Finite Categorical Projection Estimators for Noninvertible Coherence Constraints

Yu Murakami
New York General Group
March 12, 2026

Abstract

We introduce a finite-dimensional estimator for stochastic prediction under categorical coherence constraints. A task system is modeled by a finite category C, together with functors $X, Y : C \to \mathbf{FinSet}$ assigning finite input and output spaces to each task. A predictor is a family of stochastic kernels $K_c : X(c) \rightsquigarrow Y(c)$ and is coherent when $Y(u)_* \circ K_c = K_d \circ X(u)_*$ for every arrow $u : c \to d$. The coherent predictors form the end $\int_{c \in C} \mathbf{Stoch}_{\mathrm{fin}}(X(c), Y(c))$, equivalently a compact convex polytope cut out by linear equations inside a product of simplices. The contribution of the paper is a finite categorical projection estimator: empirical conditional label distributions are projected onto this coherence polytope under a weighted Euclidean norm, equivalently minimizing empirical risk for the quadratic proper scoring rule. We prove existence, uniqueness, categorical coherence by construction, deterministic nonexpansiveness in the weighted Euclidean norm, finite-sample concentration, corrected equivalence invariance with normalized object-multiplicity weights, and a two-section certificate for non-singleton feasible sets. We also explain the relationship with invariant estimation: for group actions the estimator reduces to Reynolds-style averaging, while for genuinely noninvertible categories averaging is unavailable and projection is essential.

1 Introduction

Prediction systems often contain multiple related tasks. A fine classifier may refine a coarse classifier. A prediction made on one representation may be transported to another representation. Some labels may be merged by an abstraction map, and some inputs may be coarsened by a feature map. When these transformations are part of the problem specification, it is natural to ask for predictors that respect them.

We model such situations by a finite category C. Objects are tasks or contexts, and arrows are structural transformations. Two functors $X, Y : C \to \mathbf{FinSet}$ assign finite input and output spaces. A stochastic predictor is a family of kernels $K_c : X(c) \rightsquigarrow Y(c)$. The coherence condition is

$$Y(u)_* \circ K_c = K_d \circ X(u)_*$$

for each arrow $u : c \to d$. It says that transporting an output distribution after prediction agrees with predicting after transporting the input.

This condition is strong. It is appropriate only when exact commutation expresses a genuine modeling constraint. A coherent predictor can still be inaccurate, miscalibrated, unfair, or unsafe. Coherence is one structural property, not a substitute for external evaluation.

The mathematical observation that coherent families are described by an end is standard category theory [1,2]. The contribution here is the finite estimation theory built on this observation. In particular, the paper proves that coherent predictors form an explicitly auditable compact convex polytope; projection onto this polytope is empirical risk minimization for the quadratic, or Brier, scoring rule; the projection estimator exists uniquely under nonemptiness; the estimator is coherent by construction; the projection map is 1-Lipschitz in the chosen weighted Euclidean norm; empirical concentration follows from standard multinomial concentration and nonexpansiveness; equivalence of task categories preserves the estimator only after explicit normalization of transported weights; and for group actions, the construction reduces to classical invariant averaging, while noninvertible categorical arrows require genuine convex projection.

The word auditable is used in the concrete sense that the feasible set is given by finitely many displayed linear equations and simplex inequalities; hence a proposed solution can be checked by finite arithmetic.

2 Related Work and Positioning

The equalizer formula for ends is classical [1,2]. Categorical probability and stochastic computation are treated abstractly in the Kleisli and Markov-category traditions [7–10]. The present work is not a replacement for Markov-category probability; it is orthogonal to it. Markov categories axiomatize stochastic maps and conditional independence, whereas this paper studies finite estimation under externally specified categorical coherence constraints.

Categorical approaches to learning include functorial backpropagation [4], categorical foundations of gradient-based learning [5], surveys of category theory in machine learning [6], and categorical deep learning as an architecture-level theory [18]. Recent work on categorical equivariant deep learning formulates equivariance as naturality and proves universal approximation results for category-equivariant neural networks [20]. That contribution is broader in approximation theory and different in focus from the finite convex estimator studied here.

The closest statistical neighbor is invariant estimation. Classical group-invariant estimation studies estimators constrained by symmetry and often uses averaging or conditional

expectation [11,12]. In equivariant neural networks, group or group-like symmetry constraints become linear constraints on layers or maps [13–17]. In the special case where C is a group, viewed as a one-object category, the estimator below reduces to orthogonal projection onto a fixed-point set. For finite groups with uniform weighting, this projection is a Reynolds average.

The genuinely categorical contribution occurs when C is not a groupoid. Then arrows need not be invertible, so coherence is not simply equivariance under reversible symmetries. It is a one-sided compatibility condition involving pushforwards along noninvertible maps. In that setting, there is generally no averaging operator over arrows, and convex projection is the natural finite estimator.

3 Finite Stochastic Kernels

Let **FinSet** be the category of finite sets and functions. For a finite set A, define

$$\Delta A = \{p : A \to [0,1] \mid \sum_{a \in A} p(a) = 1\}.$$

A finite stochastic kernel $K : A \rightsquigarrow B$ is a function $K : A \to \Delta B$. We write $K(b \mid a)$ for the probability of $b \in B$ given $a \in A$.

