Learning Curves for Deep Neural Networks: A Gaussian Field Theory Perspective

Omry Cohen, Or Malka, and Zohar Ringel
Racah Institute of Physics,
Hebrew University of Jerusalem,
Jerusalem 91904, Israel.
zohar.ringel@mail.huji.ac.il

Abstract

A series of recent works suggest that deep neural networks (DNNs), of fixed depth, are equivalent to certain Gaussian Processes (NNGP/NTK) in the highly over-parameterized regime (width or number-of-channels going to infinity). Other works suggest that this limit is relevant for real-world DNNs. These results invite further study into the generalization properties of Gaussian Processes of the NNGP and NTK type. Here we make several contributions along this line. First, we develop a formalism, based on field theory tools, for calculating learning curves perturbatively in one over the dataset size. For the case of NNGPs, this formalism naturally extends to finite width corrections. Second, in cases where one can diagonalize the covariance-function of the NNGP/NTK, we provide analytic expressions for the asymptotic learning curves of any given target function. These go beyond the standard equivalence kernel results. Last, we provide closed analytic expressions for the eigenvalues of NNGP/NTK kernels of depth 2 fully-connected ReLU networks. For datasets on the hypersphere, the eigenfunctions of such kernels, at any depth, are hyperspherical harmonics. A simple coherent picture emerges wherein fully-connected DNNs have a strong entropic bias towards functions which are low order polynomials of the input.

1 Introduction

Several pleasant features underlay the success of deep learning: The scarcity of bad minima encountered in their optimization [Draxler et al. (2018); Choromanska et al. (2014)], their ability to generalize well despite being heavily over-parameterized [Neyshabur et al. (2018, 2014)] and expressive [Zhang et al. (2016)], and their ability to generate internal representations which generalize across different domains and tasks [Yosinski et al. (2014); Sermanet et al. (2013)].

Due to the complexity of DNNs our current understanding of these features is still largely empirical. Notwithstanding, it was recently proven [Daniely et al. (2016); Jacot et al. (2018); Mei et al. (2018)] that keeping the DNN depth fixed and taking the number of parameters to infinity, learning in DNNs becomes analytically tractable. This is due to the fact that the networks’ parameters, in all non-linear layers, change in an important yet minor manner during training. Several subsequent works corroborated these results experimentally (Lee et al. (2018); Lee et al. (2019); Ben-David & Ringel (2019)) and verified that they hold at least for some real-world networks [Lee et al. (2019)].

The emerging picture is that DNNs trained with GD (full batch SGD) and an MSE loss are mappable to Bayesian Inference on a Gaussian Processes (GP) known as the Neural Tangent Kernel (NTK) [Jacot et al. (2018)]. Similarly DNNs trained using SGD at small learning rates are mappable to Bayesian Inference on a GP related to randomized DNN (an NNGP) [Lee et al. (2018); Ben-David & Ringel (2019)]. According to this, DNNs are merely a smart way of designing GP covariance-
functions and a computationally efficient way of performing large scale inference with GP, which circumvents the costly matrix inversions involved (see below). An additional advantage is that internal representations are accessible in the DNN context, facilitating transfer learning.

While theoretically solid, the applicability of these results to real-world networks have been questioned by Chizat & Bach (2018) and dubbed "Lazy learning". For instance in AlexNet [Krizhevsky et al. (2012)], the parameters of the first convolutional layer undergo major changes during training despite having quite a large number of channels (96). It may thus be that AlexNet operates in a qualitatively different regime or possibly that one can gradually go from lazy-learning to feature-learning without drastically changing the overall predictions of the DNN.

With these developments in mind, it would be useful to have a clear, flexible, and systematic way of calculating learning curves for NNGPs and NTKs as well as means of departing from the Gaussian limit and accounting for finite width corrections. This would help explain to what extent does the GP viewpoint explain the aforementioned "pleasant features", and to what extent is lazy-learning different than feature-learning.

In this work we make the following contributions. (1) We derive closed analytic expressions for the eigenvalues and eigenfunctions of NNGP and NTK covariance-functions on the hypersphere. This is relevant to fully-connected networks and datasets wherein each datapoint has L2 norm of 1. Employing standard equivalence kernel (EK) results, we show how this explains generalization in fully-connected networks in the large dataset limit. (2) We develop a formalism based on statistical field theory and replicas, which allows for systematic departure from the large dataset limit and the infinite-width/channels limit. (3) Our formalism is then used to derive corrections to the learning curve which go beyond EK results and allow for further systematic improvement.

2 Background

Gaussian Processes are non-parametric models on which Bayesian inference can be performed analytically, up to a matrix inversion [Rasmussen & Williams (2005)]. The dimension of this matrix is the size of the dataset. The GP itself can be viewed as a Gaussian probability distribution over a space of functions \( f(x) \). In the DNN context the function \( f(x) \) describes the output of DNN given an input \( x \) [Cho & Saul (2009)] and can be viewed as an indirect representation of DNN’s parameters. Such GPs, known as NNGP, have a clear meaning as representing the distribution over functions generated by the DNN at initialization. The NTK type GPs [Jacot et al. (2018)], have a similar yet slightly less direct interpretation as a certain sum of the NNGP along with a different NNGP wherein the DNNs activation function is replaced by its derivative. Notably the mapping between DNNs and GPs is more natural for regression problems but can also be applied to classification problems [Lee et al. (2018)].

Next we review the state-of-the-art on GP Bayesian Inference learning-curves. Learning-curves are graphs describing the generalization error as a function of data-set size \( (N) \), averaged on all possible datasets of size \( \tilde{N} \). Average learning-curves are learning-curves which are furthered averaged over all target functions/labels drawn from some prior distribution. Here we focus on explicit learning-curves wherein the resulting expression can be evaluated directly and one avoids the aforementioned matrix inversions which is computationally-costly and analytically obscure. For similar reasons we focus on cases where one has knowledge on the eigenvalues and eigenfunctions (features) of the covariance-function. Considering average learning curves, analytic expressions are scarce except for very simple GPs [Rasmussen & Williams (2005)]. Assuming that the prior over targets matches that of the GPs (and an MSE error) lower bounds were obtained in [Opper & Vivarelli (1999)]. Considering (non-averaged) learning-curves, using the Bayesian-PAC approach along with the Laplace approximation, Seeger (2003) managed to estimate learning-curves with a 7.5% error on binary MNIST with RBF covariance-functions. Considering MSE loss, large datasets \((N)\), and large noise, one has the EK result [Sollich & Williams (2005)]. The EK approximation replaces the discrete sum over data-points by an integral. This uncontrolled approximation can easily lead to false estimates (for instance at zero noise it yields zero error regardless of \( N \), see below). In contrast, here we derive EK type results and improve upon them in a systematic manner.

2
3 Field Theory Formulation of GP learning-curves

Here we develop a field theory formalism for exploring learning curves. We begin with standard definitions of GPs and Bayesian Inference on GPs. Being Gaussian, the probability distribution on function \( f(x) \) drawn from GPs are determined by its first and second moments. The first is typically taken to be zero and second is known as the covariance function \( (K_{xx} = \mathbb{E}[f(x)f(x')] \), where \( \mathbb{E} \) denotes expectation under the GP distribution). Notably, \( K_{xx} \) of both the NNGP and NTK type can be calculated analytically for many activation functions [Cho & Saul (2009); Jacot et al. (2018)]. Furthermore, Bayesian Inference on GPs drawn from DNNs is tractable [Lee et al. (2018); Cho & Saul (2009)] and explicitly given by

\[
g_* = \sum_{n,m} K_{x_n x_m} [K(D) + \sigma^2 I]^{-1}_{nm} g_m
\]

where \( x_* \) is a new datapoint, \( g_* \) is the prediction, \( g_m \) are the training targets, \( x_n \) are the training data-points, \( [K(D)]_{nm} = K_{x_n x_m} \) is the covariance-matrix (the covariance-function projected on the training dataset \( D \)), \( \sigma^2 \) is a regulator corresponding to a noisy measurement of the GP and \( I \) is the identity matrix. Some intuition for this formula can be gained by verifying that \( x_* = x_q \) yields \( g_* = g_q \) when \( \sigma^2 = 0 \).

While the above equation determines the predictions and therefore the learning-curves, it does not do so in any clear or computationally accessible manner. This fact is due to the (potentially very) large matrix inversion involved, and the additional averaging over \( D \) required.

To facilitate the analysis of Eq. 1 we turn to Feynman-path-integral, or equivalently to statistical field theory [Schulman (1996)]. These are well-studied, powerful approaches for performing integrations over a space of functions (the jargon is "paths" when \( x \) in one dimensional and "fields" when \( x \) is higher dimensional). To get some familiarity with this formalism, consider first averages over the (centered) GP itself with no dataset. Using the path-integral formalism we write it as

\[
P_0[f] = \frac{\exp \left( -\frac{1}{2} \int d\mu_x d\mu_{x'} f(x) K^{-1}(x,x') f(x') \right)}{\int Df \exp \left( -\frac{1}{2} \int d\mu_x d\mu_{x'} f(x) K^{-1}(x,x') f(x') \right)}
\]

where \( \int Df \) denotes integration over the space of functions, \( d\mu_x = P(x)dx \) is a probability measure over the inputs space and \( K^{-1}(x,x') \) is the inverse covariance function \( (\int d\mu_x K(x,x')K^{-1}(x',x'') = \delta(x-x'')/P(x)) \). To define the path-integrals one first chooses an orthonormal basis of functions \( \phi_i(x) \) (with respect to the measure \( \mu_x \)) arranged in order of likeliness \( P_0[\phi_i] \geq P_0[\phi_j] \) for \( i > j \) (note that this comparison doesn’t require calculating the path integral in (2)). Second, one expands \( f = \sum_i f_i \phi_i(x) \), and defines the path-integral as a series of simple integrals

\[
\int Df \mathcal{F}[f] = \int df_1 \int df_2 \ldots \mathcal{F} \left[ \sum_i f_i \phi_i \right]
\]

where \( \mathcal{F} \) is some functional of \( f \). Finally, one makes this last expression well-defined by taking a limit procedure where the number of integrals is gradually taken to infinity [Schulman (1996)].

Performing the above procedure we show in App. 2. that

\[
K_{x_1 x_2} = \int Df P_0[f] f(x_1) f(x_2)
\]

Notably, all other higher correlation functions split into products of the above correlation function due to standard properties of Gaussian integrals (Wick’s/Isserlis’ theorem).

