A APPENDIX

B PROOF OF EQUATION (2)

As a reminder, we consider C class labels and denote by Δ_C the C-dimensional simplex. We define the set of distributions $\mathcal{Q}$ over the C class labels by

$$\mathcal{Q} = \{q(y|\cdot) \,|\, \forall \boldsymbol{x} \in \mathcal{X}, q(y|\boldsymbol{x}) \in \Delta_C\}.$$

Consider the optimization problem

$$\min_{q \in \mathcal{Q}} \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x})} \left[D_{KL}(p(y|\boldsymbol{x}) \,\|\, q(y|\boldsymbol{x})) \right] \tag{5}$$

whose solution is straightforwardly given by the marginal distribution $\boldsymbol{x} \mapsto q^\star(y|\boldsymbol{x}) = p(y|\boldsymbol{x})$. We recall that the KL $D_{KL}(p(y|\boldsymbol{x}) \,\|\, q(y|\boldsymbol{x}))$ is defined by

$$D_{KL}(p(y|\boldsymbol{x}) \,\|\, q(y|\boldsymbol{x})) = \sum_{j=1}^{C} p_j(y|\boldsymbol{x}) \log \left(\frac{p_j(y|\boldsymbol{x})}{q_j(y|\boldsymbol{x})} \right). \tag{6}$$

For any $\boldsymbol{x}$ and $j \leq C$, we can rewrite the terms of the sum

$$p_j(y|\boldsymbol{x}) \log \left(\frac{p_j(y|\boldsymbol{x})}{q_j(y|\boldsymbol{x})} \right)$$

as

$$\mathbb{E}_{\boldsymbol{a}|\boldsymbol{x} \sim p(\boldsymbol{a}|\boldsymbol{x})} \left[p_j(y|\boldsymbol{x}, \boldsymbol{a}) \log \left(\frac{p_j(y|\boldsymbol{x})}{q_j(y|\boldsymbol{x})} \right) \right]$$

where we have used (i) the fact that $\log(\frac{p_j(y|\boldsymbol{x})}{q_j(y|\boldsymbol{x})})$ does not depend on $\boldsymbol{a}$ and (ii) the definition of the marginal distribution

$$\begin{aligned} p_j(y|\boldsymbol{x}) = & \int p_j(y|\boldsymbol{x}, \boldsymbol{a}) p(\boldsymbol{a}|\boldsymbol{x}) d\boldsymbol{a} \\ = & \mathbb{E}_{\boldsymbol{a}|\boldsymbol{x} \sim p(\boldsymbol{a}|\boldsymbol{x})} \left[p_j(y|\boldsymbol{x}, \boldsymbol{a}) \right]. \end{aligned}$$

Multiplying and dividing in the argument of the log by $p_j(y|\boldsymbol{x}, \boldsymbol{a})$, we obtain

$$\mathbb{E}_{\boldsymbol{a}|\boldsymbol{x} \sim p(\boldsymbol{a}|\boldsymbol{x})} \left[p_j(y|\boldsymbol{x}, \boldsymbol{a}) \log \left(\frac{p_j(y|\boldsymbol{x}, \boldsymbol{a})}{q_j(y|\boldsymbol{x})} \frac{p_j(y|\boldsymbol{x})}{p_j(y|\boldsymbol{x}, \boldsymbol{a})} \right) \right].$$

Summing over $j \in \{1, \ldots, C\}$ to reconstruct the KL term (6), this leads to, for any $\boldsymbol{x}$,

$$\begin{aligned} D_{KL}(p(y|\boldsymbol{x}) \,\|\, q(y|\boldsymbol{x})) = & \mathbb{E}_{\boldsymbol{a}|\boldsymbol{x}} \left[D_{KL}(p(y|\boldsymbol{x}, \boldsymbol{a}) \,\|\, q(y|\boldsymbol{x})) \right] \\ & - \mathbb{E}_{\boldsymbol{a}|\boldsymbol{x}} \left[D_{KL}(p(y|\boldsymbol{x}, \boldsymbol{a}) \,\|\, p(y|\boldsymbol{x})) \right]. \end{aligned}$$

Since the second term above *does not depend on* q, minimizing (5) is equivalent to minimizing

$$\begin{aligned} & \min_{q \in \mathcal{Q}} \quad \mathbb{E}_{\boldsymbol{x}} \left[\mathbb{E}_{\boldsymbol{a}|\boldsymbol{x}} \left[D_{KL}(p(y|\boldsymbol{x}, \boldsymbol{a}) \,\|\, q(y|\boldsymbol{x})) \right] \right] \\ = & \min_{q \in \mathcal{Q}} \quad \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{a}) \sim p(\boldsymbol{x}, \boldsymbol{a})} \left[D_{KL}(p(y|\boldsymbol{x}, \boldsymbol{a}) \,\|\, q(y|\boldsymbol{x})) \right] \end{aligned}$$

which is equal to (2) and which is, analogously to (5), minimized by the marginal distribution $\boldsymbol{x} \mapsto q^\star(y|\boldsymbol{x}) = p(y|\boldsymbol{x})$.

C HETEROSCEDASTIC MOTIVATION

We consider a simplified special case of our framework in which the conditional model $p(y|\boldsymbol{x}, \boldsymbol{a})$ is *homoscedastic* but the optimal variational distribution in the sense of Eq. 2 is *heteroscedastic*. This motivates **Het-TRAM**, in which the variational approximations $q(y|\boldsymbol{x})$ and $q(y|\boldsymbol{x}, \boldsymbol{a})$ are heteroscedastic.