The category $\mathbf{Stoch}_{\text{fin}}$ has finite sets as objects and stochastic kernels as morphisms. Composition is matrix multiplication:

$$(L \circ K)(c \mid a) = \sum_{b \in B} L(c \mid b) K(b \mid a).$$

The identity on A is the deterministic kernel $\delta_A(a' \mid a) = 1$ if $a' = a$ and 0 otherwise. Equivalently, $\mathbf{Stoch}_{\text{fin}}$ is the Kleisli category of the finite-distribution monad on **FinSet** [7,8].

Every function $f : A \to B$ induces a deterministic kernel

$$f_* : A \rightsquigarrow B, \qquad f_*(b \mid a) = \begin{cases} 1, & b = f(a), \\ 0, & b \neq f(a). \end{cases}$$

Proposition 1. *The assignment $A \mapsto A$, $f \mapsto f_*$ defines a faithful functor $(-)_* : \mathbf{FinSet} \to \mathbf{Stoch}_{\text{fin}}$.*

Proof. Let $f : A \to B$ and $g : B \to C$. For $a \in A$ and $c \in C$,

$$(g_* \circ f_*)(c \mid a) = \sum_{b \in B} g_*(c \mid b) f_*(b \mid a).$$

Only $b = f(a)$ contributes, so this equals $g_*(c \mid f(a))$, which is 1 exactly when $c = g(f(a))$. Hence $g_* \circ f_* = (g \circ f)_*$, and identities are preserved. If $f_* = g_*$, then the Dirac measures at $f(a)$ and $g(a)$ are equal for every a, so $f = g$. Thus the functor is faithful. □

4 Task Categories and Coherent Predictors

Let C be a finite category and let $X, Y : C \to \mathbf{FinSet}$ be functors. For each object c, $X(c)$ is a finite input set and $Y(c)$ is a finite output set. For each arrow $u : c \to d$, the maps $X(u) : X(c) \to X(d)$ and $Y(u) : Y(c) \to Y(d)$ transport inputs and outputs.

Definition 1. *A local stochastic predictor from X to Y is a family $K = (K_c)_{c \in C}$ where $K_c : X(c) \rightsquigarrow Y(c)$ is a stochastic kernel.*

Definition 2. *A local predictor K is categorically coherent if, for every arrow $u : c \to d$,*

$$Y(u)_* \circ K_c = K_d \circ X(u)_*.$$

Equivalently, for every $x \in X(c)$ and $y' \in Y(d)$,

$$\sum_{\substack{y \in Y(c) \\ Y(u)(y) = y'}} K_c(y \mid x) = K_d(y' \mid X(u)(x)).$$

We write $\mathrm{Coh}_C(X, Y)$ for the set of coherent predictors.

5 The End and the Coherence Polytope

For objects $c, d \in C$, define $H(c,d) = \mathbf{Stoch}_{\text{fin}}(X(c), Y(d))$. Precomposition by $X(u)_*$ and postcomposition by $Y(u)_*$ make $H : C^{\text{op}} \times C \to \mathbf{Set}$ a bifunctor.

The end

$$\int_{c \in C} H(c,c) = \int_{c \in C} \mathbf{Stoch}_{\text{fin}}(X(c), Y(c))$$

is the set of wedges satisfying the standard dinaturality equations. In this diagonal case, those equations are exactly the coherence equations above.

Theorem 1 (End description)**.** *There is a canonical equality*

$$\mathrm{Coh}_C(X, Y) = \int_{c \in C} \mathbf{Stoch}_{\text{fin}}(X(c), Y(c)).$$

Equivalently,

$$\mathrm{Coh}_C(X, Y) = \mathrm{Eq}\Big(\prod_{c \in C} \mathbf{Stoch}_{\text{fin}}(X(c), Y(c)) \rightrightarrows \prod_{u : c \to d} \mathbf{Stoch}_{\text{fin}}(X(c), Y(d)) \Big),$$

where the two maps send K to $(Y(u)_* \circ K_c)_u$ and $(K_d \circ X(u)_*)_u$.

Proof. The equalizer formula for ends gives

$$\int_c H(c,c) = \mathrm{Eq}\left(\prod_c H(c,c) \rightrightarrows \prod_{u : c \to d} H(c,d) \right).$$

For $H(c,d) = \mathbf{Stoch}_{\text{fin}}(X(c), Y(d))$, the two maps associated to $u : c \to d$ are postcomposition by $Y(u)_*$ and precomposition by $X(u)_*$. Hence an element of the equalizer is exactly a family $K_c : X(c) \rightsquigarrow Y(c)$ satisfying $Y(u)_* \circ K_c = K_d \circ X(u)_*$ for every arrow u. This is precisely $\mathrm{Coh}_C(X, Y)$. □

Theorem 2 (Coherence polytope)**.** *The set* $\mathrm{Coh}_C(X,Y)$ *is a compact convex polytope.*

Proof. For each c,

$$\mathbf{Stoch}_{\mathrm{fin}}(X(c),Y(c)) = \prod_{x\in X(c)} \Delta Y(c).$$

Each simplex is a compact convex polytope, and a finite product of compact convex polytopes is a compact convex polytope. Hence $P = \prod_{c\in C} \mathbf{Stoch}_{\mathrm{fin}}(X(c),Y(c))$ is a compact convex polytope. For every arrow $u : c \to d$, the equation $Y(u)_* \circ K_c = K_d \circ X(u)_*$ is a finite system of linear equations in the coordinates $K_c(y \mid x)$. Therefore $\mathrm{Coh}_C(X,Y)$ is the intersection of P with finitely many affine linear subspaces, hence is a compact convex polytope. □

6 Checking Coherence on Generators

Suppose C is presented by a finite graph G and relations. The relations are already imposed in C, so functors out of C automatically respect them.

