Manuscript prepared for Higher Categorical Semantics

# Generative Models and Their Symmetries

## Markov Kernels, Variational Objectives, and Higher Descent

Anonymous • 3 November 2026

**Abstract**

We give a higher categorical companion to Tianhua Chen's *The Little Book of Generative AI Foundations* (arXiv:2605.29713v2). A Markov category separates the construction of a law from the coordinates used to compute it. Relative entropy on joint laws gives the variational bound and the finite diffusion objective. Invertible pushforward gives normalizing flows, while couplings and normalized reweighting account for adversarial and energy-based models. We then form stacks of model presentations with specified latent symmetries. Observable laws and integrable variational objectives descend along covers, but a descended presentation can carry nontrivial gluing data invisible to both. A three-state finite model exhibits this obstruction over a circle. A rank-one Gaussian latent model with four latent coordinates yields the homotopy type $B\mathrm{O}(3)$ on a fixed loading orbit. Its fourth homotopy group is $\mathbb{Z}$, so a discrete quotient or an ordinary groupoid cannot recover the resulting topological family space.

**Keywords.** Markov category, disintegration, variational inference, diffusion, homotopy quotient, higher topos, probabilistic PCA.

## Contents

**1 From generative laws to spaces of presentations** **2**

**2 Kernels and evidence** **2**

**3 Linear models, PCA, and variational objectives** **4**

**4 Diffusion as path composition and reversal** **5**

**5 Exact density, adversarial comparison, and energy** **8**

**6 Model stacks and descent of observables** **9**

**7 A Gaussian example with higher homotopy** **11**

**8 Comparison, hypotheses, and scope** **13**

# 1 From generative laws to spaces of presentations

Chen's primer develops the mathematical connections among principal components, Gaussian latent variables, variational autoencoders, discrete and continuous diffusion, flows, autoregression, adversarial training, and energy-based modeling [1]. We retain this order at the level of concepts. The central change is to distinguish a probability law from a presentation of that law, and to keep the equivalences between presentations available for gluing.

Three kinds of data appear throughout. A *sampling program* composes kernels and can create dependent tuples. A *numerical objective* uses relative entropy, densities, a metric, or a test class. A *family of presentations* records how local descriptions fit together under changes of coordinates. A Markov category supplies operations of the first kind. The second kind requires analytic structure and hypotheses. The third kind is modeled by a groupoid or an object of an $\infty$-topos, according to how much topology the chosen equivalences carry.

For example, a Gaussian latent model is given by $z \sim N(0, I_k)$ and $x = Wz + b + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2 I_d)$ is independent. Its observable covariance is $WW^\top + \sigma^2 I_d$. Replacing $W$ by $Wg^{-1}$ for $g \in \mathrm{O}(k)$ gives the same observable law. A point in the quotient set records the orbit. The homotopy quotient also records the symmetry group at each loading. When loadings of rank one use four latent coordinates, the stabilizer acts on three unused coordinates. Retaining the topology of that action gives higher homotopy.

The main results are Theorems 6.1, 4.1, and 7.1. The first spells out how reparameterization-invariant laws and losses descend and computes the information forgotten by their joint map. The second identifies the reversal and relative entropy of a finite Markov path under explicit support assumptions. The third computes the orbit stack of the Gaussian example and constructs nontrivial global families with a fixed observable Gaussian law. The descent principle and chain rule are standard categorical and probabilistic tools [5, 3, 11]. The contribution is the choice of model data and equivalences, their use across the primer's constructions, and the explicit comparison of statistical equivalence with homotopy-coherent equivalence.

# 2 Kernels and evidence

## 2.1 The kernel calculus

Work first with standard Borel spaces. A kernel $K : X \rightsquigarrow Y$ assigns to each $x$ a probability $K_x$ on $Y$, measurably in $x$ when evaluated on any Borel set. The unit object is a singleton. The composite of $K$ with $L : Y \rightsquigarrow Z$ has

$$(LK)_x(C) = \int_Y L_y(C)\, K_x(\mathrm{d}y).$$

Dirac kernels give identity maps and represent measurable functions. Fubini–Tonelli gives associativity and the compatibility of composition with the tensor product of independent kernels. The copying and discarding operations are deterministic maps $\Delta_X(x) = (x, x)$ and $!_X : X \to *$. The product of the underlying measurable spaces serves as the tensor. The tensor does not satisfy the categorical product universal property for stochastic arrows, since two marginals can admit many couplings. This is the standard Markov-category setting for probability kernels [2, 3].

**Definition 2.1.** A latent presentation of an observable $X$ is a triple $(Z, \pi, K)$ consisting of a

latent space, a prior $\pi : * \rightsquigarrow Z$, and a decoder $K : Z \rightsquigarrow X$. Its joint and evidence are

$$P(\mathrm{d}z, \mathrm{d}x) = \pi(\mathrm{d}z)K_z(\mathrm{d}x), \qquad m = K\pi.$$

A variational presentation also specifies a recognition kernel $Q : X \rightsquigarrow Z$. A reparameterization from $(Z, \pi, K, Q)$ to $(Z', \pi', K', Q')$ is a Borel isomorphism $h : Z \to Z'$ satisfying

$$\pi' = h_*\pi, \qquad K'_{h(z)} = K_z, \qquad Q'_x = h_*Q_x.$$

All equalities in this definition are pointwise. A presentation without recognition omits the last equation.

Sampling the joint uses copying to retain a sampled $z$ while drawing $x$ conditional on it. In the finite case, $P(z, x) = \pi(z)K(x \mid z)$ and $m(x) = \sum_z P(z, x)$. In either case $P' = (h \times \mathrm{id})_*P$, so the evidence is preserved exactly. Two different presentations can give the same evidence even when there is no such $h$.

