LEARNING CLEAN GENERATIVE PROCESSES FROM NOISY MEASUREMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study how to learn a generative model of clean images using only noisy measurements with a known noise law. Diffusion processes and their generalizations beyond Gaussian noise require posterior predictions at every noise level. We first show that it is possible to learn these predictions from independent noisy views using Noise2Noise regression, but only at or above the noise level of the training data. To learn the predictions below that level, needed to generate clean images, we impose self-consistency: a prediction should equal its expected value after continuing the model's sampling process. We derive conditions under which self-consistency uniquely determines the continuation to the clean image distribution and verify them for Poisson, binomial, Gamma, and negative-binomial measurements. Experiments on CIFAR-10, microscopy data, and synthetic aperture radar show that the required posterior predictions can be learned from noisy views alone and that self-consistency repairs the continuation beyond the noise level of the data, which otherwise collapses.

1 INTRODUCTION

Can one learn to generate clean images from a dataset of noisy observations? This question is common in scientific imaging. Photon-counting microscopy often yields only short exposures with a few photons per pixel, because more light would damage the sample (Krull et al., 2024). Every synthetic aperture radar acquisition carries speckle, and averaging looks to reduce it costs resolution (Dalsasso et al., 2022; 2021). In cryo-electron tomography the electron beam destroys the sample, so the dose must stay low and the projections are always very noisy (Buchholz et al., 2019; Liu et al., 2022; Kishore et al., 2026). In each case the measurements follow a known measurement law, and clean images are unavailable. Diffusion models, however, are typically trained by regressing onto clean images, and here the model must generate images cleaner than any it sees during training.

Noise2Noise (Lehtinen et al., 2018) demonstrated that it is possible to learn an effective denoiser from noisy images alone, by asking the network to predict one noisy measurement of an image from another. Because the noise in the target averages out, the minimizer of the expected squared error is still the posterior mean of the clean image given the input and the network learns a denoiser without ever seeing a clean image. Later work showed that multiple noisy views can be obtained from a single one, either via adding more noise (Moran et al., 2020; Pang et al., 2021) or via splitting, not just for Gaussian measurements but also Poisson, Gamma and negative-binomial ones (Neufeld et al., 2024; Leiner et al., 2025; Monroy et al., 2025). We refer to the common mechanism behind these methods as the *N2N principle*: regression onto a noisy target learns the same posterior mean as regression onto the clean image, provided that the conditional mean of the target is the clean image and that the target is independent of the input given the clean image. The same idea trains Gaussian diffusion models from noisy data, since these models generate images using denoisers at successive noise levels (Song et al., 2021). Daras et al. (2024) follow this route: noisy images at one noise level supervise the denoisers at all noisier levels, while the consistency loss of Daras et al. (2023a) extends them to cleaner levels.

But this only solves the problem when the measurements are corrupted by Gaussian noise. For Poisson counts, radar speckle, and other measurement models, a Gaussian diffusion is no longer matched to the data: it would add Gaussian noise on top of the noise already present in the measurements, and reversing this process would lead back to the noisy measurements, not to the clean images.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Instead, we build the sampling process directly from the measurement law. We index it by an information level a that increases as the noise decreases, such as the precision of a Gaussian measurement or the exposure of a photon count. Its state at level a is distributed like a measurement at that level, and at its clean limit the state recovers the image, or a known encoding of it. Gaussian diffusion already has this form: after a change of time and scale, it can be viewed as accumulating Gaussian measurements of a single image rather than removing noise from a corrupted one. As more measurements are accumulated, their average converges to the image (Montanari, 2023), and the drift of the process is the conditional mean of the image given the measurements accumulated so far. Photons accumulate in the same way: counts arrive at rates given by the posterior mean of the image given the counts up to this point, and the counts divided by the exposure converge to a clean image (Krull et al., 2024). Section 2.1 introduces the corresponding framework for Gaussian, Poisson, binomial, Gamma, and negative-binomial measurements and identifies the posterior predictions that drive each process. Appendix C describes the five processes in detail. Because the sampler runs the measurement process itself, a run started at a measurement rather than at the empty state samples clean images from their conditional distribution given that measurement.

The N2N principle supervises only part of this process (Figure 1). Let a_{in} denote the information level of the measurements used as the network inputs. Degrading these measurements according to the measurement law provides training inputs at any level $a \leq a_{\mathrm{in}}$. Applying the N2N principle with targets transformed according to the same law then learns the required posterior predictions at these levels. No operation on a single measurement produces a cleaner measurement, and the second measurement is needed as the target. Hence, the predictions above a_{in}, which the sampler needs to reach a clean image, have no training target. Section 2.2 introduces our training strategy for these predictions. We use a property of conditional expectation: the prediction at one state of the sampling process must equal the expected prediction after continuing the process to a state with higher information. We enforce this self-consistency relation along paths generated by the model, while retaining the N2N loss where noisy training targets are available. This extends the self-consistency approach of Daras et al. (2023a; 2024) beyond Gaussian noise to the five measurement laws we consider.

This leaves a central theoretical question: why should self-consistency recover the correct continuation, rather than simply a continuation that is self-consistent with the learned model? Daras et al. (2023a) answer it for Gaussian diffusion under the assumption that the score computed from the learned denoiser is a gradient field, using tools specific to Gaussian diffusion. In Section 3, we give two arguments that do not rely on these tools. The first applies to all five measurement laws when the required posterior predictions are derived from a common function of the state. The second treats the setting used in our experiments, where the network predicts the required quantities directly, and gives uniqueness conditions for Poisson, binomial, Gamma, and negative-binomial measurements. We then connect these continuation results back to the N2N and self-consistency training objective, giving conditions under which exact minimization of these losses identifies the required continuation.

Our contributions are the following. First, we train generative processes built from five measurement laws on noisy data alone, with unbiased noisy targets for every posterior quantity they need, including the Gamma Laplace transform and the negative-binomial moments. Second, we give conditions under which self-consistency uniquely determines the continuation beyond the data, for all five laws when the predictions share one density ratio and for the four jump laws when they are predicted directly. For the jump laws, consistency paths that start at the data level suffice. Third, we show on CIFAR-10 under four measurement laws and on photon-counting microscopy that generation collapses once the sampler passes the range the data supervise, and that self-consistency recovers much of this loss. On radar data, the learned trajectory captures posterior dependence between pixels that independent per-pixel sampling misses. We discuss related work in Appendix B.

2 From noisy observations to a generative model

Let $X_0 \sim p_{\mathrm{data}}$ denote a clean image, where p_{data} is the unknown image distribution that we want to sample from. We do not observe X_0 but only noisy measurements of it, obtained under a known measurement law. Using this law we build a stochastic process $(Z_a)_{a \geq 0}$, indexed by the information level a. For a fixed image x, we write $Q_a(x, \mathrm{d}z)$ for the conditional distribution of Z_a given $X_0 = x$, which is the law of a measurement of x at level a. Higher levels carry more information about X_0:

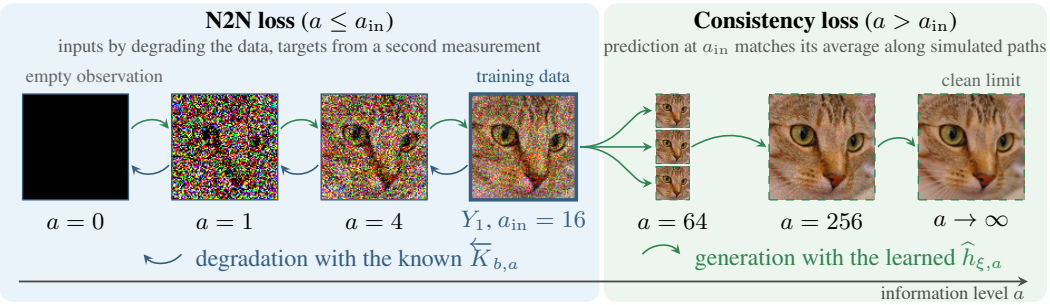


Figure 1: The Poisson observation process of one image, shown as the normalized count image N_a/a. The training data are counts at one level a_{in}. Below it, the known kernels degrade them into N2N inputs (blue). Above it, only the learned process reaches, and the consistency loss matches the prediction at a_{in} with its average along simulated paths (green, Section 2.2). Tiles above a_{in} show the true continuation of this image, and the three at $a = 64$ are independent continuations of the same measurement.

Z_0 is the empty observation, the measurements used as network inputs are at level a_{in}, and as a increases the state, normalized according to the measurement law, converges to X_0 almost surely. Our goal is to simulate this process without observing the clean image.

2.1 Observation processes and posterior predictions

We start with the Gaussian case, following Montanari (2023). Consider the noising process

$$\widetilde{Y}_t = X_0 + \widetilde{W}_t$$

where $\widetilde{W}$ is standard Brownian motion, so that $\mathrm{Var}(\widetilde{W}_t) = t\,\mathrm{Id}$. The noise level of $\widetilde{Y}_t$ is t, and its precision is $a \overset{\text{def.}}{=} 1/t$. We set $Z_a = a\widetilde{Y}_{1/a}$. By time inversion, $Z_a = aX_0 + W_a$ for a standard Brownian motion W, so

$$Q_a(x, \cdot) = \mathcal{N}(ax, a\mathrm{Id}), \qquad Z_a/a \mid X_0 = x \sim \mathcal{N}(x, a^{-1}\mathrm{Id}).$$

Thus, the variance of the normalized observation decreases with a, and Z_a/a converges almost surely to the clean image. The Poisson case is similar: $Z_a = N_a$ records the cumulative counts and

$$Q_a(x, \cdot) = \mathrm{Law}(N_a \mid X_0 = x) = \mathrm{Pois}(ax).$$

As $a \to \infty$, the normalized count image N_a/a converges almost surely to x.

The other measurement (degradation) laws in Table 1 define processes in the same way, with a the binomial detection probability, the Gamma shape, or the negative-binomial shape. Each becomes more informative as a increases and, after normalization, converges almost surely to the clean image or to a known encoding of it. The level a ranges over $(0, a_{\max})$, with $a_{\max} = 1$ for binomial observations and $a_{\max} = \infty$ for the others. Appendix C gives the details for all five laws.

Degradation. Consider first the direction from more information to less. If $a < b$, an observation at the more informative level b can be degraded to level a without knowing the clean image. For photon counts, each recorded photon is kept independently with probability a/b, so $N_a \mid N_b = n \sim \mathrm{Binomial}(n, a/b)$, a rule that does not involve X_0. Let $\overleftarrow{K}_{b,a}(z, \mathrm{d}z')$ denote this degradation kernel. It satisfies

$$Q_b(x, \cdot)\overleftarrow{K}_{b,a} = Q_a(x, \cdot), \qquad \overleftarrow{K}_{c,b}\overleftarrow{K}_{b,a} = \overleftarrow{K}_{c,a}, \qquad a < b < c, \tag{1}$$

so degradation produces the correct law at level a, directly or through an intermediate level. The kernels for all five measurement laws are given in Appendix C.2. We define the joint process through these kernels: given Z_a, the lower level states are drawn from them, independently of X_0. Hence, once Z_a is known, the states at lower levels provide no additional information about X_0:

$$\mathrm{Law}(X_0 \mid (Z_u)_{u \le a}) = \mathrm{Law}(X_0 \mid Z_a)\,.$$

The future of the process therefore depends on its past only through the current state: (Z_a) is a Markov process (Lemma C.1), and the generative dynamics below need only Z_a.

Measurement	Z_a given X_0	Dynamics given X_0	Posterior prediction
Gaussian	$\mathcal{N}(aX_0, a\mathrm{Id})$	drift X_0	$\mathbb{E}[X_0 \mid Z_a]$
Poisson	$\mathrm{Pois}(aX_0)$	pixel i gains a count at rate $X_{0,i}$	$\mathbb{E}[X_{0,i} \mid N_a]$
Binomial ($a < 1$)	$\mathrm{Binomial}(X_0, a)$	pixel i gains a count at rate $\frac{X_{0,i}-N_{a,i}}{1-a}$	$\frac{D_{i,a}(N_a)-N_{a,i}}{1-a}$
Gamma	$\mathrm{Gamma}(a, 1/X_0)$	pixel i jumps by $w > 0$ at rate $e^{-w/X_{0,i}}\,\mathrm{d}w/w$	$H_{i,a}(S_a, w)\,\mathrm{d}w/w$
Negative binomial	$\mathrm{NB}(a, q(X_0))$	pixel i jumps by $k \geq 1$ at rate $q_i(X_0)^k/k$	$M_{i,k,a}(N_a)/k$

Table 1: Observation laws and the dynamics they induce. Gamma distributions are parameterized by shape and rate. The noise acts independently across pixels given X_0, and the level a is written λ, p, L, r in the appendices. The last column replaces the function of the clean image in the dynamics by its posterior prediction, with $D_{i,a}(n) = \mathbb{E}[X_{0,i} \mid N_a = n]$, $H_{i,a}(s, w) = \mathbb{E}[e^{-w/X_{0,i}} \mid S_a = s]$, and $M_{i,k,a}(n) = \mathbb{E}[q_i(X_0)^k \mid N_a = n]$, each of which may depend on the entire observed state. The binomial image is integer-valued, and $q(X_0) \in [0, 1)^d$ is a known pixelwise encoding.

Generation. The generative process moves in the opposite direction. Starting from a state Z_a, we simulate states Z_b at higher information levels $b > a$. If X_0 were known, the measurement law would specify these dynamics. In the Gaussian case $Z_a = aX_0 + W_a$, so $\mathrm{d}Z_a = X_0\,\mathrm{d}a + \mathrm{d}W_a$. In the Poisson case, each pixel i gains a count with probability $X_{0,i}\,\mathrm{d}a$ in a short interval $\mathrm{d}a$. The dynamics depend on X_0 only through functions of it, which we index by ξ and denote $\psi_\xi(X_0)$. Since the clean image X_0 is unknown, we replace each of them by its posterior prediction

$$h_{\xi,a}(z) = \mathbb{E}[\psi_\xi(X_0) \mid Z_a = z]\,. \tag{2}$$

In the Gaussian case, $\psi_\xi(X_0) = X_0$ and the posterior mean $m_a(z) = \mathbb{E}[X_0 \mid Z_a = z]$ takes the place of X_0:

$$\mathrm{d}Z_a = m_a(Z_a)\,\mathrm{d}a + \mathrm{d}B_a,$$

where B is a Brownian motion with respect to the history of the observation process. The same posterior mean appears in the Poisson case, while other measurement laws require different functions of X_0, listed in Table 1.

2.2 Learning posterior predictions $h_{\xi,a}$ from noisy observations

Learning where data are available. Suppose that Y_1 and Y_2 are two noisy measurements of the same image, independent conditional on X_0, and let a_{in} be the information level of Y_1. Following the N2N principle, Y_1 provides the input and Y_2 the target. Except for binomial observations, the two views can also be obtained by randomly splitting a single measurement, which places each view below the level of that measurement (Appendix C.4).

Degrading Y_1 with the known kernels $\overleftarrow{K}_{a_{\mathrm{in}},a}$ gives an input Z_a at any level $a \leq a_{\mathrm{in}}$. If the clean image were available, the natural target associated with this input would be $\psi_\xi(X_0)$, since $h_{\xi,a}(Z_a) = \mathbb{E}[\psi_\xi(X_0) \mid Z_a]$. Because X_0 is not observed, we use Y_2 to construct a noisy substitute. For each quantity $\psi_\xi(X_0)$ required by the dynamics, we choose a transformation $T_\xi(Y_2)$ such that

$$\mathbb{E}[T_\xi(Y_2) \mid X_0] = \psi_\xi(X_0).$$

Both the degradation of Y_1 and the transformation of Y_2 depend on the measurement law. For instance, photon counts use $T_\xi(Y_2) = Y_2/a_{\mathrm{tar}}$, the second count image divided by its exposure a_{tar}. Appendix C.3 derives the transformations T_ξ for every prediction in Table 1.

Since Y_2 and Z_a are independent conditional on X_0, the tower property gives

$$\mathbb{E}[T_\xi(Y_2) \mid Z_a] = \mathbb{E}[\psi_\xi(X_0) \mid Z_a] = h_{\xi,a}(Z_a). \tag{3}$$

Thus, regressing $T_\xi(Y_2)$ on Z_a recovers the posterior prediction the dynamics need. We use

$$\mathcal{L}_{\mathrm{N2N}}(\widehat{h}) = \mathbb{E}_{(a,\xi)\sim\rho}\left[\left\|\widehat{h}_{\xi,a}(Z_a) - T_\xi(Y_2)\right\|^2\right], \tag{4}$$

where ρ is a training distribution over levels $a \leq a_{\mathrm{in}}$ and indices ξ. If $\mathbb{E}_{(a,\xi)\sim\rho}\,\mathbb{E}|T_\xi(Y_2)|^2 < \infty$, the excess loss $\mathcal{L}_{\mathrm{N2N}}(\widehat{h}) - \mathcal{L}_{\mathrm{N2N}}(h)$ equals $\mathbb{E}_{(a,\xi)\sim\rho}\left\|\widehat{h}_{\xi,a}(Z_a) - h_{\xi,a}(Z_a)\right\|^2$, so the N2N principle

applies to the transformed targets (Lemma C.4). Some experiments use Bregman losses with the same minimizer instead of the squared error (Table 7).

For Gaussian observations, (4) is the loss of Daras et al. (2024) up to a level-dependent weight (Appendix C.3.1).

Learning beyond the information levels covered by the data. Degradation provides training inputs only at levels $a \le a_{\mathrm{in}}$, so the N2N loss constrains $\widehat{h}_{\xi,a}$ only there. Generation, however, needs the predictions above a_{in} to approach the clean limit. For $a < b$, let $\overrightarrow{K}_{a,b}$ denote the true transition from Z_a to Z_b, so that

$$(\overrightarrow{K}_{a,b} f)(z) = \mathbb{E}[f(Z_b) \mid Z_a = z].$$

Since Z_a adds no information about X_0 once Z_b is known, the tower property gives

$$h_{\xi,a}(z) = \mathbb{E}[h_{\xi,b}(Z_b) \mid Z_a = z] = (\overrightarrow{K}_{a,b} h_{\xi,b})(z). \tag{5}$$

Thus, the prediction at state z and level a equals the expected prediction at any higher information level b, under the true process started from z. In the Gaussian case this is the self-consistency relation of Daras et al. (2023a). As there, the true transition depends on the unknown predictions, so we replace it by the transition $\widehat{\overrightarrow{K}}_{a,b}$ of the process simulated with the current model $\widehat{h}$ (Daras et al., 2023a; 2024), and define

$$\mathcal{L}_{\mathrm{cons}}(\widehat{h}) = \mathbb{E}_{(a,b,\xi)\sim\rho_{\mathrm{cons}}} \left[\left\| \widehat{h}_{\xi,a}(Z_a) - (\widehat{\overrightarrow{K}}_{a,b} \widehat{h}_{\xi,b})(Z_a) \right\|^2 \right], \tag{6}$$

where ρ_{cons} is a distribution over levels $a < b$ and indices ξ. In our experiments every path starts at the training measurement $a = a_{\mathrm{in}}$, and b ranges over $(a_{\mathrm{in}}, B]$, with B the largest level the sampler reaches. The paths are simulated with the current model in a few coarse steps, without gradients. The inner expectation is estimated from two simulated paths either through the product of their residuals, which is unbiased for (6), or through the square of their average, adding a variance term so that the true predictions no longer minimize it. Table 7 lists the settings of each experiment.

In the experiments, we first train with $\mathcal{L}_{\mathrm{N2N}}$ alone, and then refine the same model by minimizing

$$\mathcal{L}_{\mathrm{N2N}}(\widehat{h}) + \lambda_{\mathrm{cons}} \mathcal{L}_{\mathrm{cons}}(\widehat{h}), \qquad \lambda_{\mathrm{cons}} > 0. \tag{7}$$

With the product estimator, any global minimizer matches the true predictions where the N2N loss samples and is self-consistent along the simulated paths (Proposition F.1). However, this alone does not guarantee that the continuation beyond a_{in} coincides with the true predictions h. Section 3 gives conditions under which it does.

3 When consistency determines the continuation

We now give conditions under which the true predictions at one positive information level determine their continuation toward the clean limit.

Identifying the process from one information level. Assume there exists a level $a_* \in (0, a_{\mathrm{in}}]$ at which the N2N loss recovers the true posterior predictions: $\widehat{h}_{\xi,a_*} = h_{\xi,a_*}$ for all required predictions. The input Z_{a_*} is obtained from the training measurements and hence has the true law p_{a_*}.

For $b > a_*$, the true law $p_b = \mathrm{Law}(Z_b)$ satisfies $p_b \overleftarrow{K}_{b,a_*} = p_{a_*}$. For all five measurement laws, the map $\nu \mapsto \nu \overleftarrow{K}_{b,a_*}$, which sends a distribution at level b to its degraded distribution at the positive level a_*, is injective (Proposition C.2). The distribution at a higher level is therefore uniquely determined by the distribution obtained after degrading it to a_*, a property related to the classical identifiability of mixtures (Teicher, 1961).

Marginal distributions alone do not determine a process since different processes can relate their states differently across levels. The backward transitions remove this ambiguity. Indeed,

$$\mathrm{Law}(Z_a \in \mathrm{d}z, Z_b \in \mathrm{d}w) = p_b(\mathrm{d}w) \overleftarrow{K}_{b,a}(w, \mathrm{d}z), \qquad a < b.$$

The learned process is Markov as well, since its dynamics depend only on the level and the current state. A Markov process that starts from p_{a_*} and has the known kernels as its backward transitions therefore has the law of the true observation process (Theorem C.3). The question is whether consistency forces the learned process to have these backward transitions.

The equation imposed by self-consistency. The predictions $\widehat{h}$ determine the dynamics of the learned process, and hence its generator $\mathcal{A}_a[\widehat{h}]$, which gives the expected rate of change of $f(\widehat{Z}_a)$ for a test function f (Appendix C.1).

Consistency between every pair of information levels

$$\widehat{h}_{\xi,a} = \widehat{\overrightarrow{K}}_{a,b} \widehat{h}_{\xi,b}, \qquad a < b,$$

gives in the limit $b \downarrow a$

$$\big(\partial_a + \mathcal{A}_a[\widehat{h}]\big)\widehat{h}_{\xi,a} = 0, \qquad a_* < a < B. \tag{8}$$

Unique continuation asks whether the value of $\widehat{h}$ at a_* selects a unique solution of this equation.

Two arguments for unique continuation. The true predictions derive from a density ratio: for each law there is a known reference distribution r_a with $p_a(\mathrm{d}z) = F_a(z)\, r_a(\mathrm{d}z)$. For example,

$$m_a(z) = \nabla_z \log F_a(z), \qquad h_{i,a}(n) = \frac{F_a(n + e_i)}{F_a(n)}$$

for Gaussian and Poisson observations. If the learned posterior predictions also derive from one such function F_a, substituting them into (8) yields a linear equation after rescaling F_a by a positive factor depending only on a. The learned process is then a reweighting of the reference process with the known backward kernels, so it must be the true observation process (Appendix D). For Gaussian observations, this is the argument of Daras et al. (2023a), who assume that the score of the denoiser is a gradient field.

Our networks, like those of Daras et al. (2024) for Gaussian diffusions, predict the required posterior quantities directly and do not enforce a common function F. For the four jump processes, we instead prove uniqueness directly: Theorem 3.1 shows that if the learned and true predictions satisfy the same consistency equation and coincide at a_*, then, under the process-dependent bounds and regularity conditions stated in Assumption E.1, they coincide throughout $[a_*, B]$.

Theorem 3.1. *Fix $0 < a_* < B < a_{\max}$. Suppose that $\widehat{h}_{\xi,a_*} = h_{\xi,a_*}$ for every required prediction and that $\widehat{h}$ satisfies* (8) *on (a_*, B) in the sense of Appendices D and E. Assume one of the following.*

1. *All predictions derive from one density ratio function with the regularity of Assumption D.1.*
2. *The observations are Poisson, binomial, Gamma, or negative binomial, and X_0 and the learned predictions satisfy the bounds and regularity conditions of Assumption E.1, which the true predictions also meet.*

Then $\widehat{h}_{\xi,a} = h_{\xi,a}$ for $a_* \le a \le B$ and every required prediction. Further, the process generated by $\widehat{h}$ and initialized with p_{a_*} has the law of the true observation process on $[a_*, B]$.

From the training losses to the theorem. An exact global minimizer of (7) is correct where the N2N loss samples, in particular at $a_* = a_{\mathrm{in}}$, and consistent along the sampled paths (Proposition F.1). These are the two conditions of Theorem 3.1, except that the theorem needs consistency between all nearby levels, not only along the sampled paths. Daras et al. (2023a; 2024) obtain it by starting paths at every level. Our paths start only at a_{in}, which suffices for the jump laws. For count models with bounded rates, the learned transition from a_{in} to any level a is one-to-one on bounded functions, and for Gamma on bounded Lipschitz functions, so it cannot hide an inconsistency: predictions that are consistent from a_{in} to both a and a later level b are also consistent from a to b (Lemmas F.2 and F.4). We have no such result for Gaussian observations, where our argument needs paths that start at every level (Remark F.1). Under these conditions an exact minimizer has the true predictions at every level in $[a_{\mathrm{in}}, B]$ (Corollaries F.3 and F.5).

Our experiments depart from these conditions in five ways: the Poisson and negative-binomial models use the averaged, biased estimator, the consistency paths use a few coarse steps, the networks enforce only some of the bounds, the negative-binomial networks predict finitely many moments, and the radar model adds a regression on simulated states (Section 4).

4 EXPERIMENTS

Our experiments test three claims. The *anchor readout* checks that noisy views suffice to train the process where the measurements reach, $a \leq a_{\mathrm{in}}$ (Section 2.2). It generates from the empty state, stops at $a \leq a_{\mathrm{in}}$ and returns the predicted posterior mean. The *full trajectory* continues through the levels above a_{in}, which no pair of views reaches, and so tests whether self-consistency from the anchor alone supervises them (Section 3). We report both for CIFAR-10 and microscopy. On radar data, Section 4.3 tests whether the measurement-matched process carries posterior dependence between pixels (Section 2.1), against independent per-pixel sampling.

4.1 CONTINUING BEYOND THE OBSERVATIONS

For CIFAR-10, we draw two conditionally independent measurements of each training image, redrawn at every epoch, under the Poisson, binomial, Gamma and negative-binomial laws and fit a 60M-parameter U-Net to the posterior quantity each process requires. We vary the observation level a_{in}, and with it how much of the trajectory self-consistency must supervise. Redrawn views require a repeatable acquisition, so CIFAR-10 serves as a controlled study.

