ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D PHYS

**Semester Project**

# An MMD Approach to Inferring Weak-Lensing Convergence from Galaxy Sizes

Santiago Brunner
sbrunne@student.ethz.ch

Cosmology Group
Institute for Particle Physics and Astrophysics
Department of Physics, D-PHYS
ETH Zurich

**Supervision**

Prof. Dr. Alexandre Refregier
Veronika Oehl

December 8, 2025

# Abstract

Weak gravitational lensing distorts background galaxy images, providing a probe of the universe's large-scale structure. While most studies focus on galaxy shape distortions, we focus on a novel approach to estimate the convergence field using the Maximum Mean Discrepancy (MMD). We employ three kernel functions to construct linear estimators relating the MMD between local and global galaxy size distribution to the convergence. Using a simulated galaxy catologue of 31 million galaxies over 1,000 $\deg^2$, both the RBF and directed RBF achieve strong linear relationships ($R^2 \approx 0.91$) with sign assignment accuracies exceeding 70%.

Our estimators achieve pixel-level correlations of $r \approx 0.75$ with unbiased predictions. However, statistical recovery reveals limitations: predicted maps show elevated scatter (20-30%), small deviations from Gaussianity, and overestimated power spectra at the largest scales. Despite these limitations, we demonstrate that MMD-based methods can extract convergence information from galaxy size distributions.

# Contents

# 1 Introduction

Einsteins theory of general relativity describes how matter bends the structure of space and time. This curvature also affects the trajectories of photons, causing distant light sources to appear distorted. This phenomenon is known as gravitational lensing. Depending on the strength of the effect, we distinguish between weak and strong gravitational lensing. In this project, we study the weak lensing regime, focusing on light from background galaxies whose images are subtly altered as they pass through curved spacetime.

Lensing leads to two main observable effects. The shear describes the change in shape, while the convergence describes the change in size. While the shear of background galaxies was first detected in the early 1990's [10], and can now be and has been measured reliably [6, 11], here we focus on estimating the convergence. In the weak lensing regime, the effect is so subtle that it cannot be detected in a single source. Instead, we rely on statistical analysis using large samples of galaxies to identify small, systematic changes in their size distribution. This report presents an approach to estimate weak gravitational lensing convergence from the observed sizes of galaxies.

# 2 Theory

## 2.1 Weak Lensing and Power Spectra

Weak gravitational lensing describes the subtle distortions imprinted in the images of distant galaxies as their light traverses inhomogeneous gravitational fields. These distortions are quantified primarily by two effects: shear, which alters the galaxy shapes, and the convergence, which changes their apparent sizes. In this work, our attention is centered on the convergence $\kappa$. The convergence field results from the projection of the matter density along the line-of-sight and can be represented as a two-dimensional map on the sky, called convergence map. The observed angular size $\theta_{\text{obs}}$ of a galaxy, given its intrinsic size $\theta_0$, is modified by convergence through the following relation:

$$\theta_{\text{obs}} = \theta_0(1 + \kappa). \tag{1}$$

This is a first order approximation, which only holds, if $|\kappa| \ll 1$, which is true in the weak lensing regime. In this case, $\theta_{\text{obs}}$ is also independent of the shear [1].

The convergence field on larger scales can be modeled as a Gaussian random field and can be entirely characterized by its power spectrum $C_\ell$, which describes how fluctuations in convergence vary with angular scale (multipole $\ell$).

The convergence power spectrum $C_\ell^\kappa$ can be computed applying Limber's approximation in Fourier space to obtain

$$C_\ell^\kappa = \frac{9}{4}\left(\frac{H_0}{c}\right)^4 \Omega_{m,0}^2 \int_0^{\chi_\infty} d\chi \left[\frac{g(\chi)}{a(\chi)}\right]^2 P_\delta\big(k = \frac{\ell}{\chi}, \chi\big), \tag{2}$$

where $P_\delta$ is the matter power spectrum, $g(\chi)$ a weighting function called lensing efficiency, $\chi$ the comoving radial distance, and $H_0$ and $\Omega_{m,0}$ are the present value of the Hubble constant and matter density parameter, respectively [9]. Note that equation (2) holds for the auto-correlation case of the convergence field, where the same lensing efficiency $g(\chi)$ enters twice in the integral.

Since the power spectrum is directly determined by cosmological parameters, its measurement provides a powerful means to constrain those parameters and refine our understanding of the fundamental properties of our universe.

## 2.2 Maximum Mean Discrepancy

The Maximum Mean Discrepancy (MMD) is a non-parametric statistical measure that quantifies the difference between two probability distributions using kernel methods [8]. It works by mapping each distribution into a mean embedding. Given a kernel function $k$ and distributions $P$, and $Q$, their embeddings are computed as averages of the kernel evaluated at sample points. The MMD is defined as the distance between these mean embeddings:

$$\mathrm{MMD}_k(P, Q) = ||\mu_P - \mu_Q||_{\mathcal{H}}, \tag{3}$$

where $\mu_P$ and $\mu_Q$ are the kernel mean embeddings and $||\cdot||_{\mathcal{H}}$ is the norm in the reproducing kernel Hilbert space (RKHS) $\mathcal{H}$. A kernel is said to be *characteristic*, if the mean embedding is injective. A characteristic kernel ensures that the distance is zero if and only if the two distributions are the same.