Proposition 2. *A local predictor* K *is coherent if and only if* $Y(g)_* \circ K_c = K_d \circ X(g)_*$ *for every generating arrow* $g : c \to d$ *of* G*.*

Proof. One direction is immediate. Conversely, assume the equation holds for all generating arrows. For a composite $u = g_n \circ \cdots \circ g_1 : c \to d$, successive substitution gives

$$Y(u)_* \circ K_c = K_d \circ X(g_n)_* \circ \cdots \circ X(g_1)_* = K_d \circ X(u)_*.$$

If two paths represent the same arrow of C, the defining relations of C and functoriality of X, Y imply that the transported maps are equal. Thus coherence holds for every arrow. □

7 Nonemptiness and a Two-Point Certificate

Assume henceforth that $Y(c) \neq \varnothing$ for every c. Otherwise no stochastic kernel into $Y(c)$ exists for nonempty $X(c)$.

Proposition 3 (Nonemptiness from a global output section)**.** *Suppose there exists a natural transformation* $s : 1_C \Rightarrow Y$*, where* $1_C : C \to$ **FinSet** *is the constant singleton functor. Then* $\mathrm{Coh}_C(X,Y) \neq \varnothing$*.*

Proof. For each object c, let $s_c(*) \in Y(c)$ be the selected output. Define $K_c(y \mid x) = 1$ if $y = s_c(*)$ and 0 otherwise. Naturality of s gives $Y(u)(s_c(*)) = s_d(*)$ for every arrow $u : c \to d$. Hence both $Y(u)_* \circ K_c$ and $K_d \circ X(u)_*$ are the constant Dirac kernel at $s_d(*)$. Thus K is coherent. □

Proposition 4 (Two-point certificate)**.** *Suppose there exist two natural transformations* $s, t : 1_C \Rightarrow Y$ *such that* $s_c(*) \neq t_c(*)$ *for at least one object* c*. Then* $\mathrm{Coh}_C(X,Y)$ *contains at least two distinct extreme points.*

Proof. By the previous proposition, s and t define coherent deterministic predictors K^s and K^t. If $s_c(*) \neq t_c(*)$ for some c, then for any $x \in X(c)$, the corresponding rows of K^s_c and K^t_c are different Dirac distributions. Thus $K^s \neq K^t$.

Each row of K^s is a vertex of a simplex, so K^s is an extreme point of the ambient product polytope. If $K^s = \lambda A + (1-\lambda)B$ with $0 < \lambda < 1$ and $A, B \in \mathrm{Coh}_C(X,Y)$, then the same convex decomposition occurs in the ambient product polytope. Extremality there forces $A = B = K^s$. Thus K^s is extreme in $\mathrm{Coh}_C(X,Y)$, and similarly for K^t. □

This is deliberately a two-point certificate, not a dimension theorem. It proves that the feasible set is not a singleton, but it does not give an identifiability result or a lower bound on dimension.

8 Quadratic Loss and the Projection Estimator

Let empirical conditional label distributions be given by $q_c(\cdot \mid x) \in \Delta Y(c)$ for every $c \in C$ and $x \in X(c)$. Let positive weights $\mu_c(x) > 0$ be given. They may represent sample frequencies, design weights, or task priorities.

Define

$$\mathcal{H}_C = \prod_{c\in C} \prod_{x\in X(c)} \mathbb{R}^{Y(c)}.$$

For $A, B \in \mathcal{H}_C$, define

$$\langle A, B\rangle_\mu = \sum_{c\in C} \sum_{x\in X(c)} \mu_c(x) \sum_{y\in Y(c)} A_c(y \mid x) B_c(y \mid x),$$

with norm $\|A\|_\mu = \sqrt{\langle A, A\rangle_\mu}$.

Definition 3 (Categorical projection estimator)**.** *The categorical projection estimator is*

$$\widehat{K} = \operatorname*{argmin}_{K\in \mathrm{Coh}_C(X,Y)} \frac{1}{2}\|K - q\|^2_\mu .$$

Proposition 5 (Quadratic proper-scoring interpretation)**.** *For a predicted distribution* $p \in \Delta A$ *and realized label* $a \in A$*, define the quadratic score loss*

$$\ell(p,a) = \frac{1}{2}\sum_{b\in A} (p(b) - \mathbf{1}_{b=a})^2.$$

Then minimizing

$$\sum_{c,x} \mu_c(x) \sum_{y\in Y(c)} q_c(y \mid x)\ell(K_c(\cdot \mid x), y)$$

over coherent K *is equivalent to the projection estimator.*

Proof. Fix $p, q \in \Delta A$. Then

$$\sum_{a\in A} q(a)\ell(p,a) = \frac{1}{2}\sum_b p(b)^2 - \sum_b q(b)p(b) + \frac{1}{2}.$$