Table 2 compares the anchor with the full trajectory after 200 further epochs with and without consistency refinement. Without it ($\lambda_{\mathrm{cons}} = 0$), continuing past a_{in} degrades the samples to FID 35–182, as expected of a sampler that uses the network where it was never trained. With it ($\lambda_{\mathrm{cons}} > 0$), FID falls by a factor of 3.0–11.5 wherever the $\lambda_{\mathrm{cons}} = 0$ trajectory has a finite FID. The two columns differ only in the loss weight, so the gain is due to the consistency term. The binomial and Gamma models use the unbiased product estimator, whose expectation is the loss analyzed in Section 3. With it the refined trajectory beats the anchor at 25.1 and 8.0 binomial counts and for Gamma, ties at 100.0 counts and loses at 180.1 counts. The Poisson and negative-binomial models use the averaged estimator, which also penalizes the path variance and gives lower FID for Poisson. With the product estimator the Poisson trajectory reaches 5.14, 22.52 and 17.84, still far below the $\lambda_{\mathrm{cons}} = 0$ column but above the anchor at the two lower exposures (Table 9).

The trajectories are continued to different depths. Poisson and binomial run to the end of the grid, for Poisson 76 to 2900 times the anchor exposure, negative binomial to $2r_{\mathrm{obs}}$ and Gamma to $50L_{\mathrm{obs}}$. The Poisson trajectory returns its normalized count state, the others return the predicted clean image (Table 7). At five of the eleven settings, 178.4 Poisson counts, 180.1 binomial counts and all three negative-binomial shapes, the anchor has lower FID than the refined trajectory. In a single-seed Gamma run at $L_{\mathrm{obs}} = 16$, continuing to $50L_{\mathrm{obs}}$ gives FID 20.36, against 12.01 when stopping at $2L_{\mathrm{obs}}$ (Appendix H.2). The theory identifies the continuation but bounds its error growth only exponentially (Remark E.1), so it does not rule out that a long continuation loses to stopping earlier.

Misspecifying the measurement law. Two Gaussian baselines with the same architecture and the true photon rate, one with a single global variance and one after the Anscombe transform, have anchor FID 8.7–14.9 higher than the Poisson model at the two higher counts and more than six times higher at the lowest (Table 10).

The negative-binomial jump rates use all predicted posterior moments $\mathbb{E}[q_i^k \mid N_r]$. Rates derived from the predicted mean alone raise FID from 12.44 to 22.08 at $r_{\mathrm{obs}} = 50$ (10,000 samples, Appendix H.2).

4.2 PHOTON-COUNTING MICROSCOPY

CONV-PC (Krull et al., 2024) records 512 photon-counting frames of a Convallaria section, each with 0.04 to 0.07 photons per pixel on average. No clean image exists, so the only reference is the sum of all frames, itself a count image. Generative Accumulation of Photons (GAP) (Krull et al., 2024) trains on these frames by predicting the complement of a binomial subset of each one. We train the same way, following the GAP evaluation protocol, in which frames of the test field of view

			Full trajectory	
Process	Observation	Anchor	$\lambda_{\text{cons}} = 0$	+ consistency
Poisson	613.1 counts	6.17 ± 0.17	35.32 ± 3.56	3.85 ± 0.06
Poisson	178.4 counts	7.07 ± 0.01	48.19 ± 4.19	8.47 ± 0.11
Poisson	16.1 counts	16.61 ± 0.14	119.00 ± 7.35	13.13 ± 0.02
Binomial	180.1 counts	9.41 ± 0.06	61.17 ± 2.48	11.65 ± 0.12
Binomial	100.0 counts	11.68 ± 0.51	95.23 ± 7.47	11.63 ± 1.65
Binomial	25.1 counts	16.18 ± 0.24	152.38 ± 12.55	13.23 ± 0.61
Binomial	8.0 counts	22.70 ± 0.12	181.55 ± 8.38	19.68 ± 0.15
Neg. binomial	$r_{\text{obs}} = 5$	19.99 ± 0.17	66.70 ± 0.66	22.44 ± 0.31
Neg. binomial	$r_{\text{obs}} = 16$	13.54 ± 0.18	56.77 ± 5.07	15.27 ± 0.59
Neg. binomial	$r_{\text{obs}} = 50$	7.78 ± 0.15	42.66 ± 0.92	9.61 ± 1.36
Gamma	$L_{\text{obs}} = 64$	9.32 ± 0.06	—	7.06 ± 0.27

Table 2: CIFAR-10 FID from 50,000 EMA samples against the clean training set, mean $\pm$ standard deviation over two training seeds. All models train for 2000 epochs, where consistency runs use positive weight ($\lambda_{\text{cons}} > 0$) for the last 200. The anchor is a posterior readout from the $\lambda_{\text{cons}} = 0$ model, and is listed for reference. The consistency estimator is averaged for Poisson and negative binomial and the product otherwise (Appendix H.2). Counts are expected counts at a white pixel. Rescoring one checkpoint moves FID by 0.143. Without refinement the Gamma trajectory diverges (—).

are also used for training: each frame is split into a binomially thinned input and its complement, so supervision ends at the exposure of one frame while the reference has 512 times as many photons. Unlike CIFAR-10, the measurement here is fixed by the instrument, and only its split into two views is redrawn.

Table 3 repeats the CIFAR-10 comparison on 10,000 images per method. The anchor readout at the one-frame level matches reproduced GAP at 256×256 and is worse at 64×64 (114.1 against 98.1). Continuing without consistency gives FID 339.4 and 267.4 at the two resolutions. Figure 2(b) shows this failure from a measured frame, next to the refined continuation in (c). After twenty epochs of consistency refinement the same continuation reaches 44.7 ± 2.5 over three refinement seeds of one checkpoint, and every seed beats reproduced GAP at either β. At 64×64 it beats reproduced GAP under two of three renderings (Appendix H.3).

4.3 Synthetic aperture radar

Radar intensities exhibit multiplicative $\text{Gamma}(L)$ speckle at L looks. Following Dalsasso et al. (2022), we form conditionally independent half-look pairs at $L_{\text{obs}} = 1/2$ from the squared real and imaginary components of three Sentinel-1 acquisitions, after splitting each image into four sublattices to decorrelate neighboring speckle. The N2N loss therefore supervises the predictions only up to half a look, whereas radar images are usually constructed from several.

The network parameterizes each pixel's posterior as a four-component inverse-Gamma mixture, whose transform and samples are then available in closed form. We compare two samplers conditioned on the same observed half-look. The *independent sampler* draws each pixel from its predicted posterior, with no dependence between pixels. In contrast, the *joint sampler* evolves the whole image along the learned trajectory. Sampling from L_{obs} to $2L_{\text{obs}}$ generates an additional half-look, so both samplers' predictive intervals can be scored against the held-out second view. They agree closely at the pixel scale, but over 16×16 blocks the joint sampler achieves 88.6–89.2% coverage for nominal 90% intervals across two seeds, against 82.5–83.2% for the independent sampler (Table 12). At $32L_{\text{obs}}$, the independent sampler fragments bright measurements into isolated pixels, while the joint sampler produces coherent patches (Figure 3).

Besides self-consistency, refinement uses a rollout regression from simulated states to the held-out view. It stabilizes predictions at the observed level, but its optimum is the posterior given the starting

Method	FID 64^2	FID 256^2
GAP ($\beta = 0.1$)	98.1	58.4
GAP ($\beta = 0.05$)	97.2	55.8
N2N, anchor readout	114.1	55.7
N2N, full trajectory	267.4	339.4
N2N + consistency, full trajectory	93.3 ± 2.5	44.7 ± 2.5

Table 3: Generation on CONV-PC: 10,000 samples per method, fixed $[0, 4]$ rendering, reference crops from the 512-frame sums. Every method returns a predicted image. The anchor stops at one frame, the full trajectories continue to the reference exposure, and GAP accumulates under its own rule with fraction β per step. The consistency row is mean $\pm$ standard deviation over three seeds. Other rows are single runs.

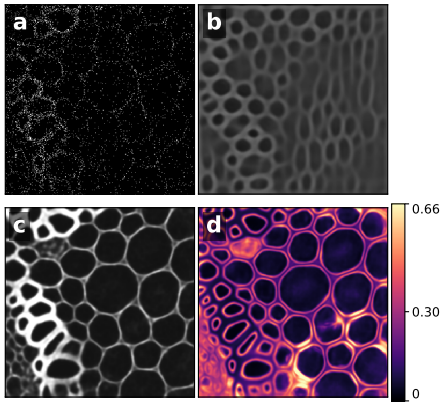


Figure 2: Continuation from one measured frame. (a) The frame, 5,020 photons in 256×256 pixels. (b) N2N continuation to the reference exposure without consistency. (c) The same continuation after consistency refinement. (d) Relative uncertainty of 500 such continuations, pixelwise standard deviation over mean.

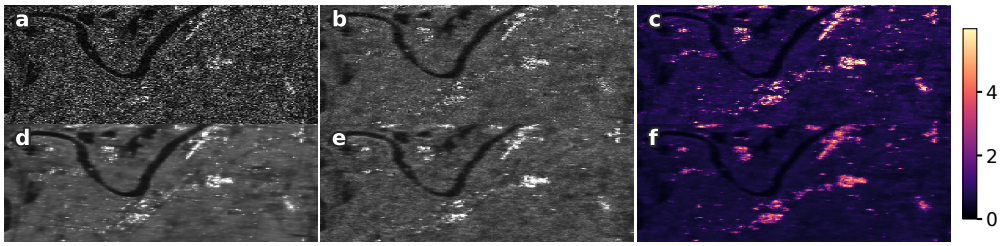


Figure 3: Independent and joint sampling at $L = 16$ ($32L_{\text{obs}}$). (a) Observed half-look view; (d) posterior mean. (b, e) One independent and one joint sample; (c, f) standard deviation over 50 draws, with a shared color bar. Reduced pixel noise reveals spatial structure, but the trajectory underestimates absolute spread beyond $16L_{\text{obs}}$ (Appendix H.4). Panels are shown at ground aspect, with range along the short axis.

view rather than the simulated state, and on its own it lowers the spread (Appendix H.4, Table 13). On a reserved reference band that extends the held-out comparison to roughly $4.5L_{\text{obs}}$, refinement raises the fraction of its own predicted per-pixel spread that the joint sampler realizes from 0.95 to about 0.985 on both seeds, against 0.96–0.97 for a zero-weight control. Coverage at the band changes by less than one point, so this measures agreement with the network's own posterior rather than accuracy.

5 DISCUSSION

A generative process built from the measurement law itself can be trained from noisy measurements alone. Noise2Noise regression supervises it up to the level of the data, and self-consistency determines its continuation toward cleaner images, uniquely under the conditions of Section 3. In the CIFAR-10 settings and on photon-counting microscopy, the consistency term repairs a continuation that otherwise collapses, and on radar data the learned trajectory captures dependence between pixels.

The theory is an identification result. The trained networks do not enforce all of its conditions, and its stability estimate grows exponentially with the image size and the length of the continuation (Remark E.1). The method also assumes a known measurement law and level and pixelwise independent noise, which for radar costs resolution. Future work could enforce one common density ratio in the network by construction and treat mixed laws such as Poisson–Gaussian.

AI use statement. In this work, we used generative AI tools for research ideation and execution: to help develop the theoretical framework, propose and refine hypotheses, formulate mathematical claims, provide ingredients for their proofs and assist in writing the proofs, design and give feedback on experiments, implement the methods, and interpret results. We have not used generative AI tools to generate synthetic data sets or to clean or reformat datasets, and translation and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to draft sections of the paper, polish its writing, write and edit code, retrieve, discover and summarize related work, and format references. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement. The appendices give the mathematical assumptions, proofs, training objectives and sampling approximations, and Table 7 lists the training and sampling settings of every experiment. The code will be shared with the reviewers and area chairs through an anonymous repository during the discussion phase.

Ethics statement. The work concerns generative modeling from noisy scientific measurements. Noisy training images may retain identifying information, and their corruption alone provides no privacy guarantee.

REFERENCES

Asad Aali, Marius Arvinte, Sidharth Kumar, and Jonathan I. Tamir. Solving inverse problems with score-based generative priors learned from noisy data. In *Asilomar Conference on Signals, Systems, and Computers*, pp. 837–843. IEEE, 2023.

David Applebaum. *Lévy Processes and Stochastic Calculus*, volume 116 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, second edition, 2009.

Weimin Bai, Yifei Wang, Wenzheng Chen, and He Sun. An expectation-maximization algorithm for training clean diffusion models from corrupted observations. *Advances in Neural Information Processing Systems*, 37:19447–19471, 2024.

Joshua Batson and Loïc Royer. Noise2Self: Blind denoising by self-supervision. In *International Conference on Machine Learning*, pp. 524–533. PMLR, 2019.

Ralph P. Boas, Jr. and David V. Widder. Functions with positive differences. *Duke Mathematical Journal*, 7(1):496–503, 1940.

Tim-Oliver Buchholz, Mareike Jordan, Gaia Pigino, and Florian Jug. Cryo-CARE: Content-aware image restoration for cryo-transmission electron microscopy data. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 502–506. IEEE, 2019.

Tianqi Chen and Mingyuan Zhou. Learning to jump: Thinning and thickening latent counts for generative modeling. In *International Conference on Machine Learning*, pp. 5367–5382. PMLR, 2023.

Emanuele Dalsasso, Loïc Denis, and Florence Tupin. SAR2SAR: A semi-supervised despeckling algorithm for SAR images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:4321–4329, 2021.

Emanuele Dalsasso, Loïc Denis, and Florence Tupin. As if by magic: Self-supervised training of deep despeckling networks with MERLIN. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

Giannis Daras, Yuval Dagan, Alex Dimakis, and Constantinos Daskalakis. Consistent diffusion models: Mitigating sampling drift by learning to be consistent. *Advances in Neural Information Processing Systems*, 36:42038–42063, 2023a.

Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alex Dimakis, and Adam Klivans. Ambient diffusion: Learning clean distributions from corrupted data. *Advances in Neural Information Processing Systems*, 36:288–313, 2023b.

Giannis Daras, Alex Dimakis, and Constantinos Daskalakis. Consistent diffusion meets Tweedie: Training exact ambient diffusion models with noisy data. In *International Conference on Machine Learning*, pp. 10091–10108. PMLR, 2024.

Giannis Daras, Adrian Rodriguez-Munoz, Adam Klivans, Antonio Torralba, and Constantinos Daskalakis. Ambient diffusion omni: Training good models with bad data. *Advances in Neural Information Processing Systems*, 38:42636–42673, 2025.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Nic Fishman, Gokul Gowri, Tanush Kumar, Jiaqi Lu, Valentin De Bortoli, Jonathan Gootenberg, and Omar Abudayyeh. Count bridges enable modeling and deconvolving transcriptomic data. In *International Conference on Learning Representations*, pp. 129437–129477, 2026.

B. Grigelionis. On representation of integer-valued random measures by means of stochastic integrals with respect to the Poisson measure. *Lithuanian Mathematical Journal*, 11(1):93–108, 1971.

Peter Holderrieth, Marton Havasi, Jason Yim, Neta Shaul, Itai Gat, Tommi Jaakkola, Brian Karrer, Ricky T. Q. Chen, and Yaron Lipman. Generator matching: Generative modeling with arbitrary Markov processes. In *International Conference on Learning Representations*, pp. 52153–52219, 2025.

Danial Hosseintabar, Fan Chen, Giannis Daras, Antonio Torralba, and Constantinos Daskalakis. DiffEM: Learning from corrupted data with diffusion models via expectation maximization. *arXiv preprint arXiv:2510.12691*, 2025.

Jiantao Jiao, Kartik Venkat, and Tsachy Weissman. Relations between information and estimation in discrete-time Lévy channels. *IEEE Transactions on Information Theory*, 63(6):3579–3594, 2017.

Bahjat Kawar, Noam Elata, Tomer Michaeli, and Michael Elad. GSURE-based diffusion model training with corrupted data. *Transactions on Machine Learning Research*, 2024.

Kwanyoung Kim and Jong Chul Ye. Noise2Score: Tweedie's approach to self-supervised image denoising without clean images. *Advances in Neural Information Processing Systems*, 34:864–874, 2021.

Vinith Kishore, Valentin Debarnot, Ricardo D Righetto, Benjamin D Engel, and Ivan Dokmanić. ICECREAM: high-fidelity equivariant cryo-electron tomography. *Biological Crystallography*, 82 (4), 2026.

Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2Void – Learning denoising from single noisy images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2124–2132. IEEE, 2019.

Alexander Krull, Hector Basevi, Benjamin Salmon, Andre Zeug, Franziska Müller, Samuel Tonks, Leela Muppala, and Aleš Leonardis. Image denoising and the generative accumulation of photons. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1517–1526. IEEE, 2024.

Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2Noise: Learning image restoration without clean data. In *International Conference on Machine Learning*, pp. 2965–2974. PMLR, 2018.

James Leiner, Boyan Duan, Larry Wasserman, and Aaditya Ramdas. Data fission: Splitting a single data point. *Journal of the American Statistical Association*, 120(549):135–146, 2025.

Brett Levac, Jon Tamir, Marcelo Pereyra, and Julián Tachella. Normalization-equivariant diffusion models: Learning posterior samplers from noisy and partial measurements. In *International Conference on Machine Learning*, 2026.

Yun-Tao Liu, Heng Zhang, Hui Wang, Chang-Lu Tao, Guo-Qiang Bi, and Z Hong Zhou. Isotropic reconstruction for electron tomography with deep learning. *Nature communications*, 13(1):6482, 2022.

Chirag Modi, Jiequn Han, Eric Vanden-Eijnden, and Joan Bruna. Generative modeling from black-box corruptions via self-consistent stochastic interpolants. In *International Conference on Learning Representations*, pp. 104078–104107, 2026.

Brayan Monroy, Jorge Bacca, and Julián Tachella. Generalized recorrupted-to-recorrupted: Self-supervised learning beyond Gaussian noise. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28155–28164. IEEE, 2025.

Andrea Montanari. Sampling, diffusions, and stochastic localization. *arXiv preprint arXiv:2305.10690*, 2023.

Nick Moran, Dan Schmidt, Yu Zhong, and Patrick Coady. Noisier2Noise: Learning to denoise from unpaired noisy data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12061–12069. IEEE, 2020.

Anna Neufeld, Ameer Dharamshi, Lucy L. Gao, and Daniela Witten. Data thinning for convolution-closed distributions. *Journal of Machine Learning Research*, 25(57):1–35, 2024.

Andrey Okhotin, Dmitry Molchanov, Vladimir Arkhipkin, Grigory Bartosh, Viktor Ohanesian, Aibek Alanov, and Dmitry P. Vetrov. Star-shaped denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 36:10038–10067, 2023.

Tongyao Pang, Huan Zheng, Yuhui Quan, and Hui Ji. Recorrupted-to-recorrupted: Unsupervised deep learning for image denoising. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2043–2052. IEEE, 2021.

Herbert Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 157–163. University of California Press, 1956.

François Rozet, Gérôme Andry, François Lanusse, and Gilles Louppe. Learning diffusion priors from observations by expectation maximization. *Advances in Neural Information Processing Systems*, 37:87647–87682, 2024.

Javier E. Santos, Zachary R. Fox, Nicholas Lubbers, and Yen Ting Lin. Blackout diffusion: Generative diffusion models in discrete-state spaces. In *International Conference on Machine Learning*, pp. 9034–9059. PMLR, 2023.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

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*, 2021.

Henry Teicher. Identifiability of mixtures. *The Annals of Mathematical Statistics*, 32(1):244–248, 1961.

David V. Widder. *The Laplace Transform*, volume 6 of *Princeton Mathematical Series*. Princeton University Press, 1941.

David V. Widder. Positive temperatures on an infinite rod. *Transactions of the American Mathematical Society*, 55:85–95, 1944.

Yuchen Zhang and Jian Zhou. Inverse flow and consistency models. In *International Conference on Machine Learning*, pp. 77152–77176. PMLR, 2025.

Mingyuan Zhou, Tianqi Chen, Zhendong Wang, and Huangjie Zheng. Beta diffusion. *Advances in Neural Information Processing Systems*, 36:30070–30095, 2023.

A GUIDE TO THE APPENDIX

Appendix B relates the method to denoising from noisy data, diffusion models trained on corrupted data, generative processes beyond Gaussian noise, and self-consistency.

Appendix C defines the five observation processes considered in the paper. For each model, we specify how observations at different information levels are related, which posterior quantities determine the generative process, and which noisy targets can be used to learn these quantities. We also prove an important identification result: once the distribution of the observation is known at any positive information level, there is at most one compatible distribution at every higher information level.

Appendices D and E then address the main continuation question. Suppose the required posterior predictions are known at some positive level a_* and satisfy self-consistency above that level. Does this force the predictions at higher information levels to be the true ones? We give two answers. Appendix D treats the case where all posterior predictions are obtained from one common density ratio. Appendix E treats the setting used in our experiments, where the network predicts the required quantities directly.

Appendix F connects these continuation results to the training objective. It states conditions under which an exact minimizer of the expected regression and consistency losses satisfies the hypotheses required by the continuation theorems. Finally, Appendix G describes the approximations introduced by numerical sampling, and Appendix H gives the experimental details and additional results.

We use the notation of the main text throughout. The kernel $\overleftarrow{K}_{b,a}$ takes an observation at a more informative level b and degrades it to a less informative level $a < b$. In the opposite direction, $\overrightarrow{K}_{a,b}$ denotes the true forward transition of the generative process. A hat denotes a learned quantity. A superscript 0 denotes a reference process introduced below. In every model, increasing the information level means that the observation contains more information about the clean image.

Result	Location
Theorem 3.1	Item 1: Corollary D.5. Item 2: Corollary E.4 on $[a_*, B]$
Noisy-target regression	Lemma C.4 and Proposition C.6
Injectivity of degradation	Proposition C.2
Identification from a positive-level marginal	Theorem C.3
Continuation from a common density ratio	Corollary D.5
Uniqueness for direct predictions	Theorem E.1 and Corollary E.4
Poisson and binomial rates with linear growth	Theorem E.6
Exact minimizers of the training objective	Proposition F.1
Training from one positive starting level	Corollaries F.3 and F.5
Poisson generation from zero	Corollary F.6

Table 4: Main statements proved in the appendix and where they are located.

Symbol	Meaning
$X_0 \sim p_{\mathrm{data}}$	Clean image and its distribution
a; λ, p, L, r	Generic information level; Poisson exposure, binomial detection probability, Gamma shape, and negative-binomial shape
Z_a, $Q_a(x, \mathrm{d}z)$, p_a	Observation at level a, its conditional law given $X_0 = x$, and its marginal law
$\overleftarrow{K}_{b,a}$	Known degradation kernel from b to $a < b$
$\overrightarrow{K}_{a,b}$, $\widehat{\overrightarrow{K}}_{a,b}$, $\overrightarrow{K}^0_{a,b}$	True, learned, and reference forward transitions
r_a, F_a, G_a	Reference marginal, density ratio, and a rescaled density ratio used in the continuation proof
$h_{\xi,a}$	Generic posterior prediction required by the dynamics
$m_a, h_{i,a}, D_{i,p}, H_{i,L}(\,\cdot\,, w), M_{i,k,r}$	Gaussian posterior mean, Poisson intensity, binomial posterior mean, Gamma posterior Laplace transform, and negative-binomial posterior moment
$\mathcal{A}_a[h]$, $\mathcal{A}^0$	Generator of the process driven by predictions h, and generator of the reference process
a_*, a_{in}, $I = [a_*, B]$	Positive level where the predictions are already identified, input level, and a finite continuation interval
S	Count states that can occur with positive probability
$\mathcal{L}_{\mathrm{N2N}}$, $\mathcal{L}_{\mathrm{cons}}$, λ_{cons}	Regression loss, consistency loss, and its positive weight
v_η, Λ, $\mathcal{V}$, $\lvert\cdot\rvert_\mu$	Jump increment, measure over possible jumps, right-hand side of the consistency equation, and weighted supremum norm

Table 5: Notation used in the appendix. The negative-binomial parameter is always denoted by $q(X_0)$; the symbol p_{data} is reserved for the clean-image distribution.

Model	Common density ratio	Direct predictions	Training from one level
Gaussian	Any image law, under Assumption D.1	Not treated here	Not treated here; see Remark F.1
Poisson	Any image law	Bounded rates and bounded images; alternatively, linearly growing rates	Bounded rates, Corollary F.3
Binomial	Any image law, with $B < 1$	Bounded counts; alternatively, linearly growing rates	Bounded rates on $B < 1$, Corollary F.3
Gamma	Joint Laplace representation and Assumption D.1	Bounded positive image values, Lipschitz dependence on the state, and a bound on large jumps	In addition, the integrability and Lipschitz conditions in Corollary F.5
Negative binomial	Any image law	$q_i(X_0) \le \bar{q} < 1$ and geometrically bounded posterior moments	Every moment index is covered, Corollary F.3

Table 6: Scope of the continuation results. These are exact statements about the true posterior predictions and expected losses. Starting the Gamma process from zero additionally requires the initialization assumption in Appendix C.1.4.

B Related work

Denoising from noisy data. Noise2Noise (Lehtinen et al., 2018) regresses one noisy measurement of an image on another and learns the posterior mean of the image given the input. Noisier2Noise and recorrupted-to-recorrupted training (Moran et al., 2020; Pang et al., 2021) obtain the two measurements from a single Gaussian one, data thinning and fission (Neufeld et al., 2024; Leiner et al., 2025) split a single Poisson, Gamma or negative-binomial measurement into independent parts, and generalized recorrupted-to-recorrupted training (Monroy et al., 2025) gives the split for exponential families together with its guarantee. Noise2Score (Kim & Ye, 2021) reaches the same posterior mean through Tweedie's formula. Blind-spot methods such as Noise2Void and Noise2Self (Krull et al., 2019; Batson & Royer, 2019) learn from single noisy images by predicting each pixel from the others, which requires noise that is independent across pixels. For radar speckle, MERLIN (Dalsasso et al., 2022) uses the real and imaginary parts of one acquisition as two views, and SAR2SAR (Dalsasso et al., 2021) applies Noise2Noise to acquisitions of the same scene at different dates. All of these methods train a denoiser at the noise level of the data. We use their splitting rules to build the N2N loss at every level the data can be degraded to, and for predictions other than the posterior mean.

Diffusion models from corrupted data. Ambient diffusion (Daras et al., 2023b), GSURE training (Kawar et al., 2024), SURE-Score (Aali et al., 2023) and expectation-maximization schemes (Rozet et al., 2024; Bai et al., 2024; Hosseintabar et al., 2025) train Gaussian diffusion models from linearly corrupted or noisy data, and Ambient Diffusion Omni (Daras et al., 2025) uses low-quality images to supervise the noisier levels of a Gaussian model. Daras et al. (2024) showed that data at one Gaussian noise level supervise the denoisers at all noisier levels, and that the consistency loss of Daras et al. (2023a) extends them to the cleaner ones. Levac et al. (2026) obtain the extension to lower noise levels from the scale equivariance of Gaussian denoisers. Inverse flows (Zhang & Zhou, 2025) and self-consistent stochastic interpolants (Modi et al., 2026) learn a transport from corrupted to clean data for continuous noise and for black-box corruptions. In all of these the sampling process is a Gaussian diffusion or a continuous interpolant. Our sampling process is the measurement process itself, so photon counts and speckled intensities are states of the process rather than outputs of a corruption channel, and a run started at a measurement samples from its posterior.

