A. Datasets

A.1. Osteoarthritis Initiative (OAI)

Description and statistics. The source of the knee x-ray dataset is the Osteoarthritis Initiative[5], which compiles radiological and clinical data on patients who have or are at high risk of developing knee osteoarthritis. We follow the dataset processing procedure used by Pierson et al. (2019) in their previous analysis. They analyzed data from the baseline visit and four follow-up timepoints (12-, 24-, 36-, and 48-month follow-ups). Two types of data from this dataset were used in our analysis: the knee x-rays, which served as the input to the neural network, and the clinical concepts associated with osteoarthritis, which were annotated by radiologists for each knee x-ray.

After filtering for observations which contain basic demographic and clinical data, the dataset contains 4,172 patients and 36,369 observations, where an observation is one knee for one patient at one timepoint. We randomly divided patients into training, validation, and test sets, with no overlap in the patient groups. Specifically, we have 21,340 observations from 2,456 people in the training set; 3,709 observations from 421 people in the validation set; and 11,320 observations from 1,295 people in the test set.

Image processing. To process the knee x-rays, each x-ray was downsampled to 512 x 512 pixels and normalized by dividing pixel values by the maximum pixel value (so all pixel values were in the range 0-1) and then z-scoring. Images were removed if they did not pass OAI x-ray image metadata quality control filters.

Clinical concept assessment and KLG merging. The primary clinical image feature used in analysis is Kellgren-Lawrence grade (KLG), a 5-level categorical variable (0 to 4) which is assessed by radiologists and used as a standard measure of radiographic osteoarthritis severity, with higher scores denoting more severe disease. In addition to KLG, each knee image is also assessed for 18 other clinical concepts (features) of osteoarthritis in various knee compartments, describing joint space narrowing (JSN), osteophytes, chondrocalcinosis, subchondral sclerosis, cysts, and attrition.

The Osteoarthritis Initiative only assessed these additional 18 clinical concepts (besides joint space narrowing, which is available for all participants) for participants with KLG ≥ 2 (a standard threshold for radiographic osteoarthritis) in at least one knee at any time point. Therefore, in their analysis (and in this paper), Pierson et al. (2019) set these clinical concepts to zero for other participants. This corresponds to assuming that that participants who were never assessed to have osteoarthritis, and thus were not assessed for other clinical concepts, did not display those features. This procedure also means it is impossible to use the clinical concepts to distinguish most x-rays with KLG $= 0$ from those with KLG $= 1$ in the dataset. To evaluate concept bottleneck models on this dataset, we therefore merged the KLG $= 0$ and KLG $= 1$ classes into a single level and translated the other KLG levels downwards by 1, leading to a 4-level categorical variable (0 to 3).

Concept processing. Some of the clinical concepts are very sparse, with almost all x-rays in the dataset showing an absence of the associated radiographic feature. We found that there were insufficient positive training examples to be able to accurately predict these concepts; moreover, including these sparse concepts in the bottleneck models lowered the accuracy of KLG prediction. We therefore filtered out the clinical concepts for which the dominant class (corresponding to an absence of the feature) represents $\geq 95\%$ of the training data.

This procedure kept 10 clinical concepts: "osteophytes femur medial", "sclerosis femur medial", "joint space narrowing medial", "osteophytes tibia medial", "sclerosis tibia medial", "osteophytes femur lateral", "sclerosis femur lateral", "joint space narrowing lateral", "osteophytes tibia lateral", and "sclerosis tibia lateral". It filtered 8 concepts: "cysts femur medial", "chondrocalcinosis medial", "cysts tibia medial", "attrition tibia medial", "cysts femur lateral", "chondrocalcinosis lateral", "cysts tibia lateral", "attrition tibia lateral".

After filtering, we z-scored the remaining clinical concepts using the training set to bring them onto the same scale.