**Proposition 2.2** (Posterior transport)**.** *If $P$ disintegrates over $m$ as $P(\mathrm{d}z, \mathrm{d}x) = m(\mathrm{d}x)R_x(\mathrm{d}z)$, then $x \mapsto h_*R_x$ is a posterior for the reparameterized presentation. In standard Borel spaces posteriors exist and agree $m$-almost surely. In the finite case the posterior is uniquely defined at each $x$ with $m(x) > 0$.*

*Proof.* Integrating $h_*R_x$ against $m$ gives the joint pushforward $P'$, as can be checked on rectangles. Standard Borel disintegration gives existence and almost-sure uniqueness [3, 4]. For finite sets the normalization is $R_x(z) = P(z, x)/m(x)$ on the support of $m$. □

The base law matters. For $Z = X = \{0, 1\}$, let $K = \mathrm{id}$ and $\pi = \delta_0$. Changing $R_1$ changes a kernel on the whole of $X$ but preserves the joint. Comparing posterior versions at 1 requires additional choices. In particular the construction provides no preferred pointwise extension along a change of data law that assigns mass to that point. The same distinction appears when reversing diffusion and factoring an autoregressive law.

## 2.2 A chain rule on joint laws

Use natural logarithms. For probabilities $\mu, \nu$, set $D_{\mathrm{KL}}(\mu\|\nu) = \int \log(\mathrm{d}\mu/\mathrm{d}\nu)\,\mathrm{d}\mu$ when $\mu \ll \nu$, and $+\infty$ otherwise. For the following formula, kernels on null sets can be set arbitrarily when doing so has no effect on the integral.

**Proposition 2.3** (Relative entropy of a disintegration)**.** *Let $P(\mathrm{d}x, \mathrm{d}z) = p(\mathrm{d}x)P_x(\mathrm{d}z)$ and $Q(\mathrm{d}x, \mathrm{d}z) = q(\mathrm{d}x)Q_x(\mathrm{d}z)$ on a standard Borel product. If $p \ll q$ and $P_x \ll Q_x$ for $p$-almost every $x$, then*

$$D_{\mathrm{KL}}(P\|Q) = D_{\mathrm{KL}}(p\|q) + \int_X D_{\mathrm{KL}}(P_x\|Q_x)\,p(\mathrm{d}x). \tag{1}$$

*If $p \not\ll q$, the left side and the marginal term are both infinite. If $p \ll q$ and conditional absolute continuity fails on a set of positive $p$-measure, the left side and conditional term are both infinite.*

*Proof.* Write $a = \mathrm{d}p/\mathrm{d}q$. Under the absolute-continuity hypotheses, integrating sections of any $Q$-null set shows $P \ll Q$. Let $f = \mathrm{d}P/\mathrm{d}Q$; integrating $f(x, z)$ over $Q_x$ gives $a(x)$ for $q$-almost every $x$. The normalized ratio $f(x, z)/a(x)$ is a conditional derivative by uniqueness of disintegrations on

the set where $a > 0$. The logarithm of the joint ratio is the sum of the marginal and conditional logarithms there. Their negative parts are integrable since $-t \log t \leq 1/e$ for $0 < t \leq 1$, both for a probability reference and after averaging over $p$. This gives (1) in the extended sense. The marginal of an absolutely continuous joint is absolutely continuous, and the same derivative construction gives conditional absolute continuity $p$-almost everywhere. These implications give the remaining claims. □

**Proposition 2.4** (Contraction under kernels)**.** *For a common kernel $K$, $D_{\mathrm{KL}}(K\mu\|K\nu) \leq D_{\mathrm{KL}}(\mu\|\nu)$. For a Borel isomorphism, the pushforward inequality is an equality.*

*Proof.* The two laws $\mu(\mathrm{d}x)K_x(\mathrm{d}y)$ and $\nu(\mathrm{d}x)K_x(\mathrm{d}y)$ have relative entropy $D_{\mathrm{KL}}(\mu\|\nu)$ by (1). Disintegrate in the other order and use nonnegativity of the conditional term. For an isomorphism apply the inequality a second time to its inverse. □

This argument uses the analytic notion of relative entropy. The copying and discarding axioms by themselves do not select that notion. Proposition 2.4 also explains why invertible changes of latent coordinate preserve variational gaps and why applying a decoder can hide differences among latent laws.

# 3 Linear models, PCA, and variational objectives

## 3.1 From orthonormal frames to Gaussian presentations

Let $x$ be centered with finite covariance $C$ on $\mathbb{R}^d$. An orthonormal $k$-frame $U$ has encoder $U^\top$, decoder $U$, and projector $P = UU^\top$. The expected squared reconstruction error is

$$\mathbb{E}\|x - Px\|^2 = \operatorname{tr} C - \operatorname{tr}(U^\top C U).$$

For eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$, writing $a_i = \|U^\top e_i\|^2$ gives $0 \leq a_i \leq 1$, $\sum a_i = k$, and $\operatorname{tr}(U^\top C U) = \sum_i \lambda_i a_i$. This is at most $\sum_{i=1}^k \lambda_i$, attained by the top eigenspace. If $\lambda_k > \lambda_{k+1}$ and $0 < k < d$, equality fixes that subspace. An eigenvalue tie can leave several minimizing subspaces. Changing the orthonormal frame by $U \mapsto Ug$ preserves its projector, and the quotient of the Stiefel manifold $\mathrm{St}_k(\mathbb{R}^d)$ by the free $\mathrm{O}(k)$-action is the Grassmannian $\mathrm{Gr}_k(\mathbb{R}^d)$. The freeness follows from $U^\top U = I$. The homotopy quotient of this free principal bundle has the homotopy type of the Grassmannian. The bundle itself need not have a global frame. A deterministic autoencoder generalizes the reconstruction factorization to nonlinear maps. A probability law on its codes requires a further choice.

For probabilistic PCA, take $z \sim N(0, I_k)$ and $x \mid z \sim N(b + Wz, \sigma^2 I_d)$, with $\sigma > 0$ [12]. Then $m = N(b, WW^\top + \sigma^2 I_d)$. The conditional density of $z$ given $x$, up to constants independent of $z$, is proportional to

$$\exp\Big(-\tfrac{1}{2} z^\top A z + \sigma^{-2} z^\top W^\top (x - b)\Big), \qquad A = I_k + \sigma^{-2} W^\top W.$$

Completing the square gives covariance $A^{-1}$ and mean $A^{-1}\sigma^{-2}W^\top(x - b)$. Under $W \mapsto Wg^{-1}$ for $g \in \mathrm{O}(k)$, the posterior mean and covariance transform by $g$ and conjugation by $g$, respectively. This is an everywhere-defined instance of Proposition 2.2, since the observable density is positive. Section 7 uses a rank-deficient loading to expose a stabilizer. The calculation applies to that probabilistic PCA model even if the loading did not arise from maximum likelihood for a given dataset.