Following a similar procedure, and denoting \( \|f\|_K^2 = \int d\mu_x d\mu_{x'} f(x) K^{-1}(x,x') f(x') \) one can show [Rasmussen & Williams (2005)]

\[
g_* = Z^{-1} \int Df \cdot f(x_*) \cdot \exp \left( -\frac{1}{2} \|f\|_K^2 - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (f(x_n) - g_n)^2 \right)
\]

\[
Z = \int Df \exp \left( -\frac{1}{2} \|f\|_K^2 - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (f(x_n) - g_n)^2 \right)
\]
where \( Z \) is known as the partition function.

The next obstacle is dealing with the fact that expressions like \( Z \) appear in the numerator and denominator of Eq. (5). To by-pass this difficulty we introduce a source term into \( Z \)

\[
Z[\alpha] = \int Df \exp \left( -\frac{1}{2} \|f\|^2_K - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (f(x_n) - g_n)^2 + \int dx f(x)\alpha(x) \right)
\]

(6)

where \( \alpha(x) \) is an arbitrary function of \( x \). Next we note that

\[
g_* = \frac{\partial}{\partial \alpha(x_\ast)} \log[Z[\alpha]]
\]

(7)

Notably, the functional derivative \( \frac{\partial}{\partial \alpha(x_\ast)} \) can be understood through a discretization limit: For discrete \( x \), \( \alpha(x) \) becomes a vector whose dimension is the number of discrete points, and the functional derivative means a standard partial derivative with respect to the discrete point \( x_\ast \).

The advantage of the above formulation follows from the identity

\[
\log[Z[\alpha]] =_{M \to 0} (Z[\alpha])^M - 1
\]

(8)

and from the use the standard "replica trick" (see for instance [Gardner & Derrida (1988)]), where we assume \( M \) to be a non-negative integer, and extrapolate the limit at \( M \to 0 \).

The generalization error is defined as \( \langle (g(x_\ast) - g_\ast(x_\ast))^2 \rangle_{x_\ast \sim \mu} \). Therefore, in order to calculate learning-curves, one needs to be able to average quantities (like \( g_\ast \) and \( g_\ast^2 \)) over all datasets of size \( N \) drawn from a probability distribution \( d\mu_x = P(x)dx \). We denote this averaging by \( \langle \ldots \rangle_{\mu,N} \). To facilitate this we begin by relaxing the requirement of fixed \( N \) and instead consider a related quantity given by the Poisson averaging of the former one

\[
\langle \ldots \rangle_{\mu,\eta} = e^{-\eta} \sum_{n=0}^{\infty} \frac{\eta^n}{n!} \langle \ldots \rangle_{\mu,n}
\]

(9)

where \( \ldots \) can be any quantity, in particular \( g_\ast \) and \( g_\ast^2 \). Taking \( \eta = N \), means we are essentially averaging over values of \( N \) which in a \( \sqrt{N} \) vicinity of \( N \). Following this we find that for a non-negative integer \( M \), \( (Z[\alpha]^M)_{\mu,\eta} \) can be re-written as

\[
(Z[\alpha]^M)_{\mu,\eta} = e^{-\eta} \int Df_1 \ldots Df_M \exp \left( -\frac{1}{2} \sum_{m=1}^{M} ||f_m||^2_K + \eta \int d\mu_x e^{-\frac{1}{2\sigma^2} \sum_{m=1}^{M} (f_m(x) - g(x))^2} + \sum_{m=1}^{M} \int dx f_m(x)\alpha(x) \right)
\]

(10)

where notably \( M \) "replicas" of \( f(x) \) emerged. Delegating the formal derivation to App. 4, here we just write the main observation which is that the \( n \)th order taylor expansion of the exponent in \( \eta \int d\mu_x \ldots \) yields \( \frac{\eta^n}{n!} \int d\mu_{x_1} \ldots d\mu_{x_n} Z[\alpha]^M \) which is exactly \( \frac{\eta^n}{n!} (Z[\alpha]^M)_{\mu,N,n} \) as expected.

Therefore, substituting Eq. (8) into Eq. (7) and averaging using Eq. (9), \( \langle g_\ast \rangle_{\mu,\eta} \) can be written as

\[
\langle g_\ast \rangle_{\mu,\eta} = \lim_{M \to 0} M^{-1} \frac{\partial}{\partial \alpha(x_\ast)} \left( (Z[\alpha]^M)_{\mu,\eta} \right)
\]

\[
= \lim_{M \to 0} M^{-1} \int Df_1 \ldots Df_M \exp \left( -\frac{1}{2} \sum_{m=1}^{M} ||f_m||^2_K + \eta \int d\mu_x e^{-\frac{1}{2\sigma^2} \sum_{m=1}^{M} (f_m(x) - g(x))^2} \right) \sum_{m=1}^{M} f_m(x_\ast)
\]

(11)

The main benefit of Eq. (11) over Eq. (1) is that it allows for a controlled expansion in \( 1/\eta \). At large \( \eta \) (or similarly large \( N \)) we expect that the fluctuations in \( f_m(x) \) would be small and centered around \( g(x) \) (from a Bayesian point of view, the posterior would become very close to the target). Indeed
such a behavior is encouraged by the term multiplied by $\eta$ in the exponent. That be the case we can systematically taylor expand

$$
\int d\mu_x e^{-\sum_{m=1}^{M} (f_m(x) - g(x))^2 / 2\sigma^2} = 1 - \int d\mu_x \sum_{m=1}^{M} (f_m(x) - g(x))^2 / 2\sigma^2 + \frac{1}{2} \int d\mu_x \left[ \sum_{m=1}^{M} (f_m(x) - g(x))^2 \right]^2 + ...
$$

(12)

as shown in App. 4, dealing with the first order term in this expansion in an exact (Gaussian) manner yields the standard EK type results. The second order term and further terms render the theory non-Gaussian and cannot be dealt with exactly but rather through standard perturbation-theory/Feynman-diagrams. In App. 4 we show that $\langle g_\ast \rangle_{\mu,\eta}$ is given by

$$
\langle g_\ast \rangle_{\mu,\eta} = f^{EK}_{\eta,\sigma^2}(x_\ast) + \frac{\eta}{\sigma^4} \int d\mu_x \left( f^{EK}_{\eta,\sigma^2}(x) - g(x) \right) \text{Cov}[x,x] \text{Cov}[x,x] + O\left( \frac{1}{\eta^3} \right)
$$

(13)

where $f^{EK}_{\eta,\sigma^2}(x)$ is the standard EK result (with $\eta$ replacing $N$) and

$$
\text{Cov}[x,y] = \frac{\int Df \cdot f(x) \cdot f(y) \cdot \exp \left( -\frac{1}{2} \|f\|^2_K - \frac{\eta}{2\sigma^2} \int d\mu_x f(x')^2 \right)}{\int Df \exp \left( -\frac{1}{2} \|f\|^2_K - \frac{\eta}{2\sigma^2} \int d\mu_x f(x')^2 \right)}
$$

Assuming we know how to diagonalize $K(x,x')$, using a set of features $\phi_i(x)$ and eigenvalues $\lambda_i$

$$
\int d\mu_x K(x,x') \phi_i(x') = \lambda_i \phi_i(x)
$$

(14)

Eq. (13) becomes

$$
\langle g_\ast \rangle_{\mu,\eta} = f^{EK}_{\eta,\sigma^2}(x_\ast) - \frac{\eta}{\sigma^4} \sum_{i,j,k} \frac{\sigma^2}{\eta} \left( \frac{1}{\lambda_j} + \frac{\eta}{\sigma^2} \right)^{-1} \left( \frac{1}{\lambda_k} + \frac{\eta}{\sigma^2} \right)^{-1} g_i \phi_j(x_\ast) \int d\mu_x \phi_i(x) \phi_j(x) \phi_k^2(x) + O\left( \frac{1}{\eta^3} \right)
$$

(15)

Notably, similar expressions for $\langle g_\ast^2 \rangle_{\mu,\eta}$ can be achieved by using two replica indices

$$
\langle g_\ast^2 \rangle_{\mu,\eta} = \lim_{M \to 0} \lim_{M' \to 0} M^{-1} M'^{-1} \frac{\partial}{\partial \alpha(x_\ast)} \left| \langle (Z[\alpha])^M (Z[\hat{a}])^{M'} \rangle_{\mu,\eta} \right|
$$

(16)

### 3.1 Finite Width

While the focus of this work is on obtaining learning-curves perturbatively in $1/N$, the field-theory formalism can be used for finite-width corrections as we briefly explain.

The NNGPs reflect the distribution over functions generated by DNNs at initialization [Cho & Saul (2009); Ben-David & Ringel (2019)]. It is Gaussian essentially because each of the output neurons sums over many uncorrelated activations from the preceding layer. The central limit theorem (CLT) thus predicts a Gaussian behavior. The careful reader might worry at this point, that while NNGPs are defined by the DNN regardless of a particular dataset, $P_0[f]$ requires knowledge of the dataset’s measure $d\mu_x$ both explicitly through the integrals and implicitly through the definition $K^{-1}$. Interestingly, as implied by Eq. 4, all correlation functions based on $P_0[f]$ do not depend on the measure hence $P_0[f]$ itself does not depend on the measure [see also Rasmussen & Williams (2005) ch. 6.1] and may represent an NNGP.

As the width ($W$) becomes finite, corrections to CLT appear. The most dominant ones, in $1/W$, would change $P_0[f]$ into $P_0[f]/(1 + \frac{1}{W} \int d\mu_x d\mu_y d\mu_z d\mu_4 U(x_1, x_2, x_3, x_4)f(x_1)f(x_2)f(x_3)f(x_4))$, where $U(x_1..x_4)$ is some function of the DNNs architecture. In principal, $U(x_1..x_4)$ can be calculated using similar techniques to Cho & Saul (2009), and is closely related to the forth cumulant. Given $U(x_1..x_4)$, perturbation theory can be carried in a straightforward manner as done here. For a study of finite-width corrections to NTKs see [Ethan & Gur-Ari (2019)]
4 Features and Eigenvalues of Fully-Connected Networks

To make the result in Eqs. (13,15) explicit and interpretable, one needs to find $\phi_i(x)$ and $\lambda_i$. This can be done most readily for the case of datasets normalized to the hypersphere ($\|x_n\| = 1$), a uniform probability measure on the hypersphere and rotation-symmetric covariance functions, meaning $K(x, x') = K(0x, O x')$ for any $O$, where $O$ is an orthogonal matrix over the space of inputs.