Some of the clinical concepts, such as joint space narrowing, are annotated with fractional grades (e.g., 1.2, 1.4, 1.6 etc.) in the dataset. These partial grades represent temporal progression and cannot be deduced by looking at a single timepoint, and they explicitly do not reflect fractional grades (e.g., 1.2 on one patient does not mean it is worse than 1.0 on another patient); we therefore truncate these fractional grades.

Reader disagreements and adjudication procedures. KLG was read by two expert readers (i.e., radiologists) for each x-ray. Discrepancies in these readings, if they met the adjudication criteria described below, were adjudicated by a third reader: if the third reading agreed with either of the existing readings, then that reading was taken to be final, and otherwise,

[5] https://nda.nih.gov/oai/

the three readers attended an adjudication session to form a consensus reading. If discrepancies were not adjudicated, the final reading was taken to be the one from the more senior reader. KLG readings were adjudicated when they disagreed on whether KLG was within 0-1 or 2-4, or when they there was a difference in the direction of change of KLG between time points.

JSN was also read by two readers, with similar adjudication procedures. Discrepancies were adjudicated if the readers did not agree on the direction of change between time points.

All other clinical concepts in our dataset were read by a single reader. For more information on the adjudication procedures, please refer to the OAI documentation on Project 15.

A.2. Caltech-UCSD Birds-200-2011 (CUB)

Description and statistics. The Caltech-UCSD Birds-200-2011 (CUB) dataset (Wah et al., 2011) comprises 11,788 photographs of birds from 200 species, with each image additionally annotated with 312 binary concepts (before processing) corresponding to bird attributes like wing color, beak shape, etc. Visibility information on each concept is also provided for each image (e.g., is the beak visible in this image?); we use this information to make our test-time intervention experiments more realistic, but not at training time. Since the original dataset only has train and test sets, we randomly split 20% of the data from the official train set to make a validation set.

Concept processing. The individual concept annotations are noisy: each annotation was provided by a single crowdworker (not a birding expert), and the concepts can be quite similar to each other, e.g., some crowdworkers might indicate that birds from some species have a red belly, while others might say that the belly is rufous (reddish-brown) instead.

To deal with this issue, we aggregate instance-level concept annotations into class-level concepts via majority voting: e.g., if more than 50% of crows have black wings in the data, then we set all crows to have black wings. This makes the approximation that all birds of the same species in the training data should share the same concept annotations. While this approximation is mostly true for this dataset, there are some exceptions due to visual occlusion, as well as sexual and age dimorphism.

After majority voting, we further filter out concepts that are too sparse, keeping only concepts (binary attributes) that are present after majority voting in at least 10 classes. After this filtering, we are left with 112 concepts.

B. Experimental details

B.1. OAI model architecture and training

The models we use to predict KLG from knee x-rays follow the hyperparameters and model setup used by Pierson et al. (2019), except for the learning rate and learning rate schedule, which we tune separately. Our models use a ResNet-18 (He et al., 2016) pretrained on ImageNet, with the last 12 convolutional layers fine-tuned on the OAI dataset.

For the bottleneck models, the ResNet-18 network extracts high-level features from the image that is used to regress to the concepts c with a single fully-connected layer. Subsequently, there is a 3-layer MLP, with a dimensionality of 50 for the first two layers, that is used to regress to the final KLG y. The standard model is similar, except without any loss term that encourages the bottleneck layer to align with the concepts.

For fine-tuning, we use a batch size of 8, with random horizontal and vertical translations as data augmentation. Network weights are optimized with Adam, with beta parameters of 0.9 and 0.999 and an initial learning rate determined by grid search over [0.00005, 0.0005, 0.005], which decays by a factor of 2 every 10 epochs. The network is trained for 30 epochs with early stopping; model weights are set at the conclusion of training to those after the epoch with lowest RMSE for KLG on the validation set.