Also,

$$\frac{1}{2}\|p - q\|_2^2 = \frac{1}{2}\sum_b p(b)^2 - \sum_b q(b)p(b) + \frac{1}{2}\sum_b q(b)^2.$$

The two expressions differ by $\frac{1}{2}-\frac{1}{2}\sum_b q(b)^2$, which depends only on q, not on p. After summing over c, x with weights $\mu_c(x)$, the total difference remains independent of K. Hence the minimizers are identical. □

9 Existence, Uniqueness, and Coherence

Theorem 3 (Existence and uniqueness)**.** *Assume* $\mathrm{Coh}_C(X,Y) \neq \varnothing$. *Then the categorical projection estimator exists and is unique.*

Proof. By Theorem 5.2, $\mathrm{Coh}_C(X,Y)$ is compact and convex. The objective $K \mapsto \frac{1}{2}\|K-q\|_\mu^2$ is continuous, so a minimizer exists. Because all weights are positive, $\langle -,-\rangle_\mu$ is an inner product. Hence the squared norm is strictly convex. If $K \neq L$ and $0 < t < 1$, then

$$\|tK+(1-t)L-q\|_\mu^2 = t\|K-q\|_\mu^2 + (1-t)\|L-q\|_\mu^2 - t(1-t)\|K-L\|_\mu^2,$$

which is strictly less than the corresponding convex combination. Thus the objective has at most one minimizer on a convex set. Since a minimizer exists, it is unique. □

Corollary 1 (Coherence by construction)**.** *The estimator* $\widehat{K}$ *is coherent.*

Proof. It is minimized over $\mathrm{Coh}_C(X,Y)$, so it belongs to $\mathrm{Coh}_C(X,Y)$. □

Corollary 2 (Exact recovery)**.** *If* $q \in \mathrm{Coh}_C(X,Y)$, *then* $\widehat{K} = q$.

Proof. If q is coherent, it is feasible. The objective at q is 0, the smallest possible squared distance. By uniqueness, $\widehat{K} = q$. □

10 Stability

The stability result below is deterministic and metric-specific. It is not a calibration guarantee or a generalization theorem.

Lemma 1 (Projection variational inequality)**.** *Let* P *be a nonempty closed convex subset of a finite-dimensional Hilbert space* $\mathcal{H}$. *Let* $p = \Pi_P(q)$ *be the metric projection of* q *onto* P. *Then* $\langle q-p, r-p\rangle \le 0$ *for every* $r \in P$. *Conversely, if* $p \in P$ *satisfies this inequality for every* $r \in P$, *then* $p = \Pi_P(q)$.

Proof. Assume $p = \Pi_P(q)$. For $r \in P$ and $t \in [0,1]$, $p_t = p + t(r-p) \in P$. The function $\phi(t) = \frac{1}{2}\|p_t - q\|^2$ has a minimum at $t = 0$. Hence $0 \le \phi'(0) = \langle p-q, r-p\rangle$, so $\langle q-p, r-p\rangle \le 0$.

Conversely, suppose the displayed inequality holds for all $r \in P$. Then

$$\|r-q\|^2 = \|r-p\|^2 + 2\langle r-p, p-q\rangle + \|p-q\|^2.$$

Since $\langle r-p, p-q\rangle = -\langle q-p, r-p\rangle \ge 0$, we obtain $\|r-q\|^2 \ge \|p-q\|^2$. Thus p minimizes distance from q over P. □

Theorem 4 (Nonexpansiveness)**.** *Let* $P = \mathrm{Coh}_C(X,Y)$. *For all empirical targets* $q, q' \in \mathcal{H}_C$,

$$\|\Pi_P(q) - \Pi_P(q')\|_\mu \le \|q-q'\|_\mu.$$

Proof. Let $p = \Pi_P(q)$ and $p' = \Pi_P(q')$. By the variational inequality,

$$\langle q-p, p-p'\rangle_\mu \ge 0, \qquad \langle q'-p', p-p'\rangle_\mu \le 0.$$

Combining gives

$$\langle q-q'-(p-p'), p-p'\rangle_\mu \ge 0.$$

Therefore

$$\|p-p'\|_\mu^2 \le \langle q-q', p-p'\rangle_\mu \le \|q-q'\|_\mu \|p-p'\|_\mu.$$

If $\|p-p'\|_\mu = 0$, the result is immediate; otherwise divide by it. □

Remark 1 (Dependence on Euclidean geometry)**.** *Theorem 10.2 is a Hilbert-space projection result. For total variation, KL, Hellinger, or other geometries, the same statement requires different arguments and may have a different Lipschitz constant or fail altogether.*

11 Statistical Concentration

Assume there is an unknown conditional distribution $p_c^\star(\cdot \mid x) \in \Delta Y(c)$ for each c, x. For each pair (c,x), observe $n_{c,x}$ independent labels sampled from $p_c^\star(\cdot \mid x)$, producing an empirical distribution $\widehat{q}_c(\cdot \mid x)$.

