Submission and Formatting Instructions for International Conference on Machine Learning (ICML 2025)

Firstname1 Lastname1 [*1] **Firstname2 Lastname2** [*1 2] **Firstname3 Lastname3** [2] **Firstname4 Lastname4** [3]
Firstname5 Lastname5 [1] **Firstname6 Lastname6** [3 1 2] **Firstname7 Lastname7** [2] **Firstname8 Lastname8** [3]
Firstname8 Lastname8 [1 2]

Abstract

1. Introduction

Matryoshka representations (Kusupati et al., 2022) build a single embedding $z = f_\theta(x) \in \mathbb{R}^D$ whose *prefixes* $z^{(m)}(x) = z_{1:d_m}(x) \in \mathbb{R}^{d_m}$, for $m \in \{1, \ldots, M\}$ with $d_1 < \cdots < d_M = D$, serve as deployable, size-adaptive features. Prior work emphasizes compression: smaller prefixes are useful in a crude sense, larger ones refine.

We adopt a broader view: the head index m is a *control axis*. Training can *schedule* along m which inductive biases the representation exhibits e.g., robustness to nuisance factors and calibrated uncertainty early, task-specific detail later. In short, Matryoshka turns dimensional order into a programmable curriculum.

Nested prefixes need not be related only through shared supervision. We can also shape *how* information accumulates as m grows. In addition to specifying the behavior for each $z^{(m)}$, we specify the inter-prefix structure (how $z^{(m)}$ should relate to its predecessor $z^{(m^-)}$ where $m^- = m - 1$). Two objects capture this: (i) a *property objective* $\mathcal{R}_m$ (desired behavior at level m), and (ii) a *consistency relation* $\mathcal{C}_m$ (accumulation across levels; e.g., monotone refinement, stability).

Let ℓ_m denote the task loss consuming $z^{(m)}$. We train a single encoder via

$$\sum_{m=1}^{M} \alpha_m \, \ell_m\big(z^{(m)}\big) + \lambda \sum_{m=1}^{M} \beta_m \, \mathcal{R}_m\big(z^{(m)}\big) + \mu \sum_{m=2}^{M} \gamma_m \, \mathcal{C}_m\big(z^{(m)}, z^{(m-1)}\big), \tag{1}$$

where $(\alpha_m, \beta_m, \gamma_m)$ schedule supervision, property, and consistency along the control axis. This cleanly separates (i) utility at each prefix (ℓ_m), (ii) *what* the prefix should be like ($\mathcal{R}_m$), and (iii) *how* information should accumulate ($\mathcal{C}_m$). Early prefixes can be regularized toward causal/coarse features (high β_m), while later coordinates capture style (higher α_m). $\mathcal{C}_m$ can enforce sparsity, orthogonality, or contracting uncertainty. The result is a single embedding *ordered with intent*.

2. Generalization Bounds

Let $(\mathcal{X}, \mathcal{Y})$ be the input-label space with finite label set $|\mathcal{Y}| = K$ and $\mathcal{P}$ be the unknown data distribution over $(\mathcal{X}, \mathcal{Y})$. We are given a training set $D = \{(x_i, y_i)_{i=1}^n\}$ drawn i.i.d from $\mathcal{P}$.
For a hard label $y \in \mathcal{Y}$ the standard multiclass cross entropy loss for prediction logits $z \in \mathbb{R}^K$ is:

$$l(z, y) = -\log(\text{softmax}(z)_y)$$

Equivalently for a target distribution $q \in \Delta^{K-1}$:

$$l(z, q) = \sum_c q_c \log(\text{softmax}(z)_c)$$

Let our model be an embedding function $f_\theta : \mathcal{X} \to \mathbb{R}^D$ followed by a linear classifier $W \in \mathbb{R}^{K \times D}$ and let parameters of the model be $\phi = (\theta, W)$.

Let $\pi_j : \mathbb{R}^D \to \mathbb{R}^{d_j}$ be the projection operator onto the first d_j coordinates, with $1 \leq j \leq m$, and $d_1 < d_2, \cdots < d_m = D$. Let $W^{(j)} = W_{:,1:d_j}$ be the corresponding truncated classifier weights.
Let $h_\phi^{(j)}(x) = W^{(j)}(\pi_j \circ f_\theta(x))$ be the logits for the j-th head logits, the full hypothesis is $h_\phi = h_\phi^{(m)}$ and let

$$p_\phi^{(j)}(.|x) = \text{softmax}(h_\phi^{(j)}(x))$$

[*]Equal contribution [1]Department of XXX, University of YYY, Location, Country [2]Company Name, Location, Country [3]School of ZZZ, Institute of WWW, Location, Country. Correspondence to: Firstname1 Lastname1 <first1.last1@xxx.edu>, Firstname2 Lastname2 <first2.last2@www.uk>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

The standard Empirical risks objective is to minimize the risk of the full dimension model:

$$\hat{R}_s(\phi) = \frac{1}{n}\sum_{i=1}^{n} l(h_\phi(x_i), y_i)$$

The Matryoshka loss in a weighted sum of empirical risks for all the heads:

$$\hat{R}_{MCE}(\phi) = \sum_{j=1}^{m} \alpha_j \left(\frac{1}{n}\sum_{i=1}^{n} l(h_\phi^{(j)}(x_i), y_i) \right)$$

with fixed weights $\alpha_j > 0$. Let $\alpha := \sum_{j=1}^{m} \alpha_j$.