### 3.2 A variational gap and its invariance

Fix a common observation reference measure $\lambda$ and a measurable likelihood density $k(x \mid z)$ for $K_z$. Let $m_\lambda(x) = \int k(x \mid z)\pi(\mathrm{d}z)$. At points with $0 < m_\lambda(x) < \infty$ define $R_x(\mathrm{d}z) = k(x \mid z)\pi(\mathrm{d}z)/m_\lambda(x)$. If $Q_x \ll R_x$ and the displayed expectations exist as finite terms, then

$$\mathcal{L}(x) = \int_Z \log k(x \mid z)\, Q_x(\mathrm{d}z) - D_{\mathrm{KL}}(Q_x \| \pi), \tag{2}$$

$$\log m_\lambda(x) = \mathcal{L}(x) + D_{\mathrm{KL}}(Q_x \| R_x). \tag{3}$$

Indeed $\log(\mathrm{d}Q_x/\mathrm{d}R_x) = \log(\mathrm{d}Q_x/\mathrm{d}\pi) - \log k(x \mid z) + \log m_\lambda(x)$ on $Q_x$-almost every $z$. The gap is zero exactly when $Q_x = R_x$. Defining the objective by $\log m_\lambda(x) - D_{\mathrm{KL}}(Q_x \| R_x)$ extends it to $-\infty$ when the gap is infinite and the log evidence is finite. This form avoids an undefined subtraction when separate likelihood and prior terms fail to be integrable.

For an empirical or population data law $\rho$, a training objective is $\int \mathcal{L}\, \mathrm{d}\rho$ on the regime where that integral exists. When evidence is positive only almost everywhere, an objective against $\rho$ needs $\rho$ to see a region on which the density and conditional versions are specified. Positive Gaussian likelihoods offer a simple sufficient case. Exact E-steps set $Q$ to the old posterior. An M-step improving $\mathcal{L}$ at fixed $Q$ improves the new log evidence because the new gap is nonnegative. The argument assumes all compared objectives exist and the old gap vanishes on the data. This is the monotonicity calculation underlying EM [13]. A VAE parameterizes both $K$ and $Q$, with no assumption that the recognition family contains the exact posterior [14]. In a Gaussian recognition family, sampling via $z = \mu_\phi(x) + L_\phi(x)\varepsilon$, $\varepsilon \sim N(0, I)$, permits pathwise differentiation under the usual domination or integrable derivative hypotheses. A training gradient additionally requires a metric or chosen coordinates on parameter space.

**Lemma 3.1** (Loss under a latent change of coordinates)**.** *Choose likelihood densities that are transported pointwise along $h$, or fix compatible versions on a set of full data measure. Under a reparameterization preserving $\rho$ and $\lambda$, the log evidence, posterior gap, and integrable expected ELBO are unchanged.*

*Proof.* The likelihood expectation is unchanged by pushforward and $m_\lambda$ agrees wherever the chosen versions agree. The posterior is transported by Proposition 2.2, or by the explicit density formula. Relative entropy is invariant under the isomorphism by Proposition 2.4. Integrate the resulting equality against $\rho$. □

A value that is invariant under reparameterization can descend even when its Euclidean gradient does not. For a local coordinate change $\phi$, derivatives transform by $D\phi^\top$. A gradient vector field therefore descends if the local metrics are transported as well. Unqualified gradient descent with fixed Euclidean metrics on every chart has no such invariance under general smooth changes of coordinates. The separation between objective and update is compatible with categorical treatments that explicitly retain update rules [15].

## 4 Diffusion as path composition and reversal

### 4.1 Reversing a finite path law

Let $X_0, \ldots, X_T$ be finite nonempty sets. Let $\rho$ on $X_0$ and forward transitions $K_t : X_{t-1} \rightsquigarrow X_t$ be strictly positive. Write $Q(x_{0:T}) = \rho(x_0) \prod_{t=1}^{T} K_t(x_t \mid x_{t-1})$ and $\mu_t = \mathrm{Law}_Q(X_t)$. The positivity

assumptions keep the proof pointwise. Restricting attention to the forward support or taking measurable versions gives related formulas with almost-sure qualifications.

**Theorem 4.1** (Reversal and path relative entropy)**.** *Define*

$$R_t(u \mid v) = \frac{\mu_{t-1}(u)K_t(v \mid u)}{\mu_t(v)}.$$

*Then* $Q(x_{0:T}) = \mu_T(x_T)\prod_{t=1}^T R_t(x_{t-1} \mid x_t)$*. For a backward model* $P_\theta = \nu(x_T)\prod_{t=1}^T S_{\theta,t}(x_{t-1} \mid x_t)$*,*

$$D_{\mathrm{KL}}(Q\|P_\theta) = D_{\mathrm{KL}}(\mu_T\|\nu) + \sum_{t=1}^{T}\sum_{v\in X_t} \mu_t(v) D_{\mathrm{KL}}(R_t(- \mid v)\|S_{\theta,t}(- \mid v)). \tag{4}$$

*The reversal of a coarse transition* $K_t \cdots K_{s+1}$ *with starting law* $\mu_s$ *is* $R_{s+1}\cdots R_t$*. Consequently reversal is contravariantly functorial for the finite category whose objects are positive states* $(X, \mu)$ *and whose arrows are kernels that preserve the specified source and target states.*

*Proof.* Each $R_t(- \mid v)$ sums to one by definition of $\mu_t$. Multiplying the formulas cancels all intermediate marginals. In the strictly positive comparison case, take logarithms of the two backward factorizations and average under $Q$. This gives the terminal term and the individual conditional terms. If $P_\theta$ vanishes on a path, either a terminal term or a conditional term is infinite. Summing the reverse factorization over the intermediate variables gives the joint of the endpoints as $\mu_t$ times the claimed composite. Uniqueness of pointwise conditioning at $\mu_t > 0$ proves the coarse-transition statement. Reversing the identity gives the identity, and applying the reversal twice recovers the kernel by the same joint-law formula. □

The choice of objects in the last sentence is essential. The same underlying forward kernel can have different Bayes reversals under different input laws. With a binary flip kernel of probability $1/4$ and $\rho(1) = 1/4$, one finds $\mu_1(1) = 3/8$, $R_1(1 \mid 1) = 1/2$, and $R_1(1 \mid 0) = 1/10$. Starting instead at the fair law gives the original flip kernel as its own reverse. Bayesian composition with state dependence has also been studied using lenses [16]; here the state is part of the object so that the supported reverse arrow is determined.