Importantly, the NNGP and NTK GPs associated with any network whose first layer is fully-connected, have this symmetry under rotations. Indeed, let $h_w(x)$ be the output vector of the first layer $[h_w(x)]_l = \phi(\sum_j w_{lj} x_j + b)$ where $x_j$ is the $j$'th component of the input vector $x$. Let $z_{w'}(h)$ be the output of the rest of the network given $h$. The covariance function of NNGPs are defined by [Cho & Saul (2009)]

$$K(x, x') = \int dw dw' P_0(w, w') z_{w'}(h_w(x)) z_{w'}(h_w(x'))$$

where $P_0(w, w')$ is a prior over the weights, typically taken to be i.i.d Gaussian for each layer $(P_0(w, w') = P_0(w) P_0(w')$ and $P_0(w) \propto e^{-\sum_i w_{ij}^2/(2\sigma_j^2)}).$ Following this one can show

$$K(0x, O x') = \int dw dw' P_0(w, w') z_{w'}(h_w(O x)) z_{w'}(h_w(O x')) = \int dw dw' P_0(w, w') z_{w'}(h_w(x)) z_{w'}(h_w(x'))$$

$$= \int dw dw' P_0(wO^T, w') z_{w'}(h_w(x)) z_{w'}(h_w(x')) = \int dw dw' P_0(w, w') z_{w'}(h_w(x)) z_{w'}(h_w(x')) = K(x, x')$$

where the second equality uses the definition of $h_w(x)$, the third results from an orthogonal change of integration variable $w \to wO$, and the forth is a property of our prior over $w$.

Considering the particular case of a depth 2 network with one hidden ReLU layer, variance $\sigma_{w_i}^2$ at the $i$th layer and no biases, we have that $K(x, x')$ in the (infinite) wide limit can be written as [Cho & Saul (2009)]

$$K(x, x') = \sigma_{w_1}^2 \sigma_{w_2}^2 \frac{1}{2\pi} \|x\| \|x'\| \left[ \frac{\pi}{2} (x \cdot x') + \sum_{k=0}^\infty \frac{\Gamma(k - \frac{1}{2})}{2\sqrt{\pi} k! (2k - 1)} (x \cdot x')^{2k} \sin \theta + (\pi - \theta) \cos \theta \right]$$

Notably, the appearance of only $\|x\| \|x'\|$ and $x \cdot x'$ and their powers in the above expansion is not a coincidence, as these are the only rotationally-invariant quantities one can make out of two vectors.

Turning to the NTK type kernels, these are always made of sums of the NNGP type GPs so they inherit this symmetry as well.

Assuming uniformly distributed (on the sphere) normalized data, the covariance function is only a function of power of $x \cdot x'$. Dot-product kernels can be diagonalized using hyperspherical harmonics $(Y_{lm}(x))$ [Azvedo & Menegatto (2015)], which are a complete (and orthonormal w.r.t a uniform measure) basis for functions on the hypersphere. For each $l$ these can be written as a sum of polynomials in the $x_j$'s of degree $l$. The extra index $m$ enumerates an orthogonal set of such polynomials. Note that usually the hyperspherical harmonics are normalized w.r.t Lebesgue measure on the hypersphere, but in this context the normalization is w.r.t a probability measure on the hypersphere. For a kernel of the form $K(x, x') = \sum_m b_m (x \cdot x')^m$ the eigenvalues are independent of $m$ and given by [Azvedo & Menegatto (2015)]

$$\lambda_l = \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi} \cdot 2^d} \sum_{s=0}^{\infty} b_{2s+l} \frac{(2s + l)!}{(2s)!} \frac{\Gamma\left(s + \frac{d}{2}\right)}{\Gamma\left(s + l + \frac{d}{2}\right)}$$

where $d$ is the dimension of the input (meaning that the inputs space is $S^{d-1}$).

Importantly, as any covariance function of an NNGP or NTK type GP for fully-connected network (regardless of depth, width ratios or non-linearity) can be expanded in this manner, we find that the eigenvalues are given by (20) and the eigenfunctions are always simply $\phi_i(x) = \phi_{lm}(x) = Y_{lm}(x)$. 

6
Notwithstanding the eigenvalues will vary between DNNs. For the case of Eq. (19) we get
\[
\lambda_{2k} = \frac{\sigma_{w_1}^2 \sigma_{w_2}^2}{2\pi} \cdot \frac{d}{8\pi} \left( \frac{\Gamma\left(k - \frac{3}{2}\right) \Gamma\left(\frac{d}{2}\right)}{\Gamma\left(k + \frac{d+1}{2}\right)} \right)^2, \quad \lambda_{2k+1} = \frac{\sigma_{w_1}^2 \sigma_{w_2}^2}{2\pi} \cdot \frac{\pi}{d} \delta_{k,0}
\] (21)

Interestingly, without biases we find that such ReLU networks do not contain any \(l = 3, 5, 7\ldots\) terms and so they are inadequate for any target function which does contain those features. The NTK covariance function of the same network has the following eigenvalues
\[
\lambda_{2k} = \frac{\sigma_{w_1}^2 \sigma_{w_2}^2}{2\pi} \cdot \frac{d(1 + 2k) + (1 - 2k)^2}{8\pi} \left( \frac{\Gamma\left(k - \frac{3}{2}\right) \Gamma\left(\frac{d}{2}\right)}{\Gamma\left(k + \frac{d+1}{2}\right)} \right)^2, \quad \lambda_{2k+1} = \frac{\sigma_{w_1}^2 \sigma_{w_2}^2}{2\pi} \cdot \frac{\pi}{d} \delta_{k,0}
\] (22)

5 Generalization in fully-connected networks

Here we combine the results of the two preceding sections to obtain a clear quantitative picture of learning curves and generalization in fully-connected networks. Let us consider first the standard EK result for fully-connected networks. We first expand our target \(g(x)\) into hyperspherical harmonics \(g(x) = \sum_{l,m} g_{lm} Y_{lm}(x)\) so the EK approximation yields
\[
g_\ast = \sum_{l,m} \frac{\lambda_l}{\lambda_l + \sigma^2/\eta} g_{lm} Y_{lm}(x_\ast)
\] (23)

Thus the GP prediction essentially cuts out all \(\lambda_l \ll \sigma^2/\eta\). Notably for input dimension \(d \gg 1\), Eqs. (21,22) both predict a very sharp decrease of \(\lambda_l \approx O(d^{-1})\) for \(l \ll d\). Consequently there is typically a well defined threshold value \(l_\ast\) for which \(\lambda_{l_\ast} \gg \sigma^2/\eta\) and \(\lambda_{l > l_\ast} \ll \sigma^2/\eta\). Taking this into account along with the fact that \(Y_{lm}\)‘s with \(l \leq l_\ast\) span all polynomials up to degree \(l_\ast\), one finds that the infinitely wide ReLU network with one hidden layer essentially finds the best polynomial of degree equal or less than \(l_\ast\) that matches the data.

One immediate use of the correction evaluated in Eq. (15), is to estimate the range of validity of the EK results. Here we find \(\eta_{\min} \approx (C_{K,\eta,\sigma^2} - \sigma^2)/\lambda_{\max}\), where \(C_{K,\eta,\sigma^2} = \sum_{l,m} \left( \frac{1}{\lambda_l} + \frac{\sigma^2}{\eta} \right)^{-1}\) (summing also over the degeneracy of \(\lambda_l\)) and \(l_{\max}\) is the \(l\) that maximizes \(\sum_m g_{lm}^2\).

This sharp cut-off on the degree of the estimated target generalizes to all fully-connected networks. Indeed note that \(K(x,x)\) is finite and therefore its integral over the hypersphere is also finite and given by \(\int d\mu K(x,x) = K(x,x) = \sum_{l,m} \lambda_l\), and \(K(x,x)\) does not depend on \(x\). The degeneracy (number of \(m\)‘s per \(l\), denoted \(\text{deg}(l)\)) is fixed from properties of hyperspherical harmonics, and equals \(\text{deg}(l) = 2^{l+1} d^d (l+1)! / (l+1)! (d-2)!\) [Frye & Efthimiou (2012)] which goes as \(O(d^l)\) for \(l \ll d\). Noting that \(\lambda_l\) are all positive, to ensure a finite \(K(x,x)\), it must be that \(\lambda_l < K(x,x)/\text{deg}(l) \leq K(x,x)\) \(O(d^{-1})\).

The following pictures emerges of what fully-connected DNNs do on datasets which are uniform on the sphere: They are simply a means of performing polynomial regression with polynomials of degree less than \(O(\log(N)/\sigma^2)\) while sparing us the computationally costly \(N\) by \(N\) matrix inversion. Below we argue that following some data augmentations, the same may very well hold for real-world datasets.

6 Numerical Verification

We turn to numerically test the accuracy of our prediction for the generalization error. We consider a target function \(g(x)\) made out of hyperspherical harmonics with \(l = 1, 2\) and the spectral weight distributed equally between the eigenspaces. We set \(||g(x)|| = 1||\).

We generated datasets \(D_N\) of \(N\) uniformly distributed points on the unit sphere \(S^{d-1}\). For each \(N = 1, 2, \ldots, N_{\max}\) we calculated the MSE of the Bayesian inference prediction of the target given \(D_N\) and \(g(D_N)\), as in Eq. (1), and averaged over many \(D_N\) draws. This quantity is the average MSE over datasets of size \(n = \langle\text{MSE}\rangle_n\). To compare with our analytic predictions, we performed Poisson averaging of \(\langle\text{MSE}\rangle_N\) in accordance with Eq. (9), truncating the summation at \(N_{\max}\) and taking \(N_{\max} - \eta_{\max} \ll \sqrt{\eta_{\max}}\) to avoid any noticeable truncation issues. The resulting quantity, \(\langle\text{MSE}\rangle_\eta\),
Figure 1: Solid lines indicate the experimental generalization error obtained by averaging the MSE of the Bayesian prediction over many datasets, them Poisson averaging with parameter $\eta$. The hyper-parameters for both kernels were chosen to be $\sigma^2_{w_1} = \sigma^2_{w_2} = 1$ and the input is in $S^{50-1}$. Dotted lines indicate the traditional EK prediction of the generalization error, which is clearly inferior to the PC prediction (dashed line). The inset on the left demonstrates an inverted learning curve and how the PC prediction follows it.

is shown in Figure 1. Our EK and pertubatively corrected EK (PC) predictions for the generalization error are plotted for comparison.

