LeJEPA for Audio: A Principled, Heuristic-Free Framework for Self-Supervised Audio Representation Learning

December 3, 2025

Ludovic TUNCAY
ludovic.tuncay@irit.fr
IRIT, Université de Toulouse,
CNRS, Toulouse INP,
Toulouse, France

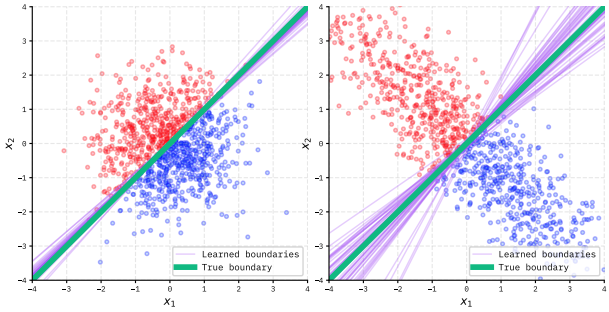
Abstract

Current Self-Supervised Learning (SSL) methods rely heavily on engineering heuristics to prevent representation collapse—such as negative sampling (Contrastive), asymmetric architectures with stop-gradients (SimSiam), or momentum encoders (BYOL, MoCo). The recent paper "LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics" (Balestriero & LeCun, 2025) introduces a theoretically grounded alternative. LeJEPA prevents collapse by enforcing a specific geometric structure on the latent space (an Isotropic Gaussian) using a novel regularization objective called SIGReg. This report analyzes the theoretical foundations of LeJEPA and proposes Audio-LeJEPA, an adaptation designed to learn robust audio representations from spectrograms without the complexity of predictor networks or momentum teachers.

1 – Introduction

The primary challenge in Joint-Embedding Architectures (JEAs) is representation collapse [1, 2, 3, 4]. This occurs when an encoder maps all inputs to a constant vector, which trivially minimizes the distance between positive pairs but destroys all semantic information.

Previous solutions introduced architectural asymmetries (Predictors [3], Momentum Teachers [4]) or objective constraints (Contrastive negatives [1], Variance-Invariance-Covariance [5]). LeJEPA [6] argues that these are ad-hoc mitigation strategies for a fundamental problem: the undefined latent geometry.



(a) Isotropic. (b) Anisotropic

Figure 1: Illustration of linear probing in Isotropic (a) versus Anisotropic (b) latent spaces. In this specific example, the linear separator shown in (a) exhibits low bias and variance. In contrast, the separator shown in (b) demonstrates high variance, being highly sensitive to the direction of the data distribution's elongation.

LeJEPA demonstrates that enforcing the embedding distribution to be an Isotropic Gaussian is necessary and sufficient to achieve three goals:

1. *Prevent collapse:* The variance is strictly non-zero.
2. *Ensure feature independence:* The axes of the latent space are de-correlated, maximizing efficiency.
3. *Minimize downstream risk:* The conditioning of the representation is provably optimal for linear probing.

Specifically, as illustrated in Figure 1, Balestriero & LeCun demonstrate that anisotropy amplifies estimator bias and variance, thereby necessitating an isotropic distribution. Furthermore, they establish that the standard Gaussian is the unique optimum for minimizing risk in non-linear probing tasks.