## 4.2 The data-conditional diffusion bound

Let $m_\theta$ be the $X_0$-marginal of $P_\theta$. Use $Q(x_{1:T} \mid x_0)$ as a recognition kernel for the latent path. For strictly positive $P_\theta$ the pointwise lower bound is

$$\mathcal{L}_\theta(x_0) = \mathbb{E}_{Q(-|x_0)}[\log P_\theta(x_{0:T}) - \log Q(x_{1:T} \mid x_0)] \le \log m_\theta(x_0).$$

The gap is the divergence to $P_\theta(x_{1:T} \mid x_0)$. Averaging the bound gives

$$-\mathbb{E}_\rho \mathcal{L}_\theta(X_0) = H(\rho) + D_{\mathrm{KL}}(Q\|P_\theta), \qquad H(\rho) = -\sum_x \rho(x)\log\rho(x).$$

The unconditional reverse $R_t$ in Theorem 4.1 need not coincide with $Q(x_{t-1} \mid x_t, x_0)$. The latter appears when one factors the forward path in reverse order conditional on the observed $x_0$. With $q_T^x = Q(X_T \mid X_0 = x)$, this gives the stepwise loss

$$\begin{aligned}-\mathcal{L}_\theta(x) = D_{\mathrm{KL}}(q_T^x\|\nu) &+ \sum_{t=2}^{T}\mathbb{E}_{Q(X_t|x)} D_{\mathrm{KL}}(Q(X_{t-1} \mid X_t, x) \,\|\, S_{\theta,t}(- \mid X_t)) \\ &- \mathbb{E}_{Q(X_1|x)}\log S_{\theta,1}(x \mid X_1).\end{aligned}$$

For $t \geq 2$, the average of a conditional divergence in this display splits further into

$$I_Q(X_{t-1}; X_0 \mid X_t) + \mathbb{E}_{\mu_t} D_{\mathrm{KL}}(R_t(- \mid X_t) \| S_{\theta,t}(- \mid X_t)).$$

This follows by adding and subtracting the logarithm of $R_t$ and using the conditional expectation over $X_0$. The first term is independent of the fitted reverse kernel. These identities separate two legitimate decompositions of the same averaged objective and explain where the training-time conditioning in a DDPM comes from [17, 1]. The Gaussian square-loss simplification in a DDPM uses additional covariance and parameterization assumptions. The path law alone does not select the simplified loss or its timestep weights.

### 4.3 Continuous dynamics and the score

A deterministic flow of maps can give pushforward on laws. A time-inhomogeneous Markov evolution instead gives kernels $K_{s,t}$ with $K_{s,u} = K_{t,u} K_{s,t}$ for $s \leq t \leq u$. To discuss density evolution, use Euclidean coordinates and assume sufficient smoothness and decay to differentiate under integration and integrate by parts. For

$$\mathrm{d}X_t = b_t(X_t)\,\mathrm{d}t + \sigma_t\,\mathrm{d}W_t, \qquad a_t = \sigma_t \sigma_t^{\top},$$

with $\sigma_t$ independent of position, the generator is $L_t f = b_t \cdot \nabla f + \frac{1}{2} a_t : \nabla^2 f$. The density satisfies

$$\partial_t p_t = L_t^* p_t = -\nabla \cdot (b_t p_t) + \tfrac{1}{2} a_t : \nabla^2 p_t.$$

With zero diffusion this reduces to the continuity equation. Under conditions ensuring a time reversal with a smooth positive density and the indicated generators, the reverse generator at reverse time $s$, with $t = T - s$, is

$$\overleftarrow{L}_s f = p_t^{-1}[L_t^*(p_t f) - f L_t^* p_t] = (-b_t + a_t \nabla \log p_t) \cdot \nabla f + \tfrac{1}{2} a_t : \nabla^2 f.$$

For the first equality, backward expectation from $t$ to $t - \epsilon$ applied to $f$ is the forward adjoint propagator on $p_{t-\epsilon} f$, divided by $p_t$. Expanding the propagator to first order and using $\partial_t p_t = L_t^* p_t$ gives the formula. The second equality follows by differentiation. A position-dependent diffusion matrix brings extra derivative terms. The positive-density and regularity hypotheses are analytic inputs to the derivation [18].

The Fokker–Planck equation also takes the form of a continuity equation with velocity $v_t = b_t - \frac{1}{2} a_t \nabla \log p_t$. If this velocity generates a flow and the associated continuity equation has a unique solution for $p_0$, the flow has the same one-time marginals as the diffusion. Its path law can differ. For Brownian motion with a smooth nondegenerate initial Gaussian, the diffusion has positive quadratic variation while the deterministic flow of the displayed velocity has finite variation on compact time intervals. This gives another instance in which an observable law, here the list of one-time laws, forgets part of a presentation.

For a corruption with conditional density $k_t(y \mid x)$ and data law $\rho$, set $p_t(y) = \int k_t(y \mid x) \rho(\mathrm{d}x)$. Under dominated differentiation and $p_t(y) > 0$,

$$\nabla_y \log p_t(y) = \mathbb{E}[\nabla_y \log k_t(y \mid X_0) \mid X_t = y]. \tag{5}$$

Differentiate the marginal density and divide by it. The factor $k_t(y \mid x)/p_t(y)$ is the posterior density relative to $\rho$. With square-integrable scores, conditional expectation is the orthogonal

projection onto functions of $X_t$, so the denoising regression has minimizer $\nabla \log p_t$. For additive noise $Y = X + \tau\varepsilon$, $\varepsilon \sim N(0, I)$ independent, the identity reads $\nabla \log p_Y(y) = (\mathbb{E}[X \mid Y = y] - y)/\tau^2$ wherever the quantities exist. Ordinary score matching also follows from integration by parts. If $p, s$ are smooth, the boundary term vanishes and expectations exist, the squared error to $\nabla \log p$, divided by two, equals $\mathbb{E}_p[\|s\|^2/2 + \nabla \cdot s]$ plus a term independent of $s$ [19]. A coordinate derivative and a metric enter these constructions. They are retained as geometric structure when comparing charts.

