KERJEPA: KERNEL DISCREPANCIES FOR EUCLIDEAN SELF-SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Breakthroughs in self-supervised Joint-Embedding Predictive Architectures (JEPAs) have established that regularizing Euclidean representations toward isotropic Gaussian priors yields provable gains in training stability and downstream generalization. We introduce a new, flexible family of Euclidean joint embedding self-supervised learning algorithms with kernel-based regularizers. One instance of this family corresponds to the SIGReg Epps-Pulley regularizer which approximates a sliced maximum mean discrepancy (MMD) with a Gaussian prior and Gaussian kernel. By expanding the class of viable kernels and priors and computing the closed-form high-dimensional limit of sliced discrepancies, we develop alternative methods with improved training stability and design flexibility and leverage those techniques to train strong vision learners at scale.

1 INTRODUCTION

Self-supervised learning (SSL) methods aim to learn useful, general representations from unlabeled data by learning the underlying structure of the data itself. The promise of these methods is that the representations serve as strong features for arbitrary downstream tasks, enabling the construction of foundation models. In this work, we study how the distribution imposed on Euclidean representations can be selected and estimated in a principled manner for visual data, allowing practitioners to make informed algorithm design decisions.

A common approach to SSL, particularly in visual domains, is to project data points, such as images, to various *views*, such as random crops, and learn embeddings that map views of the same image to similar representations. A pitfall of this framework is that it admits a trivial, pathological solution that maps every input to the same embedding. To mitigate this issue, *regularization* is crucial: by regularizing the distribution of the learned embeddings toward a fixed, user-designed *prior* distribution, representation collapse can be avoided. However, regularizing such high-dimensional distributions remains a sincere challenge, both in principle and in practice.

Historically, SSL methods have largely constrained embeddings to a hypersphere (Chen et al., 2020; Wang & Isola, 2020), a choice that stabilizes training and induces bounded gradients. More recently, there has been increasing interest in moving off the hypersphere toward learning Euclidean representations (Bardes et al., 2022; Balestriero & LeCun, 2025). In particular, LeJEPA (Balestriero & LeCun, 2025) argues that an isotropic Gaussian distribution over embeddings is optimal for downstream regression and classification tasks with weight decay, a geometry that cannot be achieved by hyperspherical representations. To induce this distribution, LeJEPA introduces SIGReg, a regularizer that compares randomly sampled univariate projections of the embedding distribution with univariate Gaussians. This construction is principled, computationally attractive, and empirically effective.

Despite these successes, the reliance on stochastic projections introduces a fundamental tension between estimation variance and computational cost. The finite approximation of the expectation over directions introduces gradient noise, while the spectral integral within each projected discrepancy is itself approximated by numerical quadrature. Increasing the number of projections reduces the first source of error but eventually undermines the linear computational scaling that slicing is intended to provide in the standard SSL regime. Current sliced approaches specialize to Gaussian priors with a fixed Gaussian kernel, motivated by a result that isotropically-Gaussian-distributed features are optimal for downstream ℓ_2-regularized tasks. Unfortunately, this ignores the distribution shift between the regularized features and those used for transfer in SSL pipelines. This rigidity limits

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.

the exploration of alternative regularizers, non-Gaussian priors, and heavy-tailed kernels which may offer superior geometric properties for downstream tasks.

In this work, we establish an equivalence between SIGReg and a specific kernel discrepancy; generalizing the kernel and the discrepancy reveals a broader family of distributional regularizers for self-supervised learning. This perspective enables normalization-free representation learning, while providing a principled framework to study alternative discrepancy constructions and their computational and statistical tradeoffs. In particular, we investigate finite-direction slicing and analytically integrated estimators, and apply such methods to larger scale datasets. Our contributions are summarized as follows:

- **Theoretical foundation.** We show that SIGReg is a special case of a broader family of kernel discrepancies, providing a unified perspective on distributional regularization in SSL.
- **Regularizer design space.** We derive algorithms induced by alternative estimators for this family of regularizers, comparing statistical and computational tradeoffs at scale.
- **Scalable training.** We achieve strong ImageNet performance with scalable student-teacher vision transformers, outperforming comparable models trained on identical data scales.

2 Related Works

While the broader goal of SSL is to learn augmentation-invariant representations, methods diverge in how they prevent representation collapse. *Contrastive* methods, such as SimCLR (Chen et al., 2020) and MoCo (He et al., 2020), avoid collapse by explicitly pushing apart the representations of distinct samples (negative pairs) while attracting positive pairs. *Non-contrastive* methods eliminate the need for negative pairs, relying instead on architectural asymmetries or statistical constraints. For instance, BYOL (Grill et al., 2020) utilizes asymmetric stop-gradient mechanisms, while Barlow Twins (Zbontar et al., 2021) and VICReg (Bardes et al., 2022) impose variance-preservation and de-correlation objectives on the embedding dimensions.

Crucial to all these frameworks is the design of data augmentations, which implicitly defines which features are learned. Techniques such as cropping, color jittering, and Gaussian blur encourage the network to discard low-level visual statistics (Moutakanni et al., 2024), while advanced strategies like multicrop (Caron et al., 2020) enforce consistency across varying spatial scales. Equally important is the standard backbone-projector architectures. Empirical success in SSL heavily relies on guillotine regularization, where the non-linear projection head is discarded after pretraining, while only the backbone features are used for downstream tasks (Bordes et al., 2023). This structural decoupling implies that the training objective, which is applied to the projections, does not directly constrain the geometry of the representations used for transfer. This misalignment draws attention to the benefits of selecting a particular geometry throughout training.

A dominant paradigm in SSL involves normalizing embeddings to the unit hypersphere, effectively fixing the scale of representation to focus learning entirely on angular relationships. This geometric constraint naturally leads to objectives that encourage the uniform distribution of features over the sphere, thereby maximizing the information entropy of the induced distribution (Wang & Isola, 2020). More sophisticated distribution matching has been approached through optimal transport, specifically using the Spherical Sliced Wasserstein (SSW) distance (Bonet et al., 2023), which leverages random projections to align the empirical distribution of embeddings with a uniform target. Similarly, kernel-based methods have been applied via Hilbert-Schmidt Independence Criterion (Li et al., 2021). These works demonstrated the utility of both slicing and kernel discrepancies, providing a foundation for their extension into other geometries.

The recent shift toward isotropic Gaussian targets in Euclidean space, exemplified by LeJEPA (Balestriero & LeCun, 2025) which builds on VICReg (Bardes et al., 2022; Sepanj & Fieguth, 2025), bridges SSL with kernel-based methods, without explicit normalization. Although LeJEPA formulates its regularizer via the *Epps-Pulley test* (Epps & Pulley, 1983) on random projections, recent analyses establish that this is equivalent to minimizing the Sliced Maximum Mean Discrepancy between the embeddings and a target Gaussian (Sriperumbudur et al., 2010; Chwialkowski et al., 2016). Based on this categorization, there is a plethora of novel discrepancies and estimators waiting to be explored (Liu et al., 2016; Korba et al., 2021; Gong et al., 2021), which we investigate in the sequel.

3 BACKGROUND

This paper studies a broad family of self-supervised learning algorithms, by carefully treating the problem of regularizing learned representations via distributional similarity estimation. In SSL, the goal is to learn an embedding function $f_\theta : \mathcal{X} \to \mathcal{Z}$ which maps inputs from space $\mathcal{X}$ (say, high-dimensional images or videos) to a (lower dimensional) latent space $\mathcal{Z} \subseteq \mathbb{R}^d$, such that the representations enable efficient downstream processing (for instance, accurate linear classification).

Leveraging known invariance structure of the data, such a feat can readily be accomplished. Suppose $\mathcal{T}$ is an augmentation map, which maps any input x to a distribution over *views* of x, which could be random crops or rotations that representations should be invariant to. Then SSL can be accomplished by minimizing a squared loss between predicted representations perturbed from the same source input, subject to regularization of the population of predicted representations *across* source inputs relative to a fixed target. This is encapsulated by the following loss $\mathcal{L}$, where $\{X_l\}_{l=1}^n \subseteq \mathcal{X}$ is a minibatch of inputs, and m is the number of perturbations applied to each source input,

$$\mathcal{L}(\theta) \triangleq \frac{1}{dnm^2} \sum_{l=1}^{n} \sum_{i,j=1}^{m} \|f_\theta(V_l^i) - f_\theta(V_l^j)\|_2^2 + \lambda \Omega(\widehat{P}, Q), \ \widehat{P} \triangleq \frac{1}{nm} \sum_{l=1}^{n} \sum_{i=1}^{m} \delta_{f_\theta(V_l^i)} \quad (1)$$

where each V_l^i is sampled independently from $\mathcal{T}(X_l)$. In (1), Q is a fixed *prior* over latent variables (say, a uniform distribution in the case of bounded latent space, or a Gaussian distribution), $\widehat{Z}$ is an empirical distribution of latents output by the model f_θ, and Ω is a regularization between latent distributions minimized when the inputs are equal. Notably, this regularization term is crucial: without it, (1) could be minimized by trivial solutions such as constant f_θ.

Designing a powerful and efficient regularizer remains an active direction of research in SSL. Traditionally, SSL methods have constrained latent representations to the sphere $\mathcal{Z} = \mathbb{S}^{d-1}$, which benefits from enforcing bounded representations useful for downstream application, but suffers from complications during training. In response to these complications, Balestriero & LeCun (2025) introduced the *LeJEPA* instantiation of (1) for Euclidean representations $\mathcal{Z} = \mathbb{R}^d$. This instantiation notably introduces the regularizer $\Omega = \mathsf{SIGReg}$ and Gaussian prior $Q = \mathcal{N}(0, \sigma^2)$, citing various theoretical justifications and practical simplicity and effectiveness.

The SIGReg regularizer breaks the curse of dimensionality involved in estimating distributional distance in $\mathscr{P}(\mathbb{R}^d)$ by computing $\Theta(1)$ random scalar projections of $\widehat{Z}$ and computing the *Epps-Pulley* test statistic via quadrature of an integral on the real line; a principled method with provable statistical guarantees. In this paper, we argue that SIGReg **is not as fundamental as it may seem**. Rather, it is built on various building blocks (namely, projection, Epps-Pulley, and quadrature) which are merely a particular means of computing the special case of a much more general class of regularizers. The remainder of the paper will delve into this more general class, providing efficient algorithms that generalize LeJEPA, and offering computational and practical benefits. The remainder of this section discusses kernel-based distance analysis methods, and the building block of LeJEPA.

3.1 MAXIMUM MEAN DISCREPANCY AND KERNEL STEIN DISCREPANCY

At the core of our investigation in this paper is the *Maximum Mean Discrepancy* (Gretton et al., 2012, MMD), a kernel-based distance between distributions. Given a kernel $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ measuring similarity between *samples*, the MMD between distributions P, Q is given by

$$\mathsf{MMD}_k^2(P, Q) = \mathbb{E}_{X,X' \sim P}[k(X, X')] + \mathbb{E}_{Y,Y' \sim Q}[k(Y, Y')] - 2\mathbb{E}_{X \sim P, Y \sim Q}[k(X, Y)]. \quad (2)$$

The choice of kernel determines a notion of similarity between embeddings. For particular *characteristic* kernels, such as the Gaussian kernel, $\mathsf{MMD}_k(P, Q)$ vanishes exactly when $P = Q$ (Gretton et al., 2012; Sriperumbudur et al., 2010). Common examples of characteristic kernels include Gaussian kernels k_{gsn} and inverse multi-quadric kernels $k_{\text{imq}}^{\alpha,\beta}$, given by

$$k_{\text{gsn}}(x, y) \triangleq \exp(-\gamma \|x - y\|_2^2) \quad \text{and} \quad k_{\text{imq}}^{\alpha,\beta}(x, y) \triangleq (1 + \alpha \|x - y\|_2^2)^{-\beta}$$

with $\gamma, \alpha, \beta > 0$. Notably, (2) admits an unbiased estimator by drawing samples from both distributions and averaging kernel evaluations between sample pairs.

The *Kernel Stein Discrepancy* (KSD) is an alternative kernel-based distance between distributions, defined through the target score $s_Q(x) = \nabla_x \log q(x)$ (Chwialkowski et al., 2016; Liu et al., 2016;

Gorham & Mackey, 2017). For a base kernel k, comparable to the kernel discussed for the MMD, the KSD constructs a kernel k_Q through the *Stein operator*, and is given by

$$\begin{aligned}\mathsf{KSD}_k^2(P,Q) &\triangleq \mathbb{E}_{X,X'\sim P}\left[k_Q(X,X')\right], \quad \text{where} \\ k_Q(x,y) &\triangleq s_Q(x)^\top k(x,y) s_Q(y) + s_Q(x)^\top \nabla_y k(x,y) + \nabla_x k(x,y)^\top s_Q(y) + \operatorname{tr}(\nabla_x \nabla_y^\top k(x,y)).\end{aligned} \tag{3}$$

By Stein's identity, KSD is equivalent to MMD with a target-dependent kernel, where the expectation is taken only under P and evaluations of the target score (Liu et al., 2016; Barp et al., 2024).

