Journal of Machine Learning Research 27 (2026) 1-62 Submitted 7/25; Revised 2/26; Published 4/26

Gradient Span Algorithms Make Predictable Progress in High Dimension

Felix Benning[*] FELIX.BENNING@GMAIL.COM
Leif Döring DOERING@UNI-MANNHEIM.DE
Department of Mathematics
University of Mannheim
68159 Mannheim, Germany

Editor: Bryon Aragam

Abstract

We prove that all 'gradient span algorithms' have asymptotically deterministic behavior on scaled Gaussian random functions as the dimension tends to infinity. This is a functional generalization of similar results for random quadratic functions and spin glasses. They explain the counterintuitive phenomenon that different training runs of many large machine learning models result in approximately equal cost curves despite random initialization on a complicated non-convex landscape. This 'predictable progress' phenomenon is exploited by the AutoML community: Since the optimization progress of a single run is already representative, multiple retries with the same hyperparameters are not necessary.

Keywords: optimization of random functions, Bayesian optimization, limit theorem, gradient span algorithm, AutoML

Contents

∗. Corresponding author, current affiliation: University of Luxembourg

1. Introduction

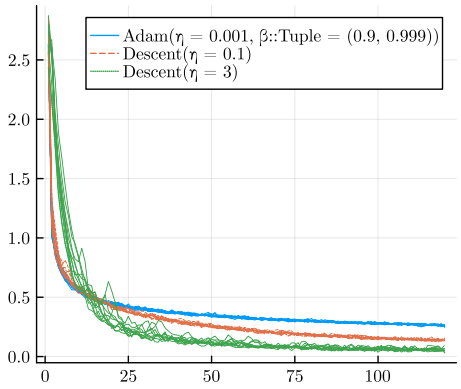
We present a theoretical explanation for a remarkable empirical phenomenon that occurs during the training of (large) machine learning models. 'Training' is the process of minimizing a cost (also known as error or risk) function $f\colon \mathbb{R}^N \to \mathbb{R}$ over parameter vectors $x \in \mathbb{R}^N$ by selecting successive parameter vectors $x_n \in \mathbb{R}^N$ based on noisy gradients at the previous parameter vectors $x_0, \dots x_{n-1}$. The parameters are typically the weights of a neural network and N is very large. The phenomenon is that over multiple initializations $x_0^{(1)}, x_0^{(2)}, \dots \in \mathbb{R}^N$ the progress a particular optimizer makes during the optimization is approximately equal over different optimization runs:

$$(f(x_0^{(i)}), f(x_1^{(i)}), \dots) \approx (f(x_0^{(j)}), f(x_1^{(j)}), \dots).$$

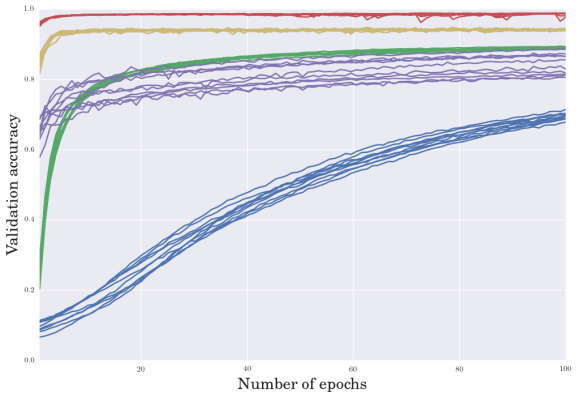
Figure 1 shows this '**predictable progress**' in practice. Using a relatively high batch size (1000) to approximate the underlying noise-less cost, we demonstrate predictable progress ourselves for the training on the MNIST dataset using a standard convolutional neural network (cf. Figure 1a). The key to predictable progress is high dimensionality of parameters (e.g. the neural network of Figure 1a has roughly $N = 2.3$ million parameters, which is still small in comparison to neural networks used for tasks beyond the "toy-problem" MNIST). Figure 1b (Klein et al., 2017) demonstrates that this phenomenon is well known and moreover relied upon for training heuristics. In Figure 2 we demonstrate this phenomenon on synthetic random functions.

The predictable progress phenomenon is certainly surprising since there is no reason to believe that function values along an optimization path should be approximately equal for different initializations. And, in view of handcrafted deterministic counterexamples, predictable progress is obviously not provable for arbitrary high-dimensional functions.

But this phenomenon clearly exists empirically, and recently predictable progress has been proven for **random quadratic functions** generated from the mean squared error applied to linear models (Paquette et al., 2022; Paquette and Trogdon, 2022; Deift and Trogdon, 2021; Pedregosa and Scieur, 2020; Scieur and Pedregosa, 2020). While quadratic functions are a simplified convex setting, these contributions offered a first explanation for the observed phenomenon of predictable behavior in high dimensions. We will extend these results to the much more general, non-convex setting of Gaussian random functions. This setting was already used in the machine learning literature for instance by Pascanu et al. (2014) and Dauphin et al. (2014) to explain why the overwhelming share of critical



(a) The plot shows an empirical approximation of the cost sequence resulting from the training of a standard convolutional neural network on the MNIST dataset (LeCun et al., 2010). We plot the values of $f(x_0^{(i)}), \ldots, f(x_{120}^{(i)})$ against the steps $0, \ldots, 120$ on the x-axis. The minimization is performed with three optimization algorithms: Adam (Kingma and Ba, 2015) (with learning rate η and momentum β) in blue and two version of gradient descent (learning rate $\eta = 0.1$ and $\eta = 3$) in red and green. Each optimizer was run 10 times from randomly selected initializations $x_0^{(i)}$ using the (random) default initialization procedure.



(b) In the plot, taken from Klein et al. (2017, Figure 3), optimization runs are grouped into 5 categories of hyperparameter settings represented by color. Each category contains 10 optimization runs using the same hyperparameter configuration. The predictable progress of the validation accuracy per configuration is used in Klein et al. (2017) to argue that it is sufficient to try a configuration once. The overall goal of Klein et al. (2017) is furthermore to fit a parametric models to the 'learning curves' in order to stop training early and switch to a different configuration if the progression does not seem promising.
The training heuristics that aspire to be 'AutoML' (i.e. automatically fit data without human intervention) are therefore built on this phenomenon. And the fact that these training heuristics are so successful demonstrates how ubiquitous predictable progress is in practice.

Figure 1: Predictable progress in machine learning practice

points are saddle points in high dimension, whereas the critical points of low dimensional GRFs are dominated by minima and maxima (Rasmussen and Williams, 2006). To lighten the technical difficulty that arises from this generalization, we do not consider stochastic gradients of sample losses. Instead we assume the algorithms operate with full gradients of the cost.

The specific modelling assumption of Pascanu et al. (2014) and Dauphin et al. (2014) were stationary isotropic GRFs, which we will extend to (non-stationary) isotropic GRFs to include the setting of random quadratic functions and more generally dot product kernels (i.e. spin glasses) which are both non-stationary (see Section A).

Similar to the convention of capital letters for random variables, we use bold font to denote random functions $\mathbf{f}$. Since our results are limit theorems in the dimension N, we mark this important dependency in the index. While the parameter sequences would also have to be indexed by the dimension N we omit this index for notational clarity. In simplified

terms we will prove the following theorem for predictable optimization in high dimension, precise theorems are given in Section 2:

> **If $\mathbf{f}_N$ is a (non-stationary) isotropic GRF on a high-dimensional domain $\mathbb{R}^N$, then the (random) sequence $\mathbf{f}_N(X_0), \mathbf{f}_N(X_1), ...$ along the (random) parameter sequence $X_0, X_1, ...$ selected by a standard first order optimization algorithm are close to a deterministic sequence $\mathfrak{f}_0, \mathfrak{f}_1, ...$ with high probability.**

To be a bit more precise, we will prove the following in Theorem 12: Suppose $X_0, X_1, ...$ is the (random) sequence of parameter points obtained by running an optimizer on the (random) function $\mathbf{f}_N \colon \mathbb{R}^N \to \mathbb{R}$ initialized at an independent, possibly random point X_0. For 'gradient span algorithms' $\mathfrak{G}$ (e.g. gradient descent) and a sequence of (non-stationary) isotropic GRFs $(\mathbf{f}_N)_{N\in\mathbb{N}}$ we construct a sequence of deterministic real numbers $\mathfrak{f}_0, \mathfrak{f}_1, ...$ such that, for $n \in \mathbb{N}$,

$$\lim_{N\to\infty} \mathbb{P}\Big(|\mathbf{f}_N(X_n) - \mathfrak{f}_n| > \epsilon\Big) = 0, \quad \forall \epsilon > 0.$$

The proof is completely constructive (Remark 15) and the limiting values $\mathfrak{f}_n$ may be computed with complexity $\mathcal{O}(n^6)$ given an algorithm $\mathfrak{G}$, the mean and covariance kernel of the random functions $\mathbf{f}_N$ and the length of the initialization vector X_0 (Remark 16). In Figure 2 we demonstrate convergence empirically by applying gradient descent to simulations of a GRF in various dimensions. In Corollary 13 we will show that an application of the stochastic triangle inequality yields approximately equal optimization progress given different initialization points (as can be seen in Figure 1 and 2).

Related work Beyond the work on random quadratic functions, our article is closely related in spirit to work from **statistical mechanics** where similar high dimensional limits are considered (referred to as the 'thermodynamic limit'). Specifically, isotropic random functions restricted to the sphere coincide with the "spherical **spin-glasses**" from statistical mechanics. This fact is not obvious, as p-spin glasses are defined explicitly as random homogeneous p-th order multivariate polynomials, i.e.[1]

$$\mathbf{f}_{N,p}(x) = \frac{1}{\sqrt{N}} \sum_{i_1,...,i_p=1}^{N} J_{i_1,...,i_p} x_{i_1} \cdots x_{i_p} = \frac{1}{\sqrt{N}} \langle J, x^{\otimes p} \rangle, \tag{1}$$

with all entries of the tensor $J = (J_{i_1,...,i_p})_{i_1,...,i_p}$ iid standard normal distributed. However it is straightforward to calculate their covariance function $\mathsf{Cov}(\mathbf{f}_{N,p}(x), \mathbf{f}_{N,p}(y)) = \frac{1}{N}\langle x, y\rangle^p$, which fully characterizes a centered Gaussian random function. And it turns out that the mixtures of independent p-spin glasses

$$\mathbf{f}_N(x) = \sum_{p=0}^{\infty} a_p \mathbf{f}_{N,p}(x), \qquad a_n \in \mathbb{R},$$

1. Spin glasses were historically defined on the domain $\{-1, +1\}^N$. Their spherical counterpart was therefore defined on the the sphere of radius $\sqrt{N}$ which contains this set. However, with the definition of a rescaled inner product (the "overlap"), they are effectively mapped to the unit sphere. Also note that the canonical "Hamiltonian" $H_{N,p}$ is of the form $H_{N,p}(x) = N\mathbf{f}_{N,p}(\sqrt{N}x)$. The outer scaling is compatible with ours, since results in statistical mechanics are proven about $H_{N,p}(x)/N$.

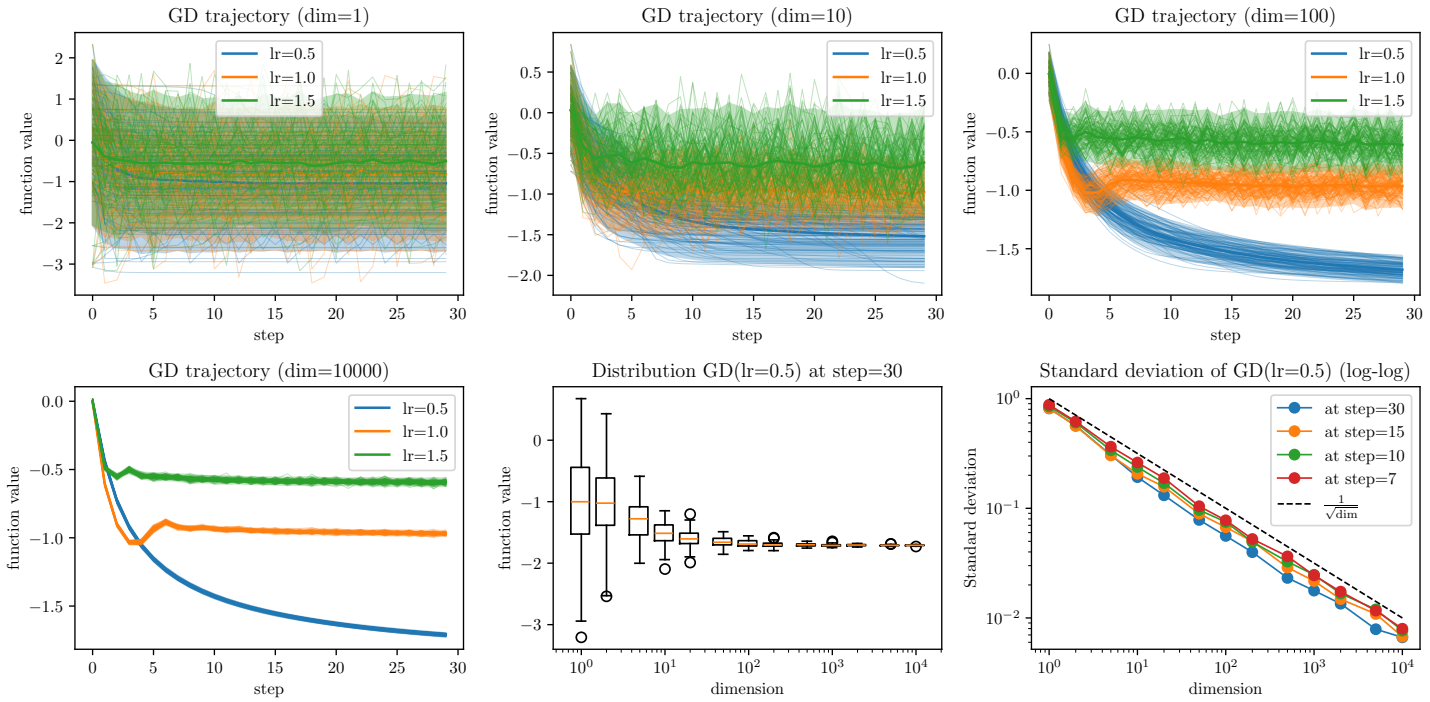


Figure 2: For each learning rate hyperparameter we simulated 100 gradient descent trajectories on a centered Gaussian random function with covariance $\mathcal{C}_{\mathbf{f}}(x, y) = \frac{1}{N} \exp(-\frac{\|x-y\|^2}{2})$. Plotted are the trajectories, their empirical mean and ribbons representing twice the empirical standard deviation. In the lower right we show how the variance of the trajectories decrease with the dimension. See Remark 16 for more details.

have the covariance $\mathcal{C}_{\mathbf{f}_N}(x, y) = \sum_{p=0}^{\infty} a_p^2 \langle x, y \rangle^p$, which exhaust all continuous isotropic covariance kernels on the sphere that are valid in all dimensions (Schoenberg, 1942).

Results about 'spin glasses', such as the celebrated Parisi-formula are therefore general results about isotropic random functions on the sphere. A recent review of results relevant for optimization can be found in (Auffinger et al., 2023). The Parisi-formula (Parisi, 1980; Talagrand, 2006; Panchenko, 2013; Huang and Sellke, 2024) provides the limiting value of the global maximum/minimum of spin glasses in the high dimensional limit. Beyond this static analysis, optimization algorithms have recently been analyzed as well. In particular, there is a hardness result (Huang and Sellke, 2022), which establishes an algorithmic barrier that does not necessarily coincide with the global minimum. Even more recently, Sellke (2024) showed for pure p-spin glasses that this algorithmic barrier may be reached by 'Langevin-dynamics', i.e. a continuous time model of stochastic gradient descent.

The tensor form of spin glasses allows for the use of message passing algorithms, specifically 'Approximate Message Passing' (AMP) (Donoho et al., 2009). AMP was originally developed to understand programs that repeatedly multiply the same random matrix to non-linear functions of the iterate. Since the same random matrix is used repeatedly, the iterate becomes correlated with the matrix and it becomes necessary to control this 'self-interaction' by conditioning. This approach has been generalized to multiple matrices in

the 'Tensor Program' series (e.g. Yang, 2021) and it has also been generalized to spin glasses made up of Tensors (Alaoui et al., 2021). By translating first order algorithms into AMP (Celentano et al., 2020), 'predictable-progress' on spin glasses may be recoverable from the general concept of 'state-evolution' (Bayati and Montanari, 2011; Javanmard and Montanari, 2013; Alaoui et al., 2021). 'State-evolution', referred to as the 'Master theorem' in (Yang, 2021) provides very general asymptotic statements about AMP algorithms. However, the random function requires a tensor representation to be expressed as AMP, which is currently unavailable for general isotropic random functions on $\mathbb{R}^N$. AMP is therefore not applicable. This may change in the future as the recent characterization of (non-stationary) isotropic covariance kernels (Benning and Schölpple, 2025) represents significant progress towards a tensor representation of general isotropic random functions.

In contrast to AMP, our proof technique does not rely on a Tensor representation and is entirely functional in nature. We have to use a similar conditioning technique to control the self-interaction, but our conditioning is closer in spirit to the one that is used in **Bayesian optimization** (BO) (e.g. Kushner, 1964; Frazier, 2018). This suggests that it may be easier to analyze Bayesian optimization algorithms using our approach. And the intuitiveness of our approach would only be obfuscated by a translation of first order algorithms into AMP.

Our proof relies on two new techniques: custom coordinate systems and a functional conditioning technique formalized by Benning (2026). These could be combined with techniques from the AMP literature to significantly strengthen our results in future work. For example, it should be possible to strengthen convergence in probability to almost sure convergence.

Organisation of the article In Section 2 we formalize the setting (gradient span optimization algorithms and (non-stationary) isotropic GRFs) and state the main result. Our asymptotic formulation of the original problem requires a dimensional scaling of the random functions that is also standard in statistical mechanics. This scaling, crucial to our approach, is discussed in Section 3. The proof of our main result is given in Section 4, first as a sketch assuming stationary isotropy and then in detail. In Appendix A we show random quadratic functions to be a special case of isotropic Gaussian random functions, in Appendix B we provide a refresher on conditional Gaussian distributions, as we rely heavily upon them, and in Appendix C we prove results that allow the strict positive definiteness assumption to be dropped in the case of *stationary* isotropy.

2. Setting and main results

To enable our asymptotic analysis, a careful setting of the scene is required. Both, the optimization algorithms considered and the distribution of $\mathbf{f}_N$ need a representation independent of the dimension N to allow for an analysis of $N \to \infty$.

2.1 The class of optimization algorithms $\mathfrak{G}$

Given a sufficiently smooth function $f : \mathbb{R}^N \to \mathbb{R}$, a naive optimization algorithm is given by gradient descent, whose evaluation points are recursively defined as

$$x_n = x_{n-1} - \alpha \nabla f(x_{n-1})$$

					preconditioning	
Example	gradient descent	momentum methods	conjugate gradient descent	adaptive scalar learning rate	Adam	Shampoo
is GSA?	✓	✓	✓	✓	✗	✗

Table 1: What algorithms are gradient span algorithms (GSA)?

with some initialization $x_0 \in \mathbb{R}^N$ and a learning rate α. The reader might want to keep gradient descent in mind as a toy example, but everything we prove holds for a much larger class of first order algorithms that contains many standard optimizers. Gradient span algorithms (GSA) are a very general class of first order algorithms which pick the n-th point x_n from the previous span of gradients

$$x_n \in \operatorname{span}\{x_0, \nabla f(x_0), \dots, \nabla f(x_{n-1})\}. \tag{2}$$

GSAs contain classic gradient descent, momentum methods such as heavy-ball momentum (Polyak, 1964) and Nesterov's momentum (e.g. Nesterov, 2018), and also the conjugate gradient method (Hestenes and Stiefel, 1952). GSAs do not only encompass minimizers but also maximizers and all sorts of other algorithms. While the initial point x_0 is typically not included in the span, we admit this generalization to allow for concepts such as 'weight normalization' (Salimans and Kingma, 2016), which project points back to the sphere (see Remark 5 for further details). Since the defining property (2) of gradient span algorithms is not sufficient to identify a particular GSA, we define a very general parametric family of GSAs in Definition 1. This family includes all the algorithms mentioned above. Importantly, the parametrization chosen does not use dimension specific information and a fixed gradient span algorithm (such as gradient descent) can therefore be used for all dimensions. In Table 1 we use a non-exhaustive list of popular algorithms to illustrate the generality of gradient span algorithms. The main limitation are preconditioning methods, which are further discussed in Remark 4.

Definition 1 (General gradient span algorithm) *For a starting point $x_0 \in \mathbb{R}^N$ and a function $f : \mathbb{R}^N \to \mathbb{R}$, a* gradient span algorithm $\mathfrak{G}$ *selects*

$$x_n := \mathfrak{G}(f, x_0, n) = h_n^{(x)} x_0 + \sum_{k=0}^{n-1} h_{n,k}^{(g)} \nabla f(x_k), \tag{3}$$

where we assume the prefactors $h_n = (h_n^{(x)}) \cup (h_{n,k}^{(g)})_{k=0,\dots,n-1}$, using the union $\cup$ to indicate a concatenation of tuples, to be functions $h_n = h_n(I_{n-1})$ of the previous dimensionless information I_{n-1} available at time n, that is

$$I_n := \Big(f(x_k) : k \le n\Big) \cup \Big(\langle v, w\rangle : v, w \in (x_0) \cup G_n\Big) \quad \textit{with} \quad G_n := \big(\nabla f(x_k) : k \le n\big).$$

The algorithm $\mathfrak{G}$ is called x_0-agnostic*, if*

- *it remains in the gradient span shifted by x_0, i.e.*

$$h_n^{(x)} = 1 \qquad \forall n \in \mathbb{N}.$$

- *The prefactors h_n do not use the inner products with x_0, that is they are functions of the reduced information*

$$I_n^{\setminus x_0} = \Big(f(X_k), \langle \nabla f(x_k), \nabla f(x_l) \rangle : k, l \leq n \Big).$$

As mentioned above the reader might just think about ordinary gradient descent, which is also x_0-agnostic. We introduce the 'x_0-agnostic' property, to allow for arbitrary initialization distributions in the stationary isotropic case. Without this property, the algorithm could, for example, use the length $\|X_0\|$ as an indicator to switch between optimizers and therefore cause non-deterministic behavior.

Remark 2 (Initial point is not special) *It should perhaps be noted, that the initial point x_0 is not the only point that may be used by gradient span algorithms, since we have*

$$\operatorname{span}\{x_0, \nabla f(x_0), \ldots, \nabla f(x_{n-1})\} = \operatorname{span}\{x_0, \ldots, x_{n-1}, \nabla f(x_0), \ldots, \nabla f(x_{n-1})\}$$

by induction over n as x_n is selected from this span.

Remark 3 (The same algorithm for all dimensions) *Since the main objective of this article is a dimensional limit statement for fixed algorithms (i.e. fixed choice of prefactors h) it is important to note that every fixed gradient span algorithm $\mathfrak{G}$ is well-defined in all dimensions because the scalar product $\langle \cdot, \cdot \rangle$ is well defined for every dimension $N \in \mathbb{N}$.*

Remark 4 (No preconditioning) *Gradient span algorithms notably exclude preconditioning methods such as Adam (Kingma and Ba, 2015) and Shampoo (Gupta et al., 2018), which apply a preconditioning matrix P_n to the gradients and thereby leave the span of gradients. The reason is related to the previous remark: We want "the same algorithm in every dimension" and the GSA family only uses dimensionless information which can be shown to converge as the dimension tends to infinity. In contrast, preconditioning methods collect information about each coordinate separately and treat them differently accordingly. This makes it much more challenging to analyze their limiting behavior in the dimension as the limiting behavior can no longer be captured by a simple number. Since Adam still seems to behave predictably in high dimension (Figure 1), it is likely possible, albeit much more challenging to analyze its limiting behavior (see also Section 5, A4).*

Remark 5 (Projection) *We claimed that the inclusion of the initial point x_0 into the span, or equivalently the prefactor $h_n^{(x)}$, would allow for algorithms projecting back to the sphere or ball. In the following, we show that our general gradient span algorithms also contain gradient span algorithms with spherical projections.*

Assume that the point x_n is defined by

$$x_n := P\tilde{x}_n \quad \text{with} \quad \tilde{x}_n = \tilde{h}_n^{(x)} x_0 + \sum_{k=0}^{n-1} \tilde{h}_{n,k}^{(g)} \nabla f(x_k),$$

where P is either a projection to the sphere or the ball. Note that this implies either a division by $\|\tilde{x}_n\|$ in case of the sphere, or a division by $\max\{\|\tilde{x_n}\|, 1\}$ in case of the ball.

