Categorical SDEs with Simplex Diffusion

Pierre H. Richemond[*] **Sander Dieleman &** **Arnaud Doucet**[*]
Google DeepMind
{richemond, sedielem, arnauddoucet}@deepmind.com

Abstract

Diffusion models typically operate in the standard framework of generative modelling by producing continuously-valued datapoints. To this end, they rely on a progressive Gaussian smoothing of the original data distribution, which admits an SDE interpretation involving increments of a standard Brownian motion. However, some applications such as text generation or reinforcement learning might naturally be better served by diffusing categorical-valued data, i.e., lifting the diffusion to a space of probability distributions. To this end, this short theoretical note proposes *simplex diffusion*, a means to directly diffuse datapoints located on an n-dimensional probability simplex. We show how this relates to the Dirichlet distribution on the simplex and how the analogous SDE is realized thanks to a multi-dimensional *Cox–Ingersoll–Ross* process (abbreviated as CIR), previously used in economics and mathematical finance. Finally, we make remarks as to the numerical implementation of trajectories of the CIR process, and discuss some limitations of our approach.

1 Introduction and background

Diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021b) are a now well-established class of generative models that find applications notably in the image (Dhariwal & Nichol, 2021; Ramesh et al., 2022; Saharia et al., 2022), video (Singer et al., 2022; Villegas et al., 2022; Ho et al., 2022), speech (Jeong et al., 2021; Huang et al., 2022) domains, and even for molecule generation (Hoogeboom et al., 2022; Corso et al., 2022). These models proceed as follows. One adds noise progressively to data using a diffusion process to transform the complex data distribution to a simple easy-to-sample distribution. The generative model is obtained by simulating an approximation of the time-reversal of this process. The resulting "denoising" process is also a diffusion whose drift depends on the logarithmic gradients of the noised data densities (Anderson, 1982), i.e. the Stein *scores*. These scores are estimated using a neural network via score matching (Hyvärinen, 2005). In the usual case where Gaussian noise is progressively added to the generative distribution, the score matching objective simply reduces to a least squares denoising term (Vincent, 2011) easily amenable to gradient descent.

In all which precedes, the datapoints are assumed to be vectors taking continuous values. Being able to proceed with diffusion when those datapoints are instead discrete-valued would further widen the applicability domain of diffusion models, in particular to language modelling (Savinov et al., 2022; Li et al., 2022; Wang et al., 2022) and even reinforcement learning (Richemond & Maginnis, 2017; Janner et al., 2022). We propose a construction of such a discrete diffusion in this short technical note. Our approach consists in directly deriving a tractable stochastic process that operates on the probability simplex itself, lifting traditional diffusion schemes to categorical distributions, rather than relying on auxiliary methods such as binary encoding (Chen et al., 2022). Because of this, we can use simplex diffusion in conjunction with the now standard mathematical machinery

[*]Equal contribution

ICML 2023 Workshop: Sampling and Optimization in Discrete Space (SODS 2023).

of diffusion models, including equivalent ODE formulation, and computation of an evidence lower bound (ELBO). Finally, we also discuss some specific limitations of our approach, namely the issues once encounters in practice when simulating high-dimensional simplices (i.e., for large values of n).

2 Simplex diffusion with the Cox–Ingersoll–Ross process

We first proceed to recall how one can sample from the Dirichlet distribution on the probability simplex using independent Gamma random variables. Then, we introduce a compatible stochastic process, the Cox–Ingersoll–Ross process.

2.1 Dirichlet distribution on the simplex

For a given integer $n \geq 2$, the $n-1$ dimensions probability simplex is the set of n-dimensional vectors $\boldsymbol{X}$ in $\mathbb{R}^n$ whose components $\boldsymbol{X} := (X_1, ..., X_n)$ satisfy $X_i \geq 0$ and $\sum_{i=1}^n X_i = 1$. A point on the simplex is hence assimilated to an n-way categorical distribution.