Suppose we have a regression dataset constructed from labels assigned by M annotators. Each annotator has their own homoscedastic Gaussian model $p(y|\boldsymbol{x}, a = m) = \mathcal{N}(\mu_{\theta_m}(\boldsymbol{x}), 1)$. Here the

Table 4: Pre-trained models used to re-label ImageNet ILSVRC12 training set and their accuracy on that training set.

Model	Training set accuracy
ResNet50V2	0.70086
ResNet101V2	0.72346
ResNet152V2	0.72738
DenseNet121	0.74782
DenseNet169	0.76184
DenseNet201	0.77344
InceptionResNetV2	0.8049
InceptionV3	0.77994
MobileNet	0.70594
MobileNetV2	0.71458
MobileNetV3Large	0.75622
MobileNetV3Small	0.68158
NASNetMobile	0.74302
VGG16	0.71178
VGG19	0.71156
Xception	0.79076

PI is a single discrete Categorical feature representing the annotator ID which takes one of M values with equal probability, $a \sim \text{Cat}(\frac{1}{M})$.

The marginal $p(y|\boldsymbol{x})$ is a Gaussian Mixture Model. We choose our variational family to be the Gaussian distribution, $q(y|\boldsymbol{x}) = \mathcal{N}(\mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x}))$. The values of μ and σ^2 that minimize Eq. 2 are: $\mu_*(\boldsymbol{x}) = \frac{1}{M}\sum_m \mu_{\theta_m}(\boldsymbol{x})$ and $\sigma^2_*(\boldsymbol{x}) = (M - \mu_*(\boldsymbol{x})) + \frac{1}{M}\sum_m \mu^2_{\theta_m}(\boldsymbol{x})$ (Lakshminarayanan et al., 2017). Crucially note that despite the conditional distribution being homoscedastic, the best variational distribution is heteroscedastic as the variance depends on the location in $\mathcal{X}$ space.

D Experimental Details

D.1 Data generation process

CIFAR-10H. We use the CIFAR-10 image as the non-privileged information $\boldsymbol{x}$. The annotator ID, the number of prior annotations the annotator has provided and the reaction time in milliseconds of the annotator, are used as privileged information $\boldsymbol{a}$. For feature pre-processing the annotator ID is one-hot encoded. The number of prior annotations and the reaction time are independently divided into 10 equally sized quantiles and the quantile ID is one-hot encoded. The image is pre-processed according the the standard MobileNet pre-processing (Howard et al., 2017).

As CIFAR-10H has on average more than 50 annotations per image and the labels are not particularly noisy. We subsample the CIFAR-10H labels by the following procedure. We keep all labels by the 41 annotators that agree with the true CIFAR-10 label less than 85% of the time. We then select a further 41 annotators from the remaining annotators. The average agreement of the bad annotators with the CIFAR-10 label is 63.3%, in the full subset of labels: 79.2% and in the full CIFAR-10H dataset: 94.9%. The subsampling procedure leaves 16,400 labels from 82 annotators while the full CIFAR-10H dataset has 514,200 labels from 2,571 annotators.

ImageNet. The annotator features are the model ID used to re-label $\boldsymbol{x}$, which is one-hot encoded and the probability of that label being sampled. See main paper for details on the sampling procedure and see Table 4 for the list of models used and their accuracy on the ImageNet training set. The pre-trained models are downloaded from tf.keras.applications[2].

[2]https://www.tensorflow.org/api_docs/python/tf/keras/applications

D.2 HYPERPARAMETERS

CIFAR-10H. For all methods $\phi(\boldsymbol{x})$ (or equivalent) is a MobileNet (Howard et al., 2017) pre-trained on ImageNet ILSVRC12, followed by a global average pooling layer and a Dense + ReLU layer with 64 units. $\psi(\boldsymbol{x}, \boldsymbol{a})$ is a two-layer MLP with 64 units per layer and ReLU activation. The first layer takes only $\boldsymbol{a}$ as an input, while the second layer takes the output of the first layer concatenated with $\phi(\boldsymbol{x})$ as input.

All models are trained for 20 epochs with the Adam optimizer with base learning rate$= 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-07$. All models are trained with L2 weight regularization with weighting $1e-3$.

Heteroscedastic models are trained using the method of Collier et al. (2021) with 4 factors for the low-rank covariance matrix approximation and a softmax temperature parameter of $\tau = 3.0$. Distilled models are also trained with a softmax temperature of $\tau = 3.0$ to smooth the teacher labels and with the distillation hyperparameter $\lambda = 0.5$ which weights the losses from the soft teacher labels and the true labels. A grid search over $\tau \in \{1.0, 2.0, 3.0, 4.0\}$ and $\lambda \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$ was conducted.

ImageNet. For all methods $\phi(\boldsymbol{x})$ (or equivalent) is a randomly initialized ResNet-50 (He et al., 2016) with the output layer removed. $\psi(\boldsymbol{x}, \boldsymbol{a})$ is a two-layer MLP with 128 units per layer and ReLU activation, the output of this MLP is concatenated with $\phi(\boldsymbol{x})$ and then passed to the output layer. The first layer of the $\psi(\boldsymbol{x}, \boldsymbol{a})$ MLP takes only $\boldsymbol{a}$ as an input, while the second layer takes the output of the first layer concatenated with $\phi(\boldsymbol{x})$ as input.

All but Het-TRAM models are trained for 90 epochs with the SGD optimizer with base learning rate$= 0.1$, decayed by a factor of 10 after 30, 60 and 80 epochs. Following Collier et al. (2021), Het-TRAM is trained for 270 epochs with the same initial learning rate and learning rate decay at 90, 180 and 240 epochs. All models are trained with L2 weight regularization with weighting $1e-4$.