But the norm

$$\|\tilde{x}_n\|^2 = (\tilde{h}_n^{(x)})^2 \underbrace{\|x_0\|^2}_{\in I_n} + 2\tilde{h}_n^{(x)} \sum_{k=0}^{n-1} \tilde{h}_{n,k}^{(g)} \underbrace{\langle \nabla f(x_k), x_0 \rangle}_{\in I_n} + \sum_{k,l=0}^{n-1} \tilde{h}_{n,k}^{(g)} \tilde{h}_{n,l}^{(g)} \underbrace{\langle \nabla f(x_k), \nabla f(x_l) \rangle}_{\in I_n}.$$

is a function of the information in I_{n-1} and can therefore be used to define

$$h_n^{(x)} := \tfrac{\tilde{h}_n^{(x)}}{\|\tilde{x}_n\|} \qquad h_{n,k}^{(g)} := \tfrac{\tilde{h}_{n,k}^{(g)}}{\|\tilde{x}_n\|}$$

in the case of the projection to the sphere and similarly for the projection to the ball.

2.2 The class of random function distributions

For a random function[2] $\mathbf{f}_N : \mathbb{R}^N \to \mathbb{R}$ on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ we denote the mean and covariance functions by

$$\mu_{\mathbf{f}_N}(x) := \mathbb{E}[\mathbf{f}_N(x)] \quad \text{and} \quad \mathcal{C}_{\mathbf{f}_N}(x,y) := \mathsf{Cov}(\mathbf{f}_N(x), \mathbf{f}_N(y)), \qquad x, y \in \mathbb{R}^N.$$

As we want to prove a limit theorem for $N \to \infty$ to make predictions about the high dimensional cost functions found in practice, the mean $\mu_{\mathbf{f}_N}$ and covariance $\mathcal{C}_{\mathbf{f}_N}$ needs to be defined for every dimension. They should also remain constant in some sense to avoid arbitrary sequences. As the domain $\mathbb{R}^N$ changes over N this requires a parametric form provided by covariance kernels. To obtain non-trivial results, some scaling is also required (discussed in Section 3).

Definition 6 (Scaled sequence of isotropic Gaussian random functions (GRF)) *A scaled sequence $(\mathbf{f}_N)_{N\in\mathbb{N}}$ of GRFs $\mathbf{f}_N : \mathbb{R}^N \to \mathbb{R}$ is called*

(i) ***(non-stationary) isotropic****, if*

$$\mu_{\mathbf{f}_N}(x) = \mu\big(\tfrac{\|x\|^2}{2}\big) \qquad \text{and} \qquad \mathcal{C}_{\mathbf{f}_N}(x,y) = \tfrac{1}{N}\kappa\big(\tfrac{\|x\|^2}{2}, \tfrac{\|y\|^2}{2}, \langle x, y\rangle\big), \quad x, y \in \mathbb{R}^N, \quad (4)$$

for continuous functions $\mu\colon \mathbb{R} \to \mathbb{R}$ and $\kappa\colon D \to \mathbb{R}$ with $D = \{\lambda \in \mathbb{R}^2_{\geq 0} \times \mathbb{R} : |\lambda_3| \leq 2\sqrt{\lambda_1\lambda_2}\} \subseteq \mathbb{R}^3$. In that case we write $\mathbf{f}_N \sim \mathcal{N}(\mu, \kappa)$.

(ii) ***stationary isotropic****, if*

$$\mu_{\mathbf{f}_N}(x) = \mu \in \mathbb{R} \qquad \text{and} \qquad \mathcal{C}_{\mathbf{f}_N}(x,y) = \tfrac{1}{N} C\big(\tfrac{\|x-y\|^2}{2}\big), \quad x, y \in \mathbb{R}^N,$$

for a continuous $C\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$. In that case we write $\mathbf{f}_N \sim \mathcal{N}(\mu, C)$.

Remark 7 (Motivation for isotropy) *While we define isotropy as a functional form of the covariance function, isotropy is best understood axiomatically as an invariance to rotation and reflection (e.g. Benning and Schölpple, 2025). To motivate this invariance in the context of predictability let us highlight a case that must be avoided.*

2. Recall that 'random process', 'stochastic process' and 'random field' are used synonymously for 'random function' in the literature, that their law is characterized by all finite dimensional marginals (e.g. Klenke, 2014, Thm. 14.36) and that Gaussian random functions are fully determined by their mean and covariance.

Think about some one-dimensional function $f_1 \colon \mathbb{R}^1 \to \mathbb{R}$ where the values along the optimization path (say of gradient descent) strongly depend on the initialization. Lift this one dimensional function into $\mathbb{R}^N$ by

$$f_N(x) := f_1(\langle v, x \rangle)$$

for some direction vector v. Then f_N certainly does not share the features of high dimensional cost functions encountered in machine learning, since the cost along the optimization path depends heavily on the initialization again. To explain the phenomenon of predictable cost sequences (cf. Figure 1) the model for cost functions therefore should not be reliant on particular directions v. Thus, we assume that directions should be exchangeable, which suggests at least rotation and reflection invariant random functions as a model.

Non-stationary isotropy as a distributional assumption for cost functions was introduced by Benning and Döring (2024) as a generalization of the common stationary isotropy assumption (e.g. Dauphin et al., 2014). This was motivated by the fact that simple linear models already break the stationary isotropy assumptions. Linear models equipped with a mean squared loss can also be used to derive random quadratic functions (Section A), which have been analyzed before and are isotropic.

Before getting to the main results, we have to define one more concept. For a given mean function μ and covariance kernel κ we are interested in a sequence of (non-stationary) isotropic random functions $\mathbf{f}_N \sim \mathcal{N}(\mu, \kappa)$ over the dimension N. But it is not clear that a function κ (resp. C) corresponds to a random function for every dimension, as it may not always define a positive definite kernel. If it does, we speak of validity in all dimensions.

Definition 8 (Valid in all dimensions) *We say κ (resp. C) is valid in all dimensions if for any dimension N there exists a (non-stationary) isotropic random function $\mathbf{f}_N \sim \mathcal{N}(\mu, \kappa)$ (respectively stationary isotropic $\mathbf{f}_N \sim \mathcal{N}(\mu, C)$).*

Since one can always restrict the domain of a random function $\mathbf{f}_N$ to a lower dimensional subspace it should be clear that the only possible type of restriction is an upper bound on the dimension. We refer the reader to Sasvári (2013, Section 3.8) for the "Schoenberg" characterization of stationary isotropic covariance kernels that are valid up to dimension N (Schoenberg, 1938). The stationary isotropic covariance functions which are valid in all dimensions are given exactly by the kernels of the form

$$C(r) = \int_{[0,\infty)} \exp(-t^2 r) \nu(dt), \qquad r \geq 0, \tag{5}$$

for some finite measure ν on $[0, \infty)$ (Sasvári, 2013, Theorem 3.8.5). The (non-stationary) isotropic kernels have only very recently been classified (Benning and Schölpple, 2025). The continuous isotropic kernels valid in all dimensions are of the form

$$\kappa(\tfrac{\|x\|}{2}, \tfrac{\|y\|}{2}, \langle x, y \rangle) = \sum_{n=0}^{\infty} \alpha_n(\|x\|, \|y\|) \langle \tfrac{x}{\|x\|}, \tfrac{x}{\|x\|} \rangle^n,$$

where the α_n are continuous positive definite kernels on $[0, \infty)$. With $\alpha_n(x, y) = a_n \|x\|^n \|y\|^n$ for $a_n \in \mathbb{R}$, dot product kernels represent a wide class of (non-stationary) isotropic random functions that are not stationary isotropic on $\mathbb{R}^N$ but also valid in all dimensions. Restricted to the sphere, dot product kernels are known as spin glasses in statistical mechanics.

2.3 Main result: predictable progress in high dimensions

Our main result proves that the optimization path $\mathbf{f}_N(X_0), \mathbf{f}_N(X_1), ...$ is asymptotically deterministic for large dimension N. In addition, we prove a similar statement about the gradient norms $\|\nabla \mathbf{f}_N(X_n)\|^2$ and more generally about their scalar products. As we assume continuity of the prefactors h_n for the gradient span algorithm $\mathfrak{G}$, they are also asymptotically deterministic due to continuous mapping. To reduce the initial complexity, we first state the simplified corollary for the stationary isotropic case. In this case some of the technical assumptions can be removed which should allow the reader to focus on the core message.

Corollary 9 (Predictable progress, isotropic case) *Let C define a stationary isotropic kernel valid in all dimensions (Definition 8) and let $\mathbf{f}_N \sim \mathcal{N}(\mu, C)$ be a sequence of scaled isotropic Gaussian random functions in N (Definition 6). Let $\mathfrak{G}$ be a general gradient span algorithm (Definition 1), where we assume that its prefactors $h_n = h_n(I_{n-1})$ are continuous in the information I_{n-1}, and utilize the most recent gradients, i.e. $h^{(g)}_{n,n-1} \neq 0$ for all $n \in \mathbb{N}$. Let the algorithm $\mathfrak{G}$ be applied to $\mathbf{f}_N$ with independent, possibly random starting point $X_0 \in \mathbb{R}^N$ and denote by X_n the points resulting from $\mathfrak{G}$. Finally, assume that $\|X_0\| = \lambda$ almost surely. Then there exist characteristic real numbers*

$$\mathfrak{f}_n = \mathfrak{f}_n(\mathfrak{G}, \mu, \kappa, \lambda) \in \mathbb{R} \quad \textit{and} \quad \mathfrak{g}_{n,k} = \mathfrak{g}_{n,k}(\mathfrak{G}, \mu, \kappa, \lambda) \in \mathbb{R},$$

such that, for all $\epsilon > 0$, $n, k \in \mathbb{N}$,

$$\lim_{N\to\infty} \mathbb{P}\big(|\mathbf{f}_N(X_n) - \mathfrak{f}_n| > \epsilon\big) = 0 \quad \textit{and}$$
$$\lim_{N\to\infty} \mathbb{P}\Big(\big|\langle \nabla \mathbf{f}_N(X_n), \nabla \mathbf{f}_N(X_k)\rangle - \mathfrak{g}_{n,k}\big| > \epsilon\Big) = 0.$$

If furthermore the algorithm is x_0-agnostic (e.g. gradient descent), then the limiting values do not depend on λ and the starting point X_0 can have arbitrary distribution independent of $\mathbf{f}_N$.

Since most algorithms are x_0-agnostic, this result explains the approximately deterministic behavior in high dimension for stationary isotropic GRFs. The general case is covered by the following remark.

Remark 10 (Initialization on the Sphere) *Initialization procedures like Glorot initialization (Glorot and Bengio, 2010) select the entries of X_0 independent, essentially identically distributed, and scaled in such a way that the norm does not diverge. The norm therefore obeys a law of large numbers and is thus plausibly deterministic in high dimension. The assumption, $\|X_0\| = \lambda$ almost surely, is therefore realistic. This explains the phenomenon of (approximately) predictable progress of optimizers started in a randomly selected initial point using Glorot initialization, as observed in Figure 1.*

The following proof of Corollary 9 should also serve as a reading aid for Theorem 12, which states the same result in greater generality. The proof shows how the general statements of Theorem 12 simplify to the more digestible statement for stationary isotropic GRFs.

Proof Observe that $\mathbf{f}_N(X_k)$ and $\langle \nabla \mathbf{f}_N(X_k), \nabla \mathbf{f}_N(X_l) \rangle$ for $k, l \leq n$ are members of the random information vector $\mathbf{I}_n$ defined in Theorem 12. Their convergence therefore follows immediately from the convergence of the information vector $\mathbf{I}_n$ proven in Theorem 12. Since all stationary isotropic random functions are (non-stationary) isotropic, this Corollary follows immediately from Theorem 12 once we verified the additional required assumptions.

The first assumption is the smoothness assumption on the covariance function (Assumption 11). It follows from the fact that *all* stationary isotropic random function, which are valid in all dimensions, are infinitely differentiable by the Schoenberg characterization (Sasvári, 2013, Theorem 3.8.5), see also (5). The second assumption is the strict positive definiteness of $(\mathbf{f}_N, \nabla \mathbf{f}_N)$, which also holds for all stationary isotropic random functions valid in all dimensions by Corollary 42 below. Note that Corollary 42 requires that the random function $\mathbf{f}_N$ is not almost surely constant. But if $\mathbf{f}_N$ were almost surely constant, then we get asymptotically deterministic behavior from the fact that $\mathbf{f}_N(x_0) \sim \mathcal{N}(\mu, \frac{1}{N} C(0))$, i.e. $\mathbf{f}_N(x_0) \to \mu$. Since $\mathbf{f}_N$ is almost surely constant, we thus obtain $\mathbf{f}_N \to \mu$ uniformly in probability. The x_0-agnostic case follows from the stationary case of Proposition 24. ■

We now come to the main theorem of the paper, predictable progress of gradient span algorithms on (non-stationary) isotropic random functions in high dimensions. While kernels of stationary isotropic GRFs that are valid in all dimensions are always smooth, and the mean function is always constant, the same may not be the case for the more general situation of (non-stationary) isotropy. We will therefore assume the following smoothness properties:

Assumption 11 (Sufficiently smooth) *With $\kappa_i(\lambda_1, \lambda_2, \lambda_3) := \frac{d}{d\lambda_i} \kappa(\lambda_1, \lambda_2, \lambda_3)$, we assume the partial derivatives*

$$\kappa_{12}, \quad \kappa_{13}, \quad \kappa_{23} \quad \text{and} \quad \kappa_{33} \tag{6}$$

of the kernel κ exist and are continuous. Furthermore, we assume the derivative of the mean μ exists and is continuous.

In Equation (10) we can see that the covariance of derivatives is directly related to the derivatives of the covariance function. For $\nabla \mathbf{f}_N$ to exist, it is therefore natural that $\mathcal{C}_{\mathbf{f}_N}$ has to be differentiable. The covariance has to be two times differentiable in a sense, as we require the following to be well defined

$$\mathsf{Cov}(\partial_i \mathbf{f}_N(x), \partial_j \mathbf{f}_N(y)) = \partial_{x_i} \partial_{x_j} \mathcal{C}_{\mathbf{f}_N}(x, y).$$

In the case of (non-stationary) isotropic functions, this requires the existence of (6) by Lemma 20. And the existence of the terms in (6) is in fact necessary and sufficient for $\nabla \mathbf{f}_N$ to exist in an L^2-sense (e.g. Gihman and Skorokhod, 1974, Ch. IV, §3, Thm. 4). Our additional assumption of continuity is slightly weaker than typical sufficient conditions for a point-wise defined, continuous version of $\nabla \mathbf{f}_N$ to exist (e.g. Adler and Taylor, 2007). Intuitively, we therefore simply assume $\nabla \mathbf{f}_N$ to be continuous almost surely.

Here is our main result:

Theorem 12 (Predictable progress, general case) *Assume κ defines a kernel valid in all dimensions (Definition 8) and let $\mathbf{f}_N \sim \mathcal{N}(\mu, \kappa)$ be a sequence of scaled (non-stationary)*

isotropic Gaussian random functions (Definition 6) in N. Assume that μ and κ are sufficiently smooth (Assumption 11) and that the covariance of $(\mathbf{f}_N, \nabla \mathbf{f}_N)$ is strictly positive definite (Definition 40). Let $\mathfrak{G}$ be a general gradient span algorithm (Definition 1), where we assume that its prefactors $h_n = h_n(I_{n-1})$ are continuous in the information I_{n-1}, and utilize the most recent gradients, i.e. $h^{(g)}_{n,n-1} \neq 0$ for all $n \in \mathbb{N}$. Let the algorithm $\mathfrak{G}$ be applied to $\mathbf{f}_N$ with independent, possibly random starting point $X_0 \in \mathbb{R}^N$ and denote by X_n the points resulting from $\mathfrak{G}$. Finally, assume that $\|X_0\| = \lambda$ almost surely. Then, for any $n \in \mathbb{N}$, the random information vector

$$\mathbf{I}_n := \Big(\mathbf{f}_N(X_k) : k \leq n\Big) \cup \Big(\langle v, w\rangle : v, w \in (x_0) \cup G_n\Big) \quad \textit{with} \quad G_n := \big(\nabla \mathbf{f}_N(X_k) : k \leq n\big).$$

converges, as the dimension increases ($N \to \infty$), in probability against a deterministic information vector $\mathscr{I}_n = \mathscr{I}_n(\mathfrak{G}, \mu, \kappa, \lambda)$.

An application of the stochastic triangle inequality yields the following corollary, which perfectly describes Figure 1.

Corollary 13 (Asymptotically identical progress over initializations) *Assume the setting of Theorem 12 and let $X_0^{(1)}$, $X_0^{(2)}$ be random initialization points selected independently from $\mathbf{f}_N$. In the non-stationary case, we additionally assume that we we have almost surely $\|X_0^{(1)}\| = \|X_0^{(2)}\| = \lambda \in \mathbb{R}$. Let the sequences $(X_n^{(i)})_{n\in\mathbb{N}}$ be generated from the initialization $X_0^{(i)}$ by the same general gradient span algorithm $\mathfrak{G}$. Then we have, for all $n \in \mathbb{N}$ and all $\epsilon > 0$,*

$$\lim_{N\to\infty} \mathbb{P}\Big(\max_{k\leq n}\Big|\mathbf{f}_N(X_k^{(1)}) - \mathbf{f}_N(X_k^{(2)})\Big| > \epsilon\Big) = 0.$$

Proof The proof is essentially an application of the stochastic triangle inequality

$$\Big\{|X - Z| > \epsilon\Big\} \subseteq \Big\{|X - Y| > \frac{\epsilon}{2}\Big\} \cup \Big\{|Y - Z| > \frac{\epsilon}{2}\Big\},$$

which yields by Theorem 12

$$\begin{aligned}
&\lim_{N\to\infty} \mathbb{P}\Big(\max_{k\leq n}\Big|\mathbf{f}_N(X_k^{(1)}) - \mathbf{f}_N(X_k^{(2)})\Big| > \epsilon\Big) \\
&\leq \lim_{N\to\infty} \sum_{k\leq n}\Big[\mathbb{P}\Big(\Big|\mathbf{f}_N(X_k^{(1)}) - \mathfrak{f}_k\Big| > \frac{\epsilon}{2}\Big) + \mathbb{P}\Big(\Big|\mathbf{f}_N(X_k^{(2)}) - \mathfrak{f}_k\Big| > \frac{\epsilon}{2}\Big)\Big] \\
&= 0.
\end{aligned}$$

■

Similar to Paquette et al. (2022), we obtain so-called asymptotically deterministic halting times as a corollary. The ϵ-halting is defined as

$$T_\epsilon := \inf\{n > 0 : \|\nabla \mathbf{f}_N(X_n)\|^2 \leq \epsilon\}.$$

We are going to prove it is asymptotically equal to the asymptotic ϵ-halting time

$$\tau_\epsilon := \inf\{n > 0 : \mathfrak{g}_n \leq \epsilon\} \in \mathbb{N},$$

where $\mathfrak{g}_n := \mathfrak{g}_{n,n}$ is the stochastic limit of $\|\nabla \mathbf{f}_N(X_n)\|^2$ in the dimension. In practical terms this means that optimization always stops at roughly the same time in high dimension.

Corollary 14 (Asymptotically deterministic halting times) *If $\epsilon \notin (\mathfrak{g}_n)_{n\in\mathbb{N}}$, then the halting time is asymptotically deterministic*

$$\lim_{N\to\infty} \mathbb{P}(T_\epsilon = \tau_\epsilon) = 1.$$

If $\epsilon = \mathfrak{g}_n$ for some n, then

$$\lim_{N\to\infty} \mathbb{P}(T_\epsilon \in [\tau_\epsilon, \tau_\epsilon^+]) = 1 \quad \textit{with} \quad \tau_\epsilon^+ := \inf\{n > 0 : \mathfrak{g}_n < \epsilon\}.$$

Proof The proof is identical to the proof of Theorem 4 of Paquette et al. (2022). ■

Remark 15 (Explicit evolution formula) *In the proof of Theorem 12 we construct explicit formulas for the conditional distribution of $(\mathbf{f}_N(X_n), \nabla \mathbf{f}_N(X_n))$. This remark summarizes these formulas. Due to its terse nature as a summary it may be difficult to read without the context of the proof and may be more helpful as an overview after the reader has familiarized themselves with the proof. In this summary we also ignore mathematical conditioning challenges.*

Let $\mathbf{f}_N \sim \mathcal{N}(\mu, \kappa)$, let the filtration be given by

$$\mathcal{F}_n := \sigma\Big(X_0, \mathbf{f}_N(X_k), \nabla \mathbf{f}_N(X_k) : k \le n\Big)$$

and let $\mathbf{v}_{[0:d_n)} = (\mathbf{v}_0, \dots, \mathbf{v}_{d_n-1})$ (notation defined in (12)) be the orthonormal basis of

$$V_n := \operatorname{span}\Big\{x_0, \nabla \mathbf{f}_N(X_k) : k < n\Big\}$$

obtained using the Gram-Schmidt process and $\mathbf{w}_{[d_n:N)}$ any orthonormal basis of $V_n^\perp$. Then for $Y_0, \dots, Y_N \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ independent of $\mathcal{F}_{n-1}$ the following equality holds in distribution:

$$\left[\begin{pmatrix} \mathbf{f}_N(X_n) \\ D_{\mathbf{v}_0}\mathbf{f}_N(X_n) \\ \vdots \\ D_{\mathbf{v}_{d_n-1}}\mathbf{f}_N(X_n) \end{pmatrix} \middle| \mathcal{F}_{n-1} \right] \overset{d}{=} \left[\begin{pmatrix} \overbrace{\mathbf{m}_n + \tfrac{1}{\sqrt{N}}[\mathbf{C}_n]^{\frac{1}{2}} \begin{pmatrix} Y_0 \\ Y_1 \\ \vdots \\ Y_{d_n} \end{pmatrix}}^{\to 0 \quad (d_n \le n \ll N)}, \\ \tfrac{\sigma_n}{\sqrt{N}} \begin{pmatrix} Y_{d_n} \\ \vdots \\ Y_{N-1} \end{pmatrix} \end{pmatrix}\right] \tag{7}$$

The terms $\mathbf{m}_n$ and $\mathbf{C}_n$ and σ_n^2 all converge as $N \to \infty$. Since there are only finitely many $(Y_0, \dots, Y_{d_n})$, the term around $\mathbf{C}_n$ becomes asymptotically irrelevant due to scaling. On the contrary σ_n remains important because there are $N - d_n$ many Y_i left such that their squared sum is of constant order. By (51) we have that this 'residual variance' is given by

$$\sigma_n^2 := \kappa_3\big(\tfrac{\|X_n\|^2}{2}, \tfrac{\|X_n\|^2}{2}, \|X_n\|^2\big) - (\boldsymbol{\Sigma}^{w,N}_{[0:n),n})^T [\boldsymbol{\Sigma}^{w,N}_{[0:n)}]^{-1} \boldsymbol{\Sigma}^{w,N}_{[0:n),n} \in \mathbb{R},$$

with $\kappa_m(\lambda_1,\lambda_2,\lambda_3) := \frac{d}{d\lambda_m}\kappa(\lambda_1,\lambda_2,\lambda_3)$ *and*

$$\Sigma^{w,N}_{[0:n),n} := \begin{pmatrix} \kappa_3\big(\frac{\|X_0\|^2}{2}, \frac{\|X_n\|^2}{2}, \langle X_0, X_n\rangle\big) \\ \vdots \\ \kappa_3\big(\frac{\|X_{n-1}\|^2}{2}, \frac{\|X_n\|^2}{2}, \langle X_{n-1}, X_n\rangle\big) \end{pmatrix}$$

$$\Sigma^{w,N}_{[0:n)} := \begin{pmatrix} \kappa_3\big(\frac{\|X_0\|^2}{2}, \frac{\|X_0\|^2}{2}, \|X_0\|^2\big) & \cdots & \kappa_3\big(\frac{\|X_0\|^2}{2}, \frac{\|X_{n-1}\|^2}{2}, \langle X_0, X_{n-1}\rangle\big) \\ \vdots & \ddots & \vdots \\ \kappa_3\big(\frac{\|X_{n-1}\|^2}{2}, \frac{\|X_0\|^2}{2}, \langle X_{n-1}, X_0\rangle\big) & \cdots & \kappa_3\big(\frac{\|X_{n-1}\|^2}{2}, \frac{\|X_{n-1}\|^2}{2}, \|X_{n-1}\|^2\big) \end{pmatrix}.$$