The Dirichlet distribution is defined over the simplex as the conjugate prior of the categorical distribution. It is a multivariate, continuous distribution, parametrized by an arbitrary vector of strictly positive scalars $\boldsymbol{\alpha}$. The Dirichlet distribution $\mathcal{D}(\boldsymbol{\alpha})$ with parameters $\boldsymbol{\alpha} := (\alpha_1, ..., \alpha_n)$, where $\alpha_1, ..., \alpha_n > 0$, has probability density function $f_{\mathcal{D}}$ (w.r.t. the standard Lebesgue measure on $\mathbb{R}^n$)

$$f_{\mathcal{D}}(x_1, \cdots, x_n; \alpha_1, \cdots, \alpha_n) = \frac{1}{Z(\boldsymbol{\alpha})} \prod_{i=1}^n x_i^{\alpha_i - 1} \tag{1}$$

with $Z(\boldsymbol{\alpha})$ a normalizing constant. In particular, the choice $\alpha_i = 1$ for all i recovers a uniform distribution over the simplex. In our construction, the Dirichlet distribution plays a role somewhat analogous to that of the Gaussian distribution in standard diffusions - in that it represents the desired stationary distribution of the diffusion process we will build below. Hence, and given its flexibility we focus on it, although other choices of simplex distributions are possible (Aitchison, 1982).

Sampling. It is well known that sampling from the Dirichlet distribution reduces to a two-step procedure: first, sampling n *independent* Gamma random variables $Y_1 \sim \mathcal{G}(\alpha_1, \beta), ..., Y_n \sim \mathcal{G}(\alpha_n, \beta)$ where α_i is their shape parameter, and β their common rate parameter. Second, normalizing those random variables to sum to 1 then yields the Dirichlet-distributed random vector

$$\boldsymbol{X} = \left(\frac{Y_1}{\sum_{i=1}^n Y_i}, ..., \frac{Y_n}{\sum_{i=1}^n Y_i} \right) \sim \mathcal{D}(\boldsymbol{\alpha}). \tag{2}$$

This result holds for any $\beta > 0$ so we can use $\beta = 1$ specifically. Taking these observations together, we now seek to find an n-dimensional stochastic process whose marginal distributions each converge to Gamma laws in the large-time limit. We exhibit such a process below.

2.2 The Cox-Ingersoll-Ross process

The Cox–Ingersoll–Ross (or CIR) process introduced in Cox et al. (1985) is a popular real-valued diffusion process used in econometrics and quantitative finance, both for yield curve (usually, the instantaneous interest rate) and stochastic equity volatility (Heston, 1993) modelling. It is an instance of *square-root diffusion* defined by the following SDE in θ_t: for any $\theta_0 \geq 0$ and $a, b, \sigma > 0$

$$\mathrm{d}\theta_t = b(a - \theta_t)\mathrm{d}t + \sigma\sqrt{\theta_t}\mathrm{d}W_t, \tag{3}$$

where $(W_t)_{t \geq 0}$ is a standard Brownian motion (or Wiener process). The solution to this SDE exists and is unique (Watanabe & Yamada, 1971), despite the non-regularity of the square root term near zero. The CIR process is ergodic, almost surely non-negative and admits as invariant limiting distribution the Gamma distribution $\mathcal{G}(2ab/\sigma^2, 2b/\sigma^2)$. If $2ab \geq \sigma^2$ and $\theta_0 > 0$, then the process is strictly positive, pathwise. For our purpose, we can set $2b = \sigma^2$ so that (3) becomes

$$\mathrm{d}\theta_t = b(a - \theta_t)\mathrm{d}t + \sqrt{2b\theta_t}\mathrm{d}W_t \tag{4}$$

and admits the Gamma distribution $\mathcal{G}(a, 1)$ as limiting distribution.