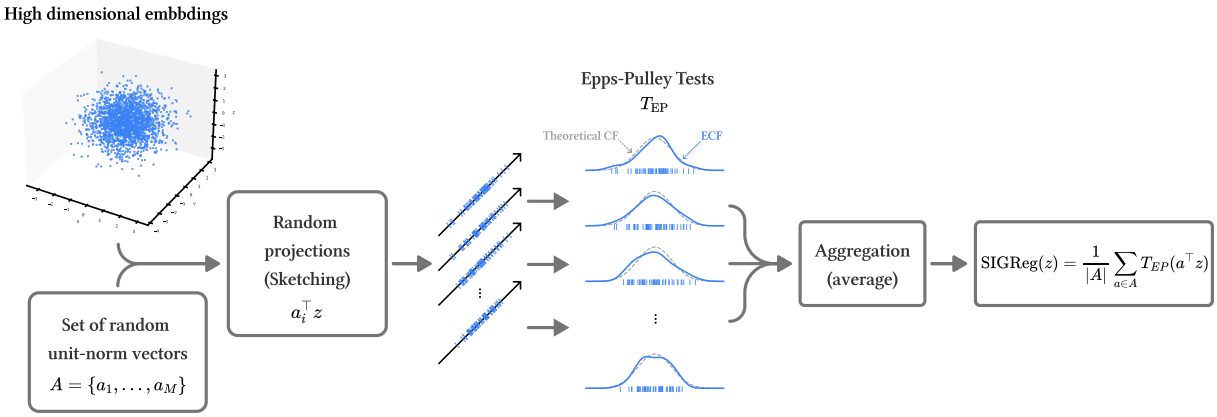


Figure 2: Schematic overview of the SIGReg process. The high-dimensional embeddings Z are projected via a random matrix A onto 1D lines. A statistical test (Epps-Pulley) is then applied to each projection to measure divergence from normality.

By explicitly regulating the geometry via SIGReg, LeJEPA removes the need for predictors, stop-gradients, and teacher networks. This results in a purely symmetric, computationally efficient architecture that is easier to train and scale.

2 – The LeJEPA Framework

2.1 – The Core Objective

LeJEPA trains a single encoder f_θ using a loss function composed of two antagonistic terms:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{\text{Sim}} + \lambda\mathcal{L}_{\text{SIGReg}}$$

The **Similarity Loss** ($\mathcal{L}_{\text{Sim}}$) enforces semantic consistency. It minimizes the Euclidean distance between embeddings of different views (x_1, x_2) of the same input. Unlike heuristic-based methods (e.g., BYOL, SimSiam), LeJEPA employs no predictor network, relying instead on a simple metric distance:

$$\mathcal{L}_{\text{Sim}} = \|f_\theta(x_1) - f_\theta(x_2)\|_2^2$$

The **Regularization Loss** ($\mathcal{L}_{\text{SIGReg}}$) enforces the geometric constraint. It ensures that the batch of embeddings converges to a standard normal distribution $\mathcal{N}(0, I)$, maximizing information capacity and minimizing correlation.

2.2 – Sketched Isotropic Gaussian Regularization (SIGReg)

Calculating the divergence between a high-dimensional embedding distribution and a target Gaussian is computationally intractable due to the "Curse of Dimensionality." LeJEPA circumvents this by leveraging the **Cramér-Wold Theorem** [7, 8, 9], which states that a high-dimensional distribution is Gaussian if and only if all its 1-dimensional projections are Gaussian.

The SIGReg algorithm proceeds in three steps (see Figure Figure 2):

1. *Random Projections (Sketching):* A random matrix A (size $K \times M$, K being the embedding dimension and M the number of projections) is generated where columns are drawn uniformly from the unit hypersphere ($\mathcal{S}^{K-1}$). The batch of embeddings Z is projected onto A: $P = Z \cdot A$.
2. *1D Statistical Test:* A differentiable test of normality is applied to the projected values. LeJEPA utilizes the Epps-Pulley Test [10, 11].
3. *Aggregation:* The final loss is the average test statistic computed across all K projections.

2.3 – The Epps-Pulley Test

The Epps-Pulley test [10, 11] is selected for its differentiability and bounded gradients. It operates in the frequency domain by comparing the Empirical Characteristic Function (ECF) of the projected data against the theoretical Characteristic Function (CF) of a standard Gaussian ($\phi(t) = e^{-t^2/2}$).