3.2 Spectral Representations of Discrepancies

Shift-invariant kernels admit equivalent frequency-domain representations via Bochner's Theorem (Rasmussen & Williams, 2005). Let $\rho_k(\omega)$ denote the spectral density associated with k and let $\phi_Q(\omega)$ denote the characteristic function (CF) of a known distribution Q. Given samples $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$, and the empirical distribution $\widehat{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$, MMD and KSD are written using the empirical characteristic function (ECF) defined as $\widehat{\phi}_n(\omega) = n^{-1}\sum_{i=1}^n \exp(\mathrm{i}\omega^\top X_i)$ and its score-weighted counterpart written as $\widehat{\phi}_{s_Q,n}(\omega) = n^{-1}\sum_{i=1}^n s_Q(X_i)\exp(\mathrm{i}\omega^\top X_i)$. Using these two empirical spectral features, MMD and KSD take the parallel forms

$$\begin{aligned}\mathsf{MMD}_k^2(\widehat{P}_n, Q) &\triangleq \textstyle\int_{\mathbb{R}^d} |\widehat{\phi}_n(\omega) - \phi_Q(\omega)|^2 \rho_k(\omega)\, \mathrm{d}\omega, \\ \mathsf{KSD}_k^2(\widehat{P}_n, Q) &\triangleq \textstyle\int_{\mathbb{R}^d} \|\widehat{\phi}_{s_Q,n}(\omega) + \mathrm{i}\omega\widehat{\phi}_n(\omega)\|^2 \rho_k(\omega)\, \mathrm{d}\omega.\end{aligned}$$

These spectral representations are useful because the frequency integrals can be approximated with deterministic quadrature or Monte Carlo frequencies, including Random Fourier Features, without pairwise kernel evaluations (Rahimi & Recht, 2007; Zhao & Meng, 2015; Huggins & Mackey, 2018).

3.3 Sliced Discrepancies

Certain distances between multi-dimensional distributions are computationally or statistically difficult to estimate, succumbing to the curse of dimensionality. The spectral representations of discrepancies in §3.2, for instance, require integrating over $\mathbb{R}^d$, which is exponentially expensive to estimate.

The concept of *slicing* or random projection can alleviate several computational challenges arising from the sample-based estimation of multivariate random variable discrepancies. Explicitly, such approaches leverage the celebrated *Cramér-Wold* theorem, which asserts that $\xi^\top X$ has the same distribution as $\xi^\top Y$ for any direction $\xi \in \mathbb{S}^{d-1}$ whenever X has the same distribution as Y. This motivates the notion of *sliced divergences*, which lift divergences on univariate distributions to multivariate distributions through the *slice operator* $\mathscr{S}$, given by $(\mathscr{S}\mathsf{d})(P,Q) \triangleq \mathbb{E}_{\xi\sim\mathrm{Unif}(\mathbb{S}^{d-1})}[\mathsf{d}(\xi_\#^\top P, \xi_\#^\top Q)]$.

As an immediate note, whenever a metric d admits an unbiased sample estimator, $\mathscr{S}\mathsf{d}$ does as well: simply sample m directions ξ independently and uniformly on $\mathbb{S}^{d-1}$, and then average the unbiased estimates $\widehat{\mathsf{d}}(\xi_\#^\top P)$ of $\mathsf{d}(\xi_\#^\top P, \xi_\#^\top Q)$. For a base divergence d with sample complexity $\beta(d,n)$ for n samples drawn from the probability distributions under comparison, this Monte-Carlo sliced distance estimator has sample complexity bounded (Nadjahi et al., 2020, Corollary S4) by

$$\beta(d,n) + \sqrt{\frac{\mathbb{E}\left[\mathrm{Var}\left\{\widehat{\mathsf{d}}(\xi_\#^\top P)\right\}\right]}{m}}.$$

3.4 The Epps-Pulley Test and LeJEPA

Having defined several forms of distribution distances, the SIGReg regularizer from LeJEPA (Balestriero & LeCun, 2025) can be formulated. SIGReg is built upon a quadrature estimate of the *Epps-Pulley* method (Epps & Pulley, 1983), given by

$$\mathsf{EP}(P; \mathcal{N}(0,\sigma^2)) \triangleq \textstyle\int_{\mathbb{R}} |\phi_P - \phi_{\mathcal{N}}|^2 \rho(\omega)\, \mathrm{d}\omega,$$

where $\phi_{\mathcal{N}}$ is the ECF of $\mathcal{N}(0,\sigma^2)$ for some $\sigma > 0$, and ρ is the spectral density of a Gaussian kernel with bandwidth $\gamma > 0$. Notably, EP is an instance of the MMD with a univariate Gaussian kernel. Consequently, multidimensional extensions can be made by employing MMD estimators with a

multivariate Gaussian kernel as well; a notable example is the BHEP test (Baringhaus & Henze, 1988). In effect, SIGReg is the EP specialized to the target $Q = \mathcal{N}(0, \sigma^2 I_d)$, given by

$$\Omega(P, \mathcal{N}(0, \sigma^2)) \triangleq \mathsf{SIGReg}(P) = (\mathscr{S}\mathsf{EP})(P; \mathcal{N}(0, \sigma^2 I_d)).$$

In practice, LeJEPA estimates SIGReg via Monte Carlo over the slicing operation as discussed in §3.3, and with quadrature on the 1-dimensional EP computations.

Notationally, it will be convenient to easily distinguish regularizers from their estimators; as such, regularizers will be written in sans-serif, while estimators are written in `monospace`. As such, `SIGReg` will refer to the instantiation of the *estimator* for SIGReg in LeJEPA, constructed by performing quadrature estimation with a finite set of anchor points.

4 General Euclidean Regularization by Distribution Matching

As shown §3.4, the SIGReg regularizer in LeJEPA defines a metric between feature distributions for preventing feature collapse, but also prescribes a *method* for estimating this metric, notably with slicing and quadrature estimation. Recognizing this distinction removes layers of complexity due to the estimation method imposed by LeJEPA, unveiling a simple and familiar regularizer at the core.

Theorem 1. *Let* SIGReg *denote the SIGReg regularizer (Balestriero & LeCun, 2025); that is*

$$\mathsf{SIGReg}(P) \triangleq \mathbb{E}_{\xi \sim \mathrm{Unif}(\mathbb{S}^{d-1})}[n^{-1}\mathsf{EP}(\xi_{\#}^{\top} P)], \quad P \in \mathscr{P}(\mathbb{R}^d).$$

It holds that SIGReg *is equivalent to a particular MMD,*

$$\mathsf{SIGReg}(P) = \mathsf{MMD}^2_{\kappa_d}(P, \mathcal{N}(0, \sigma^2 I_d)) \quad \textit{where} \quad \kappa_d(x, y) \triangleq {}_1F_1\left(\tfrac{1}{2}; \tfrac{d}{2}; -\gamma\|x - y\|_2^2\right),$$

and ${}_1F_1$ *is the* Kummer Confluent Hypergeometric function. *In particular,* κ_d *has polynomially decaying tails (as opposed to exponentially decaying tails like* k_{gsn}*).* [Proof]

Thm. 1 sheds light on the induced *geometry* imposed on the latent embeddings by the slicing operation. Notably, it imposes a *heavy tailed* similarity measure between latent embeddings, as opposed to commonly used subgaussian-tailed kernels such as k_{gsn}. Additionally, the equivalent kernel induced by slicing is *dimension-dependent*: as we scale the dimension, the slicing operation effectively decreases the *bandwidth* of the resulting kernel. As a result, it is important to consider both the choice of kernel and the use of slicing when designing distributional regularizers in high dimensional settings. The geometry induced by slicing motivates direct use of the IMQ kernel as an explicit heavy-tailed alternative in the ambient space without any additional slicing, while its dimension-dependent bandwidth suggests a simple rescaling may be sufficient. We therefore examine both sliced and non-sliced discrepancies with Gaussian and IMQ kernels experimentally, allowing us to assess the true importance of kernel tail behavior in the context of a self-supervised regularizer.

Next, we turn our attention back to estimation of the regularizer. Since Thm. 1 effectively reduces SIGReg to a particular MMD, we may turn to standard MMD estimation techniques (Gretton et al., 2012) to derive alternative estimators with different computational and statistical tradeoffs.

Theorem 2. *Let* $P \in \mathscr{P}(\mathbb{R}^d)$*, and let* $\widehat{P}_n \triangleq (X_i)_{i=1}^n \overset{\text{iid}}{\sim} P$*. Define the estimator* $\texttt{MMD-Gauss}_{\kappa_d}$ *via*

$$\texttt{MMD-Gauss}_{\kappa_d}(\widehat{P}_n) \triangleq \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} \kappa_d(X_i, X_j) - \frac{2}{n\sqrt{1+2\gamma\sigma^2}} \sum_{i=1}^n {}_1F_1\left(\tfrac{1}{2}; \tfrac{d}{2}; -\tfrac{\gamma}{1+2\gamma\sigma^2}\|X_i\|_2^2\right) + C$$

for some constant C *independent of* P*. Then* $\texttt{MMD-Gauss}_{\kappa_d}(\widehat{P}_n)$ *is an unbiased estimate of* $\mathsf{SIGReg}(P)$*, and* $|\mathbb{E}[\texttt{MMD-Gauss}_{\kappa_d}(\widehat{P}_n) - \texttt{SIGReg}(\widehat{P}_n)]| \leq O(1/n)$. [Proof]

The estimator of Thm. 2 presents interesting tradeoffs relative to the sliced Epps-Pulley estimator of LeJEPA (Balestriero & LeCun, 2025). Both estimators require samples $\widehat{P}_n$ from P. The tradeoffs lie in which additional random samples are required, as well as computational efficiency. Notably, the $\texttt{MMD-Gauss}_{\kappa_d}$ estimator avoids Monte-Carlo sampling slice directions by effectively computing the expectation over directions in closed form: this is what brings us from the simple k_{gsn} kernel to κ_d and ${}_1F_1$ evaluations. Moreover, in principle, $\texttt{MMD-Gauss}^2_{\kappa_d}$ is *unbiased*, unlike the implementation of

LeJEPA which requires approximating a Fourier integral via quadrature methods. However, these advantages come at a computational cost, which we discuss next.

In transitioning from sampling directions and evaluating univariate MMDs (e.g., Epps-Pulley tests), we perform pairwise kernel evaluations, costing $O(n^2)$ computation as opposed to the $O(n)$ complexity of LeJEPA. That said, it is worth emphasizing that this is but a design decision as opposed to an algorithmic necessity. Indeed, there is a rich literature on MMD estimation, and linear-time unbiased estimators are well-established (Gretton et al., 2012, Lemma 14), though these generally come with higher variance. We leave the exploration of such estimators for future work.

4.1 A Slice-Free Tractable Estimator

In this section, we gave an unbiased estimator that estimates the same regularization as LeJEPA's SIGReg, without necessitating random slices. Unfortunately, in doing so, an additional complication snuck in, since ${}_1F_1$ and thus the slice-induced kernel κ_d, are generally intractable to compute exactly. Fortunately, due to the following result of Knop et al. (2020), this kernel can be well approximated by a simpler characteristic kernel in the high-dimensional regime.

Theorem 3 (Knop et al. (2020), Proposition 4)**.** *It holds that*

$$ {}_1F_1\left(\tfrac{1}{2}; \tfrac{d}{2}; -c\right) - k_{\mathrm{imq}}^{4/2d-3,1/2}(c) \xrightarrow{d\uparrow\infty} 0 \quad \textit{for} \quad k_{\mathrm{imq}}^{4/2d-3,1/2}(c) \triangleq \frac{1}{\sqrt{1+\frac{4}{2d-3}c}}. $$

As a consequence, it follows immediately that $\kappa_d = k_{\mathrm{imq}}^{4\gamma/2d-3,1/2} + o(1)$.

While Knop et al. (2020) doesn't provide a quantitative convergence rate, they note that the approximation tends to be practically very useful once $d \geq 20$. This is especially relevant in our setting, where the goal is to train large generalist models, where the feature dimension is often greater than 20 by at least an order of magnitude. This motivates the following tractable `CWMMD` estimator for SIGReg.

Theorem 4. *Let* $P \in \mathscr{P}(\mathbb{R}^d)$*, and let* $\widehat{P}_n \triangleq (X_i)_{i=1}^n \overset{\text{iid}}{\sim} P$*. Define the estimator* `CWMMD` *via*

$$ \texttt{CWMMD}(\widehat{P}_n) \triangleq \frac{1}{n(n-1)}\sum_{i=1}^n \sum_{j\neq i} k_{\mathrm{imq}}^{4\gamma/2d-3,1/2}(X_i, X_j) - \frac{2}{n\sqrt{1+2\gamma\sigma^2}} \sum_{i=1}^n \frac{1}{\sqrt{1+\frac{4\gamma}{(2d-3)(1+2\gamma\sigma^2)}\|X_i\|_2^2}} + C. $$

Then $|\mathbb{E}[\texttt{CWMMD}(\widehat{P}_n) - \texttt{SIGReg}(\widehat{P}_n)]| \leq O_d(1/n)$. [Proof]

Altogether, both `CWMMD` and LeJEPA induce a small bias in regularization, and the former expends quadratic compute to eliminate the variance of Monte-Carlo slicing. With large distributed training setups, we argue that it is worth investigating which side of this tradeoff is favorable, or identifying situations when one approach may be preferred over the other. That said, it is important to note that **slicing divergences is not necessary or sufficient for breaking the curse of dimensionality**. We showed that equivalent regularizers can be derived (and indeed, generalized) without slicing at all.