Let $\widehat{K} = \Pi_P(\widehat{q})$ and $K^\star = \Pi_P(p^\star)$, where $P = \mathrm{Coh}_C(X,Y)$.

Theorem 5 (Coordinate Hoeffding bound)**.** *Let* $m = \sum_{c,x}|Y(c)|$. *Then for every* $\delta \in (0,1)$, *with probability at least* $1-\delta$,

$$\left\|\widehat{K} - K^\star\right\|_\mu \le \left(\log\frac{2m}{\delta}\sum_{c,x}\mu_c(x)\frac{|Y(c)|}{2n_{c,x}}\right)^{1/2}.$$

If $n_{\min} = \min_{c,x} n_{c,x}$, *then*

$$\left\|\widehat{K} - K^\star\right\|_\mu \le \left(\frac{\log(2m/\delta)}{2n_{\min}}\sum_{c,x}\mu_c(x)|Y(c)|\right)^{1/2}.$$

Proof. For fixed c, x, y, $\widehat{q}_c(y \mid x)$ is an average of $n_{c,x}$ Bernoulli variables with mean $p_c^\star(y \mid x)$. Hoeffding's inequality gives

$$\Pr\left(|\widehat{q}_c(y \mid x) - p_c^\star(y \mid x)| \ge \varepsilon_{c,x}\right) \le 2e^{-2n_{c,x}\varepsilon_{c,x}^2}.$$

Choose $\varepsilon_{c,x} = \sqrt{(2n_{c,x})^{-1}\log(2m/\delta)}$. A union bound over all m coordinates gives, with probability at least $1-\delta$,

$$|\widehat{q}_c(y \mid x) - p_c^\star(y \mid x)| \le \varepsilon_{c,x}$$

for all c, x, y. Hence

$$\|\widehat{q} - p^\star\|_\mu^2 \le \log \frac{2m}{\delta} \sum_{c,x} \mu_c(x) \frac{|Y(c)|}{2n_{c,x}}.$$

By nonexpansiveness, $\left\|\widehat{K} - K^\star\right\|_\mu \le \|\widehat{q} - p^\star\|_\mu$. The second bound follows from $n_{c,x} \ge n_{\min}$. □

Remark 2 (Sharper multinomial bounds)**.** *The coordinate-union proof is simple and auditable but loose. One can replace it by sharper simplex concentration inequalities, such as the Weissman–Ordentlich–Seroussi–Verdu–Weinberger bound for L_1 deviation of empirical distributions [21].*

Corollary 3 (Consistency for the projected target)**.** *If $n_{\min} \to \infty$, then $\left\|\widehat{K} - K^\star\right\|_\mu \to 0$ in probability.*

Proof. For any $\varepsilon > 0$, Theorem 11.1 makes $\Pr(\left\|\widehat{K} - K^\star\right\|_\mu > \varepsilon)$ arbitrarily small as $n_{\min} \to \infty$. □

If $p^\star \notin P$, the limit $K^\star = \Pi_P(p^\star)$ is the best coherent approximation in the quadratic geometry. The difference $p^\star - \Pi_P(p^\star)$ is the structural bias introduced by imposing exact coherence.

12 Quadratic Program and KKT Conditions

Let $z_{cxy} = K_c(y \mid x)$. The estimator solves the convex quadratic program

$$\min_z \frac{1}{2} \sum_{c,x,y} \mu_c(x)(z_{cxy} - q_c(y \mid x))^2$$

subject to $z_{cxy} \ge 0$,

$$\sum_{y \in Y(c)} z_{cxy} = 1,$$

and, for every arrow $u : c \to d$, every $x \in X(c)$, and every $y' \in Y(d)$,

$$\sum_{\substack{y \in Y(c) \\ Y(u)(y) = y'}} z_{cxy} = z_{d,X(u)(x),y'}.$$

If C is given by generators and relations, it suffices by Proposition 6.1 to impose the coherence equations for generating arrows.

Introduce Lagrange multipliers $\alpha_{c,x} \in \mathbb{R}$ for row-sum constraints, $\lambda_{u,x,y'} \in \mathbb{R}$ for coherence constraints, and $\rho_{c,x,y} \ge 0$ for nonnegativity. At the optimum, the KKT conditions are: primal feasibility; dual feasibility $\rho_{c,x,y} \ge 0$; complementary slackness $\rho_{c,x,y} z_{cxy} = 0$; and stationarity of the Lagrangian obtained by adding the row-sum and coherence equalities and subtracting $\sum_{c,x,y} \rho_{c,x,y} z_{cxy}$ from the quadratic objective. Since the objective is strictly convex and the feasible set is convex, these KKT conditions are necessary and sufficient.

The number of variables is $N = \sum_{c \in C} |X(c)|\,|Y(c)|$. If G is a generating graph for C, the number of coherence equalities imposed from generators is

$$M_{\text{coh}} = \sum_{g : c \to d \in G} |X(c)|\,|Y(d)|.$$

There are additionally $\sum_c |X(c)|$ row-sum equalities and N nonnegativity inequalities. Thus the problem scales linearly in the number of generating arrows and finite fiber sizes, although generic QP solvers may have superlinear runtime in N.