Generative processes beyond Gaussian noise. Binomial diffusion (Sohl-Dickstein et al., 2015), Blackout Diffusion (Santos et al., 2023), thinning and thickening of count images (Chen & Zhou, 2023), star-shaped and Beta diffusion (Okhotin et al., 2023; Zhou et al., 2023) and count bridges (Fishman et al., 2026) define generative processes through non-Gaussian laws, Generator Matching (Holderrieth et al., 2025) does so for arbitrary Markov processes, and stochastic localization (Montanari, 2023) gives the accumulation view of the Gaussian and Poisson processes that we use. Fishman et al. (2026) also train count bridges from aggregated counts by expectation maximization. The other methods are trained on clean data. The generative accumulation of photons (Krull et al., 2024) is trained on noisy microscopy frames by photon splitting and generates from an empty canvas, but it has no rule for the predictions at photon counts beyond those of the training frames. We use the Poisson and binomial processes of Chen & Zhou (2023) and Santos et al. (2023) and construct the Gamma and negative-binomial ones in the same way. What is new is the training from measurements at an interior level of the process, the noisy targets for predictions other than the mean, and the conditions under which consistency determines the continuation.

Self-consistency. Daras et al. (2023a) introduced the consistency loss to reduce sampling drift in Gaussian diffusion models trained on clean data, and proved that a consistent denoiser whose score is a gradient field is the denoiser of some image distribution. We keep the loss of Daras et al. (2023a) and estimate it from two simulated paths, through the product of their residuals or the square of their average (Table 7). When all predictions derive from one density ratio, the analog of their gradient-field condition, we prove that consistency and the correct predictions at one level identify the process for all five measurement laws (Appendix D). For the four jump laws we prove that consistency determines the continuation for rates predicted directly, with paths that start at the level of the data, which is how the models are trained.

C THE FIVE OBSERVATION MODELS

Let $X_0 \sim p_{\mathrm{data}}$ denote the clean image. Conditional on X_0, the measurement noise is independent across coordinates. We do not assume that the coordinates of X_0 themselves are independent.

We write e_i for the ith coordinate vector, $|n|_1 = \sum_i n_i$, $n! = \prod_i n_i!$, and $\mathbb{N} = \{0, 1, \ldots\}$. Gamma distributions are parameterized by shape and rate. Thus $\mathrm{Gamma}(L, 1/x)$ has shape L, rate $1/x$, and mean Lx.

We use a when we want to discuss the information level without committing to a particular measurement model. For the individual models, we use λ for Poisson exposure, p for binomial detection probability, L for Gamma shape, and r for negative-binomial shape. The information level can grow without bound in the Gaussian, Poisson, Gamma, and negative-binomial models. In the binomial model it instead approaches the maximum value $p = 1$. We therefore write $a_{\max} = 1$ for binomial observations and $a_{\max} = \infty$ for the other four models.

The degradation kernels introduced in the main text allow observations at different levels to be placed on one joint probability space. For example, if $a' < a < b$, we define the joint law of $(X_0, Z_b, Z_a, Z_{a'})$ by

$$p_{\mathrm{data}}(\mathrm{d}x) Q_b(x, \mathrm{d}z_b) \overleftarrow{K}_{b,a}(z_b, \mathrm{d}z_a) \overleftarrow{K}_{a,a'}(z_a, \mathrm{d}z_{a'}). \tag{9}$$

The identities

$$Q_b \overleftarrow{K}_{b,a} = Q_a, \qquad \overleftarrow{K}_{c,b} \overleftarrow{K}_{b,a} = \overleftarrow{K}_{c,a}$$

say that this construction is consistent: degrading from b to a produces the correct observation law at a, and degrading in one step or through an intermediate level gives the same result. We write

$$p_a = p_{\mathrm{data}} Q_a$$

for the marginal distribution of Z_a.

C.1 DENSITY RATIOS AND GENERATORS

A useful device in the proofs is to compare the actual observation process with a simple process that depends only on the measurement law and not on the unknown clean-image distribution. We call this the *reference process*. Let r_a denote its distribution at level a and let $\overrightarrow{K}^0_{a,b}$ denote its forward transition from a to b.

For a fixed clean image x, let

$$L^x_a(z) = \frac{\mathrm{d}Q_a(x\,\cdot)}{\mathrm{d}r_a}(z)$$

be the likelihood ratio of the observation law relative to this reference. Averaging this likelihood ratio over the unknown clean-image distribution gives

$$F_a(z) = \int L^x_a(z)\, p_{\mathrm{data}}(\mathrm{d}x).$$

The basic identities are

$$\begin{aligned} L^x_a(z) &= \frac{\mathrm{d}Q_a(x,\,\cdot)}{\mathrm{d}r_a}(z), \qquad F_a(z) = \int L^x_a(z)\, p_{\mathrm{data}}(\mathrm{d}x), \\ p_a(\mathrm{d}z) &= F_a(z) r_a(\mathrm{d}z), \qquad \pi_{a,z}(\mathrm{d}x) = \frac{L^x_a(z) p_{\mathrm{data}}(\mathrm{d}x)}{F_a(z)}. \end{aligned} \tag{10}$$

Thus F_a is the density ratio of the actual observation distribution with respect to the reference distribution. The measure $\pi_{a,z}$ is the posterior distribution of the clean image given $Z_a = z$. Whenever a ratio involving F_a appears below, it is understood only at states for which $F_a > 0$.

Lemma C.1. *Under the joint law in* (9)*, once Z_a is known, the observations at lower information levels contain no additional information about X_0. The observation process is Markov, and its forward transition satisfies*

$$\overrightarrow{K}_{a,b}(z, \mathrm{d}z') = \overrightarrow{K}^0_{a,b}(z, \mathrm{d}z') \frac{F_b(z')}{F_a(z)}, \qquad \overrightarrow{K}^0_{a,b} F_b = F_a. \tag{11}$$

For every test function for which the following expressions are defined, the generator of the observation process is

$$\mathcal{A}_a[h]f = \frac{\mathcal{A}^0(F_a f) - f\mathcal{A}^0 F_a}{F_a}. \tag{12}$$

Proof. By construction, all states below level a are obtained from Z_a through degradation kernels that do not depend on X_0. Therefore, once Z_a is known, observing the lower-level states cannot change the posterior distribution of X_0. For the same reason, the lower-level states are conditionally independent of X_0 and of all higher-level states given Z_a, so past and future are conditionally independent given the present, which is the Markov property. Bayes' rule then gives (10).

Proposition C.2 below shows that the actual observation process and the reference process have the same transitions when run backward from a higher information level to a lower one. Hence their two-level joint laws can be written as

$$p_b(\mathrm{d}z')\overleftarrow{K}_{b,a}(z', \mathrm{d}z)$$

and

$$r_b(\mathrm{d}z')\overleftarrow{K}_{b,a}(z', \mathrm{d}z),$$

respectively. Dividing the first joint law by the second and then conditioning on the state at level a gives the first identity in (11). Integrating that identity over z' gives $\overrightarrow{K}^0_{a,b} F_b = F_a$.

Finally,

$$\overrightarrow{K}_{a,b} f = \frac{\overrightarrow{K}^0_{a,b}(F_b f)}{F_a}.$$

Differentiating with respect to b at $b = a$ and using $\partial_a F_a = -\mathcal{A}^0 F_a$ gives (12). □

C.1.1 Gaussian observations

For $X_0 \in \mathbb{R}^d$, define

$$Z_a = aX_0 + W_a, \qquad Q_a(x, \cdot) = \mathcal{N}(ax, aI_d),$$

where W is Brownian motion independent of X_0.

The natural reference process is Brownian motion itself. Thus

$$r_a = \mathcal{N}(0, aI_d), \qquad \mathcal{A}^0 = \tfrac{1}{2}\Delta.$$

The density ratio is

$$F_a(z) = \mathbb{E}\exp\Big(z \cdot X_0 - \frac{a}{2}\|X_0\|^2\Big).$$

Differentiating its logarithm gives

$$m_a(z) = \nabla \log F_a(z) = \mathbb{E}[X_0 \mid Z_a = z].$$

Therefore the posterior quantity required by the Gaussian dynamics is exactly the posterior mean of the clean image. The generator is

$$\mathcal{A}_a[m]f = m_a \cdot \nabla f + \tfrac{1}{2}\Delta f.$$

For every $a > 0$, the factor $\exp(-a|X_0|^2/2)$ controls the large values of X_0. As a result, the displayed expectation and the derivatives needed below are locally finite in (a, z) for any probability distribution p_{data}.

The relation with the usual Gaussian score follows immediately. Since

$$\nabla \log p_a(z) = m_a(z) - z/a,$$

the normalized observation $Y = Z_a/a$ satisfies Tweedie's identity (Robbins, 1956; Efron, 2011)

$$\nabla_y \log p_Y(y) = a\big(\mathbb{E}[X_0 \mid Y = y] - y\big).$$

C.1.2 POISSON OBSERVATIONS

For $X_0 \in [0,\infty)^d$, let $N_{\lambda,i}$ be conditionally independent Poisson processes with rates $X_{0,i}$. Conditional on $X_0 = x$,

$$Q_\lambda(x, \{n\}) = \prod_i e^{-\lambda x_i} \frac{(\lambda x_i)^{n_i}}{n_i!}.$$

As reference process, we use independent unit-rate Poisson processes. Its distribution is

$$r_\lambda(n) = e^{-d\lambda} \frac{\lambda^{|n|_1}}{n!}.$$

The corresponding density ratio is

$$F_\lambda(n) = e^{d\lambda} \mathbb{E}\left[\prod_i X_{0,i}^{n_i} e^{-\lambda X_{0,i}}\right].$$

The reference generator is

$$\mathcal{A}^0 f(n) = \sum_i [f(n+e_i) - f(n)].$$

Using the ratio of F_λ at two neighboring count states gives

$$\begin{aligned} h_{i,\lambda}(n) &= \frac{F_\lambda(n+e_i)}{F_\lambda(n)} = \mathbb{E}[X_{0,i} \mid N_\lambda = n] = \frac{n_i+1}{\lambda} \frac{p_\lambda(n+e_i)}{p_\lambda(n)}, \\ \mathcal{A}_\lambda[h] f(n) &= \sum_i h_{i,\lambda}(n)[f(n+e_i) - f(n)]. \end{aligned} \tag{13}$$

Thus the jump rate in coordinate i is the posterior mean of the clean Poisson intensity in that coordinate.

At any positive exposure, the factor $e^{-\lambda X_{0,i}}$ makes these posterior rates finite on every state that can actually occur, without requiring moment assumptions on the clean-image distribution. This is the count-thickening process used by Chen & Zhou (2023).

C.1.3 BINOMIAL OBSERVATIONS

Here the clean image is integer valued: $X_0 \in \mathbb{N}^d$. A useful construction is to think of $X_{0,i}$ as the number of particles in coordinate i. Give each particle an independent arrival time uniformly distributed on $[0,1]$, and let $N_{p,i}$ be the number that have arrived by time p. Then

$$N_p \mid X_0 \sim \text{Binomial}(X_0, p)$$

coordinatewise.

Write

$$(x)_n = \frac{x!}{(x-n)!}$$

for $x \geq n$, and set $(x)_n = 0$ for $x < n$. Relative to the unit-rate Poisson reference,

$$F_p(n) = e^{dp} \mathbb{E}\left[\prod_i (X_{0,i})_{n_i} (1-p)^{X_{0,i}-n_i}\right], \qquad 0 < p < 1.$$

There are two closely related posterior quantities. The first is the posterior mean of the total clean count,

$$D_{i,p}(n) = \mathbb{E}[X_{0,i} \mid N_p = n].$$

The second is the rate at which new particles arrive. Since n_i particles have already arrived, this rate is

$$g_{i,p}(n) = \frac{D_{i,p}(n) - n_i}{1-p} = \frac{F_p(n+e_i)}{F_p(n)}. \tag{14}$$

The denominator $1-p$ appears because each particle that has not yet arrived has instantaneous arrival rate $1/(1-p)$. Averaging the image-dependent rate

$$\frac{X_{0,i} - n_i}{1-p}$$

over the posterior distribution of X_0 gives $g_{i,p}(n)$.

The generator is therefore

$$\mathcal{A}_p[D]f(n) = \sum_i g_{i,p}(n)[f(n+e_i) - f(n)].$$

It will be important below to distinguish the posterior count D_i from the birth rate g_i. For an arbitrary family of birth rates u_i, define

$$D_i(n,p) = n_i + (1-p)u_i(n,p)$$

and

$$\mathcal{A}[u]f = \sum_i u_i[f(\cdot + e_i) - f].$$

Since $\mathcal{A}[u]n_i = u_i$,

$$(\partial_p + \mathcal{A}[u])D_i = (1-p)(\partial_p + \mathcal{A}[u])u_i. \tag{15}$$

Hence, away from $p = 1$, imposing self-consistency on the posterior counts is equivalent to imposing it on the birth rates. The reverse, thinning direction of this process is the one used by Blackout Diffusion (Santos et al., 2023).

C.1.4 Gamma observations

For $X_0 \in (0,\infty)^d$, let $G_{L,i}$ be independent rate-one Gamma processes, independent of X_0, and define

$$S_{L,i} = X_{0,i}G_{L,i}.$$

Here a rate-one Gamma process is an increasing process whose increment over an interval of length ΔL has distribution $\mathrm{Gamma}(\Delta L, 1)$. Conditional on $X_0 = x$,

$$Q_L(x, \mathrm{d}s) = \prod_i \frac{s_i^{L-1}e^{-s_i/x_i}}{\Gamma(L)x_i^L}\,\mathrm{d}s_i, \qquad s \in (0,\infty)^d.$$

The reference process has independent rate-one Gamma coordinates. Its generator is

$$\mathcal{A}^0 f(s) = \sum_i \int_0^\infty [f(s+we_i) - f(s)]e^{-w}\frac{\mathrm{d}w}{w}. \tag{16}$$

It is convenient to write the density ratio as

$$F_L(s) = e^{\sum_i s_i}U_L(s), \qquad U_L(s) = \mathbb{E}\left[\prod_i X_{0,i}^{-L}e^{-s_i/X_{0,i}}\right]. \tag{17}$$

For every $L > 0$ and every state with $s_i > 0$, $U_L(s)$ is finite and positive for any distribution supported on finite positive image values. Indeed, for fixed $s_i > 0$, the function $x^{-L}e^{-s_i/x}$ is bounded on $(0,\infty)$.

Unlike the Gaussian and Poisson cases, the Gamma process does not require only a posterior mean. A jump of size w in coordinate i occurs at a rate determined by

$$H_{i,L}(s,w) = \mathbb{E}[e^{-w/X_{0,i}} \mid S_L = s] = e^{-w}\frac{F_L(s+we_i)}{F_L(s)}. \tag{18}$$

The generator is

$$\mathcal{A}_L[H]f(s) = \sum_i \int_0^\infty [f(s+we_i) - f(s)]H_{i,L}(s,w)\frac{\mathrm{d}w}{w}. \tag{19}$$

These same posterior Laplace transforms appear in the Gamma Lévy channel of Jiao et al. (2017). The normalized measurement has

$$\mathbb{E}[S_L/L \mid X_0] = X_0, \qquad \mathrm{Var}(S_{L,i}/L \mid X_0) = \frac{X_{0,i}^2}{L}.$$

Thus S_L/L becomes increasingly concentrated around X_0 as L increases.

The Gamma process has infinitely many small jumps on every interval, but their total size is finite. At every positive state,

$$\begin{aligned}\int_0^1 H_{i,L}(s,w)\,\mathrm{d}w &\le 1,\\ \int_1^\infty H_{i,L}(s,w)\frac{\mathrm{d}w}{w} &\le 1+\mathbb{E}[\log(1+X_{0,i})\mid S_L=s]<\infty.\end{aligned} \tag{20}$$

For the second inequality we use

$$\int_1^\infty e^{-w/x}\frac{\mathrm{d}w}{w}\le 1+\log(1+x),$$

together with the fact that $x^{-L}e^{-s_i/x}\log(1+x)$ is bounded when $L>0$ and $s_i>0$.

There is one additional issue near $L=0$. The formulas above are written for states in $(0,\infty)^d$, whereas the completely empty observation is $S_0=0$, which lies on the boundary of this state space. For every $L>0$, however, S_L lies in $(0,\infty)^d$ almost surely.

Conditional on $X_0=x$, the true process is simply

$$S_{L,i}=x_iG_{L,i}.$$

The unconditional process is obtained by averaging over $x\sim p_{\mathrm{data}}$. A theorem that identifies the process only from some positive level $a_*>0$ therefore says nothing by itself about how a learned process enters the interior from $L=0$.

Assumption C.1. When generation is started from $S_0=0$, the learned Gamma process is initialized with the correct family of small-positive-level distributions: for every $\varepsilon>0$, its state at level ε has law $p_\varepsilon=p_{\mathrm{data}}Q_\varepsilon$, and these initializations are consistent as $\varepsilon\downarrow 0$.

This is an assumption about initialization. It does not follow from a continuation theorem whose starting point is a fixed positive level. Appendix H.2 describes the initialization used in the experiments and how it departs from this assumption. For the count models, zero is an ordinary state of the process, so their initialization can be handled directly in Appendix F.4.

C.1.5 Negative-binomial observations

Let $q=q(X_0)\in[0,1)^d$ be a known encoding of the clean image. Conditional on X_0, coordinate i has jumps of size $k\ge 1$ at rate $\frac{q_i^k}{k}$. The total jump rate in that coordinate is

$$\sum_{k\ge 1}\frac{q_i^k}{k}=-\log(1-q_i).$$

At shape r, the conditional observation law is

$$Q_r(x,\{n\})=\prod_i\frac{\Gamma(n_i+r)}{n_i!\,\Gamma(r)}q_i(x)^{n_i}(1-q_i(x))^r.$$

Choose a fixed $q_0\in(0,1)$ for the reference process. Its marginal is

$$r_r(n)=\prod_i\frac{\Gamma(n_i+r)}{n_i!\,\Gamma(r)}q_0^{n_i}(1-q_0)^r,$$

and the density ratio is

$$F_r(n)=\frac{\mathbb{E}[\prod_i q_i(X_0)^{n_i}(1-q_i(X_0))^r]}{q_0^{|n|_1}(1-q_0)^{dr}}.$$

The reference generator, posterior moments, and true generator are

$$\begin{aligned}\mathcal{A}^0 f(n) &= \sum_{i=1}^{d} \sum_{k\geq 1} \frac{q_0^k}{k}[f(n+ke_i) - f(n)],\\ M_{i,k,r}(n) &= \mathbb{E}[q_i(X_0)^k \mid N_r = n] = q_0^k \frac{F_r(n+ke_i)}{F_r(n)},\\ \mathcal{A}_r[M] f(n) &= \sum_{i=1}^{d} \sum_{k\geq 1} \frac{M_{i,k,r}(n)}{k}[f(n+ke_i) - f(n)].\end{aligned} \tag{21}$$

Thus an exact sampler needs the full sequence of posterior moments $M_{i,k,r}$, not only the first one. These moments also appear in the negative-binomial Lévy channel of Jiao et al. (2017).

At every positive shape r, the total posterior jump rate is finite. To see this, note that

$$(1-q)^r |\log(1-q)|$$

is bounded on $[0, 1)$. Conditional on a fixed image, the process has finitely many jumps on every finite interval, and this remains true after averaging over the clean-image distribution.

C.2 Degradation kernels and identification of the process

We now describe what happens when an observation is moved in the direction of *less* information. The key fact is that these degradation rules do not depend on the clean image. They depend only on the measurement law.

A second key fact is that degradation from a positive level does not erase all information about the distribution at the higher level. More precisely, if two distributions at level b become the same after degradation to a positive level a, then the original distributions at level b were already the same. This is the injectivity property used later.

Proposition C.2. *Let $0 < a < b < a_{\max}$. The actual observation process and the reference process have the same backward transition $\overleftarrow{K}_{b,a}$. Coordinatewise,*

$$\begin{aligned}&\textit{Gaussian:} && Z_a \mid Z_b = z \sim \mathcal{N}((a/b)z, a(1-a/b)),\\ &\textit{Poisson and binomial:} && N_a \mid N_b = n \sim \mathrm{Binomial}(n, a/b),\\ &\textit{Gamma:} && S_a \mid S_b = s \overset{\mathrm{d}}{=} sV, \quad V \sim \mathrm{Beta}(a, b-a),\\ &\textit{Negative binomial:} && N_a \mid N_b = n \sim \mathrm{BetaBinomial}(n; a, b-a).\end{aligned} \tag{22}$$

The coordinates are conditionally independent in each of these kernels. Moreover, degradation is injective on probability distributions:

$$\nu_1 \overleftarrow{K}_{b,a} = \nu_2 \overleftarrow{K}_{b,a} \quad \Longrightarrow \quad \nu_1 = \nu_2.$$

Equivalently, the map $\nu \mapsto \nu \overleftarrow{K}_{b,a}$ is one-to-one.

Proof. We first derive the backward kernels.

For Gaussian observations,

$$Z_a - \frac{a}{b} Z_b = W_a - \frac{a}{b} W_b.$$

The quantity on the right is independent of (X_0, W_b) and is Gaussian with covariance $a(1-a/b)I_d$. This gives the first line of (22).

For Poisson and binomial observations, condition on the fact that n particles have been observed by level b. Their arrival times are independent and uniform on $[0, b]$. Each particle was therefore already present by level a with probability a/b, which gives binomial thinning.

For Gamma observations, write the level-b state as the sum of the independent increments from 0 to a and from a to b. The ratio of a Gamma variable with shape a to the sum of independent Gamma

variables with shapes a and $b-a$ has distribution $\mathrm{Beta}(a, b-a)$ and is independent of the sum. This gives $S_a = S_b V$ conditionally on S_b.

For negative-binomial observations, the increment from a to b is independent of the state at a conditional on the image. Dividing the product of the masses of independent $\mathrm{NB}(a, q)$ and $\mathrm{NB}(b-a, q)$ variables by the mass of their sum gives

$$\mathbb{P}(N_a = j \mid N_b = n) = \binom{n}{j} \frac{\mathrm{B}(a+j, b-a+n-j)}{\mathrm{B}(a, b-a)}.$$

None of these conditional distributions contains the clean image or its parameters. They therefore apply equally to the actual observation process and to the reference process. Applying the same conditioning at three levels also gives the composition property of the degradation kernels.

We next prove injectivity. The argument is different for the four types of noise, but the idea is the same: from the degraded distribution we recover a transform that uniquely determines the original distribution.

For Gaussian observations, let $\widehat{\nu}$ denote the characteristic function of the distribution at level b. Gaussian degradation evaluates this characteristic function at $(a/b)t$ and multiplies it by a Gaussian factor that is never zero. Hence the original characteristic function can be recovered.

For Poisson and binomial observations, let ϕ be the probability generating function of the law at level b, and set $\alpha = a/b$. After binomial thinning, the generating function becomes

$$\phi(1 - \alpha + \alpha z).$$

Knowing this function determines ϕ on an open interval inside $(0, 1)$. Since a probability generating function is analytic in its domain, this determines ϕ everywhere.

For Gamma observations, take logarithms:

$$\log S_a = \log S_b + \log V.$$

The characteristic function of $\log V$ is

$$\mathbb{E}[V^{it}] = \frac{\Gamma(a+it)\Gamma(b)}{\Gamma(a)\Gamma(b+it)}.$$

The Gamma function has no zeros, so this factor is never zero. Therefore the characteristic function, and hence the distribution, of $\log S_b$ can be recovered from that of $\log S_a$.

For negative-binomial observations, put $c = b - a$ and denote by ϕ_b and ϕ_a the generating functions before and after degradation. For $0 < t < 1$,

$$\phi_a(1-t) = \frac{t^{1-b}\Gamma(c)}{\mathrm{B}(a, c)} \mathcal{I}^c \left[v^{a-1} \phi_b(1-v) \right](t),$$

where

$$\mathcal{I}^c f(t) = \frac{1}{\Gamma(c)} \int_0^t (t-v)^{c-1} f(v)\, \mathrm{d}v.$$

The operator $\mathcal{I}^c$ is a fractional integral. What matters here is that it is one-to-one on locally integrable functions. Indeed, choose an integer $m > c$, apply $\mathcal{I}^{m-c}$, and then differentiate m times. The function to which it is applied is locally integrable because $a > 0$ and $|\phi_b(1-v)| \le 1$. Hence ϕ_b is uniquely recovered.

For multivariate observations, the kernels act independently on the coordinates. For the Gaussian and Gamma kernels, the joint transform factorizes, and the univariate argument applies to each factor. For thinning, the joint generating function is analytic on the open unit polydisc and known on a product of real intervals, which determines it by the identity theorem in several variables. For the negative-binomial kernel, the fractional-integral representation is applied in each coordinate in turn with the others held fixed. None of these arguments requires moment assumptions on the clean-image distribution. □

The restriction $a > 0$ is essential. At level zero, the observation is empty, so every possible higher-level distribution degrades to the same state. Degradation from zero is therefore not injective.

Theorem C.3. *Fix $0 < a_* < B < a_{\max}$. Let $(\widehat{Z}_a)_{a\in[a_*,B]}$ be a Markov process whose law at a_* is the true marginal p_{a_*}. Suppose that, for every $a_* \le a < b \le B$, its backward transition from b to a is the known degradation kernel $\overleftarrow{K}_{b,a}$. Then $(\widehat{Z}_a)$ has the same finite-dimensional distributions as the true observation process.*

Proof. Let $\widehat{p}_b$ denote the marginal law of $\widehat{Z}_b$. Since the process has the prescribed backward transition,

$$\widehat{p}_b \overleftarrow{K}_{b,a_*} = p_{a_*}.$$

The true process satisfies the same relation:

$$p_b \overleftarrow{K}_{b,a_*} = p_{a_*}.$$

By the injectivity in Proposition C.2,

$$\widehat{p}_b = p_b.$$

Thus the two processes have the same marginal distribution at every level. Now consider any finite collection of levels. Starting from the marginal at the largest level and repeatedly applying the common backward kernels determines the entire joint law. Since the candidate and true processes have the same final marginal and the same backward kernels, their joint laws agree at every finite collection of levels. □

C.3 Regression targets

We now explain how the posterior predictions required by the dynamics can be learned without observing X_0. Let Y_1 and Y_2 be two measurements of the same clean image, independent conditional on X_0. Suppose their information levels are a_{in} and a_{tar}, respectively. By degrading Y_1, we can construct an input Z_a at any level $0 < a \le a_{\text{in}}$. Let

$$W = (a, Z_a, \xi),$$

where ξ specifies which posterior quantity is being learned. The level a and the index ξ are sampled independently of the clean image and of the noise in the target view. Suppose that a transformation of the second measurement,

$$T = T_\xi(Y_2),$$

has the property

$$\mathbb{E}[T \mid X_0, W] = \psi_\xi(X_0).$$

In words, T is noisy, but its conditional mean is exactly the clean quantity whose posterior expectation we need. Conditional independence and the tower property then give

$$\mathbb{E}[T \mid W] = \mathbb{E}[\psi_\xi(X_0) \mid W] = h_{\xi,a}(Z_a).$$

Thus regression onto T learns the required posterior prediction.