B.2. CUB model architecture and training

The main architecture for fine-grained bird classification is Inception V3, pretrained on ImageNet (except for the fully-connected layers) and then finetuned end-to-end on the CUB dataset. We follow the preprocessing practices described in Cui et al. (2018). Each image used for training is augmented with random color jittering, random horizontal flip and random cropping with a resolution of 299. During inference, the original image is center-cropped and resized to 299.

For each model, we hyperparameter search on the validation set over a range of learning rates ([0.001, 0.01]), learning rate schedules (keeping learning rate constant or reducing learning rate by 10 times after every [10, 15, 20] epochs until it reaches 0.0001), and regularization strengths ([0.0004, 0.00004]), to find a good hyperparameter configuration. The best model is decided based on task accuracy (or concept accuracy for the $x \to c$ part of sequential models) on the validation set. Once we have found the best-performing hyperparameter configuration, we then retrain the model on both the train and validation sets until convergence, following Cui et al. (2018).

All training is done with a batch size of 64, and SGD with momentum of 0.9 as the optimizer. For bottleneck models, we weight each concept's contribution to the overall concept loss equally (which is in turn determined by λ for joint bottleneck models). However, the binary cross-entropy loss used for each individual concept prediction task is weighted by the ratio of class imbalance for that individual concept (which is about 1 : 9 on average) and normalized accordingly. This encourages the model to learn to predict positive concept labels, which are more rare, instead of mostly predicting negative labels.

B.3. Test-time intervention

OAI. For OAI, we use the held-out validation set to determine an input-independent ordering for concept intervention. Specifically, we use the concept labels in the validation set to intervene separately on each concept, replacing a single value in our original concept predictions with that ground truth concept. We obtain the intervention ordering by sorting the concepts in descending order of the improvement in KLG accuracy gained from intervening separately on each concept.

CUB. For CUB, the concept groups are determined by having a common prefix in the list of concept names. For example, "has_back_pattern::solid", "has_back_pattern::spotted", "has_back_pattern::striped", "has_back_pattern::multi-colored" all describe the same group that concerns back-pattern. Since all models are retrained on both train and validation sets, as described above, we do not follow the OAI procedure of determining a fixed ordering. Instead, we randomly select concept groups to intervene on at test time, using the class-level labels for all concepts within that group to replace the predicted logits. To avoid intervening on concepts that are not even visible in the image, we use the concept visibility information that comes with the official CUB dataset: for all concepts that are not visible in a given test image, their corrected values are set to 0 regardless of what the corresponding class-level labels may be.

B.4. Data efficiency

For OAI, we subsampled the training and validation data uniformly at random. For CUB, to ensure that each of the 200 classes had similar numbers of examples, we subsampled the images from each class uniformly at random. To avoid the computational load of hyperparameter searching for each model and degree of subsampling, we adopted the hyperparameters chosen for the best-performing models on the full dataset but did early stopping on the subsampled validation datasets.

B.5. Linear probes

Standard (end-to-end) models. For OAI, we separately trained linear probes on the outputs after every ResNet block and the fully-connected layers of the MLP of the standard model. The best-performing linear probe was the one trained on the output of the final ResNet block. For CUB, we ran a linear probe on the fully-connected layer of the standard model, since the $c \to y$ part of the bottleneck models are linear.

SENN. To evaluate self-explaining neural networks (SENNs) (Melis & Jaakkola, 2018), we first trained a SENN model to predict KLG on the OAI dataset and then trained linear probes on the concept layer in the SENN model. We used the open-source implementation from the authors of SENN,[6] and therefore used a classification objective for KLG prediction. To match the expressiveness of our bottleneck models, we swapped the small CNNs of the SENN concept encoder and relevance parameterizer with our ResNet-18 models. Similarly, for the decoder network in SENN, we used a more expressive decoder comprising 2 fully-connected layers with batch normalization, followed by 5 transposed convolutional layers with upsampling. The decoder was obtained by adapting a public auto-encoder implementation,[7] changing the dimensionalities of the fully-connected and transposed convolutional layers, and increasing upsampling layers to match our input image size. We set the number of concepts for SENN to 10, corresponding to the number of clinical features in OAI. The learning rate was set to 0.0005 and the batch size was set to 4, which was the maximum possible given the memory constraints. With the above settings, the experiments were ran with two different seeds.

