ON WHAT WE CAN LEARN FROM LOW-RESOLUTION DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Artificial intelligence systems typically rely on large, centrally collected datasets, a premise that does not hold in many real-world domains such as healthcare and public institutions. In these settings, data sharing is often constrained by storage, privacy, or resource limitations. For example, small wearable devices may lack the bandwidth or energy capacity needed to store and transmit high-resolution data, leading to aggregation during data collection and thus a loss of information. As a result, datasets collected from different sources may consist of a mixture of high- and low-resolution samples. Despite the prevalence of this setting, it remains unclear how informative low-resolution data is when models are ultimately evaluated on high-resolution inputs. We provide a theoretical analysis based on the Kullback–Leibler divergence that characterises how the influence of a datapoint changes with resolution, and derive bounds that relate the relative contribution of high- and low-resolution observations to the information lost under downsampling. We validate these results in simulation and further show empirically, using both a vision transformer and a convolutional neural network, that low-resolution data can improve performance when high-resolution data is scarce, across controlled benchmarks and a naturally heterogeneous medical dataset.

1 INTRODUCTION

Large-scale artificial intelligence (AI) systems rely on massive datasets, substantial storage, computation, and data-transfer resources. In many real-world settings, however, such as healthcare systems, public institutions, and distributed data ecosystems, these resources are limited, and sharing raw data across institutions is often constrained by privacy, regulatory, security, or storage requirements (El Mestari et al., 2024; Zerka et al., 2020; Fang et al., 2024; Pati et al., 2024). This makes it difficult to assemble the large centralised datasets on which modern AI systems depend.

A representative example arises in wearable devices, such as smart watches and hearing aids, which continuously collect data in everyday use (Christensen et al., 2021; Pati et al., 2024; Wani & Can, 2025). Privacy constraints, limited bandwidth, and energy resources prevent storage and transmission of raw high-resolution data. Instead, the devices only transmit feature-based representations or temporally downsampled summaries (Pontoppidan et al., 2018; Zhao & Fan, 2024). For example, a wearable device may monitor movement or acoustic conditions continuously, but if only a coarse temporal summary is transmitted, short-lived changes and other fine-grained patterns in the original signal are no longer retained, yielding the mixed-resolution setting illustrated in Figure 1. As a result, datasets collected from multiple sources may contain a mixture of high- and low-resolution datapoints. Despite the prevalence of such settings, the question of how much task-relevant information is retained in low-resolution data remains largely unexplored. Existing work on low-resolution learning has primarily focused on super-resolution or on improving robustness and efficiency at inference time, rather than understanding the role of low-resolution data during training under storage and privacy constraints (Zangeneh et al., 2020; Ge et al., 2019).

In this work, we treat low-resolution observations not as a nuisance to be corrected, but as a deliberate design choice motivated by practical constraints. We consider a setting in which the training data comprise both high- and low-resolution datapoints. Each datapoint is assumed to be generated at a single resolution and is therefore only available in either high- or low-resolution forms. We investigate how much task-relevant information is preserved under downsampling, and to what extent

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

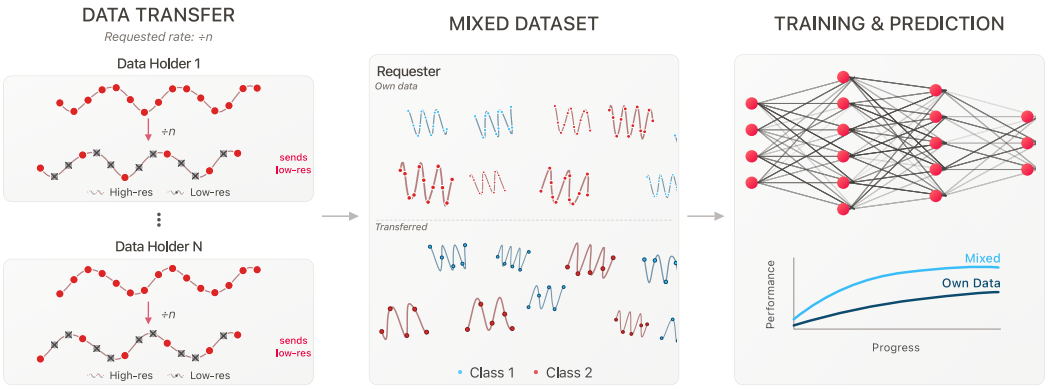

Figure 1: Illustration of the formation of a mixed-resolution dataset. A central party (the requester) possesses a limited amount of high-resolution data and requests additional data from multiple external data holders. Due to bandwidth and storage constraints, the exchanged data are provided at a lower resolution agreed upon by all parties. The resulting dataset consists of the requester's high-resolution samples combined with low-resolution data from external sources, on which model training is subsequently performed.

low-resolution data contribute to models evaluated on high-resolution inputs. This is carried out by a theoretical analysis studying how the contribution of a datapoint would differ if it were realised at a different resolution. We validate these results empirically through simulations. Complementing this analysis, we investigate the practical usefulness of low-resolution data empirically using two model architectures: a Vision Transformer (ViT) and a Convolutional Neural Network (CNN). While both architectures can accommodate inputs of varying spatial resolution, they do so in fundamentally different ways. ViTs accommodate changing input resolution through patch tokenisation and positional embedding (PE) interpolation, whereas CNNs require an adaptive pooling prior to classification. Comparing the two allows us to assess how architectural inductive biases interact with the information lost under downsampling.

Our main contributions are

1. We provide a theoretical characterisation of the relative influence of high- and low-resolution versions of a datapoint on the model parameter distribution, and derive bounds that relate this influence to the information missing from the lower-resolution representation.
2. Through experiments on standard benchmarks and on naturally heterogeneous medical data, we demonstrate that low-resolution data can improve performance on high-resolution evaluation tasks when high-resolution data is scarce.

2 Related Work

Prior methods for learning across resolutions typically either require access to samples at multiple resolutions, perform testing purely in low-resolution regimes, or focus on robustness to resolution shifts. In contrast, we study mixed-resolution training without paired observations, focusing on the performance on high-resolution inputs at test time.

Super-Resolution and Cross-Resolution Learning Super-resolution methods aim to reconstruct high-resolution images from low-resolution inputs (Saharia et al., 2022; Zamir et al., 2020; Wang et al., 2015; Hui et al., 2019) and are sometimes used as a preprocessing step to improve downstream performance (Koziarski & Cyganek, 2018; Wang et al., 2016). While effective in certain regimes, such approaches are misaligned with storage- and resource-constrained settings.

Cross-resolution learning frameworks typically rely on explicit alignment between high- and low-resolution representations. This is obtained either through the use of super-resolution pretraining in conjunction with cross-resolution feature transfer (Wang et al., 2016), employing a staged fine-to-coarse training pipeline (Peng et al., 2016), or adjusting crop-resolutions at test-time to fix the

discrepancies in resolution between train and test time (Touvron et al., 2019), all of which requires access to the original high-resolution data examples. Such assumptions are impractical when low-resolution data originates from external sources, and upscaling low-resolution samples to fit standard CNN architectures is storage-inefficient. While Koziarski & Cyganek (2018) show that accuracy degrades with decreasing resolution, they do not consider training across multiple resolutions.

Low-Resolution Recognition Low-resolution face recognition has been addressed via multidimensional scaling (Mudunuri & Biswas, 2016; Yang et al., 2018), discriminative subspace alignment with multiple CNNs (Lu et al., 2018; Zangeneh et al., 2020), simultaneous discriminant analysis (Chu et al., 2017), and knowledge distillation (Ge et al., 2019). While effective under cross-resolution mismatch, these methods focus on inference in low-resolution regimes rather than exploiting low-resolution data to improve high-resolution performance.

Learning from Compressed Representations A smaller body of work considers learning directly from compressed or downsampled inputs. RandNet (Chang et al., 2019) enables supervised learning from low-resolution data and connects to dictionary learning. Earlier theoretical work by Fabisch et al. (2013) shows that representing a network's weights using a lower-dimensional parameterisation is mathematically equivalent to projecting its inputs into a lower-dimensional space. However, these approaches do not address whether training on mixed resolutions improves generalisation to high-resolution inputs.

Multi-Resolution Training of Transformers Multi-resolution training has been explored in time-series transformers (Chen et al., 2024; Wang et al., 2024; Zhang et al., 2024), where varying patch sizes capture patterns at different temporal scales. In computer vision, ResFormer (Tian et al., 2023) enforces scale consistency by replicating batches across resolutions in conjunction with a scale consistency loss, ViTAR (Fan et al., 2024) introduces fuzzy PEs to improve robustness to resolution shifts, and NaViT (Dehghani et al., 2023) uses sequence packing during training to be able to handle different input resolutions. Although these methods improve cross-resolution robustness, they do not examine the benefit of incorporating unpaired low-resolution data during training.

CNNs and Vision Transformers CNNs (LeCun et al., 1989) process images through convolutional layers that apply learnable local filters across spatial neighbourhoods. This hierarchical structure enables the extraction of progressively more complex features, from edges and textures in early layers to high-level semantic representations in deeper layers. Weight sharing and translation equivariance provide strong inductive biases. In contrast, ViTs (Dosovitskiy et al., 2021) treat images as sequences of fixed-size patches processed by Transformer encoder layers with PEs. Self-attention enables global interactions across patches from the first layer onward, although ViTs lack the locality and equivariance biases of CNNs.

3 Problem Formulation

We study supervised learning in a setting where training data is not available at a single common resolution, while models are ultimately evaluated on high-resolution inputs only. Let $\mathcal{S}(\mathcal{D}) = \{S(d) : d \in \mathcal{D}\}$ be a stochastic process defined on a continuous domain $\mathcal{D} = [0, M]$, and let a datapoint $\boldsymbol{x}$ be a finite-dimensional representation of a realisation of the process $\mathcal{S}$. Let s and t denote two integer sampling rates where $s > t$. A high-resolution datapoint $\boldsymbol{x}_h \in \mathbb{R}^s$ is represented by s coefficients, while a low-resolution datapoint $\tilde{\boldsymbol{x}}_l \in \mathbb{R}^t$ is represented by t coefficients. For mathematical convenience and to compare the two in a common space, $\mathbb{R}^s$, we introduce an upsampling operator $f : \mathbb{R}^t \to \mathbb{R}^s$ and define $\boldsymbol{x}_l := f(\tilde{\boldsymbol{x}}_l) \in \mathbb{R}^s$. The operator preserves the alignment of the coefficients, such that, in this representation, $\boldsymbol{x}_h$ admits the decomposition: $\boldsymbol{x}_h = \boldsymbol{x}_l + \boldsymbol{x}_r$. Here $\boldsymbol{x}_r \in \mathbb{R}^s$ denotes the residual component, i.e. the information present in the high-resolution representation but absent in the low-resolution one. For simplicity, we consider processes on a one-dimensional domain, such that each datapoint is a vector. Higher-dimensional signals are handled by vectorising the coefficients.

Let $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ be a training dataset consisting of a mixture of high- and low-resolution datapoints, where each datapoint is observed at exactly one resolution. We consider a parametric model with parameters $\boldsymbol{\theta} \in \Theta$ trained using a loss function $\ell(\boldsymbol{\theta}, \boldsymbol{x})$. For notational simplicity, we

write $\ell(\boldsymbol{\theta}, \boldsymbol{x})$ as shorthand for $\ell(\boldsymbol{\theta}; \boldsymbol{x}, y)$, i.e. the target y is suppressed whenever it is fixed and unambiguous. All derivatives with respect to the input are taken with respect to the observation $\boldsymbol{x}$, not the target y. For a new datapoint, $\boldsymbol{x}_i$, we define its influence as the change in the parameter distribution induced by its inclusion:

$$\mathrm{KL}_i := \mathrm{KL}\big(p(\boldsymbol{\theta} \mid \mathcal{X}) \,\|\, p(\boldsymbol{\theta} \mid \mathcal{X} \cup \{\boldsymbol{x}_i\})\big), \tag{1}$$

where KL denotes the Kullback-Leibler divergence (Kullback & Leibler, 1951). In the following analysis, we assume $\mathrm{KL}_i > 0$ as well as treat $\boldsymbol{x}_h$ and $\boldsymbol{x}_l$ as fixed realisations of the same underlying datapoint at different resolutions.

Our central question is how the influence of a datapoint changes when it is observed at a different resolution. We characterise this difference by relating KL_h and KL_l to the residual component $\boldsymbol{x}_r$, which represents the information present in $\boldsymbol{x}_h$, but absent from $\boldsymbol{x}_l$. In contrast to prior work, we do not assume access to paired high- and low-resolution samples, nor do we require high-resolution reconstruction. Although our empirical evaluation focuses on classification, the theoretical framework permits any loss satisfying the stated regularity conditions. Dense prediction tasks, such as segmentation and localisation, require additional care because downsampling may also alter the spatial representation of the target. We therefore leave their empirical study to future work.

4 Theoretical Analysis

In this section, we theoretically analyse the effect of adding a high- versus a low-resolution datapoint to the training set. Following Sablayrolles et al. (2019) and Dong et al. (2022) we model the parameters $\boldsymbol{\theta}$ as a random variable whose distribution depends on the training set $\mathcal{X}$. Randomness arises from the training procedure in the form of stochastic gradients, random initialisation, and mini-batch sampling.

Assumption 1 (Gibbs Distribution) *Given a dataset $\mathcal{X}$ and a loss function $\ell(\boldsymbol{\theta}, \boldsymbol{x})$, the distribution of model parameters, $\boldsymbol{\theta}$, is given by*

$$p(\boldsymbol{\theta}|\mathcal{X}) = \frac{1}{Z(\mathcal{X})} \exp\left(-\frac{1}{\gamma} \sum_{\boldsymbol{x} \in \mathcal{X}} \ell(\boldsymbol{\theta}, \boldsymbol{x})\right), \tag{2}$$

where the partition function $Z(\mathcal{X}) = \int_\Theta \exp\left(-\frac{1}{\gamma} \sum_{\boldsymbol{x} \in \mathcal{X}} \ell(\boldsymbol{\theta}, \boldsymbol{x})\right) d\boldsymbol{\theta}$ is assumed to be finite.

Without loss of generality we set the temperature parameter $\gamma = 1$ for the remainder of the paper. To facilitate the analysis, we impose regularity conditions on the loss function.

Assumption 2 (Properties of the Loss Function) *For a loss function $\ell(\boldsymbol{\theta}, \boldsymbol{x})$ and constants $C_1, C_2, C_3 < \infty$*

- ***(A) Bounded Moments***

$$\mathbb{E}_{q(\boldsymbol{\theta})}\big[|\ell(\boldsymbol{\theta}, \boldsymbol{x})|\big] \le C_1, \quad \mathbb{E}_{q(\boldsymbol{\theta})}\big[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}))\big] \le C_2, \quad \mathbb{E}_{q(\boldsymbol{\theta})}\big[\ell(\boldsymbol{\theta}, \boldsymbol{x})^2\big] \le C_3$$
$$\text{for} \quad \boldsymbol{x} \in \{\boldsymbol{x}_h, \boldsymbol{x}_l\}, \qquad q(\boldsymbol{\theta}) \in \{p(\boldsymbol{\theta} \mid \mathcal{X}), p(\boldsymbol{\theta} \mid \mathcal{X} \cup \{\boldsymbol{x}_l\})\}.$$

- ***(B) Local Linearity*** *$\ell(\boldsymbol{\theta}, \boldsymbol{x})$ is twice continuously differentiable in $\boldsymbol{x}$ and locally linear (see Definition **A2** in Appendix A).*

We will now state the first result. The ratio of KL-divergences for including a point at either high- or low-resolution admits a closed-form expression as given by the following proposition (Proof in Appendix A.2).

Proposition 3 (Exact KL-divergence) *Under Assumption 1 the ratio of KL-divergences is given by:*

$$\frac{\mathrm{KL}_h}{\mathrm{KL}_l} = \frac{\log\left(\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\exp\left(-\ell(\boldsymbol{\theta}, \boldsymbol{x}_h)\right)]\right) + \mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_h)]}{\log\left(\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\exp\left(-\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)\right)]\right) + \mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)]}. \tag{3}$$

Proposition 3 characterises datapoint influence through the behaviour of the loss under the posterior distribution over parameters. A ratio above 1 means that the high-resolution point exerts a larger influence than its low-resolution counterpart, while a ratio of 1 means there is no gain in the extra information in the high-resolution representation.

Recognising the cumulant generating function (Kendall, 1943) in equation 3 reveals that datapoint influence is directly affected by higher-order moments. Approximating it to second order shows that the leading effect is governed by the variance of the loss as seen in the following proposition (proof in Appendix A.3).

Proposition 4 (Variance Approximation) *Under assumptions 1 and 2A, the ratio of KL-divergences admits the second-order approximation*

$$\frac{\mathrm{KL}_h}{\mathrm{KL}_l} \approx \frac{\mathrm{Var}_{\boldsymbol{\theta}|\mathcal{X}}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_h)]}{\mathrm{Var}_{\boldsymbol{\theta}|\mathcal{X}}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)]}. \tag{4}$$

This result provides an interpretable measure of datapoint influence. Intuitively, a datapoint must exert a larger influence on parameter distribution if the loss changes greatly across different realisations. The following proposition provides bounds of the ratio of KL-divergences, showing that the relative influence of a high-resolution datapoint is controlled by the residual component $\boldsymbol{x}_r$ (proof is found in Appendix A.4).

Proposition 5 (Ratio of KL-divergence) *Under assumptions 1 and 2, the ratio of KL-divergence satisfies the approximate bounds*

$$\left(1 - \sqrt{\frac{\|\boldsymbol{x}_r\|^2_{\Sigma_{\boldsymbol{g}}}}{\sigma_l^2}}\right)^2 \leq \frac{\mathrm{KL}_h}{\mathrm{KL}_l} \leq \left(1 + \sqrt{\frac{\|\boldsymbol{x}_r\|^2_{\Sigma_{\boldsymbol{g}}}}{\sigma_l^2}}\right)^2, \tag{5}$$

where $\sigma_l^2 = \mathrm{Var}_{\boldsymbol{\theta}|\mathcal{X}}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)], \boldsymbol{g} = \nabla_{\boldsymbol{x}} \ell(\boldsymbol{\theta}, \boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_l}$ *and* $\Sigma_{\boldsymbol{g}}$ *the covariance matrix of* $\boldsymbol{g}$ *under* $p(\boldsymbol{\theta} \mid \mathcal{X})$.