Conditional mean and variance. One can readily check that for $t > 0$

$$\mathbb{E}[\theta_t | \theta_0] = \theta_0 \exp(-bt) + a(1 - \exp(-bt)) = a + \exp(-bt)(\theta_0 - 1) \tag{5}$$

$$\text{and} \qquad \begin{aligned}\text{var}[\theta_t|\theta_0] &= 2\theta_0(\exp(-bt) - \exp(-2bt)) + a(1 - \exp(-bt))^2 \\ &= 2\theta_0(\exp(-bt) - \exp(-2bt)) + a(1 + \exp(-2bt) - 2\exp(-bt)) \\ &= a + 2\exp(-bt)(\theta_0 - a) + \exp(-2t)(a - 2\theta_0). \end{aligned} \tag{6}$$

b can be thought of as the parameter governing diffusion speed. As $t \to \infty$, we have $\mathbb{E}[\theta_t|\theta_0] \to a$ and $\text{var}[\theta_t|\theta_0] \to a$. At any point in time, the drift term in equation 3 pushes θ_t back towards its long-term average a, a phenomenon known as *mean-reversion*. For this reason b is also indicative of, and sometimes called, the speed of mean-reversion.

Density of increments. The transition density of the CIR process is available in closed-form thanks to Laplace transform techniques (Feller, 1951) and can be sampled from exactly; i.e. we have

$$\theta_t|\theta_0 \sim \frac{1 - \exp(-bt)}{2} K, \qquad K \sim \chi^2\Big(2a, 2\theta_0 \frac{\exp(-bt)}{1 - \exp(-bt)}\Big), \tag{7}$$

where $\chi^2(\nu, \mu)$ denotes the non-central chi-squared distribution with ν degrees of freedom and non-centrality parameter μ. We can write this density explicitly as

$$f(\theta_t|\theta_0) = c\exp\Big(-c(\theta_0\exp(-bt) + \theta_t)\Big)\Big(\frac{\theta_t \exp(bt)}{\theta_0}\Big)^{\frac{a-1}{2}} I_{a-1}\Big(2c\sqrt{\theta_0\theta_t\exp(-bt)}\Big), \tag{8}$$

for $c = (1-\exp(-bt))^{-1}$ and I_{a-1} being the modified Bessel function of the first kind of order $a-1$. This closed-form expression for the transition density for the CIR model makes usual denoising score matching techniques applicable, as we'll see below.

2.3 Simplex diffusion

Simplex SDE. Our original purpose is to exhibit a diffusion whose marginal distribution, in the large time limit, provides samples from a Dirichlet distribution $\mathcal{D}(\boldsymbol{\alpha})$. It follows directly from previous section that this can be achieved by simulating first n independent CIR processes Y^i in parallel, resulting in a process $\boldsymbol{Y}_t$ with values in the positive orthant, following (with $\mathrm{d}W_t^i$ the independent increments of a standard n-dimensional Brownian motion $\boldsymbol{W}_t$, so that $\langle \mathrm{d}W_t^i, \mathrm{d}W_t^j \rangle = \delta_{i,j} t$) :

$$\mathrm{d}Y_t^i = b(\alpha_i - Y_t^i)\mathrm{d}t + \sqrt{2bY_t^i}\mathrm{d}W_t^i \tag{9}$$

each Y^i thus having limiting distribution $\mathcal{G}(\alpha_i, 1)$. We then consider the normalized, unit-sum vector

$$\boldsymbol{X}_t = \left(\frac{Y_t^1}{\sum_{i=1}^n Y_t^i}, ..., \frac{Y_t^n}{\sum_{i=1}^n Y_t^i}\right) \tag{10}$$

This normalization projects $\boldsymbol{Y}_t$ from the positive orthant to the probability simplex. By construction, we have $\boldsymbol{X}_t = (X_t^1, ..., X_t^n) \sim \mathcal{D}(\boldsymbol{\alpha})$ as $t \to \infty$. This is our main result, and enables us to perform diffusion towards a vertex of the simplex (a one-hot vector, representing the state of a categorical variable) in the time-reversal process.

Now since this multidimensional SDE retains a standard Brownian increment, both the time-reversal of the SDE, and reformulation as a standard ODE for 'probability flow'-type sampling (Song et al., 2021b) proceed as usual. We detail those aspects below.