[6] https://github.com/dmelis/SENN

[7] https://github.com/arnaghosh/Auto-Encoder

C. Excess errors of independent vs. standard models

We present an analysis of the independent bottleneck model, which uses concepts at training time, versus the standard model, which does not. For simplicity, we consider a well-specified linear regression setting with normally-distributed inputs $X \in \mathbb{R}^d$, concepts $C \in \mathbb{R}^k$, and target $Y \in \mathbb{R}$:

$$X \sim N(0, \sigma_X^2 I_d) \tag{1}$$
$$C = XB + \epsilon_1, \tag{2}$$
$$Y = Cb + \epsilon_2, \tag{3}$$

where $\epsilon_1 \sim N(0, \sigma_C^2 I_k)$ and $\epsilon_2 \sim N(0, \sigma_Y^2)$. In contrast to the main text, we use capital letters for X, C, and Y here to emphasize the fact that the input, concepts, and target are random variables. In words, the input X is a normally distributed with dimension d; the concepts C of dimension k are a linear transformation of X with additive Gaussian noise; and the output Y is a scalar-valued linear transformation with additive Gaussian noise. For analytical simplicity, we require $\|b\|^2 = 1$ and $B^\top B = I_k$.

Independent bottleneck model. In this setting, the independent bottleneck model comprises two linear regression models: the first estimates the matrix B that takes $X \to C$, and the second estimates the vector b that takes $C \to Y$. For ease of analysis, we assume that each linear regression is fit using least squares on a separate dataset: the first dataset has n_1 training points in data matrices $\underline{X} \in \mathbb{R}^{n_1 \times d}$ and $\underline{C} \in \mathbb{R}^{n_1 \times k}$, and the second dataset has n_2 points in data matrices $\overline{C} \in \mathbb{R}^{n_2 \times k}$ and $\overline{Y} \in \mathbb{R}^{n_2}$. Concretely, we estimate

$$\hat{B} = \left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top \underline{C} \tag{4}$$
$$\hat{b} = \left(\overline{C}^\top \overline{C}\right)^{-1} \overline{C}^\top \overline{Y} \tag{5}$$

and then compose these estimators into the final prediction $\hat{Y}_{IB} = X\hat{B}\hat{b}$.

Standard model. In contrast, the standard model does not use concepts, and uses only one dataset with n points in $\underline{X} \in \mathbb{R}^{n \times d}$ and $\underline{Y} \in \mathbb{R}^n$. Concretely, we can express Y directly in terms of X as $Y = Xv + \epsilon$, where $v = Bb$ and $\epsilon \sim N(0, \sigma_C^2 + \sigma_Y^2)$. This gives the least squares estimate

$$\hat{v} = (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underline{Y} \tag{6}$$

and the resulting prediction $\hat{Y}_{SM} = X\hat{v}$.

Excess errors. We compare these two models using their asymptotic excess error as the number of training points $n_1 = n_2 = n$ goes to infinity, where a model's excess error is defined as how much higher its mean-squared-error is compared to the optimal estimator $\mathbb{E}[Y|X]$.

Proposition 1 (Relative excess error of independent bottleneck models vs. standard models in linear regression)**.** *Let $n_1 = n_2 = n$ tend to infinity. Then the ratio of excess errors of the independent bottleneck model to the standard model in the well-specified linear regression setting above is*

$$\lim_{n \to \infty} \frac{\mathbb{E}[(Y - \hat{Y}_{IB})^2] - \mathbb{E}[(Y - \mathbb{E}[Y|X])^2]}{\mathbb{E}[(Y - \hat{Y}_{SM})^2] - \mathbb{E}[(Y - \mathbb{E}[Y|X])^2]} \leq \frac{\frac{k}{d}\sigma_Y^2 + \sigma_C^2}{\sigma_Y^2 + \sigma_C^2}.$$

Note that asymptotic relative excess error is small—i.e., the independent bottleneck has lower excess error than the standard model—when $\frac{k}{d}$ is small and $\sigma_Y^2 \gg \sigma_C^2$. This corresponds to low dimensional concepts (relative to the input dimension) and concepts with low noise (relative to the noise in the output).