While Proposition 5 characterises the relative influence of high- and low-resolution representations, it is also useful to quantify their absolute difference. Proposition 6 provides bounds on the difference between the two KL-divergences, which measures the additional contribution of the high-resolution representation (proof in Appendix A.5).

Proposition 6 (Difference in KL-divergence) *Under assumptions 1 and 2, the difference in KL-divergences admits the approximate bounds,*

$$\frac{1}{2}\|\boldsymbol{x}_r\|^2_{\Sigma^l_{\boldsymbol{g}}} - \zeta\|\boldsymbol{x}_r\|_{\Sigma^l_{\boldsymbol{g}}} \leq \mathrm{KL}_h - \mathrm{KL}_l \leq \frac{1}{2}\|\boldsymbol{x}_r\|^2_{\Sigma^l_{\boldsymbol{g}}} + \zeta\|\boldsymbol{x}_r\|_{\Sigma^l_{\boldsymbol{g}}}, \tag{6}$$

where $\zeta = \sqrt{\mathrm{Var}_{\boldsymbol{\theta}|\mathcal{X}\cup\{\boldsymbol{x}_l\}}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))]}/\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}\cup\{\boldsymbol{x}_l\}}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))]$, $\boldsymbol{g} = \nabla_{\boldsymbol{x}} \ell(\boldsymbol{\theta}, \boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_l}$ *and* $\Sigma^l_{\boldsymbol{g}}$ *is the covariance matrix of* $\boldsymbol{g}$ *under* $p(\boldsymbol{\theta} \mid \mathcal{X} \cup \{\boldsymbol{x}_l\})$.

Taken together, propositions 5 and 6 characterise both the relative and absolute benefit of including a datapoint at higher resolution. The former shows that the relative gain depends on the directional residual norm $\|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}}$, normalised by σ_l^2. In particular, when σ_l^2 is large, the already big influence of the datapoint makes the extra coefficients in $\boldsymbol{x}_r$ less important. Proposition 6 complements this with a correction term governed by ζ. Thus the value of higher resolution depends both on the discarded information and on the sensitivity of the loss to that information.

The analysis does not impose any restriction on the resolutions present in $\mathcal{X}$, and does not depend on the particular choice of coefficients defining $\boldsymbol{x}_l$, provided alignment with $\boldsymbol{x}_h$ is preserved. $\boldsymbol{x}_l$ may therefore represent a partial feature observation with $\boldsymbol{x}_r$ being the missing features.

The bounds in Propositions 5 and 6 are obtained by controlling covariance terms via the Cauchy-Schwarz inequality (Gut, 2013), which yields worst-case bounds. If these covariance terms are known to be non-negative, the lower bounds can be improved. The resulting tighter bounds are stated in Corollary 7 with proofs in Appendix B.2.

Corollary 7 (Tighter Lower Bounds) *Under Assumptions 1 and 2, suppose additionally that* $\mathrm{Cov}_{\boldsymbol{\theta}|\mathcal{X}}\left[\boldsymbol{g}^\top \boldsymbol{x}_r,\, \ell(\boldsymbol{\theta}, \boldsymbol{x}_l)\right] \geq 0$ *and* $\mathrm{Cov}_{\boldsymbol{\theta}|\mathcal{X}\cup\{\boldsymbol{x}_l\}}\left[\boldsymbol{g}^\top \boldsymbol{x}_r,\, \exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))\right] \geq 0$*. Then propositions 5 and 6 admit the tighter lower bounds*

$$1 + \frac{\|\boldsymbol{x}_r\|^2_{\Sigma_{\boldsymbol{g}}}}{\sigma_l^2} \leq \frac{\mathrm{KL}_h}{\mathrm{KL}_l}, \qquad \tfrac{1}{2}\|\boldsymbol{x}_r\|^2_{\Sigma^l_{\boldsymbol{g}}} \leq \mathrm{KL}_h - \mathrm{KL}_l, \tag{7}$$

where $\sigma_l^2 = \mathrm{Var}_{\boldsymbol{\theta}|\mathcal{X}}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)]$*,* $\boldsymbol{g} = \nabla_{\boldsymbol{x}} \ell(\boldsymbol{\theta}, \boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_l}$*, and* $\Sigma_{\boldsymbol{g}}$ *and* $\Sigma^l_{\boldsymbol{g}}$ *are the covariance matrices of* $\boldsymbol{g}$ *under* $p(\boldsymbol{\theta} \mid \mathcal{X})$ *and* $p(\boldsymbol{\theta} \mid \mathcal{X} \cup \{\boldsymbol{x}_l\})$ *respectively.*

Under these non-negative covariance assumptions, the lower bound on the KL-ratio is strictly above 1, and the lower bound on the KL-difference is strictly positive whenever $\boldsymbol{x}_r \neq \mathbf{0}$. Thus, in this regime, the high-resolution datapoint is more influential than its low-resolution counterpart. The proofs of Propositions 5 and 6 rely on a first-order Taylor expansion. Appendix B derives corresponding bounds using a second-order expansion and shows through simulations that this refinement does not produce better bounds.

To illustrate the framework in a setting with exact KL-influence, we consider linear regression. Here the Gibbs distribution is Gaussian, and the resulting KL-influence is presented in Proposition 8 with the associated proof in Appendix B.3.

Proposition 8 (Closed-form KL-influence for Linear Regression) *For a linear regression model trained using the mean squared error loss function the model parameters, under Assumption 1, follow a normal distribution,* $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*. For a datapoint* $(\boldsymbol{x}_i, y_i)$*, the KL-influence admits the exact decomposition:*

$$\mathrm{KL}_i = -\frac{1}{2}\log(1 + 2s_i^2) - \frac{(y_i - m_i)^2}{2s_i^2 + 1} + (y_i - m_i)^2 + s_i^2, \tag{8}$$

where $m_i = \boldsymbol{x}_i^\top \boldsymbol{\mu}$ *and* $s_i^2 = \boldsymbol{x}_i^\top \boldsymbol{\Sigma} \boldsymbol{x}_i$*.*

The proposition identifies two factors controlling datapoint influence: the squared prediction residual $(y_i - m_i)^2$ and the posterior predictive uncertainty s_i^2. As $s_i^2 \to 0$, then $\mathrm{KL}_i \to 0$, whereas $\mathrm{KL}_i \to \infty$ as $s_i^2 \to \infty$. Thus a datapoint is generally more influential when the posterior is more uncertain about its prediction. This is consistent with the variance-based interpretation from Proposition 4, while providing an exact result that does not rely on a second-order truncation. For linear regression the first covariance condition of Corollary 7 can be analysed in closed form. It is violated when $\boldsymbol{x}_r$ acts sufficiently correctively of the low-resolution prediction or uncertainty. More details can be found in Appendix B.4.

4.1 Simulations of Theoretical Results

To validate the theoretical results, we conduct a series of simulations. For this, we need a dataset where a high-resolution datapoint fulfils the decomposition $\boldsymbol{x}_h = \boldsymbol{x}_l + \boldsymbol{x}_r$. To construct such a dataset, the first 100 images from each of the first two classes of CIFAR-10 are extracted. For each of the datapoints a scalogram using the Discrete Wavelet Transform (DWT) is computed with three levels using the `db2` mother wavelet (Mallat, 2009). To create different degrees of downsampled data, the bands in the scalogram containing the highest frequencies are progressively set to zero. The final datapoint, $\boldsymbol{x}_l$, is restored using the inverse DWT and projected to a 10-dimensional vector using PCA (Pearson, 1901). For more details on how this ensures that $\boldsymbol{x}_h = \boldsymbol{x}_l + \boldsymbol{x}_r$ we refer to Appendix C.1.

A neural network consisting of four linear layers with ReLU activations is then trained to distinguish between the two classes. We compute the true KL-divergences using the expression in Proposition 3, assess the approximation in Proposition 4, and compute the two bounds from propositions 5 and 6 as well as the tighter bounds in Corollary 7. Results are depicted in Figure 2 and generated by training $10,000$ models repeated over five random seeds. The black dashed line in the left panel denotes a fraction of 1, and in the right panel it denotes a difference of 0. The x-axis denotes the number of levels/frequency bands that have been removed where 0 is no removal. Exact KL-divergence values and their corresponding bounds can be found in Appendix C.2.

The variance approximation captures the overall behaviour of the KL-ratio. For both the KL-ratio and KL-difference, the lower and upper bounds become looser as the magnitude of the residual

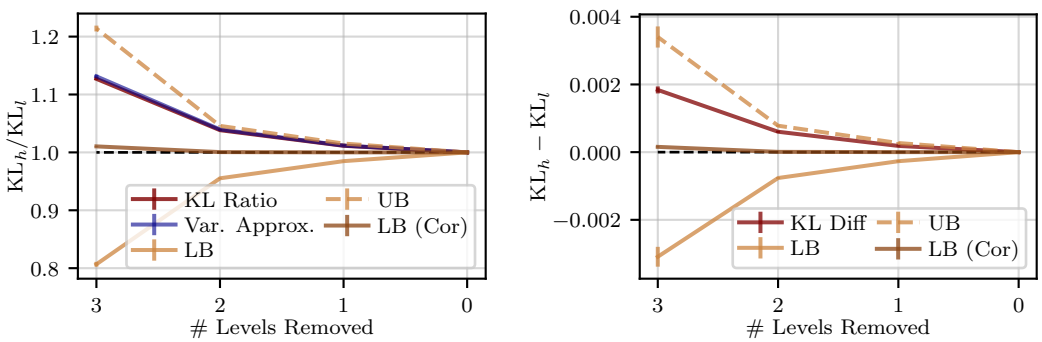

Figure 2: Simulation of bounds as a function of progressively removed high-frequency content. Error bars denote ± 5 times the standard deviation. As more high-frequency components are removed, the influence of low-resolution datapoints decreases, validating the theoretical predictions. UB = upper bound, LB = lower bound. Cor denotes bounds from Corollary 7. **Left**: True KL-divergence ratio, variance-based approximation from Proposition 4, the bounds from Proposition 5 and the tighter ratio lower bound from Corollary 7. **Right**: True KL-divergence difference, the bounds from Proposition 6, and the tighter difference lower bound from Corollary 7.

component increases. The tighter bounds introduced in Corollary 7 are indeed tighter which is consistent with the corresponding covariance terms being non-negative. Appendix C.3 shows that the variance approximation remains accurate across input dimensions and model sizes with relative error below 3% that decreases as model size grows. The appendix also shows how the bounds tighten with model size. Appendix C.4 further shows that high-resolution datapoints lying outside the support of the low-resolution distribution induce the largest parameter changes, illustrating how information discarded by downsampling governs datapoint influence.

5 Experimental Setup

We empirically evaluate whether adding low-resolution training data leads to measurable gains in practice. We do this in both controlled and naturally heterogeneous settings focusing on regimes where high-resolution training data is scarce. Models are evaluated exclusively on high-resolution inputs, and low-resolution samples are processed at the low resolution rather than being upsampled. The low-resolution inputs thus correspond to $\tilde{x}_l \in \mathbb{R}^t$ in Section 3 avoiding the storage and computational cost of embedding them in $\mathbb{R}^s$. Unless otherwise stated, results are averaged over five independent runs. Additional implementation and training details are provided in Appendix D.

Controlled Setting We first conduct experiments on CIFAR10 and CIFAR100 (Krizhevsky, 2009) to isolate the effect downsampling has on performance. This is done by both varying the proportion of high-resolution sampled as well as the resolution of the downsampled samples. This gives rise to four different types of experiments: training on a share of the high-resolution data (**Subset**); training on the same share but adding the remaining data in a low-resolution format (**Ratio**); training on a fully downsampled dataset (**Downsampled**); replacing 10% of the downsampled data with full-resolution samples (**Size**). CIFAR images have sidelength of $L = 32$. We use $L \in \{4, 8, 12, 16, 20\}$ in **Size** experiments and $L = 8$ in **Ratio** experiments. We evaluate a ViT and a modified ResNet18 with details given in Appendix D. Results on the one-dimensional AudioMNIST dataset are reported in Appendix E.

Heterogeneous Setting We model the setting from Figure 1 using the skin-lesion dataset Fed-ISIC-2019 (Tschandl et al., 2018; Codella et al., 2018; Hernández-Pérez et al., 2024; Ogier du Terrail et al., 2022). As the data is collected across six medical centres with substantial centre and label heterogeneity, this setting combines a resolution shift with a distribution shift. We select centre 1 as the requester and evaluate exclusively on its high-resolution test data. The requester then retains $\{10, 25, 50, 100\}\%$ of its local training data at 224×224 resolution while progressively receiving images from the remaining centres at a 56×56 resolution which we denote the **Grow** experiment. Donated samples are added beginning with minority classes. We report balanced accuracy due to the

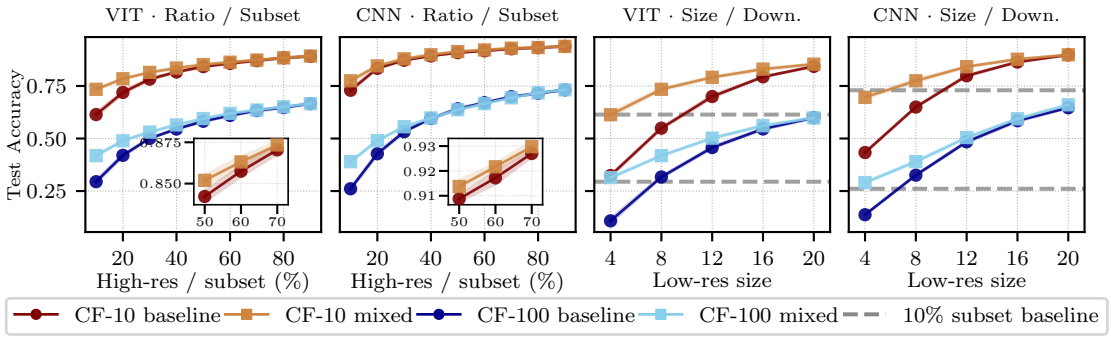

Figure 3: Effect of mixed-resolution training under different configurations. **Subset** is baseline for **Ratio**; both train on the same proportion of high-resolution samples and **Ratio** additionally includes the remaining samples in low-resolution format ($L = 8$). **Downsampled** is the baseline for **Size**, comparing a fully downsampled dataset against a mixed-resolution setting with 10% high-resolution and 90% low-resolution samples of varying sizes. The 10% subset baseline is shown in grey. CF denotes CIFAR and "Down." is short for Downsampled. Mixed-resolution training improves performance across datasets and architectures.

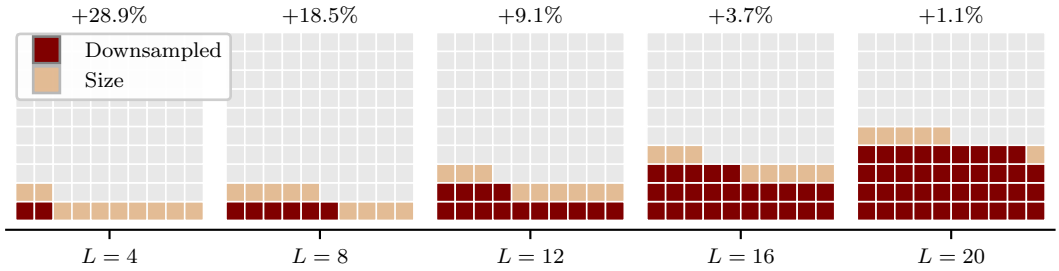

Figure 4: Trade-off between accuracy gain and storage cost for the **Size** experiment compared to the **Downsampled** on CIFAR-10 using the ViT. The 10×10 grid represents storage of the full dataset at full resolution, red cells correspond to storage of the **Downsampled** regime, and yellow cells to additional storage for the mixed-resolution (**Size**) regime. For $L = 12$, the downsampled dataset uses $12^2/32^2 = 0.14$ of full storage, while the mixed dataset uses $(12^2 \cdot 0.9 + 32^2 \cdot 0.1)/32^2 = 0.226$. Storage ratios are rounded to the nearest percentage point. The percentage above each panel denotes the relative improvement in classification accuracy achieved by the mixed approach.

strong class imbalance. Remaining details on the model architectures and optimisation strategies are reported in Appendix D.

6 Empirical Results

Controlled Setting Figure 3 depicts the results from the CIFAR datasets. The **Ratio** experiments are compared against **Subset**, while the **Size** experiments are compared against **Downsampled**. Across architectures and datasets, mixed-resolution training consistently improves performance, with gains diminishing as the proportion of high-resolution data increases. The ViT seems to benefit more for these datasets from the addition of low-resolution samples. One possible explanation is that CNNs rely more heavily on local, spatial structure, which can be disrupted by coarse downsampling. This indicates that low-resolution data can help boost performance when it originates from the same distribution. Appendix F contains different ablation studies including one that demonstrates that low-resolution data does provide a true learning signal, rather than just acting as a regulariser. While Figure 3 highlights the performance benefits of incorporating low-resolution data, it does not account for the associated storage costs. Figure 4 addresses this trade-off by visualising the additional storage required by the **Size** experiment compared with the **Downsampled** setting. The figure shows that even small amounts of high-resolution data can yield substantial improvements in accuracy.

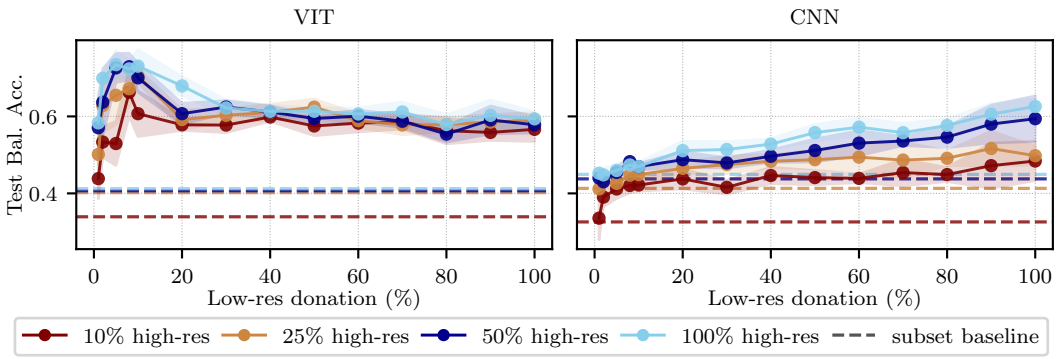

Figure 5: **Subset** and **Grow** experiments on the Fed-ISIC-2019 dataset. Each curve represents a different amount of high-resolution data from the requester, and the horizontal axis denotes the amount of low-resolution data that has been donated. Dotted lines denote corresponding high-resolution-only **Subset** baselines.