In practice, we estimate the MMD from finite samples [2]. Given data points $\{x_i\}_{i=1}^n \sim P$ and $\{y_j\}_{j=1}^m \sim Q$, the empirical MMD is:

$$\mathrm{MMD}_k^2(\hat{P}, \hat{Q}) = \frac{1}{n^2}\sum_{i,j} k(x_i, x_j) + \frac{1}{m^2}\sum_{i,j} k(y_i, y_j) - \frac{2}{nm}\sum_{i,j} k(x_i, y_j) \tag{4}$$

This approach allows comparison of complex distributions without strong assumptions, making MMD a powerful tool for statistical analysis [8, 2].

# 3 Modelling

## 3.1 Galaxy Catalogue

As already mentioned, we need a large sample of galaxies to reliably detect weak lensing effects. Since our goal is to construct an estimator for the convergence, we require a catalogue in which the convergence values for each galaxy are known. To create such a dataset, we use simulations based on the GalSBI model [5]. For each galaxy, the absolute magnitude M and the redshift z are sampled from a single Schechter luminosity function. The size of a galaxy is then determined by its half light radius $r_{50}$, i.e. the radius at which half of the galaxy's luminosity is contained. For a sampled absolute magnitude M, the size $r_{50}$ is sampled from a log-normal distribution to ensure it is always positive and matches observations. Finally, the positions of the galaxies are sampled randomly. The catalogue was split into redshift bins, from which we use the one with the largest mean in this project. The resulted catalogue used in this project consists of 31 million galaxies spread out over a patch of sky of $1,000$ deg$^2$.

## 3.2 Convergence Maps

In the above section, we have only described how we get the intrinsic galaxy properties. By constructing a convergence map we obtain the observed sizes. In this project, we model the convergence map as a Gaussian random field, with a correlation structure given by the power spectrum. To create the power spectrum, we used the Core Cosmology Library (CCL) [3] with the redshift bin obtained from the catalogue and a ΛCDM cosmology with Planck parameters [4]. In practice, we obtain the convergence map using the HEALPix python package [12], particularly the synfast method. Throughout this project, we worked with two $n_{\text{side}}$ parameters, which quantify the resolution in healpy. The coarser map used an $n_{\text{side}}$ of 64 while the finer map used 1,024. Finally, we obtain the observed angular sizes of the galaxies by applying formula (1) to each individual galaxy. I.e. each galaxy will experience a different convergence, depending on the pixel in which it is positioned.

# 4 Statistical Analysis

## 4.1 Different Convergence Maps

In our analysis, the convergence map is discretized using the HEALPix pixelization scheme, which divides the sky into equal-area pixels specified by the resolution parameter $n_{\text{side}}$. Each pixel represents a distinct region of the simulated sky and the $n_{\text{side}}$ value determines the number and size of the pixels, with higher $n_{\text{side}}$ producing finer, smaller pixels.

For the coarser convergence map ($n_{\text{side}}$=64), each pixel covers a large area, ensuring that a sufficient number of galaxies are available within each pixel for a robust statistical analysis.

To achieve greater spatial resolution reflective of real survey data, we also employ a finer convergence map ($n_{\text{side}}$=1024). In this case, galaxies are assigned convergence values corresponding to their precise pixel location in the finer map, capturing more detailed variations across the sky. However, for computational efficiency and consistency, the MMD analysis is still performed by aggregating galaxies within the larger pixels of the coarse map, while their assigned values come from the higher-resolution map. Thus, each coarse pixel contains galaxies that experience slightly different convergence values.

## 4.2 Application of MMD

The main statistical tool used in this work is the Maximum Mean Discrepancy (MMD) introduced in section 2.2. We used it to measure how the local galaxy size distributions change through lensing.

### 4.2.1 Kernel Choices

As described in section 2.2, the MMD computes the distance in the RKHS. The computed value is therefore dependent on the choice of kernel. In this project we worked with three different kernels: linear kernel, RBF kernel and a custom signature kernel defined as the product of the sign of the difference between the inputs and the RBF kernel:

$$k(x, y) = \operatorname{sgn}(x - y) \cdot \exp\left(-\frac{(x - y)^2}{2\sigma^2}\right). \tag{5}$$

Notice that the last kernel is not positive semi definite, which is in contrast to the definition of the MMD. The reason for the choice of this kernel is to be able to detect the direction of the shift of the distribution. Even though the quantity resulting from

this kernel is not an MMD in the mathematical sense, we will not make a distinction here.

The linear kernel, defined as $k(x, y) = x^\top y$, i.e. the standard inner product in the input space, was chosen, because the MMD simplifies to the shift in the means of the distributions:

$$\mathrm{MMD}^2 = (\mu_P - \mu_Q)^2$$

We could then use the linear kernel to reproduce the results of [7]. Finally, the RBF kernel was chosen for its sensitivity to non-linear changes in the distribution. This is particularly important when using the finer convergence map ($n_{\mathrm{side}} = 1024$), where galaxies within a coarse pixel experience varying convergence values, leading to more complex distributional shifts beyond simple mean changes.

### 4.2.2 Computing the MMD

The MMD is computed to quantify the difference between the global galaxy size distribution and the size distribution within individual pixels of the convergence map. Since the convergence field is constructed to have zero mean, global lensing should not alter the overall size distribution – only the local distributions within pixels are affected.