To prove this proposition, we first derive the expected errors of the independent bottleneck model and the standard model.

Lemma 1 (Risk of the independent bottleneck model)**.**

$$\mathbb{E}[(Y - \hat{Y}_{IB})^2] = \sigma_C^2 + \sigma_Y^2 + \sigma_Y^2 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_C^2} \frac{k}{n_2 - k - 1} + \sigma_C^2 \frac{d}{n_1 - d - 1} + \sigma_Y^2 \frac{1}{\sigma_X^2 + \sigma_C^2} \frac{1}{n_2 - k - 1} \sigma_C^2 \frac{d}{n_1 - d - 1}.$$

Proof. A direct calculation gives

$$\mathbb{E}[(Y - \hat{Y}_{IB})^2] = \mathbb{E}[((XBb + \epsilon_1 b + \epsilon_2) - X\hat{B}\hat{b})^2] \tag{7}$$
$$= \mathbb{E}[(\epsilon_1 b + \epsilon_2 + X(Bb - \hat{B}\hat{b}))^2] \tag{8}$$
$$= \mathbb{E}[(\epsilon_1 b + \epsilon_2)^2] + \mathbb{E}[X(Bb - \hat{B}\hat{b})(Bb - \hat{B}\hat{b})^\top X^\top] \tag{9}$$
$$= \sigma_C^2 + \sigma_Y^2 + \text{tr}\left(\mathbb{E}[X^\top X]\mathbb{E}[(Bb - \hat{B}\hat{b})(Bb - \hat{B}\hat{b})^\top]\right) \tag{10}$$
$$= \sigma_C^2 + \sigma_Y^2 + \sigma_X^2 \text{tr}\left(\mathbb{E}[(Bb - \hat{B}\hat{b})(Bb - \hat{B}\hat{b})^\top]\right), \tag{11}$$

where

$$(\hat{B}\hat{b} - Bb) = \left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top (\underline{X}B + \underline{\epsilon_1}) \left(\overline{C}^\top \overline{C}\right)^{-1} \overline{C}^\top (\overline{C}b + \overline{\epsilon_2}) - Bb \tag{12}$$
$$= \left(B + \left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top \underline{\epsilon_1}\right) \left(b + \left(\overline{C}^\top \overline{C}\right)^{-1} \overline{C}^\top \overline{\epsilon_2}\right) - Bb \tag{13}$$
$$= B\left(\overline{C}^\top \overline{C}\right)^{-1} \overline{C}^\top \overline{\epsilon_2} + \left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top \underline{\epsilon_1} b + \left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top \underline{\epsilon_1} \left(\overline{C}^\top \overline{C}\right)^{-1} \overline{C}^\top \overline{\epsilon_2}. \tag{14}$$

We need to evaluate the expectation of this expression multiplied with itself, $\mathbb{E}[(Bb - \hat{B}\hat{b})(Bb - \hat{B}\hat{b})^\top]$. Note that the cross terms will cancel since $\underline{\epsilon_1}$ and $\overline{\epsilon_2}$ are indepenent of other random variables and have mean 0, $\mathbb{E}[\underline{\epsilon_1}] = \mathbb{E}[\overline{\epsilon_2}] = 0$. This leaves three remaining direct (squared) terms, which we can evaluate separately since tr and $\mathbb{E}$ are linear operators.