Taken together, figures 3 and 4 illustrate the performance-storage trade-offs of mixed-resolution training. The **Ratio** experiment shows that, when high-resolution data is scarce, collecting more low-resolution data can be more beneficial than acquiring a few additional high-resolution samples. Conversely, the **Size** experiment shows that selectively adding high-resolution samples can be more effective than uniformly increasing the resolution of the entire dataset.

Heterogeneous Setting Figure 5 compares high-resolution baselines with progressively adding low-resolution data from external sources. The CNN generally benefits from additional donations, particularly when requester data is scarce, showing that low-resolution data can remain useful under centre and label heterogeneity. For the ViT, small donations improve performance, but the gains saturate or reverse as more is added. To disentangle the roles of resolution and distribution shift, Appendix E repeats the experiment with donor data shared at full resolution, using the same donation order. In this setting, the decline largely disappears for larger requester subsets, and is weaker for the 10% setting. Together with the controlled experiments in Figure 3, where the ViT benefits from low-resolution data from the same distribution, this suggests that neither resolution- nor distribution shift alone accounts for the degradation, but that their combination does. The CNN appears more robust to this combination. Appendix F shows that these gains persist when donor data is additionally corrupted by blur or noise.

7 Conclusion

In this work, we studied mixed-resolution training, where only a subset of the training data is available at high resolution, while the remaining data is observed in a downsampled form and evaluation is performed on high-resolution inputs. We showed theoretically that the contribution of a high-resolution datapoint relative to its low-resolution counterpart is governed both by the information missing from the low-resolution representation and by the sensitivity of the loss to that missing information. Empirically, low-resolution data consistently improves performance in controlled experiments and remains useful under the natural centre and label heterogeneity of Fed-ISIC-2019. Overall our results demonstrate low-resolution data as a useful learning resource.

Limitations

Our bounds rely on a local linear approximation that is expected to be most accurate when high- and low-resolution representations remain sufficiently close. Our experiments only consider classification. While the theory is agnostic to the choice of loss function provided it satisfies the regularity conditions, segmentation and localisation tasks require additional consideration because downsampling may alter the spatial representation of the target.

AI USE STATEMENT

In this work, we used generative AI tools to get suggestions for theoretical derivations, for creating figures, and writing/debugging code. The latter includes the suggestion of experimental parameters. We further used generative AI to search for relevant literature, improve readability and suggest structure for the paper. All AI-generated code was verified by the main author.

We take responsibility for the final content of this work, including text and claims produced with the aid of generative AI.

ETHICS STATEMENT

There are no ethical considerations for this research project.

REPRODUCIBILITY STATEMENT

The code to generate all the results are provided in the supplementary material. The proofs for any theoretical results are included in Appendices. Appendix D also provides details about the training procedure and parameters used.

REFERENCES

Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. AudioMNIST: Exploring Explainable Artificial Intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*, 361(1):418–428, January 2024. ISSN 00160032. doi: 10.1016/j.jfranklin.2023.11.038. URL https://linkinghub.elsevier.com/retrieve/pii/S0016003223007536.

Thomas Chang, Bahareh Tolooshams, and Demba Ba. Randnet: Deep Learning with Compressed Measurements of Images. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, Pittsburgh, PA, USA, October 2019. IEEE. ISBN 978-1-7281-0824-7. doi: 10.1109/MLSP.2019.8918878. URL https://ieeexplore.ieee.org/document/8918878/.

Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale Transformers with Adaptive Pathways for Time Series Forecasting, September 2024. URL http://arxiv.org/abs/2402.05956. arXiv:2402.05956 [cs].

Jeppe Høy Christensen, Gabrielle H. Saunders, Lena Havtorn, and Niels H. Pontoppidan. Real-World Hearing Aid Usage Patterns and Smartphone Connectivity. *Frontiers in Digital Health*, 3:722186, August 2021. ISSN 2673-253X. doi: 10.3389/fdgth.2021.722186. URL https://www.frontiersin.org/articles/10.3389/fdgth.2021.722186/full.

Yongjie Chu, Touqeer Ahmad, George Bebis, and Lindu Zhao. Low-resolution face recognition with single sample per person. *Signal Processing*, 141:144–157, December 2017. ISSN 01651684. doi: 10.1016/j.sigpro.2017.05.012. URL https://linkinghub.elsevier.com/retrieve/pii/S0165168417301834.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 168–172. IEEE, 2018.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, Avital Oliver, Piotr Padlewski, Alexey Gritsenko, Mario Lucic, and Neil Houlsby. Patch n' pack: Navit, a vision transformer for any aspect ratio and resolution. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 2252–2274. Curran Associates, Inc., 2023. doi: 10.52202/

075280-0106. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/06ea400b9b7cfce6428ec27a371632eb-Paper-Conference.pdf.

Tian Dong, Bo Zhao, and Lingjuan Lyu. Privacy for free: How does dataset condensation help privacy? In *International Conference on Machine Learning*, pp. 5378–5396. PMLR, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Soumia Zohra El Mestari, Gabriele Lenzini, and Huseyin Demirci. Preserving data privacy in machine learning systems. *Computers & Security*, 137:103605, February 2024. ISSN 01674048. doi: 10.1016/j.cose.2023.103605. URL https://linkinghub.elsevier.com/retrieve/pii/S0167404823005151.

Alexander Fabisch, Yohannes Kassahun, Hendrik Wöhrle, and Frank Kirchner. Learning in compressed space. *Neural Networks*, 42:83–93, June 2013. ISSN 08936080. doi: 10.1016/j.neunet.2013.01.020. URL https://linkinghub.elsevier.com/retrieve/pii/S089360801300035X.

Qihang Fan, Quanzeng You, Xiaotian Han, Yongfei Liu, Yunzhe Tao, Huaibo Huang, Ran He, and Hongxia Yang. ViTAR: Vision Transformer with Any Resolution, March 2024. URL http://arxiv.org/abs/2403.18361. arXiv:2403.18361 [cs].

Congyu Fang, Adam Dziedzic, Lin Zhang, Laura Oliva, Amol Verma, Fahad Razak, Nicolas Papernot, and Bo Wang. Decentralised, collaborative, and privacy-preserving machine learning for multi-hospital data. *eBioMedicine*, 101:105006, March 2024. ISSN 23523964. doi: 10.1016/j.ebiom.2024.105006. URL https://linkinghub.elsevier.com/retrieve/pii/S2352396424000410.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936. doi: https://doi.org/10.1111/j.1469-1809.1936.tb02137.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x.

Pedro Galeano, Esdras Joseph, and Rosa E. Lillo. The Mahalanobis Distance for Functional Data With Applications to Classification. *Technometrics*, 57(2):281–291, April 2015. ISSN 0040-1706, 1537-2723. doi: 10.1080/00401706.2014.902774. URL http://www.tandfonline.com/doi/full/10.1080/00401706.2014.902774.

Shiming Ge, Shengwei Zhao, Chenyu Li, and Jia Li. Low-Resolution Face Recognition in the Wild via Selective Knowledge Distillation. *IEEE Transactions on Image Processing*, 28(4):2051–2062, April 2019. ISSN 1057-7149, 1941-0042. doi: 10.1109/TIP.2018.2883743. URL https://ieeexplore.ieee.org/document/8554301/.

Allan Gut. *Probability: A Graduate Course*, volume 75 of *Springer Texts in Statistics*. Springer New York, New York, NY, 2013. ISBN 978-1-4614-4707-8 978-1-4614-4708-5. doi: 10.1007/978-1-4614-4708-5. URL https://link.springer.com/10.1007/978-1-4614-4708-5.

Carlos Hernández-Pérez, Marc Combalia, Sebastian Podlipnik, Noel CF Codella, Veronica Rotemberg, Allan C Halpern, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Brian Helba, et al. Bcn20000: Dermoscopic lesions in the wild. *Scientific data*, 11(1):641, 2024.

Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight Image Super-Resolution with Information Multi-distillation Network. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 2024–2032, Nice France, October 2019. ACM. ISBN 978-1-4503-6889-6. doi: 10.1145/3343031.3351084. URL https://dl.acm.org/doi/10.1145/3343031.3351084.

Maurice George Kendall. *The advanced theory of statistics*. 1943.

Michał Koziarski and Bogusław Cyganek. Impact of Low Resolution on Image Recognition with Deep Neural Networks: An Experimental Study. *International Journal of Applied Mathematics and Computer Science*, 28(4):735–744, December 2018. ISSN 2083-8492. doi: 10.2478/amcs-2018-0056. URL https://www.sciendo.com/article/10.2478/amcs-2018-0056.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951. doi: 10.1214/aoms/1177729694. URL https://doi.org/10.1214/aoms/1177729694.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4): 541–551, December 1989. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco.1989.1.4.541. URL https://direct.mit.edu/neco/article/1/4/541-551/5515.

Ze Lu, Xudong Jiang, and Alex Kot. Deep Coupled ResNet for Low-Resolution Face Recognition. *IEEE Signal Processing Letters*, 25(4):526–530, April 2018. ISSN 1070-9908, 1558-2361. doi: 10.1109/LSP.2018.2810121. URL http://ieeexplore.ieee.org/document/8303213/.

Stéphane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 2009. ISBN 978-0-12-374370-1. doi: 10.1016/B978-0-12-374370-1.X0001-8.

Sivaram Prasad Mudunuri and Soma Biswas. Low Resolution Face Recognition Across Variations in Pose and Illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5): 1034–1040, May 2016. ISSN 0162-8828, 2160-9292. doi: 10.1109/TPAMI.2015.2469282. URL http://ieeexplore.ieee.org/document/7208853/.

Kevin P. Murphy. *Probabilistic Machine Learning: An Introduction*. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, MA, 2022.

Jean Ogier du Terrail, Samy-Safwan Ayed, Edwige Cyffers, Felix Grimberg, Chaoyang He, Regis Loeb, Paul Mangold, Tanguy Marchand, Othmane Marfoq, Erum Mushtaq, et al. Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. *Advances in Neural Information Processing Systems*, 35:5315–5334, 2022.

Sarthak Pati, Sourav Kumar, Amokh Varma, Brandon Edwards, Charles Lu, Liangqiong Qu, Justin J. Wang, Anantharaman Lakshminarayanan, Shih-han Wang, Micah J. Sheller, Ken Chang, Praveer Singh, Daniel L. Rubin, Jayashree Kalpathy-Cramer, and Spyridon Bakas. Privacy preservation for federated learning in health care. *Patterns*, 5(7):100974, July 2024. ISSN 26663899. doi: 10.1016/j.patter.2024.100974. URL https://linkinghub.elsevier.com/retrieve/pii/S2666389924000825.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

Xingchao Peng, Judy Hoffman, Stella X. Yu, and Kate Saenko. Fine-to-coarse knowledge transfer for low-res image classification. In *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3683–3687, 2016. doi: 10.1109/ICIP.2016.7533047.

Niels Henrik Pontoppidan, Xi Li, Lars Bramsløw, Benjamin Johansen, Claus Nielsen, Atefeh Hafez, and Michael Kai Petersen. Data-driven hearing care with time-stamped data-logging. *Proceedings of the International Symposium on Auditory and Audiological Research*, 6:127–134, February 2018. URL https://proceedings.isaar.eu/index.php/isaarproc/article/view/2017-15.

Alexandre Sablayrolles, Matthijs Douze, Yann Ollivier, Cordelia Schmid, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, 2019. URL https://api.semanticscholar.org/CorpusID:174799799.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image Super-Resolution Via Iterative Refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, 2022. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2022.3204461. URL https://ieeexplore.ieee.org/document/9887996/.

Rui Tian, Zuxuan Wu, Qi Dai, Han Hu, Yu Qiao, and Yu-Gang Jiang. ResFormer: Scaling ViTs with Multi-Resolution Training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22721–22731, Vancouver, BC, Canada, June 2023. IEEE. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.02176. URL https://ieeexplore.ieee.org/document/10204550/.

Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. *Advances in neural information processing systems*, 32, 2019.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 180161, 2018.

Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. *Advances in Neural Information Processing Systems*, 37:36314–36341, 2024.

Zhangyang Wang, Shiyu Chang, Yingzhen Yang, Ding Liu, and Thomas S. Huang. Studying Very Low Resolution Recognition Using Deep Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4792–4800, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.518. URL http://ieeexplore.ieee.org/document/7780887/.

Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang. Deep Networks for Image Super-Resolution with Sparse Prior. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 370–378, Santiago, Chile, December 2015. IEEE. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.50. URL http://ieeexplore.ieee.org/document/7410407/.

Rızwan Uz Zaman Wani and Ozgu Can. FED-EHR: A Privacy-Preserving Federated Learning Framework for Decentralized Healthcare Analytics. *Electronics*, 14(16):3261, August 2025. ISSN 2079-9292. doi: 10.3390/electronics14163261. URL https://www.mdpi.com/2079-9292/14/16/3261.

Fuwei Yang, Wenming Yang, Riqiang Gao, and Qingmin Liao. Discriminative Multidimensional Scaling for Low-Resolution Face Recognition. *IEEE Signal Processing Letters*, 25(3):388–392, March 2018. ISSN 1070-9908, 1558-2361. doi: 10.1109/LSP.2017.2746658. URL https://ieeexplore.ieee.org/document/8022987/.

Kentaro Yoshioka. vision-transformers-cifar10: Training Vision Transformers (ViT) and related models on CIFAR-10, 2024. URL https://github.com/kentaroy47/vision-transformers-cifar10.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning Enriched Features for Real Image Restoration and Enhancement. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, volume 12370, pp. 492–511. Springer International Publishing, Cham, 2020. ISBN 978-3-030-58594-5 978-3-030-58595-2. doi: 10.1007/978-3-030-58595-2_30. URL https://link.springer.com/10.1007/978-3-030-58595-2_30. Series Title: Lecture Notes in Computer Science.

Erfan Zangeneh, Mohammad Rahmati, and Yalda Mohsenzadeh. Low resolution face recognition using a two-branch deep convolutional neural network architecture. *Expert Systems with Applications*, 139:112854, January 2020. ISSN 09574174. doi: 10.1016/j.eswa.2019.112854. URL https://linkinghub.elsevier.com/retrieve/pii/S0957417419305561.

Fadila Zerka, Samir Barakat, Sean Walsh, Marta Bogowicz, Ralph T. H. Leijenaar, Arthur Jochems, Benjamin Miraglio, David Townend, and Philippe Lambin. Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care. *JCO Clinical*

Cancer Informatics, (4):184–200, November 2020. ISSN 2473-4276. doi: 10.1200/CCI.19.00047. URL https://ascopubs.org/doi/10.1200/CCI.19.00047.

Yitian Zhang, Liheng Ma, Soumyasundar Pal, Yingxue Zhang, and Mark Coates. Multi-resolution time-series transformer for long-term forecasting. In *International conference on artificial intelligence and statistics*, pp. 4222–4230. PMLR, 2024.

Feng Zhao and Song Fan. Protecting Infinite Data Streams from Wearable Devices with Local Differential Privacy Techniques. *Information*, 15(10):630, October 2024. ISSN 2078-2489. doi: 10.3390/info15100630. URL https://www.mdpi.com/2078-2489/15/10/630.

APPENDIX

A THEORETICAL RESULTS

We first introduce the definitions and identities used in the proofs.

A.1 PRELIMINARY

Additional definitions

(A1) The Kullback-Leibler divergence for two probability distributions p and q has the following definition (Kullback & Leibler, 1951):

$$\mathrm{KL}(p \mid\mid q) = \int p(\boldsymbol{\theta}) \log \left(\frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right) d\boldsymbol{\theta}. \tag{9}$$

(A2) A function, f, is locally linear if it can be written as:

$$f(\boldsymbol{x}) = f(\boldsymbol{a}) + \nabla^\top f(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a}) + h_1(\boldsymbol{x}) \left\| \boldsymbol{x} - \boldsymbol{a} \right\|, \quad \lim_{\boldsymbol{x} \to \boldsymbol{a}} h_1(\boldsymbol{x}) = 0. \tag{10}$$

In the following derivations we consider the h_1 term to be negligible.

(A3) The Cauchy–Schwarz inequality for variances for random variables X and Y is given by (Gut, 2013):

$$|\mathrm{Cov}[X, Y]| \leq \sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]} \iff \tag{11}$$

$$-\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]} \leq \mathrm{Cov}[X, Y] \leq \sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]} \tag{12}$$

(A4) The variance of a vector product for a vector $\boldsymbol{a}$ and $\boldsymbol{x}$ is $\mathrm{Var}[\boldsymbol{a}^\top \boldsymbol{x}] = \boldsymbol{x}^\top \Sigma_{\boldsymbol{a}} \boldsymbol{x}$, where $\Sigma_{\boldsymbol{a}}$ is the covariance matrix of $\boldsymbol{a}$:

$$\begin{aligned}
\mathrm{Var}[\boldsymbol{a}^\top \boldsymbol{x}] &= \mathbb{E}\left[\left(\boldsymbol{a}^\top \boldsymbol{x} - \mathbb{E}[\boldsymbol{a}^\top \boldsymbol{x}] \right)^2 \right] \\
&= \mathbb{E}\left[\left(\boldsymbol{a}^\top \boldsymbol{x} - \mathbb{E}[\boldsymbol{a}]^\top \boldsymbol{x} \right)^2 \right] \\
&= \mathbb{E}\left[\left((\boldsymbol{a}^\top - \mathbb{E}[\boldsymbol{a}]^\top) \boldsymbol{x} \right)^2 \right] \\
&= \mathbb{E}\left[\boldsymbol{x}^\top (\boldsymbol{a}^\top - \mathbb{E}[\boldsymbol{a}]^\top)^\top (\boldsymbol{a}^\top - \mathbb{E}[\boldsymbol{a}]^\top) \boldsymbol{x} \right] \\
&= \mathbb{E}\left[\boldsymbol{x}^\top (\boldsymbol{a} - \mathbb{E}[\boldsymbol{a}])(\boldsymbol{a} - \mathbb{E}[\boldsymbol{a}])^\top \boldsymbol{x} \right] \\
&= \boldsymbol{x}^\top \mathbb{E}\left[(\boldsymbol{a} - \mathbb{E}[\boldsymbol{a}])(\boldsymbol{a} - \mathbb{E}[\boldsymbol{a}])^\top \right] \boldsymbol{x} \\
&= \boldsymbol{x}^\top \Sigma_{\boldsymbol{a}} \boldsymbol{x}
\end{aligned}$$

(A5) Let $\|\boldsymbol{x}\|_A$ denote the Mahalanobis distance from the zero vector to a point $\boldsymbol{x}$ given by $\boldsymbol{x}^\top A \boldsymbol{x} = \|\boldsymbol{x}\|_A^2$ for a positive semi-definite matrix A (Galeano et al., 2015).