Heteroscedastic models use 15 factors for the low-rank covariance matrix approximation and a softmax temperature parameter of $\tau = 1.5$. Distilled models are trained with a softmax temperature of $\tau = 3.0$ and with the distillation hyperparameter $\lambda = 0.5$. A grid search over $\tau \in \{1.0, 2.0, 3.0, 4.0\}$ and $\lambda \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$ was conducted.

E RISK ANALYSIS

Generative model and notations. We assume the following

- $\boldsymbol{a} \in \mathbb{R}^m, \boldsymbol{x} \in \mathbb{R}^d$,
- $\boldsymbol{a} \sim p(\boldsymbol{a}|\boldsymbol{x}) = \mathcal{N}(\mu(\boldsymbol{x})|\Sigma(\boldsymbol{x}))$ for some mean and covariance dependent on $\boldsymbol{x}$,
- $y = \boldsymbol{x}^\top \boldsymbol{w}^\star + \boldsymbol{a}^\top \boldsymbol{v}^\star + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

When considering n observations from this generative model, we use the matrix representations $\boldsymbol{y} \in \mathbb{R}^n, \boldsymbol{X} \in \mathbb{R}^{n \times d}$, $\boldsymbol{A} \in \mathbb{R}^{n \times m}$ and $\varepsilon \in \mathbb{R}^n$. We also write the zero-mean Gaussian vector

$$\boldsymbol{z} = (\boldsymbol{A} - \mu(\boldsymbol{X}))\boldsymbol{v}^\star + \varepsilon \in \mathbb{R}^n \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I} + \boldsymbol{\Lambda})$$

where we have defined the diagonal covariance

$$\boldsymbol{\Lambda} = \boldsymbol{\Lambda}(\boldsymbol{v}^\star, \boldsymbol{X}) = \mathrm{Diag}\big(\{(\boldsymbol{v}^\star)^\top \Sigma(\boldsymbol{x}_i)\boldsymbol{v}^\star\}_{i=1}^n\big) \in \mathbb{R}^{n \times n}.$$

We list below some notation that we will repeatedly use

- The orthogonal projector associated with $\boldsymbol{X}$:

$$\boldsymbol{\Pi}_x = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \in \mathbb{R}^{n \times n}.$$

- Similarly, the orthogonal projector associated with $\boldsymbol{A}$:

$$\boldsymbol{\Pi}_a = \boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \in \mathbb{R}^{n \times n}.$$

- The projections $\boldsymbol{X}_{a\perp} = (\boldsymbol{I} - \boldsymbol{\Pi}_a)\boldsymbol{X}$ and $\boldsymbol{A}_{x\perp} = (\boldsymbol{I} - \boldsymbol{\Pi}_x)\boldsymbol{A}$.
- The matrices: $\boldsymbol{H} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \in \mathbb{R}^{d\times n}$ and $\boldsymbol{G} = (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \in \mathbb{R}^{m\times n}$.
- The matrices above when restricted to the projections of $\boldsymbol{X}$ and $\boldsymbol{A}$ respectively, that is,
$$\boldsymbol{H}_{a\perp} = (\boldsymbol{X}_{a\perp}^\top \boldsymbol{X}_{a\perp})^{-1}\boldsymbol{X}_{a\perp}^\top \in \mathbb{R}^{d\times n} \text{ and } \boldsymbol{G}_{x\perp} = (\boldsymbol{A}_{x\perp}^\top \boldsymbol{A}_{x\perp})^{-1}\boldsymbol{A}_{x\perp}^\top \in \mathbb{R}^{m\times n}.$$

E.1 Definition of the risk

We will compare different estimators based on their different *risks*. We focus on the *fixed* design analysis (Bach, 2021), i.e., we study the errors only due to resampling the noise ε and the feature $\boldsymbol{a}$.

Given a predictor $\tau(\boldsymbol{X})$ based on the training quantities $(\boldsymbol{X}, \boldsymbol{A}, \varepsilon)$, we consider $\boldsymbol{y}' = \boldsymbol{X}\boldsymbol{w}^\star + \boldsymbol{A}'\boldsymbol{v}^\star + \varepsilon'$ (where the prime is to stress the difference with the training quantities without prime) and define the risk of τ as

$$\mathcal{R}(\tau(\boldsymbol{X}))) = \mathbb{E}_{\varepsilon'\sim p(\varepsilon'),\boldsymbol{a}'\sim p(\boldsymbol{a}'|\boldsymbol{x})}\left\{\frac{1}{n}\|\boldsymbol{y}' - \tau(\boldsymbol{X})\|^2\right\}. \tag{7}$$

Expanding the square with $\boldsymbol{y}' - \tau(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{w}^\star - \tau(\boldsymbol{X}) + \mu(\boldsymbol{X})\boldsymbol{v}^\star + \boldsymbol{z}'$, we obtain the expression

$$\mathcal{R}(\tau(\boldsymbol{X})) = \frac{1}{n}\|\boldsymbol{X}\boldsymbol{w}^\star - \tau(\boldsymbol{X}) + \mu(\boldsymbol{X})\boldsymbol{v}^\star\|^2 + \frac{1}{n}\mathrm{tr}(\sigma^2\boldsymbol{I} + \boldsymbol{\Lambda}). \tag{8}$$

Following common practices (Bach, 2021), to assess the risk, we finally take a second expectation $\mathbb{E}_{\varepsilon\sim p(\varepsilon),\boldsymbol{a}\sim p(\boldsymbol{a}|\boldsymbol{x})}[\mathcal{R}(\tau(\boldsymbol{X}))]$ with respect to the training quantities $(\boldsymbol{A}, \varepsilon)$.