4.2 A Novel Kernel Stein Regularizer

So far we have separated the regularization metric from the estimator introduced by LeJEPA, revealing alternative slice-free estimators leveraging kernel discrepancies with heavy-tailed kernels. Particularly, we showed that SIGReg is equivalent to MMD regularization with a particular kernel. In this section, we apply a similar technique to §4.1 to derive an equivalent estimator for the sliced KSD (Gong et al., 2021), which can be computed without random projections.

Theorem 5. *Let* $P \in \mathscr{P}(\mathbb{R}^d)$*, and let* $\widehat{P}_n \triangleq (X_i)_{i=1}^n \overset{\text{iid}}{\sim} P$*. Define* $\eta_{ij} \triangleq \|X_i - X_j\|^{-1}(X_i - X_j)$*, and consider the estimator* $\texttt{CWKSD} : \mathbb{R}^{d\times n} \to \mathbb{R}_+$ *given by*

$$ \begin{aligned} \texttt{CWKSD}(\widehat{P}_n) \triangleq{}& \frac{1}{n(n-1)}\sum_{i=1}^n \sum_{j\neq i} \left(2\gamma + \frac{X_i^\top X_j - \eta_{ij}^\top X_i \eta_{ij}^\top X_j}{\sigma^4(d-1)}\right) k_{\mathrm{imq}}^{4\gamma/2d-3,1/2}(X_i, X_j) \\ &+ \frac{1}{n(n-1)}\sum_{i=1}^n \sum_{j\neq i} \frac{1}{d}\left(\frac{\eta_{ij}^\top X_i \eta_{ij}^\top X_j}{\sigma^4} - \frac{X_i^\top X_j - \eta_{ij}^\top X_i \eta_{ij}^\top X_j}{\sigma^4(d-1)}\right) k_{\mathrm{imq}}^{2\gamma/d,3/2}(X_i, X_j) \\ &+ \frac{1}{n(n-1)}\sum_{i=1}^n \sum_{j\neq i} \frac{1}{d}\left(\frac{2\gamma}{\sigma^2} + 4\gamma^2\right)\|X_i - X_j\|_2^2 k_{\mathrm{imq}}^{2\gamma/d,3/2}(X_i, X_j). \end{aligned} $$

Then $\mathbb{E}[\texttt{CWKSD}(\widehat{P}_n)] = \mathscr{S}\mathsf{KSD}^2_{k_{\mathrm{gsn}}}(P, \mathcal{N}(0, \sigma^2 I_d)) + o_d(1)$. [Proof]

Thm. 5 shows that the sliced KSD^2 can be estimated without sampling random projections, by instead averaging heavy-tailed kernel evaluations on latent samples which are already present. Like `CWMMD`, variance of random projections is traded off for quadratic compute in the number of samples n.

5 METHODS

Network architecture. KerJEPA is composed of a Vision Transformer (ViT) (Kolesnikov et al., 2021) and three layer projector of hidden dimension of 4096 and GELU activations with output dimension of 256. For the ViT-S, we divide these values in half. During inference, the projector is discarded. Use of [`CLS`] or patch tokens are task dependent.

Target standardization. To preserve the statistical power of the chosen kernel over the course of training, we introduce a target standardization procedure inspired by Ermolov et al. (2021). This substitutes pairwise MSE in (1) with MSE against an aggregated target,

$$\mathcal{L}(\theta) \triangleq \frac{1}{dnm}\sum_{l=1}^{n}\sum_{i=1}^{m}\|f_\theta(V_l^i) - \mathsf{sg}(\bar{y}_l)\|_2^2 + \lambda\Omega(\widehat{P}, Q), \quad \bar{y}_l \triangleq \frac{1}{\widehat{\sigma}}(\frac{1}{m}\sum_{i=1}^{m} f_\theta(V_l^j) - \widehat{\mu}),$$

where sg is the stop-gradient operator, and where $\widehat{\mu}, \widehat{\sigma}$ are derived from EMA estimates of the global mean and variance of features (Caron et al., 2021). By incorporating standardization directly into the alignment target, we constrain the location and scale of the representation space, discouraging trivial collapse while reducing sensitivity to the kernel bandwidth.

Implementation details. We pretrain KerJEPA without labels in two settings. First, we train a ViT-S/8 on Imagenette (fast.ai, 2019) for 800 epochs using AdamW (Loshchilov & Hutter, 2019), a cosine learning rate schedule with a peak lr of 5×10^{-4}, weight decay of 5×10^{-2}, and a batch size of 256. Each sample consists of 4 global views of size 128px^2. We evaluate both MMD and KSD across a range of kernels and priors in the sliced, and unsliced setting, and analyze stability and convergence properties. We provide pseudocode for sliced regularizers in Fig. 2.

Second, we evaluate three model scales with patch size 16 on ImageNet using a target normalized self-distilled student-teacher framework with EMA, scheduled cosine-wise from 0.992 to 1.0. We fix the loss to `CWMMD` with a weight of $\lambda = 7.5$ and select the bandwidth using the one-dimensional Silverman's Rule (Silverman, 2018). We use 2 global views of size 224px^2, 10 local views of size 96px^2, and a batch size of 1024. Weight decay is warmed up from 0.05 to 0.6 for ViT-S/B and 0.8 for ViT-L over the first 75% of training, while learning rate decays from 8×10^{-4} to 1.5×10^{-4}. We apply a 0.2 learning rate scaling to the patch-convolution and layerwise decay of 0.9 for stability (Chen* et al., 2021; Oquab et al., 2024). In the multi-crop setting, the target is constructed exclusively from global views, and representations from both local and global views are aligned to this standardized target. ViT-S/B/L models are trained for 800, 400, and 250 epochs . Across settings, learning rate is warmed up on the first 10 epochs, and we use augmentations following BYOL (Grill et al., 2020) and SWAV (Caron et al., 2020). Regularization is only applied to global views and are averaged.

Evaluation protocols. We evaluate frozen backbone features using linear probing and k-nearest-neighbor (k-NN) classification. For linear probing, a classifier is trained online with random resized crops and horizontal flips, while evaluation uses center crops. We sweep learning rates $\{10^{-k}, 5\times10^{-k}\}_{k=2}^{5}$ and use no weight decay. For k-NN, classification is performed offline on unaugmented features, sweeping $k \in \{10, 20, 50, 100, 200\}$. We restrict comparisons to strong open-weight ImageNet-trained models from DINO Caron et al. (2021) and iBOT (Zhou et al., 2022), standardizing architectures and training data.

6 EXPERIMENTAL RESULTS

We organize our experiments around the progression from the theoretical properties of our discrepancy estimators to their practical performance and design choices. We first validate the predicted dependence of finite-slice estimation on representation dimensionality, showing that insufficient slicing leads to slower and less stable convergence. We then demonstrate the practical benefit of the resulting analytic formulations on ImageNet-1k, where KerJEPA learns strong visual representations across model scales. Finally, we examine the design and optimization choices that contribute to performance and stable training. We call unsliced methods those which do not apply projections and compute the discrepancy in the ambient space, sliced methods as those which compute the

Monte-Carlo estimation of the discrepancy under finite slices, and the analytically sliced as those which compute the discrepancy as the limiting case of all slices.

Slicing requirements. Finite-slice estimators introduce Monte Carlo noise whose effect depends on the dimensionality of the output space. As the number of dimensions increases, more slices are required to obtain a sufficiently accurate approximation. Fig. 1 confirms this behavior and demonstrates how insufficient numbers of slices increasingly slow convergence and introduce instability during the early stages of training. In contrast, the closed-form estimators for `CWMMD` and `CWKSD` avoid this source of Monte Carlo error and exhibit faster and more stable convergence across dimensions. These results motivate the use of the analytic formulations for our high-dimensional experiments.

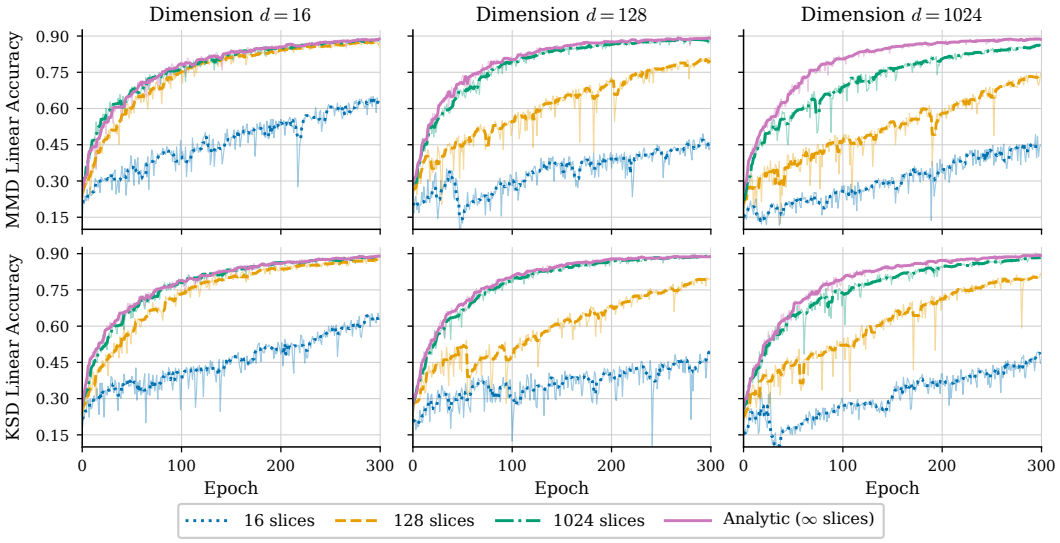

Figure 1: Impacts of slicing in various dimensions for the MMD and KSD regularizer on ImageNette, measuring test accuracy over various output dimensions. Insufficient slices slow convergence rates as a function of dimension and introduces instability in the early half of training.

Discrepancy design spaces. We examine the impacts of different discrepancies, kernels, and priors across sliced, unsliced, and analytically sliced options. Tab. 1 shows that all priors and approximation methods converge to similar results, despite their training trajectories. The framework is thus not tied to a single kernel, prior, or slicing requirement. See Fig. 2 for examples of practical sliced algorithms.

Table 1: **ImageNette regularization ablations.** Top-1 accuracy (%) across different settings on models trained for 800 epochs with dimension of 128. Unsliced heavy-tail kernels perform best.

Discrepancy	Type	Base Kernel	Prior	Slices	Linear
MMD	Unsliced	Gaussian	Gaussian	–	91.29
	Sliced (Analytic)	Gaussian	Gaussian	∞	91.13
	Sliced (Finite)	Gaussian	Gaussian	1024	91.13
	Sliced (Finite)	Gaussian	Laplace	1024	90.25
KSD	Unsliced	Gaussian	Gaussian	–	91.31
	Unsliced	Gaussian	Laplace	–	91.18
	Unsliced	IMQ	Gaussian	–	**91.90**
	Unsliced	IMQ	Laplace	–	91.12
	Sliced (Analytic)	Gaussian	Gaussian	∞	91.11
	Sliced (Finite)	Gaussian	Gaussian	1024	91.36
	Sliced (Finite)	Gaussian	Laplace	1024	90.70

Strong visual representations. We evaluate the benefit of the analytic discrepancy formulation in the self-distillation setting. KerJEPA achieves strong frozen-representation performance across varying model sizes, outperforming the corresponding DINO and iBOT baselines on both linear probing and k-NN classification (Tab. 2) on ImageNet. Outside downstream classification, KerJEPA features also enable DINO-level semantic segmentation (see Appendix D); we believe this can further be improved with the inclusion of a masked image modeling loss, which we leave to future work.

Table 2: **ImageNet-1k classification.** Top-1 accuracy (%) on frozen [CLS] features. KerJEPA outcompetes baselines and shows favorable scaling across model sizes.

Backbone	Method	k-NN	Linear
ViT-S/16	DINO	74.45	76.17
	iBOT	75.15	77.08
	KerJEPA	**76.09**	**77.85**
ViT-B/16	DINO	76.10	77.99
	iBOT	77.02	79.37
	KerJEPA	**78.95**	**80.80**
ViT-L/16	iBOT	77.94	80.87
	KerJEPA	**80.01**	**81.47**

Stability and convergence. Training in a self-distillation setting requires careful coordination of the optimization dynamics. We observe that the learning rate and weight decay schedules are closely coupled. Contrary to other methods, learning rate should not be decayed to zero, as doing so can substantially degrade representation quality, and in some cases lead to collapse. The terminal weight decay must likewise be chosen in conjunction with this minimum learning rate and that higher values of decay allowed for higher final learning rates. To maintain stability, layerwise learning-rate decay, patch scaling, projector batch normalization, and target standardization were beneficial. Across multiple runs with different random seeds, models trained without target standardization collapse, whereas applying standardization prevents these failures.

7 Discussion

This paper unifies a design space for distributional regularizers for Euclidean SSL. Our analysis and results clarify the trade-offs between computation, stability, and representation geometry in practice.

