

$$P(x_t = 1|x_0) = \frac{1}{2^t}(2^t(\frac{1+(-1)^{t-1}}{2}) + \sum_{i=1}^{t}[(-1)^i 2^{i-1}\epsilon_i^{\frac{1+(-1)^i}{2}} \prod_{j=i+1}^{t}\bar{\epsilon}_j] + x_0\prod_{j=1}^{t}\bar{\epsilon}_j)$$

In a more numerically stable form:

$$P(x_t = 1|x_0) = \frac{1+(-1)^{t-1}}{2} + \sum_{i=1}^{t}[\frac{(-1)^i}{2}\epsilon_i^{\frac{1+(-1)^i}{2}} \prod_{j=i+1}^{t}\frac{\bar{\epsilon}_j}{2}] + x_0\prod_{j=1}^{t}\frac{\bar{\epsilon}_j}{2}$$

B.2 Prior distribution

By changing the value that β_t^0, β_t^1 converge to, the prior can be made to be any probability between 0 and 1.

Now let us try and derive the prior probability more formally.

First, let us make the assumption that T is even.

From above, we have that $P(x_T = 1|x_0) = x_0\bar{\epsilon}_1 \cdots \bar{\epsilon}_T - \frac{1}{2^T}\epsilon_1^0\bar{\epsilon}_2 \cdots \bar{\epsilon}_T + \frac{1}{2^{T-1}}\epsilon_2^1\bar{\epsilon}_3 \cdots \bar{\epsilon}_T - \frac{1}{2^{T-2}}\epsilon_3^0\bar{\epsilon}_4 \cdots \bar{\epsilon}_T + \cdots + \frac{1}{2}\epsilon_T^1$.

Early terms in this sequence consist of many $\bar{\epsilon}_i$ being multiplied together. For large T, these terms contribute an infinitesimal amount to the total sum. Thus, we can consider only the *end behavior* of ϵ_t^b. We make the simplifying assumption that β_t^0, β_t^1 both approach some maximum value asymptotically, so the end behaviors of β_t^0, β_t^1 are constant. This allows us to make the following substitutions for all times t (as early times will contribute nothing to the final probability):

$\beta_t^0 := p_0, \beta_t^1 := p_1, \epsilon_t^0 := q_0 = 2(1 - p_0), \epsilon_t^1 := q_1 = 2(1 - p_1), \bar{\epsilon}_t := s = 2(1 - p_0 - p_1)$ for all t

Then our expression becomes:

$P(x_T = 1|x_0) = -q_0\frac{1}{2^T}s^{T-2+1} + q_1\frac{1}{2^{T-1}}s^{T-3+1} - q_0\frac{1}{2^{T-2}}s^{T-4+1} + \cdots + q_1\frac{1}{2}$

We rearrange the terms by those with q_0 and those with q_1, and factor out q_0 and q_1 to obtain:

$P(x_T = 1|x_0) = -q_0\frac{s}{2^2}(1 + \frac{s^2}{2^2} + \frac{s^4}{2^4} + \cdots + \frac{s^{T-2}}{2^{T-2}}) + q_1\frac{1}{2}(1 + \frac{s^2}{2^2} + \frac{s^4}{2^4} + \cdots + \frac{s^{T-2}}{2^{T-2}})$

Now the series in the parentheses are geometric series. Recall, $\sum_{i=0}^{n} r^i = \frac{1-r^{n-1}}{1-r}$. Thus, we get:

$$P(x_T = 1|x_0) = -q_0\frac{s}{2^2}\sum_{i=0}^{\frac{T}{2}-1}((\frac{s}{2})^2)^i + q_1\frac{1}{2}\sum_{i=0}^{\frac{T}{2}-1}((\frac{s}{2})^2)^i = -q_0\frac{s}{2^2}\frac{1-(\frac{s}{2})^T}{1-(\frac{s}{2})^2} + q_1\frac{1}{2}\frac{1-(\frac{s}{2})^T}{1-(\frac{s}{2})^2}$$

Now note that $(\frac{s}{2})^T \to 0$, so we get:

$P(x_T = 1|x_0) = -q_0\frac{s}{2^2}\frac{1}{1-(\frac{s}{2})^2} + q_1\frac{1}{2}\frac{1}{1-(\frac{s}{2})^2}$

Substituting back our original assumptions, we get:

$P(x_T = 1|x_0) = \frac{p_0}{p_0+p_1}$

Now let us consider the case where T is odd.

From above, we have that $P(x_T = 1|x_0) = x_0\bar{\epsilon}_1 \cdots \bar{\epsilon}_T - \frac{1}{2^T}\epsilon_1^0\bar{\epsilon}_2 \cdots \bar{\epsilon}_T + \frac{1}{2^{T-1}}\epsilon_2^1\bar{\epsilon}_3 \cdots \bar{\epsilon}_T - \frac{1}{2^{T-2}}\epsilon_3^0\bar{\epsilon}_4 \cdots \bar{\epsilon}_T + \cdots - \frac{1}{2}\epsilon_T^0 + 1$

We use the same assumptions as above for even T, and we obtain the following:

$P(x_T = 1|x_0) = -q_0\frac{1}{2}(1 + \frac{s^2}{2^2} + \frac{s^4}{2^4} + \cdots + \frac{s^{T-1}}{2^{T-1}}) + q_1\frac{1}{s}(\frac{s^2}{2^2} + \frac{s^4}{2^4} + \cdots + \frac{s^{T-1}}{2^{T-1}}) + 1$

Using the summation of a geometric series again, we get that $1 + \frac{s^2}{2^2} + \frac{s^4}{2^4} + \cdots + \frac{s^{T-1}}{2^{T-1}} = \frac{1-(\frac{s}{2})^{T+1}}{1-(\frac{s}{2})^2}$.
Again, we can assume that $(\frac{s}{2})^{T+1} \to 0$.

Then $P(x_T = 1|x_0) = -q_0\frac{1}{2}(\frac{1}{1-(\frac{s}{2})^2}) + q_1\frac{1}{s}(\frac{1}{1-(\frac{s}{2})^2} - 1) + 1$

Substituting back our original assumptions, we get:

$P(x_T = 1|x_0) = \frac{p_0}{p_0+p_1}$ (the same as when T is even)

B.3 Posterior distribution

We use Bayes' Rule: $P(x_{t-1} = 1|x_t, x_0) = \frac{P(x_t|x_{t-1}=1,x_0)P(x_{t-1}=1|x_0)}{P(x_t|x_0)}$.

We analyze each piece separately:

$P(x_t|x_{t-1} = 1, x_0) = x_t(1 - \beta_t^1) + (1 - x_t)\beta_t^1$ (if $x_t = 1$, this is the event we don't flip from 1 to 0; if $x_t = 0$, this is the event we do flip from 1 to 0).

$P(x_{t-1} = 1|x_0)$ comes directly from Equation 1.

$P(x_t|x_0) = x_tP(x_t = 1|x_0) + (1 - x_t)(1 - P(x_t = 1|x_0))$, also from Equation 1.

This gives our posterior, $q_t(x_{t-1}|x_t, x_0)$.