Pure Mathematics as a Language for Machine Learning and Data Analysis

N-20260302

§1.1 Backpropagation as cotangent pullback

The previous backpropagation calculation can be read in a categorical way without making it mystical. A forward neural network is a composite of maps between parameter spaces, activation spaces, and finally the one-dimensional loss space. If

$$L : Y \to \mathbb{R}$$

is the loss, then its differential at the output is a cotangent vector

$$dL_y \in T_y^* Y.$$

Backpropagation pulls this covector backward along the forward computational graph. In this sense the derivative operation used in training is contravariant:

$$f : X \to Y \quad \Longrightarrow \quad f^* : T^*Y \to T^*X.$$

Forward maps push tangent perturbations forward; reverse-mode automatic differentiation pulls covectors backward.

For a concrete PyTorch-shaped layer, take

$$X \in \mathbb{R}^{b\times d}, \qquad W \in \mathbb{R}^{n\times d}, \qquad Y = XW^T \in \mathbb{R}^{b\times n}.$$

In indices,

$$Y_{\alpha i} = \sum_{k=1}^{d} X_{\alpha k} W_{ik}.$$

The full Jacobian with respect to the weight matrix has entries

$$\frac{\partial Y_{\alpha i}}{\partial W_{jk}} = \delta_{ij} X_{\alpha k},$$

so its tensor shape is

$$(b, n, n, d).$$

This is the honest tensor object. Neural-network code does not materialize it because the upstream gradient

$$G = \frac{\partial L}{\partial Y} \in \mathbb{R}^{b\times n}$$

immediately contracts it:

$$\begin{aligned} \frac{\partial L}{\partial W_{jk}} &= \sum_{\alpha=1}^{b} \sum_{i=1}^{n} \frac{\partial L}{\partial Y_{\alpha i}} \frac{\partial Y_{\alpha i}}{\partial W_{jk}} \\ &= \sum_{\alpha=1}^{b} G_{\alpha j} X_{\alpha k}. \end{aligned}$$

Therefore

$$\boxed{\frac{\partial L}{\partial W} = G^T X.}$$

If one stores weights in the column convention $W_{\text{col}} \in \mathbb{R}^{d\times n}$ and writes $Y = XW_{\text{col}}$, the same contraction is

$$\frac{\partial L}{\partial W_{\text{col}}} = X^T G.$$

The categorical phrase "pull back the covector" is exactly this matrix multiplication: the high-order Jacobian collapses because it is being paired with a covector from the scalar loss.

§1.2 Geometric data analysis

The original note is an ambitious AI-assisted essay about the role of pure mathematics in machine learning. Its most useful part is not the rhetoric, but the catalogue of mathematical languages that repeatedly appear when data are no longer well described by coordinates alone.

A point cloud is often treated as a finite sample from a hidden space. Topological data analysis makes this idea precise by replacing the cloud with a filtered simplicial complex. For a scale parameter ε, one forms for instance the Vietoris–Rips complex

$$VR_\varepsilon(X) = \{\sigma \subseteq X : \ d(x,y) \le \varepsilon \text{ for all } x, y \in \sigma\}.$$

As ε grows, connected components merge, loops appear and disappear, and higher-dimensional holes may be born and die. Persistent homology records this evolution. The chain complex viewpoint is governed by the boundary identity

$$\partial_{k-1} \circ \partial_k = 0,$$

and homology is

$$H_k = \ker \partial_k / \operatorname{im} \partial_{k+1}.$$

This quotient is the mathematical form of the phrase "holes that are cycles but not boundaries."

§1.3 Manifold hypothesis and embeddings

The manifold hypothesis says that high-dimensional data often concentrate near a lower-dimensional manifold $M \subseteq \mathbb{R}^N$. If M is a smooth d-manifold, Whitney's embedding theorem says that M can be embedded in $\mathbb{R}^{2d}$, and generically in a slightly larger ambient space. In machine learning this gives a conceptual reason why dimensionality reduction is not only compression: it tries to recover the intrinsic coordinates of M.

The practical warning is that an embedding preserves the manifold as a smooth object, but not necessarily distances, probability densities, or task-relevant semantics. PCA preserves a linear subspace model; manifold learning tries to preserve local neighborhoods; representation learning tries to learn coordinates useful for a loss function.