Computational tradeoffs. While the analytically sliced variations of MMD and KSD are robust and exhibit accelerated convergence in certain scenarios, they formally incur quadratic computational complexity and rely on large-dimension asymptotes. In regimes where batch sizes do not exceed the total number of slices, the total benefit of slicing is not realized, barring any cross-device communication overhead in distributed settings. Given that these algorithms require extended training schedules, the stability of exact (or analytically sliced) methods may outweigh the theoretical speedups of finite slicing. Our theoretical results shed light on the nature of the kernels induced by slicing, particularly their heavy-tailed and dimension-dependent nature, which allows us to stabilize training and achieve faster convergence. For larger-batch settings, computation can be reduced using Random Fourier Features (Rahimi & Recht, 2007; Huggins & Mackey, 2018; Zhao & Meng, 2015), high-fidelity MMD and KSD coresets (Domingo-Enrich et al., 2023; Li et al., 2024), or B-statistic estimators (Zaremba et al., 2013) that aggregate pairwise computations across distributed devices. Our experimental results show that slicing, while asymptotically more efficient in the limit of large batch size, is less stable and slower to converge on shorter training horizons.

Representation geometry. A critical consideration for these methods is their interaction with standard SSL architectures. Training employs a backbone-projector pair, where geometry is imposed on the projector's output, yet inference relies solely on the backbone (Bordes et al., 2023). We view the projector as a non-linear pushforward map; this decoupling obscures the geometry induced at the bottleneck. Experimentally, we observe that the choice between Gaussian and Laplace priors on non-sliced experiments had negligible impact on downstream classification performance. Consequently, it remains unclear whether performance gains stem from the specific choice of isotropic prior or simply from the favorable learning dynamics of Euclidean gradients.

AI USE STATEMENT

In this work, we used generative AI tools for review and verification of text, code, and proofs. For the manuscript, we used multi-turn iterations to condense and revise human drafted content and verify redundancies across sections. For code, generative tools we used verify and test the multi-node distributed variants of each regularizer. The remainder of the code was built on open-source repositories, where individual components were integrated as needed. Proofs were double checked with generative tools, specifically for the segment on hypergeometric integrals. The remainder of the work was done by the team, which includes but is not limited to: ideation and conceptualization, base implementations, initial proofs, and drafting the manuscript. All generated content was verified and the team takes responsibility for the final content of this work.

REFERENCES

Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1948.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. ISSN 1088-6850. doi: 10.1090/s0002-9947-1950-0051437-7.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Adrien Bardes, Jean Ponce, and Yann Lecun. VICReg: Variance-Invariance-Covariance Regularization For Self-Supervised Learning. In *International Conference on Learning Representatins (ICLR)*, 2022.

L. Baringhaus and N. Henze. A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, 35(1):339–348, 1988. ISSN 1435-926X. doi: 10.1007/bf02613322.

Alessandro Barp, Carl-Johann Simon-Gabriel, Mark Girolami, and Lester Mackey. Targeted separation and convergence with kernel discrepancies. *Journal of Machine Learning Research*, 25(378): 1–50, 2024.

Clément Bonet, Paul Berg, Nicolas Courty, François Septier, Lucas Drumetz, and Minh-Tan Pham. Spherical sliced-wasserstein. In *International Conference on Learning Representatins (ICLR)*, 2023.

Florian Bordes, Randall Balestriero, Quentin Garrido, Adrien Bardes, and Pascal Vincent. Guillotine regularization: Why removing layers is needed to improve generalization in self-supervised learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, volume 119, pp. 1597–1607. PMLR, 2020.

Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning (ICML)*, volume 48, pp. 2606–2615. PMLR, 2016.

Carles Domingo-Enrich, Raaz Dwivedi, and Lester Mackey. Compress then test: Powerful kernel testing in near-linear time. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206, pp. 1174–1218. PMLR, 2023.

T. W. Epps and Lawrence B. Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726, 1983. ISSN 00063444.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3015–3024. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ermolov21a.html.

fast.ai. imagenette, 2019. URL https://github.com/fastai/imagenette.

Wenbo Gong, Yingzhen Li, and José Miguel Hernández-Lobato. Sliced kernelized Stein discrepancy. In *International Conference on Learning Representatins (ICLR)*, 2021.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning (ICML)*, volume 70, pp. 1292–1301. PMLR, 2017.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 21271–21284, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Jonathan Huggins and Lester Mackey. Random feature Stein discrepancies. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Szymon Knop, Przemyslaw Spurek, Jacek Tabor, Igor Podolak, Marcin Mazur, and Stanislaw Jastrzebski. Cramer-Wold auto-encoder. *Journal of Machine Learning Research*, 21(164):1–28, 2020.

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

Anna Korba, Pierre-Cyril Aubin-Frankowski, Szymon Majewski, and Pierre Ablin. Kernel Stein discrepancy descent. In *International Conference on Machine Learning (ICML)*, volume 139, pp. 5719–5730. PMLR, 2021.

Lingxiao Li, Raaz Dwivedi, and Lester Mackey. Debiased distribution compression. In *International Conference on Machine Learning (ICML)*, volume 235, pp. 27675–27731. PMLR, 2024.

Yazhe Li, Roman Pogodin, Danica J. Sutherland, and Arthur Gretton. Self-supervised learning with kernel dependence maximization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 15543–15556, 2021.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning (ICML)*, volume 48, pp. 276–284. PMLR, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Théo Moutakanni, Maxime Oquab, Marc Szafraniec, Maria Vakalopoulou, and Piotr Bojanowski. You don't need domain-specific data augmentations when scaling self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 116106–116125, 2024.

Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 20802–20812, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 20, 2007.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005. ISBN 9780262256834. doi: 10.7551/mitpress/3206.001.0001.

Hadi Sepanj and Paul Fieguth. Aligning feature distributions in vicreg using maximum mean discrepancy for enhanced manifold awareness in self-supervised representation learning. *Journal of Computational Vision and Imaging Systems*, 10(1):13–18, Feb. 2025. doi: 10.15353/jcvis.v10i1.10002.

B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Routledge, February 2018. ISBN 9781315140919. doi: 10.1201/9781315140919. URL http://dx.doi.org/10.1201/9781315140919.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50):1517–1561, 2010.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, volume 119, pp. 9929–9939. PMLR, 2020.

Wojciech Zaremba, Arthur Gretton, and Matthew Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, 2013.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning (ICML)*, volume 139, pp. 12310–12320. PMLR, 2021.

Ji Zhao and Deyu Meng. FastMMD: Ensemble of circular discrepancy for efficient two-sample test. *Neural Computation*, 27(6):1345–1372, 2015. doi: 10.1162/NECO_a_00732.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022.

A Shift-Invariant Kernels

Kernel methods provide a powerful framework for comparing probability distributions by embedding them into a *Reproducing Kernel Hilbert Space* (RKHS) (Aronszajn, 1950). Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a positive definite kernel with associated RKHS $\mathcal{H}$. The kernel defines a feature map $\phi : \mathbb{R}^d \to \mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$. For any probability distribution P on $\mathbb{R}^d$, its *kernel mean embedding* is defined as

$$\mu_P = \mathbb{E}_{X \sim P}[k(X, \cdot)] = \mathbb{E}_{X \sim P}[\phi(X)] \in \mathcal{H}.$$

This embedding maps a distribution into the RKHS, allowing us to measure distances between distributions via their embeddings. If the kernel k is *characteristic*, this mapping is injective: $\mu_P = \mu_Q$ if and only if $P = Q$ (Sriperumbudur et al., 2010). This property ensures that kernel-based discrepancies capture full distributional differences, not merely differences in low-order moments.

Often, one employs a continuous, shift-invariant kernel $k(x, y) = k(x - y)$, which, by *Bochner's Theorem* (Rasmussen & Williams, 2005), admits a Fourier representation on non-negative Borel measures with a *spectral density* ρ_k defined as

$$k(x - y) = \int_{\mathbb{R}^d} \exp(\mathrm{i}\omega^\top (x - y)) \rho_k(\omega) \mathrm{d}\omega.$$

A common shift-invariant kernel of interest, categorized by its fast decay profile is the Gaussian Kernel

$$k_{\mathrm{gsn}}(x, y) = \exp(-\gamma \|x - y\|_2^2), \qquad \rho_{\mathrm{gsn}}(\omega) = \frac{1}{(4\pi\gamma)^{d/2}} \exp\left(-\frac{1}{4\gamma}\|\omega\|_2^2\right).$$

Here, the bandwidth $\gamma > 0$ controls smoothness between global and local dependencies between samples.

We draw attention to a family of kernels of interest, namely, the *Inverse Multiquadric Kernel* (IMQ). Unlike the Gaussian kernel, which has exponentially decaying tails, the IMQ decays at a polynomial rate (heavy-tailed) and is defined as

$$k_{\mathrm{imq}}(x, y) = \left(1 + \alpha \|x - y\|_2^2\right)^{-\beta}, \qquad \rho_{\mathrm{imq}}(\omega) = \frac{2^{1-\beta}}{\Gamma(\beta)(2\pi\alpha)^{d/2}} \left(\frac{\|\omega\|_2}{2\sqrt{\alpha}}\right)^{\nu} K_\nu \left(\frac{\|\omega\|_2}{\sqrt{\alpha}}\right),$$

where Γ is the Gamma function and K_ν is the modified Bessel of second kind, order ν. The parameter $\beta > 0$ acts as a shape parameter that determines the heaviness of the tails (rate of polynomial decay). $\alpha > 0$ serves as the inverse length scale. The IMQ kernel can be interpreted as an infinite scale mixture of Gaussian kernels. This heavy-tailed property allows it to capture long-range dependencies that are often missed by the fixed-scale Gaussian kernel.

B Isotropic Prior Distributions

Representations are shaped using a selected prior distribution $p(x)$ as a learning target. Different isotropic priors induce different tail geometries. We incorporate these priors in different ways, either by taking expectations over $p(x)$ or by matching it's structure via the score $\nabla_x \log p(x)$ and characteristic function $\phi(\omega)$, all without having to drawing samples.

The characteristic function (CF) is defined the Fourier transform of the density with frequencies $\omega \in \mathbb{R}^d$

$$\phi_P(\omega) = \mathbb{E}_{X \sim P}\left[\exp\left(\mathrm{i}\omega^\top X\right)\right].$$

Given samples $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$, and the empirical distribution $\widehat{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$, we can write the empirical characteristic function (ECF) and its score-weighted counterpart as

$$\widehat{\phi}_n(\omega) := \frac{1}{n} \sum_{i=1}^n \exp(\mathrm{i}\omega^\top X_i), \qquad \widehat{\phi}_{s_Q, n}(\omega) := \frac{1}{n} \sum_{i=1}^n s_Q(X_i) \exp(\mathrm{i}\omega^\top X_i).$$

which enables direct matching between the ECF to a known CF. We provide examples for three isotropic priors that can be used a as regularizer below.

Table 3: Isotropic prior distributions: probability density $p(x)$, scores $\nabla_x \log p(x)$, and their characteristic function $\phi(\omega)$ used to construct different discrepancy objectives.

Distribution	Notation	$p(x)$	$\nabla_x \log p(x)$	$\phi(\omega)$
Gaussian	$\mathcal{N}(0, \sigma^2 I_d)$	$\frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|x\|_2^2\right)$	$-\frac{1}{\sigma^2}x$	$\exp\left(-\frac{1}{2}\sigma^2\|\omega\|_2^2\right)$
Laplace	$\text{Lap}(0, \sigma I_d)$	$\frac{\Gamma(d/2)}{2\pi^{d/2}\sigma^d\Gamma(d)} \exp\left(-\frac{1}{\sigma}\|x\|_2\right)$	$-\frac{1}{\sigma}\frac{x}{\|x\|_2}$	$\left(1+\sigma^2\|\omega\|_2^2\right)^{-\frac{d+1}{2}}$
Student-t	$\mathcal{T}_\nu(0, \sigma^2 I_d)$	$\frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi\sigma^2)^{d/2}} \left(1+\frac{1}{\nu\sigma^2}\|x\|_2^2\right)^{-\frac{\nu+d}{2}}$	$-\frac{\nu+d}{\nu\sigma^2+\|x\|_2^2}x$	$\frac{K_{\nu/2}(\sqrt{\nu}\sigma\|\omega\|_2)(\sqrt{\nu}\sigma\|\omega\|_2)^{\nu/2}}{\Gamma(\nu/2)2^{\nu/2-1}}$

C KERNEL DISCREPANCIES

In this section, we derive variations of the MMD and KSD with corresponding sample-based estimators, discussing the features and difficulties associated with different choices of kernels and priors. We note that all discrepancies that can be written under Bochner's theorem can be approximated via slicing paired with 1D quadrature or via RFF.

C.1 MAXIMUM MEAN DISCREPANCY

Restating the definition, the *Maximum Mean Discrepancy* (MMD) is an RKHS distance measure between probability distributions P and Q defined via kernel mean embeddings from each distribution. For a kernel k associated with RKHS $\mathcal{H}$, it is defined as

$$\mathsf{MMD}_k(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}}.$$

Equivalently, the square of this metric has the following useful form:

$$\mathsf{MMD}_k^2(P, Q) = \mathbb{E}_{X,X'\sim P}[k(X, X')] + \mathbb{E}_{Y,Y'\sim Q}[k(Y, Y')] - 2\mathbb{E}_{X\sim P, Y\sim Q}[k(X, Y)].$$

Altogether, we provide the following MMD estimators, with n, m denoting the number of samples drawn from the distribution under test and the prior distribution respectively:

1. $\widehat{\mathsf{UMMD}}^2_{k,n,m}$ (4): a general unbiased estimator for MMDs with arbitrary kernels and priors;
2. $\widehat{\mathsf{CMMD}}^2_{k,n}$ (5): a consistent estimator for shift-invariant kernels k and priors with known characteristic functions, which requires only $O(n)$ computational complexity; and
3. $\widehat{\mathsf{UGMMD}}^2_{\sigma^2,n}$ (6): an unbiased estimator for Gaussian kernels with prior $\mathcal{N}(0, \sigma^2 I_d)$, with reduced variance.