Define weighted consensus between heads as $g_\phi(.|x) \in \Delta^{K-1}$ to be the weighted geometric mean of heads posteriors (product of experts):

$$g_\phi(c|x) = \frac{\exp\left(\frac{1}{\alpha}\sum_{j=1}^{m}\alpha_j \log p_\phi^{(j)}(c|x)\right)}{\sum_{c'=1}^{K}\exp\left(\frac{1}{\alpha}\sum_{j=1}^{m}\alpha_j \log p_\phi^{(j)}(c'|x)\right)} \tag{2}$$

Let $h_g(x) \in \mathbb{R}^K$ denote any logits with $\text{softmax}(h_g(x)) = g_\phi(.|x)$, i.e. $h_g = \log g_\phi$ up to an additive constant per x.

Now for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$:

$$\begin{aligned}\sum_{j=1}^{m} \alpha_j l(h_\phi^{(j)}(x), y) = \\ \alpha l(h_g(x), y) + \sum_{j=1}^{m} \alpha_j \text{KL}(g_\phi(.|x)||p_\phi^{(j)}(.|x))\end{aligned} \tag{3}$$

e over dataset

Proof. For simplicity let's denote, $p_j := p_\phi^{(j)}(.|x)$, $g := g_\phi(.|x)$ and

$$A(c) := \frac{1}{\alpha}\sum_{j=1}^{m} \alpha_j \log p_j(c)$$

Now we have from 2:

$$g(c) = \frac{\exp A(c)}{\sum_{c'=1}^{K} \exp A(c')}$$

Let's denote the normalizing constant by $Z := \sum_{c'=1}^{K} \exp A(c')$, so:

$$\log g(c) = A(c) - \log Z$$

So now we can rewrite the first term in 2:

$$\begin{aligned}\alpha l(h_g, y) &= -\alpha \log(\text{softmax}(h_g)_y) \\ &= -\alpha \log g(y) \\ &= -\alpha A(c) + \alpha \log Z.\end{aligned}$$

Now the second term, KL sum:

$$\begin{aligned}&\sum_{j=1}^{m} \alpha_j \text{KL}(g||p_j) \\ &= \sum_{j=1}^{m} \alpha_j \sum_c g(c)(\log g(c) - \log p_j(c)) \\ &= \alpha \sum_c g(c) \log g(c) - \sum_c g(c) \sum_{j=1}^{m} \alpha_j \log p_j(c) \\ &= \alpha \sum_c g(c)(A(c) - \log Z) - \sum_c g(c)\alpha A(c) \\ &= \alpha \sum_c g(c)A(c) - \alpha \sum_c g(c)A(c) - \alpha \log Z \sum_c g(c) \\ &= -\alpha \log Z\end{aligned}$$

So now summing these two terms, yields:

$$\begin{aligned}\alpha(h_g, y) + \sum_{j=1}^{m} \alpha_j \text{KL}(g||p_j) &= -\alpha A(c) \\ &= -\sum_{j=1}^{m} \alpha_j \log p_j(y) = \sum_{j=1}^{m} \alpha_j l(h_\phi^{(j)}, y)\end{aligned}$$

□

Write the general form using bregman divergence loss/using fenchel-young losses, got the idea from chat-gpt

So for a samples $D = \{(x_i, y_i)_{i=1}^{n}\}$ the MRL loss would look like:

$$\hat{R}_{\text{MCE}}(\phi) = \alpha.\hat{R}_S^{(g)}(\phi) + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_j \text{KL}(g_\phi(.|x_i)||p_\phi^{(j)}(.|x_i)) \tag{4}$$

where $\hat{R}_{\text{MCE}}(\phi) = \frac{1}{n}\sum_{i=1}^{n} l(h_g(x_i), y_i)$ is the ERM for consensus term (geometric mean) and we define the agreement as

$$\text{Agr}_\phi(x) := \sum_{j=1}^{m} \alpha_j \text{KL}(g_\phi(.|x_i)||p_\phi^{(j)}(.|x_i)) \tag{5}$$

And the average agreement as $\hat{A}_D := \frac{1}{n}\sum_{i=1}^{n} \text{Agr}_\phi(x_i)$. Note that $\hat{A}_D = 0$ almost surely if and only if $p_\phi^{(1)} = \cdots = p_\phi^{(m)} = g_\phi$ and this decomposition holds for any weights α_j.

talk about why geometric mean difference with regular mean mixture of expert product of expert view