Since we will mostly consider differences of risks, we omit the variance term $\frac{1}{n}\mathrm{tr}(\sigma^2\boldsymbol{I} + \boldsymbol{\Lambda})$ in the equations below.

E.2 Capturing the benefit of PI without marginalization

We first describe when, in absence of any marginalization, ordinary least squares ignoring PI is worse than ordinary least squares using PI with mean imputation at prediction time.

Proposition E.1. *Assume that $\boldsymbol{X}^\top\boldsymbol{X}$ is invertible. Moreover, assume that $\boldsymbol{A}^\top\boldsymbol{A}$ and $[\boldsymbol{X}, \boldsymbol{A}]^\top[\boldsymbol{X}, \boldsymbol{A}]$ are almost surely invertible. We have that*

$$\mathbb{E}[\mathcal{R}(\tau_{\text{NO-PI}}(\boldsymbol{X}))] > \mathbb{E}[\mathcal{R}(\tau_{\text{PI}}(\boldsymbol{X}))]$$

if and only if

$$\frac{1}{n}\|(\boldsymbol{I} - \boldsymbol{\Pi}_x)\mu(\boldsymbol{X})\boldsymbol{v}^\star\|^2 + \frac{\sigma^2 d}{n} + \frac{1}{n}tr(\boldsymbol{\Pi}_x\boldsymbol{\Lambda}) > \frac{\sigma^2}{n}\mathbb{E}[\|\boldsymbol{K}\|^2]$$

with $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{H}_{a\perp} + \mu(\boldsymbol{X})\boldsymbol{G}_{x\perp}$. When $m = 1$ (i.e., $\boldsymbol{A}$ is a column vector), a sufficient condition is

$$\frac{1}{n}\|(\boldsymbol{I} - \boldsymbol{\Pi}_x)\mu(\boldsymbol{X})\boldsymbol{v}^\star\|^2 + \frac{1}{n}tr(\boldsymbol{\Pi}_x\boldsymbol{\Lambda}) > 2\mathbb{E}\left[\frac{\|\boldsymbol{\Pi}_x\boldsymbol{A}\|^2 + \|\mu(\boldsymbol{X})\|^2}{\|(\boldsymbol{I} - \boldsymbol{\Pi}_x)\boldsymbol{A}\|^2}\right] + \frac{\sigma^2 d}{n}.$$

We provide the details of the derivation of the risk for $\tau_{\text{NO-PI}}$ and τ_{PI} in Section E.2.1 and Section E.2.2 respectively. Moreover, the second part of the proposition follows from an application of Lemma E.5.

E.2.1 Ordinary least squares (no marginalization)

The solution of

$$\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2$$

is given by $\hat{\boldsymbol{w}}_0 = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$. The corresponding predictions are

$$\tau_{\text{NO-PI}}(\boldsymbol{X}) = \boldsymbol{X}\hat{\boldsymbol{w}}_0 = \boldsymbol{\Pi}_x\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}^\star + \boldsymbol{\Pi}_x\mu(\boldsymbol{X})\boldsymbol{v}^\star + \boldsymbol{\Pi}_x\boldsymbol{z}.$$

Plugging into (8), we obtain

$$\mathcal{R}(\tau_{\text{NO-PI}}(\boldsymbol{X})) = \frac{1}{n}\|(\boldsymbol{I} - \boldsymbol{\Pi}_x)\mu(\boldsymbol{X})\boldsymbol{v}^\star - \boldsymbol{\Pi}_x\boldsymbol{z}\|^2.$$

Expanding the square and using that $\mathrm{tr}(\boldsymbol{\Pi}_x) = d$, the final risk expression is

$$\begin{aligned}\mathbb{E}[\mathcal{R}(\tau_{\text{NO-PI}}(\boldsymbol{X}))] &= \frac{1}{n}\|(\boldsymbol{I} - \boldsymbol{\Pi}_x)\mu(\boldsymbol{X})\boldsymbol{v}^\star\|^2 + \frac{1}{n}\mathbb{E}[\|\boldsymbol{\Pi}_x\boldsymbol{z}\|^2] \\ &= \frac{1}{n}\|(\boldsymbol{I} - \boldsymbol{\Pi}_x)\mu(\boldsymbol{X})\boldsymbol{v}^\star\|^2 + \frac{\sigma^2 d}{n} + \frac{1}{n}\mathrm{tr}(\boldsymbol{\Pi}_x\boldsymbol{\Lambda}).\end{aligned} \tag{9}$$

E.2.2 ORDINARY LEAST SQUARES WITH PI AND MEAN IMPUTATION (NO MARGINALIZATION)

We focus on the solution of

$$\min_{\boldsymbol{w},\boldsymbol{v}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} - \boldsymbol{A}\boldsymbol{v}\|^2$$

to construct an estimator. Using Lemma E.3, we have

$$\hat{\boldsymbol{w}}_1 = \boldsymbol{H}_{a\perp}\boldsymbol{y} \quad \text{and} \quad \hat{\boldsymbol{v}}_1 = \boldsymbol{G}_{x\perp}\boldsymbol{y}.$$