The following U-statistic is an unbiased estimator Gretton et al. (2012) of the MMD, given $X_i \overset{\text{iid}}{\sim} P$ and $Y_i \overset{\text{iid}}{\sim} Q$ where

$$\begin{aligned}\widehat{\mathsf{UMMD}}^2_{k,n,m}(P, Q) = &\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} k(X_i, X_j) \\ &+ \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i} k(Y_i, Y_j) \\ &- 2\sum_{i=1}^{n}\sum_{j=1}^{m} k(X_i, Y_j).\end{aligned} \tag{4}$$

It is well-known that estimation of the MMD via this U-statistic achieves dimension-free sample complexity Gretton et al. (2012). However, there are at least two qualities of this estimator one might hope to improve:

1. It involves pairwise kernel evaluations, which scales quadratically with the number of samples drawn from the distributions under comparison; and

2. If one could compute the expectations over the prior (Q) in closed form, variance can be reduced.

With regard to the first point, we can turn to the Bochner integral formulation of the MMD, which would scale linearly with the number of samples. Particularly, we derive the CF MMD estimator (a V-estimator) below:

$$\widehat{\mathsf{CMMD}}^2_{k,n}(P,Q) = \int_{\mathbb{R}^d} \left|\phi_{\widehat{P}^n}(\omega) - \phi_Q(\omega)\right|^2 \rho_k(\omega)\mathrm{d}\omega, \quad \widehat{P}_n = \tfrac{1}{n}\textstyle\sum_{i=1}^n \delta_{X_i}, \quad X_i \overset{\text{iid}}{\sim} P, \tag{5}$$

where ϕ_P denotes the CF of the probability measure P. Notably, under this estimator, there are no pairwise comparisons being made, and the complexity scales only as $O(n)$. Here, we assume that ϕ_Q is known analytically, which is often the case for priors used in practice (e.g. Gaussian, Laplace, or Student-t priors). Note, however, that certain kernels (e.g. Student-t, IMQ) require their own quadrature for estimating their spectral density, adding additional overhead, which also scales with d.

With regard to the second criterion, computing the expectation over Q in closed form is generally intractable. Exceptionally, this can be done in the scenario of Gaussian priors under Gaussian kernels.

Proposition 1. *For any distribution $P \in \mathscr{P}(\mathbb{R}^d)$, we have*

$$\begin{aligned}\mathsf{MMD}^2_{k_{\text{gsn}}}(P, \mathcal{N}(0,\sigma^2 I_d)) = \mathbb{E}_{X,X'\sim P} \exp(-\gamma\|X - X'\|_2^2) \\ - \frac{2}{(1+2\gamma\sigma^2)^{d/2}} \mathbb{E}_{X\sim P} \exp\left(-\frac{\gamma}{1+2\gamma\sigma^2}\|X\|_2^2\right) \\ + C^{\text{gauss}}_{\gamma,\sigma,d},\end{aligned}$$

for a constant $C^{\text{gauss}}_{\gamma,\sigma,d}$ independent of P given by

$$C^{\text{gauss}}_{\gamma,\sigma,d} := \frac{1}{(1+4\gamma\sigma^2)^{d/2}}.$$

We prove this proposition below. First, we highlight the resulting unbiased U-statistic $\widehat{\mathsf{UGMMD}}^2_{\sigma^2,n}(P)$ in the case of a Gaussian prior with a Gaussian kernel—this is also known as the *BHEP* estimator Baringhaus & Henze (1988):

$$\begin{aligned}\widehat{\mathsf{UGMMD}}^2_{\sigma^2,n}(P) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j\neq i} \exp(-\gamma\|X_i - X_j\|_2^2) \\ - \frac{2}{(1+2\gamma\sigma^2)^{d/2}} \sum_{i=1}^n \exp\left(-\frac{\gamma}{1+2\gamma\sigma^2}\|X_i\|_2^2\right) \\ + C^{\text{gauss}}_{\gamma,\sigma,d},\end{aligned} \tag{6}$$

where $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P$, and where $C^{\text{gauss}}_{\gamma,\sigma,d}$ is independent of P and given in Proposition Prop. 1.

Proof. Proof of Proposition Prop. 1. Let $\mathrm{N} = \mathcal{N}(0,\sigma^2 I_d)$. Starting from equation 2, we have

$$\mathsf{MMD}^2_{k_{\text{gsn}}}(P,\mathrm{N}) = \mathbb{E}_{X,X'\sim P}[k_{\text{gsn}}(X,X')] + \underbrace{\mathbb{E}_{Y,Y'\sim \mathrm{N}}[k_{\text{gsn}}(Y,Y')]}_{\mathrm{A}} - 2\underbrace{\mathbb{E}_{(X,Y)\sim P\otimes \mathrm{N}}[k(X,Y)]}_{\mathrm{B}}.$$

We simplify terms A and B individually. Starting with A,

$$\begin{aligned}\mathrm{A} &\triangleq \mathbb{E}_{Y,Y'\sim Q}[k_{\text{gsn}}(Y,Y')] \\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \exp\left(-\gamma\|y-y'\|_2^2\right) \frac{1}{(2\pi\sigma^2)^d} \exp\left(-\frac{1}{2\sigma^2}(\|y\|_2^2 + \|y'\|_2^2)\right) \mathrm{d}y\mathrm{d}y' \\ &= \frac{1}{(1+4\gamma\sigma^2)^{d/2}} \triangleq C^{\text{gauss}}_{\gamma,\sigma,d}.\end{aligned}$$

Now, for B, we have

$$\mathbb{E}_{Y\sim \mathrm{N}}[k_{\text{gsn}}(x,Y)] = \textstyle\int_{\mathbb{R}^d} \exp(-\gamma\|x-y\|_2^2) \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|y\|_2^2\right) \mathrm{d}y.$$

This is a convolution of Gaussians, which evaluates to

$$\mathbb{E}_{Y\sim \mathrm{N}}[k_{\mathrm{gsn}}(x,Y)] = \tfrac{1}{(2\gamma\sigma^2+1)^{d/2}} \exp\left(-\tfrac{\gamma}{2\gamma\sigma^2+1}\|x\|_2^2\right).$$

Substituting into B, we have

$$\mathrm{B} \triangleq \mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim \mathrm{N}}[k_{\mathrm{gsn}}(X,Y)] = \tfrac{1}{(2\gamma\sigma^2+1)^{d/2}}\mathbb{E}_{X\sim P}\left[\exp\left(-\tfrac{\gamma}{2\gamma\sigma^2+1}\|X\|_2^2\right)\right].$$

Substituting A and B above completes the proof. □

C.2 Kernel Stein Discrepancy

Restating the definition, the *Kernel Stein Discrepancy* (KSD) Chwialkowski et al. (2016); Liu et al. (2016); Gorham & Mackey (2017); Korba et al. (2021) is an RKHS-based measure of discrepancy between a probability distribution P and a target distribution Q. For an absolutely continuous target distribution Q, let $s_Q(x) = \nabla_x \log Q(x)$ denote its score function. The Langevin Stein operator $\mathscr{A}_Q$ acts on a vector-valued test function $f : \mathbb{R}^d \to \mathbb{R}^d$ according to

$$(\mathscr{A}_Q f)(x) = s_Q(x)^\top f(x) + \nabla_x \cdot f(x).$$

Under suitable regularity conditions, Stein's identity states that $\mathbb{E}_{x\sim Q}[(\mathscr{A}_Q f)(x)] = 0$. Consequently, deviations from this identity under P provide a measure of discrepancy from Q. Taking the supremum over the unit ball of a vector-valued RKHS and applying the reproducing property yields the squared KSD:

$$\mathsf{KSD}^2(P,Q) = \mathbb{E}_{x,x'\sim P}[k_{\mathrm{stein}}(x,x')],$$

where the Stein kernel is defined by applying the Stein operator to both arguments of a positive-definite base kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$:

$$\begin{aligned} k_{\mathrm{stein}}(x,y) = {} & s_Q(x)^\top k(x,y) s_Q(y) + s_Q(x)^\top \nabla_y k(x,y) \\ & + \nabla_x k(x,y)^\top s_Q(y) + \mathrm{tr}\left(\nabla_x \nabla_y^\top k(x,y)\right). \end{aligned}$$

Importantly, the target distribution enters only through its score function, eliminating the need to sample from Q or evaluate its normalization constant. This distinguishes KSD from the general MMD formulation, which typically requires samples from both distributions or closed-form evaluations of kernel expectations under the target. The resulting discrepancy can therefore be interpreted as a target-dependent kernel discrepancy in which the base kernel is transformed according to the geometry induced by Q.

We consider variations of the KSD induced by different base kernels and target isotropic priors. Let $\{X_1, \ldots, X_n\} \subset \mathbb{R}^d$ be independent samples from an arbitrary distribution P, with empirical distribution $\widehat{P} = n^{-1}\sum_{i=1}^n \delta_{X_i}$. An unbiased U-statistic estimator of the squared KSD is given by

$$\widehat{\mathsf{KSD}}^2(P,Q) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i} k_{\mathrm{stein}}(X_i, X_j).$$

We examine Gaussian, IMQ, and general shift-invariant kernels, together with the Gaussian, Laplace, and Student-t isotropic priors in §B. The required kernel gradients and mixed Hessian traces are provided in Tab. 4. Combining these derivatives with the corresponding prior score functions yields closed-form Stein kernels for each kernel–prior combination. For radial kernels, we use the identity $\nabla_x k(x,y) = -\nabla_y k(x,y)$ to simplify the resulting expressions.

The U-statistic above requires pairwise kernel evaluations and therefore scales quadratically with the number of samples. For shift-invariant kernels, however, Bochner's theorem provides an alternative spectral representation that avoids explicit pairwise comparisons. We write the corresponding spectral KSD estimator as a function of the ECF and score-weighted ECF §B as

$$\widehat{\mathsf{CKSD}}^2_{k,n}(P,Q) = \int_{\mathbb{R}^d} \left\| \widehat{\phi}_{s_Q,n}(\omega) + \mathrm{i}\omega \widehat{\phi}_n(\omega) \right\|^2 \rho_k(\omega)\, \mathrm{d}\omega.$$

Unlike the U-statistic, this estimator does not require pairwise kernel evaluations and can therefore be evaluated in $O(n)$ time for a fixed number of frequency evaluations. As with the spectral MMD

estimator, the integral can be approximated using samples from the spectral density, providing a computationally efficient alternative for large batches. Because the score function is available in closed form for the isotropic priors considered in §B, these kernel–prior combinations can be evaluated directly, yielding a flexible family of distributional regularizers without explicit representation normalization. Additionally, we may estimate the marginals along slices, allowing for the same quadrature estimation as done in SIGReg.

We provide a table with gradients and Hessians for the Gaussian and IMQ kernels needed to assemble the discrepancy, noting that the gradients for a radial kernel are odd ($\nabla_x k(x,y) = -\nabla_y k(x,y)$).

Table 4: Gradients and Hessians for the Gaussian, IMQ, and general shift-invariant kernels..

$k(x,y)$	$\nabla_x k(x,y)$	$\mathrm{tr}(\nabla_x \nabla_y k(x,y))$
Gaussian	$-2\gamma(x-y)e^{-\gamma\|x-y\|_2^2}$	$(2\gamma d - 4\gamma^2\|x-y\|_2^2)e^{-\gamma\|x-y\|_2^2}$
IMQ	$-2\alpha\beta(x-y)(1+\alpha\|x-y\|_2^2)^{-(\beta+1)}$	$4\alpha^2\beta(\beta+1)\|x-y\|_2^2(1+\alpha\|x-y\|_2^2)^{-(\beta+2)}$
Bochner	$\int_{\mathbb{R}^d} \mathrm{i}\omega\, e^{\mathrm{i}\omega^\top(x-y)}\rho(\omega)\,d\omega$	$\int_{\mathbb{R}^d} \|\omega\|_2^2 e^{\mathrm{i}\omega^\top(x-y)}\rho(\omega)\,d\omega$

C.3 Sliced Regularization Algorithms

We provide two examples of sliced discrepancy regularization using both MMD and KSD. The MMD setting generalizes SIGReg and supports different kernels and priors

This process decomposes the computational cost into two efficient steps: projections of n embeddings of dimension d onto a set of r random directions, and quadratures to estimate the spectral integral using a fixed set of u integration knots. This reduces the complexity from $\Theta(n^2 d)$ pairwise kernel evaluations to $\Theta(nr(d+u))$, making it linear in the batch size. These fast quadrature-based methods are primarily applicable when using the Gaussian or IMQ kernels combined with Gaussian, Laplace, and Student-t priors, as their spectral density, projected marginals, and CFs are analytic and well-behaved. In the case of the IMQ kernel and Student-t CF, Bessel functions only need to be computed once on their one-dimensional marginals at the known quadrature knots on initialization. We implement the two sliced methods, termed the sliced MMDReg and sliced KSDReg, as shown in Fig. 2. and we restrict our implementation to Gaussian kernel with Gaussian and Laplace priors for simplicity, as the Gaussian kernel as weighting function allows us to approximate the integral with (symmetric) Gauss-Hermite quadratures.