Note that all the $\langle X_i, X_j\rangle$ *terms converge (Ind-III) as they are made up of components from the information vector. Consequently all the terms above converge in* N *such that* σ_n^2 *converges in* N*. Moreover*

$$\mathbf{m}_n = \begin{pmatrix} \mu\big(\frac{\|X_n\|^2}{2}\big) \\ \mu'\big(\frac{\|X_n\|^2}{2}\big)\langle X_n, \mathbf{v}_0\rangle \\ \vdots \\ \mu'\big(\frac{\|X_n\|^2}{2}\big)\langle X_n, \mathbf{v}_{d_n-1}\rangle \end{pmatrix} + \Sigma^{v,N}_{[0:n),n}{}^T[\Sigma^{v,N}_{[0:n)}]^{-1}\big(Z(\mathbf{v}_{[0:d_n)}; X_{[0:n)}) - \mu^{(v,N)}_{[0:n)}\big).$$

where the following row-major matrices are flattened into column vectors

$$Z(\mathbf{v}_{[0:d_n)}; X_{[0:n)}) = \operatorname{vec}\begin{pmatrix} \mathbf{f}_N(X_0) & \cdots & \mathbf{f}_N(X_{n-1}) \\ D_{\mathbf{v}_0}\mathbf{f}_N(X_0) & \cdots & D_{\mathbf{v}_0}\mathbf{f}_N(X_{n-1}) \\ \vdots & & \vdots \\ D_{\mathbf{v}_{d_n-1}}\mathbf{f}_N(X_0) & \cdots & D_{\mathbf{v}_{d_n-1}}\mathbf{f}_N(X_{n-1}) \end{pmatrix} \tag{8}$$

$$\mu^{(v,N)}_{[0:n)} = \operatorname{vec}\begin{pmatrix} \mu\big(\frac{\|X_0\|^2}{2}\big) & \cdots & \mu\big(\frac{\|X_{n-1}\|^2}{2}\big) \\ \mu'\big(\frac{\|X_0\|^2}{2}\big)\langle X_0, \mathbf{v}_0\rangle & \cdots & \mu'\big(\frac{\|X_{n-1}\|^2}{2}\big)\langle X_{n-1}, \mathbf{v}_0\rangle \\ \vdots & & \vdots \\ \mu'\big(\frac{\|X_0\|^2}{2}\big)\langle X_0, \mathbf{v}_{d_n-1}\rangle & \cdots & \mu'\big(\frac{\|X_{n-1}\|^2}{2}\big)\langle X_{n-1}, \mathbf{v}_{d_n-1}\rangle \end{pmatrix}.$$

With $D_\emptyset \mathbf{f}_N = \mathbf{f}_N$ *observe that the matrices above may be reindexed by* $\{\emptyset, \mathbf{v}_0, \ldots, \mathbf{v}_{d_n-1}\}$ *and* $\{X_0, \ldots, X_{n-1}\}$ *instead of* $\{0, \ldots, d_n\}$ *and* $\{0, \ldots, n-1\}$*. This makes it easier to state the corresponding covariance matrices*

$$[\Sigma^{v,N}_{[0:n)}]_{(v,x,w,y)} = \widetilde{\mathsf{Cov}}(D_v\mathbf{f}_N(x), D_w\mathbf{f}_N(y)) \qquad v, w \in \{\emptyset, \mathbf{v}_0, \ldots, \mathbf{v}_{d_n-1}\}, x, y \in X_{[0:n)}$$
$$[\Sigma^{v,N}_{[0:n),n}]_{(v,x,w)} = \widetilde{\mathsf{Cov}}(D_v\mathbf{f}_N(x), D_w\mathbf{f}_N(X_n))$$

where $\widetilde{\mathsf{Cov}}$ *treats* $\mathbf{v}_i$ *and* X_i *like deterministic variables (formal details in the proof of Theorem 12). These are flattened to match the flattened vectors above. Finally,*

$$\mathbf{C}_n = \Sigma^{v,N}_n - \Sigma^{v,N}_{[0:n),n}{}^T[\Sigma^{v,N}_{[0:n)}]^{-1}\Sigma^{v,N}_{[0:n),n},$$

where $\boldsymbol{\Sigma}_n^{v,N}$ *is the "covariance" matrix of* $(\mathbf{f}_N(X_n), D_{\mathbf{v}_0}\mathbf{f}_N(X_n), \dots, D_{\mathbf{v}_{d_n-1}}\mathbf{f}_N(X_n))$ *treating* $\mathbf{v}_i$ *and* X_n *like deterministic variables, i.e.*

$$\boldsymbol{\Sigma}_n^{v,N} = \begin{pmatrix} \widetilde{\mathsf{Cov}}(\mathbf{f}_N(X_n), \mathbf{f}_N(X_n)) & \cdots & \widetilde{\mathsf{Cov}}(\mathbf{f}_N(X_n), D_{\mathbf{v}_{d_n-1}}\mathbf{f}_N(X_n)) \\ \vdots & & \vdots \\ \widetilde{\mathsf{Cov}}(D_{\mathbf{v}_{d_n-1}}\mathbf{f}_N(X_n), \mathbf{f}_N(X_n)) & \cdots & \widetilde{\mathsf{Cov}}(D_{\mathbf{v}_{d_n-1}}\mathbf{f}_N(X_n), D_{\mathbf{v}_{d_n-1}}\mathbf{f}_N(X_n)) \end{pmatrix}.$$

Remark 16 (Computationally tracking the evolution) *Observe that the state space is essentially given by* (8)*, which contains* $n(d_n+1) \sim n^2$ *random variables. Its covariance matrix* $\boldsymbol{\Sigma}_{[0:n)}^{v,N}$ *consequently has* $\mathcal{O}(n^4)$ *entries and may be decomposed and inverted in* $\mathcal{O}(n^6)$ *time. Since* $\boldsymbol{\Sigma}_{[0:n)}^{v,N}$ *contains the entries of the previous covariance matrices, the Cholesky decomposition can be updated such that the total complexity over all iterations up to* n *remains at* $\mathcal{O}(n^6)$ *instead of growing to* $\mathcal{O}(n^7)$*. To avoid the complexity of* $\mathcal{O}(n^6 + Nn^2)$ *when* N *is really large it is important to never compute the basis vectors* $\mathbf{v}_i$*. They are simply an orthonormal basis and it suffices to compute the gradients and* X_i *in this coordinate system. Observe that by definition of* $\mathbf{v}_{d_n}$

$$D_{\mathbf{v}_{d_n}}\mathbf{f}_N(X_n) = \left\| \begin{pmatrix} D_{\mathbf{w}_{d_n}}\mathbf{f}_N(X_n) \\ \vdots \\ D_{\mathbf{w}_{N-1}}\mathbf{f}_N(X_n) \end{pmatrix} \right\| \overset{d}{=} \frac{\sigma_n}{\sqrt{N}} \sqrt{\sum_{i=d_n}^{N-1} Y_i^2} \sim \sigma_n \frac{\chi_{N-d_n}}{\sqrt{N}} \to \sigma_n,$$

where χ_k *is the chi distribution with* k *degrees of freedom. The remaining components of* $\nabla\mathbf{f}_N(X_n)$ *are already expressed in terms of the basis* $\mathbf{v}_i$*.*

The implementation used to generate Figure 2 can be found at `https://github.com/FelixBenning/pyGRF` *and we plan to publish a detailed explanation separately.*

3. A discussion of the dimensional scaling

Readers unfamiliar with the $\frac{1}{N}$ scaling of the covariance in (4) might hypothesize that this reduction of the variance simply collapses the random function $\mathbf{f}_N$ to the mean μ in the asymptotic limit. Since asymptotically deterministic behavior would then follow trivially, it is important to understand why this is not the case.

To simplify the argument consider a stationary isotropic GRF, i.e. $\mathbf{f}_N \sim \mathcal{N}(\mu, C)$, the arguments work similarly for any (non-stationary) isotropic GRF. For any *fixed* parameter x we have $\mathbf{f}_N(x) \sim \mathcal{N}(\mu, \frac{1}{N}C(0))$, so the function value $\mathbf{f}_N(x)$ indeed collapses exponentially fast to the mean by a standard Chernoff-bound

$$\mathbb{P}\big(|\mathbf{f}_N(x) - \mu| \geq t\big) \leq 2\exp\big(-N\tfrac{t}{2C(0)}\big). \tag{9}$$

While the function value $\mathbf{f}_N(x)$ collapses to the mean, we will proceed to show that the gradient $\nabla\mathbf{f}_N(x)$ does not. And if the gradient does not collapse, it is sufficient to follow the gradient to find points where $\mathbf{f}_N$ stays away from μ. So the function $\mathbf{f}_N$ does not *uniformly* collapse to the mean.

First, let us sketch how the the derivatives of random functions are related to the derivatives of the covariance, see for instance Adler and Taylor (2007, Sec. 1.4.2) for a

formal derivation. Assuming the mean to be zero, without loss of generality, we move the derivatives outside of the integral to obtain

$$\mathsf{Cov}(\partial_{x_i}\mathbf{f}_N(x), \mathbf{f}_N(y)) = \partial_{x_i}\mathbb{E}[\mathbf{f}_N(x)\mathbf{f}_N(y)] = \partial_{x_i}\mathcal{C}_{\mathbf{f}_N}(x,y).$$

Iterating on this idea yields the covariance of derivatives

$$\mathsf{Cov}(\partial_{x_i}\mathbf{f}_N(x), \partial_{y_j}\mathbf{f}_N(y)) = \partial_{x_i}\partial_{y_j}\mathbb{E}[\mathbf{f}_N(x)\mathbf{f}_N(y)] = \partial_{x_i}\partial_{y_j}\mathcal{C}_{\mathbf{f}_N}(x,y). \tag{10}$$

In the case of a stationary isotropic covariance $\mathcal{C}_{\mathbf{f}_N}(x,y) = \frac{1}{N}C\big(\frac{\|x-y\|^2}{2}\big)$ this implies

$$\partial_{x_i}\partial_{y_j}\mathcal{C}_{\mathbf{f}_N}(x,y) = \tfrac{1}{N}\Big[\underbrace{C''\big(\tfrac{\|x-y\|^2}{2}\big)(x_i-y_i)(y_j-x_j)}_{\text{(I)}} - \underbrace{C'\big(\tfrac{\|x-y\|^2}{2}\big)\delta_{ij}}_{\text{(II)}}\Big],$$

where δ_{ij} is the Kronecker delta. Part (I) is zero if $x = y$. The Kronecker delta in part (II) then implies that the entries of $\nabla\mathbf{f}(x)$ are uncorrelated and therefore independent by the Gaussian assumption. Specifically, we have

$$\partial_{x_i}\mathbf{f}_N(x) \overset{\text{iid}}{\sim} \mathcal{N}\big(0, -\tfrac{1}{N}C'(0)\big).$$

The gradient norm therefore experiences a law of large numbers

$$\|\nabla\mathbf{f}_N(x)\|^2 = \sum_{i=1}^{N}(\partial_{x_i}\mathbf{f}_N(x))^2 \underset{N\to\infty}{\overset{p}{\to}} -C'(0). \tag{11}$$

Thus, while the function value $\mathbf{f}_N(x)$ collapses to the mean exponentially fast (9), the slope of the function (counterintuitively) does not. Any other type of scaling would result in vanishing or exploding gradients. Since the supremum over gradient norms is exactly the Lipschitz constant of $\mathbf{f}_N$, this scaling is therefore necessary to stay in a class of Lipschitz functions, as is often assumed in optimization theory. And the Parisi formula (Parisi, 1980; Talagrand, 2006; Panchenko, 2013) shows in the case of spin glasses, that this scaling also stabilizes the global maximum/minimum.

Remark 17 (Isoperimetry) *In view of the counterintuitive observations above, one might ask whether the assumption of isotropy results in very peculiar Lipschitz functions. To understand why that is not the case, consider the concept of isoperimetry (e.g. Bubeck and Sellke, 2021). A random variable X on $\mathbb{R}^N$ is said to satisfy c-isoperimetry, if for any L-Lipschitz function $f : \mathbb{R}^N \to \mathbb{R}$ and any $t \geq 0$ an exponential concentration bound holds*

$$\mathbb{P}\Big(|f(X) - \mathbb{E}[f(X)]| \geq t\Big) \leq 2\exp\big(-N\tfrac{t^2}{2cL}\big).$$

Observe that this concentration bound is the mirror image of (9)*. While we consider random functions, isoperimetry is concerned with random input X to deterministic L-Lipschitz functions. In particular Gaussian or uniform input satisfies isoperimetry (Bubeck and Sellke, 2021). Intuitively, this paints the following picture of very high dimensional Lipschitz functions: Most of the function is equal to the mean except for a few peculiar points. In*

the case of a deterministic functions this requires the exclusion of specific deterministic points. In the case of random functions this implies any deterministic point is allowed but not the use of gradient information. In either case the point we pick must be 'independent' of the function. And we show in Theorem 12 that if both function and input is random, independence is essentially a sufficient condition.

Finally, note that our definition of scaled sequences of random functions can be reframed in terms of a single random function defined on $\mathbb{R}^\infty$, which is externally scaled.

Remark 18 (External scaling) *One could define a a (non-stationary) isotropic GRF* $\mathbf{f}$ *with covariance*

$$\mathcal{C}_{\mathbf{f}}(x,y) = \kappa\big(\tfrac{\|x\|^2}{2}, \tfrac{\|y\|^2}{2}, \langle x,y\rangle\big), \qquad x,y \in \mathbb{R}^\infty$$

on the space of eventually-zero sequences $\mathbb{R}^\infty$*, which can also be viewed as the union of* $\mathbb{R}^N$

$$\mathbb{R}^\infty := \bigcup_{N\in\mathbb{N}} \mathbb{R}^N \quad \textit{with} \quad \mathbb{R}^N := \{(x_i)_{i\in\mathbb{N}} \subset \mathbb{R} : x_i = 0 \quad \forall i > N\}.$$

Since the scaling $\frac{1}{\sqrt{N}}$ *would scale away any mean the function* $\mathbf{f}$ *might possess, we assume* $\mathbf{f}$ *to be* centered *and define*

$$\mathbf{f}_N(x) := \mu\big(\tfrac{\|x\|^2}{2}\big) + \tfrac{1}{\sqrt{N}}\mathbf{f}(x), \quad x \in \mathbb{R}^N.$$

The distribution of $\mathbf{f}_N$ *is then equivalent to the one in Definition 6.*

4. Proof of Theorem 12

We now present the proofs of our main result. First, we provide a sketch to motivate the overall picture (Section 4.1). After we explain how the covariance of derivatives are determined (Section 4.2), we reforge Theorem 12 into an even more general Theorem 22. This theorem is more natural to prove but requires the definition of a special orthonormal coordinate system (Definition 21). Finally, we give the main proof of Theorem 22. Proofs for some instrumental results are deferred to the appendix.

4.1 Sketch of the proof of Theorem 12

To better guide the reader, we first sketch the idea behind the proof. For simplicity, we will assume the random function to be centered and stationary isotropic $\mathbf{f}_N \sim \mathcal{N}(0, C)$.

Idea 1: Induction over Gaussian Conditionals

We interpret the objects of interest $\mathbf{g} = (\mathbf{f}_N, \nabla\mathbf{f}_N)$ (the heights joined with the gradients, sometimes called the 'jet') as an auxiliary discrete-time stochastic process $(\mathbf{g}(X_n))_{n\in\mathbb{N}_0}$ and prove the claimed convergence in N by induction over the steps n. Please pretend for now that the inputs X_n were deterministic x_n. We will address this problem in Idea 3 of the sketch. The process $(\mathbf{g}(x_n))_{n\in\mathbb{N}_0}$ is then a Gaussian process and

$$\mathbf{g}(x_{[0:n]}) := (\mathbf{g}(x_0), \ldots, \mathbf{g}(x_n))$$

is therefore a multivariate Gaussian vector. Here $\mathbf{g}$ is applied entry-wise, i.e. $\mathbf{g}(x_I) = (\mathbf{g}(x_i))_{i\in I}$ with $x_I = (x_i)_{i\in I}$, and we use the following notation for discrete ranges:

$$[n{:}m] := [n,m] \cap \mathbb{Z}, \qquad [n{:}m) := [n,m) \cap \mathbb{Z}, \qquad \text{etc.} \tag{12}$$

The reader is encourage to remember them, as we will make use of them throughout the following sections. It is well known that the conditional distribution $\mathbf{g}(x_n) \mid \mathbf{g}(x_{[0:n)})$ is then also normal distributed (cf. Theorem 38). By the induction hypothesis $\mathbf{g}(x_{[0:n)})$ is already converging in probability to something deterministic, so it is perhaps natural to decompose $\mathbf{g}(x_n)$ into

$$\mathbf{g}(x_n) = \mathbb{E}[\mathbf{g}(x_n) \mid \mathbf{g}(x_{[0:n)})] + \sqrt{\mathsf{Cov}[\mathbf{g}(x_n) \mid \mathbf{g}(x_{[0:n)})]} \begin{pmatrix} Y_0 \\ \vdots \\ Y_N \end{pmatrix} \tag{13}$$

for independent standard normal distributed Y_i independent of $\mathbf{g}(x_{[0:n)})$. The conditional expectation is then given (cf. Theorem 38) by

$$\mathbb{E}[\mathbf{g}(x_n) \mid \mathbf{g}(x_{[0:n)})] = \mathsf{Cov}(\mathbf{g}(x_{[0:n)}), \mathbf{g}(x_n))^T [\mathsf{Cov}(\mathbf{g}(x_{[0:n)}))]^{-1} \mathbf{g}(x_{[0:n)})$$

and conditional variance by

$$\begin{aligned} &\mathsf{Cov}[\mathbf{g}(x_n) \mid \mathbf{g}(x_{[0:n)})] \\ &= \mathsf{Cov}[\mathbf{g}(x_n)] - \mathsf{Cov}(\mathbf{g}(x_{[0:n)}), \mathbf{g}(x_n))^T [\mathsf{Cov}(\mathbf{g}(x_{[0:n)}))]^{-1} \mathsf{Cov}(\mathbf{g}(x_{[0:n)}), \mathbf{g}(x_n)). \end{aligned}$$

The vector $\mathbf{g}(x_{[0:n)})$ already converges by induction as the dimension N increases. What is therefore left to study are the covariance matrices.

Idea 2: Finding sparsity in the covariance matrices with a custom coordinate system

The covariance matrices which make up the conditional expectation and variance are increasing in size as the dimension N grows, because the gradient $\nabla \mathbf{f}_N$ contained in $\mathbf{g}$ increases in size. Additionally we somehow need to group the independent Y_i into something which converges. It turns out that the standard coordinate system is ill suited to achieve these goals. To get sufficiently sparse (and therefore manageable) matrices we will now phrase everything in terms of directional derivatives.

In Lemma 19 we show how the covariances of the directional derivatives of an isotropic $\mathbf{f}_N \sim \mathcal{N}(0, C)$ can be calculated explicitly. In particular we have for vectors $v, w \in \mathbb{R}^N$ and points $x, y \in \mathbb{R}^N$ with distance $\Delta = x - y$

$$\mathsf{Cov}(D_v \mathbf{f}_N(x), D_w \mathbf{f}_N(y)) = -\tfrac{1}{N}\Big[\underbrace{C''\big(\tfrac{\|\Delta\|^2}{2}\big)\langle \Delta, w\rangle\langle \Delta, v\rangle}_{\text{(I)}} + \underbrace{C'\big(\tfrac{\|\Delta\|^2}{2}\big)\langle w, v\rangle}_{\text{(II)}}\Big]. \tag{14}$$

We begin by explaining how an orthogonal basis keeps (II) sparse before we explain where the use of the standard basis breaks down for (I). For simplicity, let us forget the height $\mathbf{f}_N$ for now and pretend $\mathbf{g}$ consists only of the gradient, i.e. $\mathbf{g} = \nabla \mathbf{f}_N$. Consider the covariance matrix at a fixed point $\nabla \mathbf{f}_N(x_0)$ only. Since we only consider the derivatives at a single

point we always have $x = y = x_0$ so the distance Δ vanishes. This completely removes the first part (I) (which will later require a special basis). The second part (II) is zero if v and w are orthogonal. Assuming we use an orthonormal coordinate system (e.g. the standard basis), we therefore get

$$\mathsf{Cov}[\mathbf{g}(x_0)] = \mathsf{Cov}[\nabla \mathbf{f}_N(x_0)] = \frac{1}{N}\begin{pmatrix} -C'(0) & & \\ & \ddots & \\ & & -C'(0) \end{pmatrix} = \tfrac{-C'(0)}{N}\mathbb{I}.$$

As $\mathbf{g}(x_{[0:0)})$ is the empty set we therefore have in the first step of the induction in n (cf. Equation (13))

$$\begin{aligned} \mathbf{g}(x_0) &= \mathbb{E}[\mathbf{g}(x_0) \mid \mathbf{g}(x_{[0:0)})] + \sqrt{\mathsf{Cov}[\mathbf{g}(x_0) \mid \mathbf{g}(x_{[0:0)})]} \begin{pmatrix} Y_0 \\ \vdots \\ Y_N \end{pmatrix} \\ &= \underbrace{\mathbb{E}[\mathbf{g}(x_0)]}_{=0} + \sqrt{\mathsf{Cov}[\mathbf{g}(x_0)]} \begin{pmatrix} Y_0 \\ \vdots \\ Y_N \end{pmatrix} \\ &= \sqrt{\tfrac{-C'(0)}{N}} \begin{pmatrix} Y_0 \\ \vdots \\ Y_N \end{pmatrix}. \end{aligned} \tag{15}$$

In particular, we have by the law of large numbers,

$$\langle \nabla \mathbf{f}_N(x_0), \nabla \mathbf{f}_N(x_0) \rangle = \frac{-C'(0)}{N} \sum_{i=1}^{N} Y_i^2 \underset{N\to\infty}{\overset{p}{\to}} -C'(0).$$

For the induction start and (II) there is therefore no need to change the coordinate system. But for the induction step with $n > 1$ the situation in (I) becomes more delicate.

We now want to understand where the issues in (I) with the standard coordinate system come form. Recall that we also want to determine limiting values of $\langle \nabla \mathbf{f}_N(x_0), \nabla \mathbf{f}_N(x_1) \rangle$.

Problem: Since x_1 will be x_0 plus a gradient step, $\Delta = x_1 - x_0$ is therefore a multiple of the gradient. If we continue to use the standard coordinate system, (I) causes the covariance matrix to be dense. This is because the entries of Δ are proportional to $\partial_i \mathbf{f}_N(x_0)$, which are multiples of the Y_i and therefore almost surely never zero.