The statistic is defined as the weighted squared integral of the difference between these two functions:

$$T_{\text{EP}} = \int \left|\phi_{\text{empirical}}(t) - \phi_{\text{target}}(t)\right|^2 w(t)dt$$

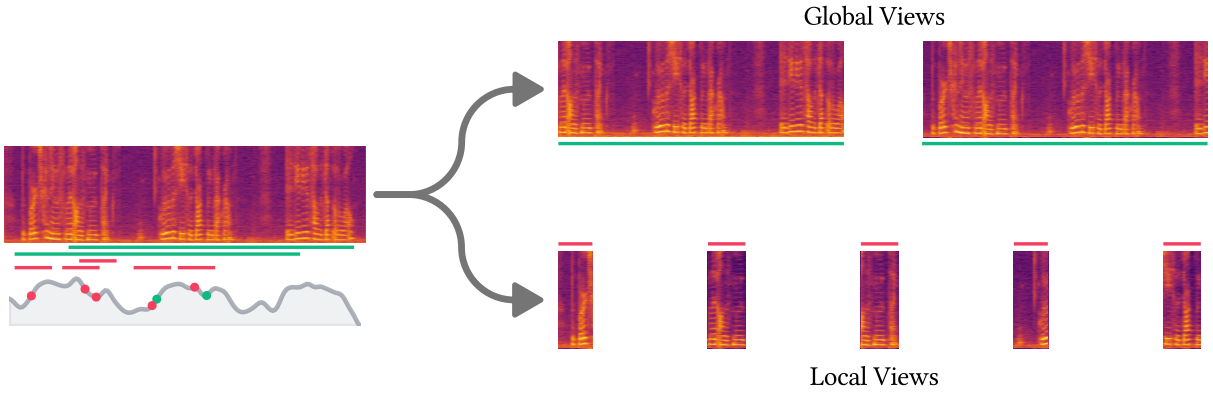


Figure 3: Schematic of the **Time-Domain Multi-Crop strategy** with **Energy-Weighted Sampling**. Unlike visual cropping, the full frequency axis is preserved. Global views (green bars) span the majority of the duration to capture semantic context. Local views (red bars) target transient acoustic events. As illustrated by the dots on the energy profile (bottom left), local view centers are sampled proportionally to signal energy to avoid training on silence.

The `num_points` parameter: In the implementation, this integral is approximated using a trapezoidal quadrature rule. The parameter `num_points` defines the resolution of this approximation:

- It represents the number of discrete nodes t sampled along the integration domain.
- A low `num_points` (e.g., 5) results in a coarse approximation, constraining mostly low-frequency components (mean, variance).
- A high `num_points` (e.g., 50) enforces matching of higher-order moments (skewness, kurtosis).
- The authors demonstrate that `num_points` $= 17$ is empirically sufficient to capture the Gaussian shape efficiently.

3 – Proposal: Audio-LeJEPA

We propose transposing LeJEPA to the audio domain to create a streamlined foundation model. This requires adapting the architecture and, crucially, the view generation strategy to suit the spectral properties of audio.

3.1 – Architecture Design

Unlike existing Audio-JEPA [12, 13] or BYOL-A [14, 15] approaches, Audio-LeJEPA removes the need for a "Target Encoder" (EMA teacher) and the "Predictor" network.

- *Backbone:* A standard Vision Transformer (e.g., ViT-B/16 [16]) or Conformer [17] processing Log-Mel Spectrograms.
- *Projector:* A 3-layer MLP mapping the backbone output to the embedding space (e.g., dim=256), identical to the vision implementation of LeJEPA.
- *Symmetry:* The architecture is fully symmetric; the exact same encoder weights θ are used to process all views.

3.2 – View Generation Strategy

Standard SSL methods in computer vision rely on random spatial cropping. However, applying 2D cropping to audio spectrograms is often destructive, as the frequency axis carries critical structural information (pitch and timbre) that must be preserved.

To address this, we propose a **Time-Domain Multi-Crop** strategy that operates exclusively along the temporal axis while ensuring acoustic relevance.