For each pixel in the coarse convergence map ($n_{\mathrm{side}} = 64$), we randomly sample 20,000 galaxy sizes both from the global catalogue and from all galaxies located within the pixel. This sample size ensures statistical significance [7] while keeping computations tractable.

For the linear kernel, we restricted our analysis to the coarse convergence map and calculated the MMD as outlined above. This served primarily as a sanity check to verify that the MMD could successfully detect shifts in the size distributions. Consequently, we did not pursue further analysis with the linear kernel.

For the RBF and custom kernels, we apply the convergence values from the finer map ($n_{\mathrm{side}} = 1024$) to assign each galaxy a more refined $\kappa$. Galaxies within a coarse pixel thus have slightly different convergence values. We compare the global size distribution to the distribution formed by all galaxies within each coarse pixel, taking into account their assigned detailed convergence values. Additionally, we compute the average convergence, $\bar{\kappa}$, in each coarse pixel to link the local MMD results to the corresponding lensing strength.

## 4.3 Constructing an Estimator

To make predictions about the convergence $\kappa$ in a given region of the sky based on the computed MMD, we require an estimator linking the two quantities. From equation (1), a linear relationship between the $\mathrm{MMD}^2$ value and $\kappa^2$ is anticipated. However, since the MMD is defined to be non-negative, it does not indicate the direction of the shift and thus cannot distinguish between positive and negative $\kappa$ on its own.

To circumvent this limitation, we assign a sign to each computed MMD based on the sign of the mean-shift between the global and local size distribution. This approach allows us to construct a "signed MMD", which captures both the magnitude and direction of the convergence-induced shift. We then use this signed MMD to perform a linear fit against the averaged convergence values, providing an estimator capable of predicting $\kappa$ in different regions of the field.

With the custom kernel, we do not have this issue as it intrinsically captures the direction of the distribution shift. In that case we already have a signed MMD and proceed as described above.

## 4.4 Recovery of a Convergence Map

Using the constructed estimator, we establish a direct relationship between the galaxy size distribution – quantified by the MMD within each coarse pixel – and the average convergence $\kappa$ in that region. According to equation (1), this relation is expected to be linear. This will be validated in section 5.

By inverting the found linear estimator, we predict the convergence $\kappa^2_{\mathrm{pred}}$ from the computed MMD values as

$$\kappa^2_{\mathrm{pred}} = \frac{\mathrm{MMD}^2 - b}{m}, \tag{6}$$

where $m$ and $b$ are the slope and intercept from the linear fit. Notably, the $\mathrm{MMD}^2$ can take negative values, allowing $\kappa^2_{\mathrm{pred}}$ to also be negative, thus resolving any ambiguity in the prediction's sign.

We calculate $\kappa^2_{\mathrm{pred}}$ for each pixel of the coarse map, enabling a pixel-by-pixel comparison between predicted and averaged true convergence values. Even if the values of the pixels don't match exactly, we might still be able to recover the underlying statistics and therefore the cosmological parameters. In particular, we want to recover the properties of a Gaussian random field with mean zero.

Furthermore, we aim to recover the convergence power spectrum, a key descriptor for the Gaussian random fields, since the two-point function fully characterizes the field. To do so, we first compute $\kappa^2_{\text{pred}}$ for each pixel as described above. We can then extract $\kappa_{\text{pred}}$ by taking the square root of $|\kappa^2_{\text{pred}}|$ and assigning the right sign, $\text{sgn}(\kappa^2_{\text{pred}})$. We can form a convergence map from $\kappa_{\text{pred}}$. This can then be used to directly get the power spectrum. Here, we used the function anafast from the HEALPix python library [12].

Given our galaxy catalogue covers only a 1,000deg$^2$ sky patch, the resulting map is masked, which limits complete power spectrum recovery. Consequently, the estimated power spectrum is a pseudo power spectrum reflecting the partial sky coverage.

# 5 Results

## 5.1 Estimators

Since our galaxy catalogue covers only a patch of the sky of 1,000 deg$^2$, and the catalogue has finitely many galaxies, there were some pixels that did not include enough galaxies to allow a statistically significant computation of the MMD. Therefore, we set a threshold for the minimal number of galaxies in a pixel of 20,000 galaxies. Pixels with less galaxies were not considered. Additionally, we also correct for too large galaxies by applying a size threshold of 5 arcsec, such that the few outliers could be excluded. Both of these mentioned thresholds follow the work in [7].

To make sure that our MMD-approach can actually detect the mean shift, we created an estimator, where the convergence in each coarser pixel ($n_{\text{side}}$=64) is constant. The resulted fit can be seen in Figure 1 and the corresponding fitting parameters in Table 1. The noise was computed with the MMD of two random samples of the global distribution, each of size 20,000. To get the mean and standard deviation, we repeated this procedure 15,000 times, resulting in:

$$\mu_{\text{noise}} = (2.93 \pm 4.16) \cdot 10^{-6} \tag{7}$$

We get a signal above the noise threshold for $\kappa^2 > 2.39 \cdot 10^{-5}$.