D Semantic Segmentation with KerJEPA

In addition to providing exceptionally strong features for downstream classification tasks, we sought out to test whether KerJEPA features also enable accurate semantic segmentation. Towards this end, we evaluated semantic segmentation on the ADE20K dataset (Zhou et al., 2017) via linear probing, and compare our results to those derived from iBOT and DINO features. The results are given in Tab. 5.

We observe that iBOT remains ahead in the context of semantic segmentation, but note that it leverages pretraining techniques, particularly a masked image modeling objective, that have been developed particularly for such a task. That said, while KerJEPA features do not match iBOT's for the purpose of semantic segmentation, they are very competitive with DINO features towards this end, which itself is a strong baseline.

Altogether, the addition of a masked image modeling objective is largely orthogonal to the developments made in this paper, which mainly deal with the regularizer. The results in Tab. 5 suggest that, when augmented with a masked image modeling objective during pretraining, KerJEPA can be very well suited to semantic segmentation as well. Since our main focus in this work was to generalize the framework introduced by LeJEPA, our study focused largely on the simplest extension of LeJEPA, but such a masked image modeling extension presents an exciting avenue for future work.

Sliced MMD as generalized SIGReg

```python
class SlicedMMReg(Module):
    def __init__(self, n_knots):
        super().__init__()
        # Full quadrature (no symmetry assumption)
        k, w = polynomial.hermite.hermgauss(n_knots)
        self.knots = tensor(k * sqrt(2)).view(1,1,-1)
        self.weights = tensor(w / sqrt(pi)).view(1,1,-1)

    def target_cf(self, w, sigma, prior):
        if prior == 'gaussian':
            return exp(-0.5 * (sigma**2) * w**2)
        elif prior == 'laplace':
            return 1.0 / (1.0 + (sigma**2) * w**2)

    def forward(self, z, slices, sigma, prior):
        # Sample slices and project
        noise = randn(slices, z.size(1))
        theta = normalize(noise, dim=1)
        proj = (z @ theta.T).unsqueeze(-1)


        # MSE ECF and CF
        args = proj * self.knots
        emp_cf = mean(torch.cos(args), dim=0)
        tgt_cf = self.target_cf(self.knots, sigma, prior)
        err = (emp_cf - tgt_cf.squeeze()).pow(2)


        # Standard integration
        weights = self.weights
        smmd = sum(err * weights, dim=-1).mean()

        return smmd
```

Sliced KSD as a Stein analogue to LeJEPA

```python
class SlicedKSDReg(Module):
    def __init__(self, n_knots):
        super().__init__()
        # Full quadrature (no symmetry assumption)
        k, w = polynomial.hermite.hermgauss(n_knots)
        self.knots = tensor(k * sqrt(2)).view(1,1,-1)
        self.weights = tensor(w / sqrt(pi)).view(1,1,-1)

    def score(self, z, sigma, prior):
        if prior == 'gaussian':
            return -z / sigma**2
        elif prior == 'laplace':
            return -normalize(z) / sigma

    def forward(self, z, slices, sigma, prior):
        # Sample slices and project
        noise = randn(slices, z.size(1))
        theta = normalize(noise, dim=1)
        proj = z @ theta.T
        s = self.score(proj, sigma, prior) @ theta.T

        # MSE Stein ECF
        args = proj * self.knots
        w = self.knots.expand_as(args)
        real = s * cos(args) - w * sin(args)
        imag = s * sin(args) + w * cos(args)
        disc = mean(real,0).pow(2) + mean(imag,0).pow(2)

        # Standard integration
        weights = self.weights
        sksd = sum(disc * weights, dim=-1).mean()

        return sksd
```

Figure 2: Pseudo-code for the finite-sliced discrepancy regularization losses for both Gaussian and Laplace priors using the Gaussian kernel. Integration is done as Gauss-Hermite quadrature approximation. Other choice of kernel may require additional quadrature estimation

Table 5: **Semantic segmentation on ADE20K.** mIoU/mAcc on frozen patch features. iBOT shows strongest performance but is trained with masked objectives.

Backbone	Method	mIoU	mAcc
ViT-S/16	iBOT	**33.01**	**42.34**
	DINO	30.29	39.05
	KerJEPA	30.67	39.05
ViT-B/16	iBOT	**38.08**	**47.41**
	DINO	34.77	43.93
	KerJEPA	34.52	43.42
ViT-L/16	iBOT	**41.02**	**50.86**
	KerJEPA	37.10	46.22

E Proofs of Theoretical Claims

Theorem 1. *Let* SIGReg *denote the SIGReg regularizer (Balestriero & LeCun, 2025); that is*

$$\mathsf{SIGReg}(P) \triangleq \mathbb{E}_{\xi\sim\mathrm{Unif}(\mathbb{S}^{d-1})}[n^{-1}\mathsf{EP}(\xi_{\#}^{\top}P)], \quad P \in \mathscr{P}(\mathbb{R}^d).$$

It holds that SIGReg *is equivalent to a particular MMD,*

$$\mathsf{SIGReg}(P) = \mathsf{MMD}^2_{\kappa_d}(P, \mathcal{N}(0,\sigma^2 I_d)) \quad \textit{where} \quad \kappa_d(x,y) \triangleq {}_1F_1\left(\tfrac{1}{2};\tfrac{d}{2};-\gamma\|x-y\|_2^2\right),$$

and ${}_1F_1$ *is the* Kummer Confluent Hypergeometric function. *In particular,* κ_d *has polynomially decaying tails (as opposed to exponentially decaying tails like* k_{gsn}*).* [Source]

Proof. As we saw in §3.4, it holds that

$$\mathsf{EP}(P) = n\mathsf{MMD}^2_{k_{\mathrm{gsn}}}(P, \mathcal{N}(0,\sigma^2)).$$

Let $X \sim P$ and $N \sim \mathcal{N}(0, \sigma^2 I_d)$. Then, by the linearity of the expectation, we have

$$\begin{aligned}\mathsf{SIGReg}(P) &= \mathbb{E}_{\xi\sim\mathcal{U}(\mathbb{S}^{d-1})}\left[\mathsf{MMD}^2_{k_{\mathrm{gsn}}}(\xi^\top_\# P, \xi^\top_\# \mathcal{N}(0,\sigma^2 I_d))\right] \\ &= \mathbb{E}_{X,X'\sim P}\mathbb{E}_{\xi\sim\mathcal{U}(\mathbb{S}^{d-1})}k_{\mathrm{gsn}}(\xi^\top X, \xi^\top X') \\ &\quad + \mathbb{E}_{Y,Y'\sim\mathcal{N}(0,\sigma^2 I_d)}\mathbb{E}_{\xi\sim\mathcal{U}(\mathbb{S}^{d-1})}k_{\mathrm{gsn}}(\xi^\top Y, \xi^\top Y') \\ &\quad - 2\mathbb{E}_{X\sim P, Y\sim\mathcal{N}(0,\sigma^2 I_d)}\mathbb{E}_{\xi\sim\mathcal{U}(\mathbb{S}^{d-1})}k_{\mathrm{gsn}}(\xi^\top X, \xi^\top Y).\end{aligned}$$

Moreover, by Nadjahi et al. (2020, Corollary S3), it holds that $(x,y) \mapsto \mathbb{E}_{\xi\sim\mathcal{U}(\mathbb{S}^{d-1})}k_{\mathrm{gsn}}(\xi^\top x, \xi^\top y) = {}_1F_1(1/2; d/2; \gamma\|x-y\|_2^2)$ is a particular solution to *Kummer's equation*; that is, ${}_1F_1(a,b,z)$ is a solution to

$$z\frac{\mathrm{d}^2 w}{\mathrm{d}z^2} + (b-z)\frac{\mathrm{d}w}{\mathrm{d}z} - \alpha w = 0.$$

Thus, substituting, we have

$$\begin{aligned}\mathsf{SIGReg}(P) &= \mathbb{E}_{X,X'\sim P}\left[{}_1F_1\left(\frac{1}{2};\frac{d}{2};\gamma\|X-X'\|_2^2\right)\right] + \mathbb{E}_{Y,Y'\sim\mathcal{N}(0,\sigma^2 I_d)}\left[{}_1F_1\left(\frac{1}{2};\frac{d}{2};\gamma\|Y-Y'\|_2^2\right)\right] \\ &\quad - 2\mathbb{E}_{X\sim P,Y\sim\mathcal{N}(0,\sigma^2 I_d)}\left[{}_1F_1\left(\frac{1}{2};\frac{d}{2};\gamma\|X-Y\|_2^2\right)\right] \\ &\equiv \mathsf{MMD}^2_{\kappa_d}(P, \mathcal{N}(0,\sigma^2 I_d)).\end{aligned}$$

This proves the first claim. Next, it is known (Abramowitz & Stegun, 1948, Chapter 13) that for fixed d,

$$\begin{aligned}&{}_1F_1\left(\frac{1}{2};\frac{d}{2};-\gamma\|x-y\|_2^2\right) \\ &\quad\sim \Gamma\left(\frac{d}{2}\right)\left[\frac{1}{\Gamma(1/2)}\exp(-\gamma\|x-y\|_2^2)\left(-\gamma\|x-y\|_2^2\right)^{(1-d)/2} + \frac{1}{\Gamma((d-1)/2)}\gamma^{-1/2}\|x-y\|_2^{-1}\right],\end{aligned}$$

where $f(\xi) \sim g(\xi)$ denotes that f is asymptotically equivalent to g in ξ, in the sense that $\lim_{\xi\to\infty} f(\xi)/g(\xi) = 1$; the expression above is stated asymptotically for $\|x-y\|_2$. Thus, directly we have that the final term dominates for large $\|x-y\|_2$, and it follows that for $d > 1$, $\kappa_d(x,y)$ decays like $\Gamma((d-1)/2)^{-1}\gamma^{-1/2}\|x-y\|_2^{-1}$. □

Theorem 2. *Let $P \in \mathscr{P}(\mathbb{R}^d)$, and let $\widehat{P}_n \triangleq (X_i)_{i=1}^n \overset{\text{iid}}{\sim} P$. Define the estimator* $\texttt{MMD-Gauss}_{\kappa_d}$ *via*

$$\begin{aligned}&\texttt{MMD-Gauss}_{\kappa_d}(\widehat{P}_n) \\ &\quad\triangleq \tfrac{1}{n(n-1)}\textstyle\sum_{i=1}^n\sum_{j\neq i}\kappa_d(X_i,X_j) - \tfrac{2}{n\sqrt{1+2\gamma\sigma^2}}\sum_{i=1}^n {}_1F_1(\tfrac{1}{2};\tfrac{d}{2};-\tfrac{\gamma}{1+2\gamma\sigma^2}\|X_i\|_2^2) + C\end{aligned}$$

for some constant C independent of P. Then $\texttt{MMD-Gauss}_{\kappa_d}(\widehat{P}_n)$ *is an unbiased estimate of* $\mathsf{SIGReg}(P)$*, and* $|\mathbb{E}[\texttt{MMD-Gauss}_{\kappa_d}(\widehat{P}_n) - \texttt{SIGReg}(\widehat{P}_n)]| \le O(1/n)$. [Source]

Proof. The proof is a direct computation. By Thm. 1, $\mathsf{SIGReg}(P) = \mathsf{MMD}^2_{\kappa_d}(P, \mathcal{N}(0,\sigma^2))$, and we will show that $\texttt{MMD-Gauss}^2_{\kappa_d}(\widehat{P}_n)$ is the (unbiased) U-statistic $\widehat{\mathsf{U\text{-}MMD}}^2_{\kappa_d}$ for $\mathsf{MMD}^2_{\kappa_d}(P, \mathcal{N}(0,\sigma^2))$. This U-statistic will be derived by integrating the contributions of $\mathcal{N}(0,1)$ in closed form instead of sampling. By Gretton et al. (2012), denoting $\mathrm{N} \triangleq \mathcal{N}(0,\sigma^2)$, we have

$$\begin{aligned}&\widehat{\mathsf{U\text{-}MMD}}^2_{\kappa_d}(\widehat{P}_n, \mathrm{N}) \\ &= \underbrace{\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}\kappa_d(X_i,X_j)}_{\mathrm{I}} + \underbrace{\iint \kappa_d(y_1,y_2)\,\mathrm{dN}(y_1)\,\mathrm{dN}(y_2)}_{\mathrm{II}\triangleq C} - \underbrace{\frac{2}{n}\sum_{i=1}^n\int\kappa_d(X_i,y)\,\mathrm{dN}(y)}_{\mathrm{III}}.\end{aligned}$$

Note that term II is totally independent of P, so we write it off as a constant C. The term I above is expressed solely in terms of pairwise kernel evaluations from samples from P, it will not be simplified further. To complete the proof, it remains to simplifty III. Unpacking κ_d into an expectation over uniform random projections, we have

$$\begin{aligned}
\text{III} &= \frac{2}{n}\sum_{i=1}^{n}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}} \exp(-\gamma(\xi^\top(X_i-y))^2)\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(\xi^\top y)^2)\,\mathrm{d}y\,\mathrm{d}\mathcal{U}(\xi) \\
&\overset{(*)}{=} \frac{2}{n}\frac{1}{\sqrt{2\gamma\sigma^2}}\sum_{i=1}^{n}\int_{\mathbb{S}^{d-1}}\exp\left(-\frac{\gamma(\xi^\top X_i)^2}{1+2\gamma\sigma^2}\right)\mathrm{d}\mathcal{U}(\xi) \\
&= \frac{2}{n\sqrt{1+2\gamma\sigma^2}}\sum_{i=1}^{n}{}_1F_1\left(\frac{1}{2};\frac{d}{2};-\frac{\gamma}{1+2\gamma\sigma^2}\|X_i\|_2\right).
\end{aligned}$$

In the previous computation, step $(*)$ simply computes the expectation of $\exp(-\gamma Z^2)$, for $Z \triangleq \xi^\top(X_i - y)$, which is a Gaussian scalar random variable. Summing I, this expression for III, and C, we find that indeed $\texttt{MMD-Gauss}^2_{\kappa_d}(\widehat{P}_n)$ is an unbiased estimator of $\mathsf{SIGReg}(P)$ as a U-statistic of $\mathsf{MMD}^2_{\kappa_d}(P, \mathcal{N}(0,\sigma^2))$.