A datapoint consists of an observation $\boldsymbol{x}$ and a target y. The loss function is written in short as $\ell(\boldsymbol{\theta}, \boldsymbol{x}) = \ell(\boldsymbol{\theta}; \boldsymbol{x}, y)$ such that the target is implicit. For all derivatives it is implied they are taken with respect to the observation, $\boldsymbol{x}$ and not the target.

Throughout the theoretical analysis, the datapoints $\boldsymbol{x}_h$ and $\boldsymbol{x}_l$ are treated as fixed, and the only source of randomness are the model parameters $\boldsymbol{\theta}$.

A.2 Proof of Proposition 3

Let $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ be a dataset consisting of samples of different resolutions. We introduce $\mathcal{X}_h = \mathcal{X} \cup \{\boldsymbol{x}_h\}$ and $\mathcal{X}_l = \mathcal{X} \cup \{\boldsymbol{x}_l\}$, where $\boldsymbol{x}_h$ and $\boldsymbol{x}_l$ denote the same underlying datapoint observed at high and low resolution, respectively.

Under Assumption 1, the model parameters conditioned on the dataset $\mathcal{X}$ follow the distribution:

$$p(\boldsymbol{\theta} \mid \mathcal{X}) = \frac{1}{Z(\mathcal{X})} \exp\left(-\frac{1}{\gamma}\sum_{i=1}^{N} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i)\right), \tag{13}$$

where, without loss of generality, we set $\gamma = 1$. For notational convenience, define $Z := Z(\mathcal{X})$, $Z_h := Z(\mathcal{X} \cup \{\boldsymbol{x}_h\})$ and $Z_l := Z(\mathcal{X} \cup \{\boldsymbol{x}_l\})$. Then the distribution of the model parameters when adding $\boldsymbol{x}_h$ is,

$$p(\boldsymbol{\theta} \mid \mathcal{X}_h) = \frac{1}{Z_h} \exp\left(-\sum_{i=1}^{N} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i) - \ell(\boldsymbol{\theta}, \boldsymbol{x}_h)\right), \tag{14}$$

The distribution can likewise be found for adding $\boldsymbol{x}_l$. It follows that the ratio between the two distributions can be written as

$$\frac{p(\boldsymbol{\theta} \mid \mathcal{X})}{p(\boldsymbol{\theta} \mid \mathcal{X}_h)} = \frac{Z_h}{Z} \exp\left(\ell(\boldsymbol{\theta}, \boldsymbol{x}_h)\right). \tag{15}$$

The KL-divergence corresponding to the inclusion of $\boldsymbol{x}_h$

$$\mathrm{KL}_h = \mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}\left[\log\left(\frac{p(\boldsymbol{\theta} \mid \mathcal{X})}{p(\boldsymbol{\theta} \mid \mathcal{X}_h)}\right)\right], \tag{16}$$

which simplifies to

$$\mathrm{KL}_h = \log\left(\frac{Z_h}{Z}\right) + \mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_h)]. \tag{17}$$

An identical argument yields

$$\mathrm{KL}_l = \log\left(\frac{Z_l}{Z}\right) + \mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)]. \tag{18}$$

To identify the ratio Z_h/Z, similarly Z_l/Z, we note that the ratio between the two distributions can be written as equation 15. Rearranging and integrating over both sides yields

$$\int p(\boldsymbol{\theta} \mid \mathcal{X}_h)\, d\boldsymbol{\theta} = \frac{Z}{Z_h} \int p(\boldsymbol{\theta} \mid \mathcal{X}) \exp\left(-\ell(\boldsymbol{\theta}, \boldsymbol{x}_h)\right) d\boldsymbol{\theta}, \tag{19}$$

hence

$$\frac{Z_h}{Z} = \mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\exp\left(-\ell(\boldsymbol{\theta}, \boldsymbol{x}_h)\right)]. \tag{20}$$

Substituting these expressions into the formulas for KL_h and KL_l yields the desired result:

$$\frac{\mathrm{KL}_h}{\mathrm{KL}_l} = \frac{\log\left(\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\exp\left(-\ell(\boldsymbol{\theta}, \boldsymbol{x}_h)\right)]\right) + \mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_h)]}{\log\left(\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\exp\left(-\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)\right)]\right) + \mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)]}. \tag{21}$$

A.3 Proof of Proposition 4

The cumulant generating function of a random variable Y is given by (Kendall, 1943)

$$K_Y(b) := \log \mathbb{E}[e^{bY}] = \sum_{n=1}^{\infty} \kappa_n \frac{b^n}{n!}, \tag{22}$$

where κ_n denotes the nth cumulant of Y. Truncating this expansion at second order yields the approximation

$$K_Y(b) \approx \kappa_1 b + \kappa_2 \frac{b^2}{2} = \mathbb{E}[Y]b + \frac{b^2}{2}\mathrm{Var}[Y]. \tag{23}$$

For a fixed datapoint $\boldsymbol{x}$, define the random variable

$$Y_{\boldsymbol{x}} := \ell(\boldsymbol{\theta}, \boldsymbol{x}), \qquad \boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathcal{X}). \tag{24}$$

Since the expectation of a function, h, is $\mathbb{E}[h(x)] = \int h(x)p(x)\,dx$ then

$$\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[h(\boldsymbol{\theta})] = \int_{\Theta} h(\boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{X})\, d\boldsymbol{\theta}, \tag{25}$$

for any integrable function h. Setting $b = -1$ and $Y = Y_{\boldsymbol{x}}$ in equation 23 yields

$$K_{Y_x}(-1) \approx -\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\ell(\boldsymbol{\theta}, \boldsymbol{x})] + \frac{1}{2}\mathrm{Var}_{\boldsymbol{\theta}|\mathcal{X}}[\ell(\boldsymbol{\theta}, \boldsymbol{x})], \tag{26}$$

since the first two cumulants satisfy $\kappa_1 = \mathbb{E}[Y]$ and $\kappa_2 = \mathrm{Var}[Y]$. Applying this to $Y_{\boldsymbol{x}_l} = \ell(\boldsymbol{\theta}, \boldsymbol{x}_l)$ and $Y_{\boldsymbol{x}_h} = \ell(\boldsymbol{\theta}, \boldsymbol{x}_h)$ results in

$$\log \mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}\left[e^{-\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)}\right] \approx -\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)\right] + \frac{1}{2}\mathrm{Var}_{\boldsymbol{\theta}|\mathcal{X}}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)\right], \tag{27}$$

$$\log \mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}\left[e^{-\ell(\boldsymbol{\theta}, \boldsymbol{x}_h)}\right] \approx -\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x}_h)\right] + \frac{1}{2}\mathrm{Var}_{\boldsymbol{\theta}|\mathcal{X}}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x}_h)\right]. \tag{28}$$

Substituting these approximations into equation 21 yields

$$\frac{\mathrm{KL}_h}{\mathrm{KL}_l} \approx \frac{\mathrm{Var}_{\boldsymbol{\theta}|\mathcal{X}}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x}_h)\right]}{\mathrm{Var}_{\boldsymbol{\theta}|\mathcal{X}}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)\right]}, \tag{29}$$

which is what we wanted to show. In the case of $\gamma \neq 1$, equation 3 becomes

$$\frac{\mathrm{KL}_h}{\mathrm{KL}_l} = \frac{\log \mathbb{E}_{\theta|\mathcal{X}}[\exp(-\frac{1}{\gamma}\ell(\theta, x_h))] + \frac{1}{\gamma}\mathbb{E}_{\theta|\mathcal{X}}[\ell(\theta, x_h)]}{\log \mathbb{E}_{\theta|\mathcal{X}}[\exp(-\frac{1}{\gamma}\ell(\theta, x_l))] + \frac{1}{\gamma}\mathbb{E}_{\theta|\mathcal{X}}[\ell(\theta, x_l)]}, \tag{30}$$

in which case we can choose $b = -1/\gamma$ to obtain equation 29 again.

A.4 Proof of Proposition 5

Starting from the result from Proposition 4:

$$\frac{\mathrm{KL}_h}{\mathrm{KL}_l} \approx \frac{\mathrm{Var}_{\boldsymbol{\theta}|\mathcal{X}}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x}_h)\right]}{\mathrm{Var}_{\boldsymbol{\theta}|\mathcal{X}}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)\right]}, \tag{31}$$

we use (**A2**) on the loss function:

$$\ell(\boldsymbol{\theta}, \boldsymbol{x}_h) = \ell(\boldsymbol{\theta}, \boldsymbol{x}_l + \boldsymbol{x}_r) \approx \ell(\boldsymbol{\theta}, \boldsymbol{x}_l) + \nabla_{\boldsymbol{x}}^{\top}\ell(\boldsymbol{\theta}, \boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_l}\boldsymbol{x}_r, \tag{32}$$

where we introduce $\boldsymbol{g} = \nabla_{\boldsymbol{x}}\ell(\boldsymbol{\theta}, \boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_l}$. A rewrite of the numerator of equation 31 yields,

$$\mathrm{Var}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_h)] = \mathrm{Var}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)] + \mathrm{Var}[\boldsymbol{g}^{\top}\boldsymbol{x}_r] + 2\mathrm{Cov}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_l), \boldsymbol{g}^{\top}\boldsymbol{x}_r]. \tag{33}$$

Lower bound the covariance using (**A3**):

$$\mathrm{Var}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_h)] \geq \mathrm{Var}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)] + \mathrm{Var}[\boldsymbol{g}^{\top}\boldsymbol{x}_r] - 2\sqrt{\mathrm{Var}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)]\mathrm{Var}[\boldsymbol{g}^{\top}\boldsymbol{x}_r]}. \tag{34}$$

Introduce $\sigma_l^2 := \mathrm{Var}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)]$ and using (**A4**):

$$\frac{\mathrm{KL}_h}{\mathrm{KL}_l} \geq \frac{\sigma_l^2 + \boldsymbol{x}_r^{\top}\Sigma_{\boldsymbol{g}}\boldsymbol{x}_r - 2\sqrt{\sigma_l^2\boldsymbol{x}_r^{\top}\Sigma_{\boldsymbol{g}}\boldsymbol{x}_r}}{\sigma_l^2}, \tag{35}$$

where using (**A5**) gives us the following lower bound:

$$\frac{\mathrm{KL}_h}{\mathrm{KL}_l} \geq \left(1 - \sqrt{\frac{\|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}}^2}{\sigma_l^2}}\right)^2. \tag{36}$$

Using (**A3**) to upper bound eq. equation 33 yields the corresponding upper bound:

$$\frac{\mathrm{KL}_h}{\mathrm{KL}_l} \leq \left(1 + \sqrt{\frac{\|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}}^2}{\sigma_l^2}}\right)^2. \tag{37}$$

A.5 Proof of Proposition 6

Starting from Definition (**A1**), we can write the difference between KL_h and KL_l as

$$\mathrm{KL}_h - \mathrm{KL}_l = \int p(\boldsymbol{\theta} \mid \mathcal{X}) \log\left(\frac{p(\boldsymbol{\theta} \mid \mathcal{X})}{p(\boldsymbol{\theta} \mid \mathcal{X}_h)}\right) d\boldsymbol{\theta} - \int p(\boldsymbol{\theta} \mid \mathcal{X}) \log\left(\frac{p(\boldsymbol{\theta} \mid \mathcal{X})}{p(\boldsymbol{\theta} \mid \mathcal{X}_l)}\right) d\boldsymbol{\theta} \tag{38}$$

$$= -\int p(\boldsymbol{\theta} \mid \mathcal{X}) \log\left(p(\boldsymbol{\theta} \mid \mathcal{X}_h)\right) d\boldsymbol{\theta} + \int p(\boldsymbol{\theta} \mid \mathcal{X}) \log\left(p(\boldsymbol{\theta} \mid \mathcal{X}_l)\right) d\boldsymbol{\theta} \tag{39}$$

$$= \int p(\boldsymbol{\theta} \mid \mathcal{X}) \log\left(\frac{p(\boldsymbol{\theta} \mid \mathcal{X}_l)}{p(\boldsymbol{\theta} \mid \mathcal{X}_h)}\right) d\boldsymbol{\theta}. \tag{40}$$

Under Assumption 1, the log term is expressed as:

$$\log\left(\frac{P(\boldsymbol{\theta} \mid \mathcal{X}_l)}{P(\boldsymbol{\theta} \mid \mathcal{X}_h)}\right) = \log\left(\frac{\frac{1}{Z_l}\exp\left(-\sum_{i=1}^{N} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i) - \ell(\boldsymbol{\theta}, \boldsymbol{x}_l)\right)}{\frac{1}{Z_h}\exp\left(-\sum_{i=1}^{N} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i) - \ell(\boldsymbol{\theta}, \boldsymbol{x}_h)\right)}\right) \tag{41}$$

$$= \log\left(\frac{Z_h}{Z_l}\right) + \left(\ell(\boldsymbol{\theta}, \boldsymbol{x}_h) - \ell(\boldsymbol{\theta}, \boldsymbol{x}_l)\right). \tag{42}$$

Introducing $\Delta(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}, \boldsymbol{x}_h) - \ell(\boldsymbol{\theta}, \boldsymbol{x}_l)$ for notational convenience. A rewrite of Z_h yields,

$$Z_h = \int \exp\left(-\sum_{i=1}^{N} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i) - \ell(\boldsymbol{\theta}, \boldsymbol{x}_l)\right) \exp\left(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l) - \ell(\boldsymbol{\theta}, \boldsymbol{x}_h)\right) d\boldsymbol{\theta} \tag{43}$$

$$= \int Z_l p(\boldsymbol{\theta} \mid \mathcal{X}_l) \exp\left(-\Delta(\boldsymbol{\theta})\right) d\boldsymbol{\theta} \tag{44}$$

$$= Z_l \mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}_l}[\exp\left(-\Delta(\boldsymbol{\theta})\right)]. \tag{45}$$

Substitute back into equation 42:

$$\log\left(\frac{Z_h}{Z_l}\right) + \Delta(\boldsymbol{\theta}) = \log\left(\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}_l}[\exp(-\Delta(\boldsymbol{\theta}))]\right) + \Delta(\boldsymbol{\theta}), \tag{46}$$

and subsequently into equation 40, the difference in KL divergence admits the following closed-form expression:

$$\mathrm{KL}_h - \mathrm{KL}_l = \int p(\boldsymbol{\theta} \mid \mathcal{X}) \left(\log\left(\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}_l}[\exp(-\Delta(\boldsymbol{\theta}))]\right) + \Delta(\boldsymbol{\theta})\right) d\boldsymbol{\theta} \tag{47}$$

$$= \log\left(\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}_l}[\exp(-\Delta(\boldsymbol{\theta}))]\right) + \mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\Delta(\boldsymbol{\theta})]. \tag{48}$$

The use of the cumulant generating function used in Proposition 4 can similarly be applied here:

$$\mathrm{KL}_h - \mathrm{KL}_l \approx \frac{1}{2}\mathrm{Var}_{\boldsymbol{\theta}|\mathcal{X}_l}[\Delta(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}_l}[\Delta(\boldsymbol{\theta})] + \mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\Delta(\boldsymbol{\theta})] \tag{49}$$

To rewrite the last term to follow the same distribution as the first two, we find the relation between the two distributions:

$$\frac{p(\boldsymbol{\theta} \mid \mathcal{X})}{p(\boldsymbol{\theta} \mid \mathcal{X}_l)} = \frac{Z_l}{Z}\exp\left(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)\right) \iff$$

$$p(\boldsymbol{\theta} \mid \mathcal{X}) = \frac{Z_l}{Z}\exp\left(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)\right) p(\boldsymbol{\theta} \mid \mathcal{X}_l).$$

Integration over $\boldsymbol{\theta}$ on both sides yields $Z_l/Z = (\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}_l}[\exp\left(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)\right)])^{-1}$, such that if we multiply both sides with $\Delta(\boldsymbol{\theta})$ and integrate we get

$$\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\Delta(\boldsymbol{\theta})] = \frac{\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}_l}[\exp\left(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)\right)\Delta(\boldsymbol{\theta})]}{\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}_l}[\exp\left(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)\right)]}. \tag{50}$$

Substitute back into equation 49:

$$\mathrm{KL}_h - \mathrm{KL}_l \approx \frac{1}{2}\mathrm{Var}_{\boldsymbol{\theta}|\mathcal{X}_l}[\Delta(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}_l}[\Delta(\boldsymbol{\theta})] + \frac{\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}_l}[\exp(\ell(\boldsymbol{\theta},\boldsymbol{x}_l))\Delta(\boldsymbol{\theta})]}{\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}_l}[\exp(\ell(\boldsymbol{\theta},\boldsymbol{x}_l))]} \tag{51}$$

$$= \frac{1}{2}\mathrm{Var}_{\boldsymbol{\theta}|\mathcal{X}_l}[\Delta(\boldsymbol{\theta})] + \frac{\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}_l}[\exp(\ell(\boldsymbol{\theta},\boldsymbol{x}_l))\Delta(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}_l}[\Delta(\boldsymbol{\theta})]\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}_l}[\exp(\ell(\boldsymbol{\theta},\boldsymbol{x}_l))]}{\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}_l}[\exp(\ell(\boldsymbol{\theta},\boldsymbol{x}_l))]} \tag{52}$$

$$= \frac{1}{2}\mathrm{Var}_{\boldsymbol{\theta}|\mathcal{X}_l}[\Delta(\boldsymbol{\theta})] + \frac{\mathrm{Cov}\left[\Delta(\boldsymbol{\theta}), \exp(\ell(\boldsymbol{\theta},\boldsymbol{x}_l))\right]}{\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}_l}[\exp(\ell(\boldsymbol{\theta},\boldsymbol{x}_l))]} \tag{53}$$