Restating theorem 1 from (McAllester, 1999): Let $\mathcal{H}$ be a hypothesis class, and for $h \in \mathcal{H}$, we have
For prior P on $\mathcal{H}$ and a posterior Q on $\mathcal{H}$, define:

$$\mathcal{R}(Q) := \mathbb{E}_{h\sim Q,(x,y)\sim\mathcal{P}}[l(h(x), y)]$$

$$\hat{\mathcal{R}}_D(Q) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{h\sim Q}[l(h(x_i), y)]$$

For any $\delta \in (0,1)$ with probability at least $1-\delta$ over the selection of the sample $D \sim \mathcal{P}^n$, for all posteriors:

$$\mathcal{R}(Q) \leq \hat{\mathcal{R}}_D(Q) + \sqrt{\frac{D_{\mathrm{KL}}(Q||P) + \ln(1/\delta)}{2n}} \tag{6}$$

So now in hierarchical setting we have multiple predictors, a natural way to define a probability disitrbution over this is using gibbs posterior which simply is sample index $J \in \{1, \cdots, m\}$ with $\mathbb{P}(J=j) = \alpha_j/\alpha$ and use head $h_\phi^{(J)}$. Denote this predictor by Q_ϕ. Then, for each (x,y) we have:

$$\begin{aligned}\mathbb{E}_J[l(h_\phi^{(J)}(x), y)] &= \frac{1}{\alpha}\sum_{j=1}^{m} \alpha_j l(h_\phi^{(j)}(x), y) \\ &= l(h_g(x), y) + \frac{1}{\alpha}\mathrm{Agr}_\phi(x)\end{aligned}$$

Averaging over dataset, we have:

$$\hat{R}_D(Q_\phi) = \hat{R}_D^{(g)}(\phi) + \frac{1}{\alpha}\hat{A}_D. \tag{7}$$

And $\hat{R}_D^{(g)}(\phi) = \frac{1}{n}\sum_{i=1}^{n} l(h_g(x_i), y_i)$

Assumptions: We assume logits are bounded $||z||_\infty \leq B$ and we use temperature-scaled softmax with $\tau \geq 1$:

$$p(c|x) = \frac{e^{z_c/\tau}}{\sum_{c'} e^{z_{c'}/\tau}}$$

It's easy to check that each class probability satisfies $p_c \geq m^* := \frac{e^{-2B/\tau}}{K} > 0$, this temperature scaling is a standard calibration that flattens overly confident probabilities (Guo et al., 2017). So we assume softmax$(h_g/\tau) = g_\phi$.

In the decomposition 2 we use reverse KL, however in PAC-bayes bound the complexity term has a forward KL between a posterior and a prior. To connect them we use the result that on a truncated simplex (i.e. all probabilities bounded away from zero and one) all f-divergences are equivalent up to constants. This means there exists a finite constant $c_\leftrightarrow$ (depending only on the minimal probability $m^* > 0$ and number of classes K) such that, for any two discrete distributions p, q with $p_c, q_c \geq m^*$:

$$\mathrm{KL}(p||q) \leq c_\leftrightarrow \mathrm{KL}(q||p)$$

This standard inequality follows (Sason & Verdú, 2016) and Intuitively, once probabilities cannot collapse to zero, forward and reverse KL are Lipschitz-equivalent.

Hierarchical PAC-Bayes bounds Under mentioned assumptions, if there exists $B > 0, \tau \geq 1$ such that $||h_\phi^{(j)}(x)||_\infty \leq B$ for all j, x. Then the probabilities lie in the m^* truncated simplex with $m^* := e^{-2B/\tau}/K$. On this set, there is a constant $c_\leftrightarrow = c_\leftrightarrow(m^*, K) \in (0, \infty)$ such that 2 holds.

Let Q_ϕ be the distribution that draws $J \in \{1, ..., m\}$ with $\mathbb{P}(J=j) = \alpha_j/\alpha$ and predicts with head $h_\phi^{(J)}$. Let P be prior on the same discrete set, same mixture before seeing D: With decomposition for Q_ϕ 2 we can control the KL term by:

$$\mathrm{KL}(Q_\phi||P) \leq \frac{c_\leftrightarrow}{\alpha}\hat{A}_D + C_0$$

for a constant $C_0 \geq 0$ independent of ϕ.

Now for any $\delta \in (0,1)$ with probability at least $1-\delta$ over D we have

$$R(Q_\phi) \leq \hat{R}_D^{(g)}(\phi) + \frac{\hat{A}_D}{\alpha} + \sqrt{\frac{\frac{c_\leftrightarrow}{\alpha}\hat{A}_D + C_0 + ln(1/\delta)}{2n}} \tag{8}$$

Proof. Apply the regular PAC-bayes bound on Q_ϕ. The empirical terms equals to $\hat{R}_D^{(g)}(\phi) + \frac{\hat{A}_D}{\alpha}$. The complexity is controlled by forward KL between the head mixture and the prior by inequality between reverse and forward KL. □