One immediately notices that both predictions converge to the experimental learning curve when $\sigma^2/\eta \ll 1$, and that the PC prediction is a notable improvement over the standard EK result. Another interesting feature is evident in the small $\eta$ regime of our results. It has been shown [Sollich (2001)] that under certain conditions generalization error can actually increase with size of the training set in the small $N$ regime, as is demonstrated by our experimental data. This effect is completely missed by the EK prediction of the learning curve, which is strictly monotonically decreasing. The corrected PC predictions shows potential of capturing it.

Notwithstanding, a real-world dataset would not be randomly distributed. Still one can perform PCA to extract its leading principal components and standardize these to have unit variance. As this all amounts to a linear transformation, it is plausible that such data augmentation would not hamper DNN performance. At the level of co-variances, this augmented dataset would now appear random and therefore one may speculate that the eigenvalues of the covariance-function sampled on the augmented dataset would resemble those obtained by random sampling. In App. 3, such a procedure is performed on the CIFAR-10 dataset. Both the random dataset and the augmented dataset are finally normalized to the sphere. Rather strikingly the difference in eigenvalues are minute. This raises hope that our results are more broadly applicable.

7 Discussion and Outlook

In this work we developed a formalism based on field theory tools for predicting learning-curves. Such a formalism may facilitate the transfer of knowledge between the physics community and ML community. Focusing on fully-connected networks and datasets which are uniform on the sphere, we derived closed expression for the features and eigenvalues of NNGP and NTK type GPs. Combining these results, along with the correspondence with DNNs, enabled us to explain why such over-parameterized networks generalize well and estimate their learning-curves. To the best of our knowledge, our learning-curves improve the current state-of-the-art on GP and DNN learning-curves. Our formalism can be used to derive further corrections and/or finite-width corrections, both of which are left for future work. Another interesting direction to pursue concerns the features and eigenvalues of CNN.
References

Azevedo, D. and Menegatto, V. A. Eigenvalues of dot-product kernels on the sphere. *ArXiv e-prints*, 2015.

Ben-David, O. and Ringel, Z. The role of a layer in deep neural networks: a Gaussian Process perspective. *arXiv e-prints*, art. arXiv:1902.02354, Feb 2019.

Chizat, L. and Bach, F. A Note on Lazy Training in Supervised Differentiable Programming. *arXiv e-prints*, art. arXiv:1812.07956, Dec 2018.

Cho, Y. and Saul, L. K. Kernel methods for deep learning. In *Proceedings of the 22Nd International Conference on Neural Information Processing Systems*, NIPS’09, pp. 342–350, USA, 2009. Curran Associates Inc. ISBN 978-1-61567-911-9. URL http://dl.acm.org/citation.cfm?id=2984093.2984132.

Choromanska, A., Henaff, M., Mathieu, M., Ben Arous, G., and LeCun, Y. The Loss Surfaces of Multilayer Networks. *arXiv e-prints*, art. arXiv:1412.0233, Nov 2014.

Daniely, A., Frostig, R., and Singer, Y. Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity. *ArXiv e-prints*, February 2016.

Draxler, F., Veschgini, K., Salmhofer, M., and Hamprecht, F. A. Essentially No Barriers in Neural Network Energy Landscape. *arXiv e-prints*, art. arXiv:1803.00885, March 2018.

Ethan, D. and Gur-Ari, G. Feynman diagrams for large width networks. 2019.

Frye, C. and Efthimiou, C. J. Spherical Harmonics in p Dimensions. *ArXiv e-prints*, May 2012.

Gardner, E. and Derrida, B. Optimal storage properties of neural network models. *Journal of Physics A Mathematical General*, 21:271–284, January 1988. doi: 10.1088/0305-4470/21/1/031.

Jacot, A., Gabriel, F., and Hongler, C. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *ArXiv e-prints*, June 2018.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

Lee, J., Sohl-dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1EA-M-0Z.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Sohl-Dickstein, J., and Pennington, J. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. *arXiv e-prints*, art. arXiv:1902.06720, Feb 2019.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences, 115*(33):E7665–E7671, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1806579115. URL https://www.pnas.org/content/115/33/E7665.

Neyshabur, B., Tomioka, R., and Srebro, N. In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning. *arXiv e-prints*, art. arXiv:1412.6614, December 2014.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. Towards Understanding the Role of Over-Parametrization in Generalization of Neural Networks. *arXiv e-prints*, art. arXiv:1805.12076, May 2018.
Opper, M. and Vivarelli, F. General bounds on bayes errors for regression with gaussian processes. In Kearns, M. J., Solla, S. A., and Cohn, D. A. (eds.), Advances in Neural Information Processing Systems 11, pp. 302–308. MIT Press, 1999. URL http://papers.nips.cc/paper/1622-general-bounds-on-bayes-errors-for-regression-with-gaussian-processes.pdf.

Rasmussen, C. E. and Williams, C. K. I. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press, 2005. ISBN 026218253X.

Schulman, L. Techniques and applications of path integration. 1996. URL https://books.google.co.il/books?id=Cuc9AQAAIAAJ.

Seeger, M. Pac-bayesian generalisation error bounds for gaussian process classification. J. Mach. Learn. Res., 3:233–269, March 2003. ISSN 1532-4435. doi: 10.1162/153244303765208386. URL https://doi.org/10.1162/153244303765208386.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. arXiv e-prints, art. arXiv:1312.6229, December 2013.

Sollich, P. Gaussian Process Regression with Mismatched Models. arXiv e-prints, art. cond-mat/0106475, Jun 2001.

Sollich, P. and Williams, C. Using the equivalent kernel to understand gaussian process regression. In Saul, L. K., Weiss, Y., and Bottou, L. (eds.), Advances in Neural Information Processing Systems 17, pp. 1313–1320. MIT Press, 2005. URL http://papers.nips.cc/paper/2676-using-the-equivalent-kernel-to-understand-gaussian-process-regression.pdf.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS’14, pp. 3320–3328, Cambridge, MA, USA, 2014. MIT Press. URL http://dl.acm.org/citation.cfm?id=2969033.2969197.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. arXiv e-prints, art. arXiv:1611.03530, November 2016.
1 Re-stating the Notation

For completeness, here we re-state the notations used in the main-text.

\(\Omega\) - Space pace of inputs.
\(x, x', x^*\) - Inputs (in \(\Omega\)).
\(\mu_x\) - Measure on \(\Omega\).
\(K(x, x')\) - Kernel function (covariance) of a Gaussian process. Assumed to be symmetric and positive-semi-definite.
\(\phi_i(x)\) - \(i\)’th eigenfunction of \(K(x, x')\). By the spectral theorem, the set \(\{\phi_i\}_{i=1}^{\infty}\) can be assumed to be orthonormal:

\[
\int_{x \in \Omega} d\mu_x \phi_i(x) \phi_j(x) = \delta_{ij}
\]

\(\lambda_i\) - \(i\)’th eigenvalue of \(K(x, x')\).

\[
\int_{x' \in \Omega} d\mu_{x'} K(x, x') \phi_i(x') = \lambda_i \phi_i(x)
\]

\(\| \cdot \|_{\mathcal{H}_K}\) - RKHS norm. If \(f(x) = \sum_i f_i \phi_i(x)\) then \(\|f\|_{\mathcal{H}_K} = \sum_i \frac{\lambda_i}{\lambda_i(x)}\) (where \(\phi_i\) is an orthonormal set). Note that this norm is independent of \(\mu_x\).

\(g(x)\) - The target function.
\(\sigma^2\) - Noise variance. The noise is assumed to be Gaussian.

\(N\) - Number of inputs in the data-set.
\(D_N\) - Data-set of size \(N\), \(D_N = \{x_1, ..., x_N\}\).
\(f^{*}_{D_N, \sigma^2}(x)\) - The prediction function.

2 Explicit path integral computations

Here we wish to prove Eq. 3 of the main text using an explicit computation of the path integral.

Denoting \(\int d\mu_x d\mu_{x'} f(x) K^{-1}(x, x') f(x')\) as \(\|f\|_{K}^2\) and noting that \(\|f\|_{K}^2 = \sum_i \frac{\lambda_i}{\lambda_i(x)}\):

\[
\int Df \cdot f(x) \cdot f(y) \cdot \exp \left( -\frac{1}{2} \|f\|_{K}^2 \right) = \frac{\int \prod_i df_i \cdot \sum_i f_i \phi_i(x) \cdot \int \prod_i df_i \cdot \sum_i f_i \phi_i(y) \cdot \exp \left( -\frac{1}{2} \sum_i \frac{\lambda_i}{\lambda_i(x)} \right)}{\int \prod_i df_i \cdot \exp \left( -\frac{1}{2} \sum_i \frac{\lambda_i}{\lambda_i(x)} \right)} = \int \prod_i df_i \cdot \exp \left( -\frac{1}{2} \sum_i \frac{\lambda_i}{\lambda_i(x)} \right) = \int \prod_i df_i \cdot \exp \left( -\frac{1}{2} \sum_i \frac{\lambda_i}{\lambda_i(x)} \right)
\]

\[
= \sum_i \int df_i \cdot \exp \left( -\frac{\lambda_i}{2\lambda_i(x)} \right) \phi_i(x) \phi_i(y) + \sum_{i \neq j} \int df_i \cdot |f| \cdot \exp \left( -\frac{\lambda_i}{2\lambda_i(x)} \right) \int df_j \cdot |f| \cdot \exp \left( -\frac{\lambda_j}{2\lambda_j(x)} \right) = \int \prod_i df_i \cdot \exp \left( -\frac{\lambda_i}{2\lambda_i(x)} \right)
\]

\[
= \sum_i \lambda_i \phi_i(x) \phi_i(y) = K(x, y)
\]

3 Relaxing the uniformity assumption

As discussed in the Numerical Verification section in the main-text in Fig. 1, we plot the comparising between an NNGP kernel sampled on random data and on properly augmented CIFAR10 data.
4 Gaussian Process Prediction as a Field Theory

Let us assume a Gaussian process (GP) with mean 0 and co-variance function $K(x,x')$. For a data-set $D_N$ of size $N$ and noisy targets $\{g(x_i)\}_{i=1}^N$, it is known that the posterior mean obtained by Bayesian inference is

$$f_{D_N,\sigma^2} = \arg\min_f \left[ \frac{1}{2} \|f\|^2_{H_K} + \sum_{x_i \in D_N} \frac{(f(x_i) - g(x_i))^2}{2\sigma^2} \right]$$