Solution: We use an adapted coordinate system. For this we select $\tilde{\mathbf{v}}_0 := \nabla \mathbf{f}_N(x_0)$ and normalize this vector $\mathbf{v}_0 = \frac{\tilde{\mathbf{v}}_0}{\|\tilde{\mathbf{v}}_0\|}$. Again, we are going to pretend that this vector was deterministic and mark this fact with the notation v_0. We then extend it by $w_{[1:N)}$ to an orthonormal basis. Since the w_i are orthogonal to the gradient span, they are orthogonal to the distances $x_0 - x_1$, i.e. $\langle \Delta, w_i \rangle = 0$ and therefore (I) is zero for almost all directions (except for v_0). As the coordinate system is orthonormal, (II) is sparse again. More generally, we chose a coordinate system $v_{[0:n)}$ capturing the span of the first n gradients $\nabla \mathbf{f}_N(x_{[0:n)})$ and therefore also the span of the first n parameters and extend this coordinate system by $w_{[n:N)}$

to an orthonormal coordinate system such that we have essentially

$$\mathbf{g}(x_{[0:n)}) = (\nabla \mathbf{f}_N(x_0), \dots, \nabla \mathbf{f}_N(x_{n-1})) = U \begin{pmatrix} D_{v_0}\mathbf{f}_N(x_0) & \cdots & D_{v_0}\mathbf{f}_N(x_{n-1}) \\ \vdots & & \vdots \\ D_{v_{n-1}}\mathbf{f}_N(x_0) & \cdots & D_{v_{n-1}}\mathbf{f}_N(x_{n-1}) \\ D_{w_n}\mathbf{f}_N(x_0) & \cdots & D_{w_n}\mathbf{f}_N(x_{n-1}) \\ \vdots & & \vdots \\ D_{w_{N-1}}\mathbf{f}_N(x_0) & \cdots & D_{w_{N-1}}\mathbf{f}_N(x_{n-1}) \end{pmatrix} U^T.$$

Here U is a basis change matrix which we are going to suppress for simplicity in the following. Note that the covariance matrix is dense for the vectors $v_{[0:n)}$ because they span the subspace of the evaluation points x_k and therefore cause (I) to be non-zero. But (I) remains zero for the vectors w_i and the covariance matrix is therefore sparse there. We want to group the chaos together (i.e. group the directional derivatives of $v_{[0:n)}$), so we view $\mathbf{g}(x_{[0:n)})$ as a row-major matrix, i.e. whenever we treat $\mathbf{g}(x_{[0:n)})$ as a vector we concatenate the rows. It then turns out that the covariance matrix is of the form

$$\mathsf{Cov}[\mathbf{g}(x_{[0:n)})] = \begin{pmatrix} \boxed{\mathsf{Cov}[D_{v_{[0:n)}}\mathbf{f}_N(x_{[0:n)})]} & & & \\ & \boxed{\Sigma_n} & & \\ & & \boxed{} & \\ & & & \ddots \\ & & & \boxed{\Sigma_{N-1}} \end{pmatrix},$$

where $\Sigma_i = \mathsf{Cov}[D_{w_i}\mathbf{f}_N(x_{[0:n)})]$ are $n \times n$ matrices and $\mathsf{Cov}[D_{[0:n)}\mathbf{f}_N(x_{[0:n)})]$ is a dense $n^2 \times n^2$ block matrix. We can assume $N - 1 > n$ without loss of generality as we let $N \to \infty$. In the proof we will make the observation that all the Σ_i are in fact equal. Note that the mixed covariance matrix $\mathsf{Cov}(\mathbf{g}(x_{[0:n)}), \mathbf{g}(x_n))$ used in the conditional expectation and covariance have similar block form and $\mathsf{Cov}[\mathbf{g}(x_n)]$ is always diagonal. The key is now to understand that as N grows to infinity, the scaling $\frac{1}{N}$ keeps the $n^2 \times n^2$ chaos in check while we get a similar law of large numbers as in the first step (cf. (15)) by grouping the identical Σ_i to get the length of the new gradient $\nabla \mathbf{f}_N(x_n)$ projected to the subspace of $w_{[n:N)}$ which is the orthogonal space of $v_{[0:n)}$. This projection makes up the new coordinate v_n and the induction continues.

Idea 3: Treating the random input with care

At the beginning of the sketch we asked the reader to ignore the fact that the evaluation points $(X_n)_{n \in \mathbb{N}_0}$ are random themselves. Moreover, we ignored that the adapted orthogonal basis $\mathbf{v}_{[0:n)}$ was constructed from random gradients $\nabla \mathbf{f}_N(X_k)$ which implies it is random as well. Instead we denoted it by $v_{[0:n)}$ and pretended it was deterministic. In the following, we sketch how to get rid of the randomness of X_n and the randomness of the coordinate system $\mathbf{v}_{[0:n)}$. It is perhaps most obvious for the random coordinate system, that

$$D_{\mathbf{v}_0}\mathbf{f}_N(x_0) = \langle \mathbf{v}_0, \nabla \mathbf{f}_N(x_0) \rangle = \|\nabla \mathbf{f}_N(x_0)\|^2$$

is certainly not Gaussian by construction (recall $\mathbf{v}_0 = \nabla \mathbf{f}_N(x_0)$). Similarly, the random points X_n will typically change the distribution of $\mathbf{f}_N(X_n)$ and taking (conditional) expectations and covariances becomes quite delicate.

Intuitively, the key to solve this issue is to condition on all the previously seen information (e.g. gradients), such that the evaluation location do become 'deterministic'. While this is already common practice in Bayesian optimization, measure theoretically this is very delicate. But for the canonical conditional Gaussian distribution (Theorem 38), this has recently been proven to work (Benning, 2026, Corollary 2.12). Essentially, if $(X_n)_{n\in\mathbb{N}_0}$ is previsible, i.e. X_{n+1} is measurable with regard to the filtration $\mathcal{F}_n = \sigma(\mathbf{g}(X_k) : k \leq n)$ the following is true

$$\mathbb{E}[h(\mathbf{g}(X_n)) \mid \mathcal{F}_{n-1}] = \Big((x_{[0:n]}) \mapsto \mathbb{E}[h(\mathbf{g}(x_n)) \mid \mathbf{g}(x_{[0:n)})]\Big)(X_{[0:n]}).$$

Applying this lemma to the evaluations points X_n is relatively straightforward. But it is not obvious how to treat the random coordinate system. The key is to turn the coordinate system into an input. The definition

$$Z(v; x) := D_v \mathbf{f}_N(x)$$

allows us to apply Corollary 2.12 from Benning (2026) to both coordinate system and points. The delicate part is that the input must be previsible with respect to the filtration $\mathcal{F}_n$. This implies the coordinate system must also be previsible! We therefore cannot use future gradients for our coordinate system. This forces us to define a custom coordinate system for every step. That means after n steps we have n different coordinate systems. We will carefully reduce this to just one coordinate system per iteration in the formal proof. Turning this strategy into a rigorous proof is a bit tedious and requires a careful induction including a few additional technical induction hypothesis.

4.2 Covariances of derivatives of (smooth) random functions

It should be clear from the sketch given in Section 4.1 that we will have to deal with covariances of derivatives of the random functions under consideration. In this section we use known results to calculate the covariance of derivatives. Swapping integration and differentiation, we have for a centered random function $\mathbf{f}_N$

$$\begin{aligned}\mathsf{Cov}(\partial_{x_i}\mathbf{f}_N(x), \mathbf{f}_N(y)) &= \mathbb{E}[\partial_{x_i}\mathbf{f}_N(x)\mathbf{f}_N(y)] = \partial_{x_i}\mathbb{E}[\mathbf{f}_N(x)\mathbf{f}_N(y)] \\ &= \partial_{x_i}\mathcal{C}_{\mathbf{f}_N}(x, y)\end{aligned}$$

so the covariance of a derivative of $\mathbf{f}_N$ with $\mathbf{f}_N$ is equal to a partial derivative of the covariance function. For the formal details consult Scheurer (2009, chapter 5), Gihman and Skorokhod (1974, Ch. IV, §3, Def. 3 and following) or Adler and Taylor (2007, Sec. 1.4.2). Similarly, other covariances can be calculated, e.g.

$$\mathsf{Cov}(\partial_{x_i}\mathbf{f}_N(x), \partial_{y_i}\mathbf{f}_N(y)) = \partial_{x_i}\partial_{y_i}\mathcal{C}_{\mathbf{f}_N}(x, y),$$

and similarly the expectation of the derivative of uncentered random function can be calculated $\mathbb{E}[\partial_{x_i}\mathbf{f}_N(x)] = \partial_{x_i}\mathbb{E}[\mathbf{f}_N(x)]$. For this reason the derivatives of the covariance function

are interesting as they represent the covariance of derivatives. Due to bilinearity of the covariance it is also possible to move directional derivatives out of the covariance.

As a warm-up, let us calculate the covariance of derivatives for isotropic random functions $\mathbf{f}_N$, which have covariance functions of the form $\mathcal{C}_{\mathbf{f}_N}(x,y) = \frac{1}{N}C\big(\frac{\|x-y\|^2}{2}\big)$.

Lemma 19 (Covariance of derivatives, stationary isotropy) *Let $\mathbf{f}_N \sim \mathcal{N}(\mu, C)$ be an isotropic random function and let $\Delta = x - y$, then*

Cov	$\mathbf{f}_N(y)$	$D_w\mathbf{f}_N(y)$
$\mathbf{f}_N(x)$	$\frac{1}{N}C\big(\frac{\|\Delta\|^2}{2}\big)$	$-\frac{1}{N}C'\big(\frac{\|\Delta\|^2}{2}\big)\langle\Delta, w\rangle$
$D_v\mathbf{f}_N(x)$	$\frac{1}{N}C'\big(\frac{\|\Delta\|^2}{2}\big)\langle\Delta, v\rangle$	$-\frac{1}{N}\Big[\underbrace{C''\big(\frac{\|\Delta\|^2}{2}\big)\langle\Delta, w\rangle\langle\Delta, v\rangle}_{(I)} + \underbrace{C'\big(\frac{\|\Delta\|^2}{2}\big)\langle w, v\rangle}_{(II)}\Big]$

Proof Straightforward application of chain and product rules with the considerations above. ■

Recall that we split the covariance of derivatives up into two parts (I) and (II) in the proof sketch. We will now outline how the same can be done for (non-stationary) isotropic kernels. Since the (non-stationary) isotropic kernel κ have multiple inputs, we further define the notation

$$\kappa_i(\lambda_1, \lambda_2, \lambda_3) := \tfrac{d}{d\lambda_i}\kappa(\lambda_1, \lambda_2, \lambda_3)$$

Lemma 20 (Covariance of derivatives, non-stationary isotropy) *Let $\mathbf{f}_N \sim \mathcal{N}(\mu, \kappa)$ be (non-stationary) isotropic (Definition 6). Then the expectation of a directional derivative is given by*

$$\mathbb{E}[D_v\mathbf{f}_N(x)] = \mu'\big(\tfrac{\|x\|^2}{2}\big)\langle x, v\rangle, \tag{16}$$

whereas, using the notation $\kappa := \kappa\big(\frac{\|x\|^2}{2}, \frac{\|y\|^2}{2}, \langle x, y\rangle\big)$ to omit inputs, the covariance of directional derivatives is given by

$$\mathsf{Cov}(D_v\mathbf{f}_N(x), \mathbf{f}_N(y)) = \frac{1}{N}\Big[\kappa_1\langle x, v\rangle + \kappa_3\langle y, v\rangle\Big] \tag{17}$$

$$\mathsf{Cov}(D_v\mathbf{f}_N(x), D_w\mathbf{f}_N(y)) = \frac{1}{N}\Big[\underbrace{\kappa_{12}\langle x, v\rangle\langle y, w\rangle + \kappa_{13}\langle x, v\rangle\langle x, w\rangle + \kappa_{32}\langle y, v\rangle\langle y, w\rangle + \kappa_{33}\langle y, v\rangle\langle y, w\rangle}_{(I)} + \underbrace{\kappa_3\langle v, w\rangle}_{(II)}\Big]. \tag{18}$$

In particular, if the directions v, w are orthogonal to the points x, y then (17) *and (I) of* (18) *are zero. (II) in turn is zero, if the directions v, w are orthogonal.*

Proof Straightforward application of chain and product rules. ■

4.3 Proof of Theorem 12

After the preparations of the previous sections we can now give the details of the proof sketched in Section 4.1. In the sketch of the proof, we highlighted the importance of an adapted coordinate system in order to keep the covariance matrices sparse. In particular, part (I) of (14) requires orthogonality to the walking directions. We also needed the coordinate system to be orthogonal to keep (II) as sparse as possible. In Idea 3, we sketched why the coordinate system had to be previsible. That is, why we needed a different coordinate system in every step. We therefore begin the proof by building these coordinate systems.

Let us start by defining the natural filtration

$$\mathcal{F}_n := \sigma\big(\mathbf{f}_N(X_k), \nabla\mathbf{f}_N(X_k) : 0 \leq k \leq n\big).$$

The following previsible *vector space of evaluation points*

$$V_n := \operatorname{span}\Big(\{x_0\} \cup \{\nabla\mathbf{f}_N(X_k) : 0 \leq k < n\}\Big), \qquad d_n := \dim(V_n), \tag{19}$$

contains $X_{[0:n]}$ by definition of the generalized gradient span algorithm (Definition 1) and is similarly measurable with respect to $\mathcal{F}_{n-1}$, i.e. previsible. Also recall that $[0\!:\!n]$ is notation for a discrete ranges

$$[n\!:\!m] := [n,m] \cap \mathbb{Z}, \qquad [n\!:\!m) := [n,m) \cap \mathbb{Z}, \qquad \text{etc.} \tag{20}$$

We will now use this chain of vector spaces to define previsible coordinate systems for every step n.

Definition 21 (Previsible orthonormal coordinate systems) *We inductively define an $\mathcal{F}_{n-1}$-measurable orthonormal basis $\mathbf{v}_{[0:d_n)}$ of V_n such that $\mathbf{v}_{[0:d_k)}$ is a basis of the space V_k.*

Induction start *Assuming $x_0 \neq 0$ and thus $d_0 = 1$, we define*

$$\mathbf{v}_0 := \frac{x_0}{\|x_0\|}.$$

In any case this results in a basis $\mathbf{v}_{[0:d_0)}$ of V_0 since $[0:d_0) = \emptyset$ if $d_0 = 0$.

Induction step (Gram-Schmidt) *Assuming we have a basis $\mathbf{v}_{[0:d_n)}$ of V_n, let us construct a basis for V_{n+1}. For this we define the following Gram-Schmidt procedure candidate*

$$\tilde{\mathbf{v}}_n := \nabla\mathbf{f}_N(X_n) - P_{V_n}\nabla\mathbf{f}_N(X_n) = \nabla\mathbf{f}_N(X_n) - \sum_{i<d_n} \langle \nabla\mathbf{f}_N(X_n), \mathbf{v}_i\rangle \mathbf{v}_i, \tag{21}$$

where P_{V_n} is the projection to V_n. If $\tilde{\mathbf{v}}_n = 0$, then $\nabla\mathbf{f}_N(X_n) \in V_n$ and we thus have $V_{n+1} = V_n$ by definition (19). In that case we already have a basis for V_{n+1}.

For any n where the dimension increases such that $d_{n+1} = d_n + 1$, we define

$$\mathbf{v}_{d_{n+1}-1} = \mathbf{v}_{d_n} := \frac{\tilde{\mathbf{v}}_n}{\|\tilde{\mathbf{v}}_n\|}. \tag{22}$$

We thus have obtained an orthonormal basis $\mathbf{v}_{[0:d_{n+1})}$ of V_{n+1} which is $\mathcal{F}_n$ measurable.

Basis extensions *With this construction of basis elements of V_n done, we $\mathcal{F}_{n-1}$-measurably select an arbitrary orthonormal basis $\mathbf{w}^{(n)}_{[d_n:N)}$ of $V_n^\perp$ to obtain an $\mathcal{F}_{n-1}$-measurable orthonormal basis*

$$B_n := (\mathbf{v}_{[0:d_n)}, \mathbf{w}^{(n)}_{[d_n:N)})$$

for every $n \in \mathbb{N}$. The coordinate systems B_n are thus previsible.

Using this specialized coordinate system we state an extension of Theorem 12. This extension looks less friendly but is more natural to prove. It implies Theorem 12, proves three additional claims and relaxes the assumptions on the prefactors of the gradient span algorithm. The additional claims cannot be stated separately as they are all proved in one laborious induction. And to prove the claim we are most interested in, i.e. (Ind-I), we also require the other claims in the induction step. Finally, with the help of Proposition 24, we can (and will) assume deterministic starting points in Theorem 22 without loss of generality.

Theorem 22 (Predictable progress [Extension of Theorem 12]) *Let κ be a kernel valid in all dimensions (Definition 8) and $\mathbf{f}_N \sim \mathcal{N}(\mu, \kappa)$ be a sequence of scaled (non-stationary) isotropic Gaussian random functions (Definition 6) in N. Assume that μ and κ are sufficiently smooth (Assumption 11) and assume the covariance of $(\mathbf{f}_N, \nabla \mathbf{f}_N)$ is strictly positive definite. Let $\mathfrak{G}$ be a general gradient span algorithm (Definition 1), which is* asymptotically *continuous and uses the most recent gradient* asymptotically *(cf. Assumption 23).*

Let $\mathfrak{G}$ be applied to $\mathbf{f}_N$ with starting points $x_0 \in \mathbb{R}^N$ such that $X_n = \mathfrak{G}(\mathbf{f}_N, x_0, n)$. We assume that the deterministic initialization point $x_0 \in \mathbb{R}^N$ is of constant length $\|x_0\| = \lambda$ over the dimension N. Using $G_n := \big(\nabla \mathbf{f}_N(X_k) : k \leq n\big)$, we further define the modified random information vector

$$\hat{\mathbf{I}}_n := \Big(\mathbf{f}_N(X_k) : k \leq n\Big) \cup \Big(\langle v, w\rangle : v \in \mathbf{v}_{[0:d_{n+1})}, w \in G_n\Big).$$

The following inductive claims hold for all $n \in \mathbb{N}$, where (Ind-II) implies that the vector $\hat{\mathbf{I}}_n$ is almost surely of constant length.

(Ind-I) ***Information convergence:*** *There exists some deterministic limiting information vector $\hat{\mathscr{I}}_n = \hat{\mathscr{I}}_n(\mathfrak{G}, \mu, \kappa)$, such that*

$$\hat{\mathbf{I}}_n \underset{N\to\infty}{\overset{p}{\to}} \hat{\mathscr{I}}_n.$$

The limiting information is split into $\hat{\mathscr{I}}_n = (\mathfrak{f}_k, \gamma_k^{(i)})_{k,i}$, where the limiting elements $\mathfrak{f}_k$ of $\mathbf{f}_N(X_k)$ were already defined in Theorem 12. We further define the limiting inner products of gradients by

$$\gamma_k^{(i)} := \lim_{N\to\infty} \langle \nabla \mathbf{f}_N(X_k), \mathbf{v}_i\rangle. \tag{23}$$

(Ind-II) ***(Asymptotic) Full rank****: If the most recent gradient is always used (and not just asymptotically), then the previsible vector space of evaluation points V_{n+1} has almost surely full rank (assuming $x_0 \neq 0$). Specifically, for all $m \leq n+1$ we have almost surely*

$$d_m = m + \mathbf{1}_{x_0 \neq 0}.$$

This always holds in the limit (even when the most recent gradient is only used asymptotically), which means that for all $k \leq n$ the Gram-Schmidt candidate $\tilde{\mathbf{v}}_k$ defined in (22) *is not zero in the limit*

$$\gamma_k^{(d_k)} = \lim_{N\to\infty} \langle \nabla \mathbf{f}_N(X_k), \mathbf{v}_{d_k} \rangle = \lim_{N\to\infty} \|\tilde{\mathbf{v}}_k\| \neq 0. \tag{24}$$

(Ind-III) ***Representation:*** *For all $k \leq m \leq n+1$ there exist limiting representation vectors*

$$y_k = y_k(\mathfrak{G}, \mu, \kappa) \quad \textit{with} \quad y_k = (y_k^{(0)}, \dots y^{(d_m-1)}) \in \mathbb{R}^{d_m}$$

of X_k, such that for all $i < d_m$ we have

$$\langle X_k, \mathbf{v}_i \rangle \underset{N\to\infty}{\overset{p}{\to}} y_k^{(i)} \quad \textit{and} \quad \|X_k\|^2 = \sum_{i=0}^{d_m-1} \langle X_k, \mathbf{v}_i \rangle^2 \underset{N\to\infty}{\overset{p}{\to}} \|y_k\|^2. \tag{25}$$

The asymptotic distances are then given for all $k, l \leq m$ by

$$\|X_k - X_l\|^2 = \sum_{i=0}^{d_m-1} \langle X_k - X_l, \mathbf{v}_i \rangle^2 \underset{N\to\infty}{\overset{p}{\to}} \rho_{kl}^2 := \|y_k - y_l\|^2. \tag{26}$$

(Ind-IV) ***The evaluation points are asymptotically different****, i.e.*

$$\rho_{kl}^2 > 0 \quad \textit{for all} \quad k, l \leq n+1 \quad \textit{with} \quad k \neq l.$$

Before we prove that Theorem 12 follows from Theorem 22 let us state the 'asymptotic continuity' and 'use of the most recent gradient' assumptions on the prefactors.

Assumption 23 (Assumptions on prefactors) *Recall, that we defined the prefactors h_n of a GSA to be functions of the information $\mathbf{I}_{n-1}$ (Definition 1). We assume that*

- *h_n is continuous in the point $\mathscr{I}_{n-1}$ for all n, and*
- *the most recent gradients are used at least in the asymptotic limit, i.e.*

$$\hat{h}_{n,n-1}^{(g)} := h_{n,n-1}^{(g)}(\mathscr{I}_{n-1}) \Big(= \lim_{N\to\infty} h_{n,n-1}^{(g)}(\mathbf{I}_{n-1}) \Big) \neq 0.$$

By Lemma 27 it is apparent, that we can equivalently assume the prefactors h_n to be functions in $\hat{\mathbf{I}}_{n-1}$ continuous in $\hat{\mathscr{I}}_{n-1}$, which use the most recent gradient in the asymptotic limit $\hat{h}_{n,n-1} = h_{n,n-1}(\hat{\mathscr{I}}_{n-1})$.

At first it might seem circular to make use of limiting elements in an assumption which is necessary to prove the limiting elements exist. But a closer inspection reveals that h_n is only used for the convergence of $\mathbf{I}_k$ to $\mathscr{I}_k$ for times $k \geq n$. We can therefore interleave this assumption on h_n with the inductive existence proof of $\mathscr{I}_{n-1}$.

The proof of Theorem 12 now follows readily from Theorem 22 via a couple of preliminary results:

Proposition 24 *We can assume without loss of generality that the independent initialization X_0 is a deterministic $x_0 \in \mathbb{R}^N$ of length $\|x_0\| = \lambda$ over all dimensions N.*

Stationary case: *Assume the random functions $\mathbf{f}_N$ are furthermore stationary isotropic and the algorithm x_0-agnostic (Definition 1). Then the limiting information is independent of λ and the random initialization X_0 does not need to satisfy $\|X_0\| = \lambda$ almost surely.*

The proof of this Proposition is accomplished via two lemmas:

Lemma 25 (The 'information' is invariant to linear isometries) *Let f be some function and $(x_0, \dots, x_n)$ evaluation points. We further define a change in coordinates*

$$g(y) := f \circ \phi \quad \textit{and} \quad y_k := \phi^{-1}(x_k) \quad \textit{for} \quad k \in \{0, \dots, n\}$$

via a linear isometry $\phi(x) = Ux$ for some orthonormal matrix U. Then the information

$$I_n := \Big(f(x_k) : k \leq n\Big) \cup \Big(\langle v, w\rangle : v, w \in (x_0) \cup G_n\Big) \quad \textit{with} \quad G_n := \big(\nabla f(x_k) : k \leq n\big)$$

is exactly equal to the information

$$J_n := \Big(g(y_k) : k \leq n\Big) \cup \Big(\langle v, w\rangle : v, w \in (y_0) \cup \tilde{G}_n\Big) \quad \textit{with} \quad \tilde{G}_n := \big(\nabla g(y_k) : k \leq n\big).$$

If ϕ is a general isometry of the form $\phi(x) = Ux + b$, then we still have equality for the the reduced information, i.e.

$$I_n^{\setminus x_0} = J_n^{\setminus y_0}.$$

Proof We have by definition

$$f(x_k) = f \circ \phi(\phi^{-1}(x_k)) = g(y_k).$$

Let us therefore turn to the inner products. Since $\phi(x) = Ux$ or $\phi(x) = Ux + b$, we have for the gradient

$$\nabla g(y) = U^T \nabla f(\phi(y)). \tag{27}$$

this implies all the inner products are equal

$$\langle \nabla g(y_k), \nabla g(y_l)\rangle = \langle U^T \nabla f(x_k), U^T \nabla f(x_l)\rangle \qquad = \langle \nabla f(x_k), \nabla f(x_l)\rangle.$$

In the linear isometry case, where $\phi(x) = Ux$, we have $y_n = U^T x_n$ and thus

$$\begin{aligned} \langle y_0, \nabla g(y_l)\rangle &= \langle U^T x_0, U^T \nabla f(x_l)\rangle &&= \langle x_0, \nabla f(x_l)\rangle \\ \langle y_0, y_0\rangle &= \langle U^T x_0, U^T x_0\rangle &&= \langle x_0, x_0\rangle. \end{aligned}$$

■

Lemma 26 (Linear isometry invariance of general gradient span algorithms) *Let $\mathfrak{G}$ be a general gradient span algorithm (Definition 1), let f be a function, x_0 a starting point. We define a change of basis*

$$g(y) := f \circ \phi \quad \textit{and} \quad y_0 := \phi^{-1}(x_0).$$

via a linear isometry $\phi(x) = Ux$ with orthonormal matrix U. Then the optimization paths

$$x_n := \mathfrak{G}(f, x_0, n) \quad \textit{and} \quad y_n := \mathfrak{G}(g, y_0, n)$$

are invariant, i.e. the simple basis change

$$y_n = \phi^{-1}(x_n)$$

is retained for all $n \in \mathbb{N}$. The same holds true for all isometries ϕ, if the algorithm is x_0-agnostic.