Now let's consider a single head hypothesis class $\mathcal{H}_1$ and define $p^* = \arg\min_{p \in \mathcal{H}} \hat{R}_D(p)$. Given $g_\phi \in \mathcal{H}_1$ we can say

$$\Delta := \hat{R}_D(p^*) - \hat{R}_D^{(g)}(\phi) \geq 0$$

Now define single head proximity to consensus as:

$$\mathrm{KLF} := \frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}(p^*(.|x_i)||g_\phi(.|x_i))$$

Then we probability at least $1-\delta$:

$$R(p^*) \leq \hat{R}_D(p^*) + \sqrt{\frac{\mathrm{KLF} + \ln(1/\delta)}{2n}}$$

3. Training Objectives

Let $\{(x_i, y_i)\}_{i=1}^n$ be the training set with $y_i \in \{1, \ldots, C\}$. The network exposes M nested heads with dimensions $d_1 < \cdots < d_M = D$. Head m produces logits $f^{(m)}(x) \in \mathbb{R}^C$ and probabilities

$$p^{(m)}(\cdot \mid x) = \mathrm{softmax}\big(f^{(m)}(x)\big).$$

Standard MRL (per-head cross-entropy).

$$\mathcal{L}_{\mathrm{MRL}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} \lambda_m \, \ell_{\mathrm{CE}}^{(m)}(x_i, y_i), \; \ell_{\mathrm{CE}}^{(m)}(x, y) = -\log p^{(m)}(y \mid x). \tag{9}$$

3.1. Logit Adjustment Across Heads

We optionally adjust each head's logits by adding log-probabilities from a *held-out* head $h(m)$:

$$\tilde{f}^{(m)}(x) = f^{(m)}(x) + \log\big(\bar{p}^{(h(m))}(\cdot \mid x) + \epsilon\big), \quad (10)$$

$$\ell^{(m)}_{\text{ADJ}}(x, y) = -\log \operatorname{softmax}\big(\tilde{f}^{(m)}(x)\big)_y. \quad (11)$$

We aggregate as in equation 9 with $\ell^{(m)}_{\text{CE}}$ replaced by $\ell^{(m)}_{\text{ADJ}}$. In our default schedule, $h(m) = \max(1, m-1)$, i.e., each head uses the immediately smaller head's probabilities (and for $m = 1$ we set the adjustment to zero). Here $\epsilon > 0$ is a numerical stabilizer and $\bar{p}^{(h)}$ denotes (optionally calibrated) probabilities from head h.

3.2. Covariance Dissimilarity Regularizer

Let $z(x) \in \mathbb{R}^D$ be the shared embedding and $Z \in \mathbb{R}^{B\times D}$ the batch matrix with rows $z(x_i)^\top$. For $m \geq 2$, define the prefix/suffix slices

$$Z^{(m)}_{\text{pre}} = Z_{[:,\,1:d_{m-1}]} \in \mathbb{R}^{B\times d_{m-1}}$$

$$Z^{(m)}_{\text{suf}} = Z_{[:,\,d_{m-1}+1:d_m]} \in \mathbb{R}^{B\times (d_m - d_{m-1})}.$$

Using batch-centered scatters

$$C(A) \;=\; \big(A - \mathbf{1}\mu^\top\big)^\top \big(A - \mathbf{1}\mu^\top\big), \mu = \tfrac{1}{B}\sum_{b=1}^{B} A_{b:}, \quad (12)$$

define the Frobenius cosine similarity

$$\cos_F(U, V) \;=\; \frac{\langle U, V\rangle_F}{\|U\|_F\,\|V\|_F + \epsilon}, \langle U, V\rangle_F = \operatorname{trace}(U^\top V). \quad (13)$$

The regularizer encourages decorrelation between increments:

$$\mathcal{R}_{\text{cov}} \;=\; \sum_{m=2}^{M} \cos_F\Big(C\big(Z^{(m)}_{\text{pre}}\big),\, C\big(Z^{(m)}_{\text{suf}}\big)\Big). \quad (14)$$

The total objective is

$$\mathcal{L}_{\text{Cov}} \;=\; \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} \lambda_m\, \ell^{(m)}_{\text{CE}}(x_i, y_i) \;+\; \lambda_{\text{cov}}\, \mathcal{R}_{\text{cov}}, \quad (15)$$

with hyperparameters $\{\lambda_m\}$ and $\lambda_{\text{cov}} > 0$. Note that equation 13 is scale-invariant, so using the unnormalized scatter in equation 12 is harmless.

3.3. Orthogonalized Matching–Pursuit (OMP) Matryoshka Head

Let $z = f_\theta(x) \in \mathbb{R}^D$ be the shared embedding and let the nested prefix sizes be $1 \leq d_1 < \cdots < d_M = D$. Denote the (raw) prefix and the new block at level m by

$$z^{(m)}_{\text{raw}} \;=\; z_{1:d_m}, \qquad \Delta^{(m)} \;=\; z_{(d_{m-1}+1):d_m} \quad (m \geq 2),$$

with $z^{(1)}_{\text{raw}} = z_{1:d_1}$ and $d_0 := 0$.