# 5 Exact density, adversarial comparison, and energy

## 5.1 Normalizing flows and autoregression

If $f : \mathbb{R}^d \to \mathbb{R}^d$ is a $C^1$ diffeomorphism, $f_*p$ is the law of $f(Z)$, and for Lebesgue densities,

$$\log p_{f(Z)}(x) = \log p_Z(f^{-1}(x)) + \log |\det Df^{-1}(x)|$$

where the logarithms are finite. The inverse determinant is a multiplicative cocycle under composition, so its log gives the sum of the layerwise corrections. For a coupling layer $(u, v) \mapsto (u, v \odot \exp s(u) + t(u))$, the Jacobian is block triangular. Its log determinant is $\sum_i s_i(u)$, and the inverse computes $v = (v' - t(u)) \odot \exp(-s(u))$. This tractability comes from the structure of the chosen deterministic isomorphisms [20, 21]. Under an invertible measurable change of coordinates, the pushforward law and KL invariance require no Jacobian. The Jacobian expresses the change relative to smooth reference volume.

A finite product law can be disintegrated successively along the projections $X_{\le t} \to X_{<t}$. On standard Borel products this yields $P(\mathrm{d}x_{1:T}) = \prod_t K_t(\mathrm{d}x_t \mid x_{<t})$, with conditional versions unique almost surely under the corresponding prefix law. Conversely, any such kernels compose to a joint. The autoregressive log likelihood is a sum wherever a compatible joint density exists and the log terms are defined. The ordering is part of the presentation. For a binary law with $P(0,0) = P(1,1) = 1/2$, either coordinate can be sampled first, then copied. Reordering preserves the joint after coordinate permutation but changes which coordinate has a stochastic conditional and which has a deterministic one. The factorization follows from disintegration. The computational restrictions imposed by an autoregressive network are additional choices.

## 5.2 Adversarial objectives and transport

A deterministic generator $G$ and a noise law $\pi$ define $\nu = G_*\pi$. Stochastic generators also define a law by kernel composition. Let $\mu$ be the data law and set $\eta = \mu + \nu$, $a = \mathrm{d}\mu/\mathrm{d}\eta$, $b = \mathrm{d}\nu/\mathrm{d}\eta$. The original minimax GAN value with unrestricted measurable discriminator $D : X \to [0, 1]$ is the supremum of

$$\int [a \log D + b \log(1 - D)] \, \mathrm{d}\eta.$$

The pointwise maximum has $D = a$ since $a + b = 1$, with limiting conventions at the endpoints. Putting $M = (\mu + \nu)/2$ gives the optimal value $-\log 4 + D_{\mathrm{KL}}(\mu \| M) + D_{\mathrm{KL}}(\nu \| M)$. This statement concerns the idealized minimax objective [22]. Restricting the discriminator class gives a different comparison that depends on the chosen tests. Mutually singular data and generator laws saturate the unrestricted value at zero even when their supports are arbitrarily close in a metric.

For a Polish metric space $(X, d)$ and probability measures of finite first moment, the Wasserstein objective

$$W_1(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \int d(x,y)\,\gamma(\mathrm{d}x,\mathrm{d}y) = \sup_{\mathrm{Lip}(f)\le 1} \left(\int f\,\mathrm{d}\mu - \int f\,\mathrm{d}\nu\right)$$

uses the coupling set and Kantorovich–Rubinstein duality [24, 23, 10]. Couplings are joint states with prescribed marginals. The cost uses the metric, and the dual uses the corresponding test class. Isometries preserve both. For $L$-Lipschitz $f$, pushing forward a coupling gives $W_1(f_*\mu, f_*\nu) \le LW_1(\mu,\nu)$. In particular, two Dirac masses at distance $\epsilon > 0$ give GAN value zero and Wasserstein distance $\epsilon$. The joint-state construction is common to both approaches while the measurements of discrepancy differ.

### 5.3 Energy, weights, and normalized laws

On a space with reference measure $\lambda$, a measurable energy $E$ yields

$$Z_E = \int e^{-E(x)}\lambda(\mathrm{d}x), \qquad p_E(\mathrm{d}x) = Z_E^{-1} e^{-E(x)}\lambda(\mathrm{d}x)$$

when $0 < Z_E < \infty$. Before normalization $e^{-E}\lambda$ is a weighted measure. Reweighting a law by $w$, when integrable and nonnegative, and then reweighting by $v$ agrees after normalization with reweighting once by $vw$, if both stages have finite positive mass. This follows by canceling the first normalizing constant. Zero weight everywhere or an infinite normalizer removes the corresponding normalized operation. Treating unnormalized weights and probabilities separately is useful for higher-order probabilistic semantics [7, 8].

Adding a constant to the energy fixes $p_E$. Conversely, equality of the two normalized laws with the same $\lambda$ implies that their finite energies differ by a constant almost everywhere on their common support. For differentiable energies on an open subset with Lebesgue reference, the score is $-\nabla E$. On a disconnected domain, equal scores allow independent constants on each connected component. For example, on the two intervals $(-2,-1)$ and $(1,2)$, scores are zero for every energy constant on each interval, but changing the difference of constants changes the relative mass of the two intervals. Thus normalization and global information enter at a stage where a local score may leave degrees of freedom. For a smooth energy on $\mathbb{R}^d$, the overdamped Langevin diffusion $\mathrm{d}X_t = -\nabla E(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t$ has $p_E$ as a stationary density under assumptions that justify the forward equation and its identification with the semigroup. Indeed $\nabla p_E = -p_E \nabla E$ makes the Fokker–Planck probability flux vanish. Stationarity alone gives no convergence rate and allows non-ergodic dynamics on disconnected domains. Noise-contrastive estimation instead compares data to a chosen noise law through a classification objective [25]. Its test and reference choices are additional to the bare energy presentation.