Time reversal. The SDE in equation (9) is of the general (vector) form

$$\mathrm{d}\boldsymbol{Y}_t = \boldsymbol{f}_t(t, \boldsymbol{Y}_t)\mathrm{d}t + \boldsymbol{G}(t, \boldsymbol{Y}_t)\mathrm{d}\boldsymbol{W}_t, \tag{11}$$

where $\boldsymbol{f}(t, \boldsymbol{Y}_t) = b(\boldsymbol{\alpha} - \boldsymbol{Y}_t)$, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)^\top$ and $\boldsymbol{G}(t, \boldsymbol{Y}_t) = \sqrt{2b} \cdot \text{diag}(\sqrt{Y_t^1}, ..., \sqrt{Y_t^n})$. Let us also introduce further notation: $p_t = \text{Law}(\boldsymbol{Y}_t)$, the law of the probability density function of $\boldsymbol{Y}_t$, and $\boldsymbol{\Sigma}(t, \boldsymbol{Y}_t) = \boldsymbol{G}(t, \boldsymbol{Y}_t)\boldsymbol{G}(t, \boldsymbol{Y}_t)^\top = 2b \cdot \text{diag}(\boldsymbol{Y}_t)$.

The *time reversal* (Anderson, 1982; Haussmann & Pardoux, 1986) of the multidimensional CIR given by equation (9) is the process $(\boldsymbol{Z}_t)_{t\in[0,T]}$ such that $\boldsymbol{Z}_t = \boldsymbol{Y}_{T-t}$ satisfies

$$\mathrm{d}\boldsymbol{Z}_t = \big[-b(\boldsymbol{\alpha} - \boldsymbol{Z}_t) + 2b \cdot \text{diag}(\boldsymbol{Z}_t)\, \nabla_{\boldsymbol{Z}} \log p_{T-t}(\boldsymbol{Z}_t) + 2b\mathbf{1}\big]\mathrm{d}t + \boldsymbol{G}(T-t, \boldsymbol{Z}_t)\mathrm{d}\boldsymbol{W}_t \tag{12}$$

with $\boldsymbol{Z}_0 \sim p_T$. In practice, we will approximate this time reversal by the diffusion

$$\mathrm{d}\boldsymbol{Z}_t = \big[-b(\boldsymbol{\alpha} - \boldsymbol{Z}_t) + 2b \cdot \text{diag}(\boldsymbol{Z}_t)\, \boldsymbol{s}_{T-t}(\boldsymbol{Z}_t) + 2b\mathbf{1}\big]\mathrm{d}t + \boldsymbol{G}(T-t, \boldsymbol{Z}_t)\mathrm{d}\boldsymbol{W}_t, \tag{13}$$

with $\boldsymbol{Z}_0 \sim p_{\text{ref}}$ where $p_{\text{ref}}(z^1, ..., z^n) = \prod_{i=1}^{n} \mathcal{G}(z^i; \alpha_i, 1)$. Here $\boldsymbol{s}_t(\boldsymbol{x})$ is a neural score network approximating $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$.

We present both the max likelihood estimator and the ODE formulation for sampling in Appendix.

Remarks on numerical simulation. The CIR process has been extensively used and studied within Monte Carlo methods (Glasserman, 2004) in quantitative finance. Care must be taken in simulating its trajectories; this can typically require an additional scalar ϵ stabilization parameter inside of the square-root diffusion term in equation 9 in order to avoid path termination due to discretization error. Another avenue is to observe that under specific conditions on their parameters, the sum of independent, squared Ornstein–Uhlenbeck processes is identical in law to a CIR process (Jamshidian, 1995); this observation relates to Bessel processes (Revuz & Yor, 2013). This enables substituting a single CIR path for multiple Ornstein-Uhlenbeck paths, trading off stability for computation.

Limitations. We might want to use our approach on very high dimensional simplices in order to simulate *one-of-many* categoricals - for instance, when modelling language tokens over a sizeable vocabulary, or in the case of a large action-space policy. This comes with practical issues, chief amongst those being the potential presence of outliers in the categorical distribution. When we draw a sample from the transition density of the CIR process for a given t, we can determine the rank of the ground truth token in the resulting (unnormalized) vector. We observed in practice that the distribution of that rank - whose closed form law involves large, and possibly intractable integrals - is extremely heavy-tailed. Informally, this can lead to noisy results. We found empirically this phenomenon to be particularly relevant in high dimensions.