Using Lemma E.4, we can simplify

$$\hat{\boldsymbol{w}}_1 = \boldsymbol{H}_{a\perp}\boldsymbol{y} = \boldsymbol{w}^\star + \boldsymbol{0} + \boldsymbol{H}_{a\perp}\varepsilon$$

and

$$\hat{\boldsymbol{v}}_1 = \boldsymbol{G}_{x\perp}\boldsymbol{y} = \boldsymbol{0} + \boldsymbol{v}^\star + \boldsymbol{G}_{x\perp}\varepsilon.$$

Since $\boldsymbol{A}$ is not available at prediction time, we impute it instead with its mean $\mu(\boldsymbol{X})$, which is assumed to be perfectly known. This leads to

$$\tau_{\text{PI}}(\boldsymbol{X}) = \boldsymbol{X}\hat{\boldsymbol{w}}_1 + \mu(\boldsymbol{X})\hat{\boldsymbol{v}}_1 = \boldsymbol{X}\boldsymbol{w}^\star + \mu(\boldsymbol{X})\boldsymbol{v}^\star + \boldsymbol{K}\varepsilon,$$

with

$$\boldsymbol{K} = \boldsymbol{X}\boldsymbol{H}_{a\perp} + \mu(\boldsymbol{X})\boldsymbol{G}_{x\perp}.$$

Plugging into (8) and taking the expectation, we obtain

$$\begin{aligned}\mathbb{E}[\mathcal{R}(\tau_{\text{PI}}(\boldsymbol{X}))] &= \frac{1}{n}\|\boldsymbol{0}\|^2 + \frac{1}{n}\mathbb{E}[\|\boldsymbol{K}\varepsilon\|^2] \\ &= \frac{\sigma^2}{n}\mathbb{E}[\|\boldsymbol{K}\|^2]. \end{aligned} \tag{10}$$

E.3 CAPTURING THE BENEFIT OF PI WITH MARGINALIZATION

We then describe when, with marginalization, ordinary least squares ignoring PI is worse than ordinary least squares using PI.

Proposition E.2. *Assume that $\boldsymbol{X}^\top\boldsymbol{X}$ is invertible. Moreover, assume that $\boldsymbol{A}^\top\boldsymbol{A}$ and $[\boldsymbol{X},\boldsymbol{A}]^\top[\boldsymbol{X},\boldsymbol{A}]$ are almost surely invertible. We have that*

$$\mathbb{E}[\mathcal{R}(\tau_{\textit{marg. NO-PI}}(\boldsymbol{X}))] > \mathbb{E}[\mathcal{R}(\tau_{\textit{marg. PI}}(\boldsymbol{X}))]$$

if and only if

$$\frac{1}{n}\|(\boldsymbol{I} - \boldsymbol{\Pi}_x)\mu(\boldsymbol{X})\boldsymbol{v}^\star\|^2 + \frac{\sigma^2 d}{n} > \frac{\sigma^2}{n}\|\mathbb{E}[\boldsymbol{L}]\|^2$$

with $\boldsymbol{L} = \boldsymbol{X}\boldsymbol{H}_{a\perp} + \boldsymbol{A}\boldsymbol{G}_{x\perp}$. When $m = 1$ (i.e., $\boldsymbol{A}$ is a column vector), a sufficient condition is

$$\frac{1}{n}\|(\boldsymbol{I} - \boldsymbol{\Pi}_x)\mu(\boldsymbol{X})\boldsymbol{v}^\star\|^2 > 2\mathbb{E}\left[\frac{\|\boldsymbol{\Pi}_x\boldsymbol{A}\|^2 + \|\boldsymbol{A}\|^2}{\|(\boldsymbol{I} - \boldsymbol{\Pi}_x)\boldsymbol{A}\|^2}\right] + \frac{\sigma^2 d}{n}.$$

We provide the details of the derivation of the risk for $\tau_{\text{marg. NO-PI}}$ and $\tau_{\text{marg. PI}}$ in Section E.3.1 and Section E.3.2 respectively. Moreover, the second part of the proposition follows from an application of Lemma E.5.

E.3.1 ORDINARY LEAST SQUARES (WITH MARGINALIZATION)

Restarting from Section E.2.1, we consider the predictions marginalized with respect to $\boldsymbol{A}$. We have

$$\tau_{\text{marg. NO-PI}}(\boldsymbol{X}) = \mathbb{E}_{\boldsymbol{a}\sim p(\boldsymbol{a}|\boldsymbol{x})}[\boldsymbol{X}\hat{\boldsymbol{w}}_0] = \boldsymbol{X}\boldsymbol{w}^\star + \boldsymbol{\Pi}_x\mu(\boldsymbol{X})\boldsymbol{v}^\star + \boldsymbol{\Pi}_x\varepsilon.$$

Plugging into (8), we obtain

$$\mathcal{R}(\tau_{\text{marg. NO-PI}}(\boldsymbol{X})) = \frac{1}{n}\|(\boldsymbol{I} - \boldsymbol{\Pi}_x)\mu(\boldsymbol{X})\boldsymbol{v}^\star - \boldsymbol{\Pi}_x\varepsilon\|^2.$$

Expanding the square and using that $\text{tr}(\boldsymbol{\Pi}_x) = d$, the final risk expression is

$$\begin{aligned}\mathbb{E}[\mathcal{R}(\tau_{\text{marg. NO-PI}}(\boldsymbol{X}))] &= \frac{1}{n}\|(\boldsymbol{I} - \boldsymbol{\Pi}_x)\mu(\boldsymbol{X})\boldsymbol{v}^\star\|^2 + \frac{1}{n}\mathbb{E}[\|\boldsymbol{\Pi}_x\varepsilon\|^2] \\ &= \frac{1}{n}\|(\boldsymbol{I} - \boldsymbol{\Pi}_x)\mu(\boldsymbol{X})\boldsymbol{v}^\star\|^2 + \frac{\sigma^2 d}{n}. \end{aligned} \tag{11}$$