As illustrated in Figure 3, we generate two types of views from a single audio clip:

1. *Global Views (Context):*
   - **Objective:** Capture the long-term semantic context of the audio (e.g., the genre of a song or the identity of a speaker).
   - **Mechanism:** We extract long segments covering the majority of the clip (e.g., 80-100% or 8-10 seconds).
   - **Augmentation:** To prevent the model from relying on trivial artifacts, we apply time-masking and mild frequency masking. We explicitly avoid pitch-shifting, as this can alter class identity in many audio domains.

2. *Local Views (Events):*
   - **Objective:** Force the model to predict the global context from short, local fragments (e.g., a single word or chord).
   - **Mechanism:** We extract short, disjoint segments (e.g., 20% or 2 seconds of the total duration).
   - **Constraint:** To preserve timbre, crops cover the full frequency range and only slice along the time axis.

3.2.1 – Energy-Weighted Sampling

A major challenge in audio SSL is the prevalence of silence or background noise. Randomly cropping a "local view" might result in a segment containing no signal, causing the model to learn nothing.

To solve this, we introduce **Energy-Weighted Sampling** for local views:

1. We compute the energy profile E_t of the spectrogram across time frames.
2. We sample a center frame t_c with probability proportional to its energy:
$$\mathbb{P}(t_c) \propto E_{t_c}$$
3. We extract a window of fixed duration $L = 2w$ around this center. To accommodate boundary conditions, we enforce a validity constraint: if the theoretical window extends beyond the signal limits $[0, T]$, the start index is clamped to the range $[0, T - L]$. This ensures the crop remains fully within the valid audio duration without introducing padding artifacts.

This ensures that local views are centered on salient acoustic events (e.g., speech onsets, musical notes) rather than silence.

3.2.2 – Alternative Strategy: Patch Masking

While Multi-Crop is our primary proposal, LeJEPA is also compatible with Masked AutoEncoder (MAE) [18, 19] style strategies. In this variant, "Global Views" are the full spectrogram with a small percentage of patches masked (e.g., 20%), while "Local Views" are heavily masked versions (e.g., 75% masked) of the same input. This trades the complexity of crop generation for the computational efficiency of processing sparse inputs.

```python
class AudioMultiCrop:
  def __init__(
    self,
    global_scale=(0.8, 1.0),
    n_global=2,
    local_scale=(0.2, 0.4),
    n_local=8,
  ):
    self.global_trans = Compose( ... )
    self.local_trans = Compose( ... )

  def __call__(self, spec):
    views = []
    # Generate Global Views
    for _ in range(self.n_global):
      views.append(
        self.global_trans(spec)
      )
    # Generate Local Views
    for _ in range(self.n_local):
      views.append(
        self.local_trans(spec)
      )
    return views
```

Listing 1: Pseudo Python code of the view generation for audio. Input is a single spectrogram; output is a list of views.

4 – Implementation Roadmap

To adapt the official LeJEPA implementation, the following modifications are required:

4.1 – Data Loader and Transform

The ImageNet loader must be replaced with an audio loader (e.g., `torchaudio` [20, 21]). The augmentation pipeline follows the Multi-Crop logic:

4.2 – Training Loop

Let x be a batch of inputs transformed into V views, consisting of V_g global views and V_l local views. The training process minimizes a joint objective at every step:

1. **Forward Pass:** All views are passed through the shared encoder f_θ to obtain embeddings $Z = \{z_v\}_{v=1}^{V}$.
2. **Target Computation:** The regression target is defined as the centroid of the global views. For a sample n, the target is:
$$\mu_n = \frac{1}{V_g} \sum_{v=1}^{V_g} z_{n,v}$$

3. **Similarity Loss:** The model minimizes the distance between **every** view (global and local) and the global centroid μ_n. This enforces that local fragments align with the global context:

$$\mathcal{L}_{\text{Sim}} = \frac{1}{V} \sum_{v=1}^{V} \left\| \mu_n - z_{n,v} \right\|_2^2$$