The MMD$^2$-values computed with the two different non-linear kernels can be seen plotted in Figure 2 as the red data points, where the x-axis is the corresponding $\kappa^2$-value with the appropriate sign. We can see that the regions further away from the origin tend to have a larger variance. This is why we decided to divide the $\kappa^2$-

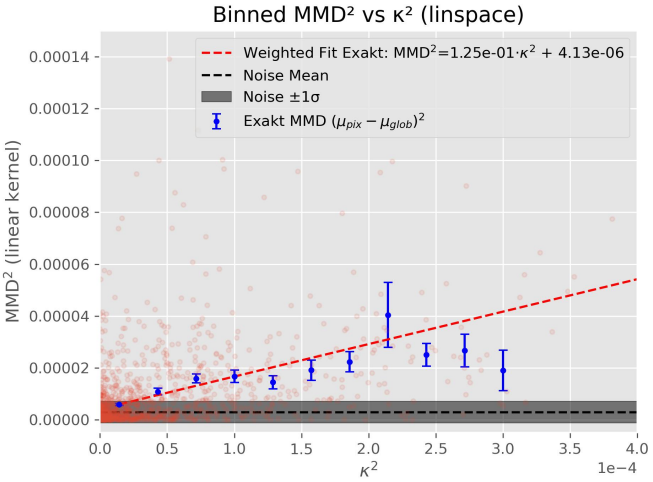



**Figure 1:** Linear estimator for the linear kernel. The red dots are the individual $\mathrm{MMD}^2$ for each pixel. The blue dots are the average $\mathrm{MMD}^2$ for each $\kappa^2$ bin.

values into equal distance bins, as already done for the linear kernel. By averaging the $\mathrm{MMD}^2$-values over the bins, we could reduce the variance, as can be seen in the blue data points. The errors of these data points were obtained by $\sigma/\sqrt{N}$, where $\sigma$ is the standard deviation of the $\mathrm{MMD}^2$-values and N is the number of values in that bin. This way, we performed a weighted linear fit assuming Gaussian noise, where each data point is weighted by the inverse of its variance. For a data point with standard error $\sigma_i$, the weight is given by $w_i = 1/\sigma_i^2$. Therefore, data points with smaller uncertainties contribute more to the fit, while points with larger uncertainties are down-weighted accordingly.

The fitting parameters for the different kernels can be found in Table 1, where the errors were obtained from the covariance matrix.

**Table 1:** Fitting parameters for the relationship between signed MMD and convergence $\kappa$ for different kernels.

| Kernel | Slope ($m$) | Intercept ($b$) |
|---|---|---|
| Linear | $(1.25 \pm 0.15) \cdot 10^{-1}$ | $(4.13 \pm 0.23) \cdot 10^{-6}$ |
| RBF | $0.288 \pm 0.013$ | $(1.61 \pm 0.25) \cdot 10^{-6}$ |
| Dir. RBF | $104.0 \pm 11.0$ | $(-2.72 \pm 0.31) \cdot 10^{-3}$ |

Notice that the values of different kernels in Table 1 cannot be compared, since the MMDs for different kernels are distances in different and unrelated RKHS.

To quantify the quality of the linear fits, we also computed the $\mathrm{R}^2$-value:

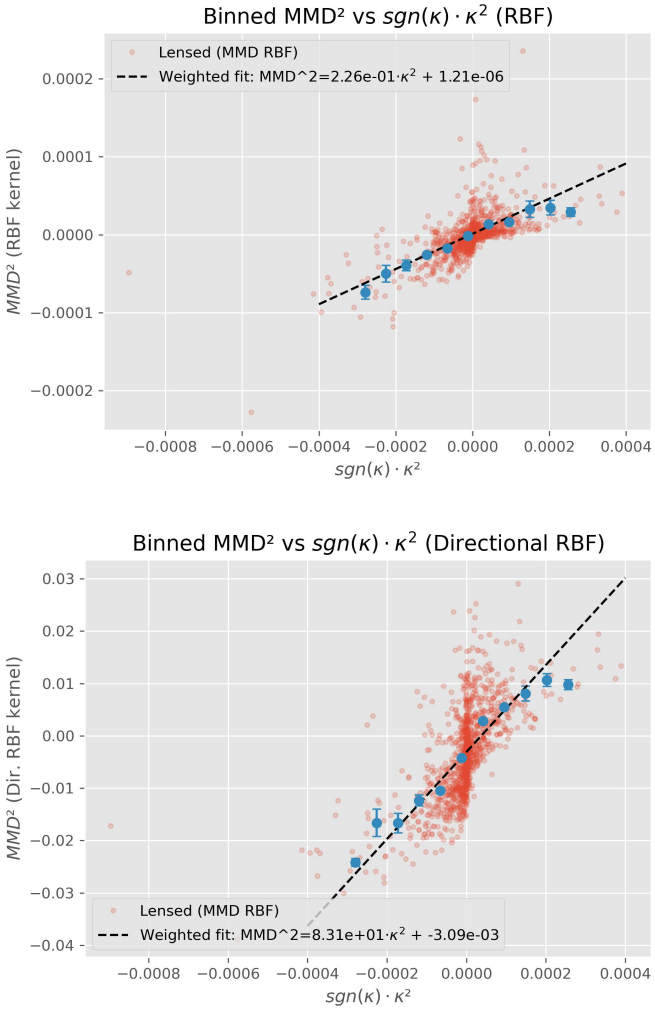



**Figure 2:** Resulted estimators (black dashed line) given the $\mathrm{MMD}^2$ for the RBF (top) and the directed RBF kernel (bottom). The blue dots are the average $\mathrm{MMD}^2$ in each $\kappa^2$ bin with its standard deviation.

$$R^2_{\text{RBF}} = 0.91 \tag{8}$$

$$R^2_{\text{Dir.RBF}} = 0.92 \tag{9}$$

Although $R^2_{\text{Dir.RBF}}$ suggests a very good fit, we can see in the bottom panel of Figure 2 that the estimator is biased as its intercept is quite large.