§1.4 Optimal transport and stochastic dynamics

For probability measures μ, ν on a metric space, the p-Wasserstein distance is

$$W_p(\mu,\nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int d(x,y)^p \, d\pi(x,y) \right)^{1/p}.$$

Here $\Pi(\mu,\nu)$ is the set of couplings with marginals μ and ν. The dual viewpoint is equally important. For $p = 1$,

$$W_1(\mu,\nu) = \sup_{\|f\|_{\mathrm{Lip}} \le 1} \left(\int f \, d\mu - \int f \, d\nu \right).$$

For $p = 2$ there is a useful one-dimensional closed form. Let

$$\mu = \mathcal{N}(a, \sigma_1^2), \qquad \nu = \mathcal{N}(b, \sigma_2^2), \qquad \sigma_1, \sigma_2 > 0.$$

In one dimension the optimal map between these two Gaussians is affine and monotone, so write

$$T(x) = \alpha x + \beta.$$

The condition $T_{\#}\mu = \nu$ forces

$$\alpha = \frac{\sigma_2}{\sigma_1}, \qquad \beta = b - \alpha a.$$

If $X = a + \sigma_1 Z$ with $Z \sim \mathcal{N}(0,1)$, then

$$T(X) = b + \sigma_2 Z.$$

Thus

$$\begin{aligned} W_2^2(\mu,\nu) &= \mathbb{E}\big[(X - T(X))^2\big] \\ &= \mathbb{E}\big[((a-b) + (\sigma_1 - \sigma_2)Z)^2\big] \\ &= (a-b)^2 + (\sigma_1 - \sigma_2)^2. \end{aligned}$$

This formula is small, but it explains a lot of geometric language around generative models: matching Gaussians means matching both their centers and their spreads.

Diffusion models can be read through stochastic differential equations

$$dX_t = b(t, X_t)\, dt + \sigma(t)\, dW_t.$$

The density p_t evolves according to a Fokker–Planck equation, and the reverse-time dynamics involve the score $\nabla_x \log p_t(x)$. Thus score-based generative modeling is an attempt to learn the vector field that reverses a noisy transport of probability mass.

For DDPMs the training objective is usually written as a variational lower bound, or in its simplified form as a denoising mean-square loss. It is not literally a Wasserstein objective. Still, each reverse step compares Gaussian conditional distributions, and for Gaussians the geometry above says that a squared displacement of means and a mismatch of variances are exactly the ingredients of W_2^2. The ELBO uses KL divergences between Gaussian conditionals; when the variance is fixed or separately prescribed, minimizing that KL has the same practical center-matching behavior as minimizing the corresponding Gaussian W_2 term. My preferred reading is therefore: optimal transport gives the geometric picture of local probability movement, while the DDPM ELBO gives the statistically convenient training objective.

§1.5 Symmetry, equivariance, and category-like thinking

If a group G acts on an input space and an output space, a map F is equivariant if

$$F(gx) = gF(x).$$

Convolutional neural networks exploit translation equivariance; graph neural networks exploit permutation equivariance; gauge-equivariant networks try to respect local choices of frame on a bundle.

Here is the matrix calculation behind the slogan. Let the graph be the cycle C_4, with adjacency matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix},$$

and let P be the permutation matrix rotating the four vertices by $90°$. Since P is a graph automorphism,

$$PA = AP.$$

The same is true for the usual normalized adjacency $\widehat{A}$, because normalization is built from A and the degree matrix, both invariant under the same rotation:

$$P\widehat{A} = \widehat{A}P.$$

A standard graph neural-network layer is

$$H^{(l+1)} = \sigma(\widehat{A}H^{(l)}W),$$

where σ acts entrywise. If the input features are rotated, then

$$\begin{aligned} F(PH^{(l)}) &= \sigma(\widehat{A}PH^{(l)}W) \\ &= \sigma(P\widehat{A}H^{(l)}W) \\ &= P\sigma(\widehat{A}H^{(l)}W) \\ &= PH^{(l+1)}. \end{aligned}$$

Equivariance is therefore not only a philosophical commitment to symmetry. In this example it is the algebraic fact that the permutation matrix commutes with the graph operator.