13 Equivalence Invariance with Normalized Transported Weights

Let $F : C \to D$ be an equivalence of finite categories. Choose an adjoint equivalence $F : C \rightleftarrows D : G$ with unit and counit natural isomorphisms $\eta : \mathrm{id}_C \Rightarrow GF$ and $\varepsilon : FG \Rightarrow \mathrm{id}_D$ satisfying the triangle identities

$$\varepsilon_{F(c)} \circ F(\eta_c) = \mathrm{id}_{F(c)}, \qquad G(\varepsilon_d) \circ \eta_{G(d)} = \mathrm{id}_{G(d)}.$$

Every equivalence can be promoted to an adjoint equivalence [1,2]. Let $X_D, Y_D : D \to$ **FinSet**, and define $X_C = X_D \circ F$ and $Y_C = Y_D \circ F$.

Definition 4 (Normalized transport of data)**.** *Let q^C, μ^C be C-side empirical data. For $d \in D$, define*

$$m_d = |\{e \in \mathrm{Ob}(D) : G(e) = G(d)\}|.$$

Define D-side weights by

$$\nu_d(x) = \frac{1}{m_d} \mu^C_{G(d)}\big(X_D(\varepsilon_d)^{-1}(x)\big),$$

and define D-side targets by

$$q^D_d(\cdot \mid x) = Y_D(\varepsilon_d)_* q^C_{G(d)}\big(\cdot \mid X_D(\varepsilon_d)^{-1}(x)\big).$$

The factor $1/m_d$ is essential: without it, the D-side objective may count equivalent copies of the same C-object multiple times.

Theorem 6 (Equivalence invariance of coherent predictors)**.** *Restriction along F induces a bijection*

$$F^* : \mathrm{Coh}_D(X_D, Y_D) \to \mathrm{Coh}_C(X_C, Y_C).$$

Proof. For $K \in \mathrm{Coh}_D(X_D, Y_D)$, define $(F^*K)_c = K_{F(c)}$. This is coherent because F preserves arrows.

For the inverse, let $M \in \mathrm{Coh}_C(X_C, Y_C)$ and define

$$K_d = Y_D(\varepsilon_d)_* \circ M_{G(d)} \circ X_D(\varepsilon_d)^{-1}_*.$$

This is well typed because $X_C(Gd) = X_D(FGd)$ and $Y_C(Gd) = Y_D(FGd)$.

For $v : d \to e$, naturality of ε gives $v \circ \varepsilon_d = \varepsilon_e \circ FG(v)$. Using coherence of M for $G(v)$,

$$Y_D(v)_* \circ K_d = Y_D(\varepsilon_e)_* \circ Y_D(FG(v))_* \circ M_{Gd} \circ X_D(\varepsilon_d)_*^{-1}$$
$$= Y_D(\varepsilon_e)_* \circ M_{Ge} \circ X_D(FG(v))_* \circ X_D(\varepsilon_d)_*^{-1}.$$

Naturality of ε also implies $FG(v) \circ \varepsilon_d^{-1} = \varepsilon_e^{-1} \circ v$, so the last line is $K_e \circ X_D(v)_*$. Thus K is coherent.

The triangle identity $\varepsilon_{F(c)} \circ F(\eta_c) = \mathrm{id}_{F(c)}$ implies $F(\eta_c) = \varepsilon_{F(c)}^{-1}$. Using coherence of M for η_c, one obtains $K_{F(c)} = M_c$. Conversely, starting with coherent K on D, coherence with respect to $\varepsilon_d : FGd \to d$ gives $Y_D(\varepsilon_d)_* \circ K_{FGd} = K_d \circ X_D(\varepsilon_d)_*$, so the inverse construction returns K_d. Therefore F^* is a bijection. □

Theorem 7 (Equivalence invariance of the normalized projection estimator)**.** *Let q^D, ν be obtained from q^C, μ^C by normalized transport. Let*

$$\widehat{K}_C = \Pi_{\mathrm{Coh}_C(X_C,Y_C)}(q^C), \qquad \widehat{K}_D = \Pi_{\mathrm{Coh}_D(X_D,Y_D)}(q^D),$$

where the C-side norm is defined by μ^C and the D-side norm by ν. Then $F^*\widehat{K}_D = \widehat{K}_C$.

Proof. Let $K \in \mathrm{Coh}_D(X_D, Y_D)$ and set $M = F^*K$. By the previous theorem, every coherent K on D is determined by M via

$$K_d = Y_D(\varepsilon_d)_* \circ M_{G(d)} \circ X_D(\varepsilon_d)_*^{-1}.$$

By normalized transport, q_d^D is transported from $q_{G(d)}^C$ by the same bijections. Deterministic pushforward along a bijection only permutes coordinates of probability vectors, so squared Euclidean distances are preserved after identifying $x' = X_D(\varepsilon_d)^{-1}(x)$.