In section 4.3 we described the procedure of assigning the right sign to the $\text{MMD}^2$-values, either by the mean-shift in the case of the RBF kernel or intrinsically through the custom kernel. We can now evaluate how many signs were correctly assigned, i.e. match the corresponding sign of $\kappa$. From the 1,145 pixels over which the MMDs were computed, the mean-shift assigned 852 signs correctly, while for the custom kernel it were 817, corresponding to 74.41% and 71.35%, respectively.

## 5.2 Recovery of the Convergence Map

We first obtained the results on pixel-level. The results can be seen in Figure 3 for the RBF on the top and the custom kernel at the bottom. For each kernel, we can see in the top left plot the pixel wise comparison between the true and predicted $\kappa$ for each coarse pixel. The red dashed line is the 1:1 line and corresponds to a perfect prediction. Both kernels achieve similar correlations as described by their Pearson correlation coefficient (PCC) $r \approx 0.75$. The top right panel shows the distribution of residuals. Both estimators lead to a residual distribution with mean very close to zero, but specially for the RBF kernel, the distribution is clearly shifted towards negative values. The lower left panel shows the residuals plotted against the true convergence values. For the RBF, the residuals are constant along the values of $\kappa_{\text{true}}$, although not symmetric around zero. There are more negative residuals (706 of 1,145), but the positive residuals are scattered further from zero. In contrast the residuals of the directional RBF kernel are mostly positive for small $\kappa_{\text{true}}$, and get more negative with increasing $\kappa_{\text{true}}$.

Finally, the bottom right panel summarizes the pixel-level statistics.

Figure 4 shows the true, predicted and residual convergence maps from left to right for the RBF kernel (top row) and the custom kernel (bottom row). Grey pixels either lie outside our sky patch or do not contain sufficiently many galaxies to compute a statistically significant MMD. The true convergence map exhibits the expected random structure of a Gaussian random field. The predicted maps broadly reproduce this structure, though systematic deviations are visible in the upper region


**Figure 3:** Pixel-level validation of convergence recovery for the RBF kernel (top) and directed RBF kernel (bottom). Upper left: Predicted vs true convergence with 1:1 line. Upper right: Distribution of residuals with Gaussian fit overlay. Bottom left: Residuals as a function of true convergence. Bottom right: Summary statistics.

of the patch, where the predicted values tend to be too large. This over-prediction is clearly seen in the residual maps (right panels), which show predominantly red (positive residual) pixels in that region. Elsewhere, the residuals are mostly small and randomly distributed, indicating good recovery.

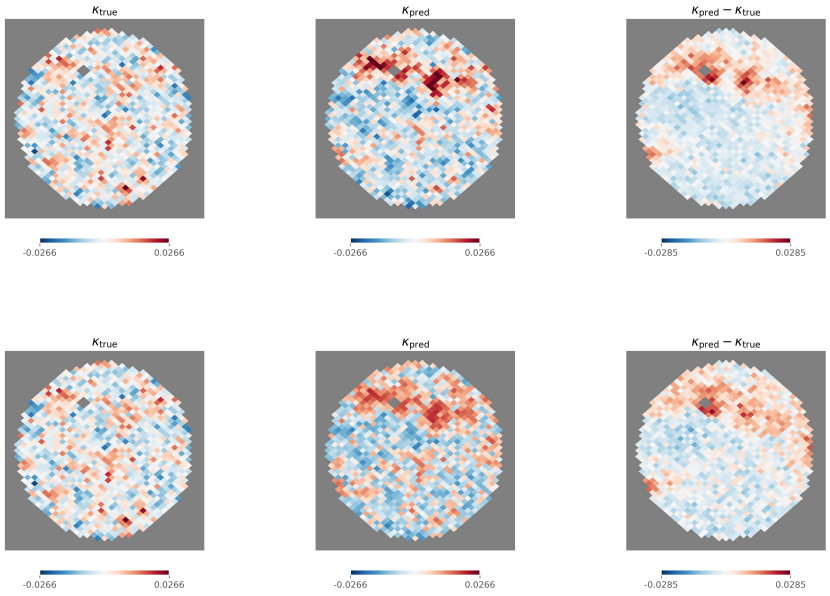



**Figure 4:** Spatial distribution of convergence for the RBF kernel (top row) and directed RBF (bottom row). Left: True convergence. Center: Predicted convergence. Right: Residuals ($\kappa_{\text{pred}} - \kappa_{\text{true}}$).

Now we turn to the recovery of the statistics. The true convergence map is a Gaussian random field with zero mean. We assess whether the predicted maps recover this Gaussianity. Table 2 summarizes the first four moments of the predicted convergence distributions compared to the true map. Both predicted maps achieve means close to zero confirming no overall bias. However, they show strong deviations from Gaussianity in the higher-order moments.