E.3.2 ORDINARY LEAST SQUARES WITH PI AND MARGINALIZATION

Restarting from Section E.2.2, we consider the predictions marginalized with respect to $\boldsymbol{A}$. In particular, we do not impute $\boldsymbol{A}$ by its mean but rather directly take the expectation over $\boldsymbol{A}$. We have

$$\tau_{\text{marg. PI}}(\boldsymbol{X}) = \mathbb{E}_{\boldsymbol{a}\sim p(\boldsymbol{a}|\boldsymbol{x})}[\boldsymbol{X}\hat{\boldsymbol{w}}_1 + \boldsymbol{A}\hat{\boldsymbol{v}}_1] = \boldsymbol{X}\boldsymbol{w}^\star + \mu(\boldsymbol{X})\boldsymbol{v}^\star + \mathbb{E}_{\boldsymbol{a}\sim p(\boldsymbol{a}|\boldsymbol{x})}[\boldsymbol{L}]\varepsilon,$$

with

$$\boldsymbol{L} = \boldsymbol{X}\boldsymbol{H}_{a\perp} + \boldsymbol{A}\boldsymbol{G}_{x\perp}.$$

Plugging into (8) and taking the expectation, we obtain

$$\begin{aligned}\mathbb{E}[\mathcal{R}(\tau_{\text{marg. PI}}(\boldsymbol{X}))] &= \frac{1}{n}\|\boldsymbol{0}\|^2 + \frac{1}{n}\mathbb{E}[\|\mathbb{E}_{\boldsymbol{a}\sim p(\boldsymbol{a}|\boldsymbol{x})}[\boldsymbol{L}]\varepsilon\|^2] \\ &= \frac{\sigma^2}{n}\|\mathbb{E}_{\boldsymbol{a}\sim p(\boldsymbol{a}|\boldsymbol{x})}[\boldsymbol{L}]\|^2. \end{aligned} \tag{12}$$

E.4 TECHNICAL LEMMAS

Lemma E.3. *Assume that both $\boldsymbol{X}^\top\boldsymbol{X}$ and $\boldsymbol{A}^\top\boldsymbol{A}$ are invertible. Moreover, assume that both $\boldsymbol{X}_{a\perp}^\top\boldsymbol{X}_{a\perp}$ and $\boldsymbol{A}_{x\perp}^\top\boldsymbol{A}_{x\perp}$ are invertible.*

We can write the solution of

$$\min_{\boldsymbol{w},\boldsymbol{v}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} - \boldsymbol{A}\boldsymbol{v}\|^2$$

as

$$\hat{\boldsymbol{w}} = \boldsymbol{H}_{a\perp}\boldsymbol{y} \quad \textit{and} \quad \hat{\boldsymbol{v}} = \boldsymbol{G}_{x\perp}\boldsymbol{y}.$$

Proof. The proof follows by applying inversion formula for the block matrix

$$\boldsymbol{Q} = \begin{bmatrix}\boldsymbol{X}^\top\boldsymbol{X} & \boldsymbol{X}^\top\boldsymbol{A} \\ \boldsymbol{A}^\top\boldsymbol{X} & \boldsymbol{A}^\top\boldsymbol{A}\end{bmatrix}$$

where $\boldsymbol{X}_{a\perp}^\top\boldsymbol{X}_{a\perp}$ and $\boldsymbol{A}_{x\perp}^\top\boldsymbol{A}_{x\perp}$ are the two Schur complements of $\boldsymbol{X}^\top\boldsymbol{X}$ and $\boldsymbol{A}^\top\boldsymbol{A}$. Under the assumptions of the lemma, the matrix is $\boldsymbol{Q}$ is invertible. □

Lemma E.4. *We have the following properties*

- $\boldsymbol{H}_{a\perp}\boldsymbol{X} = (\boldsymbol{X}_{a\perp}^\top\boldsymbol{X}_{a\perp})^{-1}\boldsymbol{X}^\top(\boldsymbol{I} - \boldsymbol{\Pi}_a)\boldsymbol{X} = (\boldsymbol{X}_{a\perp}^\top\boldsymbol{X}_{a\perp})^{-1}(\boldsymbol{X}_{a\perp}^\top\boldsymbol{X}_{a\perp}) = \boldsymbol{I}$,
- $\boldsymbol{H}_{a\perp}\boldsymbol{A} = (\boldsymbol{X}_{a\perp}^\top\boldsymbol{X}_{a\perp})^{-1}\boldsymbol{X}^\top(\boldsymbol{I} - \boldsymbol{\Pi}_a)\boldsymbol{A} = \boldsymbol{0}$.

Conversely, we have

- $\boldsymbol{G}_{x\perp}\boldsymbol{A} = (\boldsymbol{A}_{x\perp}^\top\boldsymbol{A}_{x\perp})^{-1}\boldsymbol{A}^\top(\boldsymbol{I} - \boldsymbol{\Pi}_x)\boldsymbol{A} = (\boldsymbol{A}_{x\perp}^\top\boldsymbol{A}_{x\perp})^{-1}(\boldsymbol{A}_{x\perp}^\top\boldsymbol{A}_{x\perp}) = \boldsymbol{I}$,
- $\boldsymbol{G}_{x\perp}\boldsymbol{X} = (\boldsymbol{A}_{x\perp}^\top\boldsymbol{A}_{x\perp})^{-1}\boldsymbol{A}^\top(\boldsymbol{I} - \boldsymbol{\Pi}_x)\boldsymbol{X} = \boldsymbol{0}$.