## 6 Model stacks and descent of observables

### 6.1 Charts and allowed equivalences

There are two different categorical levels in the construction. One level has sample spaces and stochastic transitions. Its composition accounts for drawing samples. The other level has presentations as points and specified invertible reparameterizations as paths. Its composition accounts for changes of description. Taking the nerve of the ordinary kernel category provides

coherent composition in the first sense but retains discrete mapping spaces. The examples below introduce higher homotopy through the topology of the allowed reparameterizations.

Let $\mathcal{X}$ be an $\infty$-topos, and let $G$ be a group object acting on $M \in \mathcal{X}$. Think of $M$ as a family of presentations with a chosen format and $G$ as the changes of latent coordinate. Write $M//G = |M \times G^{\bullet}|$ for the realization of the action groupoid. A map from $U$ to $M//G$ is a $U$-family of presentations. Here $U$ is a test object, such as a smooth parameter manifold. Pullback in $U$ restricts the family. The notation concerns the external stack of descriptions. It does not posit a probability monad on every $\infty$-topos.

For one concrete analytic model of $M$, fix finite sets $X, Z$. A strictly positive prior, decoder, and recognition matrix give a point of a finite product of open simplices. Take the represented smooth sheaf of this manifold on smooth test manifolds and allow the finite permutation group of $Z$ to act on rows and columns. The evidence map to the open simplex on $X$ is polynomial. At a fixed data law $\rho$, the expected ELBO is smooth on the positive domain and is invariant under those permutations. Thus both maps are morphisms of smooth sheaves. In Gaussian parameter families one can use the represented sheaf of the Gaussian mean and covariance parameters for evidence, and any smooth invariant loss on its specified domain. The rank-one orbit below requires only the constant exact-posterior objective. For merely measurable models the same conclusions require a chart object and a target of invariants on which the maps are defined. Pointwise density formulas alone do not supply these objects.

**Theorem 6.1** (Descent of laws and objectives)**.** *Let $d : M \to D$ record an observable law and let $l : M \to L$ record a loss. Suppose that together they define a map of action groupoids to the constant groupoid on $A = D \times L$. In particular, strict invariance on a sheaf of models and its action group suffices. Let $\overline{a} : M//G \to A$ denote the induced map.*

1. *For an effective epimorphism $V \to U$, the space $\mathrm{Map}_{\mathcal{X}}(U, M//G)$ is the totalization of $\mathrm{Map}_{\mathcal{X}}(V^{\times_U (\bullet+1)}, M//G)$.*

2. *For any $B \to A$ there is a canonical equivalence*

$$B \times_A (M//G) \;\simeq\; (B \times_A M)//G,$$

   *where the action on $B$ is trivial. Consequently families with a prescribed law and loss satisfy the same descent over the slice $\mathcal{X}_{/A}$.*

3. *A map $U \to M//G$ amounts to a principal $G$-torsor $P \to U$ together with a compatible equivariant map $P \to M$. A lift $U \to M$ of the family through the atlas is equivalent to a choice of section of the torsor.*

*Proof.* By effectiveness, the realization of the Cech nerve of $V \to U$ is $U$. Mapping out of this colimit into $M//G$ yields the first equivalence, including its higher coherence data. Invariance gives an augmentation of the action nerve to $A$, so it can be realized in $\mathcal{X}_{/A}$. Colimits in an $\infty$-topos are universal. Pulling this realization back along $B \to A$ therefore gives the realization of the pulled-back action nerve, which is the claimed homotopy quotient. The same Cech argument applies in the slice.

For the third claim, the action nerve is a groupoid object and hence effective in an $\infty$-topos [5, Theorem 6.1.0.6]. Pulling back $M \to M//G$ along $U \to M//G$ gives an effective epimorphism $P \to U$ whose pair groupoid is the pullback of the action groupoid. In particular the canonical

map $P \times G \to P \times_U P$ is an equivalence. This is the torsor property. Conversely, an equivariant map from a torsor to $M$ gives a map of action nerves. Realizing and using $P//G \simeq U$ produces the family. These operations are inverse by the pullback description and descent. A section is precisely a lift to the pulled-back atlas and hence to $M$. □

In item 2, fixing a value of the joint observable $(d, l)$ means taking its homotopy fiber. Given a specified invariant $a : U \to A$, the space of presentations over $a$ is $\mathrm{Map}_{\mathcal{X}_{/A}}(U, M//G)$. Thus equality of observable laws and losses records only the base of this map. The fiber can contain presentations with automorphisms, with nontrivial transitions on overlaps, or with higher families of equivalences. Lemma 3.1 supplies the invariance hypothesis for ELBOs when the analytic versions and integrability assumptions are met. For metrics and test classes, one uses charts preserving that structure or allows it to vary compatibly as part of the presentation.

**Proposition 6.2** (Stabilizer on a transitive orbit)**.** *Suppose an orbit has an atlas $O = G/H$ for a subgroup object $H \subseteq G$, with $G \to O$ locally split in the cases of sheaves considered here and $G \times_O G \simeq G \times H$. Then $O//G \simeq BH$. In particular a constant law on the orbit can have a noncontractible space of presentations.*

*Proof.* By Theorem 6.1, a map into $O//G$ is a $G$-torsor with an equivariant map to $G/H$. Pull back the map on the distinguished coset to get an $H$-torsor. This can be checked after locally trivializing the $G$-torsor and lifting the local map to $G/H$ along $G \to G/H$. Conversely, an $H$-torsor $Q$ induces a $G$-torsor $Q \times^H G$ with a compatible map to $G/H$, using consistent left and right action conventions. Local trivializations show these operations are inverse. The functor of $H$-torsors is represented by $BH = */\!/H$, which proves the equivalence by Yoneda. □

### 6.2 A finite model with nontrivial gluing

Let $Z = \{1, 2, 3\}$, $X = \{0, 1\}$, and take

$$\pi = (1/4,\ 1/4,\ 1/2), \qquad K(1 \mid -) = (3/4,\ 3/4,\ 1/4).$$

The evidence is fair. With $Q$ equal to the posterior, its $x = 1$ row is $(3/8, 3/8, 1/4)$, and its $x = 0$ row is $(1/8, 1/8, 3/4)$. The ELBO is $-\log 2$ at either observation. The swap of states $1, 2$ preserves all three kernels or states in the presentation, while a permutation moving state 3 fails to preserve the prior. The automorphism group of the presentation is exactly $C_2$.