Lemma C.4. *Assume $\mathbb{E}|T|^2 < \infty$. For every square-integrable predictor $\widehat{h}(W)$,*

$$\mathbb{E}|T - \widehat{h}(W)|^2 = \mathbb{E}|T - h(W)|^2 + \mathbb{E}|\widehat{h}(W) - h(W)|^2, \qquad h(W) = \mathbb{E}[T \mid W]. \tag{23}$$

Consequently, h is the unique minimizer of the expected squared loss, up to events of probability zero. The same statement holds for vector targets with the squared Euclidean norm.

Proof. Write

$$T - \widehat{h} = (T - h) + (h - \widehat{h})$$

and expand the square. The cross term vanishes after conditioning on W because

$$\mathbb{E}[T - h \mid W] = 0.$$

This leaves exactly the two terms on the right-hand side of (23). □

For Gaussian, Poisson, and binomial observations, unbiased targets for the posterior mean are, respectively,

$$\frac{Y_2}{a_{\text{tar}}}, \qquad \frac{Y_2}{\lambda_{\text{tar}}}, \qquad \frac{Y_2}{p_{\text{tar}}}.$$

Their coordinatewise conditional variances are

$$\frac{1}{a_{\text{tar}}}, \qquad \frac{X_{0,i}}{\lambda_{\text{tar}}}, \qquad \frac{X_{0,i}(1 - p_{\text{tar}})}{p_{\text{tar}}}.$$

To ensure that the total expected regression loss is finite, these quantities must also be integrable over the distribution of images and over the information levels sampled during training.

C.3.1 GAUSSIAN REGRESSION AND DOUBLE TWEEDIE

In the Gaussian case, the noisy target can also be constructed from a single noisy image. Let

$$Y = X_0 + \sigma_{\mathrm{obs}}\varepsilon,$$

and, for a larger noise level $\sigma > \sigma_{\mathrm{obs}}$, form

$$\widetilde{Y} = Y + \sqrt{\sigma^2 - \sigma_{\mathrm{obs}}^2}\,\varepsilon',$$

where ε and ε' are independent standard Gaussian vectors. Set

$$\alpha = \frac{\sigma_{\mathrm{obs}}^2}{\sigma^2}.$$

A direct Gaussian-conditioning calculation gives

$$\mathbb{E}[Y \mid X_0, \widetilde{Y}] = (1-\alpha)X_0 + \alpha\widetilde{Y}.$$

Therefore

$$T = \frac{Y - \alpha\widetilde{Y}}{1-\alpha}$$

satisfies

$$\mathbb{E}[T \mid X_0, \widetilde{Y}] = X_0.$$

It is therefore a valid noisy target for the clean posterior mean. For a predictor f,

$$\mathbb{E}\big|(1-\alpha)f(\widetilde{Y}) + \alpha\widetilde{Y} - Y\big|^2 = (1-\alpha)^2\mathbb{E}|f(\widetilde{Y}) - T|^2. \tag{24}$$

This is the noisy-target form of the double-Tweedie objective of Daras et al. (2024). At a fixed noise level, the two losses differ only by the positive factor $(1-\alpha)^2$ and therefore have the same minimizer of the expected loss. If different noise levels are mixed in one objective, this factor changes how much each level is weighted, but it does not change the optimal predictor at any individual level when that predictor can be represented by the model. The information coordinate used in our observation process is

$$a = \sigma^{-2}, \qquad Z_a = a\widetilde{Y}.$$

The formula above requires $\sigma > \sigma_{\mathrm{obs}}$. At equality, one either takes a limit or uses an independent target measurement.

C.3.2 GAMMA TARGETS AND THEIR LOSS

For the Gamma process, the required posterior quantity is

$$H_{i,L}(s,w) = \mathbb{E}[e^{-w/X_{0,i}} \mid S_L = s].$$

We therefore need a noisy target whose conditional expectation is $e^{-w/X_{0,i}}$. For a target observation with shape $r > 0$ and for a queried jump size $w \geq 0$, define

$$T_{w,r}(v) = \begin{cases} (1 - w/v)^{r-1}, & v > w, \\ 0, & 0 < v \leq w. \end{cases}$$

Here $r = L_{\mathrm{tar}}$ is the shape of one coordinate of the target measurement. Notice that $T_{0,r} = 1$, while for $r = 1$,

$$T_{w,1}(v) = \mathbf{1}_{\{v>w\}}.$$

Proposition C.5. *If $V \mid X_0 = x \sim \mathrm{Gamma}(r, 1/x)$, then*

$$\mathbb{E}[T_{w,r}(V) \mid x] = e^{-w/x}$$

for every $r > 0$. For a fixed $w > 0$, the conditional second moment of $T_{w,r}(V)$ is finite if and only if $r > 1/2$. When $r > 1/2$, writing $\vartheta = w/x$,

$$\mathrm{Var}(T_{w,r}(V) \mid x) = \frac{e^{-\vartheta}}{\Gamma(r)} \int_0^\infty u^{2r-2}(u+\vartheta)^{1-r}e^{-u}\,\mathrm{d}u - e^{-2\vartheta}. \tag{25}$$

For every fixed $r > 1/2$, the second moment is bounded uniformly over all $x > 0$ and $w \geq 0$.

Proof. Multiplying $T_{w,r}$ by the Gamma density simplifies the power of v:

$$\mathbb{E}[T_{w,r}(V) \mid x] = \frac{1}{\Gamma(r)x^r} \int_w^\infty (v-w)^{r-1} e^{-v/x} \, \mathrm{d}v.$$

With the change of variable $u = v - w$, this becomes $e^{-w/x}$, which proves unbiasedness. For the second moment, set $v = w + xu$. After this substitution, the integrand in (25) behaves like u^{2r-2} as $u \downarrow 0$ whenever $w > 0$. This is integrable at zero exactly when

$$2r - 2 > -1,$$

or equivalently $r > 1/2$. If $r \geq 1$, then $0 \leq T_{w,r} \leq 1$, so the second moment is immediately bounded. If $1/2 < r < 1$, use $(u+\vartheta)^{1-r} \leq u^{1-r} + \vartheta^{1-r}$ to obtain

$$\mathbb{E}[T_{w,r}(V)^2 \mid x] \leq e^{-\vartheta} \left[1 + \frac{\Gamma(2r-1)}{\Gamma(r)} \vartheta^{1-r} \right].$$

The right-hand side is uniformly bounded in $\vartheta \geq 0$. At $w = 0$, the target is identically equal to one. □

When $r \leq 1/2$, squared-error regression is not appropriate because the target has infinite second moment for $w > 0$. We instead use the I-divergence

$$D_{\mathrm{I}}(t, u) = t \log(t/u) - t + u,$$

with the convention $0 \log 0 = 0$.

Proposition C.6. *Let*

$$T = T_{w,r}(V), \qquad m = \mathbb{E}[T \mid W] \in (0, 1],$$

and let $\widehat{h}(W)$ be finite and positive. Then

$$\mathbb{E} D_{\mathrm{I}}(T, \widehat{h}) = \mathbb{E} D_{\mathrm{I}}(T, m) + \mathbb{E} D_{\mathrm{I}}(m, \widehat{h}). \tag{26}$$

For every fixed $r > 0$, the optimal expected loss is finite, and m is the unique minimizer up to events of probability zero.

Proof. We first check that the expected loss at the optimum is finite. When $r \geq 1$, $T \in [0, 1]$, so

$$|T \log T| \leq 1/e.$$

When $0 < r < 1$, the same change of variable as above gives, with $G \sim \mathrm{Gamma}(r, 1)$,

$$\mathbb{E}[T \log T \mid x] = (1-r) e^{-\vartheta} \mathbb{E} \log(1 + \vartheta/G).$$

Using

$$\log(1 + \vartheta/G) \leq \log(1+\vartheta) + \log(1 + 1/G),$$

we obtain the bound

$$(1-r)(e^{-1} + C_r), \qquad C_r = \mathbb{E} \log(1 + 1/G) < \infty.$$

Hence $\mathbb{E}|T \log T| < \infty$ uniformly over x and w. Conditioning on W then gives

$$\mathbb{E} D_{\mathrm{I}}(T, m) = \mathbb{E}[T \log T] - \mathbb{E}[m \log m] < \infty.$$

Expanding the conditional loss around $m = \mathbb{E}[T \mid W]$ gives (26). Since $D_{\mathrm{I}}(m, \widehat{h}) \geq 0$ and is zero exactly when $\widehat{h} = m$, the minimizer is unique. □

C.3.3 NEGATIVE-BINOMIAL MOMENT TARGETS

For the negative-binomial process, the exact jump rates require the posterior moments

$$\mathbb{E}[q_i(X_0)^k \mid N_r].$$

The following target learns these moments directly. For target shape $r > 0$ and integer $k \geq 1$, define

$$T_{k,r}(v) = \mathbf{1}_{\{v \geq k\}} \prod_{j=0}^{k-1} \frac{v-j}{v+r-1-j}. \tag{27}$$

If $V \mid q \sim \mathrm{NB}(r, q)$, then $\mathbb{E}[T_{k,r}(V) \mid q] = q^k$. Moreover,

$$0 \leq T_{k,r}(v) \leq C_{k,r} := \max\left\{1, \frac{k!\,\Gamma(r)}{\Gamma(k+r)}\right\}.$$

To verify unbiasedness, multiply the target by the negative-binomial mass. The factors in the product cancel, leaving

$$\sum_{v \geq k} \frac{\Gamma(v-k+r)}{(v-k)!\,\Gamma(r)} q^v (1-q)^r = q^k.$$

When $r \geq 1$, every factor in (27) is at most one. When $r < 1$, the product decreases as v increases beyond k, so its maximum occurs at $v = k$. This proves the stated bound and, in particular, shows that squared loss is finite for every fixed k and $r > 0$. If the moment index itself is sampled from an unbounded set, say $K \in \{1, 2, \ldots\}$, a sufficient condition for finite total regression risk is

$$\mathbb{E}[C_{K,r}^2] < \infty.$$

The special case $k = 1$ learns

$$\mathbb{E}[q_i(X_0) \mid N_r].$$

This should not be confused with the normalized count V/r, whose conditional mean is

$$\frac{q}{1-q}.$$

If the encoding is linear, $q_i(X_0) = cX_{0,i}$, then $T_{1,r}/c$ is an unbiased target for the clean image coordinate. Applying a nonlinear transformation to V/r does not in general produce an unbiased image target.

C.4 SPLITTING A SINGLE MEASUREMENT

The previous discussion assumed that two conditionally independent noisy views Y_1 and Y_2 are available. For Gaussian, Poisson, Gamma, and negative-binomial observations, two such views can also be created by randomly splitting a single acquisition. These are instances of Gaussian recorruption and data thinning (Pang et al., 2021; Neufeld et al., 2024; Leiner et al., 2025).

Proposition C.7. *Let*

$$a_{\mathrm{in}}, a_{\mathrm{tar}} > 0, \qquad a_{\mathrm{in}} + a_{\mathrm{tar}} = a_{\mathrm{acq}},$$

and suppose

$$Y_{\mathrm{acq}} \mid X_0 \sim Q_{a_{\mathrm{acq}}}(X_0, \cdot).$$

The following coordinatewise constructions produce two observations Y_1 and Y_2 that are independent conditional on X_0, with

$$Y_1 \sim Q_{a_{\mathrm{in}}}(X_0, \cdot), \qquad Y_2 \sim Q_{a_{\mathrm{tar}}}(X_0, \cdot):$$

$$\begin{aligned}
\textit{Gaussian:} \quad & Y_1 = \frac{a_{\mathrm{in}}}{a_{\mathrm{acq}}} Y_{\mathrm{acq}} + \sqrt{\frac{a_{\mathrm{in}} a_{\mathrm{tar}}}{a_{\mathrm{acq}}}}\,\varepsilon, && Y_2 = Y_{\mathrm{acq}} - Y_1, \\
\textit{Poisson:} \quad & Y_1 \mid Y_{\mathrm{acq}} \sim \mathrm{Binomial}(Y_{\mathrm{acq}}, a_{\mathrm{in}}/a_{\mathrm{acq}}), && Y_2 = Y_{\mathrm{acq}} - Y_1, \\
\textit{Gamma:} \quad & Y_1 = V Y_{\mathrm{acq}}, \quad V \sim \mathrm{Beta}(a_{\mathrm{in}}, a_{\mathrm{tar}}), && Y_2 = (1-V) Y_{\mathrm{acq}}, \\
\textit{Negative binomial:} \quad & Y_1 \mid Y_{\mathrm{acq}} \sim \mathrm{BetaBinomial}(Y_{\mathrm{acq}}; a_{\mathrm{in}}, a_{\mathrm{tar}}), && Y_2 = Y_{\mathrm{acq}} - Y_1.
\end{aligned}$$

Here $\varepsilon \sim \mathcal{N}(0, I_d)$. All auxiliary random variables are independent across coordinates and independent of X_0.

Proof. For Gaussian observations, a direct calculation shows that, conditional on X_0, Y_1 and Y_2 have the required means and variances and have zero cross-covariance. Since they are jointly Gaussian, zero cross-covariance implies independence. For the other three models, conditional on X_0 the observation at level a_{acq} can be written as a sum of two independent observations at levels a_{in} and a_{tar}:

$$Q_{a_{\mathrm{acq}}} = Q_{a_{\mathrm{in}}} * Q_{a_{\mathrm{tar}}}.$$

The splitting rule displayed above is exactly the conditional distribution of the first summand given their sum, as derived in Proposition C.2. Reversing this conditioning recovers the independent pair. □

There is a practical tradeoff in how the split is chosen. If a Poisson acquisition at level a_{acq} is split so that the input has level $a < a_{\mathrm{acq}}$, then the target has level $a_{\mathrm{acq}} - a$. Its normalized variance is

$$\frac{X_0}{a_{\mathrm{acq}} - a},$$

which diverges as a approaches the acquisition level. Thus inputs very close to the original acquisition leave very little information for the independent target view. For binomial observations, an analogous split does not produce conditionally independent views of the clean count. We therefore use independent acquisitions for that model.

C.5 Reachable states and clean limits

For the count models, not every count vector needs to have positive probability under the clean-image distribution. It is useful to identify the set of states that can actually occur.

Proposition C.8. *For Poisson, binomial, and negative-binomial observations, define*

$$S = \{n \in \mathbb{N}^d : p_a(n) > 0\}.$$

This set does not depend on the positive information level $a \in (0, a_{\max})$. For binomial observations,

$$n \in S \quad \Longleftrightarrow \quad p_{\mathrm{data}}\{X_0 \geq n\} > 0,$$

where the inequality is coordinatewise. For Poisson observations, $n \in S$ exactly when the image distribution assigns positive probability to images satisfying $X_{0,i} > 0$ for every coordinate with $n_i > 0$. For negative-binomial observations, the same statement holds with $q_i(X_0) > 0$ in place of $X_{0,i} > 0$. The set S contains the zero vector and is downward closed: if $n \in S$ and $0 \leq m \leq n$ coordinatewise, then $m \in S$. The true density ratio F_a is positive on S and zero outside S. If the reference process is stopped when it attempts to leave S, its transform by F and its backward transitions inside S are unchanged.

Proof. The characterization of S follows directly from the conditional likelihoods. At every positive level, those likelihoods are positive for exactly the states described in the proposition. Hence the support does not depend on the information level. Each support is downward closed because lowering one or more counts cannot turn a possible state into an impossible one. The zero state has positive conditional probability for every admissible clean image. The reference mass $r_a(n)$ is strictly positive on all of $\mathbb{N}^d$. Therefore

$$F_a(n) = \frac{p_a(n)}{r_a(n)}$$

is positive exactly on S. Reference count processes move only upward coordinatewise. Hence, if a reference path ends in the downward-closed set S, it could not have left S at an earlier time. Stopping the reference process when it exits S therefore does not change paths that remain in S, or the transition probabilities between states in S. Formula (11) then gives the same transformed process on S. □

We finally record the limit in which the observation becomes clean.

Proposition C.9. *For every admissible clean-image distribution, the observation processes satisfy, almost surely,*

$$\frac{Z_a}{a} \longrightarrow X_0, \qquad \frac{N_\lambda}{\lambda} \longrightarrow X_0, \qquad N_p \longrightarrow X_0, \\ \frac{S_L}{L} \longrightarrow X_0, \qquad \frac{N_r}{r+N_r} \longrightarrow q(X_0). \tag{28}$$

The limits are taken as $a, \lambda, L, r \to \infty$ and $p \uparrow 1$. The last ratio is understood coordinatewise.

Proof. For the Gaussian process,

$$\frac{Z_a}{a} = X_0 + \frac{W_a}{a},$$

and Brownian motion satisfies $W_a/a \to 0$ almost surely. Conditional on a fixed clean image, the usual strong law for a Poisson process gives

$$N_\lambda/\lambda \to X_0.$$

The same argument for the Gamma process gives

$$S_L/L \to X_0.$$

In the binomial construction, each coordinate contains only finitely many particles. Every particle has an arrival time strictly below one almost surely, so

$$N_p \to X_0$$

as $p \uparrow 1$. For the negative-binomial process, the expected count added per unit shape in coordinate i is

$$\sum_{k\geq 1} k\frac{q_i^k}{k} = \sum_{k\geq 1} q_i^k = \frac{q_i}{1-q_i}.$$

Therefore

$$\frac{N_{r,i}}{r} \longrightarrow \frac{q_i}{1-q_i}.$$

Applying the continuous map $u \mapsto u/(1+u)$ gives

$$\frac{N_{r,i}}{r+N_{r,i}} \longrightarrow q_i.$$

These statements hold with conditional probability one for every admissible clean image. Averaging over X_0 therefore gives the claimed almost-sure limits without requiring moments of p_{data}. □

For the negative-binomial model, the limiting quantity is the encoded image $q(X_0)$ rather than X_0 itself. Recovering the clean image therefore requires the known encoding q to be invertible on the support of the data. The experimental choice

$$q(X_0) = 0.9X_0$$

has this property.

D Continuation from a common density ratio

We now turn to the first continuation argument. Fix a finite interval

$$I = [a_*, B], \qquad 0 < a_* < B < a_{\max}.$$

Suppose that, instead of predicting each required posterior quantity independently, all predictions are obtained from a single positive function F_a. For Gaussian observations, the corresponding prediction is

$$\nabla \log F_a.$$

For the jump models, the required predictions are obtained from ratios such as $F_a(z+v)/F_a(z)$, together with the fixed model-dependent factors given in Appendix C.1. We denote all these predictions by h^F. We also define the associated generator through (12), even when F is only

a candidate function and not yet known to be the true density ratio. The infinitesimal form of self-consistency is

$$\big(\partial_a + \mathcal{A}_a[\widehat{h}]\big)\widehat{h}_{\xi,a} = 0. \tag{29}$$

This equation says that, along the process generated by the predictions, the expected value of each posterior prediction does not change as the information level increases. Written directly in terms of $\widehat{h}$, this is a nonlinear equation. The point of the common-F representation is that this nonlinearity disappears: after a scalar rescaling, F satisfies a linear equation for the reference process. We then use that equation to recover the correct backward degradation kernels.

D.1 Consistency gives a linear equation

Define

$$V_a(z) = \frac{(\partial_a + \mathcal{A}^0)F_a(z)}{F_a(z)}.$$

If F already satisfied the linear reference equation

$$(\partial_a + \mathcal{A}^0)F_a = 0,$$

then V would be identically zero. Thus V measures precisely how far a candidate F is from satisfying that equation. The symbol V here is a scalar function. It should not be confused with the vector field $\mathcal{V}$ used later for directly predicted quantities.

Lemma D.1. *Let*

$$R_v(a,z) = \frac{F_a(z+v)}{F_a(z)}.$$

Assume that the derivatives and generator integrals below exist absolutely. Then

$$\big(\partial_a + \mathcal{A}_a[h^F]\big)R_v(a,z) = R_v(a,z)\big[V_a(z+v) - V_a(z)\big]. \tag{30}$$

For the Brownian reference, if

$$m = \nabla \log F,$$

then

$$\partial_a m + (m\cdot\nabla)m + \tfrac{1}{2}\Delta m = \nabla V. \tag{31}$$

Consequently, if the predictions generated by F satisfy the consistency equation (29)*, then for each fixed information level a,*

$$V_a(z) = c(a)$$

is independent of the state z. If

$$c \in C((a_*, B)) \cap L^1(I),$$

define

$$G_a(z) = \exp\!\left(-\int_{a_*}^{a} c(t)\,\mathrm{d}t\right) F_a(z). \tag{32}$$

Then G produces exactly the same posterior predictions as F, and

$$(\partial_a + \mathcal{A}^0)G_a = 0$$

on (a_*, B).

Proof. We first consider a jump-model ratio

$$R_v(a,z) = F_a(z+v)/F_a(z).$$

Multiply the left-hand side of (30) by $F_a(z)$. Using

$$F_a(z)R_v(a,z) = F_a(z+v)$$

and the translation structure of the reference generator gives

$$(\partial_a + \mathcal{A}^0)F_a(z+v) - R_v(a,z)(\partial_a + \mathcal{A}^0)F_a(z).$$

Dividing by $F_a(z)$ yields (30). For Gaussian observations, write

$$m = \nabla \log F.$$

A direct calculation gives

$$V = \partial_a \log F + \tfrac{1}{2}\Delta \log F + \tfrac{1}{2}|\nabla \log F|^2.$$

Taking the spatial gradient of this identity gives (31). We now impose self-consistency. For Gaussian observations, (29) and (31) imply

$$\nabla V_a = 0.$$

Hence V_a is constant in the state. For a jump model, (30) shows that self-consistency forces

$$V_a(z + v) = V_a(z)$$

whenever v is an allowed jump. In the count models, every reachable state can be connected to zero by a sequence of unit coordinate steps. Therefore V_a has the same value at every reachable state. For Gamma observations, given any two positive states, one can find a third state lying coordinatewise above both; each original state can be connected to that common state by positive coordinate increments. Thus V_a is again constant. We may therefore write

$$V_a(z) = c(a).$$

Multiplying F_a by a scalar depending only on a does not change $\nabla \log F_a$ and does not change any ratio $F_a(z + v)/F_a(z)$. Hence G in (32) gives the same posterior predictions. The derivative of the scalar factor cancels $c(a)$, leaving

$$(\partial_a + \mathcal{A}^0)G_a = 0.$$

□

For binomial observations, we apply this argument to the birth rates rather than directly to the posterior counts. Equation (15) shows that the two consistency equations are equivalent for $p < 1$. For negative-binomial observations, the ratios corresponding to unit jumps already connect the reachable count states and are therefore enough to show that V_a is state independent. For the count models, F is extended by zero outside the reachable set S, and the reference process is stopped when it attempts to leave S. Equation (30) is used only on edges contained in S. This zero extension is consistent because the reference paths move only upward. Finally, the requirement $c \in L^1(I)$ is a genuine endpoint condition. Pointwise continuity of F inside the interval does not by itself imply that the integral of $c(a)$ remains finite near a_* or B.

D.2 From the differential equation to reference transitions

The previous subsection gives the differential equation

$$(\partial_a + \mathcal{A}^0)G_a = 0.$$

What we need for the continuation argument is the corresponding identity over finite intervals:

$$\overrightarrow{K}^0_{a,b} G_b = G_a, \qquad a_* \le a \le b \le B. \tag{33}$$

In words, if the reference process starts from a state at level a, moves to level b, and we evaluate G_b at the resulting state, then its conditional expectation is G_a at the starting state. Equation (33) is sometimes called space–time harmonicity. We prove it separately for the count, Brownian, and Gamma reference processes because the regularity issues are different.

D.2.1 Count references

Proposition D.2. *Let Π be a nonnegative kernel on a countable state space S with $\Pi(z, S) \le 1$, and let*

$$\mathcal{A}^0 = \kappa(\Pi - \mathrm{Id}), \qquad 0 < \kappa < \infty.$$

Suppose

$$G : I \times S \to [0, \infty)$$

is continuous in the information level at every state, is C^1 in the level on (a_*, B), and satisfies

$$\Pi G_a(z) < \infty, \qquad (\partial_a + \mathcal{A}^0)G_a(z) = 0.$$

Then $\overrightarrow{K}^0_{a,b} G_b = G_a$, where

$$\overrightarrow{K}^0_{a,b} = e^{-\kappa(b-a)} \sum_{j=0}^{\infty} \frac{[\kappa(b-a)]^j}{j!} \Pi^j. \tag{34}$$

No uniform bound on G over the state space is required.

Proof. Set

$$T = B - a_*, \qquad g(t, z) = e^{\kappa t} G_{B-t}(z).$$

The differential equation for G becomes

$$g'(t) = \kappa \Pi g(t).$$

Since $g \geq 0$ and Π is nonnegative, $g'(t) \geq 0$. We want to prove the power-series representation

$$g(t, \cdot) = \sum_{j=0}^{\infty} \frac{[\kappa(t-s)]^j}{j!} \Pi^j g(s, \cdot), \qquad 0 \leq s \leq t \leq T. \tag{35}$$

Once this is established, undoing the change of variables gives (33). The only subtlety is that G is not assumed uniformly bounded over the countable state space. We therefore justify the Taylor expansion using positivity. For a step $h > 0$, define the kth forward difference

$$d_k(t) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} g(t + jh).$$

Each finite sum satisfies

$$d_k' = \kappa \Pi d_k.$$

If $d_k \geq 0$, then $d_k' \geq 0$, so

$$d_{k+1}(t) = d_k(t + h) - d_k(t) \geq 0.$$

Since $d_0 = g \geq 0$, induction gives nonnegative forward differences of every order. A classical result of Boas & Widder (1940) then implies that, in the interior of the interval, g is smooth and all of its derivatives are nonnegative. Such functions admit their Taylor representation (Widder, 1941, Chapter IV). Repeated differentiation of $g' = \kappa \Pi g$ gives

$$g^{(j)} = \kappa^j \Pi^j g.$$

To justify passing derivatives through Π, note that the relevant difference quotients are nonnegative and can be dominated by a fixed difference quotient at a slightly larger time. Starting from $\Pi g < \infty$, dominated convergence then proves inductively that every required integral is finite and that the displayed derivative identity holds. Substituting these derivatives into the Taylor expansion gives (35) for interior times. To include the left endpoint $s = 0$, fix a later interior time and let $s \downarrow 0$. Monotonicity gives termwise convergence, while the series is dominated by its value at a slightly later time, which is finite. The right endpoint is handled similarly by continuity and monotone convergence. This proves the finite-interval identity on all of I. □

For the Poisson and binomial references,

$$\kappa = d.$$

For the negative-binomial reference,

$$\kappa = -d \log(1 - q_0).$$

If the reference process is stopped when it attempts to leave the reachable set S, some rows of Π have total mass below one. This is allowed by the proposition.

D.2.2 The Brownian reference

Proposition D.3. *Let*

$$G : I \times \mathbb{R}^d \to [0, \infty)$$

be continuous. Assume that, in the interior of I, it has one continuous derivative in the information level and two continuous spatial derivatives, and satisfies

$$\partial_a G + \tfrac{1}{2} \Delta G = 0.$$

Then

$$\overrightarrow{K}^0_{a,b} G_b = G_a$$

for the Brownian reference transitions.