Given a minibatch with prefix matrix $Z_p \in \mathbb{R}^{B\times d_{m-1}}$ and suffix matrix $Z_s \in \mathbb{R}^{B\times(d_m - d_{m-1})}$, define the ridge projector

$$P_\lambda(Z_p) \;=\; Z_p\big(Z_p^\top Z_p + \lambda I_{d_{m-1}}\big)^{-1} Z_p^\top, \qquad \lambda > 0,$$

and the residual operator

$$\mathcal{R}_\lambda(Z_p, Z_s) \;=\; Z_s - P_\lambda(Z_p)\, Z_s.$$

At level $m \geq 2$, we orthogonalize the new block against the current prefix by

$$\widehat{\Delta}^{(m)} \;=\; \mathcal{R}_\lambda\big(Z_p, Z_s\big) \quad \text{(row-wise, per sample).} \quad (16)$$

The orthogonalized prefix feature used at head m is the concatenation

$$\widehat{z}^{(1)} \;=\; z^{(1)}_{\text{raw}}, \widehat{z}^{(m)} \;=\; \big[\, z_{1:d_{m-1}}\,;\, \widehat{\Delta}^{(m)} \big] \in \mathbb{R}^{d_m} \quad (m \geq 2). \quad (17)$$

For $\lambda \to 0$ and $\operatorname{rank}(Z_p) = d_{m-1}$, equation 16 coincides with the ordinary orthogonal projector, recovering a matching–pursuit style update that removes the component of the new block lying in $\operatorname{span}(Z_p)$.

Implementation details mirrored from code. (i) The ridge solve uses a numerically stable linear system $A^\star = \big(Z_p^\top Z_p + \lambda I\big)^{-1} Z_p^\top Z_s$ and computes the projection as $Z_p A^\star$ (no explicit matrix inverse). (ii) The residualization in equation 16 is performed *without tracking gradients* and in full precision to avoid AMP-induced instabilities; its output is cast back to the model dtype.

4. Out of Distribution Generalization

Setup. Let $f_\theta : \mathcal{X} \to \mathbb{R}^D$ be an encoder; for input x, write $z = f_\theta(x) \in \mathbb{R}^D$. For Matryoshka heads with prefix dimensions $\mathcal{D} = \{D_1 < \cdots < D_M\}$, define the head-restricted, ℓ_2-normalized embedding

$$\tilde{z}^{(m)}(x) \;=\; \frac{z_{1:D_m}}{\|z_{1:D_m}\|} \in \mathbb{R}^{D_m}.$$

Episode construction (shared for all K). Fix a shot count $K \in \{0, 1, 3, 5, 10, 20\}$. For each class c with at least K labeled samples in the target dataset, choose a *support* set $\mathcal{S}_c$ of size K (random without replacement), and let the remaining labeled samples of class c form the *query* set $\mathcal{Q}_c$. Let $\mathcal{C}_K = \{c : |\mathcal{S}_c| = K,\, |\mathcal{Q}_c| \geq 1\}$ be the participating classes, and $\mathcal{Q} = \bigcup_{c\in\mathcal{C}_K} \mathcal{Q}_c$ the query pool. (Classes that would yield zero queries are excluded from $\mathcal{C}_K$.)

Nearest Class Mean (NCM) for $K \geq 1$. For head m and each $c \in \mathcal{C}_K$, form the prototype

$$\mu^{(m)}_c \;=\; \frac{1}{K}\sum_{x\in\mathcal{S}_c} \tilde{z}^{(m)}(x), \qquad \hat{\mu}^{(m)}_c \;=\; \frac{\mu^{(m)}_c}{\big\|\mu^{(m)}_c\big\|}.$$

Score a query $x \in \mathcal{Q}$ by cosine similarity to prototypes:

$$s_{\text{NCM}}^{(m)}(x,c) \;=\; \left\langle \tilde{z}^{(m)}(x),\, \hat{\mu}_c^{(m)} \right\rangle$$
$$\hat{y}^{(m)}(x) \;=\; \arg\max_{c \in \mathcal{C}_K} s_{\text{NCM}}^{(m)}(x,c).$$

Restricted $K{=}0$ baseline (same queries/labels). For $K = 0$, we *do not* form prototypes. Instead, we evaluate the trained classifier head(s) on the *same query indices* $\mathcal{Q}$ and restrict predictions to the *same label set* $\mathcal{C}_1$ (from the $K{=}1$ split). Let $L^{(m)}(x) \in \mathbb{R}^{C_{\text{full}}}$ be the logits of head m; define

$$\tilde{L}^{(m)}(x) \;=\; L^{(m)}(x)\big|_{\mathcal{C}_1} \in \mathbb{R}^{|\mathcal{C}_1|},$$
$$\hat{y}_{K=0}^{(m)}(x) \;=\; \arg\max_{c \in \mathcal{C}_1} \tilde{L}_c^{(m)}(x).$$

This "restricted zero-shot" makes $K{=}0$ directly comparable to $K{\ge}1$: same queries, same label universe.