The first term is

$$\text{tr}\left(\mathbb{E}\left[B\left(\overline{C}^\top \overline{C}\right)^{-1} \overline{C}^\top \overline{\epsilon_2 \epsilon_2}^\top \overline{C} \left(\overline{C}^\top \overline{C}\right)^{-1} B^\top\right]\right) \tag{15}$$
$$= \text{tr}\left(\mathbb{E}\left[\overline{C}^\top \left(\overline{C}^\top \overline{C}\right)^{-1} B^\top B \left(\overline{C}^\top \overline{C}\right)^{-1} \overline{C}^\top\right] \mathbb{E}\left[\overline{\epsilon_2 \epsilon_2}^\top\right]\right) \tag{16}$$
$$= \text{tr}\left(\mathbb{E}\left[\overline{C}^\top \left(\overline{C}^\top \overline{C}\right)^{-1} I_k \left(\overline{C}^\top \overline{C}\right)^{-1} \overline{C}^\top\right] \sigma_Y^2 I_{n_2}\right) \tag{17}$$
$$= \sigma_Y^2 \text{tr}\left(\mathbb{E}\left[\left(\overline{C}^\top \overline{C}\right)^{-1}\right]\right). \tag{18}$$

The expression within the above expectation is distributed as an inverse Wishart distribution, and therefore

$$\sigma_Y^2 \text{tr}\left(\mathbb{E}\left[\left(\overline{C}^\top \overline{C}\right)^{-1}\right]\right) \tag{19}$$
$$= \sigma_Y^2 \text{tr}\left(\frac{\mathbb{E}\left[C^\top C\right]^{-1}}{n_2 - k - 1}\right) \tag{20}$$
$$= \sigma_Y^2 \frac{1}{\sigma_X^2 + \sigma_C^2} \frac{k}{n_2 - k - 1}, \tag{21}$$

where the last equality comes from $\mathbb{E}[C^\top C] = \sigma_X^2 B^\top B + \sigma_C^2 I_k = (\sigma_X^2 + \sigma_C^2) I_k$.

The second term follows a similar calculation:

$$\text{tr}\left(\mathbb{E}\left[\left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top \underline{\epsilon_1} b b^\top \underline{\epsilon_1}^\top \underline{X} \left(\underline{X}^\top \underline{X}\right)^{-1}\right]\right) \tag{22}$$
$$= \text{tr}\left(\mathbb{E}\left[\underline{X}^\top \left(\underline{X}^\top \underline{X}\right)^{-1} \left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top\right] \mathbb{E}\left[\underline{\epsilon_1} b b^\top \underline{\epsilon_1}^\top\right]\right) \tag{23}$$
$$= \sigma_C^2 \text{tr}\left(\mathbb{E}\left[\left(\underline{X}^\top \underline{X}\right)^{-1}\right]\right) \tag{24}$$
$$= \sigma_C^2 \frac{1}{\sigma_X^2} \frac{d}{n_1 - d - 1}, \tag{25}$$

where the second equality follows because $\underline{\epsilon_1} b$ is normally distributed with mean 0 and covariance $\sigma_C^2 I_{n_1}$.

The third term is

$$\text{tr}\left(\mathbb{E}\left[\left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top \underline{\epsilon_1} \left(\overline{C}^\top \overline{C}\right)^{-1} \overline{C}^\top \overline{\epsilon_2 \epsilon_2}^\top \overline{C} \left(\overline{C}^\top \overline{C}\right)^{-1} \underline{\epsilon_1}^\top \underline{X} \left(\underline{X}^\top \underline{X}\right)^{-1}\right]\right) \tag{26}$$

$$= \text{tr}\left(\mathbb{E}\left[\overline{C} \left(\overline{C}^\top \overline{C}\right)^{-1} \underline{\epsilon_1}^\top \underline{X} \left(\underline{X}^\top \underline{X}\right)^{-1} \left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top \underline{\epsilon_1} \left(\overline{C}^\top \overline{C}\right)^{-1} \overline{C}^\top \overline{\epsilon_2 \epsilon_2}^\top\right]\right) \tag{27}$$