Using (**A2**) on the loss function:

$$\ell(\boldsymbol{\theta},\boldsymbol{x}_h) \approx \ell(\boldsymbol{\theta},\boldsymbol{x}_l) + \nabla_{\boldsymbol{x}}^\top \ell(\boldsymbol{\theta},\boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_l}\boldsymbol{x}_r = \ell(\boldsymbol{\theta},\boldsymbol{x}_l) + \boldsymbol{g}^\top\boldsymbol{x}_r, \tag{54}$$

introducing $\boldsymbol{g} = \nabla_{\boldsymbol{x}}\ell(\boldsymbol{\theta},\boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_l}$ such that $\Delta(\boldsymbol{\theta}) = \boldsymbol{g}^\top\boldsymbol{x}_r$. From this point onward in the derivation all expectations, variances and covariances are taken with respect to $P(\boldsymbol{\theta} \mid \mathcal{X} \cup \{\boldsymbol{x}_l\})$, and we will thus drop subscripts. We then obtain the following lower bound:

$$\begin{aligned}
\mathrm{KL}_h - \mathrm{KL}_l &\approx \tfrac{1}{2}\mathrm{Var}[\boldsymbol{g}^\top\boldsymbol{x}_r] + \frac{\mathrm{Cov}[\boldsymbol{g}^\top\boldsymbol{x}_r, \exp(\ell(\boldsymbol{\theta},\boldsymbol{x}_l))]}{\mathbb{E}[\exp(\ell(\boldsymbol{\theta},\boldsymbol{x}_l))]} \\
&\overset{\textbf{A3}}{\geq} \tfrac{1}{2}\mathrm{Var}[\boldsymbol{g}^\top\boldsymbol{x}_r] - \frac{\sqrt{\mathrm{Var}[\exp(\ell(\boldsymbol{\theta},\boldsymbol{x}_l))]}}{\mathbb{E}[\exp(\ell(\boldsymbol{\theta},\boldsymbol{x}_l))]} \cdot \sqrt{\mathrm{Var}[\boldsymbol{g}^\top\boldsymbol{x}_r]} \\
&\overset{\textbf{A4}+\textbf{A5}}{=} \tfrac{1}{2}\|\boldsymbol{x}_r\|^2_{\Sigma_{\boldsymbol{g}}^l} - \zeta\|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}^l},
\end{aligned}$$

where $\zeta = \sqrt{\mathrm{Var}_{\boldsymbol{\theta}|\mathcal{X}\cup\{\boldsymbol{x}_l\}}[\exp(\ell(\boldsymbol{\theta},\boldsymbol{x}_l))]}/\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}\cup\{\boldsymbol{x}_l\}}[\exp(\ell(\boldsymbol{\theta},\boldsymbol{x}_l))]$. The upper bound is found by upper bounding the covariance term using equation **A3** which results in

$$\mathrm{KL}_h - \mathrm{KL}_l \leq \tfrac{1}{2}\|\boldsymbol{x}_r\|^2_{\Sigma_{\boldsymbol{g}}^l} + \zeta\|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}^l}. \tag{55}$$

B Additional Theoretical Results

For notational simplicity, we add some additional definitions:

1. For a matrix $H \in \mathbb{R}^{s\times s}$ define $\Sigma_H := \mathrm{Cov}[\mathrm{vec}(H)]$
2. $Q_H(\boldsymbol{x}_r) := \left(\mathrm{vec}\left(\boldsymbol{x}_r\boldsymbol{x}_r^\top\right)^\top \Sigma_H \mathrm{vec}\left(\boldsymbol{x}_r\boldsymbol{x}_r^\top\right)\right)^{1/2}$

such that

$$\mathrm{Var}[\boldsymbol{x}_r^\top H\boldsymbol{x}_r] = Q_H(\boldsymbol{x}_r)^2 \tag{56}$$

B.1 Second Order Approximations of Proofs

Here we provide proofs of Proposition 5-6 using a second order Taylor expansion.

Proposition 5 A second order Taylor expansion of the loss function yields:

$$\ell(\boldsymbol{\theta},\boldsymbol{x}_h) = \ell(\boldsymbol{\theta},\boldsymbol{x}_l+\boldsymbol{x}_r) \approx \ell(\boldsymbol{\theta},\boldsymbol{x}_l) + \nabla_{\boldsymbol{x}}^\top\ell(\boldsymbol{\theta},\boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_l}\boldsymbol{x}_r + \tfrac{1}{2}\boldsymbol{x}_r^\top\nabla_{\boldsymbol{x}}^2\ell(\boldsymbol{\theta},\boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_l}\boldsymbol{x}_r \tag{57}$$

$$= \ell(\boldsymbol{\theta},\boldsymbol{x}_l) + \boldsymbol{g}^\top\boldsymbol{x}_r + \tfrac{1}{2}\boldsymbol{x}_r^\top H\boldsymbol{x}_r, \tag{58}$$

where $\boldsymbol{g} = \nabla_{\boldsymbol{x}}\ell(\boldsymbol{\theta},\boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_l}$ and $H := \nabla_{\boldsymbol{x}}^2\ell(\boldsymbol{\theta},\boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_l}$. The numerator of eq. equation 31 is thus given by,

$$\begin{aligned}
\mathrm{Var}[\ell(\boldsymbol{\theta},\boldsymbol{x}_h)] &= \mathrm{Var}[\ell(\boldsymbol{\theta},\boldsymbol{x}_l)] + \mathrm{Var}[\boldsymbol{g}^\top\boldsymbol{x}_r] + \mathrm{Var}\left[\tfrac{1}{2}\boldsymbol{x}_r^\top H\boldsymbol{x}_r\right] \\
&\quad + 2\mathrm{Cov}[\ell(\boldsymbol{\theta},\boldsymbol{x}_l), \boldsymbol{g}^\top\boldsymbol{x}_r] + 2\mathrm{Cov}\left[\ell(\boldsymbol{\theta},\boldsymbol{x}_l), \tfrac{1}{2}\boldsymbol{x}_r^\top H\boldsymbol{x}_r\right] + 2\mathrm{Cov}\left[\boldsymbol{g}^\top\boldsymbol{x}_r, \tfrac{1}{2}\boldsymbol{x}_r^\top H\boldsymbol{x}_r\right].
\end{aligned}$$

Introducing $\sigma_l^2 := \mathrm{Var}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)]$. Apply (**A4**) and using (**A3**) to lower bound all covariances:

$$\begin{aligned}\mathrm{Var}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_h)] &\geq \sigma_l^2 + \boldsymbol{x}_r^\top \Sigma_{\boldsymbol{g}} \boldsymbol{x}_r + \tfrac{1}{4}\mathrm{vec}(\boldsymbol{x}_r \boldsymbol{x}_r^\top)^\top \Sigma_H \mathrm{vec}(\boldsymbol{x}_r \boldsymbol{x}_r^\top) - 2\sqrt{\sigma_l^2 \boldsymbol{x}_r^\top \Sigma_{\boldsymbol{g}} \boldsymbol{x}_r} \\ &\quad - \sqrt{\sigma_l^2 \mathrm{vec}(\boldsymbol{x}_r \boldsymbol{x}_r^\top)^\top \Sigma_H \mathrm{vec}(\boldsymbol{x}_r \boldsymbol{x}_r^\top)} - \sqrt{\boldsymbol{x}_r^\top \Sigma_{\boldsymbol{g}} \boldsymbol{x}_r \mathrm{vec}(\boldsymbol{x}_r \boldsymbol{x}_r^\top)^\top \Sigma_H \mathrm{vec}(\boldsymbol{x}_r \boldsymbol{x}_r^\top)} \\ &\overset{(\mathbf{A5})}{=} \sigma_l^2 + \|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}}^2 + \tfrac{1}{4} Q_H(\boldsymbol{x}_r)^2 \\ &\quad - 2\sigma_l \|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}} - \sigma_l Q_H(\boldsymbol{x}_r) - \|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}} Q_H(\boldsymbol{x}_r).\end{aligned}$$

Rearranging gives us the final lower bound:

$$\frac{\mathrm{KL}_h}{\mathrm{KL}_l} \geq \frac{\left(\sigma_l - \|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}}\right)^2 + \frac{1}{4} Q_H(\boldsymbol{x}_r)^2 - \left(\sigma_l Q_H(\boldsymbol{x}_r) + \|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}} Q_H(\boldsymbol{x}_r)\right)}{\sigma_l^2}. \tag{59}$$

Using (**A3**) to upper bound all covariances results in:

$$\frac{\mathrm{KL}_h}{\mathrm{KL}_l} \leq \frac{\left(\sigma_l + \|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}}\right)^2 + \frac{1}{4} Q_H(\boldsymbol{x}_r)^2 + \left(\sigma_l Q_H(\boldsymbol{x}_r) + \|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}} Q_H(\boldsymbol{x}_r)\right)}{\sigma_l^2}. \tag{60}$$

Proposition 6 A second order Taylor expansion of the loss function yields:

$$\begin{aligned}\ell(\boldsymbol{\theta}, \boldsymbol{x}_h) = \ell(\boldsymbol{\theta}, \boldsymbol{x}_l + \boldsymbol{x}_r) &\approx \ell(\boldsymbol{\theta}, \boldsymbol{x}_l) + \nabla_{\boldsymbol{x}}^\top \ell(\boldsymbol{\theta}, \boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_l} \boldsymbol{x}_r + \tfrac{1}{2} \boldsymbol{x}_r^\top \nabla_{\boldsymbol{x}}^2 \ell(\boldsymbol{\theta}, \boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_l} \boldsymbol{x}_r \qquad (61) \\ &= \ell(\boldsymbol{\theta}, \boldsymbol{x}_l) + \boldsymbol{g}^\top \boldsymbol{x}_r + \tfrac{1}{2} \boldsymbol{x}_r^\top H \boldsymbol{x}_r, \qquad (62)\end{aligned}$$

where $\boldsymbol{g} = \nabla_{\boldsymbol{x}} \ell(\boldsymbol{\theta}, \boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_l}$ and $H := \nabla_{\boldsymbol{x}}^2 \ell(\boldsymbol{\theta}, \boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_l}$. Dropping subscripts we can rewrite equation 52

$$\begin{aligned}\mathrm{KL}_h - \mathrm{KL}_l &\approx \tfrac{1}{2}\mathrm{Var}[\Delta(\boldsymbol{\theta})] - \mathbb{E}[\Delta(\boldsymbol{\theta})] + \frac{\mathbb{E}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))\Delta(\boldsymbol{\theta})]}{\mathbb{E}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))]} \\ &= \tfrac{1}{2}\mathrm{Var}\left[\boldsymbol{g}^\top \boldsymbol{x}_r + \tfrac{1}{2} \boldsymbol{x}_r^\top H \boldsymbol{x}_r\right] - \left(\mathbb{E}\left[\boldsymbol{g}^\top \boldsymbol{x}_r + \tfrac{1}{2} \boldsymbol{x}_r^\top H \boldsymbol{x}_r\right]\right) + \frac{\mathbb{E}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))(\boldsymbol{g}^\top \boldsymbol{x}_r + \frac{1}{2} \boldsymbol{x}_r^\top H \boldsymbol{x}_r)]}{\mathbb{E}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))]}.\end{aligned}$$

Applying (**A4**):

$$\begin{aligned}\mathrm{KL}_h - \mathrm{KL}_l &\approx \tfrac{1}{2}\|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}}^2 + \tfrac{1}{8} Q_H(\boldsymbol{x}_r)^2 + \tfrac{1}{2}\mathrm{Cov}[\boldsymbol{g}^\top \boldsymbol{x}_r, \boldsymbol{x}_r^\top H \boldsymbol{x}_r] - \mathbb{E}[\boldsymbol{g}^\top \boldsymbol{x}_r] \\ &\quad - \tfrac{1}{2}\mathbb{E}[\boldsymbol{x}_r^\top H \boldsymbol{x}_r] + \frac{\mathbb{E}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))\boldsymbol{g}^\top \boldsymbol{x}_r] - \frac{1}{2}\mathbb{E}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))\boldsymbol{x}_r^\top H \boldsymbol{x}_r]}{\mathbb{E}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))]}.\end{aligned}$$

By lower bounding the first covariance term using (**A3**), we can rewrite the first three terms:

$$\begin{aligned}\mathrm{KL}_h - \mathrm{KL}_l &\geq \tfrac{1}{2}\left(\|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}} - \tfrac{1}{2} Q_H(\boldsymbol{x}_r)\right)^2 + \frac{\mathbb{E}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))\boldsymbol{g}^\top \boldsymbol{x}_r] - \mathbb{E}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))]\mathbb{E}[\boldsymbol{g}^\top \boldsymbol{x}_r]}{\mathbb{E}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))]} \\ &\quad + \frac{\frac{1}{2}\mathbb{E}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))\boldsymbol{x}_r^\top H \boldsymbol{x}_r] - \frac{1}{2}\mathbb{E}[\boldsymbol{x}_r^\top H \boldsymbol{x}_r]\mathbb{E}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))]}{\mathbb{E}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))]} \\ &= \tfrac{1}{2}\left(\|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}} - \tfrac{1}{2} Q_H(\boldsymbol{x}_r)\right)^2 + \frac{\mathrm{Cov}\left[\boldsymbol{g}^\top \boldsymbol{x}_r, \exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))\right]}{\mathbb{E}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))]} \\ &\quad + \frac{\frac{1}{2}\mathrm{Cov}\left[\boldsymbol{x}_r^\top H \boldsymbol{x}_r, \exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))\right]}{\mathbb{E}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))]}.\end{aligned}$$

Applying (**A3**) to lower bound the other covariances and (**A5**) to rewrite the norms:

$$\begin{aligned}\mathrm{KL}_h - \mathrm{KL}_l &\geq \tfrac{1}{2}\left(\|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}} - \tfrac{1}{2} Q_H(\boldsymbol{x}_r)\right)^2 \\ &\quad - \frac{\|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}} \sqrt{\mathrm{Var}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))]}}{\mathbb{E}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))]} - \frac{\frac{1}{2} Q_H(\boldsymbol{x}_r)\sqrt{\mathrm{Var}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))]}}{\mathbb{E}[\exp(\ell(\boldsymbol{\theta}, \boldsymbol{x}_l))]},\end{aligned}$$

which gives us the following lower bound:

$$\mathrm{KL}_h - \mathrm{KL}_l \geq \tfrac{1}{2}\left(\|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}} - \tfrac{1}{2} Q_H(\boldsymbol{x}_r)\right)^2 - c\left(\|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}} + \tfrac{1}{2} Q_H(\boldsymbol{x}_r)\right), \tag{63}$$

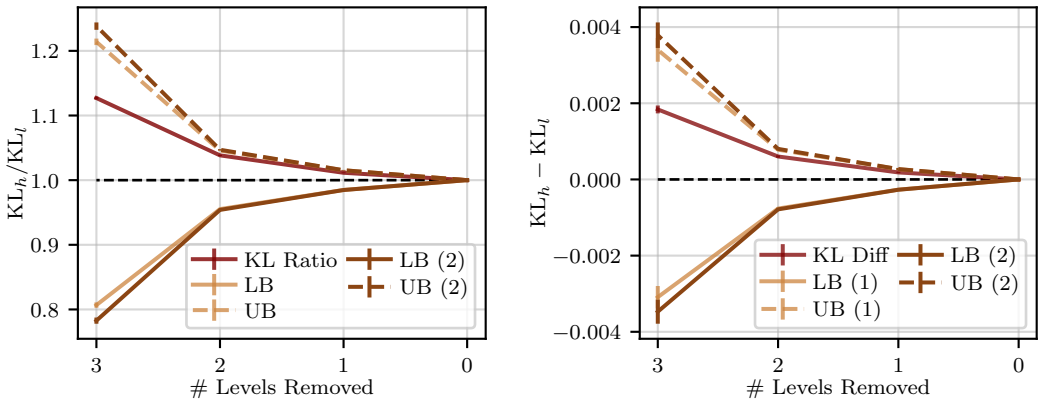

Figure 6: Simulation of bounds as a function of progressively removed high-frequency content. Error bars again denote five standard deviations. LB = Lower Bound, UB = Upper Bound with the order of approximation in parenthesis. There is no added benefit of going from a first to a second order Taylor expansion.

where $c = \sqrt{\mathrm{Var}_{\boldsymbol{\theta}|\mathcal{X}\cup\{\boldsymbol{x}_l\}}[\exp(\ell(\boldsymbol{\theta},\boldsymbol{x}_l))]}/\mathbb{E}[\exp(\ell(\boldsymbol{\theta},\boldsymbol{x}_l))]$. Had (**A3**) been used to upper bound all covariances, we end up with the following upper bound:

$$\mathrm{KL}_h - \mathrm{KL}_l \leq \tfrac{1}{2}\left(\|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}} + \tfrac{1}{2}Q_H(\boldsymbol{x}_r)\right)^2 + c\left(\|\boldsymbol{x}_r\|_{\Sigma_{\boldsymbol{g}}} + \tfrac{1}{2}Q_H(\boldsymbol{x}_r)\right), \tag{64}$$

Figure 6 augments Figure 2 with the proofs using the second order approximations. There is not an added benefit of using a second order Taylor expansion for either bound.

B.2 Proof of Corollary 7

The tighter lower bound for Proposition 5 follows directly from equation 33 lower-bounding the covariance with zero.

Likewise the tighter lower bound for Proposition 6 follows from lower bounding the covariance in equation 53 with zero.