Metrics (per head m). For Top-k accuracy,

$$\text{Acc}_{\text{Top-}k}^{(m)}(K) \;=\; \frac{1}{|\mathcal{Q}|} \sum_{x \in \mathcal{Q}} \mathbf{1}\Big\{\, y(x) \in \text{Top-}k\big(S_K^{(m)}(x,\cdot)\big) \Big\},$$

where

$$S_K^{(m)}(x,c) \;=\; \begin{cases} s_{\text{NCM}}^{(m)}(x,c), & K \ge 1,\ c \in \mathcal{C}_K, \\ \tilde{L}_c^{(m)}(x), & K = 0,\ c \in \mathcal{C}_1. \end{cases}$$

We report curves over $K \in \{0,1,3,5,10,20\}$ for each $D_m \in \mathcal{D}$.

5. Label Correction

6. Nested representaitons

6.1. Nested Linear Representations and Gram Matrices

Let $X \in \mathbb{R}^{n \times p}$ be the data matrix (with rows $x_i^\top$) and let $W \in \mathbb{R}^{p \times d}$ be a linear feature map. Define the (full) representation matrix

$$R_1 := XW \in \mathbb{R}^{n \times d}.$$

We split the d features into two blocks of equal size, $d = d_2 + d_3$ with $d_2 = d_3 = d/2$, writing

$$W = \begin{bmatrix} W^{(2)} & W^{(3)} \end{bmatrix}, \qquad W^{(2)} \in \mathbb{R}^{p \times d_2},\ W^{(3)} \in \mathbb{R}^{p \times d_3}.$$

Then

$$R_1 = XW = \begin{bmatrix} R_2 & R_3 \end{bmatrix},$$
$$R_2 := XW^{(2)} \in \mathbb{R}^{n \times d_2},\ R_3 := XW^{(3)} \in \mathbb{R}^{n \times d_3}.$$

Define the sample–sample Gram (or representation covariance) matrices

$$S_1 := R_1 R_1^\top \in \mathbb{R}^{n \times n}, \qquad S_2 := R_2 R_2^\top \in \mathbb{R}^{n \times n}.$$

Proposition 6.1 (Decomposition of the Gram matrix)**.** *With the notation above,*

$$S_1 = S_2 + S_3, \qquad \textit{where } S_3 := R_3 R_3^\top \succeq 0.$$

Proof. We have

$$R_1 = [R_2\ R_3], \qquad R_1^\top = \begin{bmatrix} R_2^\top \\ R_3^\top \end{bmatrix}.$$

Thus

$$S_1 = R_1 R_1^\top = [R_2\ R_3] \begin{bmatrix} R_2^\top \\ R_3^\top \end{bmatrix} = R_2 R_2^\top + R_3 R_3^\top = S_2 + S_3.$$

Each matrix of the form $R_k R_k^\top$ is a Gram matrix and hence positive semidefinite. Therefore $S_3 \succeq 0$. □

Inclusion and Loewner order. From $S_1 = S_2 + S_3$ with $S_3 \succeq 0$ we obtain

$$S_1 \succeq S_2 \quad \text{(Loewner order)}.$$

Equivalently, for all $v \in \mathbb{R}^n$,

$$v^\top S_1 v \;\ge\; v^\top S_2 v.$$

In this sense, S_2 is "contained" in S_1.

Moreover,

$$\text{range}(S_2) = \text{span}\{\text{columns of } R_2\} \subseteq \text{span}\{\text{columns of } R_1\} = \text{range}(S_1),$$

and

$$\text{rank}(S_2) \;\le\; \text{rank}(S_1).$$

Energy decomposition. Using $\text{tr}(RR^\top) = \|R\|_F^2$, we have

$$\text{tr}(S_1) = \text{tr}(S_2) + \text{tr}(S_3),$$

so the total feature "energy" (squared Frobenius norm of R_1) decomposes additively into contributions from the two halves of the feature map.

6.2. Nested Predictions

Consider scalar predictions for simplicity. Let

$$w_1 \in \mathbb{R}^d$$

denote the weight vector for the full model, and partition it as

$$w_1 = \begin{bmatrix} w_2 \\ w_3 \end{bmatrix}, \qquad w_2 \in \mathbb{R}^{d_2},\ w_3 \in \mathbb{R}^{d_3}.$$

Define the prediction vectors (logits) on the n samples as

$$\hat{y}^{(1)} := R_1 w_1 \in \mathbb{R}^n, \qquad \hat{y}^{(2)} := R_2 w_2 \in \mathbb{R}^n, \qquad \hat{y}^{(3)} := R_3 w_3 \in \mathbb{R}^n.$$

Then

$$\begin{aligned}\hat{y}^{(1)} &= R_1 w_1 = [R_2 \; R_3] \begin{bmatrix} w_2 \\ w_3 \end{bmatrix} \\ &= R_2 w_2 + R_3 w_3 = \hat{y}^{(2)} + \hat{y}^{(3)}.\end{aligned}$$

Hence, at the logit level we always have the additive decomposition

$$\boxed{\hat{y}^{(1)} = \hat{y}^{(2)} + \hat{y}^{(3)}}.$$

If probabilities are obtained via a non-linear link function σ, for example

$$p^{(k)} := \sigma\big(\hat{y}^{(k)}\big),$$

then the linear decomposition holds in *logit space*, not in probability space in general.