Using the normalized weight $\nu_d(x) = m_d^{-1}\mu_{G(d)}^C(x')$, we can write

$$\left\|K - q^D\right\|_{\nu,D}^2 = \sum_{d \in D} \frac{1}{m_d} D_{Gd}(M, q^C),$$

where

$$D_c(M, q^C) = \sum_{x \in X_C(c)} \mu_c^C(x) \sum_{y \in Y_C(c)} (M_c(y \mid x) - q_c^C(y \mid x))^2.$$

Group the sum by $c = Gd$. For each fixed c, exactly m_d objects d have $Gd = c$, and each contributes the same term with factor $1/m_d$. Therefore the total contribution is exactly one copy of the c-term. Hence

$$\left\|K - q^D\right\|_{\nu,D}^2 = \left\|F^*K - q^C\right\|_{\mu,C}^2.$$

Thus restriction along F identifies the D-side objective on coherent predictors with the C-side objective. Since F^* is a bijection of feasible sets, minimizers correspond. □

Without normalization, if D contains two equivalent copies of a C-object and both are given the full C-side weight, the D-side objective counts that object twice. If different C-objects are duplicated with different multiplicities, the restricted minimizer can be a different weighted projection. Normalization prevents this object-copy bias.

14 The Group Case and Reynolds Averaging

Let G be a finite group viewed as a one-object category. Let X, Y be finite G-sets. A kernel $K : X \rightsquigarrow Y$ is coherent exactly when

$$Y(g)_* \circ K = K \circ X(g)_*$$

for all $g \in G$. This is ordinary equivariance.

The group acts linearly on the ambient vector space of matrices by

$$(g \cdot K) = Y(g)_* \circ K \circ X(g^{-1})_*.$$

The coherent predictors are exactly the fixed points of this action, intersected with the stochastic-row constraints. If the weighted Euclidean norm is G-invariant, the orthogonal projection onto the fixed-point subspace is the Reynolds operator

$$\mathcal{R}(K) = \frac{1}{|G|} \sum_{g \in G} g \cdot K.$$

Because the action permutes rows and columns, $\mathcal{R}$ preserves row-stochasticity. Hence in the group case the categorical projection estimator reduces to classical invariant averaging.

For a non-groupoid category, arrows need not have inverses. Then $X(u^{-1})$ is meaningless, and the Reynolds average has no analogue. The coherence equations remain linear, but projection onto their intersection with the product of simplices is the relevant replacement.

15 Worked Numerical Example

Let C have two objects f, c and one nonidentity arrow $a : f \to c$. Let

$$X(f) = \{1, 2, 3\}, \quad X(c) = \{A, B\},$$

with $X(a)(1) = A$, $X(a)(2) = A$, and $X(a)(3) = B$. Let

$$Y(f) = \{r, s, t\}, \quad Y(c) = \{U, V\},$$

with $Y(a)(r) = U$, $Y(a)(s) = U$, and $Y(a)(t) = V$.

Use unit weights. Suppose

$$q_f(\cdot \mid 1) = (0.6, 0.3, 0.1), \quad q_f(\cdot \mid 2) = (0.2, 0.1, 0.7),$$
$$q_f(\cdot \mid 3) = (0.1, 0.2, 0.7),$$
$$q_c(\cdot \mid A) = (0.4, 0.6), \quad q_c(\cdot \mid B) = (0.8, 0.2).$$

Coherence requires

$$K_c(U \mid A) = K_f(r \mid 1) + K_f(s \mid 1)$$
$$= K_f(r \mid 2) + K_f(s \mid 2).$$

The empirical data violate this because $0.6 + 0.3 = 0.9$, $0.2 + 0.1 = 0.3$, and $q_c(U \mid A) = 0.4$.

Let $\theta_A = K_c(U \mid A)$. For fixed θ_A, the closest fine distribution to $(0.6, 0.3, 0.1)$ subject to $r + s = \theta_A$ is

$$\left(0.6 + \frac{\theta_A - 0.9}{2},\ 0.3 + \frac{\theta_A - 0.9}{2},\ 1 - \theta_A\right),$$

provided the entries are nonnegative. The closest fine distribution to $(0.2, 0.1, 0.7)$ subject to $r + s = \theta_A$ is

$$\left(0.2 + \frac{\theta_A - 0.3}{2},\ 0.1 + \frac{\theta_A - 0.3}{2},\ 1 - \theta_A\right).$$

The resulting objective is

$$\frac{3}{2}(\theta_A - 0.9)^2 + \frac{3}{2}(\theta_A - 0.3)^2 + 2(\theta_A - 0.4)^2.$$

Differentiating gives $3(\theta_A - 0.9) + 3(\theta_A - 0.3) + 4(\theta_A - 0.4) = 0$, so $\theta_A = 0.52$. Thus

$$\widehat{K}_c(\cdot \mid A) = (0.52, 0.48),$$

$$\widehat{K}_f(\cdot \mid 1) = (0.41, 0.11, 0.48),$$
$$\widehat{K}_f(\cdot \mid 2) = (0.31, 0.21, 0.48).$$

The nonnegativity caveat is satisfied here. In examples where the displayed affine formulas leave the simplex, the constrained projection lands on the boundary and the simple closed form must be replaced by the full quadratic program.

For B, let $\theta_B = K_c(U \mid B)$. The fine empirical U-mass is $0.1 + 0.2 = 0.3$, while the coarse empirical U-mass is 0.8. The objective is $\frac{3}{2}(\theta_B - 0.3)^2 + 2(\theta_B - 0.8)^2$. Differentiating gives $3(\theta_B - 0.3) + 4(\theta_B - 0.8) = 0$, so

$$\theta_B = \frac{4.1}{7} \approx 0.585714.$$

Therefore

$$\widehat{K}_c(\cdot \mid B) \approx (0.585714, 0.414286),$$

and

$$\widehat{K}_f(\cdot \mid 3) \approx (0.242857, 0.342857, 0.414286).$$

The example shows the estimator performing a nontrivial compromise among incompatible fine and coarse empirical predictors.