Proof. Fix $b > a_*$ and reverse the level variable:

$$u(t, z) = G_{b-t}(z).$$

Then

$$\partial_t u = \tfrac{1}{2}\Delta u.$$

Thus u is a nonnegative solution of the ordinary heat equation, with initial value

$$u(0, z) = G_b(z).$$

The representation theorem for nonnegative heat-equation solutions (Widder, 1944) says that such a solution is obtained by convolving an initial measure with the Gaussian heat kernel. Because G_{b-t} converges locally uniformly to G_b as $t \downarrow 0$, that initial measure is simply

$$G_b(z)\,\mathrm{d}z.$$

Therefore

$$G_a = \overrightarrow{K}^0_{a,b} G_b$$

for every $a_* < a \le b$. It remains only to include $a = a_*$. Fatou's lemma gives

$$\overrightarrow{K}^0_{a_*,b} G_b \le G_{a_*} < \infty.$$

For Gaussian kernels whose variances lie between $s_0/2$ and s_0, the smaller-variance density is bounded by a fixed multiple of the variance-s_0 density. Taking

$$s = b - a, \qquad s_0 = b - a_*,$$

therefore supplies an integrable dominating function. Dominated convergence allows $a \downarrow a_*$ and extends the equality to the endpoint. □

D.2.3 The Gamma reference

The Gamma case is more technical because the reference process has infinitely many small jumps. The additional assumption below says that, after removing the simple factor $e^{\sum_i s_i}$, the candidate function is a Laplace transform of a positive measure. The true Gamma density ratio has exactly this form.

Proposition D.4. *Let $G : I \times (0,\infty)^d \to (0,\infty)$ be continuous in the information level at every state and C^1 in the level in the interior. Suppose that each coordinate integral in $\mathcal{A}^0 G_a$ converges absolutely and*

$$(\partial_a + \mathcal{A}^0) G_a = 0$$

on (a_*, B). Assume also that, for every level a, there is a positive measure ρ_a on $[0,\infty)^d$ such that

$$u_a(s) := e^{-\sum_i s_i} G_a(s) = \int_{[0,\infty)^d} e^{-s\cdot y}\,\rho_a(\mathrm{d}y). \tag{36}$$

Then

$$\overrightarrow{K}^0_{a,b} G_b = G_a$$

for the rate-one Gamma reference process.

Proof. The proof has three steps. First, we translate the differential equation for G into an evolution equation for the measures ρ_a. Because the Laplace transform uniquely determines a positive measure, it is enough to determine how ρ_a changes with a. Frullani's identity gives

$$\int_0^\infty (e^{-wy_i} - e^{-w})\frac{\mathrm{d}w}{w} = -\log y_i.$$

For $y_i > 1$, the positive logarithmic part is integrable because, for every fixed positive s,

$$(\log y_i)^+ e^{-s\cdot y} \le C_s e^{-(s/2)\cdot y}.$$

On each of the two sets $\{y_i > 1\}$ and $\{y_i < 1\}$, the integrand in Frullani's identity has one sign. Tonelli's theorem can therefore be applied separately to its positive and negative parts. The assumed absolute convergence of the Gamma generator then also implies integrability of the negative

logarithmic part. In particular, at every interior level the measure ρ_a has no mass on a coordinate face $\{y_i = 0\}$. Define

$$J(y) = \sum_i \log y_i.$$

The differential equation for G becomes

$$\partial_a u_a(s) = \int J(y) e^{-s \cdot y} \rho_a(\mathrm{d}y). \tag{37}$$

Second, we turn this differential identity into an equation for the measures themselves. Fix $s_0 \in (0, \infty)^d$ and define the finite measure

$$\sigma_a(\mathrm{d}y) = e^{-s_0 \cdot y} \rho_a(\mathrm{d}y).$$

Write

$$P_a = \int \sum_i (\log y_i)^+ \, \mathrm{d}\sigma_a, \qquad N_a = \int \sum_i (-\log y_i)^+ \, \mathrm{d}\sigma_a.$$

The preceding exponential bound, together with continuity of $u_a(s_0/2)$ on the compact interval I, makes P_a uniformly bounded. Integrating

$$\partial_a u_a(s_0) = P_a - N_a$$

and using continuity at the endpoints gives

$$\int_{a_*}^{B} \int \sum_i |\log y_i| \, \sigma_a(\mathrm{d}y) \, \mathrm{d}a < \infty. \tag{38}$$

Now integrate (37) between levels a and b, evaluate it at $s = s_0 + q$ with $q \geq 0$, and use Fubini's theorem. Uniqueness of Laplace transforms of finite measures gives

$$\sigma_b = \sigma_a + \int_a^b J \sigma_t \, \mathrm{d}t. \tag{39}$$

The logarithmic integrability bound (38) allows this identity, initially obtained in the interior, to be extended to the endpoints. On a compact box $[\varepsilon, R]^d$, the multiplier $J(y)$ is bounded, so the measure equation is simply a linear ordinary differential equation. Its unique solution is

$$\sigma_b = e^{(b-a)J} \sigma_a.$$

Letting $\varepsilon \downarrow 0$ and $R \uparrow \infty$ yields

$$\rho_b(\mathrm{d}y) = \left(\prod_i y_i \right)^{b-a} \rho_a(\mathrm{d}y). \tag{40}$$

Third, we substitute this relation into the Gamma reference transition. The increment of the reference process between levels a and b has independent $\mathrm{Gamma}(b-a, 1)$ coordinates. Its exponential transform at the required argument contributes the factor

$$\left(\prod_i y_i \right)^{-(b-a)}.$$

Therefore, by Tonelli's theorem and (40),

$$\begin{aligned} \overrightarrow{K}^0_{a,b} G_b(s) &= e^{\sum_i s_i} \int e^{-s \cdot y} \left(\prod_i y_i \right)^{-(b-a)} \rho_b(\mathrm{d}y) \\ &= e^{\sum_i s_i} \int e^{-s \cdot y} \rho_a(\mathrm{d}y) \\ &= G_a(s). \end{aligned}$$

This is the desired finite-interval identity. □

D.3 Assumptions and continuation

We collect the regularity requirements used above. Their purpose is only to justify the differentiations, integrals, endpoint limits, and passage from the differential equation to the finite-interval relation.

Assumption D.1. On $I = [a_*, B]$, assume the following.

1. All required posterior predictions are obtained from one positive function F using the formulas of Appendix C.1. For the count models, F is set to zero outside the reachable set S.
2. In the interior of I, all derivatives and integrals used in Lemma D.1 exist. For jump processes, the generator integrals involving F, the shifted values of F, and the required ratios converge absolutely. For the Brownian process, the spatial derivatives through order three and the mixed level–space derivatives appearing in (31) are continuous.
3. The scalar function c from Lemma D.1 satisfies
$$c \in C((a_*, B)) \cap L^1(I).$$
The rescaled function G from (32) satisfies the hypotheses of the appropriate finite-interval result: Proposition D.2, D.3, or D.4.
4. For Gaussian and Gamma observations, F and the corresponding posterior predictions are continuous in the state at every level, including a_* and B.

The true density ratio satisfies these requirements on every finite positive-level interval. Indeed, its transition identity (11) implies
$$(\partial_a + \mathcal{A}^0)F_a = 0,$$
so the corresponding scalar c is identically zero. The explicit likelihood formulas above justify the required derivatives and integrals at positive levels. In the Gamma case,
$$e^{-\sum_i s_i} F_L(s) = U_L(s)$$
is precisely a joint Laplace transform, as required by Proposition D.4.

Corollary D.5. *Consider any of the five observation models. Suppose that the candidate predictions are generated from a common function F, agree with the true posterior predictions at level a_*, and satisfy* (29) *on (a_*, B). Under Assumption D.1, the candidate predictions equal the true predictions at every level in $I = [a_*, B]$. Moreover, the process generated by these predictions, when initialized with the true marginal p_{a_*}, has the same law as the true observation process on I.*

Proof. By Lemma D.1, self-consistency produces a rescaled positive function G satisfying the linear reference equation. The appropriate proposition from the previous subsection then gives
$$\overrightarrow{K}^0_{a,b} G_b = G_a.$$
At the starting level a_*, equality of the posterior predictions determines G_{a_*} up to one positive multiplicative constant. For Gaussian observations, equality of $\nabla \log G_{a_*}$ and the true logarithmic gradient determines the function up to such a constant. For the jump models, equality of the neighboring-state ratios determines the same constant by following paths through the state space. Choose the constant so that
$$G_{a_*} r_{a_*} = p_{a_*}.$$
Now define candidate forward transitions by
$$\widehat{\overrightarrow{K}}_{a,b}(z, \mathrm{d}w) = \overrightarrow{K}^0_{a,b}(z, \mathrm{d}w) \frac{G_b(w)}{G_a(z)}. \tag{41}$$
Because $\overrightarrow{K}^0_{a,b} G_b = G_a$, these are probability kernels. The same identity also gives their composition law. If the process starts from $G_{a_*} r_{a_*}$, its marginal at level b is $G_b r_b$. More importantly, when the corresponding two-level joint law is conditioned backward, the factors of G cancel. The backward transition is exactly the known degradation kernel $\overleftarrow{K}_{b,a}$. For the count models with a stopped

reference process, the same calculation is performed on the reachable set S. Since paths move only upward, a path cannot leave S and later return to it. Theorem C.3 now applies. It gives

$$G_b r_b = p_b$$

for every $b \in I$. Hence G_b is the true density ratio r_b-almost everywhere. For count models, the reference distribution has positive mass at every reachable state, so equality holds at every state in S. For Gaussian and Gamma observations, the reference distribution has full support and the density ratios are continuous, so equality extends to every state. The posterior predictions obtained from G are therefore exactly the true ones. The resulting forward transitions are also the true transitions. Thus the candidate process and the true observation process have the same law. Since the candidate predictions equal the true ones, this is also the process generated by the candidate predictions whenever the generator determines its process, as it does for bounded count rates and for the Gamma construction of Lemma E.3. □

Applying the corollary on every finite interval identifies the process at all higher information levels and therefore gives the clean limits in Proposition C.9 when the process is initialized at p_{a_*}. This does not by itself explain how to start from the completely empty observation. That separate issue is handled in Appendices F.4 and G.3.

E Uniqueness for direct predictions

The common-density-ratio result is mathematically convenient, but our experimental networks do not use this parameterization. They predict the required posterior quantities directly. We therefore need a second continuation argument. The idea is simple: consider two families of predictions that agree at a_* and both satisfy the same consistency equation. Under suitable bounds, the difference between them cannot grow away from zero. Hence the continuation is unique. The assumptions needed for this argument depend on the observation model. For the count models, they control the jump rates. For Gamma observations, they control both small jumps and the tail of the jump-size distribution.

E.1 A general uniqueness estimate

Let the outputs be indexed by ξ, and choose positive weights μ_ξ. For a family $u = (u_\xi)$, define

$$\|u\|_\mu = \sup_{z,\xi} \frac{|u_\xi(z)|}{\mu_\xi}.$$

All evolution equations in this section are understood in integrated form at every state and every output index. The following theorem gives the estimate we will apply to each observation model.

Theorem E.1. *Let u and v satisfy*

$$u_a = u_{a_*} + \int_{a_*}^{a} \mathcal{V}_t(u_t)\,\mathrm{d}t$$

on $I = [a_*, B]$, and suppose v satisfies the same equation. Assume

$$u_{a_*} = v_{a_*}.$$

Let

$$\delta(a) = \|u_a - v_a\|_\mu$$

be measurable and bounded. Suppose

$$\|\mathcal{V}_a(u_a) - \mathcal{V}_a(v_a)\|_\mu \le \ell(a)\,\omega(\|u_a - v_a\|_\mu),$$

where $\ell \ge 0$ is integrable and ω is continuous, nondecreasing, satisfies

$$\omega(0) = 0, \qquad \omega(r) > 0 \quad \textit{for } r > 0,$$

and obeys

$$\int_{0^+} \frac{\mathrm{d}r}{\omega(r)} = \infty.$$

Then $u_a = v_a$ for every $a \in I$.

Proof. Subtract the two integrated equations. After dividing by the output weights and taking the supremum,

$$\delta(a) \le \int_{a_*}^{a} \ell(t)\omega(\delta(t))\,\mathrm{d}t.$$

For $\varepsilon > 0$, define

$$v_\varepsilon(a) = \varepsilon + \int_{a_*}^{a} \ell(t)\omega(\delta(t))\,\mathrm{d}t.$$

Then

$$\delta(a) \le v_\varepsilon(a)$$

and, almost everywhere,

$$v_\varepsilon'(a) \le \ell(a)\omega(v_\varepsilon(a)).$$

Dividing by $\omega(v_\varepsilon)$ and integrating gives

$$\int_{\varepsilon}^{v_\varepsilon(a)} \frac{\mathrm{d}r}{\omega(r)} \le \int_{a_*}^{a} \ell(t)\,\mathrm{d}t.$$

If $\delta(a)$ were positive at some level, then $v_\varepsilon(a)$ would remain bounded away from zero as $\varepsilon \downarrow 0$. The left-hand side would then diverge because

$$\int_{0^+} \frac{\mathrm{d}r}{\omega(r)} = \infty,$$

while the right-hand side is finite. This contradiction shows that $\delta(a) = 0$ for every a. □

The same estimate also gives a stability statement when the starting value and the consistency equation are only approximately correct.

Remark E.1. *Suppose*

$$\partial_a \widehat{u} = \mathcal{V}_a(\widehat{u}) + \mathrm{res}_a, \qquad \partial_a u = \mathcal{V}_a(u).$$

Then the argument above begins from

$$\delta(a) \le \varepsilon + \int_{a_*}^{a} \ell(t)\omega(\delta(t))\,\mathrm{d}t,$$

where

$$\varepsilon = \|\widehat{u}_{a_*} - u_{a_*}\|_\mu + \int_{a_*}^{B} \|\mathrm{res}_t\|_\mu\,\mathrm{d}t.$$

If $\omega(r) = r$, the ordinary Grönwall bound gives

$$\sup_I \delta \le \varepsilon \exp\left(\int_I \ell\right).$$

For the Gamma modulus used below,

$$\omega(r) = r\big(1 + \log^+(1/r)\big), \qquad \log^+ x = \max\{0, \log x\},$$

and for constant $\ell = K$, one obtains

$$\sup_I \delta \le e\,\varepsilon^{\exp[-K(B-a_*)]}$$

whenever the right-hand side is at most one. These are uniform bounds over states and output indices. A small average *training loss does not imply that the quantity ε above is small. Moreover, the constants grow with the image size. For the count models $\ell = 4J$ with J proportional to d (Appendix E.3), so the factor $\exp(4J(B - a_*))$ is exponential in the number of pixels and in the length of the continuation. For Gamma, the constant K of Lemma E.2 is proportional to d, since its proof sums over the d coordinates. These estimates are therefore qualitative for images, and they can deteriorate further as the continuation interval approaches the clean limit.*

E.2 ADMISSIBLE PREDICTIONS

For each observation model we now state the bounds under which the preceding uniqueness argument will be applied. Fix

$$I = [a_*, B].$$

For the count models, let S be the reachable set from Proposition C.8. Any jump leaving S is assigned rate zero. When convenient, we extend such rates by zero to states outside S.

Assumption E.1. On I, assume the following model-specific conditions. The constants may depend on I, and they are chosen large enough to contain both the true predictions and the candidate predictions. Count-model predictions are measurable in the information level for every fixed state and output index.

1. *Poisson.* Assume
$$X_0 \in [0, M]^d \qquad \text{almost surely},$$
and
$$0 \le \widehat{h}_{i,a}(n) \le R_I$$
on $I \times S$, where $M \le R_I < \infty$. We write
$$u_{i,a} = \widehat{h}_{i,a}.$$

2. *Binomial.* Assume
$$X_0 \in \{0, \dots, M\}^d \qquad \text{almost surely},$$
and $B < 1$. Suppose
$$n_i \le \widehat{D}_{i,a}(n) \le R_I,$$
where $M \le R_I < \infty$. Define the corresponding birth rate
$$u_{i,a}(n) = \frac{\widehat{D}_{i,a}(n) - n_i}{1 - a}.$$
In this setting the reachable set S is finite.

3. *Negative binomial.* Assume
$$q(X_0) \in [0, \bar{q}]^d \qquad \text{almost surely},$$
with $\bar{q} < 1$. Choose
$$\bar{q} \le \beta_I < 1, \qquad \beta_I > 0,$$
and suppose
$$0 \le u_{i,k,a}(n) := \widehat{M}_{i,k,a}(n) \le \beta_I^k$$
for every $k \ge 1$.

4. *Gamma.* Assume
$$X_0 \in [x_{\min}, x_{\max}]^d \qquad \text{almost surely},$$
where
$$0 < x_{\min} \le x_{\max} < \infty.$$
Write
$$u_{i,w,a}(s) = \widehat{H}_{i,a}(s, w).$$
Assume the candidate is measurable in a, continuous in (s, w), and satisfies
$$\begin{aligned} 0 \le \widehat{H}_{i,a}(s, w) &\le \min\{1, C_I(1 + w)^{-\alpha_I}\}, \\ |\widehat{H}_{i,a}(s, w) - \widehat{H}_{i,a}(s', w)| &\le \mathrm{Lip}_I\, |s - s'|_1, \end{aligned} \tag{42}$$
for finite constants $C_I \ge 1$, $\alpha_I > 0$, and Lip_I.

With these definitions, the consistency equations for all four jump models take the same form:

$$\partial_a u_{\xi,a}(z) = \mathcal{V}(u_a)_\xi(z) := -\int u_{\eta,a}(z)\big[u_{\xi,a}(z+v_\eta) - u_{\xi,a}(z)\big]\,\Lambda(\mathrm{d}\eta). \tag{43}$$

The notation in this formula specializes as follows. For Poisson and binomial observations,

$$\eta = i, \qquad v_i = e_i,$$

and the integral is simply a sum over coordinates. For negative-binomial observations,

$$\eta = (i,k), \qquad v_{i,k} = ke_i, \qquad \Lambda\{(i,k)\} = \frac{1}{k}.$$

For Gamma observations,

$$\eta = (i,w), \qquad v_{i,w} = we_i,$$

and Λ is counting measure over i multiplied by $\mathrm{d}w/w$. For binomial observations, equation (15) is what converts consistency of the posterior counts into the common birth-rate equation above.

E.3 Bounds for the consistency equation

We now verify the estimate required by Theorem E.1. For the count models, choose output weights $\mu_\xi > 0$ so that every candidate in the class satisfies

$$\|u\|_\mu \le 1.$$

Define

$$J = \int \mu_\eta\,\Lambda(\mathrm{d}\eta).$$

When $J < \infty$, compare the two products in (43) by adding and subtracting the mixed term with one factor from u and one from v. This gives

$$|\mathcal{V}(u) - \mathcal{V}(v)|_\mu \le 4J|u-v|_\mu, \qquad |u|_\mu, |v|_\mu \le 1. \tag{44}$$

One factor of $2J$ comes from changing the jump rate, and another $2J$ comes from changing the increment of the predicted quantity. The useful choices of weights are

Model	μ	J
Poisson	R_I	dR_I
Binomial	$R_I/(1-B)$	$dR_I/(1-B)$
Negative binomial	$\mu_{i,k} = \beta_I^k$	$-d\log(1-\beta_I)$.

If necessary, R_I can be enlarged so that it is strictly positive. The same bounds imply absolute convergence of the generator sums. Thus, for all three count models, Theorem E.1 applies with

$$\omega(r) = r, \qquad \ell = 4J.$$

For Gamma observations, this argument cannot be used directly because

$$\int_0^1 \frac{\mathrm{d}w}{w} = \infty.$$

The Gamma process has infinitely many small jumps, so we must use the fact that a small jump changes a Lipschitz prediction only by an amount proportional to its size.

Lemma E.2. *Suppose u and v take values in $[0,1]$, are Lipschitz in the state with common constant L_0 for the norm $|\cdot|_1$, and satisfy the common tail bound*

$$u, v \le C(1+w)^{-\alpha},$$

with $C \ge 1$ and $\alpha > 0$. Then

$$|\mathcal{V}(u) - \mathcal{V}(v)|_\infty \le K\delta\big(1 + \log^+(1/\delta)\big), \qquad \delta = |u-v|_\infty. \tag{45}$$

where K depends only on d, C, α, L_0. The right-hand side is defined to be zero when $\delta = 0$.

Proof. Fix an output (j, ζ) and a state s. For each jump coordinate i, the difference between the two generator integrands can be split into two parts:

$$T_1 = \int_0^\infty |(u-v)_i(s,w)||u_j(s+we_i,\zeta) - u_j(s,\zeta)|\frac{\mathrm{d}w}{w},$$
$$T_2 = \int_0^\infty v_i(s,w)|(u-v)_j(s+we_i,\zeta) - (u-v)_j(s,\zeta)|\frac{\mathrm{d}w}{w}.$$

Assume first that $0 < \delta \leq 1$. We have

$$|(u-v)_i(s,w)| \leq \min\{\delta, 2Cw^{-\alpha}\},$$

while Lipschitz continuity gives

$$|u_j(s+we_i,\zeta) - u_j(s,\zeta)| \leq \min\{L_0 w, 1\}.$$

Similarly,

$$v_i(s,w) \leq \min\{1, Cw^{-\alpha}\},$$

and

$$|(u-v)_j(s+we_i,\zeta) - (u-v)_j(s,\zeta)| \leq \min\{2L_0 w, 2\delta\}.$$

We now separate small and large jumps. For T_1, split first at $w = 1$, and in the tail split again where

$$\delta = 2Cw^{-\alpha}, \qquad w = (2C/\delta)^{1/\alpha}.$$

This gives

$$T_1 \leq L_0\delta + \frac{\delta}{\alpha}\left(\log\frac{2C}{\delta} + 1\right).$$

For T_2, the small-jump contribution is split where

$$L_0 w = \delta$$

when that point lies below one. The resulting bound is

$$T_2 \leq 2\delta\left(1 + \log^+\frac{L_0}{\delta}\right) + \frac{2C\delta}{\alpha}.$$

If $L_0 = 0$, the small-jump term vanishes. Summing these estimates over $i = 1, \ldots, d$ and absorbing fixed constants into K gives (45). □

For Gamma predictions we therefore use the ordinary supremum norm, corresponding to $\mu \equiv 1$. The function

$$\omega(r) = r\big(1 + \log^+(1/r)\big)$$

satisfies

$$\int_{0^+} \frac{\mathrm{d}r}{\omega(r)} = \infty.$$

Hence it meets the condition of Theorem E.1. The same estimates also explain why the Gamma generator is well defined: Lipschitz continuity controls the singular factor $\mathrm{d}w/w$ near $w = 0$, while the polynomial envelope controls the large-w tail. Joint continuity in (s, w) allows the supremum over states and jump sizes to be reduced to a countable dense subset, which ensures measurability as a function of the information level.

E.4 The true predictions and the generated process

We next verify that the true posterior predictions satisfy the same conditions and the same consistency equation. For the Poisson model, bounded images imply

$$0 \leq h_i \leq M.$$

For the binomial model,

$$n_i \leq D_i \leq M.$$

For the negative-binomial model,

$$0 \leq M_{i,k} \leq \bar{q}^k.$$

Differentiating the likelihood formulas from Appendix C.1 gives (43). The image bounds justify the differentiation and the infinite negative-binomial sum. Rates of jumps that leave the reachable set vanish by Proposition C.8. For Gamma observations, bounded image values give

$$H_{i,a}(s,w) \le e^{-w/x_{\max}},$$

which in particular satisfies a polynomial tail bound of the form required in (42). Let

$$\Delta = x_{\min}^{-1} - x_{\max}^{-1}.$$

Bayes' rule gives

$$\partial_{s_i} H_{j,a}(s,\zeta) = -\operatorname{Cov}\Big(X_{0,i}^{-1}, e^{-\zeta/X_{0,j}} \mid S_a = s\Big).$$

A covariance between random variables taking values in intervals of lengths c_1 and c_2 has absolute value at most $c_1 c_2/4$. Applying this bound here gives

$$|\partial_{s_i} H_{j,a}(s,\zeta)| \le \frac{\Delta}{4} \min\{1, \Delta\zeta e^{-\zeta/x_{\max}}\}. \tag{46}$$

Thus the required dependence on the observed state is Lipschitz. Differentiating with respect to the information level instead gives

$$\partial_a H_{j,a}(s,\zeta) = -\operatorname{Cov}\left(\sum_i \log X_{0,i}, e^{-\zeta/X_{0,j}} \,\middle|\, S_a = s\right).$$

Use

$$\log x = \int_0^\infty (e^{-w/x} - e^{-w}) \frac{\mathrm{d}w}{w}$$

and the posterior identity

$$H_{i,a}(s,w) H_{j,a}(s + we_i, \zeta) = \mathbb{E}[e^{-w/X_{0,i}} e^{-\zeta/X_{0,j}} \mid S_a = s].$$

Substitution gives exactly the Gamma version of (43). The bound (46) and the exponential tail justify interchanging the integral and the conditional expectation. For the count processes, bounded jump rates are enough to construct a unique Markov process by uniformization. Gamma requires a separate argument because there are infinitely many small jumps.