On the site of opens in $S^1$, take the locally constant stack determined by this model and its allowed automorphisms. The map $S^1 \to S^1$, $z \mapsto z^2$, is a nontrivial principal $C_2$-bundle. It supplies a global presentation in the stack whose charts are the finite model above. It has no global labeling compatible with the transition functions, since a section of the double cover would have to satisfy $2 \deg(s) = 1$. The trivial bundle has the same observable law and ELBO, with a global labeling. This distinction is already visible to a 1-groupoid. A plain set of equivalence classes of local models loses it. The example also shows that stack descent yields a global family even when the family admits no global representative in the chosen atlas.

## 7 A Gaussian example with higher homotopy

Fix $w \in \mathbb{R}^d \setminus \{0\}$, $\sigma > 0$, and $b \in \mathbb{R}^d$. Use four latent coordinates with standard Gaussian prior and initial loading $W_0 = we_1^\top$. The group $G = \mathrm{O}(4)$ acts on loadings by $g \cdot W = Wg^{-1}$. Let

$O = \{W_0 g^{-1} : g \in G\}$, a single orbit, and use exact recognition. All its loadings have observable law

$$m = N(b, ww^\top + \sigma^2 I_d).$$

In particular the pointwise log evidence and the expected exact ELBO are constant on the orbit for any fixed data law of finite second moment.

**Theorem 7.1** (A fixed Gaussian law with higher presentation space)**.** *The smooth action stack for $O$ and $G$ on the site of smooth manifolds and open covers is equivalent to the classifying stack of smooth $\mathrm{O}(3)$-torsors. For the topological action, its Borel homotopy quotient has the homotopy type $B\mathrm{O}(3)$, with*

$$\pi_4(B\mathrm{O}(3)) \cong \mathbb{Z}.$$

*There are smooth families over $S^4$ with the same evidence and exact ELBO as the trivial family which admit no global lift through $O$.*

*Proof.* Since $w \neq 0$, $W_0 g^{-1} = W_0$ is equivalent to $ge_1 = e_1$. The stabilizer $H$ consists of the orthogonal transformations of $e_1^\perp$ and is isomorphic to $\mathrm{O}(3)$. The orbit consists of $wv^\top$ for $v \in S^3$. The map $\mathrm{O}(4) \to S^3$, $g \mapsto ge_1$, is a principal $H$-bundle with smooth local sections, as one sees by completing $v$ to a local orthonormal frame. Proposition 6.2 now gives the smooth stack claim.

For the topological quotient use a contractible free $G$-space $EG$. Since $O = G/H$, one has $EG \times_G O \cong EG/H$, a model for $BH$. The long exact sequence of $H \to EH \to BH$ gives $\pi_4 BH \simeq \pi_3 H$, based in the identity component of $H$. That component is $\mathrm{SO}(3)$. Unit quaternions form the universal double cover $S^3 \to \mathrm{SO}(3)$ via conjugation on imaginary quaternions, giving $\pi_3 \mathrm{SO}(3) \cong \pi_3 S^3 \cong \mathbb{Z}$.

For a concrete smooth family, glue trivial $\mathrm{SO}(3)$-bundles on two hemispherical neighborhoods of $S^4$ using the unit-quaternion double cover on the equator, extended over its collar. Regard the result as an $H$-torsor by extension of structure group. The clutching map represents a generator of $\pi_3 H$ and so the $H$-torsor is nontrivial [6]. A lift of its classifying map through $O$ would be a smooth map $v : S^4 \to S^3$. The corresponding $H$-torsor would be the pullback of $\mathrm{O}(4) \to S^3$ along $v$. The bundle $\mathrm{O}(4) \to S^3$ has a global smooth section, using left multiplication by the unit quaternion $v$ to take 1 to $v$ in $\mathbb{R}^4$. Every such pullback is therefore trivial, contradicting the clutching construction. □

The last part illustrates a distinction between the two torsors in the proof. Under $O//G \simeq BH$ the family determines an $H$-torsor. Pulling back the atlas $O \to O//G$ gives the induced $G$-torsor. A lift trivializes the latter. The existence of a section of $G \to O$ in this example ensures that any lifted family also has trivial $H$-torsor. This implication need not hold for a general homogeneous orbit.

**Proposition 7.2** (Orthogonal stabilizers by rank)**.** *For $W : \mathbb{R}^k \to \mathbb{R}^d$ of rank $r$, the stabilizer of $W$ for the right orthogonal action is isomorphic to $\mathrm{O}(k-r)$. Its topological orbit homotopy quotient is $B\mathrm{O}(k-r)$.*

*Proof.* The equality $Wg^{-1} = W$ implies $gW^\top = W^\top$, so $g$ fixes the image of $W^\top$ pointwise. Conversely, this pointwise identity implies the original equality. Orthogonal transformations fixing that subspace are precisely the transformations of its orthogonal complement, $\ker W$. The orbit is the homogeneous space of the compact Lie group by this closed stabilizer. Applying the preceding Borel-quotient calculation proves the claim. □

A full-column-rank loading has trivial stabilizer for this specified action. Rank deficiency gives an explicit source of isotropy, independent of the numerical likelihood. Other classes of changes of coordinates can produce a different stack. The homotopy computation uses the Lie group topology. Treating its real points as a discrete group would give a 1-type and remove the fourth homotopy group. The smooth stack represents families over smooth test manifolds, whereas $B\mathrm{O}(3)$ in the homotopy computation is a space built from the topological group. The clutching construction exhibits their relation without identifying those two objects.