Lemma E.5. *Assume $m = 1$, i.e., $\boldsymbol{A}$ is a column vector. We have*

$$\mathbb{E}[\|\boldsymbol{K}\|^2] \leq 2d + 2\mathbb{E}\left[\frac{\|\boldsymbol{\Pi}_x\boldsymbol{A}\|^2 + \|\mu(\boldsymbol{X})\|^2}{\|(\boldsymbol{I} - \boldsymbol{\Pi}_x)\boldsymbol{A}\|^2}\right].$$

Similarly, it holds that

$$\|\mathbb{E}[\boldsymbol{L}]\|^2 \leq 2d + 2\mathbb{E}\left[\frac{\|\boldsymbol{\Pi}_x\boldsymbol{A}\|^2 + \|\boldsymbol{A}\|^2}{\|(\boldsymbol{I} - \boldsymbol{\Pi}_x)\boldsymbol{A}\|^2}\right].$$

Proof. We start by splitting the term into

$$\|\boldsymbol{K}\|^2 \leq 2\|\boldsymbol{X}\boldsymbol{H}_{a\perp}\|^2 + 2\|\mu(\boldsymbol{X})\boldsymbol{G}_{x\perp}\|^2.$$

Notice that $\boldsymbol{H}_{a\perp}\boldsymbol{H}_{a\perp}^\top = (\boldsymbol{X}_{a\perp}^\top\boldsymbol{X}_{a\perp})^{-1}$ and similarly $\boldsymbol{G}_{x\perp}\boldsymbol{G}_{x\perp}^\top = (\boldsymbol{A}_{x\perp}^\top\boldsymbol{A}_{x\perp})^{-1}$.

Since $\|\boldsymbol{M}\|^2 = \text{tr}(\boldsymbol{M}^\top\boldsymbol{M})$, we have

$$\|\boldsymbol{K}\|^2 \leq 2\text{tr}((\boldsymbol{X}^\top\boldsymbol{X})(\boldsymbol{X}_{a\perp}^\top\boldsymbol{X}_{a\perp})^{-1}) + 2\text{tr}(\mu(\boldsymbol{X})^\top\mu(\boldsymbol{X})(\boldsymbol{A}_{x\perp}^\top\boldsymbol{A}_{x\perp})^{-1}).$$

By definition of $\boldsymbol{A}_{x\perp}$, when $m = 1$, we have

$$(\boldsymbol{A}_{x\perp}^\top \boldsymbol{A}_{x\perp})^{-1} = \frac{1}{\|(\boldsymbol{I} - \boldsymbol{\Pi}_x)\boldsymbol{A}\|^2}.$$

For the term $(\boldsymbol{X}_{a\perp}^\top \boldsymbol{X}_{a\perp})^{-1}$, the Sherman–Morrison formula leads to

$$(\boldsymbol{X}_{a\perp}^\top \boldsymbol{X}_{a\perp})^{-1} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} + \frac{1}{1 - \boldsymbol{b}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{b}} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{b}\boldsymbol{b}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$$

with $\boldsymbol{b} = 1/\|\boldsymbol{A}\| \cdot \boldsymbol{X}^\top \boldsymbol{A} \in \mathbb{R}^d$. Further simplifying, we obtain

$$\operatorname{tr}((\boldsymbol{X}^\top \boldsymbol{X})(\boldsymbol{X}_{a\perp}^\top \boldsymbol{X}_{a\perp})^{-1}) = \operatorname{tr}\Big(\boldsymbol{I} + \frac{\boldsymbol{\Pi}_x \boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{\Pi}_x}{\|\boldsymbol{A}\|^2 - \|\boldsymbol{\Pi}_x \boldsymbol{A}\|^2}\Big) = d + \frac{\|\boldsymbol{\Pi}_x \boldsymbol{A}\|^2}{\|(\boldsymbol{I} - \boldsymbol{\Pi}_x)\boldsymbol{A}\|^2}.$$

For the second part of the proof, we start by applying Jensen inequality:

$$\|\mathbb{E}[\boldsymbol{L}]\|^2 \leq \mathbb{E}[\|\boldsymbol{L}\|^2].$$

The rest of the proof follows along the same arguments, replacing $\mu(\boldsymbol{X})$ by $\boldsymbol{A}$. □

F Related work table

Table 5: Comparison to related work.

Method	$p(\boldsymbol{a}\|\boldsymbol{x})$ Required	Training	Test Cost	Weight Sharing	Approximate $p(y\|\boldsymbol{x})$
Imputation	×	1 model, 1 step	= NO PI	✓	×
Distillation (Lopez-Paz et al., 2015)	×	2 models, 2 steps	= NO PI	×	×
Het. Dropout (Lambert et al., 2018)	×	1 model, 1 step	= NO PI	✓	✓
MIML-FCN+ (Yang et al., 2017)	×	1 model, 1 step	= NO PI	×	×
Full marginalization	✓	1 model, 1 step	$\mathcal{O}(S *$ NO PI)	✓	✓
TRAM (Ours)	×	1 model, 1 step	= NO PI	✓	✓
Het-TRAM (Ours)	×	1 model, 1 step	= NO PI	✓	✓
Distilled-TRAM (Ours)	×	2 models, 2 steps	= NO PI	✓	✓

G Two-step TRAM: ImageNet scale representation learning experiment

We conduct experiment to test two things: 1) does the one-step TRAM procedure, introduced in §3.2, which is easier for practitioners to implement, approximate the two-step TRAM procedure well and 2) can the results of the toy represrntation learning experiment, §2.2, be replicated in a larger scale setting.