**Table 2:** Summary statistics of the convergence distribution for the true map, RBF and directed RBF kernels.

| | Mean | Std. Dev. | Skewness | Kurtosis |
|---|---|---|---|---|
| True | $-6.7 \cdot 10^{-4}$ | $7.1 \cdot 10^{-3}$ | $1.2 \cdot 10^{-1}$ | $6.6 \cdot 10^{-2}$ |
| RBF | $-1.2 \cdot 10^{-3}$ | $8.9 \cdot 10^{-3}$ | $6.5 \cdot 10^{-1}$ | $5.1 \cdot 10^{-1}$ |
| Dir. RBF | $-5.4 \cdot 10^{-4}$ | $9.3 \cdot 10^{-3}$ | $2.4 \cdot 10^{-1}$ | $-1.3$ |

The $\kappa$ distribution of the recovered convergence maps can be found in Figure 5, where the red histogram corresponds to the true convergence map for comparison.

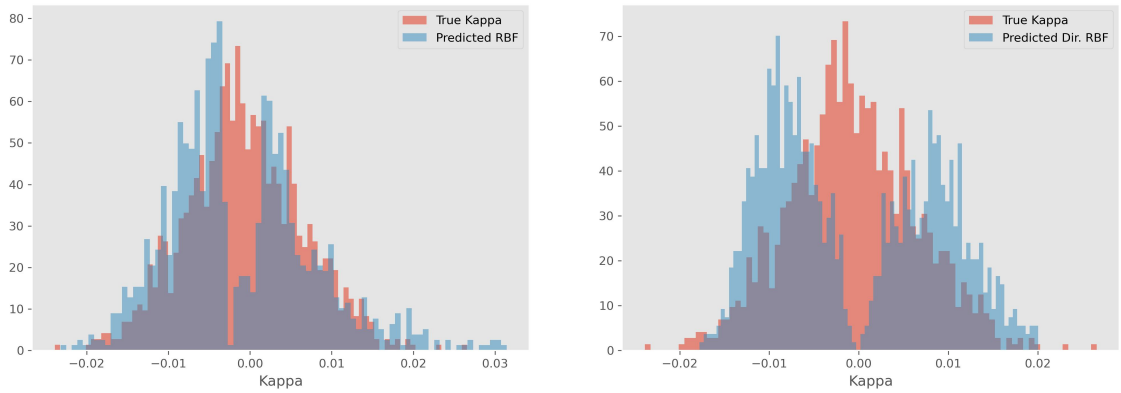



**Figure 5:** Distribution of convergence values (blue) for the RBF kernel (left) and directed RBF kernel (right). Red histograms show the true convergence distribution for comparison.

A notable feature is the artificial gap near zero in both predicted distributions. Neither kernel can reliably produce values close to zero, resulting in a bimodal appearance. This gap arises because the MMD-approach cannot detect very small changes in the distribution: even pixels with small $|\kappa|$ are pushed toward slightly positive or negative predictions, creating a forbidden region around $\kappa \approx 0$. This effect, while visible in the histograms, represents the inherent difficulty in finding the true convergences when the lensing signal is comparable to statistical noise.

Finally, we obtained the power spectra of the recovered convergence maps. Figure 6 shows the comparison between the recovered spectra from both kernels (blue and orange), the power spectrum from a perfect recovery (purple), and the pseudo-power spectrum (red) expected from our partial sky coverage. Both the RBF and custom kernels produce power spectra lying significantly above the perfect recovery spectrum on large scales. For higher multipoles (from $\ell \approx 40$) the shape of the different spectra match.

# 6 Discussion

An important part of this project was the sign assignment for the $\text{MMD}^2$, since this is crucial for an unambiguous convergence estimator. In section 5, we saw that both approaches, the mean shift and the custom kernel, can indeed predict the correct sign reliably in over 70% of the cases. Even though the mean-shift approach worked a bit better, it is certainly a remarkable result that the custom kernel achieves comparable results. However, both methods misassign approximately 25-30% of pixels. This can be attributed to several factors. First, in regions where $\kappa$ is close

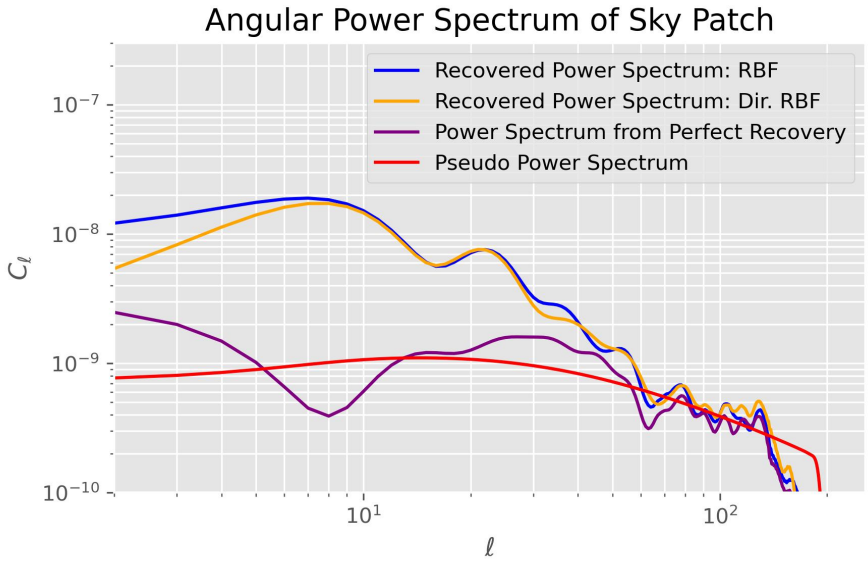



**Figure 6:** Angular power spectrum of the recovered convergence maps compared to the true power spectrum (purple) and the pseudo-power spectrum (red) accounting for partial sky coverage. Both the RBF (blue) and directed RBF (orange) kernels overestimate the power on small scales.

to zero, the lensing-induced shift in the size distributions is small compared to the scatter, making sign determination very noisy. Second, the finite number of galaxies per pixel introduces statistical fluctuations that can obscure the true direction of the shift.