Finally, we note that while the interpretation of noisy vectors as unnormalized probability distributions via a Dirichlet prior is useful to build intuition, it is not rigorous. When one considers the posterior distribution at token level $p_t(\boldsymbol{x}_0|\boldsymbol{x}_t)$, where $\boldsymbol{x}_0$ is a one-hot vector representing a token, and $\boldsymbol{x}_t$ is the noisy unnormalized probability input vector, we can apply Bayes' rule and get

$$p_t(\boldsymbol{x}_0|\boldsymbol{x}_t) = \frac{p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)p(\boldsymbol{x}_0)}{\sum_{\boldsymbol{x}_0} p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)p(\boldsymbol{x}_0)} \tag{14}$$

thus showing that $p(\boldsymbol{x}_0|\boldsymbol{x}_t)$ is actually nonlinear in $\boldsymbol{x}_t$.

Related and alternative approaches. The Cox–Ingersoll–Ross process is seldom used in machine learning. Similar derivations to ours nonetheless previously appeared in Baker et al. (2018), where a CIR process is also used to approximate a Dirichlet distribution, but in a Bayesian inference context, with the very different purpose of obviating discretization error in stochastic gradient MCMC (Welling & Teh, 2011; Ma et al., 2015). Other stochastic processes than the CIR can be built that admit the Dirichlet distribution as a limiting distribution. Evans (2003) considers functions of the components of a multivariate Brownian motion running on a hypersphere. When those functions are all identically a squaring, by construction the squared components sum to 1 and can thus represent a categorical probability vector. In that setting the invariant distribution of the squared-components vector is proven to be symmetric Dirichlet with parameter $1/2$. Unlike ours, that approach is however not fully compatible with standard diffusion score matching, since the transition density of the Brownian motion on the sphere is to our knowledge not known in closed form - it is merely possible to sample from (Mijatovic et al., 2020). Other choices than a Dirichlet limiting distribution are also possible, even as it represents a reasonable and flexible prior family; Aitchison (1982) proposes a generic *log-ratio transform* projecting unconstrained, multivariate distributions defined on $\mathbb{R}^n$ onto the simplex. Separately, Lafferty & Lebanon (2005) perform an asymptotic expansion of the heat kernel on statistical manifolds (including an approximation of the simplex), with application to the multinomial family of distributions towards text classification.

3 Conclusion

We have introduced *simplex diffusion*, a simple method that uses a multi-dimensional Cox-Ingersoll-Ross process, via a unit-sum normalization of its time marginals, to diffuse categorical distributions directly on the probability simplex. Our approach is tractable and compatible with the tools of standard stochastic calculus central to diffusion models. Further research will involve operationalizing and evaluating deep learning models that leverage this principle.

References

John Aitchison. *The Statistical Analysis of Compositional Data*. Chapman and Hall, 1982.

Brian D O Anderson. Reverse-time diffusion equation models. *Stochastic Processes and Their Applications*, 12:313–326, 1982.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 2021.

Jack Baker, Paul Fearnhead, Emily Fox, and Christopher Nemeth. Large-scale stochastic sampling from the probability simplex. *Advances in Neural Information Processing Systems*, 2018.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and A. Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 2022.

Ting Chen, Ruixiang Zhang, and Geoffrey E. Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *ArXiv*, abs/2208.04202, 2022.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and T. Jaakkola. DiffDock: Diffusion steps, twists, and turns for molecular docking. *ArXiv*, abs/2210.01776, 2022.

John C Cox, Jonathan E Ingersoll Jr, and Stephen A Ross. A theory of the term structure of interest rates. *Econometrica*, 2:385–407, 1985.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 2021.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H. Richemond, A. Doucet, Robin Strudel, Chris Dyer, Conor Durkan, Curtis Hawthorne, Rémi Leblond, Will Grathwohl, and Jonas Adler. Continuous diffusion for categorical data. *ArXiv*, abs/2211.15089, 2022.

Steven N Evans. Diffusions on the simplex from Brownian motions on hypersurfaces. *Lecture Notes - Monograph Series - Statistics and Science: A Festschrift for Terry Speed*, pp. 35–48, 2003.

William Feller. Two singular diffusion problems. *Annals of Mathematics*, 54:173, 1951.