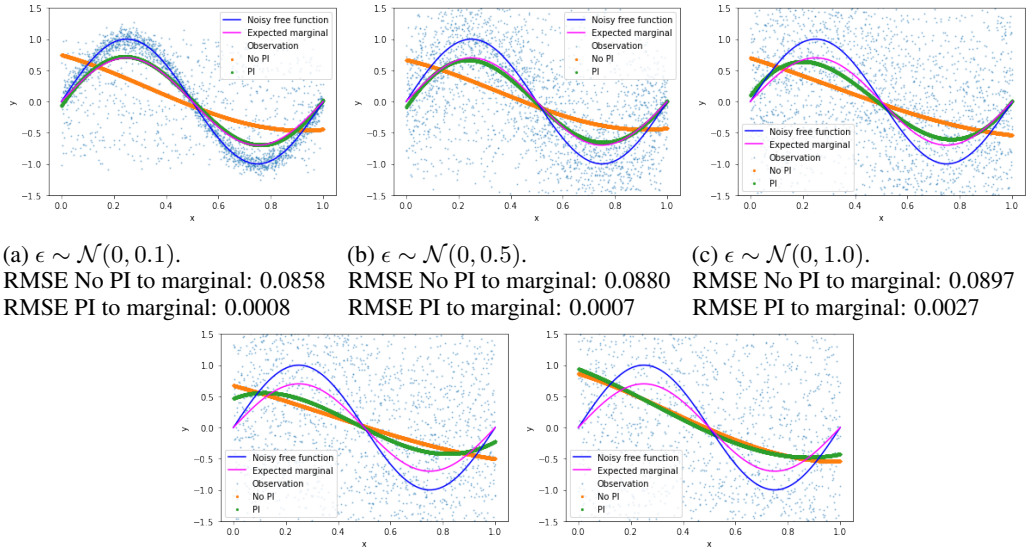
We train a feature extractor with and without access to PI on ImageNet, following the same procedure, architecture and dataset used in the main paper. We then freeze the feature extractor and train a single dense/linear layer with softmax activation on top of the fixed features. We then evaluate the efficacy of these features trained with and without PI using this "linear probe" evaluation widely used in the representation learning literature (Chen et al., 2020).

The results are presented in Table 6. We see that the simpler single-step TRAM method approximates the more complicated two-step TRAM method very well. In addition we see that the features learned by the network with access to PI which are then frozen and evaluated using a linear probe protocol perform better in terms of accuracy and log-likelihood.

H Toy experiment: vary ϵ

Table 6: Two-step TRAM: scaling up our toy representation learning experiment. ImageNet validation set negative log-likelihood and accuracy. Averaged over 10 training runs $\pm$ 1 std. dev.

METHOD	↓NLL	↑ACCURACY
ONE-STEP NO PI	1.264 ± 0.007	71.7 ± 0.2
TWO-STEP NO PI	1.265 ± 0.008	71.7 ± 0.3
ONE-STEP TRAM	1.225 ± 0.006	72.5 ± 0.2
TWO-STEP TRAM	1.226 ± 0.002	72.7 ± 0.2



(a) $\epsilon \sim \mathcal{N}(0, 0.1)$.
RMSE No PI to marginal: 0.0858
RMSE PI to marginal: 0.0008

(b) $\epsilon \sim \mathcal{N}(0, 0.5)$.
RMSE No PI to marginal: 0.0880
RMSE PI to marginal: 0.0007

(c) $\epsilon \sim \mathcal{N}(0, 1.0)$.
RMSE No PI to marginal: 0.0897
RMSE PI to marginal: 0.0027

(d) $\epsilon \sim \mathcal{N}(0, 1.5)$.
RMSE No PI to marginal: 0.0841
RMSE PI to marginal: 0.0472

(e) $\epsilon \sim \mathcal{N}(0, 2.0)$.
RMSE No PI to marginal: 0.0977
RMSE PI to marginal: 0.0977

Figure 4: Varying the influence of ϵ on our motivating toy experiment.

We vary the standard deviation of ϵ used in our motivating toy experiment, §2.2. The results can be seen graphically in Fig. 4. Fig. 4 also contains the average RMSE to the true marginal across the data points plotted. The graphical and numerical results demonstrate that even for large levels of noise PI aids with representation learning but as expected, as the level of noise grows the advantage of using PI diminishes as it becomes increasingly difficult to distinguish irreducible noise from noise which can be explained away with PI.

I IMAGENET EXPERIMENT PI ABLATION

We run an ablation, removing PI feature: the probability of the label assigned by the model from the PI set. We are thus left with just one PI feature, the one-hot encoded ID of the model that produced the label.

We see the results in Table 7. As expected (and predicted by our theoretical analysis), removing informative PI reduces the effectiveness of TRAM. Nonetheless, TRAM with the reduced PI feature set still outperforms the No PI baseline, with accuracy and log-likelihood lying between the No PI and full PI feature set TRAM methods.

Table 7: ImageNet ablation with reduced PI feature set. ImageNet validation set negative log-likelihood and accuracy. Averaged over 10 training runs $\pm$ 1 std. dev.

METHOD	↓NLL	↑ACCURACY
NO PI	1.264 ± 0.007	71.7 ± 0.2
TRAM WITH FULL PI SET	1.225 ± 0.006	72.5 ± 0.2
TRAM WITH REDUCED PI SET	1.246 ± 0.004	72.0 ± 0.2