Proof We proceed by induction over n, where the induction start is obvious.

For the induction step $(n-1) \to n$ we use the induction claim $y_k = \phi^{-1}(x_k)$ for all $k \leq n-1$ to obtain that the information I_{n-1} is invariant (by Lemma 25). This implies by definition of the gradient span algorithm and (27)

$$y_n = h_n^{(x)} y_0 + \sum_{k=0}^{n-1} h_{n,k}^{(g)} \nabla g(y_k) = U^T \Big(h_n^{(x)} x_0 + \sum_{k=0}^{n-1} h_{n,k}^{(g)} \nabla f(x_k) \Big) = \phi^{-1}(x_n),$$

where the invariance of the information was used implicitly as the h_n are functions of I_{n-1}.

In the x_0-agnostic case, the induction step is almost the same except we have for isometries $\phi(x) = Ux + b$ that $y_k = \phi^{-1}(x_k) = U^T(x_k - b)$ and thus

$$y_n = y_0 + \sum_{k=0}^{n-1} h_{n,k}^{(g)} \nabla g(y_k) = U^T \Big(x_0 - b + \sum_{k=0}^{n-1} h_{n,k}^{(g)} \nabla f(x_k) \Big) = \phi^{-1}(x_n),$$

using the fact that $h_n^{(x)} = 1$. We similarly use that the reduced information $I_{n-1}^{\setminus x_0}$ is retained for the prefactors. ■

We are now ready to prove that we can assume without loss of generality that the initialization is deterministic.

Proof [Proof of Proposition 24] Since we have

$$\mathbb{P}\big(\hat{\mathbf{I}}_n \in A\big) = \mathbb{E}\Big[\mathbb{P}\big(\hat{\mathbf{I}}_n \in A \mid X_0\big)\Big],$$

it is sufficient to show that $\mathbb{P}(\hat{\mathbf{I}}_n \in A \mid X_0 = x_0)$ is only dependent on $\|x_0\| = \lambda$, which is constant over the distribution of X_0.

For any x_0, y_0 with $\|x_0\| = \|y_0\| = \lambda$, there exists a linear isometry ϕ such that $x_0 = \phi^{-1}(y_0)$.

Let $\mathbf{I}_n = \mathbf{I}_n(\mathbf{f}_N, y_0)$ be the information vector generated from running the gradient span algorithm on $\mathbf{f}_N$ with starting point y_0 for n steps. Since $\mathbf{f}_N \overset{(d)}{=} \mathbf{f}_N \circ \phi$ in distribution, due to (non-stationary) isotropy (cf. Definition 6), we have

$$\mathbf{I}_n(\mathbf{f}_N, y_0) \overset{(d)}{=} \mathbf{I}_n(\mathbf{f}_N \circ \phi, y_0) = \mathbf{I}_n(\mathbf{f}_N, x_0)$$

where we have used Lemma 25 and Lemma 26 for the last equation. With the note that $\hat{\mathbf{I}}_n$ is a deterministic map of $\mathbf{I}_n$ as it only requires Gram-Schmidt orthogonalization as outlined in Definition 21 we can conclude this proof

$$\begin{aligned}\mathbb{P}\Big(\hat{\mathbf{I}}_n(\mathbf{f}_N, X_0) \in A \mid X_0 = y_0\Big) &= \mathbb{P}\Big(\hat{\mathbf{I}}_n(\mathbf{f}_N, y_0) \in A\Big) \\ &= \mathbb{P}\Big(\hat{\mathbf{I}}_n(\mathbf{f}_N, x_0) \in A\Big) \\ &= \mathbb{P}\Big(\hat{\mathbf{I}}_n(\mathbf{f}_N, X_0) \in A \mid X_0 = x_0\Big).\end{aligned}$$

In the stationary case with x_0-agnostic algorithm, we can make use of arbitrary isometries ϕ. In particular we can chose $\phi(x) = x - x_0$ that maps x_0 to zero. With the same arguments as above (using the x_0-agnostic version of Lemma 25 and Lemma 26) we get

$$\mathbb{P}\Big(\hat{\mathbf{I}}_n(\mathbf{f}_N, X_0) \in A \mid X_0 = x_0\Big) = \mathbb{P}\Big(\hat{\mathbf{I}}_n(\mathbf{f}_N, X_0) \in A \mid X_0 = 0\Big).$$

The distribution is thus completely independent of x_0. In particular, this forces the limiting values to be independent of x_0 and therefore also independent of λ. ■

Lemma 27 *For any fixed x_0, $\mathbf{I}_n$ is a continuous function of $\hat{\mathbf{I}}_n$ for all $n \in \mathbb{N}$.*

Proof Let us first recall the definition of $\mathbf{I}_n$ and $\hat{\mathbf{I}}_n$. With $G_n := \big(\nabla \mathbf{f}_N(X_k) : k \leq n\big)$ we have in a direct comparison:

$$\begin{aligned}\mathbf{I}_n &:= \Big(\mathbf{f}_N(X_k) : k \leq n\Big) \cup \Big(\langle v, w\rangle : v, w \in (x_0) \cup G_n\Big) \\ \hat{\mathbf{I}}_n &:= \Big(\mathbf{f}_N(X_k) : k \leq n\Big) \cup \Big(\langle v, w\rangle : v \in \mathbf{v}_{[0:d_{n+1})}, w \in G_n\Big).\end{aligned}$$

As the identity is a continuous map, we simply map the function values $\mathbf{f}_N(X_k)$ from $\hat{\mathbf{I}}_n$ to itself in $\mathbf{I}_n$. We therefore only need to find a way to continuously construct the inner products of $\mathbf{I}_n$ from $\hat{\mathbf{I}}_n$. Since $\nabla \mathbf{f}_N(X_k)$ is contained in V_{n+1} by its definition (19) and $\mathbf{v}_{[0:d_{n+1})}$ is a basis of V_{n+1} by construction (Definition 21), we have for all $k, l \leq n$

$$\begin{aligned}\langle \nabla \mathbf{f}_N(X_k), \nabla \mathbf{f}_N(X_l)\rangle &= \Big\langle \sum_{i=1}^{d_{n+1}-1} \langle \nabla \mathbf{f}_N(X_k), \mathbf{v}_i\rangle \mathbf{v}_i, \sum_{j=1}^{d_{n+1}-1} \langle \nabla \mathbf{f}_N(X_l), \mathbf{v}_j\rangle \mathbf{v}_j \Big\rangle \\ &= \sum_{i=0}^{d_{n+1}-1} \underbrace{\langle \nabla \mathbf{f}_N(X_k), \mathbf{v}_i\rangle}_{\in \hat{\mathbf{I}}_n} \underbrace{\langle \nabla \mathbf{f}_N(X_l), \mathbf{v}_i\rangle}_{\in \hat{\mathbf{I}}_n} \qquad (28)\end{aligned}$$

Since d_{n+1} is almost surely constant by (Ind-II)[3], this covers most of the inner products of $\mathbf{I}_n$. What is left are the inner products using x_0. If $x_0 = 0$, then all those inner products are zero and the zero map does the job.

3. (Ind-II) and (Ind-IV) have to be sacrificed if one wanted to get rid of the strict positive definiteness assumption in future work. This argument then has to be replaced using an upper bound on the d_{n+1} and Lemma 30.

In the following we therefore assume $x_0 \neq 0$. Because x_0 is deterministic, we do not need to construct $\|x_0\|^2 = \lambda^2$ from $\hat{\mathbf{I}}_n$ as a constant map does the job. Since $x_0 \neq 0$ implies $\mathbf{v}_0 = \frac{x_0}{\|x_0\|}$ by construction (Definition 21), we have for all $k \leq n$

$$\langle x_0, \nabla \mathbf{f}_N(X_k) \rangle = \|x_0\| \underbrace{\langle \mathbf{v}_0, \nabla \mathbf{f}_N(X_k) \rangle}_{\in \hat{\mathbf{I}}_n}.$$

We have thus continuously constructed all inner products in $\mathbf{I}_n$ form $\hat{\mathbf{I}}_n$. ■

Here is how the main theorem of the paper follows from Theorem 22.

Proof [Proof of Theorem 12] By Proposition 24, we can assume without loss of generality that the initial point is deterministic. Since the other assumptions of Theorem 22 are the same (or even more general), we only need to prove that for every $n \in \mathbb{N}$ there exists some $\mathscr{I}_n$ such that

$$\mathbf{I}_n \underset{N \to \infty}{\overset{p}{\to}} \mathscr{I}_n.$$

As $\mathbf{I}_n$ is a continuous function of $\hat{\mathbf{I}}_n$ by Lemma 27 this follows immediately from continuous mapping by the inductive claim (Ind-I) of Theorem 22. ■

The rest of the section is an inductive proof of Theorem 22.

4.4 Proof of Theorem 22

The heart of the proof will be a lengthy induction. Before, we want to address the additional assumptions of

1. representable limit points (Ind-III) and
2. the claim that these points are different (Ind-IV),

which we introduced in Theorem 22 but did not use in the proof of Theorem 12. Together with the strictly positive definite covariance of $(\mathbf{f}_N, \nabla \mathbf{f}_N)$ these assumptions will allow us to argue for converging entries of covariance matrices and invertible limiting covariance matrices, which are used in the conditional expectation and conditional variance.

4.4.1 Complexity reduction

Before we start the induction, we will perform some complexity reductions.

1. We show that (Ind-III) follows from (Ind-I) in Lemma 28.
2. We show that (Ind-IV) follows from (Ind-II) and (Ind-I) in Lemma 29.
3. We reduce the work necessary to prove (Ind-I).

In the actual induction we will therefore be able to focus on the claims (Ind-I) and (Ind-II).

Lemma 28 *For all fixed $n \in \mathbb{N}$, the convergence of information (Ind-I) implies the representation (Ind-III).*

Proof Assuming (Ind-I) we have that $\hat{\mathbf{I}}_n \to \hat{\mathscr{I}}_n$. By Lemma 27 this also implies $\mathbf{I}_n \to \mathscr{I}_n$. Since the prefactors h_m are functions of $\mathbf{I}_{m-1}$ continuous in $\mathscr{I}_{m-1}$, this implies for all $m \leq n+1$ that the prefactors converge

$$h_m = h_m(\mathbf{I}_{m-1}) \underset{N\to\infty}{\overset{p}{\to}} h_m(\mathscr{I}_{m-1}) =: \hat{h}_m.$$

Recall that by (23) of (Ind-I) we have for all $k \leq n$ and all $i < d_{n+1}$

$$\langle \nabla \mathbf{f}_N(X_k), \mathbf{v}_i \rangle \underset{N\to\infty}{\overset{p}{\to}} \gamma_k^{(i)}$$

for some $\gamma_k^{(i)} \in \mathbb{R}$. For all $m \leq n+1$ we therefore have by definition of X_m (Definition 1)

$$\begin{aligned} \langle X_m, \mathbf{v}_i \rangle &= h_m^{(x)} \langle x_0, \mathbf{v}_i \rangle + \sum_{k=0}^{m-1} h_{m,k}^{(g)} \langle \nabla \mathbf{f}_N(X_k), \mathbf{v}_i \rangle \\ \underset{N\to\infty}{\overset{p}{\to}} \; y_m^{(i)} \quad &:= \hat{h}_m^{(x)} \|x_0\| \delta_{0i} + \sum_{k=0}^{m-1} \hat{h}_{m,k}^{(g)} \gamma_k^{(i)}, \end{aligned} \tag{29}$$

where δ_{ij} denotes the Kronecker delta. Since X_k is contained in the vector space of evaluation points V_m for all $k \leq m$ by definition (19), its norms converges

$$\|X_k\|^2 = \sum_{i=0}^{d_m-1} \langle X_k, \mathbf{v}_i \rangle^2 \underset{N\to\infty}{\overset{p}{\to}} \sum_{i=0}^{d_m-1} (y_k^{(i)})^2 = \|y_k\|^2,$$

and likewise their distances for all $k, l \leq m$

$$\|X_k - X_l\|^2 = \sum_{i=0}^{d_m-1} \langle X_k - X_l, \mathbf{v}_i \rangle^2 \underset{N\to\infty}{\overset{p}{\to}} \rho_{kl}^2 := \|y_k - y_l\|^2.$$

This proves the limiting representation (Ind-III). ■

In the following we will prove, assuming (Ind-I) and (Ind-II), that the limiting distances ρ_{kl} are greater zero, i.e. (Ind-IV). The main ingredients are (Ind-II) and the assumed asymptotic use of the last gradient.

Lemma 29 *For all fixed $n \in \mathbb{N}$, the convergence of information (Ind-I) together with the asymptotic full rank (Ind-II) imply asymptotically different evaluation points (Ind-IV).*

Proof We will proceed by induction over m, where we assume the claim to be shown for all $k, l \leq m \leq n+1$. The induction start $m = 0$ is trivial since a single point is always distinct. For the induction step we assume to have the statement for m, that is

$$\rho_{kl} > 0 \quad \text{for all} \quad k, l \leq m \quad \text{with} \quad k \neq l.$$

To show the statement for $m+1 \leq n+1$, we only need to check the distances to the point y_{m+1} as we have the others by induction.

To show that y_{m+1} is distinct from any y_k with $k \leq m$, we use the fact that the most recent gradient is used asymptotically by assumption of the Theorem 22 (cf. Assumption 23), i.e.

$$\hat{h}^{(g)}_{m+1,m} = h^{(g)}_{m+1,m}(\mathscr{I}_m) \neq 0. \tag{30}$$

The claim (Ind-II) ensures that V_{m+1} has a larger dimension than V_m, that is $d_{m+1} = d_m + 1$. The last basis element of V_{m+1} is thus given by $\mathbf{v}_{d_{m+1}-1} = \mathbf{v}_{d_m}$. By (21) this element is produced from the last gradient $\nabla \mathbf{f}_N(X_m)$, which cannot be used for X_k with $k \leq m$ but must be used for X_{m+1} asymptotically by (30). Therefore X_{m+1} asymptotically contains a component of $\mathbf{v}_{d_m}$, that no other X_k has, which translates to the inequality of the asymptotic representations y_{m+1} and y_k.

Formally, since $X_0, \dots X_m$ is contained in V_m spanned by $\mathbf{v}_{[0:d_m)}$, we have for all $k \leq m$

$$y^{(d_m)}_k \overset{(29)}{=} \lim_{N\to\infty} \langle X_k, \mathbf{v}_{d_m} \rangle = 0. \tag{31}$$

For the asymptotic representation of the last evaluation point X_{m+1} on the other hand, we have

$$y^{(d_m)}_{m+1} \overset{(29)}{=} \hat{h}^{(x)}_{m+1} \|x_0\| \underbrace{\delta_{0 d_m}}_{=0} + \sum_{k=0}^{m} \hat{h}^{(g)}_{m+1,k} \gamma^{(d_m)}_k = \hat{h}^{(g)}_{m+1,m} \gamma^{(d_m)}_m. \tag{32}$$

The last equation is due to (33). This follows from the fact that for all $k < m$ the gradient $\nabla \mathbf{f}_N(X_k)$ is contained in V_m spanned by $\mathbf{v}_{[0:d_m)}$. By definition of $\gamma^{(i)}_k$ (23) we therefore have

$$\gamma^{(d_m)}_k \overset{(23)}{=} \lim_{N\to\infty} \langle \mathbf{f}_N(X_k), \mathbf{v}_{d_m} \rangle = 0. \tag{33}$$

Putting (31) and (32) together we have

$$\rho^2_{(m+1)k} = \|y_{m+1} - y_k\|^2 \geq (y^{(d_m)}_{m+1} - y^{(d_m)}_k)^2 = (\hat{h}^{(g)}_{m+1,m} \gamma^{(d_m)}_m)^2 > 0,$$

where the last inequality is due to the asymptotic use of the most recent gradient (30) and the limiting full rank claim (24) of (Ind-II). ■

In (Ind-I) we claim convergence of $\hat{\mathbf{I}}_n$, where $\hat{\mathbf{I}}_n$ contains inner products of the form

$$\langle \mathbf{v}_i, \nabla \mathbf{f}_N(X_k) \rangle$$

for $k \leq n$ and $i < d_{n+1}$. The following lemma essentially implies that the restriction on i was unnecessary. So when we increase $n-1$ to n in the induction step, we do not have to revisit the inner products of old gradients and can focus solely on $\nabla \mathbf{f}_N(X_n)$.

Lemma 30 *For all $k \in \mathbb{N}$ and $i \geq d_{k+1}$ we have*

$$\langle \nabla \mathbf{f}_N(X_k), \mathbf{v}_i \rangle \underset{N\to\infty}{\overset{p}{\to}} 0 = \gamma^{(i)}_k.$$

Proof For all $i \geq d_{k+1}$ the basis vector $\mathbf{v}_i$ is constructed to be orthogonal to $\nabla \mathbf{f}_N(X_k)$ contained in V_{k+1} spanned by $\mathbf{v}_{[0:d_{k+1})}$. This implies

$$\lim_{N\to\infty} \langle \nabla \mathbf{f}_N(X_k), \mathbf{v}_i \rangle = 0 = \gamma_k^{(i)}$$

where the last equation is simply the definition of $\gamma_k^{(i)}$ of (23). ■

Remark 31 (Gamma are triangular) *Lemma 30 can be visualized using a triangular matrix as follows. With the definition*

$$\mathbb{R}^n := \{(x_i)_{i\in[0:\infty)} : x_i = 0 \quad \forall i \geq n\},$$

we can view $\mathbb{R}^m$ as a subspace of $\mathbb{R}^n$ for $m \leq n$. The vector

$$\gamma_k := (\gamma_k^{(i)})_{i\in[0:d_{k+1})} \in \mathbb{R}^{d_{k+1}},$$

is then (by Lemma 30) also a member of $\mathbb{R}^m$ for $m \geq d_{k+1}$. Concatenating the vectors

$$\gamma_{[0:n]} = \begin{pmatrix} \gamma_0^{(0)} & \dots & \gamma_n^{(0)} \\ \vdots & & \vdots \\ \gamma_0^{(d_{n+1}-1)} & \dots & \gamma_n^{(d_{n+1}-1)} \end{pmatrix} = \begin{pmatrix} \gamma_0^{(0)} & \dots & \gamma_{n-1}^{(0)} & \gamma_n^{(0)} \\ \gamma_0^{(1)} & \dots & \gamma_{n-1}^{(1)} & \gamma_n^{(1)} \\ 0 & & & \\ \vdots & \ddots & & \vdots \\ 0 & \dots & 0 & \gamma_n^{(d_{n+1}-1)} \end{pmatrix}$$

therefore results in a upper triangular matrix above an offset diagonal, since we always have $d_k \leq k+1$ due to the definition of V_k (19).

Note that, for visualization purposes, we assumed $d_{n+1} = d_n + 1$ in the second representation. If the dimension stays constant at some times k, then the zeros encroach above this diagonal.

4.4.2 Induction start with $n = 0$

We start the induction by proving the claim for $n = 0$. We have structured the induction start similar to the induction step. We therefore suggest the reader to familiarize themselves with this strategy here, as it is easier to get lost in the details of the induction step.

By Lemma 28 and Lemma 29 we only need to prove (Ind-I) and (Ind-II). For (Ind-I) we need to prove that

$$\hat{\mathbf{I}}_0 = \big(\mathbf{f}_N(x_0)\big) \cup \big(\langle \nabla \mathbf{f}_N(x_0), v\rangle : v \in \mathbf{v}_{[0:d_1)}\big) = \big(\mathbf{f}_N(x_0), \langle \nabla \mathbf{f}_N(x_0), \mathbf{v}_{[0:d_1)}\rangle\big)$$

converges to a limiting $\hat{\mathscr{I}}_0 = (\mathfrak{f}_0, \gamma_0^{(0)}, \dots, \gamma_0^{(d_1-1)})$ in probability. Note that $d_1 \leq 2$ as the vector space of evaluation points is defined in (19) to be previsible, that is

$$V_1 = \operatorname{span}\{x_0, \nabla \mathbf{f}_N(x_0)\}.$$

So we have $d_1 - 1 \leq 1$ depending on whether the starting point x_0 is zero. The vector

$$\gamma_0 = (\gamma_0^{(i)})_{i<d_1} = (\gamma_0^{(0)}, \dots, \gamma_0^{(d_1-1)})$$

might therefore collapse to a single entry $\gamma_0^{(0)}$. The approach we will now take mirrors the approach in the induction step. Beyond the pedagogical benefit, this order is also quite natural for the induction start when using the notation we introduced.

Step 1 First we prove that

$$\big(\mathbf{f}_N(x_0)\big) \cup \big(\langle \nabla \mathbf{f}_N(x_0), v\rangle : v \in \mathbf{v}_{[0:d_0)}\big)$$

converges. We highlight that the limit d_0 is purposefully different from the range of $\hat{\mathbf{I}}_0$.

Step 2 We will then prove (Ind-II), which ensures the dimension actually increases

$$d_1 = d_0 + 1.$$

Step 3 Finally we prove that $\langle \nabla \mathbf{f}_N(x_0), \mathbf{v}_{d_0}\rangle = \langle \nabla \mathbf{f}_N(x_0), \mathbf{v}_{d_1-1}\rangle$ converges. This adds the missing index from the first step.

While we could prove (Ind-II) before (Ind-I) in the induction start, we will require the results of *Step 1* for (Ind-II) in the induction step. The element $\mathbf{v}_{d_0}$ is not only different from $\mathbf{v}_{[0:d_0)}$ in the sense that the dimension increase needs to be shown, it is also the first truly random basis element here. In the induction step the difference will be between the previsible basis elements and the last non-previsible element, which needs to be treated differently.

Step 1 Since our random function is (non-stationary) isotropic with $\mathbf{f}_N \sim \mathcal{N}(\mu, \kappa)$, we have

$$\mathbf{f}_N(x_0) \sim \mathcal{N}\Big(\mu\big(\tfrac{\|x_0\|^2}{2}\big), \frac{1}{N}\kappa\big(\tfrac{\|x_0\|^2}{2}, \tfrac{\|x_0\|^2}{2}, \|x_0\|^2\big)\Big).$$

this immediately implies convergence of the first component

$$\mathbf{f}_N(x_0) \underset{N\to\infty}{\overset{p}{\to}} \mu\big(\tfrac{\|x_0\|^2}{2}\big) = \mu\big(\tfrac{\lambda^2}{2}\big) =: \mathfrak{f}_0.$$

Let us now turn to the convergence of the inner products. We consider two cases.

Case ($x_0 \neq 0$): In this case we have $\mathbf{v}_0 = \frac{x_0}{\|x_0\|}$ by Definition 21. Therefore $\mathbf{v}_0$ is deterministic and by an application of Lemma 20 we have

$$D_{\mathbf{v}_0}\mathbf{f}_N(x_0) \sim \mathcal{N}\Big(\mu'\big(\tfrac{\|x_0\|^2}{2}\big)\|x_0\|, \frac{1}{N}\big[(\kappa_{12}+\kappa_{13}+\kappa_{32}+\kappa_{33})\|x_0\|^2 + \kappa_3\big]\Big),$$

using the notation

$$\kappa := \kappa\big(\tfrac{\|x_0\|^2}{2}, \tfrac{\|x_0\|^2}{2}, \|x_0\|^2\big) = \kappa\big(\tfrac{\lambda^2}{2}, \tfrac{\lambda^2}{2}, \lambda^2\big)$$

to omit inputs to the kernel κ. But this immediately implies

$$\langle \nabla \mathbf{f}_N(x_0), \mathbf{v}_0\rangle = D_{\mathbf{v}_0}\mathbf{f}_N(x_0) \underset{N\to\infty}{\overset{p}{\to}} \mu'\big(\tfrac{\|x_0\|^2}{2}\big)\|x_0\| =: \gamma_0^{(0)}.$$

As $d_0 = \dim(V_0) = \dim(\operatorname{span}(x_0)) = 1$, we have covered all elements in $\mathbf{v}_{[0:d_0)}$.

Case ($x_0 = 0$): In this case, $d_0 = \dim(V_0) = 0$ and $\mathbf{v}_{[0:d_0)}$ is empty. We therefore do not have to do anything.