Norm-squared decomposition. The squared prediction norm satisfies

$$\begin{aligned}\left\|\hat{y}^{(1)}\right\|_2^2 &= \left\|\hat{y}^{(2)} + \hat{y}^{(3)}\right\|_2^2 \\ &= \left\|\hat{y}^{(2)}\right\|_2^2 + \left\|\hat{y}^{(3)}\right\|_2^2 + 2\left\langle \hat{y}^{(2)}, \hat{y}^{(3)} \right\rangle.\end{aligned}$$

We obtain a Pythagorean relation

$$\left\|\hat{y}^{(1)}\right\|_2^2 = \left\|\hat{y}^{(2)}\right\|_2^2 + \left\|\hat{y}^{(3)}\right\|_2^2$$

if and only if

$$\left\langle \hat{y}^{(2)}, \hat{y}^{(3)} \right\rangle = 0,$$

i.e., if the contributions from the two halves of the features are orthogonal in $\mathbb{R}^n$. In general, the cross-term $2\langle \hat{y}^{(2)}, \hat{y}^{(3)} \rangle$ does not vanish.

6.3. Nested Hypothesis Classes

Define the hypothesis classes associated with the two feature blocks as

$$\begin{aligned}\mathcal{H}_2 &:= \left\{ x \mapsto w_2^\top R_2(x) : w_2 \in \mathbb{R}^{d_2} \right\}, \\ \mathcal{H}_1 &:= \left\{ x \mapsto w_1^\top R_1(x) : w_1 \in \mathbb{R}^{d} \right\}.\end{aligned}$$

Any predictor realizable by the smaller feature set can be realized by the larger feature set by padding the weights with zeros:

$$R_2 w_2 = R_1 \begin{bmatrix} w_2 \\ 0 \end{bmatrix}.$$

Thus

$$\mathcal{H}_2 \subseteq \mathcal{H}_1.$$

Consequently, for any loss functional $\mathcal{L}$ (e.g. population risk under a fixed data distribution),

$$\inf_{h \in \mathcal{H}_1} \mathcal{L}(h) \;\le\; \inf_{h \in \mathcal{H}_2} \mathcal{L}(h),$$

i.e., the best achievable loss with the richer feature set is no worse than with the smaller one.

6.4. Monotone Kernel Sequences

The previous construction naturally extends to multiple nested blocks. Suppose we have blocks $R_{(k)} \in \mathbb{R}^{n \times d_k}$ and define

$$R_{\le K} := [R_{(1)} \; R_{(2)} \; \cdots \; R_{(K)}], \qquad S_{\le K} := R_{\le K} R_{\le K}^\top.$$

Then

$$S_{\le K} = \sum_{k=1}^{K} R_{(k)} R_{(k)}^\top,$$

so

$$S_{\le 1} \preceq S_{\le 2} \preceq \cdots \preceq S_{\le K}$$

in the Loewner order. This yields a monotone sequence of kernels, where each additional feature block contributes a positive semidefinite increment to the Gram matrix.

Acknowledgements

Do not include acknowledgements in the initial version of the paper submitted for blind review.

If a paper is accepted, the final camera-ready version can (and usually should) include acknowledgements. Such acknowledgements should be placed at the end of the section, in an unnumbered section that does not count towards the paper page limit. Typically, this will include thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.

Impact Statement

Authors are **required** to include a statement of the potential broader impact of their work, including its ethical aspects and future societal consequences. This statement should be in an unnumbered section at the end of the paper (co-located with Acknowledgements – the two may appear in either order, but both must be before References), and does not count toward the paper page limit. In many cases, where the ethical impacts and expected societal implications are those that are well established when advancing the field of Machine Learning, substantial discussion is not required, and a simple statement such as the following will suffice:

> "This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here."

The above statement can be used verbatim in such cases, but we encourage authors to think about whether there is content which does warrant further discussion, as this statement will be apparent if the paper is later flagged for ethics review.