$$= \sigma_Y^2 \text{tr}\left(\mathbb{E}\left[\underline{\epsilon_1}^\top \underline{X} \left(\underline{X}^\top \underline{X}\right)^{-1} \left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top \underline{\epsilon_1} \left(\overline{C}^\top \overline{C}\right)^{-1}\right]\right) \tag{28}$$

$$= \sigma_Y^2 \frac{1}{\sigma_X^2 + \sigma_C^2} \frac{1}{n_2 - k - 1} \text{tr}\left(\mathbb{E}\left[\underline{\epsilon_1}^\top \underline{X} \left(\underline{X}^\top \underline{X}\right)^{-1} \left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top \underline{\epsilon_1}\right]\right) \tag{29}$$

$$= \sigma_Y^2 \frac{1}{\sigma_X^2 + \sigma_C^2} \frac{1}{n_2 - k - 1} \sigma_C^2 \text{tr}\left(\mathbb{E}\left[\underline{X} \left(\underline{X}^\top \underline{X}\right)^{-1} \left(\underline{X}^\top \underline{X}\right)^{-1} \underline{X}^\top\right]\right) \tag{30}$$

$$= \sigma_Y^2 \frac{1}{\sigma_X^2 + \sigma_C^2} \frac{1}{n_2 - k - 1} \sigma_C^2 \frac{1}{\sigma_X^2} \frac{d}{n_1 - d - 1}. \tag{31}$$

Putting the three terms together,

$$\text{tr}\left(\mathbb{E}[(\hat{B}\hat{b} - Bb)(\hat{B}\hat{b} - Bb)^\top]\right) \tag{32}$$

$$= \sigma_Y^2 \frac{1}{\sigma_X^2 + \sigma_C^2} \frac{k}{n_2 - k - 1} + \sigma_C^2 \frac{1}{\sigma_X^2} \frac{d}{n_1 - d - 1} + \sigma_Y^2 \frac{1}{\sigma_X^2 + \sigma_C^2} \frac{1}{n_2 - k - 1} \sigma_C^2 \frac{1}{\sigma_X^2} \frac{d}{n_1 - d - 1}, \tag{33}$$

so the expected squared error is

$$\mathbb{E}[(Y - \hat{Y}_{IB})^2] \tag{34}$$

$$= \sigma_C^2 + \sigma_Y^2 + \sigma_Y^2 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_C^2} \frac{k}{n_2 - k - 1} + \sigma_C^2 \frac{d}{n_1 - d - 1} + \sigma_Y^2 \frac{1}{\sigma_X^2 + \sigma_C^2} \frac{1}{n_2 - k - 1} \sigma_C^2 \frac{d}{n_1 - d - 1}. \tag{35}$$

□

Lemma 2 (Risk of the standard model)**.**

$$\mathbb{E}[(Y - \hat{Y}_{SM})^2] = \sigma_C^2 + \sigma_Y^2 + \frac{d(\sigma_C^2 + \sigma_Y^2)}{n - d - 1}.$$

Proof. A direct calculation gives

$$\mathbb{E}[(Y - \hat{Y}_{SM})^2] = \mathbb{E}[((Xv + \epsilon) - X\hat{v})^2] \tag{36}$$

$$= \mathbb{E}[(\epsilon + X(v - \hat{v}))^2] \tag{37}$$

$$= \mathbb{E}[\epsilon^2] + \mathbb{E}[X(v - \hat{v})(v - \hat{v})^\top X^\top] \tag{38}$$

$$= \sigma_C^2 + \sigma_Y^2 + \text{tr}\left(\mathbb{E}[X^\top X]\mathbb{E}[(v - \hat{v})(v - \hat{v})^\top]\right) \tag{39}$$

$$= \sigma_C^2 + \sigma_Y^2 + \sigma_X^2 \text{tr}\left(\mathbb{E}[(v - \hat{v})(v - \hat{v})^\top]\right). \tag{40}$$