B.3 Proof of Proposition 8

The loss function used in linear regression is the Mean Squared Error (MSE) which is defined as

$$MSE(\boldsymbol{\theta};\boldsymbol{x},y) := (y - \boldsymbol{x}^\top\boldsymbol{\theta})^2. \tag{65}$$

Inserting this into Gibbs posterior from Assumption 1:

$$p(\boldsymbol{\theta}\mid\mathcal{X}) = \frac{1}{Z(\mathcal{X})}\exp\left(-\sum_{i=1}^{N}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\theta})^2\right) = \mathcal{N}(\boldsymbol{\theta}\mid\boldsymbol{\mu},\Sigma), \tag{66}$$

thus $\boldsymbol{\theta}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Introducing $z_i := \boldsymbol{x}_i^\top\boldsymbol{\theta}\sim\mathcal{N}(m,s^2)$ where (Murphy, 2022)

$$m = \boldsymbol{x}_i^\top\boldsymbol{\mu}, \qquad s^2 = \boldsymbol{x}_i^\top\boldsymbol{\Sigma}\boldsymbol{x}_i. \tag{67}$$

Insert the definition of z_i in the exponential term from equation 3:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\exp\left(-(y_i - \boldsymbol{x}_i^\top\boldsymbol{\theta})^2\right)] &= \mathbb{E}_{z_i\sim\mathcal{N}(m,s^2)}[\exp\left(-(y_i - z_i)^2\right)] \\
&= \int \exp\left(-(y_i - z_i)^2\right)\frac{1}{\sqrt{2\pi s^2}}\exp\left(-\frac{(z_i - m)^2}{2s^2}\right)dz_i \\
&= \frac{1}{\sqrt{2\pi s^2}}\int \exp\left(-(y_i - z_i)^2 - \frac{(z_i - m)^2}{2s^2}\right)dz_i.
\end{aligned}$$

Rewriting the exponent:

$$(y - z_i)^2 + \frac{(z_i - m)^2}{2s^2} = y^2 + z_i^2 - 2yz_i + \frac{z_i^2 + m^2 - 2z_i m}{2s^2}$$
$$= az_i^2 + bz_i + c,$$

where $a = 1 + 1/(2s^2)$, $b = -2y - m/s^2$ and $c = y^2 + m^2/(2s^2)$. The exponent can thus be rewritten to

$$-(az_i^2 + bz_i + c) = -a\left(z_i + \frac{b}{2a}\right)^2 + \frac{b^2}{4a} - c$$
$$= -a'(z_i - b')^2 + c',$$

where $a' = a$, $b' = -b/(2a)$ and $c' = b^2/(4a) - c$. Substitute back into the exponential term:

$$\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\exp\left(-(y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta})^2\right)] = \frac{1}{\sqrt{2\pi s^2}} \int \exp\left(-a'(z_i - b')^2 + c'\right) dz_i$$
$$= \frac{1}{\sqrt{2\pi s^2}} \exp(c') \int \exp\left(-a'(z_i - b')^2\right) dz_i$$
$$= \frac{1}{\sqrt{2\pi s^2}} \exp\left(c'\right) \sqrt{\frac{\pi}{a'}}$$

Inserting the definitions of c' and a':

$$\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\exp\left(-(y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta})^2\right)] = \frac{\sqrt{\pi}}{\sqrt{2\pi s^2\left(1 + \frac{1}{2s^2}\right)}} \exp\left(\frac{b^2}{4a} - c\right)$$
$$= \frac{1}{\sqrt{2s^2 + 1}} \exp\left(\frac{(-2y_i - \frac{m}{s^2})^2}{4a} - (y_i^2 + \frac{m^2}{2s^2})\right)$$
$$= \frac{1}{\sqrt{2s^2 + 1}} \exp\left(\frac{4y^2 + \frac{m^2}{s^4} + \frac{4y_i m}{s^2}}{4a} - (y_i^2 + \frac{m^2}{2s^2})\right)$$
$$= \frac{1}{\sqrt{2s^2 + 1}} \exp\left(\frac{1}{a}\left(y_i^2 + \frac{m^2}{4s^4} + \frac{y_i m}{s^2}\right) - (y_i^2 + \frac{m^2}{2s^2})\right).$$

Since $1/a = 2s^2/(2s^2 + 1)$ then the exponent, c' can be rewritten as such:

$$c' = \frac{2s^2}{2s^2 + 1}\left(y_i^2 + \frac{m^2}{4s^4} + \frac{y_i m}{s^2}\right) - \left(y_i^2 + \frac{m^2}{2s^2}\right)$$
$$= \frac{2s^2 y_i^2}{2s^2 + 1} + \frac{m^2}{(2s^2 + 1)2s^2} + \frac{2my_i}{1 + 2s^2} - y_i^2 - \frac{m^2}{2s^2}$$
$$= y_i^2\left(\frac{2s^2}{2s^2 + 1} - 1\right) + \frac{2my}{2s^2 + 1} + \frac{m^2}{2s^2}\left(\frac{1}{2s^2 + 1} - 1\right)$$
$$= -\frac{y_i^2}{2s^2 + 1} + \frac{2y_i m}{2s^2 + 1} - \frac{m^2}{2s^2 + 1}$$
$$= \frac{1}{2s^2 + 1}\left(-y_i^2 - m^2 + 2y_i m\right)$$
$$= -\frac{(y_i - m)^2}{2s^2 + 1},$$

and thus

$$\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\exp\left(-(y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta})^2\right)] = \frac{1}{\sqrt{1 + 2s^2}} \exp\left(-\frac{(y_i - m)^2}{2s^2 + 1}\right) \Rightarrow$$
$$\log \mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\exp\left(-(y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta})^2\right)] = -\frac{1}{2}\log(1 + 2s^2) - \frac{(y_i - m)^2}{2s^2 + 1}.$$

Computing the other term from equation 3:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_i)] &= \mathbb{E}_{\boldsymbol{\theta}|\mathcal{X}}[(y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta})^2] \\ &= \mathbb{E}_{z_i \sim \mathcal{N}(m, s^2)}[(y_i - z_i)^2] \\ &= y_i^2 + \mathbb{E}_{z_i \sim \mathcal{N}(m, s^2)}[z_i^2] - 2y_i \mathbb{E}_{z_i \sim \mathcal{N}(m, s^2)}[z_i] \\ &= y_i^2 + m^2 + s^2 - 2y_i m \\ &= (y_i - m)^2 + s^2.\end{aligned}$$

KL_i can thus be written as

$$\mathrm{KL}_i = -\frac{1}{2}\log(1 + 2s^2) - \frac{(y_i - m)^2}{2s^2 + 1} + (y_i - m)^2 + s^2 \tag{68}$$

We have for different limit values of s^2:

$$\mathrm{KL}_i \to \begin{cases} 0 & \text{for } s^2 \to 0 \\ \infty & \text{for } s^2 \to \infty \end{cases}, \tag{69}$$

which means the influence of a datapoint $\boldsymbol{x}_i$ can be expressed as such:

$$\mathrm{KL}_i = \mathcal{O}(s^2), \tag{70}$$

for the case of linear regression. This aligns well with Proposition 4.

B.4 Analytical Covariance Assumption

The covariance assumption for the fraction of KL-divergences looks the following:

$$\mathrm{Cov}_{\boldsymbol{\theta}|\mathcal{X}}[\boldsymbol{g}^\top \boldsymbol{x}_r, \ell(\boldsymbol{\theta}, \boldsymbol{x}_l)] \geq 0. \tag{71}$$

First we derive $\boldsymbol{g}$:

$$\boldsymbol{g} := \nabla_{\boldsymbol{x}} \ell(\boldsymbol{\theta}, \boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_l} = \nabla_{\boldsymbol{x}} (y - \boldsymbol{x}^\top \boldsymbol{\theta})^2\big|_{\boldsymbol{x}=\boldsymbol{x}_l} \tag{72}$$

$$= -2(y - \boldsymbol{x}_l^\top \boldsymbol{\theta})\boldsymbol{\theta}. \tag{73}$$

Inserting in the covariance, and introducing $z_i := \boldsymbol{x}_i^\top \boldsymbol{\theta}$

$$\begin{aligned}\mathrm{Cov}_{\boldsymbol{\theta}|\mathcal{X}}[\boldsymbol{g}^\top \boldsymbol{x}_r, \ell(\boldsymbol{\theta}, \boldsymbol{x}_l)] &= \mathrm{Cov}_{\boldsymbol{\theta}|\mathcal{X}}[-(2(y - \boldsymbol{x}_l^\top \boldsymbol{\theta})\boldsymbol{\theta})^\top \boldsymbol{x}_r, (y - \boldsymbol{x}_l^\top \boldsymbol{\theta})^2] \\ &= \mathrm{Cov}_{z_l \sim \mathcal{N}(m, s^2)}[-2(y - z_l)\boldsymbol{\theta}^\top \boldsymbol{x}_r, (y - z_l)^2] \\ &= \mathrm{Cov}_{(z_l, z_r) \sim \mathcal{N}(\boldsymbol{m}, \Sigma_{lr})}[-2(y - z_l)z_r, (y - z_l)^2].\end{aligned}$$

Introduce the residual $r_i := y - z_i$:

$$\begin{aligned}\mathrm{Cov}_{\boldsymbol{\theta}|\mathcal{X}}[\boldsymbol{g}^\top \boldsymbol{x}_r, \ell(\boldsymbol{\theta}, \boldsymbol{x}_l)] &= -2\mathrm{Cov}_{(z_l, z_r) \sim \mathcal{N}(\boldsymbol{m}, \Sigma_{lr})}[r_l z_r, r_l^2] \\ &= -2\left(\mathbb{E}_{(z_l, z_r) \sim \mathcal{N}(\boldsymbol{m}, \Sigma_{lr})}[r_l^3 z_r] - \mathbb{E}_{(z_l, z_r) \sim \mathcal{N}(\boldsymbol{m}, \Sigma_{lr})}[r_l z_r]\mathbb{E}_{z_l \sim \mathcal{N}(m, s^2)}[r_l^2]\right).\end{aligned}$$

Now we have these three terms that we will look at one-by-one:

1. $\mathbb{E}_{z_l \sim \mathcal{N}(m, s^2)}[r_l^2]$ has been derived above, as this is just the expected loss:

$$\mathbb{E}_{z_l \sim \mathcal{N}(m, s^2)}[r_l^2] = (y - m_l)^2 + s_l^2$$

2. $\mathbb{E}_{(z_l, z_r) \sim \mathcal{N}(m, \Sigma_{lr})}[r_l z_r]$ use the rule about the product about distribution of two normally distributed variables

$$\begin{aligned}\mathbb{E}_{(z_l, z_r) \sim \mathcal{N}(\boldsymbol{m}, \Sigma_{lr})}[r_l z_r] &= \mathbb{E}[y z_r] - \mathbb{E}[z_l z_r] \\ &= y m_r - (m_l m_r + \boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r) \\ &= (y - m_l) m_r - \boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r\end{aligned}$$

3. $\mathbb{E}_{(z_l,z_r)\sim\mathcal{N}(\boldsymbol{m},\Sigma_{\boldsymbol{\theta}})}[r_l^3 z_r]$. Introduce $u = z_l - m_l$, $v = z_r - m_r$ and $a = y - m_l$. Then

$$\begin{aligned}
\mathbb{E}_{(z_l,z_r)\sim\mathcal{N}(\boldsymbol{m},\Sigma_{lr})}[r_l^3 z_r] &= \mathbb{E}[(y-(m_l+u))^3(m_r+v)] \\
&= \mathbb{E}[(a-u)^3(m_r+v)] \\
&= \mathbb{E}[(a^3 - 3a^2u + 3au^2 - u^3)(m_r+v)] \\
&= \mathbb{E}[a^3 m_r - 3a^2 u m_r + 3au^2 m_r - u^3 mr \\
&\quad + a^3 v - 3a^2 uv + 3au^2 v - u^3 v] \\
&= a^3 m_r + 3as_l^2 m_r - 3a^2\mathbb{E}[uv] - \mathbb{E}[u^3 v] \\
&= a^3 m_r + 3as_l^2 m_r - 3a^2 \boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r - 3s_l^2 \boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r \\
&= m_r((y-m_l)^3 + 3(y-m_l)s_l^2) - 3\boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r((y-m_l)^2 + s_l^2)
\end{aligned}$$

Joining the three terms:

$$\begin{aligned}
\mathrm{Cov}_{\boldsymbol{\theta}|\mathcal{X}}[\boldsymbol{g}^\top \boldsymbol{x}_r, \ell(\boldsymbol{\theta}, \boldsymbol{x}_l)] &= -2\left(\mathbb{E}_{(z_l,z_r)\sim\mathcal{N}(\boldsymbol{m},\Sigma_{lr})}[r_l^3 z_r] - \mathbb{E}_{(z_l,z_r)\sim\mathcal{N}(\boldsymbol{m},\Sigma_{lr})}[r_l z_r]\mathbb{E}_{z_l\sim\mathcal{N}(m,s^2)}[r_l^2]\right) \\
&= -2\big(m_r((y-m_l)^3 + 3(y-m_l)s_l^2) - 3\boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r((y-m_l)^2 + s_l^2) \\
&\quad - ((y-m_l)^2 + s_l^2)\cdot((y-m_l)m_r - \boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r))\big) \\
&= -2\big(m_r(y-m_l)^3 + 3m_r(y-m_l)s_l^2 - 3\boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r(y-m_l)^2 - 3\boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r \\
&\quad - (y-m_l)^3 m_r + \boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r(y-m_l)^2 - s_l^2(y-m_l)m_r + s_l^2 \boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r\big).
\end{aligned}$$

Group terms by m_r and $\Sigma_{\boldsymbol{\theta}}$:

$$\begin{aligned}
\mathrm{Cov}_{\boldsymbol{\theta}|\mathcal{X}}[\boldsymbol{g}^\top \boldsymbol{x}_r, \ell(\boldsymbol{\theta}, \boldsymbol{x}_l)] &= -2\Big(m_r\big((y-m_l)^3 + 3(y-m_l)s_l^2 - (y-m_l)^3 - s_l^2(y-m_l)\big) \\
&\quad + \boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r\big(-3(y-m_l)^2 - 3s_l^2 + (y-m_l)^2 + s_l^2\big)\Big) \\
&= -2\big(2m_r(y-m_l)s_l^2 - \boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r(2(y-m_l)^2 + 2s_l^2)\big),
\end{aligned}$$

hence the covariance is given by

$$\mathrm{Cov}_{\boldsymbol{\theta}|\mathcal{X}}[\boldsymbol{g}^\top \boldsymbol{x}_r, \ell(\boldsymbol{\theta}, \boldsymbol{x}_l)] = -4m_r(y-m_l)s_l^2 + 4\boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r((y-m_l)^2 + s_l^2). \tag{74}$$

That means in order for the assumption to hold, the following must be true:

$$\boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r((y-m_l)^2 + s_l^2) \geq m_r(y-m_l)s_l^2. \tag{75}$$

In order to understand this covariance assumption better, we conduct some simulations. That requires we generate some data that can be considered of high- and low-resolution in the linear regression case.

B.4.1 Generate Data for Linear Regression

To study when the covariance condition holds, we require high- and low-resolution versions of the same datapoint in a linear regression setting. For each experiment, we generate 200 signals of length 20 and fit a linear regession model with a bias term and one weight per coefficient. Signals are generated in the frequency domain by sampling coefficients from a normal distribution and scaling them by

$$h(f) = \frac{1}{(1+f)^\alpha}, \quad \alpha = 2, \tag{76}$$

where f is frequency index, before being transformed back to the signal domain. Low-resolution samples are obtained by low-pass filtering with a Kaiser window. The scaling of the high-frequencies ensures that a moderate amount of energy is removed by the low-pass filter and that label information is not too concentrated in high-frequency components. Targets are generated as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{\eta}$, where $\boldsymbol{\theta}^*$ is sampled from a normal distribution and $\eta \sim \mathcal{N}(0, 0.5^2)$. From the resulting Gaussian posterior presented in Proposition 8, we compute m_l, m_r, s_l^2 and $\boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r$ for each datapoint. The procedure is repeated 2000 times.

B.4.2 Plotting the Covariance Simulations

We begin with rewriting equation 74 to make it dependent on fewer parameters:

$$\mathrm{Cov}_{\boldsymbol{\theta}|\mathcal{X}}[\boldsymbol{g}^\top \boldsymbol{x}_r, \ell(\boldsymbol{\theta}, \boldsymbol{x}_l)] = -4m_r(y - m_l)s_l^2 + 4\boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r((y - m_l)^2 + s_l^2) \iff \tag{77}$$

$$\frac{\mathrm{Cov}[\boldsymbol{g}^\top \boldsymbol{x}_r, \ell(\boldsymbol{\theta}, \boldsymbol{x}_l)]}{4s_l^4} = \rho(u^2 + 1) - \nu u, \tag{78}$$

where

$$\rho = \frac{\boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r}{s_l^2}, \qquad \nu = \frac{m_r}{s_l}, \qquad u = \frac{(y - m_l)}{s_l}. \tag{79}$$

Further introducing

$$r = \frac{\nu}{\rho} = \frac{m_r s_l}{\boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r}, \tag{80}$$

such that

$$\frac{\mathrm{Cov}[\boldsymbol{g}^\top \boldsymbol{x}_r, \ell(\boldsymbol{\theta}, \boldsymbol{x}_l)]}{4s_l^4} = \rho(u^2 + 1) - \rho r u \tag{81}$$

$$= \rho(u^2 + 1 - ru). \tag{82}$$

The sign of the covariance is therefore given by

$$\mathrm{sign}(\mathrm{Cov}[\boldsymbol{g}^\top \boldsymbol{x}_r, \ell(\boldsymbol{\theta}, \boldsymbol{x}_l)]) = \mathrm{sign}(\rho)\mathrm{sign}(u^2 - ru + 1). \tag{83}$$

The sign changes where $u^2 - ru + 1 = 0$ i.e. where $r = u + 1/u$. Since $|u + 1/u| \geq 2$, then the polynomial is strictly positive when $|r| < 2$. In this region, the sign of the covariance, is determined by ρ alone, and the condition holds if and only if $\boldsymbol{x}_l^\top \Sigma_{\boldsymbol{\theta}} \boldsymbol{x}_r \geq 0$. When $|r| > 2$, the polynomial is negative for u between the roots $(r \pm \sqrt{r^2 - 4})/2$, so the sign flips within this band.

The two cases are most easily interpreted directly from equation 78, that shows how the assumption holds if and only if $\rho(u^2 + 1) \geq \nu u$. The two terms reflect two ways in which the residual component can correct the low-resolution prediction. The term νu is positive when the mean residual prediction m_r has the same sign as the low-resolution error, $y - m_l$, i.e. when including $\boldsymbol{x}_r$ moves the mean prediction towards the target. The term ρ is negative when the predictions of the low-resolution and the residual components are negatively correlated. The assumption is therefore violated when the residual component corrects the low-resolution mean prediction or its uncertainty (given by ρ). In these cases the high-resolution datapoint can have a lower loss variance than the low-resolution counterpart.

Figure 7 visualises these regimes for the synthetic data described in the previous section. The left panel shows r against u at 58% energy removed, together with the curve $r = u + 1/u$, with points coloured by the normalised covariance. The middle panel follows five datapoints as energy is progressively removed, and the right panel shows the fraction of datapoints violating the assumption as a function of removed energy. Violations become more frequent as more energy is discarded, remaining below 20% up to roughly 58% energy removed and reaching about 35% at the highest levels. When the assumption is violated, only the tighter bounds of Corollary 7 cease to apply, but the bounds from Proposition 5 still hold up.