Lemma E.3. *Let*

$$H_{i,a}(s,w) \in [0,1]$$

be measurable on $I \times (0,\infty)^d \times (0,\infty)$. Suppose there are envelopes $\overline{H}_i(w) \in [0,1]$ and finite constants M_H, L_H such that

$$H_{i,a}(s,w) \le \overline{H}_i(w), \qquad \sum_i \int_0^\infty \overline{H}_i(w)\mathrm{d}w \le M_H, \tag{47}$$

and

$$\sum_i \int_0^\infty |H_{i,a}(s,w) - H_{i,a}(s',w)|\,\mathrm{d}w \le L_H |s - s'|_1. \tag{48}$$

Then, for every initial distribution on $(0,\infty)^d$, these jump rates define an increasing càdlàg Markov process with generator (19)*. Its paths have finite total variation on I. Its law is unique among solutions of the Poisson-driven equation below when all jump sizes $w > 0$ are retained.*

Proof. The construction is easiest if we first ignore extremely small jumps. Let $\mathcal{N}_i$ be independent Poisson random measures with intensity

$$\mathrm{d}a\,\mathrm{d}w\,\mathrm{d}u/w,$$

independent of the initial state. For $\varepsilon > 0$, retain only jumps of size at least ε and define

$$S_t^\varepsilon = S_{a_*} + \sum_i e_i \int_{(a_*,t]\times[\varepsilon,\infty)\times(0,\infty)} w \mathbf{1}_{\{u \le H_{i,a}(S_{a-}^\varepsilon, w)\}} \mathcal{N}_i(\mathrm{d}a, \mathrm{d}w, \mathrm{d}u). \tag{49}$$

Only Poisson points with $u \le \overline{H}_i(w)$ can be accepted. For $w \ge \varepsilon$, their total rate is at most M_H/ε. Hence there are only finitely many candidate events on a finite interval, and the cutoff process can be

constructed by processing these events in chronological order. We next let $\varepsilon \downarrow 0$. Construct all cutoff processes with the same Poisson random measures. If $0 < \varepsilon' < \varepsilon$, the jumps omitted by the larger cutoff have expected total size at most

$$d\varepsilon(t - a_*).$$

The integrated Lipschitz condition controls the additional difference created because the two processes can accept different later jumps. Using the compensation formula,

$$\mathbb{E} \sup_{a_* \le r \le t} \|S_r^\varepsilon - S_r^{\varepsilon'}\|_1 \le d\varepsilon(t - a_*) + L_H \int_{a_*}^t \mathbb{E} \sup_{a_* \le r \le a} \|S_r^\varepsilon - S_r^{\varepsilon'}\|_1 \, \mathrm{d}a.$$

Grönwall's inequality gives

$$\mathbb{E} \sup_{a_* \le r \le t} \|S_r^\varepsilon - S_r^{\varepsilon'}\|_1 \le d\varepsilon(t - a_*) e^{L_H(t - a_*)}.$$

Thus the cutoff paths form a Cauchy family in expected uniform distance. A subsequence converges uniformly almost surely to an increasing càdlàg process. The same estimate allows passage to the limit in the driving equation, so the limit retains all jump sizes $w > 0$. The limit is a measurable function of the initial condition and of the driving Poisson measures. If another process is driven by the same measures, the same estimate compares it with every cutoff approximation, and letting $\varepsilon \downarrow 0$ shows that the two solutions coincide. This gives pathwise uniqueness for the construction and hence uniqueness in law. Restarting the same construction from any intermediate level gives the Markov property. Finally, (47) implies that the expected total size of all jumps added over I is at most

$$M_H(B - a_*),$$

so the paths have finite variation. This is a standard Poisson-random-measure construction; see, for example, Applebaum (2009) (Chapters 2 and 6). □

Corollary E.4. *Suppose Assumption E.1 holds on every interval $[a_*, B]$ with $B < a_{\max}$. If a candidate prediction agrees with the true prediction at a_* and satisfies* (43) *in integrated form on every such interval, then*

$$\widehat{h}_{\xi,a} = h_{\xi,a}$$

for every

$$a \in [a_*, a_{\max}).$$

The jump process generated from these predictions and initialized with p_{a_*} has the same path law as the true observation process and therefore has the clean limits in Proposition C.9. For Gamma observations, the identified process is the one constructed in Lemma E.3.

Proof. On every finite interval, the true and candidate predictions belong to the same admissible class and solve the same integrated consistency equation with the same value at a_*. For Poisson, binomial, and negative-binomial observations, the Lipschitz estimate (44) and Theorem E.1 give equality. For Gamma observations, use Lemma E.2 in the same theorem. Thus the candidate and true predictions agree on every finite interval. For the count models, the rates are now exactly the true rates. Uniformization therefore gives the true process law. For Gamma observations, the common kernel is the true H. Its exponential tail gives

$$M_H = dx_{\max}.$$

Integrating (46) over jump size gives

$$L_H = \frac{d\Delta^2 x_{\max}^2}{4}.$$

The true Gamma observation process has jump compensator

$$\sum_i H_{i,a}(S_{a-}, w) \, \frac{\mathrm{d}a \, \mathrm{d}w}{w}$$

for jumps of size we_i. A standard representation theorem for point processes with absolutely continuous compensators allows these jumps to be generated, on an enlarged probability space, by the thinning construction (49) (Grigelionis, 1971). Lemma E.3 then identifies its law with the process constructed from the learned kernel. Equality of process laws gives the clean limits. □

E.5 Poisson and binomial rates with linear growth

The bounded-rate assumptions above are sufficient but can be stronger than necessary. In Poisson and binomial models, natural birth rates may grow with the current count. We therefore give a second argument that allows linear growth. The idea is to recover the common function F from the directly predicted birth rates. For this to be possible, increasing coordinate i and then coordinate j must give the same product of rate ratios as increasing j first and then i. We first show that self-consistency preserves this path-independence property. Let $S \subseteq \mathbb{N}^d$ be nonempty and downward closed. Let $b_{i,a}(n) \geq 0$ denote a birth rate. Set it to zero if $n + e_i \notin S$, and extend the rates by zero outside S. The consistency equation is

$$\partial_a b_{i,a}(n) + \sum_k b_{k,a}(n)\big[b_{i,a}(n + e_k) - b_{i,a}(n)\big] = 0. \tag{50}$$

For Poisson observations, b_i is the predicted posterior intensity. For binomial observations,

$$b_i = \frac{\widehat{D}_i - n_i}{1 - a}.$$

Lemma E.5. *Suppose b is continuous in the information level at every state, satisfies* (50) *in integrated form, and obeys the linear-growth bound*

$$0 \leq b_{i,a}(n) \leq K(1 + |n|_1) \qquad \text{on } I \times S \tag{51}$$

for some finite K. Define

$$C_{ij}(n, a) = b_{i,a}(n)b_{j,a}(n + e_i) - b_{j,a}(n)b_{i,a}(n + e_j).$$

Thus $C_{ij} = 0$ says that the product of rates obtained by taking an i-step followed by a j-step is the same as the product obtained in the opposite order. If

$$C_{ij}(n, a_*) = 0$$

for every i, j, n, then

$$C_{ij}(n, a) = 0$$

throughout I. In addition, every rate that is strictly positive at a_* remains strictly positive on I.

Proof. Because the right-hand side of (50) is continuous in a, each rate is C^1 in the interior of I. Suppress the level index temporarily and write

$$\beta(n) = \sum_k b_k(n).$$

Differentiating the definition of C_{ij} and using (50) gives

$$\begin{aligned} \partial_a C_{ij}(n) = {} & [\beta(n) - \beta(n + e_i + e_j)]C_{ij}(n) \\ & + \sum_k \big[b_i(n + e_j)C_{kj}(n) - b_j(n + e_i)C_{ki}(n)\big] \\ & - b_i(n) \sum_k C_{kj}(n + e_i) + b_j(n) \sum_k C_{ki}(n + e_j). \end{aligned} \tag{52}$$

The identity can be verified by expanding the right-hand side. Because the rates are extended by zero outside S, the same expression remains valid at the boundary. For $N \geq 0$, define

$$M_N(a) = \max_{|n|_1 \leq N, i, j} |C_{ij}(n, a)|.$$

The linear-growth bound implies that the sum of the absolute coefficients in (52) is at most

$$10dK(N + 1),$$

and the terms involving C occur only at states with size at most $N + 1$. Therefore, if $C(\,\cdot\,, a_0) = 0$, then with $c = 10dK$,

$$M_N(a) \leq c(N + 1) \int_{a_0}^{a} M_{N+1}(t)\,\mathrm{d}t.$$

Iterating this estimate m times over an interval of length τ gives

$$M_N(a) \le \frac{(c\tau)^m (N+m)!}{m!N!}\, 2K^2(N+m+2)^2.$$

Using

$$\frac{(N+m)!}{m!N!} \le 2^{N+m}$$

gives

$$M_N(a) \le 2^N (2c\tau)^m\, 2K^2(N+m+2)^2.$$

If $2c\tau < 1$, the right-hand side tends to zero as $m \to \infty$. Hence $C = 0$ on every sufficiently short interval. Covering I with finitely many such intervals proves $C = 0$ everywhere. If $K = 0$, the conclusion is immediate. Once $C = 0$, equation (50) simplifies to

$$\partial_a b_{i,a}(n) = b_{i,a}(n)[\beta_a(n) - \beta_a(n+e_i)].$$

This is a scalar linear equation for each fixed edge. Its integrating-factor formula shows that a rate that starts strictly positive cannot become zero at a later level. □

Theorem E.6. *Consider the Poisson or binomial observation process with reachable state space S. Suppose candidate birth rates*

$$b_{i,a}(n)$$

are continuous in the information level, vanish on edges leaving S, satisfy (50) *in integrated form, obey* (51)*, and agree with the true rates at a_*. Then the candidate rates equal the true rates throughout I. The process generated by these rates and initialized with p_{a_*} has the true path law.*

Proof. At level a_*, the true birth rates are ratios of the true density ratio. They are positive on every edge inside S and satisfy

$$C_{ij} = 0.$$

Lemma E.5 shows that both properties remain true at every later level. We can therefore reconstruct a single function F from the rates. Let

$$c_* = \frac{p_{a_*}(0)}{r_{a_*}(0)} > 0$$

and define

$$F_a(0) = c_* \exp\!\left(\int_{a_*}^{a} [d - \beta_t(0)]\, \mathrm{d}t\right), \qquad F_a(n+e_i) = b_{i,a}(n) F_a(n). \tag{53}$$

Every coordinatewise path from zero to a state $n \in S$ stays inside S. Any two such paths can be transformed into each other by repeatedly swapping adjacent coordinate steps. The identity $C_{ij} = 0$ says exactly that each such swap leaves the product of rates unchanged. Hence the value of $F_a(n)$ does not depend on the chosen path. The function F is therefore well defined and positive on S. Set it to zero outside S. At level a_*, its value at zero and all of its neighboring-state ratios agree with those of the true density ratio. Thus

$$F_{a_*} r_{a_*} = p_{a_*}.$$

For an edge inside S,

$$\partial_a \log b_{i,a}(n) = \beta_a(n) - \beta_a(n+e_i).$$

Summing this identity along any path from zero to n gives

$$\partial_a \log F_a(n) = d - \beta_a(n).$$

For the unit-rate Poisson reference stopped on leaving S,

$$\mathcal{A}^0 F_a(n) = F_a(n)[\beta_a(n) - d],$$

and

$$\Pi F_a(n) = d^{-1} F_a(n) \beta_a(n) < \infty.$$

Hence

$$(\partial_a + \mathcal{A}^0) F_a = 0.$$

Proposition D.2 converts this differential equation into the finite-interval relation

$$\overrightarrow{K}^0_{a,b} F_b = F_a.$$

The transformed process defined by (41) therefore has exactly the known backward degradation kernels. Theorem C.3 gives

$$F_a r_a = p_a.$$

Since the reference distribution has positive mass at every reachable state, the ratios of F agree pointwise with the true ratios. Hence all candidate birth rates are the true ones. Finally, the linear-growth condition prevents explosion. The total jump rate is bounded by

$$dK(1 + |n|_1),$$

so the number of jumps is dominated by a standard linear birth process, which is nonexplosive. The directly constructed pure-birth process therefore has the same law as the transformed process above. □

The same result can be applied on every finite interval if the constant in the linear-growth bound is allowed to depend locally on the interval. The condition is an assumption on the candidate class; we do not claim that it holds for every possible clean-image distribution.

F From the training objective to continuation

The continuation theorems begin from two properties:

1. the learned posterior predictions equal the true predictions at some positive level a_*; and
2. the predictions satisfy self-consistency at the higher levels under consideration.

We now explain when these properties follow from exact minimization of the expected training objective. It is important to distinguish two questions. First, what does a zero expected loss imply on the states and levels that are actually sampled by the loss? Second, when does that sampled set cover enough of the state space and information levels to invoke the continuation theorems?

F.1 Exact minimizers and training support

For a candidate predictor $\widehat{h}$, let

$$\widehat{\overrightarrow{K}}_{a,b}$$

denote the transition of the process generated by that predictor. For a starting state z, levels $a < b$, and prediction index ξ, define the consistency residual

$$\Delta_\xi(a, b, z) = \widehat{h}_{\xi,a}(z) - \big(\widehat{\overrightarrow{K}}_{a,b}\widehat{h}_{\xi,b}\big)(z). \tag{54}$$

The ideal consistency loss is

$$\mathcal{L}_{\mathrm{cons}}(\widehat{h}) = \mathbb{E}|\Delta_\xi(a, b, z)|^2.$$

The sampling distribution of (a, b, z, ξ) is allowed to depend on the candidate model, because z or the endpoint may come from trajectories generated by that model. We assume that the conditional expectations in (54) are finite. On finite intervals, the bounded prediction classes used below guarantee this automatically.

Proposition F.1. *Suppose that the candidate class contains the true posterior predictions h, that their regression risk is finite, and that the process generated by h has the true transitions. Use the squared regression loss of Lemma C.4, except for Gamma targets with infinite second moment, where the I-divergence loss of Proposition C.6 is used. Then every global minimizer of*

$$\mathcal{L}_{\mathrm{N2N}}(\widehat{h}) + \lambda_{\mathrm{cons}}\mathcal{L}_{\mathrm{cons}}(\widehat{h}), \qquad \lambda_{\mathrm{cons}} > 0, \tag{55}$$

has the following two properties. First,

$$\widehat{h} = h$$

almost surely under the sampling distribution used by the regression loss. Second,

$$\Delta = 0$$

almost surely under the consistency sampling distribution generated by that candidate. If the prediction error and consistency residual are continuous, then these identities hold at every point in the support of the corresponding sampling distributions.

Proof. For squared regression, Lemma C.4 gives

$$\mathcal{L}_{\mathrm{N2N}}(\widehat{h}) - \mathcal{L}_{\mathrm{N2N}}(h) = \mathbb{E}|\widehat{h} - h|^2 \geq 0.$$

Equality holds exactly when the prediction is correct almost surely under the regression sampling law. The Gamma I-divergence decomposition gives the same conclusion. The true posterior predictions satisfy the tower-property consistency identity, so

$$\mathcal{L}_{\mathrm{cons}}(h) = 0.$$

Hence the true model reaches the value $\mathcal{L}_{\mathrm{N2N}}(h)$ in the combined objective. No candidate can go below this value, because the excess regression loss and the consistency loss are both nonnegative. Therefore every global minimizer must simultaneously attain the minimum regression loss and zero consistency loss. This remains true even when the distribution used to sample the consistency term depends on the candidate itself. Finally, if a continuous nonnegative function is zero almost surely under a measure, it must be zero throughout the support of that measure. □

This is the statement about the expected training objective used in the main text. For bounded count rates, the required existence and uniqueness of the true process follow from uniformization. For Gamma observations under the conditions below, they follow from Lemma E.3 and Corollary E.4. For Gaussian observations, a standard sufficient condition is uniqueness in law for

$$\mathrm{d}Z_a = m_a(Z_a)\,\mathrm{d}a + \mathrm{d}B_a.$$

For example, local Lipschitz continuity and at most linear growth of m are sufficient.

Coverage at the starting level. The proposition gives equality only on the support of the regression sampling distribution. To obtain exact agreement at the continuation level a_*, that support must reach a_*. One possibility is to place positive probability directly on level a_*, with full conditional support over the relevant states and prediction indices. Another is to sample levels approaching a_* and assume continuity of the predictions, so that equality extends to a_* by taking a limit. We refer to either situation as *coverage at* a_*. For Gamma observations, training must also cover every queried jump size $w > 0$, and continuity in w is used to extend almost-sure equality to all jump sizes. For negative-binomial observations, every coordinate and every moment index $k \geq 1$ must be covered. The total regression risk must remain finite, as discussed in Appendix C.3.3.

Remark F.1. *Suppose that*

$$\Delta_\xi(a, b, z) = 0$$

for every state and prediction index and for every sufficiently short interval $a < b$ inside I. This follows, for example, if the consistency sampling law has full support over such intervals and the residual is continuous. Assume also that the candidate predictions lie in the domain of the generator, in the sense that

$$\lim_{\varepsilon \downarrow 0} \frac{\widehat{\overrightarrow{K}}_{a,a+\varepsilon} \widehat{h}_{\xi,a+\varepsilon} - \widehat{h}_{\xi,a}}{\varepsilon}(z) = \big(\partial_a + \mathcal{A}_a[\widehat{h}]\big)\widehat{h}_{\xi,a}(z).$$

Dividing the finite-interval consistency identity by ε and letting $\varepsilon \downarrow 0$ immediately gives (29)*. This argument applies to all five models, but it requires consistency paths to start throughout the continuation interval. In the next two subsections we show that, for the jump models under stronger assumptions, it is enough to start every consistency path from one fixed positive level a_*.*

F.2 One starting level for count models

Assume now that every consistency trajectory begins at the same level a_*. The consistency loss then directly enforces

$$\widehat{h}_{\xi,a_*} = \widehat{\overrightarrow{K}}_{a_*,b}\widehat{h}_{\xi,b},$$

but it does not directly compare the predictions at two later levels a and b with $a_* < a < b$. For bounded count rates, this missing relation follows because the learned forward transition is one-to-one on bounded functions. This is a different injectivity property from Proposition C.2. There, we showed that the *known degradation kernel* is injective on probability distributions. Here, we show that the *learned forward transition* is injective on bounded test functions.

Lemma F.2. *Let S be countable and let measurable jump rates $q_a(z, z') \geq 0$ satisfy*

$$\sup_{a\in I, z\in S} \sum_{z'\neq z} q_a(z, z') \leq q_{\max} < \infty.$$

Let $\widehat{\overrightarrow{K}}_{a,b}$ be the resulting transition operators and $\mathcal{A}_a$ the corresponding generator. Then $\widehat{\overrightarrow{K}}_{a,b}$ is injective on $\ell^\infty(S)$. Moreover, suppose that a uniformly bounded family u_a satisfies

$$\widehat{\overrightarrow{K}}_{a_*,b}u_b = u_{a_*}, \qquad b \in I. \tag{56}$$

Then consistency holds between every pair of intermediate levels:

$$\widehat{\overrightarrow{K}}_{a,b}u_b = u_a, \qquad u_b(z) - u_a(z) = -\int_a^b (\mathcal{A}_t u_t)(z)\,\mathrm{d}t. \tag{57}$$

Proof. The uniform rate bound allows the process to be constructed by uniformization. In particular,

$$\|\mathcal{A}_a f\|_\infty \leq 2q_{\max}\|f\|_\infty.$$

Fix b and a bounded function f, and set

$$v_s = \widehat{\overrightarrow{K}}_{s,b}f.$$

Conditioning on the first jump gives the backward equation

$$v_s = f + \int_s^b \mathcal{A}_t v_t\,\mathrm{d}t.$$

Suppose

$$v_a = \widehat{\overrightarrow{K}}_{a,b}f = 0.$$

Subtracting the backward equation at level a gives

$$v_s = -\int_a^s \mathcal{A}_t v_t\,\mathrm{d}t.$$

Hence

$$\|v_s\|_\infty \leq 2q_{\max}\int_a^s \|v_t\|_\infty\,\mathrm{d}t.$$

Because S is countable, the supremum is measurable in t. Grönwall's inequality yields

$$v_s = 0$$

for every $s \in [a, b]$, and therefore

$$f = v_b = 0.$$

This proves injectivity. Now assume (56). By the semigroup property,

$$\widehat{\overrightarrow{K}}_{a_*,a}\left(\widehat{\overrightarrow{K}}_{a,b}u_b - u_a\right) = 0.$$

Injectivity of $\widehat{\overrightarrow{K}}_{a_*,a}$ implies

$$\widehat{\overrightarrow{K}}_{a,b}u_b = u_a.$$

Thus consistency holds between all intermediate levels. Taking

$$u_s = \widehat{\overrightarrow{K}}_{s,B}u_B$$

and applying the backward equation gives the second identity in (57). □

The count-model classes in Assumption E.1 satisfy the required uniform rate bound. The corresponding $q_{\max}$ can be taken to be the value J computed in Appendix E.3.

Corollary F.3. *Consider Poisson, binomial, or negative-binomial observations and fix $a_* > 0$. Assume the following.*

1. *On every finite continuation interval, the candidate class satisfies Assumption E.1, contains the true predictions, and is continuous in the information level at every state and prediction index.*
2. *The noisy regression targets are unbiased, the true regression risk is finite, and the regression sampling law has coverage at a_* as described in Appendix F.1.*
3. *Every consistency trajectory starts at a_*. The joint sampling law of the starting state, endpoint level, and prediction index has full support on*

$$S \times (a_*, a_{\max}) \times \Xi.$$

For negative-binomial observations, Ξ contains every coordinate and every moment index $k \geq 1$.

Then every global minimizer of (55) *equals the true posterior predictions on the entire continuation range*

$$[a_*, a_{\max}).$$

The process generated by that minimizer and initialized with p_{a_*} has the true path law on every finite interval.

Proof. Fix an arbitrary

$$B < a_{\max}.$$

Proposition F.1 gives two facts: the candidate equals the truth at a_*, and its consistency residual is zero on the support of the consistency sampling law. We next extend the zero residual from almost every sampled endpoint to every endpoint in $[a_*, B]$. Bounded rates imply that the transition law depends continuously on the endpoint in total variation. The predictions are pointwise continuous in the information level. Therefore the residual is continuous in the endpoint. Since the sampling law has full support,

$$\widehat{\overrightarrow{K}}_{a_*,b}\widehat{h}_{\xi,b} = \widehat{h}_{\xi,a_*}$$

holds for every $b \leq B$, every reachable state, and every prediction index. Apply Lemma F.2 separately to each output. It gives consistency between every pair of intermediate levels and hence the integrated version of (43). For the binomial model, the lemma first gives the integrated equation for the posterior-count prediction. Equation (15) converts it into the corresponding equation for the birth rates. Corollary E.4 now identifies the candidate predictions with the true ones on $[a_*, B]$ and identifies the process law. Since B was arbitrary, the conclusion holds throughout $[a_*, a_{\max})$. □

F.3 One starting level for Gamma observations

The same idea works for Gamma observations, but an additional complication appears. The Gamma process has infinitely many small jumps, so its generator is not bounded on the entire space of bounded functions. We instead work with bounded Lipschitz functions on $(0, \infty)^d$. Denote this class by BL, and measure Lipschitz constants using the norm $|\cdot|_1$. For such a function, a jump of size w changes its value by at most a constant times w, which cancels the singular factor $1/w$ in the Gamma jump measure near zero.

Lemma F.4. *Suppose the Gamma jump kernels satisfy Lemma E.3 with constants M_H and L_H. Then, for $a \leq b$ in I, the following hold.*

1. *The transitions preserve bounded Lipschitz functions, and*

$$\|\widehat{\overrightarrow{K}}_{a,b}g\|_\infty \leq \|g\|_\infty,$$

while

$$\mathrm{Lip}(\widehat{\overrightarrow{K}}_{a,b}g) \leq e^{L_H(b-a)}\,\mathrm{Lip}(g).$$

2. *If* $g \in \mathrm{BL}$ *and*

$$v_s = \widehat{\overrightarrow{K}}_{s,b} g,$$

then

$$v_s = g + \int_s^b \mathcal{A}_t v_t \, \mathrm{d}t,$$

and

$$\|\mathcal{A}_t v_t\|_\infty \le d \, \mathrm{Lip}(v_t) + 2M_H \|v_t\|_\infty.$$

3. $\widehat{\overrightarrow{K}}_{a,b}$ *is injective on* BL.

4. *If a family* $u_a \in \mathrm{BL}$ *satisfies* (56) *for every* $a \in I$*, then* (57) *holds.*

Proof. We first prove the Lipschitz estimate. Start two copies of the Gamma process from states x and x' and drive them with the same Poisson random measures. The integrated Lipschitz bound on the jump kernels gives

$$\mathbb{E}\|S_b^x - S_b^{x'}\|_1 \le \|x - x'\|_1 + L_H \int_a^b \mathbb{E}\|S_t^x - S_t^{x'}\|_1 \, \mathrm{d}t.$$

Grönwall's inequality yields

$$\mathbb{E}\|S_b^x - S_b^{x'}\|_1 \le e^{L_H(b-a)} \|x - x'\|_1.$$

Applying a Lipschitz function g gives the claimed Lipschitz bound for $\widehat{\overrightarrow{K}}_{a,b} g$. The supremum bound follows simply because a Markov transition averages values of g and therefore cannot increase their maximum absolute value. For later use, define

$$c_g = e^{L_H(B-a_*)} \, \mathrm{Lip}(g).$$

Next consider the generator. For any $f \in \mathrm{BL}$, split the jump integral at $w = 1$. For $w < 1$,

$$|f(s + we_i) - f(s)| \le \mathrm{Lip}(f) w.$$

For $w \ge 1$,

$$|f(s + we_i) - f(s)| \le 2\|f\|_\infty.$$

Therefore

$$|\mathcal{A}_t f|_\infty \le d \, \mathrm{Lip}(f) + 2M_H |f|_\infty. \tag{58}$$

To derive the backward equation rigorously, first remove all jumps smaller than $\varepsilon \in (0, 1)$. The remaining process has finite total rate bounded by

$$q_\varepsilon = d \log(1/\varepsilon) + M_H,$$

so its ordinary finite-rate backward equation is valid. Let

$$v_s^\varepsilon = \widehat{\overrightarrow{K}}^{\varepsilon}_{s,b} g.$$

The coupling argument above gives

$$\mathrm{Lip}(v_s^\varepsilon) \le c_g$$

uniformly in ε. The cutoff estimate from Lemma E.3 also gives

$$\|v_t^\varepsilon - v_t\|_\infty \le d(B - a_*) c_g \varepsilon.$$

We now pass the cutoff backward equation to the limit. Write

$$\mathcal{A}_t^\varepsilon v_t^\varepsilon - \mathcal{A}_t v_t = \mathcal{A}_t(v_t^\varepsilon - v_t) - (\mathcal{A}_t - \mathcal{A}_t^\varepsilon) v_t^\varepsilon.$$

The second term is at most

$$d\varepsilon c_g.$$

For

$$\chi = v_t^\varepsilon - v_t,$$

we have both a supremum bound and a Lipschitz bound, so

$$|\chi(s + we_i) - \chi(s)| \le \min\{2c_g w, 2\|\chi\|_\infty\}.$$

Splitting the jump integral at the point where these two bounds meet gives

$$\|\mathcal{A}_t\chi\|_\infty \le 2d\|\chi\|_\infty \left[1 + \log^+\left(\frac{c_g}{\|\chi\|_\infty}\right)\right] + 2M_H\|\chi\|_\infty.$$

This tends to zero as $|\chi|_\infty \to 0$. If $c_g = 0$, then g is constant and the claim is immediate. Bound (58) provides an integrable dominating function in time. Hence the finite-rate backward equation converges to the desired backward equation for the full Gamma process. We next prove injectivity. Suppose

$$\widehat{\overrightarrow{K}}_{a,b} g = 0.$$

Let

$$v_s = \widehat{\overrightarrow{K}}_{s,b} g, \qquad \phi(s) = \|v_s\|_\infty.$$

Subtracting the backward equations at a and s gives

$$v_s = -\int_a^s \mathcal{A}_t v_t \, \mathrm{d}t.$$

This time, split each Gamma generator at a small threshold ε. Jumps below ε contribute at most

$$dc_g\varepsilon,$$

while jumps above ε have finite total rate q_ε and contribute at most $2q_\varepsilon\phi(t)$. Therefore

$$\phi(s) \le dc_g\varepsilon(s - a) + 2q_\varepsilon \int_a^s \phi(t) \, \mathrm{d}t.$$

Grönwall's inequality gives

$$\phi(s) \le dc_g(s - a)e^{2M_H(s-a)}\varepsilon^{1-2d(s-a)}.$$

If

$$s - a \le \theta < \frac{1}{2d},$$

the exponent of ε is positive. Letting $\varepsilon \downarrow 0$ gives

$$v_s = 0.$$

The same argument can then be restarted from the end of this short interval. Repeating it finitely many times reaches b, and therefore

$$g = v_b = 0.$$

This proves injectivity. Finally, suppose (56) holds. Preservation of BL and the semigroup property give

$$\widehat{\overrightarrow{K}}_{a_*,a}\left(\widehat{\overrightarrow{K}}_{a,b} u_b - u_a\right) = 0.$$

Injectivity implies

$$\widehat{\overrightarrow{K}}_{a,b} u_b = u_a.$$

Thus consistency holds between every two intermediate levels. Writing

$$u_s = \widehat{\overrightarrow{K}}_{s,B} u_B$$

and using the backward equation gives (57). □

Corollary F.5. *Suppose*

$$X_0 \in [x_{\min}, x_{\max}]^d \qquad \textit{almost surely},$$

and fix $a_* > 0$. Assume the following.