For a data-set $D_N$ of size $N$ and noisy targets $\{g(x_i)\}_{i=1}^N$, we present the GP canonical partition function:

$$Z_{D_N,\sigma^2} [\alpha(x)] \overset{def}{=} \int Df \exp \left( -\frac{1}{2} \|f\|^2_{H_K} + \int \alpha(x) f(x) dx - \sum_{x_i \in D_N} \frac{(f(x_i) - g(x_i))^2}{2\sigma^2} \right)$$

Where the $Df$ notation stands for path integral. Notice that the functional derivative of $\log (Z_{D_N,\sigma^2} [\alpha(x)])$ w.r.t $\alpha(x)$ at $\alpha(x) = 0$ yields:

$$\frac{\partial}{\partial \alpha(x)} \bigg|_{\alpha(x)=0} \log (Z_{D_N,\sigma^2} [\alpha(x)]) = \frac{1}{Z_{D_N,\sigma^2} [\alpha(x) = 0]} \cdot \frac{\partial}{\partial \alpha(x)} \bigg|_{\alpha(x)=0} (Z_{D_N,\sigma^2} [\alpha(x)]) =$$

$$= \int Df \cdot f(x^*) \exp \left( -\frac{1}{2} \|f\|^2_{H_K} - \sum_{x_i \in D_N} \frac{(f(x_i) - g(x_i))^2}{2\sigma^2} \right) = \arg\min_{f_{x^*}} \left[ \frac{1}{2} \|f\|^2_{H_K} + \sum_{i=1}^N \frac{(f(x_i) - g(x_i))^2}{2\sigma^2} \right]$$

where the last equality is due to the fact that for Gaussian distributions, the expected value coincides with the most probable value. Therefore, the exact bayesian inference mean:

$$f_{D_N,\sigma^2} (x^*) = \frac{\partial}{\partial \alpha(x)} \bigg|_{\alpha(x)=0} \log (Z_{D_N,\sigma^2} [\alpha(x)])$$
4.1 Canonical Ensemble Formalism

for evaluating the quality of a certain GP, we’re interested in the average prediction for all the data-sets of size \( N \), meaning:

\[
f_{C}^{N,\sigma^2}(x^*) \overset{def}{=} \langle f_{D_N,\sigma^2}(x^*) \rangle_{D_N \sim \mu^\eta_N} = \int d\mu_{x_1} \int d\mu_{x_2} \cdots \int d\mu_{x_N} f_{D_N=(x_1,\ldots,x_N),\sigma^2}(x^*). 
\]

Using the replica trick we obtain:

\[
f_{C}^{N,\sigma^2}(x^*) = \lim_{M \to 0} \frac{\partial}{\partial \alpha(x^*)} \left|_{\alpha(x^*)=0} \frac{\langle Z_{D_N,\sigma^2}^M[\alpha(x)] \rangle_{D_N \sim \mu^\eta_N} - 1}{M} \right.
\]

Now, let us calculate \( \langle Z_{D_N,\sigma^2}^M[\alpha(x)] \rangle_{D_N \sim \mu^\eta_N} \) for an integer \( M \):

\[
Z_{D_N,\sigma^2}^M[\alpha(x)] = \int \cdots \int \prod_{j=1}^M Df_j \exp \left( -\frac{1}{2} \sum_{j=1}^M \|f_j\|_{K,H}^2 + \sum_{j=1}^M \int \alpha(x) f_j(x) dx - \sum_{j=1}^M \sum_{x_i \in D_N} (f_j(x_i) - g(x_i))^2 \right)
\]

\[
Z_{N,M,\sigma^2} = \langle Z_{D_N,\sigma^2}^M[\alpha(x)] \rangle_{D_N \sim \mu^\eta_N} = \int d\mu_{x_1} \cdots \int d\mu_{x_N} Z_{D_N,\sigma^2}^M[\alpha(x)] = \int \cdots \int \prod_{j=1}^M Df_j \exp \left( -\frac{1}{2} \sum_{j=1}^M \|f_j\|_{K,H}^2 + \sum_{j=1}^M \int \alpha(x) f_j(x) dx \right) \langle \exp \left( -\sum_{j=1}^M (f_j(x) - g(x))^2 \right) \rangle_{x \sim \mu^\eta_x}^N
\]

so

\[
f_{C}^{N,\sigma^2}(x^*) = \lim_{M \to 0} \frac{\partial}{\partial \alpha(x^*)} \left|_{\alpha(x^*)=0} \frac{\langle Z_{D_N,\sigma^2}^M[\alpha(x)] \rangle_{D_N \sim \mu^\eta_N}}{M} \right.
\]

4.2 Grand Canonical Ensemble Formalism

We now wish to allow fluctuations in the value of \( N \), meaning averaging over \( f_{C}^{N,\sigma^2}(x^*) \) for different values of \( N \). The motivation is to simplify the calculations, while averaging around a relatively confined region of \( N \)’s. Let us average the canonical prediction while weighting \( N \) according to Poisson distribution with expected value \( \eta \):

\[
f_{\eta,\sigma^2}^{GC}(x^*) \overset{def}{=} \sum_{N=0}^\infty \frac{e^{-\eta N}}{N!} f_{C}^{N,\sigma^2}(x^*),
\]

and defining:

\[
Z_{\eta,M,\sigma^2}[\alpha(x)] = \sum_{N=0}^\infty \frac{e^{-\eta N}}{N!} \langle Z_{D_N,\sigma^2}^M[\alpha(x)] \rangle_{D_N \sim \mu^\eta_N}
\]
we get:

\[ f_{\eta,\sigma^2}^G(x^*) = \left. \frac{\partial}{\partial \alpha(x^*)} \right|_{\alpha(x)=0} \lim_{M \to 0} \frac{Z_{\eta,M,\sigma^2}[\alpha(x)]}{M} \]

That is, the functional derivative w.r.t. \( \alpha(x) \) at \( \alpha(x) = 0 \) yields the average prediction, averaged over different data-set sizes (the canonical averaging) and different data-sets for each size (the grand canonical averaging).

For a given \( \eta \), the standard deviation of \( N \) is \( \eta \), so the relative error is \( \frac{1}{\sqrt{\eta}} \), decreases with \( \eta \).

Substituting \( \langle Z_{D_N,s^2}[\alpha(x)] \rangle_{D_N \sim \mu_x^N} \) in the expression for \( Z_{\eta,M,\sigma^2}[\alpha(x)] \) we obtain:

\[
\langle Z_{D_N,s^2}[\alpha(x)] \rangle_{D_N \sim \mu_x^N} = \sum_{N=0}^{\infty} \frac{e^{-\eta N}}{N!} \int \ldots \int_{M}^M Df_{j} \exp \left( -\frac{1}{2} \sum_{j=1}^{M} \|f_j\|_{H_{f_j}}^2 + \sum_{j=1}^{M} \int \alpha(x) f_j(x) \, dx \right)
\]

\[
\exp \left( -\sum_{j=1}^{M} \left( f_j(x) - g(x) \right)^2 \frac{1}{2\sigma^2} \right) \right)_{x \sim \mu_x} ^N = \exp \left( -\frac{1}{2} \sum_{j=1}^{M} \|f_j\|_{H_{f_j}}^2 + \sum_{j=1}^{M} \int \alpha(x) f_j(x) \, dx \right)
\]

\[
\sum_{N=0}^{\infty} \frac{e^{-\eta N}}{N!} \langle \exp \left( -\sum_{j=1}^{M} \left( f_j(x) - g(x) \right)^2 \frac{1}{2\sigma^2} \right) \rangle_{x \sim \mu_x} ^N = e^{-\eta} \int \ldots \int_{M}^M Df_{1} \ldots Df_{M}
\]

\[
\exp \left( -\frac{1}{2} \sum_{j=1}^{M} \|f_j\|_{H_{f_j}}^2 + \sum_{j=1}^{M} \int \alpha(x) f_j(x) \, dx + \eta \exp \left( -\frac{1}{2} \sum_{j=1}^{M} \left( f_j(x) - g(x) \right)^2 \frac{1}{2\sigma^2} \right) \right)_{x \sim \mu_x}
\]

### 4.3 Deriving the Equivalence Kernel using the Grand Canonical Formalism

We wish to get rid of the exponent inside the exponent. Expending it using (first order) Taylor series:
\[ Z_{\eta,M,\sigma^2} [\alpha(x)] = \]
\[ = e^{-\eta} \int \prod_{j=1}^{M} Df_j \]
\[ \exp \left( -\frac{1}{2} \sum_{j=1}^{M} \| f_j \|_{\mathcal{H}_K}^2 + \sum_{j=1}^{M} \int \alpha(x) f_j(x) \, dx + \eta \left( \exp \left( -\frac{1}{2} \sum_{j=1}^{M} \frac{(f_j(x) - g(x))^2}{2\sigma^2} \right) \right)_{x \sim \mu_x} \right) \approx \]
\[ = e^{-\eta} \int \prod_{j=1}^{M} Df_j \]
\[ \exp \left( -\frac{1}{2} \sum_{j=1}^{M} \| f_j \|_{\mathcal{H}_K}^2 + \sum_{j=1}^{M} \int \alpha(x) f_j(x) \, dx + \eta \left( 1 - \sum_{j=1}^{M} \frac{(f_j(x) - g(x))^2}{2\sigma^2} \right)_{x \sim \mu_x} \right) = \]
\[ = \left[ \int Df \exp \left( -\frac{1}{2} \| f \|_{\mathcal{H}_K}^2 + \int \alpha(x) f(x) \, dx - \eta \left( \frac{(f(x) - g(x))^2}{2\sigma^2} \right)_{x \sim \mu_x} \right) \right]^M \]
\[ \overset{\text{def}}{=} (Z_{\eta,\sigma^2} [\alpha(x)])^M \]
\[ Z_{\eta,\sigma^2} [\alpha(x)] = \int Df \exp \left( -\frac{1}{2} \| f \|_{\mathcal{H}_K}^2 + \int \alpha(x) f(x) \, dx - \frac{\eta}{2\sigma^2} \int d\mu_x (f(x) - g(x))^2 \right) \]

and without any additional approximation:

\[ f_{\eta,\sigma^2}^{GC} (x^*) = \frac{\partial}{\partial \alpha(x^*)} \left. \lim_{M \to 0} \frac{Z_{\eta,M,\sigma^2} [\alpha(x)] - 1}{M} \right|_{\alpha(x)=0} \approx \frac{\partial}{\partial \alpha(x^*)} \left. \lim_{M \to 0} \frac{(Z_{\eta,\sigma^2} [\alpha(x)])^M - 1}{M} \right|_{\alpha(x)=0} = \]
\[ = \frac{\partial}{\partial \alpha(x^*)} \left. \log (Z_{\eta,\sigma^2} [\alpha(x)]) \right|_{\alpha(x)=0} = \frac{1}{Z_{\eta,\sigma^2} [\alpha(x) = 0]} \cdot \frac{\partial}{\partial \alpha(x^*)} (Z_{\eta,\sigma^2} [\alpha(x)]) = \]
\[ = \frac{1}{Z_{\eta,\sigma^2} [\alpha(x) = 0]} \int Df \cdot f(x^*) \]
\[ \exp \left( -\frac{1}{2} \| f \|_{\mathcal{H}_K}^2 + \int \alpha(x) f(x) \, dx - \frac{\eta}{2\sigma^2} \int d\mu_x (f(x) - g(x))^2 \right) \bigg|_{\alpha(x)=0} = \]
\[ = \int Df \cdot f(x^*) \cdot \exp \left( -\frac{1}{2} \| f \|_{\mathcal{H}_K}^2 - \frac{\eta}{2\sigma^2} \int d\mu_x (f(x) - g(x))^2 \right) = \]
\[ = \int Df \exp \left( -\frac{1}{2} \| f \|_{\mathcal{H}_K}^2 - \frac{\eta}{2\sigma^2} \int d\mu_x (f(x) - g(x))^2 \right) = \]
\[ = \arg \min_{f_{x^*}} \left[ \frac{1}{2} \| f \|_{\mathcal{H}_K}^2 + \frac{\eta}{2\sigma^2} \int d\mu_x (f(x) - g(x))^2 \right] \overset{\text{def}}{=} f_{\eta,\sigma^2}^{EK} (x^*) \]

and that is exactly the result for the equivalence kernel, where \( \eta \) is the data-set size (we regarded it as the mean of the data-set size).

Let us derive it explicitly. For \( f(x) = \sum_i f_i(x) \) and \( g(x) = \sum_i g_i(x) \):
\[ f_{\eta, \sigma^2}^{EK}(x^*) = \int Df \cdot f(x^*) \exp \left( -\frac{1}{2} \|f\|_{H_K}^2 - \frac{\eta}{2\sigma^2} \int d\mu_x (f(x) - g(x))^2 \right) = \int Df \exp \left( -\frac{1}{2} \|f\|_{H_K}^2 - \frac{\eta}{2\sigma^2} \int d\mu_x (f(x) - g(x))^2 \right) = \int \Pi_i df_i \cdot \sum_i f_i \phi_i (x^*) \cdot \exp \left( -\frac{1}{2} \sum_i \left( \frac{f_i^2}{\lambda_i} + \frac{\eta}{2\sigma^2} (f_i - g_i)^2 \right) \right) = \sum_i \phi_i (x^*) \cdot \frac{df_i \cdot f_i \cdot \exp \left( -\frac{f_i^2}{2\lambda_i} - \frac{\eta}{2\sigma^2} (f_i - g_i)^2 \right)}{\int df_i \exp \left( -\frac{f_i^2}{2\lambda_i} - \frac{\eta}{2\sigma^2} (f_i - g_i)^2 \right)} = \sum_i \frac{\lambda_i g_i \phi_i (x^*)}{\lambda_i + \frac{\eta}{\sigma^2}} \cdot \int df_i \exp \left( -\frac{f_i^2}{2\lambda_i} - \frac{\eta}{2\sigma^2} (f_i - g_i)^2 \right) \phi_i (x) \phi_j (y) \] 

\[ \{ f(x) \cdot f(y) \}_0 = \frac{\int Df \cdot f(x) \cdot f(y) \exp \left( -\frac{1}{2} \|f\|_{H_K}^2 - \frac{\eta}{2\sigma^2} \int d\mu_x (f(x) - g(x))^2 \right)}{\int Df \exp \left( -\frac{1}{2} \|f\|_{H_K}^2 - \frac{\eta}{2\sigma^2} \int d\mu_x (f(x) - g(x))^2 \right)} = \int \Pi_i df_i \cdot \sum_{i,j} f_i f_j \phi_i (x) \phi_j (y) \cdot \exp \left( -\frac{1}{2} \sum_i \left( \frac{f_i^2}{\lambda_i} + \frac{\eta}{2\sigma^2} (f_i - g_i)^2 \right) \right) = \sum_i \frac{\lambda_i g_i \phi_i (x)}{\lambda_i + \frac{\eta}{\sigma^2}} \cdot \left( \frac{\lambda_j g_j}{\lambda_j + \frac{\eta}{\sigma^2}} \right) \int df_i \exp \left( -\frac{f_i^2}{2\lambda_i} - \frac{\eta}{2\sigma^2} (f_i - g_i)^2 \right) \phi_i (x) \phi_j (y) \] 

\[ \langle f(x) \rangle_0 \langle f(y) \rangle_0 = \sum_i \frac{\lambda_i g_i \phi_i (x)}{\lambda_i + \frac{\eta}{\sigma^2}} \cdot \left( \frac{\lambda_j g_j}{\lambda_j + \frac{\eta}{\sigma^2}} \right) \int df_i \exp \left( -\frac{f_i^2}{2\lambda_i} - \frac{\eta}{2\sigma^2} (f_i - g_i)^2 \right) \phi_i (x) \phi_j (y) \] 

Therefore:

\[ \text{Cov} \left[ f(x), f(y) \right] = \sum_i \left( \frac{1}{\lambda_i} + \frac{\eta}{\sigma^2} \right)^{-1} \phi_i (x) \phi_i (y) \]

and we see that the correlations are \( O \left( \frac{1}{\eta} \right) \).

For rotationally invariant kernel, we get that

\[ \text{Var} \left[ f(x) \right] = \sum_{l=0}^{\infty} \sum_{m=0}^{\deg(l)} \left( \frac{1}{\lambda_l} + \frac{\eta}{\sigma^2} \right)^{-1} Y_{l,m}^2 (x) \overset{df}{=} Y_{l,m}^2 (x) \overset{\text{def}}{=} C_{K, \eta, \sigma^2} \]

is independent of \( x \) since

\[ \sum_{m=0}^{\deg(l)} Y_{l,m}^2 (x) = \deg (l) \]
\[ C_{K, \eta, \sigma^2} = \sum_{l=0}^{\infty} \sum_{m=0}^{\text{deg}(l)} \lambda_l \]

### 4.5 Pertubative Correction for the Equivalence Kernel

#### 4.5.1 Averaging \( f \)

Going to the next order in the expansion:

\[
Z_{\eta, M, \sigma^2} [\alpha (x)] =
\]

\[
e^{-\eta} \int_{\mathcal{H}_K} \prod_{j=1}^{M} Df_j
\]

\[
\exp \left( -\frac{1}{2} \sum_{j=1}^{M} \| f_j \|^2_{\mathcal{H}_K} + \sum_{j=1}^{M} \int \alpha (x) f_j (x) \, dx \right)
\]

\[
\approx e^{-\eta} \int_{\mathcal{H}_K} \prod_{j=1}^{M} Df_j
\]

\[
\exp \left( -\frac{1}{2} \sum_{j=1}^{M} \| f_j \|^2_{\mathcal{H}_K} + \sum_{j=1}^{M} \int \alpha (x) f_j (x) \, dx \right)
\]

\[
\exp \left( \eta \left( 1 - \sum_{j=1}^{M} \frac{(f_j (x) - g(x))^2}{2\sigma^2} + \frac{1}{2} \left( \sum_{j=1}^{M} \frac{(f_j (x) - g(x))^2}{2\sigma^2} \right)^2 \right)_{x \sim \mu_x} \right)
\]

\[
= \int_{\mathcal{H}_K} \prod_{j=1}^{M} Df_j
\]

\[
\exp \left( \sum_{j=1}^{M} \left( -\frac{1}{2} \| f_j \|^2_{\mathcal{H}_K} + \int \alpha (x) f_j (x) \, dx - \frac{\eta}{2\sigma^2} \int d\mu_x (f_j (x) - g(x))^2 \right) \right)
\]

\[
\exp \left( \frac{\eta}{8\sigma^4} \sum_{j=1}^{M} \sum_{l=1}^{M} \int d\mu_x (f_j (x) - g(x))^2 \cdot (f_l (x) - g(x))^2 \right)
\]

Note that:

\[
f_{\eta, \sigma^2}^{GC} (x^*) = \frac{\partial}{\partial \alpha (x^*)} \left|_{\alpha(x)=0} \right. \lim_{M \to 0} \frac{Z_{\eta, M, \sigma^2} [\alpha (x)] - 1}{M} = \]

\[
= \lim_{M \to 0} \left[ (Z_{\eta, \sigma^2}^{\mathcal{E}K} [\alpha (x) = 0])^M \cdot \lim_{M \to 0} \left[ \frac{\partial}{\partial \alpha (x^*)} \left|_{\alpha(x)=0} \right. \frac{Z_{\eta, M, \sigma^2} [\alpha (x)]}{M} \right] \right] = \]

\[
= \lim_{M \to 0} \left[ \left( Z_{\eta, \sigma^2}^{\mathcal{E}K} [\alpha (x) = 0] \right)^M \cdot \lim_{M \to 0} \left[ \frac{\partial}{\partial \alpha (x^*)} \left|_{\alpha(x)=0} \frac{Z_{\eta, M, \sigma^2} [\alpha (x)]}{M} \right] \right] \right] = \]
Calculating the first order perturbative corrections:

\[ f_{\eta,\sigma^2}^{GC}(x^*) = \lim_{M \to 0} \frac{\partial}{\partial \alpha(x)} \bigg|_{\alpha(x) = 0} \frac{Z_{\eta,M,\sigma^2} \ [\alpha(x)]}{M \cdot (Z_{\eta,\sigma^2}^{E\!K} \ [\alpha(x) = 0])^M} = \]

\[ = \lim_{M \to 0} \frac{1}{M \cdot Z_{\eta,\sigma^2}^{E\!K}[0]^M} \int \cdots \int \prod_{j=1}^{M} Df_j \exp \left( -\frac{\eta}{2\sigma^2} \sum_{j=1}^{M} \int d\mu_x (f_j(x) - g(x))^2 - \frac{\eta}{2\sigma^2} \sum_{j=1}^{M} \int d\mu_x (f_j(x) - g(x))^2 \right) \]