Step 2 To prove the dimension increases, we consider the Gram-Schmidt candidate

$$\begin{aligned}\tilde{\mathbf{v}}_0 &:= \nabla \mathbf{f}_N(x_0) - P_{V_0}\mathbf{f}_N(x_0) = \nabla \mathbf{f}_N(x_0) - \sum_{i<d_0} \langle \nabla \mathbf{f}_N(X_0), \mathbf{v}_i\rangle \mathbf{v}_i \\ &= \sum_{i=d_0}^{N-1} \langle \nabla \mathbf{f}_N(X_0), \mathbf{w}_i^{(0)}\rangle \mathbf{w}_i^{(0)},\end{aligned}$$

where $\mathbf{w}^{(0)}_{[d_0:N)}$ is the basis defined to be orthogonal to V_0 (Definition 21). Since V_0 is deterministic, this orthogonal basis is also deterministic. In the induction step, both will only be previsible. Note that in the case $x_0 = 0$, $\tilde{\mathbf{v}}$ is simply the entire gradient by definition of d_0.

Since the $\mathbf{w}_i^{(0)}$ are deterministic, and orthogonal to x_0 in either case, we can apply Lemma 20 to obtain

$$D_{\mathbf{w}_i^{(0)}}\mathbf{f}_N(x_0) \overset{\text{iid}}{\sim} \mathcal{N}\Big(0, \frac{\kappa_3}{N}\Big).$$

This implies that there exist $Y_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ such that

$$D_{\mathbf{w}_i^{(0)}}\mathbf{f}_N(x_0) = \sqrt{\frac{\kappa_3}{N}} Y_i.$$

But since $d_0 \leq 1$ and in particular d_0 is finite, this implies with the law of large numbers

$$\|\tilde{\mathbf{v}}_0\|^2 = \sum_{i=d_0}^{N-1} \langle \nabla \mathbf{f}_N(x_0), w_i\rangle^2 = \kappa_3 \cdot \frac{1}{N} \sum_{i=d_0}^{N-1} Y_i^2 \underset{N\to\infty}{\overset{p}{\to}} \kappa_3. \tag{34}$$

Since we have $\kappa_3 > 0$ by Lemma 45, $\|\tilde{\mathbf{v}}_0\|^2$ is almost surely strictly greater than zero, which implies $\dim(V_1) > \dim(V_0)$ almost surely. Additionally, the dimension also increases in the limit, i.e. $\lim_{N\to\infty} \|\tilde{\mathbf{v}}_0\| = \kappa_3 > 0$. We have therefore shown all of (Ind-II).

Step 3 We have by definition of $\mathbf{v}_{d_0}$ in (22) and the definition of $\tilde{\mathbf{v}}_0$

$$\langle \nabla \mathbf{f}_N(x_0), \mathbf{v}_{d_0}\rangle = \Big\langle \nabla \mathbf{f}_N(x_0), \tfrac{\tilde{\mathbf{v}}_0}{\|\tilde{\mathbf{v}}_0\|}\Big\rangle = \|\tilde{\mathbf{v}}_0\| \underset{N\to\infty}{\overset{p}{\to}} \sqrt{\kappa_3} =: \gamma_0^{(d_0)}.$$

This finishes the last step and therefore the induction start.

4.4.3 Induction step $(n-1) \to n$

We now get to the main body of the proof. Before we start, let us recapitulate the lemmas we proved for complexity reduction. By Lemma 28 and Lemma 29 it is sufficient to prove the statements (Ind-I) and (Ind-II), as the others follow.

With $G_n := \big(\nabla \mathbf{f}_N(X_k) : k \leq n\big)$ the modified information $\hat{\mathbf{I}}_n$ of the information convergence (Ind-I) was given by

$$\hat{\mathbf{I}}_n := \Big(\mathbf{f}_N(X_k) : k \leq n\Big) \cup \Big(\langle v, w\rangle : v \in \mathbf{v}_{[0:d_{n+1})}, w \in G_n\Big).$$

By induction we already have $\hat{\mathbf{I}}_{n-1} \to \hat{\mathscr{I}}_{n-1}$. And due to our discussion of $\langle \mathbf{v}_i, \nabla \mathbf{f}_N(X_k)\rangle$ for $i \geq d_{n+1} \geq d_{k+1}$ in Lemma 30 we therefore only need to prove

$$\begin{pmatrix} \mathbf{f}_N(X_n) \\ \langle \mathbf{v}_{[0:d_{n+1})}, \nabla \mathbf{f}_N(X_n)\rangle \end{pmatrix} = \begin{pmatrix} \mathbf{f}_N(X_n) \\ D_{\mathbf{v}_0}\mathbf{f}_N(X_n) \\ \vdots \\ D_{\mathbf{v}_{d_{n+1}-1}}\mathbf{f}_N(X_n) \end{pmatrix} \underset{N\to\infty}{\overset{p}{\to}} \begin{pmatrix} \mathfrak{f}_n \\ \gamma_n \end{pmatrix}, \tag{35}$$

where we used the definition $\gamma_n = (\gamma_n^{(i)})_{i\in[0:d_{n+1})}$ from Remark 31. In this remark we have also explained how $\gamma_k \in \mathbb{R}^{d_{k+1}}$ can be viewed as a member of $\mathbb{R}^{d_{n+1}}$ and laid out the concatenated vectors $\gamma_{[0:n]}$ in triangular matrix form. In the following we will arrange the random information into a matching matrix form by the definition of an auxiliary function Z.

This function has multiple purposes: First, it turns the direction vectors of the directional derivatives into inputs for an application of the previsible sampling result (Benning, 2026, Corollary 2.12), which we explained in Idea 3 of the proof sketch (Section 4.1). Second, it structures the information vector $\hat{\mathbf{I}}_n$ into a better readable matrix form for human digestion. Third, it allows us to rearrange the elements of the column vectors we concatenate into row vectors. This effectively rearranges the covariance matrix of Z into a more sparse block form.

With some abuse of notation we define Z for a varying number of direction vectors $w_1, \ldots, w_m \in \mathbb{R}^N$ and evaluation points $x_1, \ldots, x_k \in \mathbb{R}^N$ for $m \leq N$ and $k \leq n$

$$Z(w_1, \ldots, w_m; x_0, \ldots, x_k) := \begin{pmatrix} \mathbf{f}_N(x_0) & \ldots & \mathbf{f}_N(x_k) \\ D_{w_1}\mathbf{f}_N(x_0) & \cdots & D_{w_1}\mathbf{f}_N(x_k) \\ \vdots & & \vdots \\ D_{w_m}\mathbf{f}_N(x_0) & \cdots & D_{w_m}\mathbf{f}_N(x_k) \end{pmatrix}.$$

The number of inputs is therefore not variable and we separate their type by a semicolon. Formally one would have to consider slices of Z, but since projections are measurable this is not an issue. Note that we are not interested in this matrix as an operator. The two dimensional layout is only used for better readability and we will generally view it as a vector. For this purpose we treat the matrix as 'row-major', i.e. whenever we treat it like a vector, we concatenate the rows. This grouping is purposefully different from the column layout of the γ_k as it will enable a block matrix sparsity in the covariance matrices later on.

By induction we have information convergence (Ind-I) for $n-1$, which can be expressed using Z as

$$Z(\mathbf{v}_{[0:d_n)}; X_{[0:n)}) = \begin{pmatrix} \mathbf{f}_N(X_0) & \cdots & \mathbf{f}_N(X_{n-1}) \\ D_{\mathbf{v}_0}\mathbf{f}_N(X_0) & \cdots & D_{\mathbf{v}_0}\mathbf{f}_N(X_{n-1}) \\ \vdots & & \vdots \\ D_{\mathbf{v}_{d_n-1}}\mathbf{f}_N(X_0) & \cdots & D_{\mathbf{v}_{d_n-1}}\mathbf{f}_N(X_{n-1}) \end{pmatrix} \underset{N\to\infty}{\overset{p}{\to}} \begin{pmatrix} \mathfrak{f}_{[0:n)} \\ \gamma_{[0:n)} \end{pmatrix}. \tag{36}$$

Observe that the induction step simply adds the additional column given in (35) to the right of the matrix, where we use Lemma 30 to argue that we can arbitrarily extend the number of rows.

We will now follow the same strategy as in the induction start:

Step 1 First, we prove the convergence of the 'new column'

$$Z(\mathbf{v}_{[0:d_n)}; X_n) = \begin{pmatrix} \mathbf{f}_N(X_n) \\ \langle \mathbf{v}_{[0:d_n)}, \nabla \mathbf{f}_N(X_n) \rangle \end{pmatrix} \underset{N\to\infty}{\overset{p}{\to}} \begin{pmatrix} \mathfrak{f}_n \\ \gamma_n^{([0:d_n))} \end{pmatrix}$$

Step 2 We prove (Ind-II), which ensures asymptotically

$$d_{n+1} = d_n + 1.$$

Step 3 We finally prove convergence of the 'new corner element'

$$Z(\mathbf{v}_{d_n}; X_n) = \langle \mathbf{v}_{d_n}, \nabla \mathbf{f}_N(X_n) \rangle \underset{N\to\infty}{\overset{p}{\to}} \gamma_n^{(d_n)}.$$

The reason for this split are twofold. First, we have that $\mathbf{v}_{[0:d_n)}$ is previsible, that is $\mathcal{F}_{n-1}$-measurable, while $\mathbf{v}_n$ is not. $\mathbf{v}_n$ must therefore be treated differently. Second, we need to construct $\mathbf{v}_{d_n}$ from $\tilde{\mathbf{v}}_n$, which will naturally prove (Ind-II) before we get to the convergence of the inner product in *Step 3*.

This strategy can get a bit lost, as we will spend the majority of our time with *Step 1* before wrapping up *Step 2* and *Step 3* fairly quickly. An additional reason for this strategy to get lost is, that we will not just consider $(\mathbf{f}_N(x_n), \langle \mathbf{v}_{[0:d_n)}, \nabla \mathbf{f}_N(X_n) \rangle)$, but instead consider the conditional distribution of $(\mathbf{f}_N(x_n), \langle B_n, \nabla \mathbf{f}_N(X_n) \rangle)$ for the entire previsible basis $B_n = (\mathbf{v}_{[0:d_n)}, \mathbf{w}^{(n)}_{[d_n:N)})$. Later, we will aggregate the directional derivatives in the directions $\mathbf{w}_i^n$ into $\tilde{\mathbf{v}}_n$, which means that we work towards *Step 1* and *Step 2* simultaneously. The first objective will be, to apply the previsible sampling result (Benning, 2026, Corollary 2.12) such that we can treat the previsible inputs as deterministic.

Getting rid of the random input In the proof sketch we outlined how we needed to treat the random input with care (cf. 4.1, Idea 3). In particular we need to consider the (basis; point) pairs

$$(B_0; X_0), \cdots, (B_n; X_n) := \Big(\mathbf{v}_{[0:d_0)}, \mathbf{w}^{(0)}_{[d_0:N)}; X_0\Big), \cdots, \Big(\mathbf{v}_{[0:d_n)}, \mathbf{w}^{(n)}_{[d_n:N)}; X_n\Big).$$

Recall that we defined the basis B_k in Definition 21 to be previsible just like X_k, i.e. measurable with regard $\mathcal{F}_{k-1}$. Moreover, since all basis changes are previsible and invertible it is straightforward to check inductively that there exists a measurable bijective map between different basis representations of the derivative information up to n

$$\big(\mathbf{f}_N(X_k), \nabla \mathbf{f}_N(X_k)\big)_{k<n} \leftrightsquigarrow \big(Z(B_k; X_k)\big)_{k<n}.$$

This implies that their generated sigma algebras are the same

$$\mathcal{F}_{n-1} \overset{\text{def.}}{=} \sigma\Big(\big(\mathbf{f}_N(X_k), \nabla \mathbf{f}_N(X_k)\big), k \leq n-1\Big) = \sigma\big(Z(B_k; X_k), k \leq n-1\big).$$

By Corollary 2.12 from Benning (2026) we then have, for the canonical conditional Gaussian distribution (Theorem 38) and all bounded measurable h,

$$\mathbb{E}\Big[h\big(Z(B_n; X_n)\big) \mid \mathcal{F}_{n-1}\Big] = F(B_{[0:n]}; X_{[0:n]}),$$

for the function F defined using the canonical conditional Gaussian distribution (Theorem 38) with deterministic evaluation points x_k and basis b_k as

$$F(b_{[0,n]}; x_{[0:n]}) := \mathbb{E}\Big[h\big(Z(b_n; x_n)\big) \mid Z(b_0; x_0), \dots, Z(b_{n-1}; x_{n-1})\Big].$$

The above is abuse of notation: We explicitly need the canonical conditional Gaussian distribution for Corollary 2.12 from Benning (2026) to be applicable. This notation for the conditional expectation is however only defined up to zero sets. As we have no interest in any other conditional distribution than the canonical one, we hope that this notation is more helpful than confusing.

In other words, Corollary 2.12 from Benning (2026) justifies treating the inputs $(B_k; X_k)$ as deterministic when calculating the conditional distribution. But keeping track of $n+1$ different basis B_k for every evaluation point X_k is very inconvenient. So our next goal is to reduce the number of basis to one. For this we optimistically define for a single basis b

$$G(b; x_{[0:n]}) := \mathbb{E}\Big[h\big(Z(b; x_n)\big) \mid Z(b; x_0), \dots, Z(b; x_{n-1})\Big].$$

We are now going to observe that F is constant in all but the most recent basis b_n, i.e.

$$F(b_{[0:n]}; x_{[0:n]}) = G(b_n; x_{[0:n]}). \tag{37}$$

That is because the sigma algebras generated by different basis representations are identical as they can be bijectively translated into each other, i.e. for any b

$$\sigma(Z(b_0; x_0), \dots, Z(b_{n-1}; x_{n-1})) = \sigma(Z(b; x_0), \dots, Z(b; x_{n-1})).$$

In particular this is true for $b = b_n$ and thus we have (37), i.e.

$$\begin{aligned}&\mathbb{E}\Big[h\big(Z(b_n; x_n)\big) \mid Z(b_0; x_0), \dots, Z(b_{n-1}; x_{n-1})\Big] \\ &= \mathbb{E}\Big[h\big(Z(b_n; x_n)\big) \mid Z(b_n; x_0), \dots, Z(b_n; x_{n-1})\Big].\end{aligned}$$

Since we only need to consider the most recent basis from now on, we drop the index and write $\mathbf{w}_k := \mathbf{w}_k^{(n)}$ and summarize our result as

$$\mathbb{E}\Big[h\big(Z(\mathbf{v}_{[0:d_n)}, \mathbf{w}_{[d_n:N)}; X_n)\big) \mid \mathcal{F}_{n-1}\Big] = G(\mathbf{v}_{[0:d_n)}, \mathbf{w}_{[d_n:N)}; X_0, \dots, X_n). \tag{38}$$

Recall that we use the semicolon to separate the basis elements from the evaluation points, which should be treated as concatenated vectors respectively.

In essence, by definition of G we can now treat our evaluation points $X_{[0:n]}$ and coordinate system $B_n = (\mathbf{v}_{[0:d_n)}, \mathbf{w}_{[d_n:N)})$ as deterministic when calculating the conditional distribution.

The conditional distribution is known in the Gaussian case Since Z is a Gaussian random function, we have for non-random input (b, x) that its conditional distribution is also Gaussian (cf. Theorem 38). Equation (38) translates this result to the random input and we can therefore conclude that

$$Z(B_n; X_n) \mid \mathcal{F}_{n-1} \sim \mathcal{N}\Big(\mathbb{E}[Z(B_n; X_n) \mid \mathcal{F}_{n-1}], \mathsf{Cov}[Z(B_n; X_n) \mid \mathcal{F}_{n-1}]\Big).$$

In particular there exist independent $Y_1, \dots, Y_N \sim \mathcal{N}(0,1)$ independent of $\mathcal{F}_{n-1}$ such that we have in distribution

$$Z(B_n; X_n) = \mathbb{E}[Z(B_n; X_n) \mid \mathcal{F}_{n-1}] + \sqrt{\mathsf{Cov}[Z(B_n; X_n) \mid \mathcal{F}_{n-1}]} \begin{pmatrix} Y_0 \\ \vdots \\ Y_N \end{pmatrix}, \tag{39}$$

where we denote the cholesky decomposition of a matrix A by the squareroot $\sqrt{A}$.

Remark 32 *As we only wish to prove convergence in probability to deterministic numbers, this distributional equality will be enough for our purposes, but if the covariance was invertible one could get a almost sure equality (cf. Remark 39). Later, we will also see that the covariance is at least asymptotically invertible due to strict positive definiteness of Z and asymptotically different evaluation points (Ind-IV).*

Calculating the conditional expectation and covariance Since the first two moments can be calculated by treating the inputs as deterministic (cf. Equation (38)), we can calculate the conditional expectation and variance using the well known formulas for Gaussian conditionals (cf. Theorem 38)

$$\mathbb{E}[Z(B_n; X_n) \mid \mathcal{F}_{n-1}] = \mu_n^N + \mathbf{\Sigma}_{[0:n),n}^{N\ T} [\mathbf{\Sigma}_{[0:n)}^N]^{-1} (Z(B_n; X_{[0:n)}) - \mu_{[0:n)}^N) \tag{40}$$

$$\mathsf{Cov}[Z(B_n; X_n) \mid \mathcal{F}_{n-1}] = \tfrac{1}{N}\left[\mathbf{\Sigma}_n^N - \mathbf{\Sigma}_{[0:n),n}^{N\ T} [\mathbf{\Sigma}_{[0:n)}^N]^{-1} \mathbf{\Sigma}_{[0:n),n}^N\right], \tag{41}$$

where we define the mixed covariance by

$$\tfrac{1}{N}\mathbf{\Sigma}_{[0:n),n}^N := \mathcal{C}_Z(B_n; X_{[0:n)}, X_n) \qquad \mathcal{C}_Z(b; x_{[0:n)}, x_n) := \mathsf{Cov}(Z(b; x_{[0:n)}), Z(b; x_n)),$$

the autocovariance matrices by

$$\begin{aligned} \tfrac{1}{N}\mathbf{\Sigma}_{[0:n)}^N &:= \mathcal{C}_Z(B_n; X_{[0:n)}) & \mathcal{C}_Z(b; x_{[0:n)}) &:= \mathsf{Cov}[Z(b; x_{[0:n)})] \\ \tfrac{1}{N}\mathbf{\Sigma}_n^N &:= \mathcal{C}_Z(B_n; X_n) & \mathcal{C}_Z(b; x_n) &:= \mathsf{Cov}[Z(b; x_n)] \end{aligned}$$

and the expectations by

$$\begin{aligned} \mu_n^N &:= \mu_Z(B_n; X_n) & \mu_Z(b; x_n) &:= \mathbb{E}[Z(b; x_n)] \\ \mu_{[0:n)}^N &:= \mu_Z(B_n; X_n) & \mu_Z(b; x_{[0:n)}) &:= \mathbb{E}[Z(b; x_{[0:n)})] \end{aligned}$$

The detour over the functions $\mathcal{C}_Z$ and μ_Z was necessary to define the unconditional covariance matrices properly, because inputs may only be treated as deterministic in conditional distributions. So in general, since B_n and X_n are random variables, we have for the unconditional covariance

$$\mathcal{C}_Z(B_n; X_n) \neq \mathsf{Cov}[Z(B_n; X_n)].$$

To ensure the inputs are treated as deterministic as Equation (38) demands, we therefore had to make sure the expectation was already applied before we plugged in our random input.

Note that we have moved the dimensional scaling $\frac{1}{N}$ of the covariances (cf. Lemma 20) outside of our definition of $\mathbf{\Sigma}^N_{[0:n)}$, $\mathbf{\Sigma}^N_{[0:n),n}$ and $\mathbf{\Sigma}^N_n$, as we are now going to prove their entries, and the entries of $\mu^N_{[0:n)}$ and μ^N_n, converge. This will eventually allow us to prove that the conditional expectation and covariance will converge, which leads to the stochastic convergence of Z we want to obtain for (Ind-I). Moving the dimensional scaling out also made the dimensional scaling of the conditional covariance in (41) much more visible.

Covariance matrix entries converge Recall that Z is made up of evaluations of $\mathbf{f}_N$ and $D_v\mathbf{f}_N$ for directions v. Thus the entries of its covariance matrices are given by Lemma 20, which we restate here for your convenience.

Lemma 33 (Covariance of derivatives, non-stationary isotropy) *Let $\mathbf{f}_N \sim \mathcal{N}(\mu, \kappa)$ be (non-stationary) isotropic (Definition 6). Then the expectation of a directional derivative is given by*

$$\mathbb{E}[D_v\mathbf{f}_N(x)] = \mu'\big(\tfrac{\|x\|^2}{2}\big)\langle x, v\rangle, \tag{16}$$

whereas, using the notation $\kappa := \kappa\big(\frac{\|x\|^2}{2}, \frac{\|y\|^2}{2}, \langle x, y\rangle\big)$ to omit inputs, the covariance of directional derivatives is given by

$$\mathsf{Cov}(D_v\mathbf{f}_N(x), \mathbf{f}_N(y)) = \frac{1}{N}\Big[\kappa_1\langle x, v\rangle + \kappa_3\langle y, v\rangle\Big] \tag{17}$$

$$\mathsf{Cov}(D_v\mathbf{f}_N(x), D_w\mathbf{f}_N(y)) = \frac{1}{N}\Big[\underbrace{\kappa_{12}\langle x, v\rangle\langle y, w\rangle + \kappa_{13}\langle x, v\rangle\langle x, w\rangle + \kappa_{32}\langle y, v\rangle\langle y, w\rangle + \kappa_{33}\langle y, v\rangle\langle y, w\rangle}_{(I)} + \underbrace{\kappa_3\langle v, w\rangle}_{(II)}\Big]. \tag{18}$$

In particular, if the directions v, w are orthogonal to the points x, y then (17) and (I) of (18) are zero. (II) in turn is zero, if the directions v, w are orthogonal.

Recall that we have (25) of (Ind-III) for $n-1$ by induction assumption, i.e. we have for all $k \leq n$ and all $i < d_n$

$$\langle X_k, \mathbf{v}_i\rangle \xrightarrow[N\to\infty]{p} y_k^{(i)} \quad \text{and} \quad \|X_k\|^2 = \sum_{i=0}^{d_n-1}\langle X_k, \mathbf{v}_i\rangle^2 \xrightarrow[N\to\infty]{p} \|y_k\|^2.$$

Since the X_k for $k \leq n$ are contained in V_n orthogonal to $\mathbf{w}_{[d_n:N)}$ we also have for all $i \geq d_n$

$$\langle X_k, \mathbf{w}_i\rangle = 0. \tag{42}$$

Put together, we have that all the inner products $\langle X_k, v\rangle$ for $k \leq n$ and $v \in B_n$ converge.

The entries of $\mathbf{\Sigma}^N_{[0:n)}$, $\mathbf{\Sigma}^N_{[0:n),n}$, $\mathbf{\Sigma}^N_n$, $\mu^N_{[0:n)}$ and μ^N_n calculated with Lemma 20 therefore all converge in probability by the continuous mapping theorem, since κ and μ are sufficiently smooth by Assumption 11 used in Theorem 22. Observe that it was very important to remove the dimensional scaling $\frac{1}{N}$ of (17) and (18) from the covariance matrices $\mathbf{\Sigma}^N_{[0:n)}$, $\mathbf{\Sigma}^N_{[0:n),n}$ and $\mathbf{\Sigma}^N_n$, as their entries would otherwise all converge to zero. As we want to invert $\mathbf{\Sigma}^N_{[0:n)}$ this would have been very inconvenient.

We further note that the sizes of these covariance matrices change with the dimension, because the number of directional derivatives increases with N which increases the size of $Z(\mathbf{v}_{[0:d_n)}, \mathbf{w}_{[d_n:N)}; X_{[0:n)})$ and therefore the size of its covariance matrix. The convergence of their entries is therefore not yet sufficient for a limiting object to be well defined.

Splitting the increasing matrices into block matrices of constant size This is the heart of the proof, which relies heavily on our custom coordinate system $B_n = (\mathbf{v}_{[0:d_n)}, \mathbf{w}_{[d_n:N)})$. To understand this, let us focus on the covariance of derivatives given in (18) of Lemma 20, i.e.

$$\mathsf{Cov}(D_v \mathbf{f}_N(x), D_w \mathbf{f}_N(y)) = \frac{1}{N}\Big[\underbrace{\kappa_{12}\langle x,v\rangle\langle y,w\rangle + \kappa_{13}\langle x,v\rangle\langle x,w\rangle + \kappa_{32}\langle y,v\rangle\langle y,w\rangle + \kappa_{33}\langle y,v\rangle\langle y,w\rangle}_{\text{(I)}} + \underbrace{\kappa_3\langle v,w\rangle}_{\text{(II)}}\Big]. \tag{18}$$

Notice that for the $\mathbf{w}_i$ the part (I) is always zero by (42). This is why we defined $\mathbf{w}_{[d_n:n)}$ to be orthogonal to V_n in Definition 21. Since we also defined our basis to be an orthonormal basis, (II) is only non-zero when covariances of the directional derivatives in the same direction are taken.