C Simulation Details

C.1 Synthetic Data via Wavelets

To empirically validate the bounds derived in Proposition 5 and 6, we require data points that satisfy the decomposition:

$$\boldsymbol{x}_h = \boldsymbol{x}_l + \boldsymbol{x}_r. \tag{84}$$

We construct such data using the Discrete Wavelet Transform (DWT). The DWT provides an exact, non-redundant representation of a signal using a finite set of coefficients and enables lossless

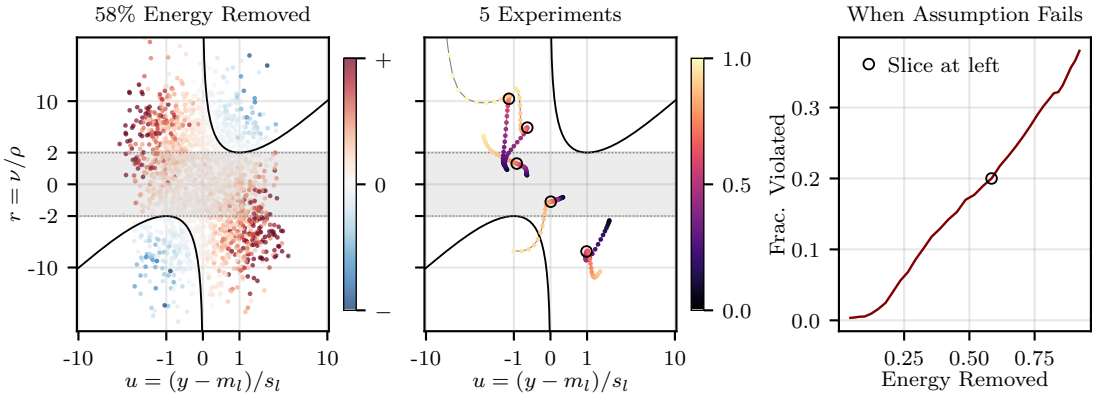

Figure 7: The covariance condition of Corollary 7 for linear regression. **Left**: datapoints at 58% energy removed, plotted by normalised residual $u = (y - m_l)/s_l$ and $r = \nu/\rho$, coloured by $\mathrm{Cov}/(4s_l^4)$. The black curve $r = u + 1/u$ marks where the sign flips. **Middle**: trajectories of five datapoints as energy is removed with black corresponding to not energy removal. The black circle marks the slice seen on the left. **Right**: fraction of datapoints violating the condition as a function of energy removed; the circle marks the slice shown on the left.

reconstruction via its inverse transform (Mallat, 2009). It also yields a multi-resolution decomposition with basis functions localized in both time and frequency.

Given a signal $\boldsymbol{x} \in \mathbb{R}^s$, the DWT produces a set of coefficients $\{a_J, d_J, d_{J-1}, \ldots, d_1\}$, where a_J denotes the approximation coefficients capturing low-frequency content, and d_j denotes the detail coefficients capturing progressively higher frequency components. We construct a low-resolution signal $\boldsymbol{x}_l$ by applying the inverse DWT using only a subset of coefficients (e.g., a_J and selected d_j). The residual component $\boldsymbol{x}_r$ is obtained by applying the inverse DWT to the complementary set of detail coefficients. Since the inverse DWT is linear and the coefficient sets form a disjoint partition, their reconstructions add exactly to the original signal, yielding equation 84. As PCA is also a linear transform, the subsequent projection keeps the additive decomposition.

Figure 8 illustrates this construction. On the left we have a scalogram of a synthetic signal, and the right shows exactly what parts of the time series belong to the respective levels in the scalogram.

Figure 9 demonstrates what the removal of a different number of levels has of impact on a synthetic time series and a CIFAR10 image.

C.2 Exact Values of Bounds

Table 1 contains the actual KL values, variance approximations and bounds.

C.3 Approximation Accuracy and Bound Tightness

We further investigate the accuracy of the approximation in Proposition 4 and the tightness of propositions 5.

Accuracy of Variance Approximation We first study the accuracy of the variance approximation in Proposition 4. Instead of projecting the synthetic data into a fixed 10-dimensional space, we vary the projected input dimension from 2 to 100. For each input dimension, we compute the relative error

$$e_r = \frac{\mathrm{KL}_h/\mathrm{KL}_l - \mathrm{Var}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_h)]/\mathrm{Var}[\ell(\boldsymbol{\theta}, \boldsymbol{x}_l)]}{\mathrm{KL}_h/\mathrm{KL}_l}, \tag{85}$$

where the low-resolution input has had three wavelet bands removed. We evaluate this quantity for models with 2, 3, and 4 hidden layers. The results can be seen in Figure 10. The relative error decreases as model size increases, and remains below 3% even for the 2-layer model. This indicates that the variance approximation provides a good description of the KL-ratio.

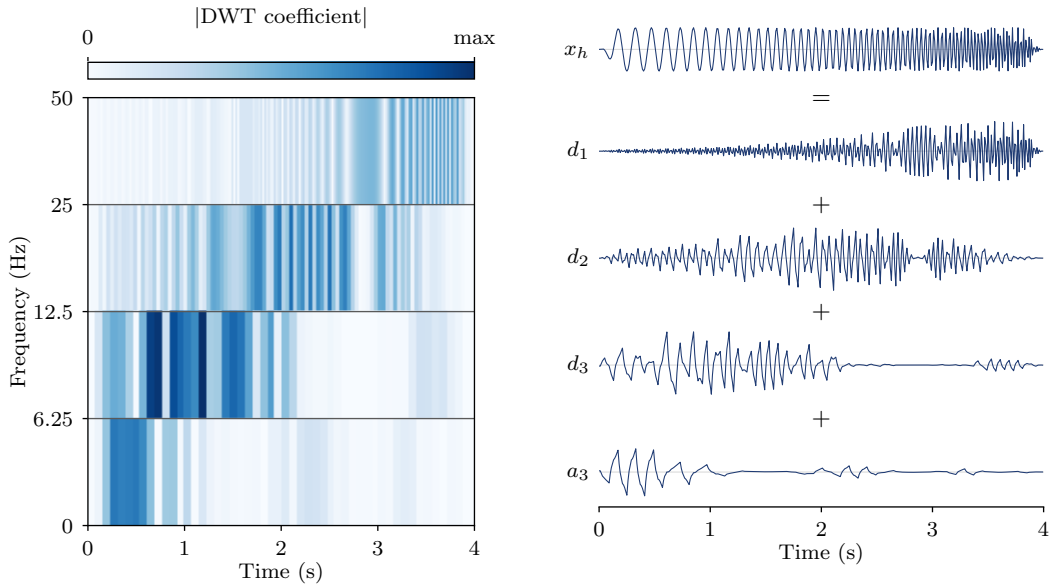

Figure 8: Wavelet construction of the synthetic data. **Left**: magnitude of the db2 DWT coefficients of $\boldsymbol{x}_h$. **Right**: $\boldsymbol{x}_h$ and the reconstruction from each band in isolation. The coefficient sets are disjoint and the inverse DWT is linear, so the four reconstructions sum exactly to $\boldsymbol{x}_h$, which is the decomposition in equation 84.

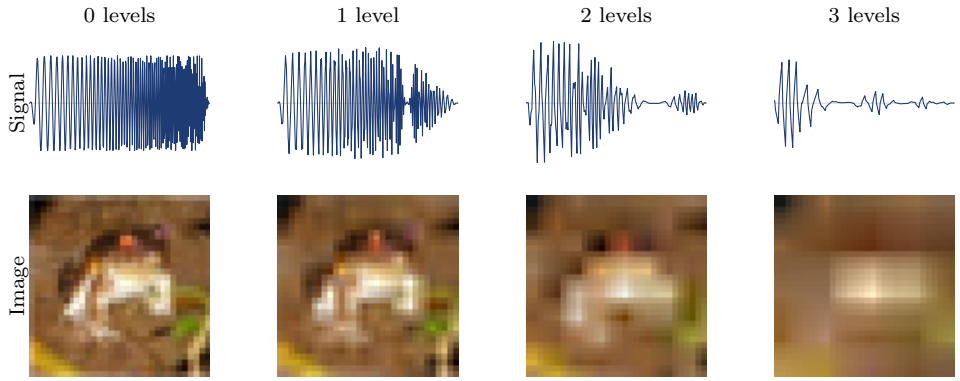

Figure 9: Effect of zeroing the k finest detail bands of the db2 DWT, for $k = 0, \ldots, 3$, on the synthetic 1D signal (top) and on a CIFAR-10 image (bottom). This is the operation used to construct the low-resolution inputs $\boldsymbol{x}_l$.

Tightness of the KL-ratio bounds. To study when the bounds in Proposition 5 and Corollary 7 are tight, we define the gap between the true KL-ratio and a bound as our measure of tightness. For lower bounds, this is the difference between the true value and the bound; for upper bounds, we report the absolute difference. Figure 11 shows this analysis for projected input dimensions from 2 to 100 and for networks with 2, 3, and 4 hidden layers.

In the top panels, three wavelet bands are removed and the resulting gap is plotted as a function of input dimension. The corresponding low-resolution loss variance σ_l^2 is overlaid, since it appears in all KL-ratio bounds. In the bottom panels, we fix a 2-layer network and vary the number of removed frequency bands. The bars show the corresponding bound gaps, while the directional residual norm $\|\boldsymbol{x}_r\|_{\Sigma_g}$ is shown on the secondary axis. The results show that the bounds become looser as more residual information is removed, and tighter as model size increases.

Table 1: Exact values of KL-ratio, KL-diff, and the bounds as seen in Figure 2. Levels refer to the number of frequency bands removed in the scalogram generated from the DWT. LB = lower bound, UB = upper bound, and Cor refers to the tighter bounds in Corollary 7.

# Levels	KL Ratio	Ratio LB	Ratio UB	Ratio LB (Cor)
Levels 0	$1.0 \pm 0.00 \cdot 10^{0}$	$1.0 \pm 0.00 \cdot 10^{0}$	$1.0 \pm 0.00 \cdot 10^{0}$	$1.0 \pm 0.00 \cdot 10^{0}$
Levels 1	$1.01 \pm 1.10 \cdot 10^{-4}$	$0.98 \pm 9.00 \cdot 10^{-5}$	$1.02 \pm 9.00 \cdot 10^{-5}$	$1.00 \pm 0.00 \cdot 10^{0}$
Levels 2	$1.04 \pm 2.00 \cdot 10^{-4}$	$0.96 \pm 1.80 \cdot 10^{-4}$	$1.05 \pm 1.80 \cdot 10^{-4}$	$1.00 \pm 0.00 \cdot 10^{0}$
Levels 3	$1.13 \pm 4.30 \cdot 10^{-4}$	$0.81 \pm 8.20 \cdot 10^{-4}$	$1.21 \pm 1.01 \cdot 10^{-3}$	$1.01 \pm 9.00 \cdot 10^{-5}$
	KL Diff	**Diff LB**	**Diff UB**	**Diff LB (Cor)**
0	$0.00 \cdot 10^{0} \pm 0.00 \cdot 10^{0}$	$0.00 \cdot 10^{0} \pm 0.00 \cdot 10^{0}$	$0.00 \cdot 10^{0} \pm 0.00 \cdot 10^{0}$	$0.00 \cdot 10^{0} \pm 0.00 \cdot 10^{0}$
1	$1.80 \cdot 10^{-4} \pm 0.00 \cdot 10^{0}$	$-2.70 \cdot 10^{-4} \pm 1.00 \cdot 10^{-5}$	$2.70 \cdot 10^{-4} \pm 1.00 \cdot 10^{-5}$	$0.00 \cdot 10^{0} \pm 0.00 \cdot 10^{0}$
2	$6.00 \cdot 10^{-4} \pm 1.00 \cdot 10^{-5}$	$-7.60 \cdot 10^{-4} \pm 1.00 \cdot 10^{-5}$	$7.80 \cdot 10^{-4} \pm 1.00 \cdot 10^{-5}$	$1.00 \cdot 10^{-5} \pm 0.00 \cdot 10^{0}$
3	$1.84 \cdot 10^{-3} \pm 2.00 \cdot 10^{-5}$	$-3.09 \cdot 10^{-3} \pm 6.00 \cdot 10^{-5}$	$3.40 \cdot 10^{-3} \pm 6.00 \cdot 10^{-5}$	$1.60 \cdot 10^{-4} \pm 0.00 \cdot 10^{0}$

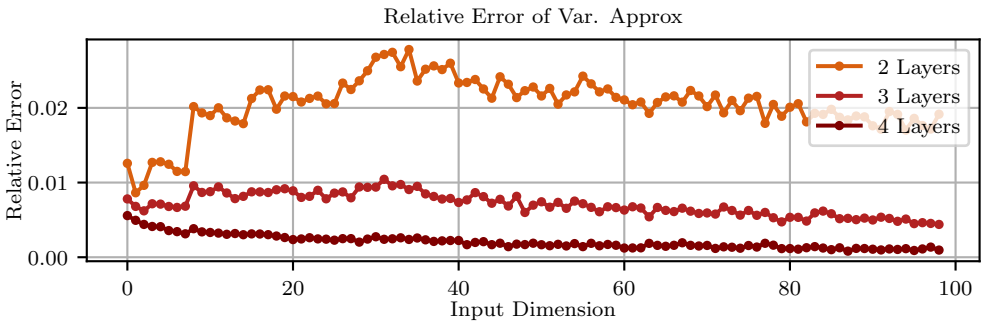

Figure 10: Relative error of the variance approximation in Proposition 4 as a function of input dimension and model size. Three levels of the scalogram are removed from the low-resolution datapoint.

C.4 Toy Example

To illustrate when a high-resolution datapoint can be more informative than its low-resolution counterpart, we consider a simple toy example. The left panel of Figure 12 shows that progressively downsampled CIFAR10 representations occupy a more concentrated region after PCA projection. Motivated by this assumption, we construct a two-class synthetic dataset in which low-resolution points have one eighth of the variance of high-resolution points. We fit linear discriminant analysis (Fisher, 1936) and scale each point by the parameter change induced by its inclusion in the training set. The high-resolution points extending beyond the support of the low-resolution distribution generally produce larger parameter changes, illustrating how information discarded by downsampling can govern datapoint influence.

D Experimental Details

D.1 Training Procedure

We process minibatches proportionally to the data mixture: i.e. if the high-resolution ratio is 80%, four high-resolution batches are processed for every low-resolution batch. Training uses the AdamW optimiser with constant weight decay of 0.05, an initial learning rate of 5×10^{-4}, and a learning-rate schedule consisting of 20 warm-up epochs followed by cosine annealing. Experiments run for 100 epochs in total. Gradients are clipped to a maximum norm of 1.0. Experiments on CIFAR are run with a batchsize of 128, 64 for AudioMNIST and 32 for Fed-ISIC-2019.

Data augmentation for the CIFAR and Fed-ISIC-2019 datasets consists of random crops, horizontal flips, random erasing, and colour jitter. For AudioMNIST, we use random time shifts, random erasing, and magnitude warping. For both CIFAR and AudioMNIST, CutMix and Mixup are applied with

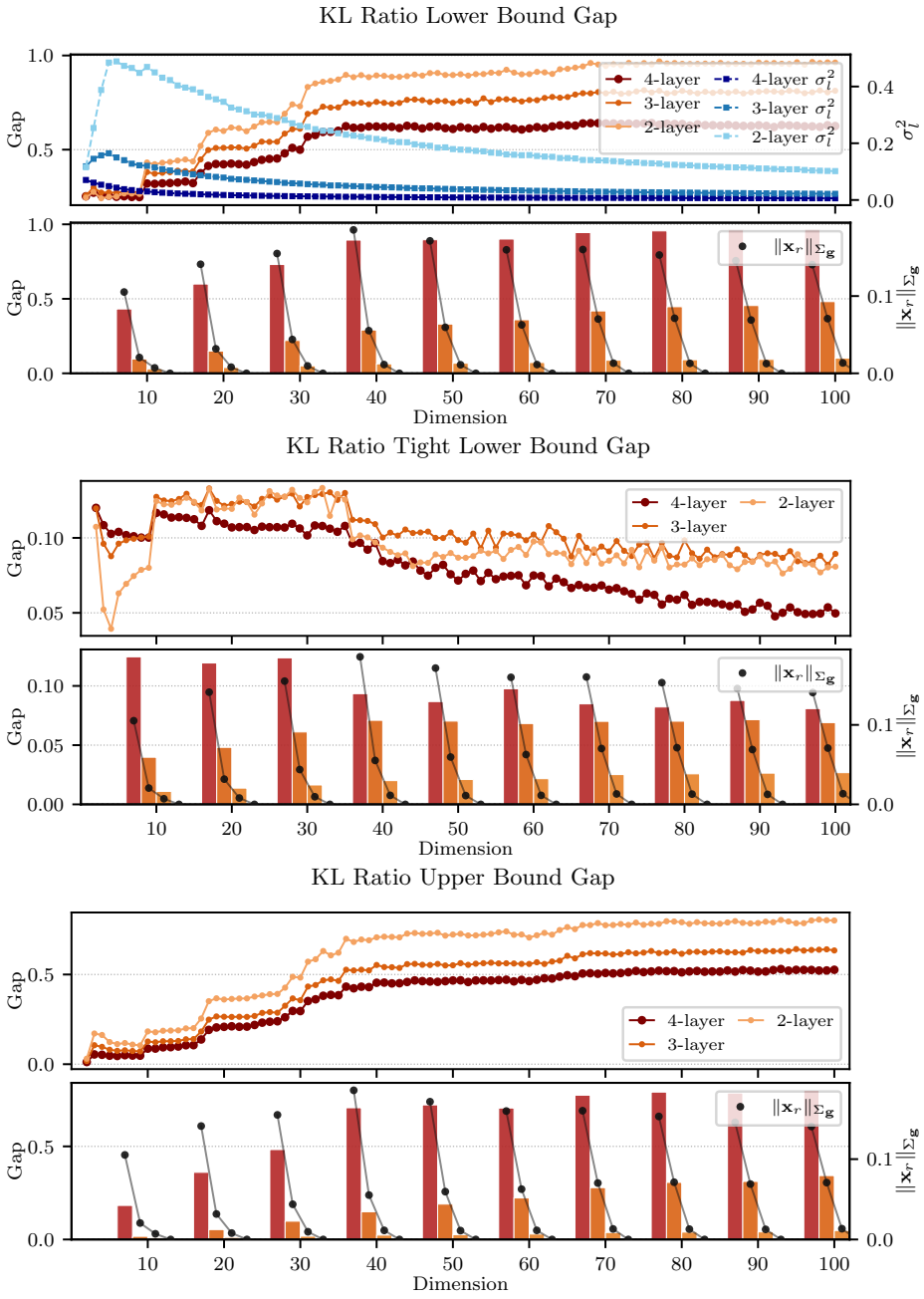

Figure 11: Tightness of the KL-ratio bounds from Proposition 5 and Corollary 7 under varying projected input dimension and removed frequency content. Top panels: gap between the true KL-ratio and each bound for 2-, 3-, and 4-layer networks after removing three wavelet bands. Bottom panels: corresponding bound gaps for a 2-layer model under different levels of removed frequency content, with the four bars representing, from left to right, the removal of three, two, one, and zero bands. The directional residual norm $\|x_r\|_{\Sigma_g}$ is overlaid on the secondary axis. The bounds become looser as more residual energy is removed.