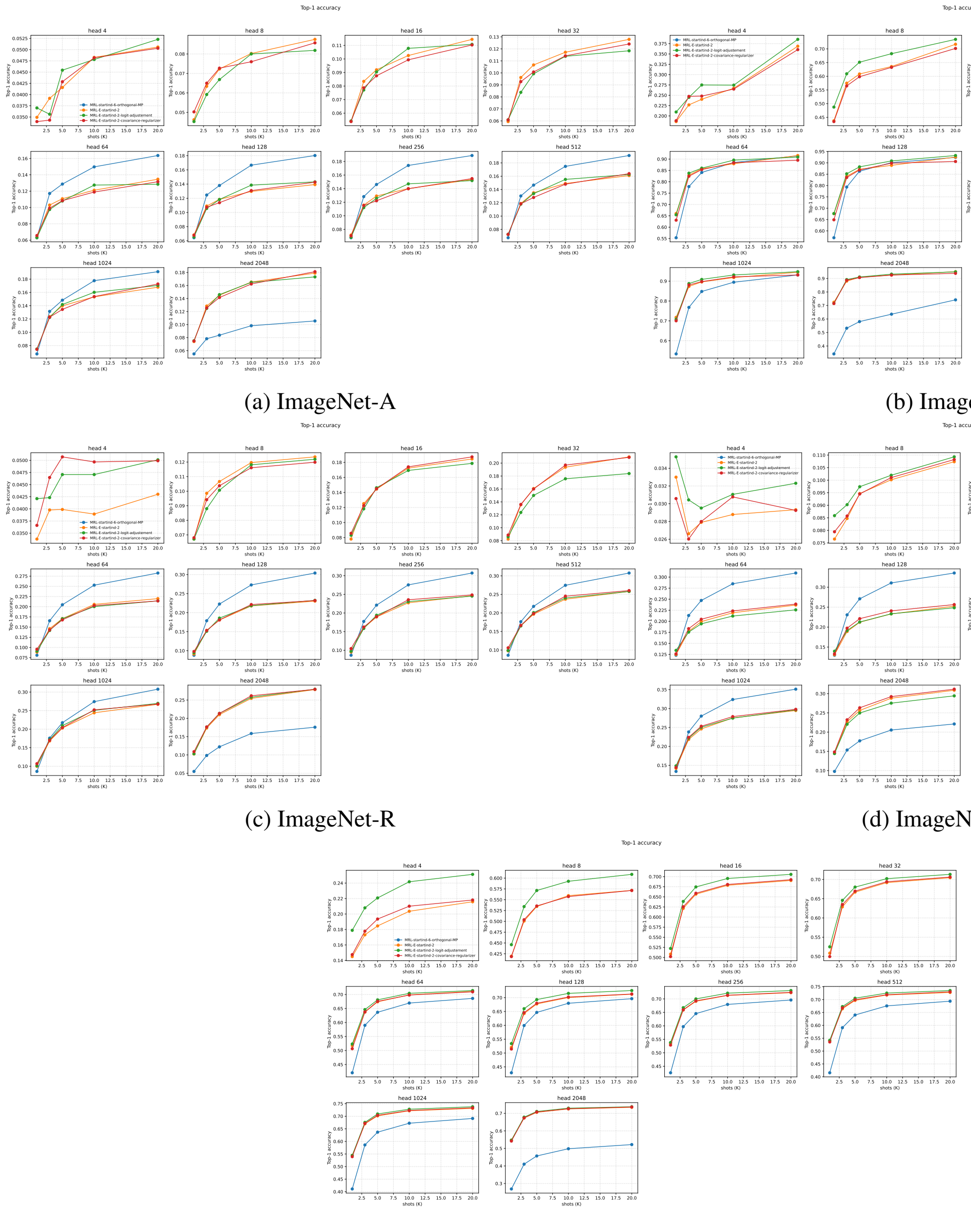
References

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P., et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

McAllester, D. A. Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pp. 164–170, 1999.

Sason, I. and Verdú, S. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.



(a) ImageNet-A

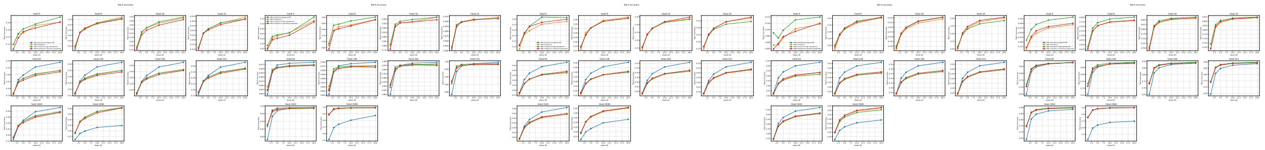
(b) ImageNet-O

(c) ImageNet-R

(d) ImageNet-Sketch

(e) Validation

Figure 1. **Top-1 accuracy** vs. shots $K \in \{0, 1, 3, 5, 10, 20\}$ for each Matryoshka head D_m. Each panel shows per-head curves for all compared models on a given target dataset.



(a) ImageNet-A

(b) ImageNet-O

(c) ImageNet-R

(d) ImageNet-Sketch

(e) Validation

Figure 2. **Top-5 accuracy** vs. shots $K \in \{0, 1, 3, 5, 10, 20\}$ for each Matryoshka head D_m, same layout as Fig. 1.

A. You *can* have an appendix here.

You can have as much text here as you want. The main body must be at most 8 pages long. For the final version, one more page can be added. If you want, you can use an appendix like this one.

The `\onecolumn` command above can be kept in place if you prefer a one-column appendix, or can be removed if you prefer a two-column appendix. Apart from this possible change, the style (font size, spacing, margins, page numbering, etc.) should be kept the same as the main body.