\[ = \lim_{M \to 0} \frac{1}{M \cdot Z_{\eta,\sigma^2}^{E\!K}(0)^M} \int \cdots \int \prod_{j=1}^{M} Df_j \exp \left( -\frac{\eta}{2\sigma^2} \sum_{j=1}^{M} \int d\mu_x (f_j(x) - g(x))^2 - \frac{\eta}{2\sigma^2} \sum_{j=1}^{M} \int d\mu_x (f_j(x) - g(x))^2 \right) \]

\[ = \lim_{M \to 0} \frac{1}{M} \left\langle \left( 1 + \frac{\eta}{8\sigma^2} \sum_{j=1}^{M} \sum_{i=1}^{M} \int d\mu_x (f_j(x) - g(x))^2 \cdot (f_i(x) - g(x))^2 \right) \cdot \sum_{i=1}^{M} f_i(x^*) \right\rangle_{f_1,\ldots,f_M \sim E\!K} + O \left( \frac{1}{\eta^3} \right) \]

\[ = \langle f(x^*) \rangle_0 + \lim_{M \to 0} \frac{1}{M} \left\langle \left( \frac{\eta}{8\sigma^2} \sum_{j=1}^{M} \sum_{i=1}^{M} \int d\mu_x (f_j(x) - g(x))^2 \cdot (f_i(x) - g(x))^2 \right) \cdot \sum_{i=1}^{M} f_i(x^*) \right\rangle_{f_1,\ldots,f_M \sim E\!K} + O \left( \frac{1}{\eta^4} \right) \]

\[ + \lim_{M \to 0} \frac{1}{M} \frac{\eta}{8\sigma^2} \int d\mu_x \left\langle \sum_{j=1}^{M} \sum_{l=1}^{M} \sum_{i=1}^{M} (f_j(x) - g(x))^2 \cdot (f_l(x) - g(x))^2 f_i(x^*) \right\rangle_{f_1,\ldots,f_M \sim E\!K} + O \left( \frac{1}{\eta^3} \right). \]

Calculating the correction:

\[ \left\langle \sum_{j=1}^{M} \sum_{l=1}^{M} \sum_{i=1}^{M} (f_j(x) - g(x))^2 \cdot (f_l(x) - g(x))^2 f_i(x^*) \right\rangle_{f_1,\ldots,f_M \sim E\!K} = \]

\[ = M \left\langle (f(x) - g(x))^4 f(x^*) \right\rangle_0 + M (M - 1) \left[ 2 \left\langle (f(x) - g(x))^2 \right\rangle_0 \left\langle (f(x) - g(x))^2 f(x^*) \right\rangle_0 + \left\langle (f(x) - g(x))^4 \right\rangle_0 (f(x^*))_0 \right] \]

\[ + M (M - 1) (M - 2) \left\langle (f(x) - g(x))^2 \right\rangle_0^2 \left\langle f(x^*) \right\rangle_0 \]

Note that we eventually divide by \( M \) and take the limit \( M \to 0 \), so we only care about \( O(M) \) terms.
\[ f_{\eta,\sigma^2}^{GC}(x^*) = f_{\eta,\sigma^2}^{EK}(x^*) \]

\[ + \frac{\eta}{8\sigma^4} \int d\mu_x \left[ \left\langle (f(x) - g(x))^4 f(x^*) \right\rangle_0 - 2 \left\langle (f(x) - g(x))^2 f(x^*) \right\rangle_0 \left\langle (f(x) - g(x))^2 \right\rangle_0 \right. \]

\[ \left. - \left\langle (f(x) - g(x))^4 \right\rangle_0 (f(x^*))_0 + 2 \left\langle (f(x) - g(x))^2 \right\rangle_0 (f(x^*))_0 \right] + O\left( \frac{1}{\eta^2} \right) \]

These correlations can be calculated using Feynman diagrams, since the free theory (EK) is quadratic (Gaussian):

\[ \left\langle (f(x) - g(x))^4 f(x^*) \right\rangle_0 = \]

\[ = 3 f_{\eta,\sigma^2}^{EK}(x^*) \text{Var}[f(x)]^2 + 6 f_{\eta,\sigma^2}^{EK}(x^*) \left( f_{\eta,\sigma^2}^{EK}(x) - g(x) \right)^2 \text{Var}[f(x)] \]

\[ + f_{\eta,\sigma^2}^{EK}(x^*) \left( f_{\eta,\sigma^2}^{EK}(x) - g(x) \right)^4 + 4 \left( f_{\eta,\sigma^2}^{EK}(x) - g(x) \right)^3 \text{Cov}[f(x), f(x^*)] \]

\[ + 12 \left( f_{\eta,\sigma^2}^{EK}(x) - g(x) \right) \text{Var}[f(x)] \text{Cov}[f(x), f(x^*)] \]

\[ \left\langle (f(x) - g(x))^4 \right\rangle_0 (f(x^*))_0 = \]

\[ = 3 f_{\eta,\sigma^2}^{EK}(x^*) \text{Var}[f(x)]^2 \]

\[ + 6 f_{\eta,\sigma^2}^{EK}(x^*) \left( f_{\eta,\sigma^2}^{EK}(x) - g(x) \right)^2 \text{Var}[f(x)] + f_{\eta,\sigma^2}^{EK}(x^*) \left( f_{\eta,\sigma^2}^{EK}(x) - g(x) \right)^4 \]

\[ \left\langle (f(x) - g(x))^2 \right\rangle_0 \left\langle (f(x) - g(x))^2 f(x^*) \right\rangle_0 = \]

\[ = 2 \text{Var}[f(x)] \text{Cov}[f(x), f(x^*)] \left( f_{\eta,\sigma^2}^{EK}(x) - g(x) \right) + f_{\eta,\sigma^2}^{EK}(x^*) \text{Var}[f(x)]^2 \]

\[ + 2 \text{Cov}[f(x), f(x^*)] \left( f_{\eta,\sigma^2}^{EK}(x) - g(x) \right)^3 \]

\[ + 2 f_{\eta,\sigma^2}^{EK}(x^*) \text{Var}[f(x)] \left( f_{\eta,\sigma^2}^{EK}(x) - g(x) \right)^2 + f_{\eta,\sigma^2}^{EK}(x^*) \left( f_{\eta,\sigma^2}^{EK}(x) - g(x) \right)^4 \]

\[ \left\langle (f(x) - g(x))^2 \right\rangle_0 (f(x^*))_0 = \]

\[ = f_{\eta,\sigma^2}^{EK}(x^*) \text{Var}[f(x)]^2 \]

\[ + 2 f_{\eta,\sigma^2}^{EK}(x^*) \text{Var}[f(x)] \left( f_{\eta,\sigma^2}^{EK}(x) - g(x) \right)^2 + f_{\eta,\sigma^2}^{EK}(x^*) \left( f_{\eta,\sigma^2}^{EK}(x) - g(x) \right)^4 \]

Summing everything up:

\[ \left\langle (f(x) - g(x))^4 f(x^*) \right\rangle_0 - \left\langle (f(x) - g(x))^4 \right\rangle_0 (f(x^*))_0 \]

\[ - 2 \left\langle (f(x) - g(x))^2 \right\rangle_0 \left\langle (f(x) - g(x))^2 f(x^*) \right\rangle_0 + 2 \left\langle (f(x) - g(x))^2 \right\rangle_0 \left( f(x^*) \right)_0 = \]

\[ = 8 \left( f_{\eta,\sigma^2}^{EK}(x) - g(x) \right) \text{Var}[f(x)] \text{Cov}[f(x), f(x^*)] \]

and all the bubble diagrams cancel as expected.

So we get:
\[ f_{\eta, \sigma^2}^{GC}(x^*) = f_{\eta, \sigma^2}^{EK}(x^*) + \frac{\eta}{\sigma^4} \int d\mu_x \left( f_{\eta, \sigma^2}^{EK}(x) - g(x) \right) \text{Var} \left( f(x) \right) \text{Cov} \left( f(x), f(x^*) \right) + O \left( \frac{1}{\eta^3} \right) \]

Substituting the expressions for the variance and covariance:

\[ f_{\eta, \sigma^2}^{GC}(x^*) = f_{\eta, \sigma^2}^{EK}(x^*) - \frac{\eta}{\sigma^4} \sum_{i,j,k} \frac{\sigma^2}{\eta} \left( \frac{1}{\lambda_i} + \frac{\eta}{\sigma^2} \right)^{-1} \left( \frac{1}{\lambda_j} + \frac{\eta}{\sigma^2} \right)^{-1} g_i \phi_j(x^*) \int d\mu_x \phi_i(x) \phi_j(x) \phi_k^2(x) + O \left( \frac{1}{\eta^3} \right) \]

4.5.2 Averaging \( f^2 \)

This time we must use two different replica indices:

\[
\langle [f_{D_N, \sigma^2}^{*}(x^*])^2 \rangle_{D_N \sim \mu_N^N} = \left( \left[ \frac{\partial \log \left( Z_{D_N, \sigma^2} \left[ \alpha(x) \right] \right)}{\partial \alpha(x^*)} \right]_{\alpha(x)=0} \right)^2_{D_N \sim \mu_N^N} = \left( \lim_{M \to 0} \frac{1}{M} \cdot \frac{\partial Z_{D_N, \sigma^2} \left[ \alpha(x) \right]}{\partial \alpha(x^*)} \right) \bigg|_{\alpha(x)=0} \right)^2_{D_N \sim \mu_N^N} = \lim_{M \to 0} \lim_{M \to 0} \frac{1}{MM} \cdot \int Df_1 \cdots \int Df_M \int D\tilde{f}_1 \cdots \int D\tilde{f}_M \exp \left( -\frac{1}{2} \sum_{m=1}^{M} \| f_m \|_{H_K}^2 - \frac{1}{2} \sum_{m=1}^{M} \| \tilde{f}_m \|_{H_K}^2 \right) \sum_{m=1}^{M} f_m(x^*) \sum_{m=1}^{M} \tilde{f}_m(x^*) \left\langle \exp \left( -\sum_{m=1}^{M} \frac{(f_m(x) - g(x))^2}{2\sigma^2} - \sum_{m=1}^{M} \frac{(\tilde{f}_m(x) - g(x))^2}{2\sigma^2} \right) \right\rangle_{z \sim \mu_x}^N \]