As we treat our Z matrix (36) as row-major, the directional derivatives are grouped by direction $\mathbf{w}_i$ in Z, which therefore results in the following block matrix structure.

$$\begin{aligned}
\tfrac{1}{N}\boldsymbol{\Sigma}^N_{[0:n)} &= \mathcal{C}_Z(\mathbf{v}_{[0:d_n)}, \mathbf{w}_{[d_n:N)}; X_{[0:n)}) \\
&= \begin{pmatrix} \mathcal{C}_Z(\mathbf{v}_{[0:d_n)}; X_{[0:n)}) & & & \\ & \mathcal{C}_{D_{\mathbf{w}_{d_n}}\mathbf{f}_N}(X_{[0:n)}) & & \\ & & \ddots & \\ & & & \mathcal{C}_{D_{\mathbf{w}_{N-1}}\mathbf{f}_N}(X_{[0:n)}) \end{pmatrix} \\
&=: \frac{1}{N}\begin{pmatrix} \boxed{\boldsymbol{\Sigma}^{v,N}_{[0:n)}} & & & \\ & \boxed{\boldsymbol{\Sigma}^{w,N}_{[0:n)}} & & \\ & & \ddots & \\ & & & \boxed{\boldsymbol{\Sigma}^{w,N}_{[0:n)}} \end{pmatrix},
\end{aligned} \tag{43}$$

For $\boldsymbol{\Sigma}^{w,N}_{[0:n)}$ to be well defined, we require the blocks $\mathcal{C}_{D_{\mathbf{w}_i}\mathbf{f}_N}(X_{[0:n)})$ to not depend on i. But since (I) is zero and $\langle \mathbf{w}_i, \mathbf{w}_i\rangle = 1$ we have by (18) of Lemma 20

$$\begin{aligned}
&\mathcal{C}_{D_{\mathbf{w}_i}\mathbf{f}_N}(X_{[0:n)}) \\
&= \frac{1}{N}\begin{pmatrix} \kappa_3\big(\frac{\|X_0\|^2}{2}, \frac{\|X_0\|^2}{2}, \langle X_0, X_0\rangle\big) & \cdots & \kappa_3\big(\frac{\|X_0\|^2}{2}, \frac{\|X_{n-1}\|^2}{2}, \langle X_0, X_{n-1}\rangle\big) \\ \vdots & & \vdots \\ \kappa_3\big(\frac{\|X_{n-1}\|^2}{2}, \frac{\|X_0\|^2}{2}, \langle X_{n-1}, X_0\rangle\big) & \cdots & \kappa_3\big(\frac{\|X_{n-1}\|^2}{2}, \frac{\|X_{n-1}\|^2}{2}, \langle X_{n-1}, X_{n-1}\rangle\big) \end{pmatrix} \\
&=: \tfrac{1}{N}\boldsymbol{\Sigma}^{w,N}_{[0:n)}.
\end{aligned}$$

In particular there is no dependence on $\mathbf{w}_i$ so $\boldsymbol{\Sigma}^{w,N}_{[0:n)}$ is well defined. Since these block matrices are of constant size, they converge if all their finitely many entries converge. But we already argued that the entries converge (in the previous paragraph) and we therefore

have[4]

$$\boldsymbol{\Sigma}^{v,N}_{[0:n)} \xrightarrow[N\to\infty]{p} \Sigma^{v,\infty}_{[0:n)} \in \mathbb{R}^{n(d_n+1)\times n(d_n+1)} \quad \text{and} \quad \boldsymbol{\Sigma}^{w,N}_{[0:n)} \xrightarrow[N\to\infty]{p} \Sigma^{w,\infty}_{[0:n)} \in \mathbb{R}^{n\times n}.$$

For the mixed covariance we similarly have

$$\begin{aligned}
\tfrac{1}{N}\boldsymbol{\Sigma}^{N}_{[0:n),n} &= \mathcal{C}_Z(\mathbf{v}_{[0:d_n)}, \mathbf{w}_{[d_n:N)}; X_{[0:n)}, X_n) \\
&= \begin{pmatrix} \mathcal{C}_Z(\mathbf{v}_{[0:n)}; X_{[0:n)}, X_n) & & & \\ & \mathcal{C}_{D_{w_{d_n}}\mathbf{f}_N}(X_{[0:n)}, X_n) & & \\ & & \ddots & \\ & & & \mathcal{C}_{D_{w_{N-1}}\mathbf{f}_N}(X_{[0:n)}, X_n) \end{pmatrix} \\
&=: \frac{1}{N} \begin{pmatrix} \boxed{\boldsymbol{\Sigma}^{v,N}_{[0:n),n}} & & & \\ & \boxed{\boldsymbol{\Sigma}^{w,N}_{[0:n),n}} & & \\ & & \ddots & \\ & & & \boxed{\boldsymbol{\Sigma}^{w,N}_{[0:n),n}} \end{pmatrix},
\end{aligned} \tag{44}$$

with a similar argument why $\boldsymbol{\Sigma}^{w,N}_{[0:n),n}$ is well defined as for $\boldsymbol{\Sigma}^{w,N}_{[0:n)}$. And again by the discussion of the previous paragraph establishing convergence of the entries, we have that these block matrices of constant size converge

$$\boldsymbol{\Sigma}^{v,N}_{[0:n),n} \xrightarrow[N\to\infty]{p} \Sigma^{v,\infty}_{[0:n),n} \in \mathbb{R}^{n(d_n+1)\times(d_n+1)} \quad \text{and} \quad \boldsymbol{\Sigma}^{w,N}_{[0:n),n} \xrightarrow[N\to\infty]{p} \Sigma^{w,\infty}_{[0:n),n} \in \mathbb{R}^{n\times 1}.$$

We can also split up the autocovariance $\boldsymbol{\Sigma}^{N}_{n}$ in a similar fashion with

$$\boldsymbol{\Sigma}^{v,N}_{n} \xrightarrow[N\to\infty]{p} \Sigma^{v,\infty}_{n} \in \mathbb{R}^{(d_n+1)\times(d_n+1)} \quad \text{and} \quad \boldsymbol{\Sigma}^{w,N}_{n} \xrightarrow[N\to\infty]{p} \Sigma^{w,\infty}_{n} \in \mathbb{R}^{1\times 1}.$$

Finally, the expectation functions can also be split into a block containing the directional derivatives $\mathbf{v}_{[0:d_n)}$ spanning V_n and the expectations of directional derivatives in the $\mathbf{w}_i$ directions. More specifically we have

$$\mu^{v,N}_{n} = \mu_Z(\mathbf{v}_{[0:d_n)}; X_n) \xrightarrow[N\to\infty]{p} \mu^{v,\infty}_{n} \in \mathbb{R}^{d_n+1} \quad \text{and} \quad \mu^{v,N}_{[0:n)} \xrightarrow[N\to\infty]{p} \mu^{v,\infty}_{[0:n)} \in \mathbb{R}^{n(d_n+1)}, \tag{45}$$

4. The increasing number of identical $\boldsymbol{\Sigma}^{w,N}_{[0:n)}$ will drive the law of large numbers of

$$\|P_{V_n^\perp}\nabla\mathbf{f}_N(x_0)\|^2 = \sum_{i=d_n}^{N-1} (D_{\mathbf{w}_i}\mathbf{f}_N(x_0))^2$$

similar to the first step (cf. Equation (34)). At the moment they are still matrices but this will change when combined with the mixed covariances $\boldsymbol{\Sigma}^{w,N}_{[0:n),n}$ finally resulting in (53).

which converge since the inner products and norms used in (16) converge. Since the directions $\mathbf{w}_i$ are selected orthogonal to all evaluation points $X_n \in V_n$, the inner product in the expectation of the directional derivative (16) is always zero and we therefore have

$$\mu_n^{w,N} = \mu_Z(\mathbf{w}_{[d_n:N)}; X_n) = 0 \in \mathbb{R} \quad \text{and} \quad \mu_{[0:n)}^{w,N} = 0 \in \mathbb{R}^n. \tag{46}$$

Convergence of the conditional expectation Let us take a step back and review what we have done and want to do. We have argued that $Z(B_n; X_{[0:n)})$ is conditionally Gaussian and that it is therefore enough to understand the conditional expectation and covariance (cf. (39)). We have then applied a well known result about the conditional distribution of Gaussian random variables to obtain explicit formulas for the conditional expectation (40) and covariance (41) made up of unconditional covariance matrices. We proved that their entries converged and split these covariance matrices into block form, such that these blocks of constant size converge in probability. What is left to do, is to put these results together to prove convergence results about $Z(B_n; X_{[0:n)})$. We start with the conditional expectation and then get to $Z(B_n; X_{[0:n)})$ itself by a consideration of the conditional covariance.

So let us take a look at the conditional expectation. Since the $\mathbf{w}_i$ are by definition orthonormal to the previsible running span of evaluation points V_n (19), we have $D_{\mathbf{w}_k}\mathbf{f}_N(X_{[0:n)}) = 0$. By applying the block matrix structure (43), (44) to the the representation of the conditional expectation (40) we therefore get using (46)

$$\mathbb{E}[D_{\mathbf{w}_i}\mathbf{f}_N(X_n) \mid \mathcal{F}_{n-1}] = \mu_n^{w,N} + {\boldsymbol{\Sigma}_{[0:n),n}^{w,N}}^T [\boldsymbol{\Sigma}_{[0:n)}^{w,N}]^{-1}(D_{\mathbf{w}_k}\mathbf{f}_N(X_{[0:n)}) - \mu_{[0:n)}^{w,N}) = 0. \tag{47}$$

The only interesting part of the vector $\mathbb{E}[Z(\mathbf{v}_{[0:d_n)}, \mathbf{w}_{[d_n:N)}; X_n) \mid \mathcal{F}_{n-1}]$ are therefore the first d_n entries $\mathbb{E}[Z(\mathbf{v}_{[d_n:n)}; X_n) \mid \mathcal{F}_{n-1}]$. For this we use the formula for the conditional expectation (40) and our block matrix decompositions (43), (44) again to obtain

$$\mathbb{E}[Z(\mathbf{v}_{[0:d_n)}; X_n) \mid \mathcal{F}_{n-1}] = \mu_n^{(v,N)} + {\boldsymbol{\Sigma}_{[0:n),n}^{v,N}}^T [\boldsymbol{\Sigma}_{[0:n)}^{v,N}]^{-1}(Z(\mathbf{v}_{[0:d_n)}; X_{[0:n)}) - \mu_{[0:n)}^{(v,N)}). \tag{48}$$

Since $Z(\mathbf{v}_{[0:d_n)}; X_{[0:n)})$ converges in probability by the induction assumptions (Ind-I) summarized in (36), and since we have spent the previous paragraphs proving that the covariance matrices and expectations converge, it almost seems like we are done. But while the conditional expectation (40) can handle non-invertible matrices with a generalized inverse (Theorem 38), an application of continuous mapping requires continuity and the matrix inverse is only continuous at invertible matrices.

Lemma 34 *$\Sigma_{[0:n)}^{v,\infty}$ is strictly positive definite and therefore invertible.*

Proof The essential ingredients are that $(\mathbf{f}_N, \nabla\mathbf{f}_N)$ is strictly positive definite by assumption of Theorem 22 and that the evaluation points are asymptotically different $\rho_{ij} > 0$ by (Ind-IV). The technical complications are that $\Sigma_{[0:n)}^{v,\infty}$ is only defined as the limit of covariance matrices (which are themselves not necessarily strictly positive definite themselves) and that this limit involves changing the domain of $(\mathbf{f}_N, \nabla\mathbf{f}_N)$ as we increase its dimension.

Let us first address the problem of the changing domain. Since we are only interested in the block matrix of the derivatives in the directions contained in V_n and the evaluation points $X_{[0:n)}$ are also contained in V_n we can map them via an isometry to $\mathbb{R}^{d_n}$ which

retains all distances and inner products and therefore retains the covariance matrices $\mathbf{\Sigma}^{v,N}_{[0:n)}$ (cf. Lemma 20). Note that the dimensional scaling $\frac{1}{N}$ is not an issue here, as we have already removed the scaling from $\mathbf{\Sigma}^{v,N}_{[0:n)}$. Therefore $\mathbf{\Sigma}^{v,N}_{[0:n)}$ can be viewed as a sequence of covariance matrices of $(\mathbf{f}_N, \nabla \mathbf{f}_N)$ with $\mathbf{f}_N : \mathbb{R}^{d_n} \to \mathbb{R}$. This solves the problem of the changing domain.

Finally, we need that the limiting matrix $\Sigma^{v,\infty}_{[0:n)}$ is in fact a covariance matrix of $(\mathbf{f}_N, \nabla \mathbf{f}_N)$ and not just a limit of covariance matrices. For this we apply (Ind-III), which ensures that the limiting scalar products (25) and distances (26) which make up $\Sigma^{v,\infty}_{[0:n)}$ (cf. Lemma 20) are realizable by points $y_0, \dots, y_{n-1} \in \mathbb{R}^{d_n}$. $\Sigma^{v,\infty}_{[0:n)}$ is therefore a covariance matrix of $(\mathbf{f}_N, \nabla \mathbf{f}_N)$ viewed as a random function with domain $\mathbb{R}^{d_n}$. Since the distances ρ_{ij} are positive by (Ind-IV), the asymptotic representations y_k are distinct and their covariance matrix $\Sigma^{v,\infty}_{[0:n)}$ is therefore strictly positive definite by the strict positive definiteness of $(\mathbf{f}_N, \nabla \mathbf{f}_N)$. ■

The matrix inverse is therefore continuous in $\Sigma^{v,\infty}_{[0:n)}$ and by the convergence of the matrices in probability and the convergence of $Z(\mathbf{v}_{[0:d_n)}; X_{[0:n)})$ by induction assumptions (Ind-I) for $n-1$ restated in (36), the conditional expectation (48) converges in probability

$$\begin{aligned}
&\mathbb{E}[Z(\mathbf{v}_{[0:d_n)}; X_n) \mid \mathcal{F}_{n-1}] \\
&\xrightarrow[N\to\infty]{p} \mu^{v,\infty}_n + {\Sigma^{v,\infty}_{[0:n),n-1}}^T [\Sigma^{v,\infty}_{[0:n)}]^{-1} \left(\begin{pmatrix} \mathfrak{f}_{[0:n)} \\ \gamma_{[0:n)} \end{pmatrix} - \mu^{v,\infty}_{[0:n)} \right) =: \begin{pmatrix} \mathfrak{f}_n \\ \gamma_n^{([0:d_n))} \end{pmatrix}.
\end{aligned} \tag{49}$$

Together with the conditional expectation of the directional derivatives in the directions $\mathbf{w}_i$ in (47), we have now fully determined the conditional expectation.

An analysis of the conditional variance will immediately give us an understanding of the distribution and we can therefore put these considerations together to obtain convergence of Z itself.

***Step 1*: Convergence of $Z(\mathbf{v}_{[0:d_n)}; X_n)$** Recall, we outlined at the start of the induction that we would first prove convergence of $Z(\mathbf{v}_{[0:d_n)}; X_n)$ (i.e. *Step 1*) before proving V_{n+1} to asymptotically have full rank (*Step 2*) and the convergence of the new corner element $D_{\mathbf{v}_{d_n}} \mathbf{f}_N(X_n)$ (*Step 3*).

In (49) we have found the limit of the conditional expectation of $Z(\mathbf{v}_{[0:d_n)}; X_n)$. This limit is also going to be the limit of the new column $Z(\mathbf{v}_{[0:d_n)}; X_n)$ itself. To prove this we only need to control the variance. For this purpose we recall that the unconditional variance is given by

$$\mathbf{\Sigma}^N_n = \begin{pmatrix} \mathbf{\Sigma}^{v,N}_n & \\ & \kappa_3\left(\frac{\|X_n\|^2}{2}, \frac{\|X_n\|^2}{2}, \|X_n\|^2\right) \mathbb{I}_{(N-d_n)\times(N-d_n)} \end{pmatrix}.$$

The conditional covariance matrix is therefore given by

$$\begin{aligned}
&\mathsf{Cov}[Z(\mathbf{v}_{[0:d_n)}, \mathbf{w}_{[d_n:N)}; X_n) \mid \mathcal{F}_{n-1}] \\
&= \frac{1}{N} \begin{bmatrix} \mathbf{\Sigma}^{v,N}_n - {\mathbf{\Sigma}^{v,N}_{[0:n),n}}^T [\mathbf{\Sigma}^{v,N}_{[0:n)}]^{-1} \mathbf{\Sigma}^{v,N}_{[0:n),n} & \\ & \sigma^2_{w,d} \mathbb{I}_{(N-d_n)\times(N-d_n)} \end{bmatrix},
\end{aligned} \tag{50}$$

where the conditional variance in the direction of the $\mathbf{w}_i$ is given by

$$\sigma^2_{w,N} \coloneqq \kappa_3\big(\tfrac{\|X_n\|^2}{2}, \tfrac{\|X_n\|^2}{2}, \|X_n\|^2\big) - \mathbf{\Sigma}^{w,N}_{[0:n),n}{}^T[\mathbf{\Sigma}^{w,N}_{[0:n)}]^{-1}\mathbf{\Sigma}^{w,N}_{[0:n),n}. \tag{51}$$

We will need $\sigma^2_{w,N}$ for *Step 2* and *Step 3*, but for now it is not important. Due to the diagonal structure we have for our new column $Z(\mathbf{v}_{[0:d_n)}; X_n)$ by (39)

$$\begin{aligned}
&Z(\mathbf{v}_{[0:d_n)}, X_n)\\
&= \mathbb{E}[Z(\mathbf{v}_{[0:d_n)}, X_n) \mid \mathcal{F}_{n-1}] + \sqrt{\mathsf{Cov}[Z(\mathbf{v}_{[0:d_n)}, X_n) \mid \mathcal{F}_{n-1}]} \begin{pmatrix} Y_0 \\ \vdots \\ Y_{d_n} \end{pmatrix}\\
&= \mathbb{E}[Z(\hat{v}_{[0:d_n)}, X_n) \mid \mathcal{F}_{n-1}] + \frac{1}{\sqrt{N}} \sqrt{\mathbf{\Sigma}^{v,N}_n - \mathbf{\Sigma}^{v,N}_{[0:n),n}{}^T[\mathbf{\Sigma}^{v,N}_{[0:n)}]^{-1}\mathbf{\Sigma}^{v,N}_{[0:n),n}} \begin{pmatrix} Y_0 \\ \vdots \\ Y_{d_n} \end{pmatrix}\\
&\xrightarrow[N\to\infty]{p} \begin{pmatrix} \mathfrak{f}_n \\ \gamma_n^{([0:d_n))} \end{pmatrix}.
\end{aligned} \tag{52}$$

This is exactly the convergence of the 'new column' required for *Step 1*.

***Step 2*: The rank of V_{n+1}** The last two steps follow fairly quickly. Recall that V_{n+1} is the *previsible* running span of evaluation points defined in (19). Since it is previsible, it only includes gradients up to time n and the Gram-Schmidt candidate (21) of its most recent addition is therefore given by

$$\tilde{\mathbf{v}}_n \coloneqq \nabla \mathbf{f}_N(X_n) - P_{V_n}\nabla \mathbf{f}_N(X_n) = \sum_{k=d_n}^{N-1} \langle \mathbf{w}_i, \nabla \mathbf{f}_N(X_n)\rangle \mathbf{w}_i = \sum_{k=d_n}^{N-1} D_{\mathbf{w}_i}\mathbf{f}_N(X_n)\mathbf{w}_i,$$

where P_{V_n} is the projection onto V_n.

Now we naturally want to analyze the directional derivatives $D_{\mathbf{w}_i}\mathbf{f}_N(X_n)$ which make up $\tilde{\mathbf{v}}_n$. By representation (39) and our formula for the conditional covariance (50) we have in distribution

$$D_{\mathbf{w}_i}\mathbf{f}_N(X_n) = \underbrace{\mathbb{E}[D_{\mathbf{w}_i}\mathbf{f}_N(X_n) \mid \mathcal{F}_{n-1}]}_{\overset{(47)}{=}0} + \sqrt{\tfrac{1}{N}\sigma_{w,N}}Y_{i+1}.$$

We now already see where our law of large numbers is going to come from. But we first need to take a closer look at the residual variance $\sigma^2_{w,N}$ defined in (51). We get convergence in probability of the residual variance to a value strictly greater zero

$$\begin{aligned}
\sigma^2_{w,N} &= \big(\kappa_3\big(\tfrac{\|X_n\|^2}{2}, \tfrac{\|X_n\|^2}{2}, \|X_n\|^2\big) - \mathbf{\Sigma}^{w,N}_{[0:n),n}{}^T[\mathbf{\Sigma}^{w,N}_{[0:n)}]^{-1}\mathbf{\Sigma}^{w,N}_{[0:n),n}\big)\\
&\xrightarrow[N\to\infty]{p} \big(\kappa_3\big(\tfrac{\|y_n\|^2}{2}, \tfrac{\|y_n\|^2}{2}, \|y_n\|^2\big) - \Sigma^{w,\infty}_{[0:n),n}{}^T[\Sigma^{w,\infty}_{[0:n)}]^{-1}\Sigma^{w,\infty}_{[0:n),n}\big) =: \sigma^2_{w,\infty} > 0,
\end{aligned} \tag{53}$$

using the convergence of the block matrices and the following lemma.

Lemma 35 $\Sigma^{w,\infty}_{[0:n)}$ *is strictly positive definite and* $\sigma^2_{w,\infty} > 0$.

Proof We are going to show with a similar argument as in Lemma 34 that the matrix

$$\Sigma^{w,\infty}_{[0:n]} := \begin{bmatrix} \Sigma^{w,\infty}_{[0:n)} & \Sigma^{w,\infty}_{[0:n),n} \\ \Sigma^{w,\infty}_{[0:n),n}{}^T & \kappa_3\big(\frac{\|y_n\|^2}{2}, \frac{\|y_n\|^2}{2}, \|y_n\|^2\big) \end{bmatrix}$$

is strictly positive definite. Before we do so, let us quickly argue why this finishes the proof. Since $\Sigma^{w,\infty}_{[0:n]}$ is then strictly positive definite, it has a cholesky decomposition L such that

$$\Sigma^{w,\infty}_{[0:n]} = LL^T = \begin{bmatrix} L_n & 0 \\ l^T & \sigma \end{bmatrix} \begin{bmatrix} L_n^T & l \\ l & \sigma \end{bmatrix},$$

which implies that L_n is the cholesky decomposition of $\Sigma^{w,\infty}_{[0:n)}$, $l = L_n^{-1}\Sigma^{w,\infty}_{[0:n),n}$ and

$$\sigma = \sqrt{\kappa_3\big(\tfrac{\|y_n\|^2}{2}, \tfrac{\|y_n\|^2}{2}, \|y_n\|^2\big) - \Sigma^{w,\infty}_{[0:n),n}{}^T [\Sigma^{w,\infty}_{[0:n)}]^{-1} \Sigma^{w,\infty}_{[0:n),n}} = \sigma_{w,\infty}.$$

Since we have

$$\det(L) = \sigma_{w,\infty} \det(L_n),$$

strict positive definiteness of $\Sigma^{w,\infty}_{[0:n]}$ and therefore $0 \neq \det(\Sigma^{w,\infty}_{[0:n]}) = \det(L)^2$ implies

$$\sigma^2_{w,\infty} > 0 \quad \text{and} \quad \det(L_n) \neq 0.$$

But $\det(L_n) \neq 0$ also implies that $\Sigma^{w,\infty}_{[0:n)}$ has to be strictly positive definite, since L_n is its cholesky decomposition.

What is therefore left to prove is the strict positive definiteness of $\Sigma^{w,\infty}_{[0:n]}$. For this note that

$$\mathbf{\Sigma}^{w,N}_{[0:n]} := \begin{bmatrix} \mathbf{\Sigma}^{w,N}_{[0:n)} & \mathbf{\Sigma}^{w,N}_{[0:n),n} \\ \mathbf{\Sigma}^{w,N}_{[0:n),n}{}^T & \kappa_3\big(\frac{\|X_n\|^2}{2}, \frac{\|X_n\|^2}{2}, \|X_n\|^2\big) \end{bmatrix}$$

is the plug-in covariance matrix of $D_{\mathbf{w}_i}\mathbf{f}_N(X_{[0:n]})$. Where we use the term "plug-in" covariance to say that we treat the evaluation points $X_{[0:n]}$ and the direction $\mathbf{w}_i$ as deterministic. Since $X_{[0:n]}$ are contained in V_n, we again map them isometrically to $\mathbb{R}^{d_n}$. But this time we view $\mathbb{R}^{d_n}$ as a subspace of $\mathbb{R}^{d_n+1}$ and map the additional vector $\mathbf{w}_i$ to e_{d_n+1} such that $\mathbf{\Sigma}^{w,N}_{[0:n]}$ is a covariance matrix of $\nabla\mathbf{f}_N$ in $\mathbb{R}^{d_n+1}$.