Paul Glasserman. *Monte Carlo Methods in Financial Engineering*, volume 53. Springer, 2004.

Ulrich G Haussmann and Etienne Pardoux. Time reversal of diffusions. *The Annals of Probability*, pp. 1188–1205, 1986.

Steven Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6:327–343, 1993.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *ArXiv*, abs/2210.02303, 2022.

Emiel Hoogeboom, Victor Garcia Satorras, Clement Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *ICML*, 2022.

Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. ProDiff: Progressive fast diffusion model for high-quality text-to-speech. *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *The Journal of Machine Learning Research*, 6:695–709, December 2005. ISSN 1532-4435.

Farshid Jamshidian. A simple class of square-root interest-rate models. *Applied Mathematical Finance*, 2:61–72, 1995.

Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *ICML*, 2022.

Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim. Diff-TTS: A denoising diffusion model for text-to-speech. In *Interspeech*, 2021.

John D. Lafferty and Guy Lebanon. Diffusion kernels on statistical manifolds. *J. Mach. Learn. Res.*, 6:129–163, 2005.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. Diffusion-LM improves controllable text generation. *ArXiv*, abs/2205.14217, 2022.

Yian Ma, Yi-An Ma, Tianqi Chen, and Emily B. Fox. A complete recipe for stochastic gradient MCMC. *Advances in Neural Information Processing Systems*, 2015.

Aleksandar Mijatovic, Veno Mramor, and Gerónimo Uribe Bravo. An algorithm for simulating Brownian increments on a sphere. *Journal of Physics A: Mathematical and Theoretical*, 54, 2020.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *ArXiv*, abs/2204.06125, 2022.

Daniel Revuz and Marc Yor. *Continuous Martingales and Brownian Motion*, volume 293. Springer Science & Business Media, 2013.

Pierre H. Richemond and Brendan Maginnis. On Wasserstein reinforcement learning and the Fokker-Planck equation. *ArXiv*, abs/1712.07185, 2017.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022.

Nikolay Savinov, Junyoung Chung, Mikolaj Binkowski, Erich Elsen, and Aäron van den Oord. Step-unrolled denoising autoencoders for text generation. *International Conference on Learning Representations*, 2022.

Uriel Singer, Adam Polyak, Thomas Hayes, Xiaoyue Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. *ArXiv*, abs/2209.14792, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 2019.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Haoran Sun, Lijun Yu, Bo Dai, Dale Schuurmans, and Hanjun Dai. Score-based continuous-time discrete diffusion models. *ArXiv*, abs/2211.16750, 2022.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and D. Erhan. Phenaki: Variable length video generation from open domain textual description. *ArXiv*, abs/2210.02399, 2022.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Rose E. Wang, Esin Durmus, Noah D. Goodman, and Tatsunori Hashimoto. Language modeling via stochastic processes. *ArXiv*, abs/2203.11370, 2022.

Shinzo Watanabe and Toshio Yamada. On the uniqueness of solutions of stochastic differential equations II. *Journal of Mathematics of Kyoto University*, 11:553–563, 1971.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. *International Conference on Machine Learning*, 2011.

Appendices

Here we include derivations both for max likelihood training and the ODE formulation for sampling.

Max likelihood training. This form also lends itself to computation of an evidence lower bound. Maximizing the likelihood of the data is equivalent to minimizing the KL divergence between the terminal time marginal induced by our SDE and the data distribution, which we compute exactly as in (Song et al., 2021a, Section 4). Let $\mathcal{P}$ and $\hat{\mathcal{P}}$ the path measures corresponding respectively to equations (12) and (13). Then by Girsanov theorem, the KL-divergence $\mathrm{KL}(\mathcal{P}||\hat{\mathcal{P}})$ satisfies

$$\mathrm{KL}(\mathcal{P}||\hat{\mathcal{P}}) = \mathrm{KL}(p_T||p_{\mathrm{ref}}) + \Delta I \tag{15}$$