1. *On every finite interval $I = [a_*, B]$, the candidate class satisfies Assumption E.1 with $\alpha_I > 1$ and also satisfies*
$$\sum_i \int_0^\infty \left|\widehat{H}_{i,a}(s,w) - \widehat{H}_{i,a}(s',w)\right| \mathrm{d}w \le L_I |s - s'|_1. \tag{59}$$
The class contains the true predictions. Candidates are continuous in the information level and jointly continuous in (a, s, w).

2. *Regression uses squared loss with a fixed target shape larger than $1/2$, or I-divergence with any fixed positive target shape. The true regression risk is finite, and the regression law has coverage at a_*.*

3. *Every consistency trajectory starts at a_*. The joint sampling distribution of starting state, endpoint level, coordinate, and jump size has full support on*
$$(0,\infty)^d \times (a_*, \infty) \times \{1, \dots, d\} \times (0, \infty).$$

Then every global minimizer of (55) *equals the true Gamma posterior predictions for every $a \ge a_*$, every positive state, every coordinate, and every jump size $w > 0$. The process generated from that minimizer and initialized with p_{a_*} has the true path law on every finite interval.*

Proof. Fix
$$I = [a_*, B].$$
Since $\alpha_I > 1$, the tail bound in (42) is integrable with respect to $\mathrm{d}w$. Together with (59), this verifies the assumptions of Lemma E.3 for every candidate in the class. The true Gamma kernel generates the true observation process. Proposition F.1 therefore gives exact agreement with the true prediction at a_* and zero consistency residual almost surely under the candidate's consistency sampling distribution. We next extend this almost-sure identity to every state, endpoint, and queried jump size. The residual is jointly continuous in those variables: the transition operators preserve Lipschitz bounds, the candidate is jointly continuous and bounded, and the Gamma process is stochastically continuous at every fixed level. Dominated convergence together with the coupling estimate from Lemma F.4 gives the required continuity. Because the sampling law has full support, zero residual almost surely therefore implies
$$\widehat{\overrightarrow{K}}_{a_*,b} \widehat{H}_{j,b}(\,\cdot\,,\zeta) = \widehat{H}_{j,a_*}(\,\cdot\,,\zeta)$$
for every admissible tuple (b, j, ζ). For fixed (j, ζ), the prediction as a function of the state belongs to BL. Lemma F.4 then gives the integrated consistency equation between all intermediate levels. Corollary E.4 identifies the predictions and the process law on I. Since B was arbitrary, the conclusion holds for every $a \ge a_*$. □

The assumptions above do contain the true Gamma kernel. Its exponential tail is stronger than any of the required polynomial tail bounds. Moreover, integrating (46) over w gives
$$L_I = \frac{d\Delta^2 x_{\max}^2}{4}$$
in (59). The joint continuity requirement is stated explicitly because it is what allows an identity that holds almost surely under a joint sampling distribution to be extended to every tuple in its support.

F.4 Generation from the empty state

The continuation theorems begin at a positive level a_*. Full generation, however, begins from the empty or uninformative state. The regression loss can identify the true predictions directly at the levels covered by the noisy observations, namely
$$0 < a \le a_{\mathrm{in}}.$$
The continuation argument identifies them for
$$a \ge a_*.$$
If
$$a_* \le a_{\mathrm{in}},$$
these two ranges overlap and together cover every positive information level.

Corollary F.6. *Suppose*

$$X_0 \in [0, M]^d \qquad \text{almost surely}.$$

Assume the count-model class conditions of Corollary F.3, with continuity at every positive level. Suppose the regression targets are unbiased and have finite true risk, and the regression sampling law has full support on

$$(0, a_{\mathrm{in}}] \times S \times \{1, \dots, d\}.$$

Assume also the consistency coverage of Corollary F.3 at some

$$0 < a_* \le a_{\mathrm{in}}.$$

Then every global minimizer of (55) *equals the true Poisson jump rates on*

$$(0, \infty) \times S.$$

The process generated by these rates and started from zero is nonexplosive, has the true Poisson observation law, and satisfies

$$\frac{N_\lambda}{\lambda} \longrightarrow X$$

almost surely for a random image $X \sim p_{\mathrm{data}}$.

Proof. Proposition F.1, full support of the regression law, and continuity imply that the learned rates equal the true rates throughout

$$(0, a_{\mathrm{in}}].$$

Corollary F.3 gives equality throughout

$$[a_*, \infty).$$

Since $a_* \le a_{\mathrm{in}}$, these intervals overlap and cover all positive levels. The true total Poisson rate is bounded by

$$dM.$$

Uniformization therefore gives a unique nonexplosive process starting from the zero state. The value assigned to the rate at the single level $a = 0$ has no effect on the law of the process. The clean limit then follows from Proposition C.9. □

The same argument applies to binomial and negative-binomial observations under their respective class assumptions. For the negative-binomial model, every moment index $k \ge 1$ must be covered. For binomial observations, the argument is applied on finite intervals below $p = 1$. For Gamma observations, starting from zero additionally requires the entrance assumption in Assumption C.1.

G Sampling the learned process

The continuation theorems above concern the exact stochastic processes: the information level is continuous, all jump sizes are retained, and the process is followed to its theoretical clean limit. The numerical samplers used in the experiments make several additional approximations. They use finitely many information levels, may truncate small or large jumps, and stop at a finite information level. We describe these approximations here so that they are not confused with the exact theoretical statements.

G.1 Finite steps

For Gaussian observations, we discretize the stochastic differential equation with Euler–Maruyama:

$$z_{j+1} = z_j + \Delta a_j \widehat{m}_{a_j}(z_j) + \sqrt{\Delta a_j}\, \eta_j, \qquad \eta_j \sim \mathcal{N}(0, I_d).$$

For Poisson observations, a tau-leap freezes the current posterior rate over one step:

$$n_{j+1,i} = n_{j,i} + P_{j,i}, \qquad P_{j,i} \sim \mathrm{Pois}\big(\Delta\lambda_j \widehat{h}_{i,\lambda_j}(n_j)\big),$$

independently across coordinates. This frozen-rate approximation does not reproduce the exact conditional variance of the increment. Under the true process,

$$\begin{aligned}&\mathrm{Var}(N_{\lambda_j+\Delta\lambda_j,i} - N_{\lambda_j,i} \mid N_{\lambda_j} = n_j)\\ &\quad = \Delta\lambda_j h_{i,\lambda_j}(n_j) + \Delta\lambda_j^2 \,\mathrm{Var}(X_{0,i} \mid N_{\lambda_j} = n_j).\end{aligned}$$

The Poisson tau-leap reproduces the first term but omits the second, which comes from uncertainty in the latent clean intensity during the step. For binomial observations, the frozen birth rate is

$$\frac{\widehat{D}_{i,p_j}(n_j) - n_{j,i}}{1 - p_j}.$$

This rate must be nonnegative. A Poisson approximation to the number of births can jump outside the reachable state space even though the exact binomial process cannot. Projecting or clipping such a state back into the reachable set changes the transition law. The experiments instead plug the predicted count into the exact binomial increment,

$$n_{j+1,i} = n_{j,i} + \mathrm{Binomial}\Big(\mathrm{round}\big((\widehat{D}_{i,p_j}(n_j) - n_{j,i})_+\big),\ \frac{p_{j+1} - p_j}{1 - p_j}\Big),$$

which cannot overshoot the predicted count but, like the Poisson tau-leap, ignores the uncertainty in $X_{0,i}$ during the step. We do not prove convergence of these finite-step schemes for arbitrary learned prediction functions.

G.2 Sampling jump sizes

For Gamma observations, keeping only jumps of size

$$w \geq \varepsilon$$

turns the infinite-activity process into a finite-rate process. The total rate is finite under either the true integrability bound (20) or the learned tail bound (42). Suppose, for example, that

$$X_{0,i} \leq x_{\max}.$$

Then the true posterior transform satisfies

$$H_{i,L}(s, w) \leq e^{-w/x_{\max}}.$$

If a learned kernel satisfies the same bound, one can simulate candidate jumps in coordinate i from the dominating measure

$$e^{-w/x_{\max}} \frac{\mathrm{d}w}{w}$$

and accept a proposed jump of size w with probability

$$e^{w/x_{\max}} \widehat{H}_{i,L}(s, w),$$

evaluated at the current information level and state. Updating the state after every accepted event gives an exact simulation of the ε-cutoff process. The effect of discarding small jumps can be bounded locally. If $0 \leq H_i \leq 1$, then jumps smaller than ε contribute at most

$$dT\varepsilon$$

to the expected total added state over an interval of length T. For a Lipschitz test function, their direct contribution to the generator is at most

$$d\varepsilon \,\mathrm{Lip}(f).$$

If the integrated state-Lipschitz assumption of Lemma E.3 also holds, the coupling argument in that lemma gives the stronger pathwise estimate

$$\mathbb{E} \sup_{0 \leq t \leq T} \|S_{a+t} - S^{\varepsilon}_{a+t}\|_1 \leq d\varepsilon T e^{L_H T}.$$

Without such a stability assumption, controlling the jumps omitted in the current step is not enough to control their effect on all future state-dependent rates. For negative-binomial observations, suppose

$$q_i \leq \bar{q} < 1.$$

Then

$$M_{i,k,r} \leq \bar{q}^k.$$

An exact event sampler for any learned kernel obeying the same bound can therefore use the following rejection scheme. Propose events at total rate

$$-d \log(1 - \bar{q}).$$

Choose the coordinate uniformly. Then choose the jump size k with probability

$$\frac{\bar{q}^k}{k[-\log(1-\bar{q})]}.$$

Accept the proposed jump with probability

$$\frac{\widehat{M}_{i,k,r}(n)}{\bar{q}^k},$$

evaluated at the current level and state. This construction uses all $k \geq 1$ and therefore does not require a moment cutoff. The image experiments instead freeze rates on a finite grid and predict only finitely many moments. For the true process, omitting all jumps with $k > k_{\max}$ removes at most

$$d \sum_{k>k_{\max}} \frac{\bar{q}^k}{k}$$

of the total jump intensity and at most

$$\frac{d\bar{q}^{k_{\max}+1}}{1-\bar{q}}$$

of the expected count added per unit shape. For

$$\bar{q} = 0.9, \qquad k_{\max} = 30,$$

these quantities are approximately 0.4% of the total jump intensity and 4% of the expected count rate at a pixel with $q_i = \bar{q}$. They are smaller for darker pixels. These calculations quantify what is removed directly by truncating the true jump kernel. They do not, by themselves, give a bound on the distance between the final generated image distributions, because errors can propagate through later state-dependent rates.

G.3 Starting and stopping

Generation begins from the least informative state and passes through the range in which noisy regression directly supervises the model. For Gaussian, Poisson, binomial, and negative-binomial processes, zero is an ordinary initial state. For Gamma observations, zero lies on the boundary of the state space used in the interior jump formulation, so starting from zero requires the entrance assumption in Appendix C.1.4. In practical Gaussian implementations one often begins at a large but finite noise level rather than at the exact infinite-noise limit. This is another numerical approximation. At the other end of the trajectory, stopping at a finite information level does not produce an exactly clean observation. For example, conditional on $X_0 = x$,

$$\operatorname{Var}\left(\frac{N_\lambda}{\lambda}\middle| X_0 = x\right) = \frac{x}{\lambda},$$

and

$$\operatorname{Var}\left(\frac{S_L}{L}\middle| X_0 = x\right) = \frac{x^2}{L}.$$

Thus the normalized state still contains measurement noise at every finite λ or L. An alternative is to output the posterior mean of the clean image at the stopping state. This removes conditional measurement noise by averaging, but it has a different distribution from the clean data at finite information level. Indeed, whenever $X_{0,i}$ has finite variance,

$$\operatorname{Var}(\mathbb{E}[X_{0,i} \mid Z_a]) = \operatorname{Var}(X_{0,i}) - \mathbb{E}\operatorname{Var}(X_{0,i} \mid Z_a).$$

The posterior mean therefore has lower variance than the clean image unless the remaining posterior uncertainty is zero. Throughout the experiments, we call this posterior-mean output the *readout*. We state explicitly when a reported metric is computed from the readout rather than from the normalized stochastic state.

H EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

H.1 NETWORKS AND ESTIMATES OF THE CONSISTENCY LOSS

The CIFAR-10 experiments use a U-Net with approximately 60 million parameters. The Poisson and binomial networks predict posterior means. The negative-binomial network additionally predicts the posterior moments needed for its jump rates. The radar experiments use a U-Net with approximately 9 million parameters. At every pixel it predicts the parameters of the mixture model described in Appendix H.4. Our implementation builds on the Poisson-jump code of Chen & Zhou (2023).[1]

The networks predict the required posterior quantities directly. They do not enforce the common-density-ratio representation studied in Appendix D. For Poisson and binomial observations, the outputs are clipped to a fixed nonnegative range. This gives a uniform output bound. The binomial implementation, however, does not explicitly represent the reachable set S, so it does not guarantee that rates on forbidden edges vanish, as required by Assumption E.1. For negative-binomial observations, the output for the kth moment is a sigmoid multiplied by β^k. Thus the network satisfies a geometric upper bound in k. The implementation represents only finitely many moments, whereas the theorem uses all $k \geq 1$.

For radar, the mixture parameterization makes each pixelwise prediction a valid Laplace transform. It does not by itself ensure the uniform tail and state-Lipschitz conditions of Corollary F.5, which we have not verified over the full trained network class. That corollary permits all consistency paths to start from one observed Gamma level, provided its remaining assumptions hold. The experiments therefore test the method under approximations that the exact continuation results do not cover: finite parameterizations, finite coverage of levels, states and jump sizes in training, estimated losses and finite-step simulation.

Monte Carlo estimators. To estimate consistency, start from a state z at level a and draw P conditionally independent trajectories from the current model to level b. For a vector-valued prediction, write the residual along path p as

$$r_p = \widehat{h}_{\xi,a}(z) - \widehat{h}_{\xi,b}(Z_b^{(p)}).$$

We consider two estimators:

$$\widehat{\ell}_{\mathrm{avg}} = \left| \frac{1}{P} \sum_{p=1}^{P} r_p \right|^2, \qquad \widehat{\ell}_{\mathrm{prod}} = \langle r_1, r_2 \rangle. \tag{60}$$

Conditional on the starting state, levels, and output index,

$$\mathbb{E}\widehat{\ell}_{\mathrm{avg}} = |\mathbb{E}r_1|^2 + \frac{1}{P} \operatorname{tr} \operatorname{Cov}(r_1), \qquad \mathbb{E}\widehat{\ell}_{\mathrm{prod}} = |\mathbb{E}r_1|^2. \tag{61}$$

The ideal loss is $\|\mathbb{E}r_1\|^2$, the squared norm of the conditional mean residual. The averaged estimator adds a Monte Carlo variance term that can remain positive even for exactly self-consistent predictions. At finite P, its expected loss can therefore have a different minimizer from the ideal loss. The product estimator is unbiased for the ideal loss value, assuming finite second moments and conditionally independent exact trajectories from a fixed model.

H.2 CIFAR-10 WITH POISSON, BINOMIAL, NEGATIVE-BINOMIAL AND GAMMA OBSERVATIONS

The CIFAR-10 experiments let us vary the measurement law and observation level while keeping the underlying image distribution fixed. They test both the importance of matching the process to the measurements and the effect of consistency refinement beyond the range supervised by noisy views.

Networks, views and training. All CIFAR-10 models use the same U-Net of approximately 60 million parameters and the same training budget. For each 32×32 image, we draw two conditionally independent noisy views, use one as the network input and the other as the regression target, and

[1] https://github.com/tqch/poisson-jump

	Poisson	Binomial	Neg. binomial	Gamma	Conv-PC	SAR
N2N loss	I-divergence, δ_t-weighted	I-divergence	squared, variance-weighted	BCE + squared mean	I-divergence, δ_t-weighted	I-divergence kernel, power mean ($\alpha = \frac{1}{2}$)
N2N levels	uniform t, atom t_{obs}/T	uniform t, atom t_{obs}/T	power law, atom 0.2	$L_{\mathrm{obs}}u^3$	uniform $t \geq t_{\min}$; refinement adds atom $t_{\min}/T$	$L_{\mathrm{obs}}u^{3/2}$
Estimator	averaged	product	averaged	product	product	product
Path steps ($P = 2$)	4	4	64	16	4	24 ($P = 4$)
Path endpoints	uniform $t < t_{\mathrm{obs}}$	uniform $t < t_{\mathrm{obs}}$	$(r_{\mathrm{obs}}, 2r_{\mathrm{obs}}]$	log-uniform $(L_{\mathrm{obs}}, 50L_{\mathrm{obs}}]$	uniform $t < t_{\min}$	log-uniform $(L_{\mathrm{obs}}, 50L_{\mathrm{obs}}]$
Paths per device batch	≤ 8	≤ 8	≤ 8	4	≤ 8	4, every third step
λ_{cons}	1	1	1	0.3	0.003	1
Gradient at anchor	yes	yes	yes	detached	yes	detached
Refinement	200 epochs from 1800	200 from 1800	200 from 1800	200 from 1800	20 epochs from 100	40 epochs (2×20)
Sampler steps	1000	1000	2000 (power grid)	250	1000	400 to $2L_{\mathrm{obs}}$
Trajectory output	state N_λ/λ	predicted image	predicted mean at $2r_{\mathrm{obs}}$	predicted mean at $50L_{\mathrm{obs}}$	predicted image	–
Extra term	–	–	–	–	–	rollout regression

Table 7: Training and sampling settings. The CIFAR-10 models share the U-Net, batch size 128, learning rate $2 \cdot 10^{-4}$ and EMA decay 0.9999. Paths start at the anchor in every experiment. Only images whose drawn level lies above the anchor carry a path, so the Poisson and binomial models average $8t_{\mathrm{obs}}/T$ paths per device batch and the negative binomial 1.6. The estimator was chosen per family on FID (Table 9). The Conv-PC model is a 35M-parameter U-Net trained for 100 epochs of 500 steps on 128-pixel crops of single frames, with batch size 32, learning rate 10^{-4} and EMA decay 0.9999. Its anchor is $t_{\min} = 634$ of 1000, where the input holds 98% of a frame.

redraw both at every epoch. Base training runs for 1800 epochs and is followed by 200 epochs of consistency refinement. Each training seed is refined from its own base checkpoint. We compute FID from 50,000 samples generated with an exponential moving average of the weights, using the clean CIFAR-10 training set as the reference. Poisson and binomial trajectories use 1000 sampling steps, and negative-binomial trajectories use 2000.

Because both views are redrawn, these experiments assume that the acquisition can be repeated. In the microscopy and radar experiments the recorded measurement is fixed, and only its split into training views is redrawn.

Observation levels. The implementation indexes the process by a timestep t that decreases as the observation becomes more informative. After inversion and rescaling, this is equivalent to the convention used in the theoretical framework, where instead of timestep t, we use the information level a. Table 8 gives every level of Table 2 in the parameters of its own law.

The Poisson process accumulates counts at rate $\lambda\alpha_t$ per pixel with $\lambda = 4.65 \cdot 10^4$, so the counts quoted in the main text are $\lambda\alpha_{t_{\rm obs}}$ at a white pixel. The binomial process places the particles of the 8-bit clean image, so a white pixel holds $N = 255$ particles and its expected count is $255\, p_{t_{\rm obs}}$.

Negative-binomial observations have no comparable count axis, because the factor $(1-q)^2$ in the delta-method variance makes bright pixels much quieter and dark pixels much noisier than a Poisson observation of the same nominal budget. We place them instead by the root mean squared error of the best unbiased single-view estimator of X_0 under the CIFAR-10 pixel law. On that scale $r_{\rm obs} = 5$, 16 and 50 are as informative as 15.5, 55.2 and 178.3 Poisson counts, so the outer two negative-binomial rows of Table 2 are matched to its Poisson rows at 16.1 and 178.4 counts.

Gamma observations are indexed by the number of looks. The best unbiased single-view estimator is S_L/L, whose relative error $1/\sqrt{L_{\rm obs}}$ is 12.5% at $L_{\rm obs} = 64$ and, unlike the Poisson case, is the same at every pixel. A Poisson observation with 64 counts at a white pixel has the same relative error there, so the Gamma row lies between the 16.1- and 178.4-count rows. This equivalence holds only at a white pixel.

Process	Level	Law parameter	Counts at a white pixel
Poisson	$t = 440$	$\alpha = 1.3 \cdot 10^{-2}$	610
Poisson	$t = 500$	$\alpha = 3.8 \cdot 10^{-3}$	180
Poisson	$t = 600$	$\alpha = 3.5 \cdot 10^{-4}$	16
Binomial	$t = 132$	$p = 0.71$	180
Binomial	$t = 226$	$p = 0.39$	100
Binomial	$t = 364$	$p = 0.098$	25
Binomial	$t = 448$	$p = 0.031$	8

Table 8: Observation levels behind Table 2, as parameters of the observation law: the counts are $\lambda\alpha$ with $\lambda = 4.65 \cdot 10^4$ for the Poisson process and $255\, p$ for the binomial one. Table 2 quotes these to one decimal because they are exact values set by the schedule. The negative-binomial and Gamma settings have no count axis and are placed on an equal-accuracy scale in the text.

The families continue their full trajectories to different levels, which should be kept in mind when comparing rows of Table 2. The Poisson and binomial trajectories run to the end of the grid, roughly 76 times the exposure of the 613-count anchor and 2900 times that of the 16-count anchor. The negative-binomial trajectories stop at $2r_{\rm obs}$, twice their observed shape, and the Gamma trajectories at $50L_{\rm obs}$. In each family, this is the largest level that the consistency paths reach during refinement.

Simulating the consistency loss. Each training image draws one timestep t. If $t \geq t_{\rm obs}$, the image contributes an N2N term at t, with input $\mathrm{Bin}(Y_1, \alpha_t/\alpha_{\rm obs})$. Otherwise it contributes an N2N term at the anchor, with input Y_1, and a consistency path from Y_1 to level t. Every path therefore starts at the anchor. For the Poisson and binomial models the N2N levels carry an atom of mass $t_{\rm obs}/T$ at the anchor, and path endpoints are uniform over the timesteps no pair of views reaches. No gradient

flows through the simulated paths. Table 7 lists for each experiment the number of path steps, the paths per batch, the consistency weight and whether the anchor prediction is detached.

Choice of consistency estimator. Table 9 evaluates both estimators of (60) at every Poisson and binomial setting. Within each family the same estimator wins at every setting, and the winner differs between the two families. The main text selects the estimator per family. With $P = 2$ paths, the Monte Carlo variance term in $\widehat{\ell}_{\mathrm{avg}}$ is about 0.1% of the loss on the unrefined checkpoints, at the trajectory depths that contribute most to training. On the refined checkpoints it grows to 8–24%, so the two estimators differ appreciably by the end of refinement. We do not have an explanation for this reversal.

Model	Observation	Averaged estimator	Product estimator
Poisson, $t = 440$	613.1 counts	$\mathbf{3.85 \pm 0.06}$	5.14 ± 0.25
Poisson, $t = 500$	178.4 counts	$\mathbf{8.47 \pm 0.11}$	22.52 ± 0.82
Poisson, $t = 600$	16.1 counts	$\mathbf{13.13 \pm 0.02}$	17.84 ± 1.21
Binomial, $t = 132$	180.1 counts	13.95	$\mathbf{11.65 \pm 0.12}$
Binomial, $t = 226$	100.0 counts	16.81	$\mathbf{11.63 \pm 1.65}$
Binomial, $t = 364$	25.1 counts	27.16	$\mathbf{13.23 \pm 0.61}$
Binomial, $t = 448$	8.0 counts	36.43	$\mathbf{19.68 \pm 0.15}$

Table 9: Full-trajectory CIFAR-10 FID after 200 epochs of consistency refinement, using the two estimators in (60). Bold marks the lower FID. The averaged estimator wins at all three Poisson exposures, by 1.3, 14.1 and 4.7 FID, and the product estimator wins at all four binomial ones, by 2.3 to 16.8. Poisson rows and the product column are mean $\pm$ standard deviation over two training seeds, each refining its own 1800-epoch checkpoint. The binomial averaged column is a single seed.

Misspecified measurement laws. The two Gaussian baselines of Table 10 share the architecture and budget of the Poisson model and differ only in the law they assume. The direct baseline uses one global standard deviation $\sigma_{\mathrm{obs}} = \sqrt{\bar{x}/\lambda_{\mathrm{obs}}}$, the Poisson standard deviation evaluated at the mean CIFAR-10 pixel value $\bar{x} = 0.4734$ instead of at each pixel. The Anscombe baseline works on $A(y) = 2\sqrt{\lambda_{\mathrm{obs}} y + 3/8}$, rescaled so that a white pixel maps to one. The same approximation then gives $\sigma_{\mathrm{obs}} = 1/(2\sqrt{\lambda_{\mathrm{obs}} + 3/8})$. This baseline returns to intensity space through the algebraic inverse of A, clipped to $[0, 1]$, which is biased. Both baselines use the true photon rate λ_{obs}. Each stops at the first level whose trajectory noise exceeds its observation noise, $(1 - \bar{\alpha}_t)/\bar{\alpha}_t \geq \sigma^2_{\mathrm{obs}}$, and returns the anchor readout there. The Poisson model is better by 8.7–14.9 FID at the two higher counts, and at the lowest count the baselines' FID is more than six times larger.

A single-view model makes the most extreme misspecification: it treats the observation as clean. With the same architecture and 1800 epochs, a diffusion model trained on single views as if they were clean images gives FID 148.09 at 25 binomial counts, against 17.38 with two views. Under Gaussian diffusion at 178 Poisson counts it gives 69.42, against 7.19. Both gaps, 130.7 and 62.2 FID, arise at the observed level, before any continuation.

	Counts at a white pixel		
Law assumed for the measurements	613.1	178.4	16.1
Poisson (ours)	$\mathbf{5.64 \pm 0.06}$	$\mathbf{7.19 \pm 0.01}$	$\mathbf{16.29 \pm 0.38}$
Gaussian, one variance	17.33 ± 0.51	17.89 ± 0.22	109.84 ± 2.50
Gaussian after Anscombe	20.58 ± 0.75	15.87 ± 0.80	107.29 ± 1.97

Table 10: A misspecified measurement law, at the anchor readout. FID from 50,000 samples, mean $\pm$ standard deviation over two training seeds. All three models are 1800-epoch checkpoints.