4. **SIGReg Loss:** The geometric regularization is applied to the batch of embeddings for each view independently to prevent collapse.

$$\mathcal{L}_{\text{SIGReg}} = \frac{1}{V} \sum_{v=1}^{V} \text{SIGReg}(z_v)$$

5. **Total Objective:** The final loss is a weighted sum, backpropagated through f_θ:

$$\mathcal{L}_{\text{Total}} = (1 - \lambda)\mathcal{L}_{\text{Sim}} + \lambda \mathcal{L}_{\text{SIGReg}}$$

4.3 – Hyperparameters

Based on the stability analysis in the LeJEPA paper, we recommend the following starting configuration:

- λ: 0.05 (Recommended default for balancing prediction and regularization).
- Batch Size: ≥ 128 (Required for stable statistical estimation in SIGReg).
- Projection Heads: 1024 slices for SIGReg.
- Epps-Pulley `num_points`: 17.

5 – Conclusion

Adopting LeJEPA for audio offers a unique opportunity to simplify the current "zoo" of Audio SSL methods. By replacing the fragile tuning of EMA schedules and predictor architectures with a rigorous geometric regularization (SIGReg), we can achieve robust, general-purpose audio representations. The proposed "Time-Domain Multi-Crop" strategy effectively translates the spatial invariance of vision models into temporal invariance suitable for audio, ensuring the model learns to associate local transient events with long-term semantic contexts.

Bibliography

[1] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision."

[2] X. Chen and K. He, "Exploring Simple Siamese Representation Learning."

[3] J.-B. Grill *et al.*, "Bootstrap your own latent: A new approach to self-supervised Learning."

[4] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning."

[5] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning."

[6] R. Balestriero and Y. LeCun, "LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics."

[7] H. Cramér and H. Wold, "Some Theorems on Distribution Functions," *Journal of the London Mathematical Society*, no. 4, pp. 290–294, Oct. 1936,

[8] M. Samanta, "Non-parametric estimation of conditional quantiles," *Statistics & Probability Letters*, vol. 7, no. 5, pp. 407–412, Apr. 1989,

[9] J. A. Cuesta-Albertos, R. Fraiman, and T. Ransford, "A Sharp Form of the Cramér–Wold Theorem," *Journal of Theoretical Probability*, vol. 20, no. 2, pp. 201–209, June 2007,

[10] T. W. Epps and L. B. Pulley, "A Test for Normality Based on the Empirical Characteristic Function," *Biometrika*, vol. 70, no. 3, pp. 723–726, 1983,

[11] B. Ebner and N. Henze, "Bahadur Efficiencies of the Epps–Pulley Test for Normality," *Journal of Mathematical Sciences*, vol. 273, no. 5, pp. 861–870, July 2023,

[12] L. Tuncay, E. Labbé, E. Benetos, and T. Pellegrini, "Audio-JEPA: Joint-Embedding Predictive Architecture for Audio Representation Learning," in *2025 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, Nantes, France, June 2025, pp. 1–5.

[13] Z. Fei, M. Fan, and J. Huang, "A-JEPA: Joint-Embedding Predictive Architecture Can Listen."

[14] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "BYOL for Audio: Self-Supervised Learning for General-Purpose Audio Representation," in *2021 International Joint Conference on Neural Networks (IJCNN)*, July 2021, pp. 1–8.

[15] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "BYOL for Audio: Exploring Pre-Trained General-Purpose Audio Representations," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 31, pp. 137–151, Nov. 2022,

[16] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale."

[17] A. Gulati *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition."

[18] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners."

[19] P.-Y. Huang *et al.*, "Masked Autoencoders that Listen."

[20] Y.-Y. Yang *et al.*, "TorchAudio: Building Blocks for Audio and Speech Processing."

[21] J. Hwang *et al.*, "TorchAudio 2.1: Advancing speech recognition, self-supervised learning, and audio processing components for PyTorch."