Since the MMD measures distances on different and unrelated RKHS for different kernels, we cannot compare the fitting parameters of the estimators directly. Instead, we focus on the $R^2$ score to asses the quality of the fits. For both kernels, we used the same convergence map to construct the respective estimators. While Figure 2 shows the expected linear relation for small $|\kappa^2|$, the deviation from the linear fit increases slightly as $|\kappa^2|$ grows. This deviation is similar for both kernels, leading also to very similar $R^2$ scores.

The pixel-level analysis demonstrates that the MMD-based estimator successfully captures the overall trend of the convergence field. Both kernels achieve strong correlations ($r \approx 0.75$) between predicted and true convergence values. The residual analysis confirms estimators with mean residuals consistent with zero. However, the spatial distribution of the residuals (Figure 4) reveals systematic over-prediction in the upper region of the sky batch. Notice that the only grey pixels containing insufficiently many galaxies are in that exact region. A possible reason for the large deviation from the true map could therefore be a lower number of galaxies in that

region.

The artificial band near zero in the scatter plots of Figure 3 reflects the incapability of measuring small $\mathrm{MMD}^2$ values, creating a forbidden region for predictions close to zero.

Beyond pixel-level accuracy, the recovery of the correct statistical properties is crucial for cosmological applications. We assess this through the distribution moments and the angular power spectrum.

The statistical moments reveal both successes and limitations of the MMD-based estimator. The recovery of near-zero mean for both kernels (Table 2) confirms the absence of overall bias, which is critical for unbiased cosmological inference. However, the predicted maps show a higher standard deviation (approximately 20-30%) than the true map. This indicates that the estimator adds scatter beyond the intrinsic convergence field, reducing the signal-to-noise ratio of the recovered map. The higher-order moments reveal more significant deviations from Gaussianity. Both kernels show elevated skewness and excess kurtosis that deviate substantially from the near-zero values expected for a Gaussian distribution. These departures have multiple origins: the artificial gap near zero creates a bimodal-like distribution, while the systematic over-prediction in certain spatial regions contributes to the asymmetry.

The recovered power spectra show the limitations of our estimators on large scales, as they significantly overestimate the power. This deviation arises from our finite sky patch, that only has limited information about these scales. Unfortunately, we were not able to recover multiple convergence maps to assess the uncertainty of our estimators. But it is expected that the uncertainty is very large for small $\ell$, which is consistent with the results in [7]. Other factors contributing to this excess power are the increased number of pixels with similar, small $\kappa_{\mathrm{pred}}$, due to the fact that the pixels with $\kappa_{\mathrm{true}} \approx 0$ are pushed away from zero. Also, the over-prediction in certain spatial regions introduces coherent structures that are not present in the true convergence field, contributing correlated power rather than random noise.

Importantly, both kernels produce nearly identical results across all metrics despite their different sign assignment approaches, suggesting that the observed limitations are systematic features of the MMD-based framework rather than kernel-specific issues. While the power spectrum on large scales could not be recovered, the match on smaller scales is encouraging.

# 7 Conclusion and Outlook

In this work, we developed an MMD-based estimator for the weak lensing convergence field from galaxy size distributions. We employed three kernels: linear, RBF, and a custom directed RBF kernel. The linear kernel validated our approach, confirming consistency with previous work [7]. The RBF kernel achieved $R^2 = 0.91$, demonstrating a strong linear relationship between signed $MMD^2$ and convergence. The directed RBF kernel achieved $R^2 = 0.92$ while intrinsically capturing the sign of $\kappa$ without requiring separate mean-shift computation.

Both sign assignment methods correctly identified the convergence direction in over 70% of pixels, with misassignments primarily occurring in low-$|\kappa|$ regions where statistical noise dominates.

On the pixel level, both kernels achieved strong correlations ($r \approx 0.75$) with unbiased predictions, successfully capturing local convergence trends. However, statistical recovery revealed some limitations: the predicted maps showed elevated scatter (20-30% higher standard deviation), slight deviations from Gaussianity, and overestimated power spectra, particularly at large scales where our finite sky patch provides limited constraints. Many of these limitations stem from the failed detection of small $|\kappa|$, which creates an artificial gap near $\kappa \approx 0$. Developing kernels or methods that better recover low-convergence regions could significantly improve statistical recovery.

We have demonstrated that the MMD provides a viable framework for convergence estimation, extending beyond mean-based approaches to capture more general distributional changes. Several directions could improve and extend this work. The directed RBF kernel represents a first attempt at intrinsic sign capture; future work could explore alternative asymmetric kernels, kernel combinations, or learned kernels optimized for convergence estimation. Our 1,000 $deg^2$ analysis reflects current survey scales, though upcoming surveys with >10,000 $deg^2$ coverage would better constrain large-scale power. Future work should incorporate realistic observational effects such as size measurement errors to develop estimators applicable to real data.

# References

[1] Matthias Bartelmann and Matteo Maturi. Weak gravitational lensing, 2016.

[2] Farah Cherfaoui, Hachem Kadri, Sandrine Anthoine, and Liva Ralaivola. A Discrete RKHS Standpoint for Nyström MMD. working paper or preprint, April 2022.