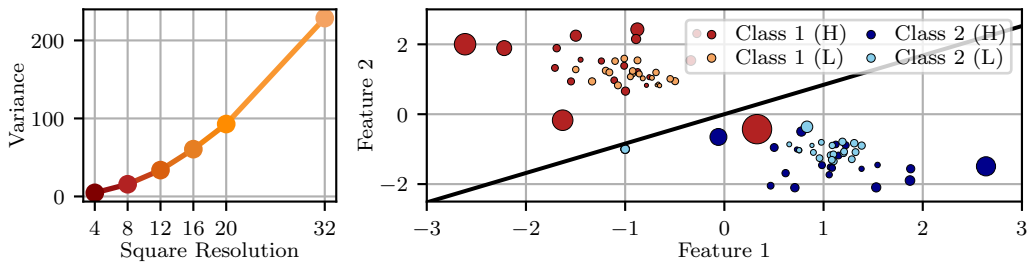

Figure 12: **Left**: Variance of low-dimensional representations as a function of image resolution. Computed from PCA projections of progressively downsampled images from CIFAR10. **Right**: Synthetic two-class data separated using linear discriminant analysis. Marker size given by the magnitude of the change in model parameters when including a datapoint in the training set. The black line is the decision boundary given by training on all points.

probability 0.5, except for low-resolutions sizes that constitute less than 10% of the high-resolution size. CutMix and Mixup are turned off for the Fed-ISIC-2019 dataset, due to the high class imbalance.

D.2 Experimental Setup

For the transformer-based experiments, we modify the ViT architecture from Yoshioka (2024). For CIFAR datasets and Fed-ISIC-2019 we use a depth of 8, an attention head number of 6. CIFAR use a patch size of 4 and Fed-ISIC-2019 uses a patch size of 14. For AudioMNIST, we use the same ViT but modify it to handle 1D inputs and choose a depth of 4, an attention head number of 4, and a patch size of 8. The network is configured with a combination of fixed sinusoidal PEs and learnable ones.

CIFAR and Fed-ISIC-2019 in CNN-based experiments are conducted using the ResNet-18 architecture, where the initial max-pooling layer is removed and first kernel size is changed from 7 to 3 to better accommodate smaller input resolutions. For AudioMNIST we modify the convolutional architecture presented in the original paper to include an adaptive pooling layer such that it can accommodate differently sized inputs.

For the Fed-ISIC-2019 dataset 30% low-resolution means that 30% of the total data pool held by the non-requester data holders is donated.

E Additional Empirical Results

We perform the same Ratio and Size experiments on AudioMNIST (Becker et al., 2024) to test whether the findings extend to one-dimensional signals. We use a full sampling rate of 4000 Hz (downsampled from 8000 Hz) and low-resolution rates of $\{240, 560, 1000, 1560\}$ Hz, with 240 Hz used in the Ratio experiments. To avoid overly high accuracy at low subsets, the AudioMNIST training set is reduced from $18,000$ to $5,000$. The results are reported in Figure 13.

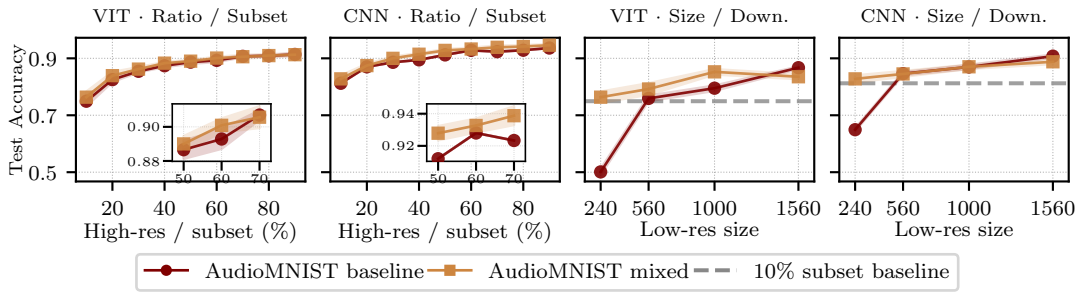

Figure 13: The experiments reported in Figure 3 for AudioMNIST.

Figure 14 shows the Grow experiment of Figure 5 with donor data added at full resolution, keeping all other settings, including the minority-first donation order, unchanged. Unlike in Figure 5 the ViT does not decline in balanced accuracy at larger requester subsets; for 100% requester data, the performance increases steadily with the addition of donor data. When training on a 10% requester subset, the performance still peaks at small donations and then decreases somewhat, suggesting that the shifted label distribution of the donors can outweigh the benefit of additional data. Since the donation order is identical in both experiments, class ordering alone cannot explain the decline observed in Figure 5. Together with the controlled experiments in Figure 3, which show that the ViT benefits from low-resolution data without distribution shift, this indicates that the degradation in Figure 5 arises from the combination of resolution and distribution shift rather than from either alone.

The CNN shows a similar but weaker early peak in the 10% setting, and is overall more robust to changes in both resolution and label distribution than the ViT.

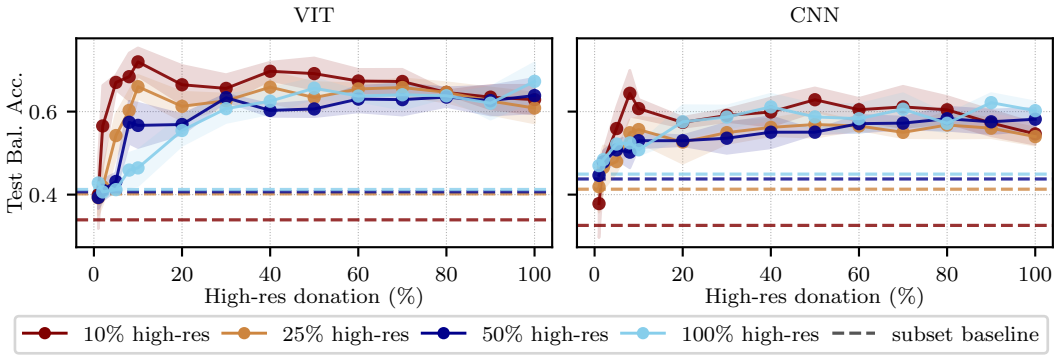

Figure 14: The experiment reported in Figure 5 but for full-resolution donor data (224×224). Everything else is the same including the donation order.

F Ablation Studies

F.1 Weighting of Low-resolution Loss

Since evaluation is performed exclusively on high-resolution inputs, we anneal the contribution of the low-resolution branch during training while keeping the high-resolution weight fixed at $w_{\text{high}} = 1$. We consider two schedules for weighting the low-resolution loss.

1-phase Cosine Schedule Let $p = e/E \in [0, 1]$ denote the training progress at epoch e out of a total of E epochs. We compute a cosine annealing coefficient w, which is used to update the low-resolution loss weight:

$$w = \tfrac{1}{2}\left(1 + \cos(\pi p)\right), \quad w_{\text{low}} = \text{ratio}_{\text{low}} \cdot (0.2 + 0.8\, w). \tag{86}$$

This schedule ensures that low-resolution samples initially contribute in proportion to their dataset share, after which their contribution decays smoothly to one fifth of that share.

2-phase Cosine Schedule Let $p = e/E \in [0, 1]$ denote the training progress at epoch e out of a total of E epochs.

The schedule consists of two phases with transition point $p^\star = 0.7$ and floor $a = 0.05$. The low-resolution weight is defined as

$$w_{\text{low}}(p) = \text{ratio}_{\text{low}} \cdot \begin{cases} a + (1-a)\,\frac{1}{2}(1 + \cos(\pi p/p^\star))\,, & 0 \le p \le p^\star, \\ a\,\frac{1}{2}(1 + \cos(\pi(p-p^\star)/(1-p^\star)))\,, & p^\star < p \le 1. \end{cases} \tag{87}$$

Thus, during the first 70% of training the low-resolution branch decays smoothly from its data-proportional value $\text{ratio}_{\text{low}}$ to a small floor $a \cdot \text{ratio}_{\text{low}}$, after which it is annealed continuously to

Table 2: Ablation study on the low-resolution weighting for selected ratio experiments on the CIFAR10 dataset. "Ours" refers to the results seen in 3 with the ViT using the schedule in equation 86 and the CNN the one seen in equation 87. Equal weighting means both high- and low-resolution losses are weighted with 1.

	ViT		**CNN**	
Ratio	Ours	Equal Weighting	Ours	Equal Weighting
10%	73.4 ± 0.6	69.2 ± 0.8	77.5 ± 0.9	75.1 ± 0.7
30%	81.5 ± 0.3	78.9 ± 0.6	87.9 ± 0.4	87.1 ± 0.4
50%	85.2 ± 0.4	84.4 ± 0.3	91.4 ± 0.3	90.9 ± 0.4
70%	87.3 ± 0.2	87.3 ± 0.5	93.0 ± 0.2	92.8 ± 0.1
90%	89.2 ± 0.3	89.1 ± 0.4	93.8 ± 0.2	93.8 ± 0.1

zero over the remaining 30%. This yields a smooth transition from joint mixed-resolution training to pure high-resolution fine-tuning.

To account for different numbers of high- and low-resolution mini-batches per epoch, we rescale both branch weights as

$$w_{\text{low}} \leftarrow w_{\text{low}} \cdot \frac{0.5}{n_{\text{low}}/(n_{\text{low}} + n_{\text{high}})}, \qquad w_{\text{high}} \leftarrow w_{\text{high}} \cdot \frac{0.5}{n_{\text{high}}/(n_{\text{low}} + n_{\text{high}})} \tag{88}$$

where n_{low} and n_{high} denote the numbers of low- and high-resolution mini-batches per epoch.

All experiments on CIFAR-100 use the 1-phase cosine schedule. On CIFAR-10 and AudioMNIST, CNN-based models performed better under the 2-phase schedule, whereas ViT-based models continued to use the 1-phase schedule.

Table 2 shows an ablation study on the CIFAR10 dataset on selected Ratio experiments. The schedule is compared to weighting the high- and low-resolution loss equally at a value of 1. The proposed schedules especially make a difference for the lower ratios. For ratios of 70% or 90%, high-resolution data already dominates and it thus matters less what the weighting of low-resolution loss is.

F.2 Scale Consistency Loss

Tian et al. (2023) propose a scale-consistency loss to encourage representations learned at different input resolutions to be aligned during training. Concretely, the scale-consistency loss is defined as a smooth-$l1$ loss applied to whitened feature representations. For ViTs, this loss operates on the `cls`-tokens corresponding to different resolutions:

$$\mathcal{L}_{\text{scale}} = \sum_{i=1}^{r-1} \mathcal{L}_{\ell_1}\big(\texttt{cls}_{i+1}, \texttt{cls}_i\big), \tag{89}$$

where r denotes the number of resolutions used during training, and the sum enforces consistency between adjacent resolution levels, with higher resolutions acting as teachers for lower ones. To assess whether this auxiliary loss is beneficial in our setting, we construct paired high- and low-resolution inputs by first performing a forward pass on the high-resolution samples, then downsampling these samples and passing them through the model again. This yields aligned representations at the two resolutions, to which the scale-consistency loss can be applied. We incorporate this into the loss function for selected Ratio experiments with a weighting coefficient of 0.5. For CNN architectures, the loss is computed analogously using the post-pooled feature vectors, ensuring comparability across model classes. The results are listed in Table 3. For both architectures the addition of this auxiliary loss does not yield any significant improvements.

F.3 Regularisation

To demonstrate that the low-resolution data acts as more than a mere regulariser, we perform some additional experiments on the CIFAR10 dataset with 10% high-resolution data. For both the Ratio

Table 3: Ablation study on the scale consistency loss proposed by Tian et al. (2023) on selected Ratio experiments on the CIFAR10 dataset. Loss is added with a coefficient of 0.5 and applied to the `cls`-token for the ViT experiments and the post-pooled feature vector for the CNN experiments. "Ours" refer to the results seen in Figure 3.

	ViT		**CNN**	
Ratio	Ours	W. Scale Consistency	Ours	W. Scale Consistency
10%	73.4 ± 0.6	71.0 ± 0.5	77.5 ± 0.9	77.7 ± 0.7
30%	81.5 ± 0.3	79.6 ± 0.5	87.9 ± 0.4	87.9 ± 0.2
50%	85.2 ± 0.4	84.1 ± 0.8	91.4 ± 0.3	91.3 ± 0.3
70%	87.3 ± 0.2	87.3 ± 0.4	93.0 ± 0.2	92.9 ± 0.1
90%	89.2 ± 0.3	89.3 ± 0.4	93.8 ± 0.2	93.9 ± 0.1

and Subset experiments, we sweep the weight decay coefficient over $\lambda \in \{10^{-3}, 5 \cdot 10^{-3}, 10^{-2}, 5 \cdot 10^{-2}, 10^{-1}, 0.25, 0.5, 0.75, 1, 1.5, 2, 2.5\}$. Results for both the ViT and CNN are shown in Figure 15. All results are averaged across five random seeds.

Two observations can be made. First, no weight decay setting in the Subset experiment matches the performance achieved by the best Ratio configuration. This suggests that the gains from incorporating low-resolution data cannot be explained solely by stronger regularisation through weight decay. Second, the ViT is considerably more sensitive to the choice of weight decay than the CNN. Performance improves rapidly as λ increases from the smallest values and peaks around $\lambda = 0.05$, after which it gradually declines. This sensitivity indicates that weight decay should ideally be tuned separately for each architecture and experimental setting.

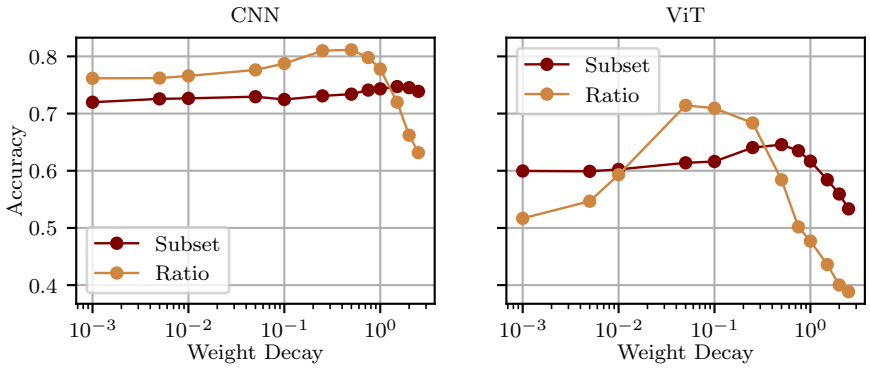

Figure 15: Ablation over weight decay values for the CIFAR10 dataset with 10% high resolution data.

F.4 Robustness to Corruptions

The preceding experiments assume that low-resolution observations differ from high-resolution observations only through downsampling. In practice, data collected or transmitted under resource constraints may also be affected by acquisition noise or blur. We therefore evaluate whether the benefit of donated low-resolution data persist under three corruptions: Gaussian blur, additive Gaussian noise, and Poisson noise. Each corruption is applied at mild and strong intensity prior to downsampling. Examples are shown in Figure 16.

We consider the Fed-ISIC-2019 setting in which the requester trains on 10% of its local data at high resolution and receives all available donor data at low resolution. Table 4 compares the corrupted conditions with both the high-resolution-only Subset baseline and the corresponding clean low-resolution donation setting. Corruption generally reduces the benefit of the donated data relative to clean low-resolution inputs. Nevertheless, every corrupted condition outperforms the high-resolution-

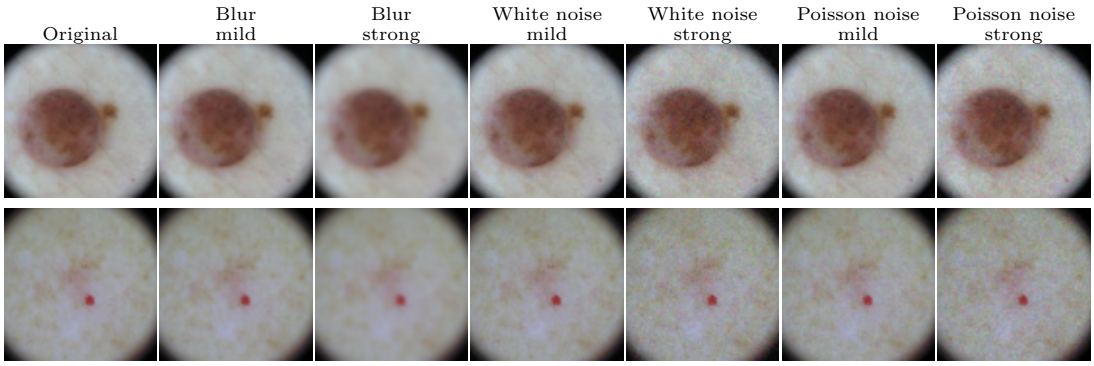

Figure 16: Examples of the acquisition corruptions applied before downsampling. Gaussian blur uses radii of 1.0 and 2.5 pixels for the mild and strong settings, respectively. Additive Gaussian noise uses standard deviations corresponding to 3% and 8% of the 8-bit intensity range. Poisson noise uses photon-count scales of 1.0 and 0.3, where a smaller scale produces stronger shot noise.

Condition	ViT	CNN
Subset baseline (HR only)	0.339	0.326
Clean donation (LR)	0.566	0.484
Gaussian noise (mild)	0.556	0.470
Gaussian noise (strong)	0.550	0.470
Gaussian Blur (mild)	0.581	0.475
Gaussian Blur (strong)	0.561	0.480
Poisson noise (mild)	0.564	0.481
Poisson noise (strong)	0.543	0.470

Table 4: Balanced accuracy on Fed-ISIC-2019 when the requester trains on 10% of its local high-resolution data and receives all donor data at low resolution. Corruptions are applied to donor images before downsampling. The high-resolution-only Subset baseline and clean low-resolution donation setting are included for comparison. Despite degradation relative to clean low-resolution data, all corrupted donation settings outperform the Subset baseline.

only baseline for both architectures. These results indicate that the gains from low-resolution donation are not restricted to ideal downsampling and remain present under moderate acquisition degradation.