## 8 Comparison, hypotheses, and scope

Table 1 lists what the kernel interpretation retains and what is supplied by each analytic model. The chapter numbers refer to the revised source [1].

| Primer topic | Structural operation | Analytic or geometric input |
|---|---|---|
| PCA and PPCA (1–2) | Retract, latent joint, pushforward | Inner product, finite covariance, Gaussian prior and noise |
| VAE and EM (2–3) | Recognition kernel, posterior, joint relative entropy | Likelihood versions, integrability, optimization or derivative assumptions |
| Discrete diffusion (4) | Markov path, state-indexed reversal | Support, chosen terminal law, Gaussian assumptions for square loss |
| Dynamics and scores (5–6) | Evolution of states and conditional expectation | Generator, reference volume, differentiability and regularity of reversal |
| Flows and autoregression (7) | Invertible pushforward or prefix disintegration | Jacobians, tractable layers or chosen ordering and versions |
| GAN and WGAN (8) | Observable generator law and comparison via tests or couplings | Discriminator class or metric and moment bounds |
| Energy models (8) | Reweighting followed by normalization | Reference measure and positive finite partition function |

Table 1: The structures used in the categorical reading.

The measurable and smooth roles should be distinguished. The standard Borel kernel model gives disintegration and has no built-in function objects for unrestricted higher-order probability. Quasi-Borel spaces supply a cartesian closed setting and a probability monad for higher-order probabilistic programs [7]. Commutative monads account for composing independent distributions in appropriate settings [8, 9]. These constructions address different requirements from descent of model presentations. In particular a bare $\infty$-topos does not supply the analytic ingredients used in Sections 2–5. A generalization to a new site requires a chosen class of stochastic programs, an invariant chart map, and the hypotheses for any claimed numerical operation. Our examples specify those ingredients directly.

The distinction between reparameterization and observational equivalence is relevant to identifiability. If $K\pi = K'\pi'$, no latent equivalence need exist. Even when parameters do lie in the same orbit, stabilizers can act nontrivially on the space of descriptions. Passing to the evidence therefore forgets more than just the names of latent variables. The same issue applies to recognition families with extra symmetries or to different autoregressive factorizations that happen to give the same law. For optimization, the loss in a descent statement is a value. Claims about a learning update require compatible derivative, metric, or optimizer data and an equivariance proof for the update

itself.

The standard results used here include stochastic kernel composition, disintegration, relative entropy identities, the classical generative objectives, and effectiveness of groupoid objects. The additional results in the manuscript concern the joint descent of a law and an objective for specified model symmetries, the resulting relative fibers, and worked models that distinguish trivial and nontrivial gluing. The finite and Gaussian examples give independent exact checks of the claims. The continuous-time derivation assumes a well-defined reversal with regular densities and does not establish an SDE existence theorem. No claim about statistical consistency, training convergence, or benchmark performance follows from the stack construction.

**Provenance and use of automation.** This is an independent mathematical treatment of the cited primer. The primer's author is cited for the source material and is not represented as a coauthor. OpenAI Codex assisted source retrieval, mathematical drafting, and typesetting. The submission uses the byline Anonymous. The evidence in the manuscript consists of proofs, algebraic examples, and comparison with the cited literature.

## References

[1] T. Chen. *The Little Book of Generative AI Foundations: An Intuitive Mathematical Primer*. arXiv:2605.29713v2, 8 September 2026. https://arxiv.org/abs/2605.29713v2

[2] M. Giry. A categorical approach to probability theory. In *Categorical Aspects of Topology and Analysis*, LNM 915, 68–85, 1982. https://doi.org/10.1007/BFb0092872

[3] T. Fritz. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics* 370, 107239, 2020. https://arxiv.org/abs/1908.07021

[4] O. Kallenberg. *Foundations of Modern Probability*. 3rd ed., Springer, 2021. https://doi.org/10.1007/978-3-030-61871-1

[5] J. Lurie. *Higher Topos Theory*. Princeton University Press, 2009. https://arxiv.org/abs/math/0608040

[6] A. Hatcher. *Vector Bundles and K-Theory*. Version 2.2, 2017, Section 1.2. https://pi.math.cornell.edu/~hatcher/VBKT/VBpage.html

[7] C. Heunen, O. Kammar, S. Staton, and H. Yang. A convenient category for higher-order probability theory. LICS, 2017. https://arxiv.org/abs/1701.02547

[8] A. Kock. Commutative monads as a theory of distributions. *Theory and Applications of Categories* 26, 97–131, 2012. https://arxiv.org/abs/1108.5952

[9] T. Fritz and P. Perrone. Bimonoidal structure of probability monads. *Electronic Notes in Theoretical Computer Science* 341, 97–114, 2018. https://arxiv.org/abs/1804.03527

[10] T. Fritz and P. Perrone. A probability monad as the colimit of spaces of finite samples. *Theory and Applications of Categories* 34, 170–220, 2019. https://arxiv.org/abs/1712.05363

[11] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. 2nd ed., Wiley, 2006. https://doi.org/10.1002/047174882X

[12] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *JRSS B* 61, 611–622, 1999. https://doi.org/10.1111/1467-9868.00196

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *JRSS B* 39, 1–22, 1977. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

[14] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. ICLR, 2014. https://arxiv.org/abs/1312.6114

[15] B. Fong, D. I. Spivak, and R. Tuyéras. Backprop as functor: A compositional perspective on supervised learning. LICS, 2019. https://arxiv.org/abs/1711.10455

[16] T. St Clere Smithe. Bayesian updates compose optically. arXiv:2006.01631, 2020. https://arxiv.org/abs/2006.01631

[17] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. NeurIPS, 2020. https://arxiv.org/abs/2006.11239

[18] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. ICLR, 2021. https://arxiv.org/abs/2011.13456

[19] A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *JMLR* 6, 695–709, 2005. https://www.jmlr.org/papers/v6/hyvarinen05a.html

[20] D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. ICML, 2015. https://arxiv.org/abs/1505.05770

[21] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *JMLR* 22(57), 1–64, 2021. https://arxiv.org/abs/1912.02762

[22] I. J. Goodfellow et al. Generative adversarial nets. NeurIPS, 2014. https://arxiv.org/abs/1406.2661

[23] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. ICML, 2017. https://proceedings.mlr.press/v70/arjovsky17a.html

[24] C. Villani. *Optimal Transport: Old and New*. Springer, 2009. https://doi.org/10.1007/978-3-540-71050-9

[25] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. AISTATS, 2010. https://proceedings.mlr.press/v9/gutmann10a.html