For the final claim, it was shown in (Balestriero & LeCun, 2025, Theorem 6) that

$$\left|\mathbb{E}\left[\texttt{SIGReg}(\widehat{P}_n) - \mathsf{SIGReg}(P)\right]\right| \le O(1/n).$$

Therefore, we immediately have

$$\begin{aligned}
&\left|\mathbb{E}\left[\texttt{MMD-Gauss}^2_{\kappa_d}(\widehat{P}_n) - \texttt{SIGReg}(\widehat{P}_n)\right]\right| \\
&\quad= \left|\mathbb{E}\left[\texttt{MMD-Gauss}^2_{\kappa_d}(\widehat{P}_n) - \mathsf{SIGReg}(P)\right] + \mathbb{E}\left[\mathsf{SIGReg}(P) - \texttt{SIGReg}(\widehat{P}_n)\right]\right| \\
&\quad= \left|\mathbb{E}\left[\mathsf{SIGReg}(P) - \texttt{SIGReg}(\widehat{P}_n)\right]\right| \le O(1/n),
\end{aligned}$$

which completes the proof. □

Theorem 4. *Let $P \in \mathscr{P}(\mathbb{R}^d)$, and let $\widehat{P}_n \triangleq (X_i)_{i=1}^n \overset{\text{iid}}{\sim} P$. Define the estimator* CWMMD *via*

$$\begin{aligned}
&\texttt{CWMMD}(\widehat{P}_n) \\
&\quad\triangleq \tfrac{1}{n(n-1)}\textstyle\sum_{i=1}^n\sum_{j\neq i} k_{\text{img}}^{4\gamma/2d-3,1/2}(X_i,X_j) - \tfrac{2}{n\sqrt{1+2\gamma\sigma^2}}\sum_{i=1}^n \tfrac{1}{\sqrt{1+\frac{4\gamma}{(2d-3)(1+2\gamma\sigma^2)}\|X_i\|_2^2}} + C.
\end{aligned}$$

Then $|\mathbb{E}[\texttt{CWMMD}(\widehat{P}_n) - \texttt{SIGReg}(\widehat{P}_n)]| \le O_d(1/n)$. [Source]

Proof. The proof follows simply by substituting the result of Thm. 3 into the result of Thm. 2 and applying the triangle inequality. Particularly, defining $\alpha \triangleq \frac{4\gamma}{(2d-3)(1+2\gamma\sigma^2)}$, by Thm. 3,

$$\begin{aligned}
&\left|\texttt{CWMMD}(\widehat{P}_n) - \texttt{MMD-Gauss}^2_{\kappa_d}(\widehat{P}_n)\right| \\
&\le \left|\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}\left[k_{\text{img}}^{4\gamma/2d-3,1/2}(X_i,X_j) - {}_1F_1\left(\frac{1}{2};\frac{d}{2};-\gamma\|X_i-Y_i\|_2^2\right)\right]\right| \\
&\quad+\left|\frac{2}{n\sqrt{1+2\gamma\sigma^2}}\sum_{i=1}^n\left[\frac{1}{\sqrt{1+\alpha\|X_i\|_2^2}} - {}_1F_1\left(\frac{1}{2};\frac{d}{2};-\frac{\gamma}{1+2\gamma\sigma^2}\|X_i\|_2^2\right)\right]\right| \\
&\le \frac{1}{n(n-1)}n(n-1)o_d(1) + \frac{2}{n\sqrt{1+2\gamma\sigma^2}}no_d(1) \\
&\le o_d(1).
\end{aligned}$$

Then, we have

$$\begin{aligned}
&|\mathbb{E}[\texttt{CWMMD}(\widehat{P}_n) - \texttt{SIGReg}(\widehat{P}_n)]| \\
&\quad = |\mathbb{E}[\texttt{CWMMD}(\widehat{P}_n) - \texttt{MMD-Gauss}^2_{\kappa_d}(\widehat{P}_n) + \texttt{MMD-Gauss}^2_{\kappa_d}(\widehat{P}_n) - \texttt{SIGReg}(\widehat{P}_n)]| \\
&\quad \leq |\mathbb{E}[\texttt{CWMMD}(\widehat{P}_n) - \texttt{MMD-Gauss}^2_{\kappa_d}(\widehat{P}_n)]| + |\mathbb{E}[\texttt{MMD-Gauss}^2_{\kappa_d}(\widehat{P}_n) - \texttt{SIGReg}(\widehat{P}_n)]| \\
&\quad \leq o_d(1) + O(1/n) \\
&\quad \leq O_d(1/n),
\end{aligned}$$

where the penultimate step holds by Thm. 2. □

Theorem 5. *Let $P \in \mathscr{P}(\mathbb{R}^d)$, and let $\widehat{P}_n \triangleq (X_i)_{i=1}^n \overset{\text{iid}}{\sim} P$. Define $\eta_{ij} \triangleq \|X_i - X_j\|^{-1}(X_i - X_j)$, and consider the estimator $\texttt{CWKSD} : \mathbb{R}^{d\times n} \to \mathbb{R}_+$ given by*

$$\begin{aligned}
\texttt{CWKSD}(\widehat{P}_n) \triangleq\ & \tfrac{1}{n(n-1)} \textstyle\sum_{i=1}^n \sum_{j\neq i} \left(2\gamma + \frac{X_i^\top X_j - \eta_{ij}^\top X_i \eta_{ij}^\top X_j}{\sigma^4(d-1)}\right) k_{\text{imq}}^{4\gamma/2d-3,1/2}(X_i, X_j) \\
&+ \tfrac{1}{n(n-1)} \textstyle\sum_{i=1}^n \sum_{j\neq i} \tfrac{1}{d}\left(\frac{\eta_{ij}^\top X_i \eta_{ij}^\top X_j}{\sigma^4} - \frac{X_i^\top X_j - \eta_{ij}^\top X_i \eta_{ij}^\top X_j}{\sigma^4(d-1)}\right) k_{\text{imq}}^{2\gamma/d,3/2}(X_i, X_j) \\
&+ \tfrac{1}{n(n-1)} \textstyle\sum_{i=1}^n \sum_{j\neq i} \tfrac{1}{d}\left(\frac{2\gamma}{\sigma^2} + 4\gamma^2\right) \|X_i - X_j\|_2^2 k_{\text{imq}}^{2\gamma/d,3/2}(X_i, X_j).
\end{aligned}$$

Then $\mathbb{E}[\texttt{CWKSD}(\widehat{P}_n)] = \mathscr{S}\mathsf{KSD}^2_{k_{\text{gsn}}}(P, \mathcal{N}(0, \sigma^2 I_d)) + o_d(1)$. [Source]

Proof. This result is a direct consequence of the expression for the sliced squared KSD given and proved in Lem. 1. Indeed, CWKSD is simply the U-estimator for the sliced square KSD, noting that

$$\begin{aligned}
{}_1F_1\left(\frac{1}{2}; \frac{d}{2}; -\gamma\|x - x\|_2^2\right) &= k_{\text{imq}}^{4\gamma/2d-3,1/2}(x, x') + o_d(1) \\
{}_1F_1\left(\frac{3}{2}; \frac{d}{2} + 1; -\gamma\|x - x\|_2^2\right) &= k_{\text{imq}}^{2\gamma/d,3/2}(x, x') + o_d(1)
\end{aligned}$$

as described earlier by Knop et al. (2020). □

To complete the proof of Thm. 5, we must establish an expression for the sliced squared KSD, which was used to derive the U-estimator above.

Lemma 1. *Let $\eta_{x'\to x} \triangleq \|x - x'\|_2^{-1}(x - x')$. Then the sliced KSD with kernel k_{gsn} is given by*

$$\begin{aligned}
&\mathscr{S}\mathsf{KSD}^2_{k_{\text{gsn}}}(P, \mathcal{N}(0, \sigma^2 I_d)) \\
&\quad = \mathbb{E}_{X,X'\sim P}\left[\left(2\gamma + \frac{X^\top X' - \eta_{X'\to X}^\top X \eta_{X'\to X}^\top X'}{\sigma^4(d-1)}\right) {}_1F_1\left(\frac{1}{2}; \frac{d}{2}; -\gamma\|X - X'\|_2^2\right)\right] \\
&\qquad + \mathbb{E}_{X,X'\sim P}\left[\frac{1}{d}\left(\frac{\eta_{X'\to X}^\top X \eta_{X'\to X}^\top X'}{\sigma^4} - \frac{X^\top X' - \eta_{X'\to X}^\top X \eta_{X'\to X}^\top X'}{\sigma^4(d-1)}\right) {}_1F_1\left(\frac{3}{2}; \frac{d}{2} + 1; -\gamma\|X - X'\|_2^2\right)\right] \\
&\qquad + \mathbb{E}_{X,X'\sim P}\left[\frac{1}{d}\left(\frac{2\gamma}{\sigma^2} + 4\gamma^2\right) \|X - X'\|_2^2 {}_1F_1\left(\frac{3}{2}; \frac{d}{2} + 1; -\gamma\|X - X'\|_2^2\right)\right].
\end{aligned}$$

Proof. This proof will leverage several hypergeometric integral computations that are given in §F. Let $\mathrm{N} \sim \mathcal{N}(0, \sigma^2 I_d)$. Note that for any $\xi \in \mathbb{S}^{d-1}$, we have $\text{law}(\xi^\top \mathrm{N}) = \mathcal{N}(0, \sigma^2)$. Moreover, it is readily derived that the kernel k_Q for the KSD as written in equation 3 for prior $Q = \mathcal{N}(0, \sigma^2)$ and base kernel k_{gsn} is given by

$$k_Q(u, v) = \left[\frac{uv}{\sigma^4} - \left(\frac{2\gamma}{\sigma^2} + 4\gamma^2\right)(u - v)^2 + 2\gamma\right] \exp(-\gamma(u - v)^2).$$

Recall that $\mathrm{d}\mathcal{U}(\mathbb{S}^{d-1})(\xi) = \frac{\pi^{d/2}}{\Gamma(d/2)}\mathrm{d}\xi \triangleq C_d \mathrm{d}\xi$. Now, the sliced squared KSD takes the following form,

$$\begin{aligned}
&\mathscr{S}\mathsf{KSD}^2_{k_{\mathrm{gsn}}}(P, \mathcal{N}(0, \sigma^2 I_d)) \\
&\qquad = C_d \mathbb{E}_{X,X'\sim P} \int_{\mathbb{S}^{d-1}} \left(\frac{\xi^\top X \xi^\top X'}{\sigma^4} + 2\gamma \right) \exp(-\gamma(\xi^\top (X - X'))^2) \mathrm{d}\xi \\
&\qquad\quad - C_d \left(\frac{2\gamma}{\sigma^2} + 4\gamma^2 \right) \mathbb{E}_{X,X'\sim P} \int_{\mathbb{S}^{d-1}} (\xi^\top (X - X'))^2 \exp(-\gamma(\xi^\top (X - X'))^2) \mathrm{d}\xi \\
&\qquad \triangleq C_d \mathbb{E}_{X,X'\sim P} \left[\mathrm{I}(X, X') + \left(\frac{2\gamma}{\sigma^2} + 4\gamma^2 \right) \mathrm{II}(X, X') \right].
\end{aligned}$$

Recall the functions J_1, J_2 defined in §F. Beginning with I, we have

$$\mathrm{I}(x, x') = \underbrace{\int_{\mathbb{S}^{d-1}} \frac{\xi^\top x \xi^\top x'}{\sigma^4} \exp(-\gamma(\xi^\top (x - x'))^2) \mathrm{d}\xi}_{\mathrm{I}_1(x,x')} + 2\gamma J_1(\gamma \|x - x'\|_2^2).$$