16 Soft Coherence

Hard coherence may be too restrictive. A soft version minimizes

$$\frac{1}{2}\left\|K - q\right\|_\mu^2 + \frac{\lambda}{2} \sum_{u:c \to d} \omega_u \left\|Y(u)_* \circ K_c - K_d \circ X(u)_*\right\|^2$$

over all local stochastic predictors K, where $\lambda > 0$. As $\lambda \to \infty$, minimizers converge, under standard compactness arguments, to minimizers of the hard constrained problem whenever the hard feasible set is nonempty. Thus hard categorical projection can be understood as the infinite-penalty limit of soft coherence regularization.

17 Modeling Scope

Categorical coherence is a structural prior. It is most appropriate when the transformations in C are part of the semantic specification of the prediction task. For example, if a coarse label is definitionally the image of a fine label, then a coherent predictor ensures that fine and coarse predictions cannot contradict each other under that definition.

Exact coherence can be inappropriate when different resolutions are generated by different mechanisms, when measurement noise is transformation-dependent, or when coarse labels have information not recoverable by aggregating fine labels. In such cases, the soft estimator of the previous section or a different probabilistic model may be preferable.

18 Conclusion

This paper defined and analyzed a finite categorical projection estimator for stochastic prediction under noninvertible coherence constraints. Its novelty lies not in the end formula itself, but in combining finite categorical coherence with convex projection, quadratic proper scoring, stability, concentration, normalized equivalence invariance, and explicit computation.

For group actions, the estimator specializes to invariant averaging. For noninvertible finite categories, it yields a projection method where averaging is unavailable. This is the central distinction: categorical coherence includes equivariance as a special case but also covers one-sided abstraction and refinement constraints.

References

[1] S. Mac Lane. *Categories for the Working Mathematician*. 2nd ed. Springer, 1998.

[2] E. Riehl. *Category Theory in Context*. Dover, 2016.

[3] S. Awodey. *Category Theory*. 2nd ed. Oxford University Press, 2010.

[4] B. Fong, D. I. Spivak, and R. Tuyeras. Backprop as Functor: A Compositional Perspective on Supervised Learning. *Proceedings of LICS*, 2019.

[5] G. S. H. Cruttwell, B. Gavranovic, N. Ghani, P. Wilson, and F. Zanasi. Categorical Foundations of Gradient-Based Learning. *Programming Languages and Systems*. Springer, 2022.

[6] D. Shiebler. Category Theory in Machine Learning. arXiv:2106.07032, 2021.

[7] M. Giry. A Categorical Approach to Probability Theory. In *Categorical Aspects of Topology and Analysis*, LNCS 915. Springer, 1982.

[8] T. Fritz. A Synthetic Approach to Markov Kernels, Conditional Independence and Theorems on Sufficient Statistics. *Advances in Mathematics*, 370, 2020.

[9] K. Cho and B. Jacobs. Disintegration and Bayesian Inversion via String Diagrams. *Mathematical Structures in Computer Science*, 29(7), 2019.

[10] B. Jacobs. *Introduction to Coalgebra: Towards Mathematics of States and Observation*. Cambridge University Press, 2016.

[11] M. L. Eaton. *Group Invariance Applications in Statistics*. Institute of Mathematical Statistics, 1989.

[12] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. 2nd ed. Springer, 1998.

[13] T. Cohen and M. Welling. Group Equivariant Convolutional Networks. *Proceedings of ICML*, 2016.

[14] R. Kondor and S. Trivedi. On the Generalization of Equivariance and Convolution in Neural Networks to the Action of Compact Groups. *Proceedings of ICML*, 2018.

[15] M. Weiler and G. Cesa. General E(2)-Equivariant Steerable CNNs. *NeurIPS*, 2019.

[16] M. Finzi, M. Welling, and A. G. Wilson. A Practical Method for Constructing Equivariant Multilayer Perceptrons for Arbitrary Matrix Groups. *Proceedings of ICML*, 2021.

[17] T. S. Cohen, M. Geiger, J. Kohler, and M. Welling. Spherical CNNs. *ICLR*, 2018.

[18] B. Gavranovic, P. Lessard, A. Dudzik, T. von Glehn, J. G. M. Araujo, and P. Velickovic. Position: Categorical Deep Learning is an Algebraic Theory of All Architectures. arXiv:2402.15332, 2024.

[19] B. Fong and D. I. Spivak. *Seven Sketches in Compositionality*. Cambridge University Press, 2019.

[20] Y. Maruyama. Categorical Equivariant Deep Learning: Category-Equivariant Neural Networks and Universal Approximation Theorems. arXiv:2511.18417, 2025.

[21] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger. Inequalities for the L_1 Deviation of the Empirical Distribution. Hewlett-Packard Laboratories Technical Report HPL-2003-97R1, 2003.

[22] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[23] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. 2nd ed. Springer, 2017.

[24] W. Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301), 1963.