The original essay also speculates about topos-theoretic or categorical semantics for AI. A cautious reconstruction is this: category theory is useful when one needs to track structure-preserving maps, functorial changes of representation, and compositional semantics. It does not by itself explain intelligence. Its honest role is to provide a disciplined language for invariance, abstraction, and transport of structure.

§1.6 Further direction

The note is best read as a map of mathematical languages for data: topology sees holes, geometry sees coordinates and curvature, measure theory sees distributions, dynamics sees evolution, algebra sees symmetry, and category theory sees structure-preserving translation. The useful question is not which language is fashionable, but which structure is actually stable under the transformations used by the model.

§1.7 Backpropagation and differential geometry

Backpropagation is reverse-mode differentiation. Geometrically, forward computation pushes tangent vectors forward, while backpropagation pulls cotangent vectors backward through the computation graph.

§1.8 Manifold hypothesis and representation learning

The manifold hypothesis says high-dimensional data may concentrate near lower-dimensional geometric structures. Embeddings, charts, curvature, and local neighborhoods then become relevant to learning representations.

§1.9 Symmetry and equivariance

If a task has a symmetry group G, an equivariant model satisfies

$$f(gx) = gf(x).$$

This reduces sample complexity by building the correct transformation law into the model.

§1.10 Optimal transport and distributions

Optimal transport compares probability measures by the cost of moving mass. It connects geometry, PDE, and machine learning through distances between distributions.

§1.11 Mathematical structures for learning systems

Pure mathematics contributes language for structure: cotangent pullback for gradients, manifolds for data geometry, equivariance for symmetry, optimal transport for distributions, and category-like compositionality for architectures.

§1.12 Rebuilt reading path: what pure mathematics contributes

There are several layers.

(i) **Geometry** studies the shape of the representation space: distances, curvature, geodesics, embeddings, and coordinate charts.

(ii) **Topology** studies features stable under continuous deformation: connected components, loops, voids, and persistence across scales.

(iii) **Measure theory** studies data as distributions rather than isolated points.

(iv) **Dynamical systems** study learning or generation as evolution of states or densities.

(v) **Algebra and category theory** study invariance, symmetry, and composition of structure-preserving maps.

Pure mathematics does not magically explain AI. It supplies languages in which different kinds of stability can be stated precisely.

§1.13 Topological data analysis in usable form

Given a finite metric point cloud X, persistent homology constructs a family of complexes K_ε indexed by a scale parameter. For the Vietoris–Rips complex,

$$\{x_0, \dots, x_k\} \in VR_\varepsilon(X) \quad \Longleftrightarrow \quad d(x_i, x_j) \le \varepsilon \text{ for all } i, j.$$

The homology group is

$$H_k(K_\varepsilon) = \ker \partial_k / \operatorname{im} \partial_{k+1}.$$

This quotient is the mathematical form of the phrase "holes that are cycles but not boundaries."

§1.14 Wasserstein geometry and stochastic dynamics

For probability measures μ, ν on a metric space,

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int d(x,y)^p \, d\pi(x,y) \right)^{1/p}.$$

The coupling π is the hidden plan. In generative modeling, one often tries to construct a transport from a simple distribution to a complicated one. Normalizing flows do this by invertible maps, diffusion models by stochastic dynamics, and optimal transport by variational minimization.

A diffusion process has the schematic form

$$dX_t = b(t, X_t)\, dt + \sigma(t)\, dW_t.$$

The reverse process depends on the score field $\nabla_x \log p_t(x)$. Training a score network is therefore an attempt to approximate a time-dependent vector field on probability density.

§1.15 Category language without exaggeration

A category consists of objects and morphisms, with associative composition and identity morphisms. A functor preserves this structure. Category theory is useful when the same construction can be performed across many spaces in a way compatible with maps. For example, homology is functorial:

$$f : X \to Y \quad \Longrightarrow \quad H_k(f) : H_k(X) \to H_k(Y).$$

The honest categorical contribution is functoriality and compositionality, not a mystical explanation of intelligence.