Using the closed form expression for $\mathrm{I}_1(x, x')$ derived in Lem. 5, we have

$$\begin{aligned}
\mathrm{I}(x, x') &= \left(2\gamma + \frac{x^\top x' - e_1^\top x e_1^\top x'}{\sigma^4 (d-1)} \right) J_1(\gamma \|x - x'\|_2^2) \\
&\quad + \left(\frac{e_1^\top x e_1^\top x'}{\sigma^4} - \frac{x^\top x' - e_1^\top x e_1^\top x'}{\sigma^4 (d-1)} \right) J_2(\gamma \|x - x'\|_2^2),
\end{aligned}$$

where $e_1 = \|x - x'\|_2^{-1}(x - x')$. Next, by Lem. 6, we have

$$\mathrm{II}(x, x') = \|x - x'\|_2^2 J_2(\gamma \|x - x'\|^2).$$

Putting these together, and substituting the closed forms for J_1, J_2 due to Lems. 2 and 3, we have

$$\begin{aligned}
&\mathscr{S}\mathsf{KSD}^2(P, \mathcal{N}(0, \sigma^2)) \\
&\quad = C_d \mathbb{E}_{X,X'\sim P} \left[2 \tfrac{\pi^{d/2}}{\Gamma(d/2)} \left(2\gamma + \tfrac{X^\top X' - e_1^\top X e_1^\top X'}{\sigma^4 (d-1)} \right) {}_1F_1\left(\tfrac{1}{2}; \tfrac{d}{2}; -\gamma \|X - X'\|_2^2 \right) \right] \\
&\qquad + C_d \mathbb{E}_{X,X'\sim P} \left[\tfrac{\pi^{d/2}}{\Gamma(d/2+1)} \left(\tfrac{e_1^\top X e_1^\top X'}{\sigma^4} - \tfrac{X^\top X' - e_1^\top X e_1^\top X'}{\sigma^4 (d-1)} \right) {}_1F_1\left(\tfrac{3}{2}; \tfrac{d}{2} + 1; -\gamma \|X - X'\|_2^2 \right) \right] \\
&\qquad + C_d \mathbb{E}_{X,X'\sim P} \left[\tfrac{\pi^{d/2}}{\Gamma(d/2+1)} \left(\tfrac{2\gamma}{\sigma^2} + 4\gamma^2 \right) \|X - X'\|_2^2 {}_1F_1\left(\tfrac{3}{2}; \tfrac{d}{2} + 1; -\gamma \|X - X'\|_2^2 \right) \right].
\end{aligned}$$

The claim follows by expanding C_d and simplifying, noting that

$$\frac{\Gamma(d/2)}{\Gamma(d/2+1)} = \frac{\Gamma(d/2)}{\frac{d}{2}\Gamma(d/2)} = \frac{2}{d}.$$

□

F Hypergeometric Integrals

In this section, we compute various integrals that will be useful throughout the text. Henceforth, we fix $u, v \in \mathbb{R}^d$ and $\gamma > 0$, and we define the following integrals:

$$J_1(c) := \int_{\mathbb{S}^{d-1}} \exp(-c\xi_1^2)\, \mathrm{d}\xi, \quad c \in \mathbb{R}.$$

$$J_2(c) := \int_{\mathbb{S}^{d-1}} \xi_1^2 \exp(-c\xi_1^2)\, \mathrm{d}\xi, \quad c \in \mathbb{R}.$$

Lemma 2. *We have that*

$$J_1(c) = \frac{2\pi^{d/2}}{\Gamma(d/2)} {}_1F_1\left(\frac{1}{2}; \frac{d}{2}; -c \right).$$

Proof. This is the well-known closed form of the Gaussian integral on the sphere. □

Lemma 3. *We have that*

$$J_2(c) = \frac{\pi^{d/2}}{\Gamma(d/2+1)} {}_1F_1\left(\frac{3}{2}; \frac{d}{2}+1; -c\right).$$

Proof. This follows directly from Lem. 6. □

Lemma 4. *Suppose u, v are orthogonal to e_1, the first unit basis vector. Then*

$$\int_{\mathbb{S}^{d-1}} (\xi^\top u)(\xi^\top v)\exp(-c\xi_1^2)\,\mathrm{d}\xi = (u^\top v)\frac{J_1(c)-J_2(c)}{d-1}.$$

Proof. Firstly, note that for $i, j \neq 1$ and $i \neq j$, we have

$$\int_{\mathbb{S}^{d-1}} \xi_i\xi_j \exp(-c\xi_1^2)\,\mathrm{d}\xi = 0$$

since the integrand is odd. Moreover, by symmetry, it is clear that

$$\int_{\mathbb{S}^{d-1}} \xi_i^2 \exp(-c\xi_1^2)\,\mathrm{d}\xi = \int_{\mathbb{S}^{d-1}} \xi_j^2 \exp(-c\xi_1^2)\,\mathrm{d}\xi$$

for any $i, j \geq 2$. Since for any $\xi \in \mathbb{S}^{d-1}$ we have by definition $\sum_{i=1}^d \xi_i^2 = 1$, we have

$$\begin{aligned} J_1(c) &= \sum_{i=1}^d \int_{\mathbb{S}^{d-1}} \xi_i^2 \exp(-c\xi_1^2)\,\mathrm{d}\xi \\ &= J_2(c) + (d-1)\int_{\mathbb{S}^{d-1}} \xi_i^2 \exp(-c\xi_1^2)\,\mathrm{d}\xi \qquad i \geq 2 \end{aligned}$$

$$\int_{\mathbb{S}^{d-1}} \xi_i^2 \exp(-c\xi_1^2)\,\mathrm{d}\xi = \frac{J_1(c)-J_2(c)}{d-1}.$$

We can now relate this result to the integral we wish to compute. Particularly, we have

$$\begin{aligned} \int_{\mathbb{S}^{d-1}} (\xi^\top u)(\xi^\top v)\exp(-c\xi_1^2)\,\mathrm{d}\xi &= \sum_{i=2}^d \sum_{j=2}^d \int_{\mathbb{S}^{d-1}} u_i v_j \xi_i \xi_j \exp(-c\xi_1^2)\,\mathrm{d}\xi \\ &= \sum_{i=2}^d u_i v_i \int_{\mathbb{S}^{d-1}} \xi_i^2 \exp(-c\xi_1^2)\,\mathrm{d}\xi \\ &= u^\top v \frac{J_1(c)-J_2(c)}{d-1}. \end{aligned}$$

□

Lemma 5. *Let $x, x' \in \mathbb{R}^d$. Then define $\mathrm{I}_1(x, x')$ via*

$$\mathrm{I}_1(x,x') = \int_{\mathbb{S}^{d-1}} (\xi^\top x)(\xi^\top x')\exp(-\gamma(\xi^\top(x-x'))^2)\mathrm{d}\xi.$$

It holds that

$$\mathrm{I}_1(x,x') = (x^\top x' - (e_1^\top x)(e_1^\top x'))\frac{J_1(r\gamma^2)-J_2(r\gamma^2)}{d-1} + (e_1^\top x)(e_1^\top x')J_2(r\gamma^2).$$

Proof. Let $\delta = x - x'$, and consider a coordinate system in which δ is parallel to the first basis vector denoted e_1. Denoting by $x_\perp$ the component of x orthogonal to e_1, we have

$$\begin{aligned} \xi^\top x &= \xi_1(e_1^\top x) + \xi^\top x_\perp \\ \xi^\top x' &= \xi_1(e_1^\top x') + \xi^\top x'_\perp. \end{aligned}$$

Letting $r = \|x - x'\|_2$, we have

$$\begin{aligned}\mathrm{I}_1(x, x') = &\underbrace{\int_{\mathbb{S}^{d-1}} \xi_1^2 (e_1^\top x)(e_1^\top x') \exp(-\gamma r^2 \xi_1^2) \mathrm{d}\xi}_{f_1(x,x')} \\ &+ \int_{\mathbb{S}^{d-1}} \xi_1 (e_1^\top x)(\xi^\top x'_\perp) \exp(-\gamma r^2 \xi_1^2) \mathrm{d}\xi \\ &+ \int_{\mathbb{S}^{d-1}} \xi_1 (e_1^\top x')(\xi^\top x_\perp) \exp(-\gamma r^2 \xi_1^2) \mathrm{d}\xi \\ &+ \underbrace{\int_{\mathbb{S}^{d-1}} (\xi^\top x_\perp)(\xi^\top x'_\perp) \exp(-\gamma r^2 \xi_1^2) \mathrm{d}\xi}_{f_2(x,x')} .\end{aligned}$$

Note that the integrands in the middle two terms are odd, so those integrals vanish. Moreover, we have

$$f_1(x, x') \equiv (e_1^\top x)(e_1^\top x') J_2(\gamma r^2).$$

Furthermore, by Lem. 4, we have

$$f_2(x, x') = x_\perp^\top x'_\perp \frac{J_1(\gamma r^2) - J_2(\gamma r^2)}{d-1}.$$

Altogether, we have shown that

$$\mathrm{I}_1(x, x') = (x^\top x' - (e_1^\top x)(e_1^\top x')) \frac{J_1(r\gamma^2) - J_2(r\gamma^2)}{d-1} + (e_1^\top x)(e_1^\top x') J_2(r\gamma^2).$$

□

Lemma 6. *Let $x, x' \in \mathbb{R}^d$. Then it holds that*

$$\begin{aligned}\mathrm{II}(x, x') &:= \int_{\mathbb{S}^{d-1}} (\xi^\top (x - x'))^2 \exp(-\gamma (\xi^\top (x - x'))^2) \mathrm{d}\xi \\ &= \|x - x'\|_2^2 \frac{\pi^{d/2}}{\Gamma(d/2 + 1)} {}_1F_1\left(\frac{3}{2}; \frac{d}{2} + 1; -\gamma \|x - x'\|_2^2\right).\end{aligned}$$

Proof. Let $\delta = x - x'$. Immediately, overloading notation, we have $\mathrm{II}(x, x') \equiv \mathrm{II}(\delta)$, where

$$\mathrm{II}(\delta) := \int_{\mathbb{S}^{d-1}} (\xi^\top \delta)^2 \exp(-\gamma (\xi^\top \delta)^2) \mathrm{d}\xi.$$

Note that II is rotationally symmetric. So, without loss of generality, we can take $\delta = r e_1$, where $r = \|\delta\|_2$ and e_1 is the first standard unit vector in $\mathbb{R}^d$. Then, we have

$$\begin{aligned}\mathrm{II}(\delta) &= \int_{\mathbb{S}^{d-1}} (r\xi_1)^2 \exp(-\gamma (r\xi_1)^2) \mathrm{d}\xi \\ &= r^2 \int_{\mathbb{S}^{d-1}} \xi_1^2 \exp(-\gamma (r\xi_1)^2) \mathrm{d}\xi.\end{aligned}$$

Now we've expressed II as an integral of a function depending on only one coordinate over the unit sphere. Switching to polar coordinates, defining $\xi_1 = \cos\phi$, we have

$$\begin{aligned}\mathrm{II}(\delta) &= r^2 \mathrm{Area}(\mathbb{S}^{d-2}) \int_0^\pi \cos^2\phi \exp(-\gamma (r\cos\phi)^2) \sin^{d-2}(\phi) \mathrm{d}\phi \\ &= r^2 \frac{2\pi^{(d-1)/2}}{\Gamma((d-1)/2)} \int_0^\pi \cos^2\phi \exp(-\gamma (r\cos\phi)^2) \sin^{d-2}(\phi) \mathrm{d}\phi.\end{aligned}$$

Now, writing $u := \xi_1$, we have $\sin^{d-2}(\phi)\mathrm{d}\phi = -(1-u^2)^{(d-3)/2}\mathrm{d}u$, so that

$$\begin{aligned}\mathrm{II}(\delta) &= r^2 \frac{2\pi^{(d-1)/2}}{\Gamma((d-1)/2)} \int_{-1}^{1} u^2 (1-u^2)^{(d-3)/2} \exp(-\gamma r^2 u^2)\mathrm{d}u \\ &= 2r^2 \frac{2\pi^{(d-1)/2}}{\Gamma((d-1)/2)} \int_{0}^{1} u^2 (1-u^2)^{(d-3)/2} \exp(-\gamma r^2 u^2)\mathrm{d}u.\end{aligned}$$

We will now express this in terms of ${}_1F_1$. Recall that

$${}_1F_1(a;b;z) = \frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)} \int_0^1 t^{a-1}(1-t)^{b-a-1} e^{zt}\mathrm{d}t.$$

To relate $\mathrm{II}(\delta)$ to ${}_1F_1$, we write $t := u^2$, so that $\mathrm{d}u = \frac{1}{2}t^{-1/2}\mathrm{d}t$, yielding

$$\mathrm{II}(\delta) = r^2 \frac{2\pi^{(d-1)/2}}{\Gamma((d-1)/2)} \int_0^1 t^{1/2}(1-t)^{(d-3)/2} \exp(-\gamma r^2 t)\mathrm{d}t.$$

Now let $a = 3/2$, $b = d/2 + 1$, and $z = -\gamma r^2$, so that

$$\begin{aligned}\mathrm{II}(\delta) &= r^2 \frac{2\pi^{(d-1)/2}}{\Gamma((d-1)/2)} \int_0^1 t^{a-1}(1-t)^{b-a-1} \exp(zt)\mathrm{d}t \\ &= r^2 \frac{2\pi^{(d-1)/2}}{\Gamma((d-1)/2)} \left[\left(\frac{\Gamma(d/2+1)}{\Gamma(3/2)\Gamma((d-3)/2)} \right)^{-1} {}_1F_1\left(\frac{3}{2}; \frac{d}{2}+1; z\right) \right] \\ &= \|x - x'\|_2^2 \frac{\pi^{d/2}}{\Gamma(d/2+1)} {}_1F_1\left(\frac{3}{2}; \frac{d}{2}+1; -\gamma\|x-x'\|_2^2\right).\end{aligned}$$

□