with the integral difference ΔI given by

$$\begin{aligned}\Delta I &= \frac{1}{2}\mathbb{E}_{\mathcal{P}}\left[\int_0^T \left\|\boldsymbol{\Sigma}(T-t,\boldsymbol{Z}_t)\nabla_{\boldsymbol{Z}}\log p_{T-t}(\boldsymbol{Z}_t) - \boldsymbol{\Sigma}(T-t,\boldsymbol{Z}_t)\boldsymbol{s}_{T-t}(\boldsymbol{Z}_t)\right\|^2_{\boldsymbol{\Sigma}^{-1}(T-t,\boldsymbol{Z}_t)}\mathrm{d}t\right] \\ &= \frac{1}{2}\mathbb{E}_{\mathcal{P}}\left[\int_0^T \left\|\nabla_{\boldsymbol{Y}}\log p_t(\boldsymbol{Y}_t) - \boldsymbol{s}_t(\boldsymbol{Y}_t)\right\|^2_{\boldsymbol{\Sigma}(t,\boldsymbol{Y}_t)}\mathrm{d}t\right],\end{aligned}$$

where we use the notation $||\boldsymbol{x}||_{\boldsymbol{A}} = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$. Now thanks to the denoising score matching trick, we get that, up to the additive constant (w.r.t. optimization) term $\mathrm{KL}(p_T||p_{\mathrm{ref}})$,

$$\begin{aligned}\mathrm{KL}(\mathcal{P}||\hat{\mathcal{P}}) &\equiv \frac{1}{2}\mathbb{E}_{\mathcal{P}}\left[\int_0^T \left\|\nabla_{\boldsymbol{Y}}\log p_{t|0}(\boldsymbol{Y}_t|\boldsymbol{Y}_0) - \boldsymbol{s}_t(\boldsymbol{Y}_t)\right\|^2_{\boldsymbol{\Sigma}(t,\boldsymbol{Y}_t)}\mathrm{d}t\right] \\ &\equiv \mathbb{E}_{\mathcal{P}}\left[\int_0^T b\Big(\nabla_{\boldsymbol{Y}}\log p_{t|0}(\boldsymbol{Y}_t|\boldsymbol{Y}_0) - \boldsymbol{s}_t(\boldsymbol{Y}_t)\Big)^\top \mathrm{diag}(\boldsymbol{Y}_t)\Big(\nabla_{\boldsymbol{Y}}\log p_{t|0}(\boldsymbol{Y}_t|\boldsymbol{Y}_0) - \boldsymbol{s}_t(\boldsymbol{Y}_t)\Big)\mathrm{d}t\right]\end{aligned}$$

ODE formulation for sampling. The ODE formulation consists in finding an ODE

$$\frac{\mathrm{d}\boldsymbol{Y}_t}{\mathrm{d}t} = \tilde{\boldsymbol{f}}_t(t,\boldsymbol{Y}_t) \tag{16}$$

that admits the same temporal marginals as the solution of equation (9). Using the formulation in Song et al. (2021b), or simply by applying Ito's lemma, one gets:

$$\tilde{\boldsymbol{f}}_t(t,\boldsymbol{Y}_t) = \boldsymbol{f}_t(t,\boldsymbol{Y}_t) - \frac{1}{2}\nabla\cdot\boldsymbol{\Sigma}(t,\boldsymbol{Y}_t) - \frac{1}{2}\boldsymbol{\Sigma}(t,\boldsymbol{Y}_t)\nabla_{\boldsymbol{Y}}\log p_t(\boldsymbol{Y}_t), \tag{17}$$

which in our case results in

$$\frac{\mathrm{d}\boldsymbol{Y}_t}{\mathrm{d}t} = b(\boldsymbol{\alpha} - \mathbf{1} - \boldsymbol{Y}_t - \mathrm{diag}(\boldsymbol{Y}_t)\nabla_{\boldsymbol{Y}}\log p_t(\boldsymbol{Y}_t)) \tag{18}$$

This highlights another benefit of the ODE formulation: we can simulate the ODE in the log-domain and get an equation of the form $\frac{\mathrm{d}\log \boldsymbol{y}}{\mathrm{d}t} = -b(1 + \nabla_{\boldsymbol{y}}\log p_t(\boldsymbol{y}))$, to promote numerical stability.