[3] Nora Elisa Chisari, David Alonso, Elisabeth Krause, C. Danielle Leonard, Philip Bull, Jeremy Neveu, Antonio Villarreal, Sukhdeep Singh, Thomas McClintock, John Ellison, Zilong Du, Joe Zuntz, Alexander Mead, Shahab Joudaki, Christiane S. Lorenz, Tilman Troester, Javier Sanchez, Francois Lanusse, Mustapha Ishak, Renee Hlozek, Jonathan Blazek, Jean-Eric Campagne, Husni Almoubayyed, Tim Eifler, Matthew Kirby, David Kirkby, Stephane Plaszczynski, Anze Slosar, Michal Vrastil, and Erika L. Wagoner. Core cosmology library: Precision cosmological predictions for lsst. *The Astrophysical Journal Supplement Series*, 242(1):2, May 2019.

[4] Planck Collaboration. Planck 2018 results: Vi. cosmological parameters. *Astronomy amp; Astrophysics*, 641:A6, September 2020.

[5] Silvan Fischbacher, Tomasz Kacprzak, Luca Tortorelli, Beatrice Moser, Alexandre Refregier, Patrick Gebhardt, and Daniel Gruen. Galsbi: phenomenological galaxy population model for cosmology using simulation-based inference. *Journal of Cosmology and Astroparticle Physics*, 2025(06):007, June 2025.

[6] Nick Kaiser, Gordon Squires, and Tom Broadhurst. A method for weak lensing observations. *The Astrophysical Journal*, 449:460, August 1995.

[7] Noah Kirchhoff. Estimating weak lensing convergence from galaxy sizes. Semester Project, ETH Zurich, Cosmology Group, Institute for Particle Physics and Astrophysics, Department of Physics, D-PHYS, May 2025.

[8] Zahra Mehraban and Alois Pichler. Quantization of probability measures in maximum mean discrepancy distance, 2025.

[9] Alexandre Refregier. Weak gravitational lensing by large-scale structure. *Annual Review of Astronomy and Astrophysics*, 41(1):645–668, September 2003.

[10] J. A. Tyson, F. Valdes, and R. A. Wenk. Detection of Systematic Gravitational Lens Galaxy Image Alignments: Mapping Dark Matter in Galaxy Clusters. , 349:L1, January 1990.

[11] Angus H. Wright, Benjamin Stoelzner, Marika Asgari, Maciej Bilicki, Benjamin Giblin, Catherine Heymans, Hendrik Hildebrandt, Henk Hoekstra, Benjamin Joachimi, Konrad Kuijken, Shun-Sheng Li, Robert Reischke, Maximilian von Wietersheim-Kramsta, Mijin Yoon, Pierre Burger, Nora Elisa Chisari, Jelte de Jong, Andrej Dvornik, Christos Georgiou, Joachim Harnois-Deraps, Priyanka Jalan, Anjitha John William, Shahab Joudaki, Giorgio Francesco Lesci, Laila Linke, Arthur Loureiro, Constance Mahony, Matteo Maturi, Lance Miller, Lauro Moscardini, Nicola R. Napolitano, Lucas Porth, Mario Radovich, Peter Schneider, Tilman Troester, Edwin Valentijn, Anna Wittje, Ziang Yan, and Yun-Hao Zhang. Kids-legacy: Cosmological constraints from cosmic shear with the complete kilo-degree survey. *Astronomy & Astrophysics*, October 2025.

[12] Andrea Zonca, Leo Singer, Daniel Lenz, Martin Reinecke, Cyrille Rosset, Eric Hivon, and Krzysztof Gorski. healpy: equal area pixelization and spherical harmonics transforms for data on the sphere in python. *Journal of Open Source Software*, 4:1298, 03 2019.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

## Declaration of originality

The signed declaration of originality is a component of every written paper or thesis authored during the course of studies. **In consultation with the supervisor**, one of the following two options must be selected:

☐ I hereby declare that I authored the work in question independently, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used no generative artificial intelligence technologies[1].

☒ I hereby declare that I authored the work in question independently. In doing so I only used the authorised aids, which included suggestions from the supervisor regarding language and content and generative artificial intelligence technologies. The use of the latter and the respective source declarations proceeded in consultation with the supervisor.

**Title of paper or thesis**:

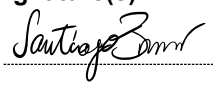
An MMD Approach to Inferring Weak-Lensing Convergence from Galaxy Sizes

**Authored by**:
*If the work was compiled in a group, the names of all authors are required.*

| **Last name(s):** | **First name(s):** |
|---|---|
| Brunner | Santiago |
| | |
| | |
| | |

With my signature I confirm the following:
- I have adhered to the rules set out in the Citation Guidelines.
- I have documented all methods, data and processes truthfully and fully.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for originality.

| **Place, date** | **Signature(s)** |
|---|---|
| Zurich, 08.12.2025 |  |
| | |
| | |
| | *If the work was compiled in a group, the names of all authors are required. Through their signatures they vouch jointly for the entire content of the written work.* |

[1] For further information please consult the ETH Zurich websites, e.g. https://ethz.ch/en/the-eth-zurich/education/ai-in-education.html and https://library.ethz.ch/en/researching-and-publishing/scientific-writing-at-eth-zurich.html (subject to change).