Calibration of the continued prediction. We measure calibration of the posterior-mean prediction by the slope

$$\frac{\mathrm{Cov}(\widehat{x}_0, x_0)}{\mathrm{Var}(\widehat{x}_0)},$$

which equals one when the target calibration relation holds. States are drawn from the forward kernel at the level in question, on 2048 held-out CIFAR-10 test images, so levels cleaner than the observation can be reached even though no pair of views supervises them. We compare the base model, the refined model and a zero-weight control, trained for the same 200 epochs with $\lambda_{\mathrm{cons}} = 0$. The control and the refined model are the two full-trajectory columns of Table 2.

Inside the supervised range every model is within $2 \cdot 10^{-3}$ of one, before and after refinement. Beyond the observed level the models differ. At 178 Poisson counts the slope of the base model is off by 3.5–$5.5 \cdot 10^{-3}$ and that of the zero-weight control by 5.3–$9.0 \cdot 10^{-3}$, while that of the refined model is off by 3.6–$5.7 \cdot 10^{-4}$, an order of magnitude less. At 100 binomial counts the three deviations are 0.95–$1.8 \cdot 10^{-2}$, 0.40–$1.3 \cdot 10^{-2}$ and 1.2–$1.8 \cdot 10^{-3}$. At the measured levels the refined slope is about 3 to 15 times closer to one than that of the zero-weight control, in both settings. Additional training without the consistency term leaves the deviation at the order of the base model, so the improvement comes from the term.

For the Gamma models the same comparison can be made in FID. At the anchor level the zero-weight model gives 9.32 and the refined model 8.70, a gain of 0.62 ± 0.11. The 200 additional epochs alone change FID by 0.02, and the two seeds disagree in sign.

Jump rates for the negative binomial. The negative-binomial models use the linear encoding $q(X_0) = 0.9X_0$ and draw two conditionally independent views from $\mathrm{NB}(r_{\mathrm{obs}}, q(X_0))$. The network predicts the posterior moments

$$M_{i,k,r}(n) = \mathbb{E}[\, q_i(X_0)^k \mid N_r = n \,], \qquad k = 1, \ldots, 66,$$

which the sampler needs for its jump rates, using the unbiased targets in (27). A separate output serves as the final image readout. It is trained on $T_{1,r_{\mathrm{obs}}}(Y_2)$ and divided by 0.9 at sampling time, so its regression target in expectation is $\mathbb{E}[X_0 \mid N_r]$, which the rescaled $k = 1$ moment also estimates. The two agree to within 0.01 dB in reconstruction quality. Sampling uses tau-leaping on a power-law grid of 2000 shape steps, which captures 63 of the modes of a finite-mixture oracle on 64 images, against 38 for a uniform grid of the same length.

A geometric weighting of the moment loss in k leaves large k weakly supervised. The models reported in Table 2 instead weight each moment target by its estimated variance, which supervises every k. They are sampled with every k, as the continuation theorem prescribes, and without a jump-size cutoff. On the refined checkpoint at $r_{\mathrm{obs}} = 50$, FID is flat above a cutoff of $k_{\max} = 10$: 13.28, 12.47 and 12.44 for $k_{\max} = 10$, 30 and 66. The geometrically weighted model needs a cutoff at $k_{\max} = 30$, because its weakly supervised large-k outputs produce isolated bright pixels. With this cutoff it reaches 5.41 at the observed level, against 7.78 for the variance-weighted model. We report the variance-weighted model because a cutoff changes the process that the theory specifies.

If the rate for each jump size is instead computed from the first moment as $\widehat{q}^{\,k}/k$, FID is about nine points worse at every cutoff above $k_{\max} = 5$. These rates carry no information beyond M_1, and their FID stops improving at 22.1. With the network and the sampling seed fixed, the same comparison on the geometrically weighted model at 1200 epochs and $k_{\max} = 30$ gives 7.88 with the predicted moments and 13.97 with rates computed from the mean. Two unit-jump processes matched only to the mean give 189 and 469. The predicted higher moments thus have a much larger effect than the cutoff.

Entrance level of the Gamma process. The Gamma process reaches $S_0 = 0$ only as $L \downarrow 0$: for every $\varepsilon > 0$, the state S_ε is almost surely positive. Unlike for the count models (Appendix F.4), the reverse process must therefore start at a positive level, and we choose this level from the training range. The network is trained on $[\varepsilon_0, L_{\mathrm{obs}}]$ with $\varepsilon_0 = 10^{-4} L_{\mathrm{obs}}$, where the zero state is no longer typical. We start from $S = 0$ at $10^{-5} L_{\mathrm{obs}}$, where it still is, and advance to ε_0 with the exact kernel for a clean-image law that has independent pixels. Every step above ε_0 uses the learned kernel. This falls short of Assumption C.1, which requires the law $p_{\mathrm{data}} Q_\varepsilon$. The entrance laws we construct are consistent as $\varepsilon \downarrow 0$ and have the per-pixel marginals of p_{data}, but not its spatial dependence.

Two measurements indicate that this substitution has little effect. First, the entrance state carries a fraction $\varepsilon_0/L_{\mathrm{obs}} = 10^{-4}$ of the mass of the anchor state, and replacing its spatial content by uniform, shuffled or smooth-noise scale fields changes the result only within noise. Second, the independent-pixel law has an 8×8 block correlation of 0.125, the value for white noise, while the generated samples reach 0.807 against 0.799 for real data. The learned kernel produces this structure. The independent-pixel law is used only for sampling, and training is Noise2Noise on two noisy Gamma views throughout.

The Gamma trajectories are read out at $50L_{\mathrm{obs}}$ because the consistency endpoints are drawn log-uniformly on $[L_{\mathrm{obs}}, 50L_{\mathrm{obs}}]$, so 82% of them lie beyond $2L_{\mathrm{obs}}$. An earlier readout at $2L_{\mathrm{obs}}$ changes FID in a direction that depends on the exposure. At $L_{\mathrm{obs}} = 64$, the readout at $50L_{\mathrm{obs}}$ gives 7.06 ± 0.27 and the readout at $2L_{\mathrm{obs}}$ gives 7.19 ± 0.19. The difference is smaller than the 0.143 change from redrawing the 50,000 samples. At $L_{\mathrm{obs}} = 16$ the same comparison gives 20.36 and 12.01. The continuation from $2L_{\mathrm{obs}}$ to $50L_{\mathrm{obs}}$ thus improves the samples at the higher exposure and degrades them at the lower one. The two refinements are identical apart from the exposure scaling, so the reversal comes from the exposure. We do not report $L_{\mathrm{obs}} = 16$ as a row, since it has a single training seed.

Seeds and FID resampling. Every configuration in Table 2 is trained with two seeds, each refined from its own 1800-epoch checkpoint, and the table reports their mean and standard deviation. The only single-seed entry is the binomial averaged-estimator column of Table 9. The seed spread does not include the variation of FID across sample sets. Scoring one checkpoint on a fresh set of 50,000 samples changes FID by 0.143. A standard deviation smaller than this reflects only the agreement between the two seeds.

H.3 Photon-counting microscopy

Generation. The Conv-PC dataset contains repeated microscopy frames consisting of photon counts. Each observed frame is randomly split into a thinned input and its complement. This matches the subset-training construction used by Generative Accumulation of Photons Krull et al. (2024). As in GAP, training includes the frames of the test field of view, which supplies reference crops and the frame shown in Figure 2. For the main generation experiment, we draw 10,000 samples from each model. All methods are rendered using the same fixed intensity range $[0, 4]$ and are evaluated with the same clean-FID pipeline. Starting from one base checkpoint, we run 20 epochs of consistency refinement for three random seeds. We use the product estimator, consistency weight 0.003, two simulated paths per update, and four sampling steps beyond the observed level. Generation starts from zero counts and continues through the information level corresponding to one observed frame. Estimator and weight were chosen on a pilot run with 256 samples, where the averaged estimator reaches FID 74.6 and the product estimator 46.7, against 58.4 for the anchor readout.

At 64×64 resolution, consistency refinement improves over our reproduction of GAP when either a fixed rendering range or a percentile computed from the reference distribution is used. If each image is clipped using its own percentiles, the ordering reverses. The published GAP FIDs, 134.20 and 86.82, are not directly comparable because the public archive and the PNG rendering differ from the original.

All scores above use the predicted image at the final state, whereas the reference crops are count images. With Poisson noise at the reference exposure added to every predicted image, the anchor readout scores 42.4 and the refined trajectory 35.6 at 256×256. Part of the FID of the predicted images therefore reflects the noise texture of the reference. With the added noise, the continuation without consistency still collapses (FID 227), and GAP scores 73.6.

Conditional continuation. The 512-frame reference in Figure 4 is not a clean image: it averages 39.5 counts per pixel in the displayed crop, so it carries 16% shot noise at its mean and more in the dark lumens. With every image divided by its mean, the anchor readout and GAP's estimate differ from the reference by 0.49 and 0.51 RMSE, a single continuation from either method by 0.63 and 0.64, and the N2N-only continuation by 1.19. Since one frame of $5{,}020$ photons does not determine the scene, a sample should differ from the reference by about the posterior spread. The errors are indeed largest where the spread over 500 continuations is large: 1.06 error and 0.44 spread on the bright walls, against 0.34 and 0.14 in the lumens.

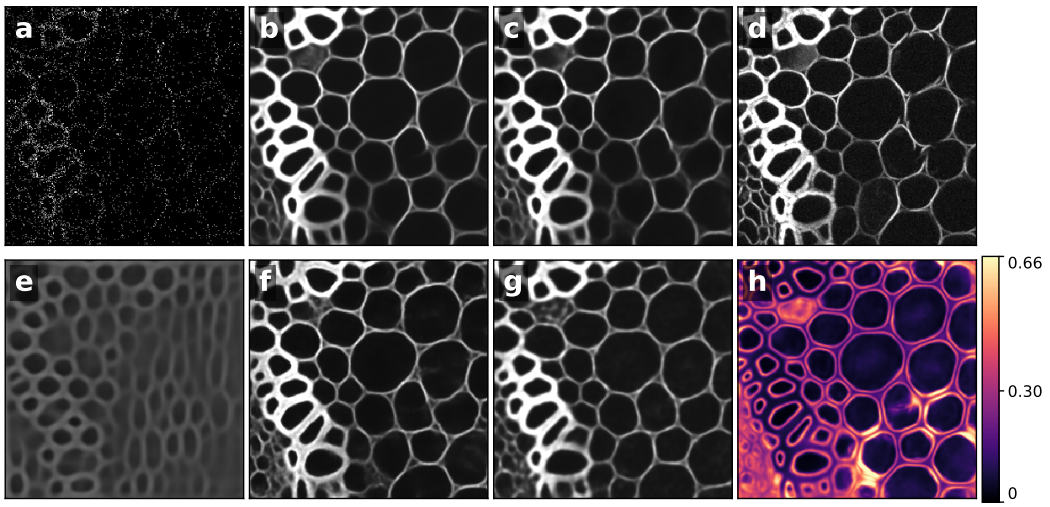


Figure 4: Continuation from one measured frame. (a) One measured frame from the test field of view, 5,020 photons in 256×256 pixels. (b) Anchor readout of the N2N model. (c) GAP's estimate from the same frame. (d) The 512-frame reference. (e) N2N continuation to the reference exposure without consistency. (f) GAP continuation under its own rule, accumulating to 0.32 photons per pixel. (g) Continuation with consistency refinement. (h) Relative uncertainty of 500 such continuations: pixelwise standard deviation divided by pixelwise mean, color scale from 0 to 0.66. Panels (b)–(g) are scaled by their mean and clipped at four.

Panel (h) shows the spread of the predicted image at the endpoint, relative to its pixelwise mean. It concentrates on the wall boundaries and on the dim cluster at the upper left. Count images at the reference exposure spread more, because they add Poisson shot noise, which accounts for 23% of their variance on the walls and 58% in the lumens. Without the shot noise, their relative spread is 0.22, against 0.19 for the predicted image.

Restoration and data access. The restoration experiment follows the transductive protocol used by GAP. The released GAP training split contains frames from the field of view used for testing. Under the matched Noise2Noise protocol, all frames from this field of view may therefore also be used during training. We evaluate all 512 frames using GAP's scale-normalized PSNR. With each frame randomly split for Noise2Noise training, three training seeds give 25.081, 24.993 and 25.038 dB, against 24.917 dB for our reproduction of GAP, a mean improvement of 0.12 ± 0.04 dB. When both models instead use 511-frame leave-one-out targets, the Noise2Noise advantage is 1.031 dB (Table 11). If Noise2Noise is trained only on the other three fields of view, excluding the test field of view entirely, the PSNR falls to 23.168 ± 0.005 dB.

Method	PSNR (dB)
GAP, fission	24.917 ± 0.005
N2N, fission, three seeds	25.037 ± 0.044
N2N, 511-frame target, 256-pixel crops	26.836 ± 0.010
GAP, 511-frame target	25.805 ± 0.009

Table 11: Single-frame Conv-PC restoration under transductive training. Rows with three training seeds report the standard deviation across seeds. The other rows report the standard error across the 512 evaluated frames. The published GAP result for the fission setting is 24.84 dB.

H.4 Sentinel-1 radar

Views and preprocessing. We model the sum of L radar looks as

$$S_L \mid X_0 \sim \mathrm{Gamma}(L, 1/X_0).$$

The normalized intensity S_L/L has conditional mean X_0 and relative conditional variance $1/L$. The data consist of three Sentinel-1 interferometric-wide single-look-complex acquisitions from 2024. After TOPS deramping, the squared real and imaginary components provide two half-looks. Following Dalsasso et al. (2022), we model them as $S^{(i)} \mid X_0 \sim \mathrm{Gamma}(1/2, 1/X_0)$, so $L_{\mathrm{obs}} = 1/2$. Under this model, the ratio $S^{(1)}/(S^{(1)} + S^{(2)})$ has a beta distribution with variance 0.125. The empirical variance is 0.1247. After correcting for image texture, the estimated conditional correlation between the two views is $+0.003$.

Raw neighboring radar samples still have appreciable speckle correlation: approximately 0.31 in range and 0.44 in azimuth. To reduce this dependence, we separate each image into the four interlaced 2×2 sublattices. This keeps all pixels but increases the spacing between neighboring samples in each resulting image. The empirical adjacent-sample correlation then falls to approximately 0.04. The split is needed because the generative model assumes conditionally independent Gamma noise across pixels. Without it, the network can learn spatially correlated speckle from the data, which the sampling model does not reproduce. In our experiments this mismatch prevents the sampled images from reproducing the scene structure reliably. The resulting dataset contains 25,920 patches of size 128×128, of which ten percent are held out. After the sublattice split, the ground spacing is approximately 4.7×27.8 m.

Pixelwise posterior predictions. At every pixel i, the network represents the Gamma posterior transform as a four-component mixture:

$$\widehat{H}_{i,L}(s, w) = \sum_{j=1}^{4} \omega_{i,j}(L, s) \left(1 + \frac{w}{\kappa_{i,j}(L, s)}\right)^{-\alpha_{i,j}(L,s)}. \tag{62}$$

The weights $\omega_{i,j}$ are nonnegative and sum to one, and $\alpha_{i,j}$ and $\kappa_{i,j}$ are positive. For fixed network outputs, this function is the Laplace transform of a Gamma mixture for $1/X_{0,i}$, or equivalently of an inverse-Gamma mixture for $X_{0,i}$. Although the output is pixelwise, all mixture parameters can depend on the complete input patch. Uniform bounds $\alpha_{i,j} \geq \alpha_0 > 0$ and $\kappa_{i,j} \leq \kappa_{\max} < \infty$ would give the polynomial large-w decay that (42) requires. We do not enforce them. The remaining assumptions of the Gamma continuation theorem, uniform Lipschitz continuity in the observed state and the integrated bound (59), do not follow from the mixture representation either.

Training uses the I-divergence with the Gamma target constructed from the second half-look, with target shape $r = 1/2$. By Proposition C.5, this target has infinite second moment for every $w > 0$, which rules out squared-error regression. The optimal I-divergence risk is finite by Proposition C.6. Since the one-dimensional Gamma family with known shape is complete, this target is the only integrable unbiased estimator of $e^{-w/X_{0,i}}$ based on one observation, so no unbiased target from the same measurement has a finite second moment.

To reduce the effect of extreme target values, we average the target over a 2×2 window. The average still has infinite second moment, and it is biased where the reflectivity X_0 changes within the window: a plain 2×2 average reduces the mean intensity of the brightest 0.1% of pixels from 0.77 to 0.62 relative to the reference. We therefore average the target only for jump sizes w below a local, state-dependent scale and leave larger jump sizes unsmoothed.

On held-out patches, the probability integral transform (PIT) of the second half-look under the predicted law has Kolmogorov–Smirnov (KS) distance 0.005, against 0.25 for a simple Gamma plug-in predictor centered on the noisy observation. Without the target averaging, the pooled distance increases to 0.016. Pixel-scale distances of this size should be read as comparisons between predictors. About 1% of the observed pixel values are clipped at their lower bound, which puts a small atom in the empirical distribution that no continuous predictive law has, and the resulting spike in the PIT is of the same order as the distances themselves. A clipped value moves a 16×16 block sum by about 10^{-7} in relative terms, so block-sum distances are unaffected. Comparisons that score two predictors against the same target are unaffected at every scale. Nominal 50%, 80%, 90% and 95% predictive intervals have empirical coverage 49.9%, 80.2%, 90.2% and 95.2%.

Refinement from an observed half-look. All results below use two refinement seeds of a single base model. We study conditional trajectories that start from the first observed half-look $S^{(1)}$. Refinement simulates additional looks and lasts 40 epochs, run as two consecutive 20-epoch stages,

the second of which widens the simulated range. The refinement objective has two terms. In the consistency term, the prediction at the starting state is detached from differentiation. The rollout regression uses half of each training batch. Starting from the first half-look, we simulate the current model to a level $L \in (L_{\mathrm{obs}}, 50L_{\mathrm{obs}}]$ drawn log-uniformly, and regress the prediction at the simulated state on the target constructed from the second half-look. This target remains unbiased at the simulated state, because the two views are conditionally independent given the reflectivity.

Three ranges of levels are involved. The Noise2Noise loss supervises levels up to L_{obs}, and the rollout regression places the network on its own states up to $50L_{\mathrm{obs}}$. The held-out tests are at $2L_{\mathrm{obs}}$, against the second view, and at about $4.5L_{\mathrm{obs}}$, against the reserved reference band.

The rollout regression does not supervise the law of the trajectory. The rollout draws its randomness independently of the second view given the first, so the optimum of the regression at a simulated state is the posterior prediction given the *starting* view, which is broader than the posterior given the simulated state. We keep the term because it stabilizes the prediction at the observed level, and we evaluate the simulated increments separately.

We normalize increments by $L - L_{\mathrm{obs}}$. For the true Gamma process,

$$\mathbb{E}\left[\frac{S_L - S^{(1)}}{L - L_{\mathrm{obs}}}\middle| S^{(1)}\right] = \mathbb{E}[X_0 \mid S^{(1)}].$$

After refinement, the mean normalized increment of the learned trajectory agrees with the network's predicted posterior mean to within 4% through $16L_{\mathrm{obs}}$. At $L = 2L_{\mathrm{obs}}$, the true increment $S_{2L_{\mathrm{obs}}} - S^{(1)}$ has, conditional on X_0, exactly the same distribution as the second half-look. For 16×16 block sums, the second-view PIT of the learned trajectory has KS distance 0.048 and 0.030 on the two seeds (Table 12). Because of the clipped pixel values, we do not read the pixel column of that table as a property of the trajectory.

We compare second moments of the simulated increments with the Gamma scaling, using a reference texture variance estimated from the cross-covariance of the two half-looks. The posterior mean alone would underestimate this variance by roughly a factor of four. On this heavy-tailed field, the cross-covariance estimate is itself biased downward, most strongly at small block sizes. At the 4-, 8- and 16-pixel block scales, the second moments follow the Gamma scaling to within approximately 10% for $L \leq 8L_{\mathrm{obs}}$, the range we checked. We do not report the pixel scale over this range. At $L = 16L_{\mathrm{obs}}$ the fine-scale variance falls below the reference. Since the reference is biased downward, a ratio against it would understate this gap, and we do not quantify it.

Spatial prediction. The pixelwise posterior mixture in (62) specifies one-pixel conditional distributions, but it does not by itself specify how uncertainty is shared across pixels. We therefore evaluate spatial dependence by predicting sums over blocks of the held-out second half-look. We compare two samplers. The first draws one reflectivity value independently at every pixel from the predicted pixelwise mixture and then draws an independent Gamma half-look conditional on those reflectivities. The second uses the learned conditional trajectory, which can propagate dependence across pixels through the network's state-dependent rates. For each of the two refinement seeds, both methods use 32 Monte Carlo draws for each of 256 held-out patches. Table 12 reports the KS distance and the coverage of nominal 90% intervals.

At the single-pixel level, the two methods have similar coverage. At 16×16 block scale, however, independent-pixel sampling gives only 82.5–83.2% coverage, whereas the conditional trajectory gives 88.6–89.2%. The trajectory thus better captures spatial uncertainty at the tested block scales. This does not show that the two samplers have the same one-pixel marginal distributions at a finite stopping level.

At the 16×16 scale, the PIT of the conditional trajectory is skewed to one side while its interval widths stay accurate. Its KS distance therefore grows with block size, but its coverage does not fall. The lowest of ten PIT bins holds 13.3% and 11.9% of the mass on the two seeds, against 10% expected, so the real block sums tend to lie below the generated draws. Only the part of this excess below the 5% point reduces the coverage of the 90% interval, which is consistent with the 88–89% in Table 12. The independent sampler has a larger excess in the same bin, 17.4% and 18.6%. Block-sum calibration also varies between adjacent checkpoints of a single run, and the training loss does not identify the best checkpoint. We select the checkpoint on a held-out slice disjoint from the one we report.

Seed	Sampler	Pixel	4×4	8×8	16×16
1	Chain	0.004/89.9	0.013/89.9	0.026/89.6	0.048/88.6
1	Independent pixels	0.002/90.3	0.029/88.3	0.056/85.9	0.100/82.5
0	Chain	0.004/89.8	0.013/89.5	0.019/89.4	0.030/89.2
0	Independent pixels	0.002/90.3	0.031/88.3	0.057/86.1	0.088/83.2

Table 12: Prediction of held-out Sentinel-1 second-view block sums. Each entry reports KS distance / empirical coverage of the nominal 90% interval, in percent. For each refinement seed, both samplers use 32 draws on 256 patches drawn from the training run's own held-out split, disjoint from every patch the model was trained on.

Ablation of the refinement terms. To separate the consistency term from the rollout regression, we branch four models from the same unrefined checkpoint and refine each for 40 epochs in a single stage, with each term switched on or off. The budget matches that of the reported model, but none of the four is the reported model. For each, we measure the fraction of the network's own per-pixel posterior spread that the trajectory realizes at the reserved reference band.

We use this ratio because absolute coverage at the band is hard to interpret. The band is narrower than the training views, so it resolves a somewhat smoother field, and its effective number of looks, about 1.8, is known only approximately. These effects and the clipped pixel values are shared by all four models and cancel in the ratio. For the same reason, only changes in coverage at the band are informative. Coverage is 91.4% and 91.1% on the two seeds before refinement and 90.8% on both after, against a nominal 90%. Refinement thus changes coverage by less than a point, while the realized spread increases from 0.947 to about 0.985.

Refinement	Seed 1	Seed 0
None, unrefined checkpoint	0.947	0.947
Neither term, same budget	0.969	0.964
Rollout regression only	0.916	0.934
Consistency only	0.990	0.992
Both	0.986	0.984

Table 13: Fraction of the network's own per-pixel posterior spread that the sampled trajectory realizes at the reserved reference band, as a ratio of log-variances over 512 held-out patches. A value of 1 means the trajectory expresses exactly the spread the network predicts. All refined rows branch from the same unrefined checkpoint and use the reported model's budget of 40 epochs in a single stage. The second row is the zero-weight control, whose seed-0 run stopped early at 30 of 40 epochs on the training criterion.

Consistency alone raises the ratio from 0.947 to 0.990 and 0.992, whereas rollout regression alone lowers it to 0.916 and 0.934, below the unrefined checkpoint. With both terms the ratio is 0.986 and 0.984. The zero-weight control, trained for the same budget without either term, reaches 0.969 and 0.964, so extra training alone gives roughly 40 to 50% of the gain from consistency. Consistency alone is, however, less stable at the observed level. Its residual mean moves across seeds by more than that of the other refined models, while the combination stays inside on both. We therefore report the model with both terms. The ablation shows that the consistency term supplies the supervision beyond the observed level, while the rollout regression stabilizes the observed level.

Despeckling. We score an estimator $\widehat{X}_0$ by the residual ratio $S^{(2)}/(L_{\mathrm{obs}}\widehat{X}_0)$ on the same 256 held-out patches. Its mean is 1 when the estimate is calibrated, in the sense $\mathbb{E}[X_0 \mid \widehat{X}_0] = \widehat{X}_0$. The usual unbiasedness condition $\mathbb{E}[\widehat{X}_0 \mid X_0] = X_0$ is a different requirement, and the two differ by Jensen's inequality. The 5% null floor of its KS distance is 0.0007. Table 14 compares our posterior mean with a 7×7 Lee filter.

Longer chains. In one additional experiment with a single training seed, both refinement terms are applied out to $50L_{\mathrm{obs}}$. At 8×8 block scale, coverage of the held-out second view is 89.0%,

Estimator of X_0	residual mean	residual KS
Ours, posterior mean	$\mathbf{1.003}$	$\mathbf{0.014}$
Lee 7×7	0.707	0.141

Table 14: Despeckling on held-out Sentinel-1 patches. The residual ratio $S^{(2)}/(L_{\mathrm{obs}}\widehat{X}_0)$ has mean 1 for a calibrated estimate. Its 5% null KS floor is 0.0007.

compared with 85.0% for independent pixelwise draws. The mean simulated increment stays within 5% of the network's posterior mean through $32L_{\mathrm{obs}}$, and within 8% at $50L_{\mathrm{obs}}$. Variance at the 8- and 16-pixel block scales remains close to the reference through $L = 25$, that is $50L_{\mathrm{obs}}$. At the pixel scale, the variance falls clearly below the reference beyond $16L_{\mathrm{obs}}$. As discussed above, the cross-covariance estimate of the reference is unreliable on this field, so we do not quantify the deficit.

For a full radar burst of size $12{,}920 \times 753$, we draw 50 conditional samples at $L = 25$. The median ratio between the sample mean and the predicted posterior mean is 1.03. The corresponding ratio of standard deviations is 0.74, and on a crop of the same burst the ratio of generated to predicted 90% interval widths is 0.81. Even at $L = 25$, the Gamma observation model still contains non-negligible measurement noise: its conditional speckle variance is 4% of squared reflectivity. Calibration against the held-out second view also does not show that the chain has the marginal distributions of the target clean-image posterior at 25 looks.