Since

$$\hat{v} - v = (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top (\underline{X}v + \underline{\epsilon}) - v \tag{41}$$

$$= (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underline{\epsilon}, \tag{42}$$

we have

$$\text{tr}\left(\mathbb{E}[(v-\hat{v})(v-\hat{v})^\top]\right) = \text{tr}\left(\mathbb{E}[(\underline{X}^\top\underline{X})^{-1}\underline{X}^\top\underline{\epsilon}\underline{\epsilon}^\top\underline{X}(\underline{X}^\top\underline{X})^{-1}]\right) \tag{43}$$

$$= \text{tr}\left(\mathbb{E}[\underline{X}(\underline{X}^\top\underline{X})^{-1}(\underline{X}^\top\underline{X})^{-1}\underline{X}^\top]\mathbb{E}[\underline{\epsilon}\underline{\epsilon}^\top]\right) \tag{44}$$

$$= (\sigma_C^2+\sigma_Y^2)\text{tr}\left(\mathbb{E}[(\underline{X}^\top\underline{X})^{-1}]\right) \tag{45}$$

$$= (\sigma_C^2+\sigma_Y^2)\frac{1}{\sigma_X^2}\frac{d}{n-d-1}. \tag{46}$$

Plugging this back into the expression for $\mathbb{E}[(Y-\hat{Y}_{SM})^2]$ yields

$$\mathbb{E}[(Y-\hat{Y}_{SM})^2] = \sigma_C^2+\sigma_Y^2+\frac{d(\sigma_C^2+\sigma_Y^2)}{n-d-1}. \tag{47}$$

□

Proof of Proposition 1. Note that the optimal estimator has risk

$$\mathbb{E}[(Y-\mathbb{E}[Y|X])^2] = \mathbb{E}[\epsilon^2] \tag{48}$$

$$= \sigma_C^2+\sigma_Y^2. \tag{49}$$

Thus, from Lemmas 1 and 2, the ratio of excess errors is

$$\frac{\mathbb{E}[(Y-\hat{Y}_{IB})^2]-\mathbb{E}[(Y-\mathbb{E}[Y|X])^2]}{\mathbb{E}[(Y-\hat{Y}_{SM})^2]-\mathbb{E}[(Y-\mathbb{E}[Y|X])^2]} \tag{50}$$

$$= \frac{n-d-1}{d(\sigma_C^2+\sigma_Y^2)}\left(\sigma_Y^2\frac{\sigma_X^2}{\sigma_X^2+\sigma_C^2}\frac{k}{n_2-k-1}+\sigma_C^2\frac{d}{n_1-d-1}+\sigma_Y^2\frac{1}{\sigma_X^2+\sigma_C^2}\frac{1}{n_2-k-1}\sigma_C^2\frac{d}{n_1-d-1}\right). \tag{51}$$

Taking the limit as n goes to infinity and letting $n_1=n_2=n$ gives the desired result

$$\lim_{n\to\infty}\frac{\mathbb{E}[(Y-\hat{Y}_{IB})^2]-\mathbb{E}[(Y-\mathbb{E}[Y|X])^2]}{\mathbb{E}[(Y-\hat{Y}_{SM})^2]-\mathbb{E}[(Y-\mathbb{E}[Y|X])^2]} = \frac{\sigma_Y^2}{\sigma_C^2+\sigma_Y^2}\frac{\sigma_X^2}{\sigma_X^2+\sigma_C^2}\frac{k}{d}+\frac{\sigma_C^2}{\sigma_C^2+\sigma_Y^2} \tag{52}$$

$$\leq \frac{\frac{k}{d}\sigma_Y^2+\sigma_C^2}{\sigma_Y^2+\sigma_C^2}. \tag{53}$$

□