Averaging w.r.t poisson distribution:
\[
\left\langle \left\langle [f_{D_N,\sigma^2}(x^*)]^2 \right\rangle \right\rangle_{D_N \sim \mu^N} = \sum_{N=0}^{\infty} \frac{e^{-\eta} \eta^N}{N!} \left\langle \left\langle [f_{D_N,\sigma^2}(x^*)]^2 \right\rangle \right\rangle_{D_N \sim \mu^N} = \\
= \lim_{M \to 0} \lim_{\lambda \to 0} \frac{1}{MM} \cdot \int Df_1 \cdots \int Df_M \int D\tilde{f}_1 \cdots \int D\tilde{f}_M \\
\exp \left( -\eta \cdot \frac{1}{2} \sum_{m=1}^{M} \|f_m\|_{\mathcal{H}_K}^2 \right) \sum_{m=1}^{M} \left\langle \left\langle \frac{\tilde{f}_m(x) - g(x)}{2\sigma^2} \right\rangle \right\rangle_{x \sim \mu^N} \\
\exp \left( \eta \left\langle \left\langle \frac{(f_m(x) - g(x))^2}{2\sigma^2} \right\rangle \right\rangle_{x \sim \mu^N} \right) \sum_{m=1}^{M} f_m(x) \sum_{m=1}^{M} \tilde{f}_m(x) \\
= \lim_{M \to 0} \lim_{\lambda \to 0} \frac{1}{MM} \cdot \int Df_1 \cdots \int Df_M \int D\tilde{f}_1 \cdots \int D\tilde{f}_M \\
\exp \left( -\frac{1}{2} \sum_{m=1}^{M} \|f_m\|_{\mathcal{H}_K}^2 \right) \sum_{m=1}^{M} \left\langle \left\langle \frac{\tilde{f}_m(x) - g(x)}{2\sigma^2} \right\rangle \right\rangle_{x \sim \mu^N} \\
\exp \left( \frac{\eta}{2} \left\langle \left\langle \frac{(f_m(x) - g(x))^2}{2\sigma^2} \right\rangle \right\rangle_{x \sim \mu^N} \right) \sum_{m=1}^{M} f_m(x) \sum_{m=1}^{M} \tilde{f}_m(x) \\
= \left( f_{E_{\eta,\sigma^2}}(x^*) \right)^2 \\
+ \lim_{M \to 0} \lim_{\lambda \to 0} \frac{1}{MM} \cdot \frac{\eta}{8\sigma^4} \int d\mu_x \left\langle \left\langle \left( \sum_{a=1}^{M} (f_a(x) - g(x))^2 \right) \left( \sum_{b=1}^{M} (\tilde{f}_b(x) - g(x))^2 \right) \right\rangle_0 \sum_{c=1}^{M} f_c(x) \sum_{d=1}^{M} \tilde{f}_d(x) \right\rangle_0 \\
= \left( f_{E_{\eta,\sigma^2}}(x^*) \right)^2 \\
+ \lim_{M \to 0} \lim_{\lambda \to 0} \frac{1}{MM} \cdot \frac{\eta}{4\sigma^4} \int d\mu_x \left\langle \left\langle \left( \sum_{a=1}^{M} (f_a(x) - g(x))^2 \right) \left( \sum_{b=1}^{M} (\tilde{f}_b(x) - g(x))^2 \right) \right\rangle_0 \sum_{c=1}^{M} f_c(x) \sum_{d=1}^{M} \tilde{f}_d(x) \right\rangle_0 \\
= \left( f_{E_{\eta,\sigma^2}}(x^*) \right)^2 + \frac{\eta}{4\sigma^4} \int d\mu_x \left( \lim_{M \to 0} \frac{1}{M} \left\langle \left\langle \left( \sum_{a=1}^{M} (f_a(x) - g(x))^2 \right) \right\rangle_0 \sum_{b=1}^{M} f_b(x) \sum_{c=1}^{M} f_c(x) \right\rangle_0 f_{E_{\eta,\sigma^2}}(x^*) \right) \\
+ \frac{1}{8\sigma^4} \left( \lim_{M \to 0} \frac{1}{M} \left( \sum_{a=1}^{M} (f_a(x) - g(x))^2 \right) \right) \text{as we saw in (f)}
and we’re left with:

\[
\lim_{M \to 0} \frac{1}{M} \left< \sum_{a=1}^{M} (f_a(x) - g(x))^2 \sum_{b=1}^{M} f_b(x^*) \right> = \\
\lim_{M \to 0} \frac{1}{M} \left[ M \left< (f(x) - g(x))^2 f(x^*) \right> + M(M-1) \left< (f(x) - g(x))^2 \right> \right] = \\
= \left< (f(x) - g(x))^2 f(x^*) \right> - \left< f(x^*) \right> \left< (f(x) - g(x))^2 \right> = 2 \left< f_{EK}^* - g(x) \right> \text{Cov} [f(x), f(x^*)]
\]

but this correction gives \( O \left( \frac{1}{\eta^3} \right) \) so:

\[
\left< \left[ f_{D, \sigma^2}^* \right]^2 \right>_1 = \left( \sum_{l,m} g_{l,m} \right) \text{Cov} [f(x), f(x^*)] + O \left( \frac{1}{\eta^3} \right)
\]

and notably

\[
\left< f^2 \right> = \left< f \right>^2 + O \left( \frac{1}{\eta^3} \right)
\]

### 4.6 Perubative Correction for Rotationally Invariant Kernel

We now wish to evaluate this expression for a rotationally invariant kernel and a uniform measure on the hypersphere. This simplifies the expression for \( f \) to:

\[
f_{GC}^* (x^*) = \\
f_{\eta, \sigma^2}^* (x^*) + \frac{\eta}{\sigma^2} \int \mu_x (f_{\eta, \sigma^2} (x) - g(x)) \text{Var} [f(x)] \text{Cov} [f(x), f(x^*)] + O \left( \frac{1}{\eta^3} \right) = \\
f_{\eta, \sigma^2}^* (x^*) + \frac{\eta}{\sigma^2} C_{\eta, \sigma^2} \int \mu_x (f_{\eta, \sigma^2}^* (x) - g(x)) \text{Cov} [f(x), f(x^*)] = \\
f_{\eta, \sigma^2}^* (x^*) = \frac{\eta}{\sigma^2} C_{\eta, \sigma^2} \sum_{i,j} \frac{\sigma^2}{\lambda_i + \frac{\sigma^2}{\eta}} \left( \frac{1}{\lambda_i} + \frac{\eta}{\sigma^2} \right)^{-1} g_i \phi_j (x^*) \int \mu_x \phi_i (x) \phi_j (x) + O \left( \frac{1}{\eta^3} \right) = \\
= f_{\eta, \sigma^2}^* (x^*) = C_{\eta, \sigma^2} \sum_{l,m} \frac{g_l m}{\sigma^2 \left( 2 + \frac{\lambda_l \eta}{\sigma^2} + \frac{\sigma^2}{\lambda_l \eta} \right)} Y_{l,m}^2 (x^*) + O \left( \frac{1}{\eta^3} \right).
\]

The expression for \( \left< f^2 \right> \) is:

\[
\left< \left[ f_{D, \sigma^2}^* \right]^2 \right>_1 = \left( f_{\eta, \sigma^2}^* (x^*) \right)^2 - 2 f_{\eta, \sigma^2}^* (x^*) C_{\eta, \sigma^2} \sum_{l,m} \frac{g_l m}{\sigma^2 \left( 2 + \frac{\lambda_l \eta}{\sigma^2} + \frac{\sigma^2}{\lambda_l \eta} \right)} Y_{l,m}^2 (x^*) + O \left( \frac{1}{\eta^3} \right)
\]
5 Various insights

5.1 Correction means worse generalization

The correction always means worse generalization than what the EK suggests. Indeed

\[ f_{\eta,\sigma^2}^{GC}(x^*) = f_{\eta,\sigma^2}^{EK}(x^*) - C_{K,\eta,\sigma^2} \sum_{l,m} \frac{g_{l,m}}{\sigma^2} \left( 2 + \frac{\lambda_l \eta}{\sigma^2} + \frac{\sigma^2}{\lambda_l \eta} \right) Y_{l,m}(x^*) + O \left( \frac{1}{\eta^3} \right) = \]

\[ = \sum_{l,m} \frac{\lambda_l}{\lambda_l + \frac{\sigma^2}{\eta}} g_{l,m} Y_{l,m}(x^*) - C_{K,\eta,\sigma^2} \sum_{l,m} \frac{g_{l,m}}{\sigma^2} \left( 2 + \frac{\lambda_l \eta}{\sigma^2} + \frac{\sigma^2}{\lambda_l \eta} \right) Y_{l,m}(x^*) + O \left( \frac{1}{\eta^3} \right) = \]

\[ = \sum_{l,m} \left( \frac{\lambda_l}{\lambda_l + \frac{\sigma^2}{\eta}} - \frac{C_{K,\eta,\sigma^2}}{\sigma^2} \left( 2 + \frac{\lambda_l \eta}{\sigma^2} + \frac{\sigma^2}{\lambda_l \eta} \right) \right) g_{l,m} Y_{l,m}(x^*) \]

\[ < \frac{\lambda_l}{\lambda_l + \frac{\sigma^2}{\eta}} < 1 \]

5.2 Range of validity of EK

We define the validity of EK by comparing the EK result with the correction we derived earlier. When the two are comparable taking only the EK result is unjustified.

\[ \frac{1}{1 + \frac{\sigma^2}{\lambda_l \eta}} - \frac{C_{K,\eta,\sigma^2}}{\sigma^2} \left( 2 + \frac{\lambda_l \eta}{\sigma^2} + \frac{\sigma^2}{\lambda_l \eta} \right) \approx 0 \]

denoting \( \frac{\sigma^2}{\lambda_l \eta} = x \) and \( \frac{C_{K,\eta,\sigma^2}}{\sigma^2} = a \):

\[ \frac{1}{1 + x} - \frac{ax}{(x + 1)^2} = 0 \rightarrow ax = x + 1 \rightarrow x = \frac{1}{a - 1} \]

meaning that

\[ \eta \approx \frac{C_{K,\eta,\sigma^2} - \sigma^2}{\lambda_l} \]

where \( l \) is the most meaningful angular momentum in \( g \),

\[ l = \max \arg \left\{ \sum_{m=0}^{N(d,l)} g_{l,m}^2 \right\} \]