Now we finish the proof with the same limiting argument as in Lemma 34. Since $\nabla\mathbf{f}_N$ is strictly positive definite by assumption, we get that $\Sigma^{w,\infty}_{[0:n]}$ is strictly positive definite as the covariance matrix of $(\partial_{d_n+1}\mathbf{f}_N)$ at the points $y_{[0:n]} = (y_0, \dots, y_n) \subseteq \mathbb{R}^{d_n+1}$ of (Ind-III). These are the limiting representations and non of the points y_k are equal by (Ind-IV). ■

Using the convergence of the residual variance $\sigma^2_{w,N} \to \sigma^2_{w,\infty}$ allows us to finally make our law of large numbers argument

$$\|\tilde{\mathbf{v}}_n\|^2 = \sum_{i=d_n}^{N-1} (D_{\mathbf{w}_i}\mathbf{f}_N(X_n))^2 = \frac{\sigma^2_{w,N}}{N} \sum_{i=d_n+1}^{N} Y_i^2 \underset{N\to\infty}{\overset{p}{\to}} \sigma^2_{w,\infty} > 0. \tag{54}$$

This implies that V_n has full rank asymptotically (Ind-II).

Assuming the last gradient is always used (and not just in the asymptotic limit), we can get $d_{n+1} = d_n + 1$ almost surely, since $\sigma^2_{w,N} > 0$ almost surely is sufficient for $\|\tilde{\mathbf{v}}_n\|^2 > 0$ almost surely. The use of the most recent gradient ensures the points X_k are different by an inductive argument similar to Lemma 29. This in turn ensures, using the strict positive definiteness of $(\mathbf{f}_N, \nabla \mathbf{f}_N)$, that $\sigma^2_{w,N} > 0$ almost surely with a similar argument as in Lemma 35.

***Step 3*: Convergence of the new corner element** The last step follows immediately by definition of $\mathbf{v}_{d_n} = \frac{\tilde{\mathbf{v}}_n}{\|\tilde{\mathbf{v}}_n\|}$ in Definition 21 and (54)

$$D_{\mathbf{v}_{d_n}} \mathbf{f}_N(X_n) = \left\langle \nabla \mathbf{f}_N(X_n), \tfrac{\tilde{\mathbf{v}}_n}{\|\tilde{\mathbf{v}}_n\|} \right\rangle = \|\tilde{\mathbf{v}}_n\| \underset{N\to\infty}{\xrightarrow{p}} \sigma_{w,\infty} =: \gamma_n^{(d_n)} > 0.$$

5. Discussion and Outlook

In this section we want to discuss limitations, possible generalizations and extensions of our results. Let us begin by with the only structural assumption:

S1. **Isotropy.** This assumption is the key argument to get independent, identically distributed partial derivatives and thereby a law of large numbers of the gradient norm. Isotropic kernels are characterized by the property $\mathcal{C}(x, y) = \mathcal{C}(Ux, Uy)$ for any orthogonal matrix U (Benning and Schölpple, 2025). It is natural to ask how small the set of transformations U may be before our results break down. In place of rotation invariance, one might consider exchangeable directions for example (which is captured by the set of dimension permutations). In the first step this would likely yield exchangeable partial derivatives and and thereby a law of large numbers. However our proof approach would already break down in the second step as we are using an adapted coordinate system and not the standard basis. We are doubtful whether these results even hold in a more general setting.

Next we want to discuss the assumptions that may be removed or relaxed. We sort them by our current perception of their difficulty in increasing order:

A1. **Strict positive definiteness and use of the most recent gradient**. The strict positive definiteness was used in the proof of Theorem 22 to show that the limiting covariance matrices are invertible. Without this assumption one would either have to work with generalized matrix inverses and prove that these are still continuous in their entries or use an approach that avoids matrix inverses altogether such as describing the conditional distribution using characteristic functions.

Note that we needed the evaluation points to be different (Ind-IV) for strict positive definiteness to apply and we used the assumption that the most gradient was used for this purpose. This also ensured that the span of gradients V_n had full rank (Ind-II). Removing the strict positive definiteness and the use of the most recent gradient by the algorithm would therefore also remove these corollary implications.

A2. **Finite time horizon.** We only show convergence of the iterates up to a fixed time horizon n. This is structural to our qualitative approach as we need convergence of the previous step to show convergence of the next step. There are two potential ways to allow $n \to \infty$ jointly with the dimension N:

   (a) **Quantitative convergence and stability results.** The proof would be similar to proving convergence of an ODE discretization. In every step errors are accumulated and passed through continuous functions (which would need to be Lipschitz or similar in a quantitative setting). This error accumulation would need to happen slowly enough that the overall error still converges even if the number of error accumulation steps grows to infinity. One can observe empirically that the variance does not increase with the number of steps n (cf. Figure 2), which means that this is likely possible.

   (b) **Big n handover.** Since the iterates converge for every finite but possible very large time horizon n, it is perhaps possible to prove that for very large n it is overwhelmingly likely that the iterates are captured by a local minimum and are therefore constant anyway. This approach is likely significantly harder: One would first need to prove that the optimizer reaches a certain level at time n, which requires **explicit bounds on the** $\mathfrak{f}_n$ and then combine this with topological results about GRFs to show that at a certain level the function only has convex regions close to minima. These topological results are only partially available for the special case of spin glasses (e.g. Auffinger and Zeng, 2023).

A3. **Gaussian random function.** Our argument is based on a law of large numbers applied to the norm of gradients. For this we needed that the *squared* directional derivatives are uncorrelated. We obtained this result from the fact that the directional derivatives themselves are uncorrelated. In the Gaussian case they are thereby independent and their squares are therefore also uncorrelated. To remove the Gaussian assumption one would therefore have to obtain uncorrelated squares directly from isotropy. The difficulty then lies in pushing this argument through the steps, which may turn out to be straightforward or very difficult.

A4. **Gradient Span Algorithms.** In practice many algorithms leave the linear span of gradients with the use of preconditioning matrices. Example include Shampoo (Gupta et al., 2018) or Adam (Kingma and Ba, 2015), which uses a diagonal preconditioning matrix of entry-wise learning rates. While it is likely that similar results hold for these type of algorithms based on their empirical behavior (cf. Figure 1), it will be very difficult to prove this theoretically. Observe that a key step in our proof is to show that the 'learning rates' converge. More specifically we show that the coefficients of the evaluation locations with respect to the adapted span converge. This is necessary to ensure a deterministic behavior of the optimizer in the limit. For algorithms using preconditioners a similar result would need to be proven. However it is unclear in what sense a preconditioning matrix of increasing size may converge. Random matrix theory answers this question by stating that the distribution of eigenvalues converges. This approach loses the order of eigenvalues however, which is a problem here since they are highly correlated to the gradient realizations of the random function. Only

after a sense in which preconditioning matrices may converge is established one can hope to extend our results to this generality.

Perhaps the key lies in the fact that we are not considering arbitrary preconditioning matrices. Specifically, the authors conjecture in a different paper (Benning and Döring, 2024, Appendix E.1) that the purpose of the preconditioning matrix is to turn an anisotropic Random Function into an isotropic one and that preconditioning is unnecessary on isotropic functions.

Perhaps even more interesting than the relaxation of assumptions would be the following extensions:

E1. An analysis of **stochastic gradient descent** and similar stochastic algorithms which do not have access to the true cost but rather noisy evaluations of $\mathbf{f}_N$. It is again unclear in what sense one would obtain convergence, since we cannot expect convergence of the coordinates in the adapted span. Perhaps it is possible to prove convergence against a distribution of coordinates in the adapted span and thereby obtain convergence against a distribution of function values. Maybe ideas from Langevin dynamics can be reused here.

E2. **Bounds on the limiting function values** $\mathfrak{f}_n$ for a given optimizer and random function distribution would be extremely useful for comparing optimizers. Perhaps it is even possible to find the 'best optimizer' by optimizing over these limiting function values. This would be the most useful extension but also the most difficult. The optimizer that is one-step optimal is 'Random Function Descent' (Benning and Döring, 2024), but even the two-step optimal optimizer is very difficult to find as one has to find structure in in increasingly large covariance matrices.

Acknowledgements

This research was supported by the RTG 1953, funded by the German Research Foundation (DFG). The authors would like to thank Yan Fyodorov for his hospitality and helpful discussions at King's College. We thank him along with Mark Sellke and Antoine Maillard for their time and valuable insights on spin glasses.

References

Robert J. Adler and Jonathan E. Taylor. *Random Fields and Geometry*. Springer Monographs in Mathematics. Springer New York, New York, NY, 2007. ISBN 978-0-387-48112-8. doi: 10.1007/978-0-387-48116-6.

Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Optimization of mean-field spin glasses. *The Annals of Probability*, 49(6):2922–2960, November 2021. ISSN 0091-1798, 2168-894X. doi: 10.1214/21-AOP1519.

Antonio Auffinger and Qiang Zeng. Complexity of Gaussian Random Fields with Isotropic Increments. *Communications in Mathematical Physics*, 402(1):951–993, August 2023. ISSN 1432-0916. doi: 10.1007/s00220-023-04739-0.

Antonio Auffinger, Andrea Montanari, and Eliran Subag. Optimization of Random High-Dimensional Functions: Structure and Algorithms. In *Spin Glass Theory and Far Beyond*, pages 609–633. WORLD SCIENTIFIC, 5 Toh Tuck Link, Singapore, February 2023. ISBN 978-981-12-7391-9. doi: 10.1142/9789811273926_0029.

Mohsen Bayati and Andrea Montanari. The Dynamics of Message Passing on Dense Graphs, with Applications to Compressed Sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, February 2011. ISSN 1557-9654. doi: 10.1109/TIT.2010.2094817.

Felix Benning. Measure Theory of Conditionally Independent Random Function Evaluation. February 2026. doi: 10.48550/arXiv.2504.08513.

Felix Benning and Leif Döring. Random Function Descent. In *Advances in Neural Information Processing Systems*, volume 37, pages 111248–111298, Vancouver, Canada, December 2024. Curran Associates, Inc.

Felix Benning and Max David Schölpple. Schoenberg characterization of continuous non-stationary isotropic positive definite kernels. June 2025. doi: 10.48550/arXiv.2506.22048.

Sebastien Bubeck and Mark Sellke. A Universal Law of Robustness via Isoperimetry. In *Advances in Neural Information Processing Systems*, volume 34, pages 28811–28822, Virtual Event, 2021. Curran Associates, Inc.

Michael Celentano, Andrea Montanari, and Yuchen Wu. The estimation error of general first order methods. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 1078–1141. PMLR, July 2020.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gerard Ben Arous, and Yann LeCun. The Loss Surfaces of Multilayer Networks. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 192–204, San Diego, CA, USA, February 2015. PMLR.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, volume 27, Montréal, Canada, 2014. Curran Associates, Inc.

Percy Deift and Thomas Trogdon. The conjugate gradient algorithm on well-conditioned Wishart matrices is almost deterministic. *Quarterly of Applied Mathematics*, 79(1):125–161, March 2021. ISSN 0033-569X, 1552-4485. doi: 10.1090/qam/1574.

David L. Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, November 2009. doi: 10.1073/pnas.0909892106.

Morris L. Eaton. *Multivariate Statistics: A Vector Space Approach*, volume 53 of *Lecture Notes-Monograph Series*. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2007. ISBN 978-0-940600-69-0. doi: 10.1214/lnms/1196285102.

Peter I. Frazier. Bayesian Optimization. In *Recent Advances in Optimization and Modeling of Contemporary Problems*, INFORMS TutORials in Operations Research, chapter 11, pages 255–278. INFORMS, Phoenix, Arizona, USA, October 2018. ISBN 978-0-9906153-2-3. doi: 10.1287/educ.2018.0188.

Iosif Il'ich Gihman and Anatoliĭ Vladimirovich Skorokhod. *The Theory of Stochastic Processes I.* Classics in Mathematics. Springer, Berlin, Heidelberg, 1974. ISBN 978-3-540-20284-4 978-3-642-61943-4. doi: 10.1007/978-3-642-61943-4.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, Sardinia, Italy, March 2010. JMLR Workshop and Conference Proceedings.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018.

Magnus R. Hestenes and Eduard Stiefel. Methods of Conjugate Gradients for Solving Linear Systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, December 1952. doi: 10.6028/jres.049.044.

Brice Huang and Mark Sellke. Tight Lipschitz Hardness for optimizing Mean Field Spin Glasses. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 312–322, October 2022. doi: 10.1109/FOCS54457.2022.00037.

Brice Huang and Mark Sellke. A Constructive Proof of the Spherical Parisi Formula, May 2024.

Adel Javanmard and Andrea Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144, December 2013. ISSN 2049-8764. doi: 10.1093/imaiai/iat004.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, 2015.

Aaron Klein, Stefan Falkner, Jost Tobias Springenberg, and Frank Hutter. Learning Curve Prediction with Bayesian Neural Networks. In *International Conference on Learning Representations*, 2017.

Achim Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer, London, 2014. ISBN 978-1-4471-5360-3 978-1-4471-5361-0. doi: 10.1007/978-1-4471-5361-0.

H. J. Kushner. A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise. *Journal of Basic Engineering*, 86(1):97–106, March 1964. ISSN 0021-9223. doi: 10.1115/1.3653121.

Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. THE MNIST DATABASE of handwritten digits, 2010.

Yurii Evgen'evič Nesterov. *Lectures on Convex Optimization*. Springer Optimization and Its Applications; Volume 137. Springer, Cham, second edition edition, 2018. ISBN 978-3-319-91578-4. doi: 10.1007/978-3-319-91578-4.

Dmitry Panchenko. *The Sherrington-Kirkpatrick Model*. Springer Monographs in Mathematics. Springer, New York, NY, 2013. ISBN 978-1-4614-6288-0 978-1-4614-6289-7. doi: 10.1007/978-1-4614-6289-7.

Courtney Paquette, Bart van Merriënboer, Elliot Paquette, and Fabian Pedregosa. Halting Time is Predictable for Large Models: A Universality Property and Average-Case Analysis. *Foundations of Computational Mathematics*, 23(2):597–673, February 2022. ISSN 1615-3383. doi: 10.1007/s10208-022-09554-y.

Elliot Paquette and Thomas Trogdon. Universality for the Conjugate Gradient and MINRES Algorithms on Sample Covariance Matrices. *Communications on Pure and Applied Mathematics*, 76(5):1085–1136, September 2022. ISSN 1097-0312. doi: 10.1002/cpa.22081.

G. Parisi. A sequence of approximated solutions to the S-K model for spin glasses. *Journal of Physics A: Mathematical and General*, 13(4):L115, April 1980. ISSN 0305-4470. doi: 10.1088/0305-4470/13/4/009.

Razvan Pascanu, Yann N. Dauphin, Surya Ganguli, and Yoshua Bengio. On the saddle point problem for non-convex optimization, May 2014.

Fabian Pedregosa and Damien Scieur. Acceleration through spectral density estimation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7553–7562, Virtual Event (formerly Vienna), November 2020. PMLR.

Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, January 1964. ISSN 0041-5553. doi: 10.1016/0041-5553(64)90137-5.

Carl Edward Rasmussen and Christopher K.I. Williams. *Gaussian Processes for Machine Learning*. Number 3 in Adaptive Computation and Machine Learning. MIT Press, Cambridge, Massachusetts, 2 edition, 2006. ISBN 0-262-18253-X.

Tim Salimans and Durk P Kingma. Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks. In *Advances in Neural Information Processing Systems*, volume 29, Barcelona, Spain, 2016. Curran Associates, Inc.

Zoltán Sasvári. *Multivariate Characteristic and Correlation Functions*. Number 50 in De Gruyter Studies in Mathematics. Walter de Gruyter, Berlin/Boston, March 2013. ISBN 978-3-11-022399-6.

Michael Scheurer. *A Comparison of Models and Methods for Spatial Interpolation in Statistics and Numerical Analysis*. PhD thesis, Göttingen, 2009.

Isaac J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

Isaac J. Schoenberg. Positive definite functions on spheres. *Duke Mathematical Journal*, 9(1):96–108, March 1942. ISSN 0012-7094, 1547-7398. doi: 10.1215/S0012-7094-42-00908-6.

Damien Scieur and Fabian Pedregosa. Universal Average-Case Optimality of Polyak Momentum. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8565–8572, Virtual Event (formerly Vienna), November 2020. PMLR.

Mark Sellke. The Threshold Energy of Low Temperature Langevin Dynamics for Pure Spherical Spin Glasses. *Communications on Pure and Applied Mathematics*, 77(11), March 2024. doi: 10.1002/cpa.22197.

Michael L. Stein. *Interpolation of Spatial Data*. Springer Series in Statistics. Springer, New York, NY, 1999. ISBN 978-1-4612-7166-6 978-1-4612-1494-6. doi: 10.1007/978-1-4612-1494-6.

Michel Talagrand. The Parisi Formula. *Annals of Mathematics*, 163(1):221–263, 2006. ISSN 0003-486X.

Greg Yang. Tensor Programs III: Neural Matrix Laws, May 2021.

Appendix A. Random quadratic functions are isotropic

The goal of this section is twofold: First, we want to motivate that the cost functions found in machine learning should be expected to be non-stationary. Second, we want to demonstrate how random linear models result in (isotropic) random quadratic functions. This suggests that the isotropy assumption may be plausible.

Subject to strong simplifying assumptions, p-layer neural networks may also be related to p-spin glasses (1) (Choromanska et al., 2015), i.e. isotropic random functions on the sphere. Much more realistic assumptions are sufficient to relate linear models to random quadratic functions. More specifically, the random functions considered by Paquette et al. (2022), etc. are of the form

$$\mathbf{f}_N(x) = \frac{1}{2n}\|\mathbf{A}(x-\tilde{\mathbf{x}})+\eta\|^2 = \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{A}_{i,\cdot}(x-\tilde{\mathbf{x}})+\eta_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{f}_N^{(i)}(x),$$

where $\mathbf{A} \in \mathbb{R}^{n\times N}$ is a random data matrix, $\tilde{\mathbf{x}} \in \mathbb{R}^N$ is a random vector representing the true signal and η is noise. The last equation shows how $\mathbf{f}_N$ can be decomposed into n stochastic losses $\mathbf{f}_N^{(i)}$, where n is the amount of data.

Lemma 36 (Quadratic function representation) *Assume* $\mathbf{A}$, $\tilde{\mathbf{x}}$ *and* η *to be independent, where the noise* η *itself is a vector of iid variables with zero mean and variance* σ_η^2.

Further assume the isotropic feature model of Paquette et al. (2022), i.e. the entries of $\mathbf{A}$ *are iid with zero mean and variance* $\sigma_{\mathbf{A}}^2$*. Then* $\mathbf{f}_N$ *can be represented as*

$$\mathbf{f}_N(x) = \mathbf{f}_N^\infty(x) + \frac{1}{n}\sum_{i=1}^{n} \underbrace{(\mathbf{f}_N^{(i)}(x) - \mathbf{f}_N^\infty(x))}_{=:\epsilon_i(x)} \quad \textit{with} \quad \mathbf{f}_N^\infty(x) = \tfrac{\sigma_{\mathbf{A}}^2}{2}\|x-\tilde{\mathbf{x}}\|^2 + \tfrac{\sigma_\eta^2}{2},$$

where $\mathbf{f}_N^\infty$ *is the 'infinite data limit' and* ϵ_i *is the stochastic noise, which is identically distributed. Conditional on the true signal* $\tilde{\mathbf{x}}$*, the* ϵ_i *are also independent with zero mean. The noise* ϵ_i *is then also unconditionally uncorrelated and has zero mean.*

Proof Before confirming it has the desired properties, we simply define the infinite data limit to be

$$\begin{aligned}
\mathbf{f}_N^\infty(x) &:= \mathbb{E}[\mathbf{f}_N^{(i)}(x) \mid \tilde{\mathbf{x}}] = \mathbb{E}\Big[\tfrac{1}{2}(x-\tilde{\mathbf{x}})^T\mathbf{A}_{i,\cdot}^T\mathbf{A}_{i,\cdot}(x-\tilde{\mathbf{x}}) + \mathbf{A}_{i,\cdot}(x-\tilde{\mathbf{x}})\eta_i + \tfrac{1}{2}\eta_i^2 \mid \tilde{\mathbf{x}}\Big] \\
&= \tfrac{1}{2}(x-\tilde{\mathbf{x}})^T \underbrace{\mathbb{E}\big[\mathbf{A}_{i,\cdot}^T\mathbf{A}_{i,\cdot}\big]}_{\sigma_{\mathbf{A}}^2\mathbb{I}}(x-\tilde{\mathbf{x}}) + \tfrac{1}{2}\underbrace{\mathbb{E}\big[\eta_i^2\big]}_{\sigma_\eta^2} \\
&= \tfrac{\sigma_{\mathbf{A}}^2}{2}\|x-\tilde{\mathbf{x}}\|^2 + \tfrac{\sigma_\eta^2}{2}.
\end{aligned}$$

By the independence of $\mathbf{A}$, $\tilde{\mathbf{x}}$, and η and the independence of the entries of $\mathbf{A}$, conditionally on $\tilde{\mathbf{x}}$, the $\mathbf{f}_N^{(i)}(x)$ are independent. Since $\mathbf{f}_N^\infty$ is deterministic, conditional on $\tilde{\mathbf{x}}$ by definition, the ϵ_i are therefore independent, conditionally on $\tilde{\mathbf{x}}$. That the conditional mean is zero follows from our definition of $\mathbf{f}_N^\infty$. The unconditional statements follow from the tower property of the conditional expectation. ■

Since we do not consider stochastic optimization algorithms (like stochastic gradient descent) but full gradient optimization in this paper, a comparison requires a comparison with the underlying infinite data limit $\mathbf{f}_N^\infty$, which can be rewritten as

$$\mathbf{f}_N^\infty(x) = \underbrace{\tfrac{\sigma_\eta^2}{2} + \tfrac{\sigma_{\mathbf{A}}^2}{2}\big(\mathbb{E}[\|\tilde{\mathbf{x}}\|^2] + \|x\|^2\big)}_{\text{deterministic}} + \sigma_{\mathbf{A}}^2\langle x, \tilde{\mathbf{x}}\rangle + \underbrace{\tfrac{\sigma_{\mathbf{A}}^2}{2}\big(\|\tilde{\mathbf{x}}\|^2 - \mathbb{E}[\|\tilde{\mathbf{x}}\|^2]\big)}_{\text{random constant}}.$$

We are now going to assume that the signal $\tilde{\mathbf{x}}$ is centered (which represents a translation of the coordinate system such that $\mathbb{E}[\tilde{\mathbf{x}}]$ is the origin). Let us assume assume, like Paquette et al. (2022, Assumption 1), that the entries $\tilde{\mathbf{x}}^{(i)}$ are independent with variance $\frac{R^2}{N}$. The law of large numbers then implies

$$\tfrac{\sigma_{\mathbf{A}}^2}{2}\big(\|\tilde{\mathbf{x}}\|^2 - \mathbb{E}[\|\tilde{\mathbf{x}}\|^2]\big) \underset{N\to\infty}{\overset{p}{\to}} 0.$$

Hence, the random constant disappears in the high-dimensional limit. Since constants are also irrelevant for optimization there are two good reasons why this random constant can be dropped without loss of generality when analyzing optimization on $\mathbf{f}_N^\infty$. But then, assuming $\tilde{\mathbf{x}}$ to be Gaussian, the infinite data limit

$$\hat{\mathbf{f}}_N^\infty(x) := \mu\big(\tfrac{\|x\|^2}{2}\big) + \sigma_{\mathbf{A}}^2\langle x, \tilde{\mathbf{x}}\rangle, \quad \text{with} \quad \mu\big(\tfrac{\|x\|^2}{2}\big) := \tfrac{\sigma_\eta^2}{2} + \tfrac{\sigma_{\mathbf{A}}^2}{2}\big(\mathbb{E}[\|\tilde{\mathbf{x}}\|^2] + \|x\|^2\big),$$

is a (non-stationary) isotropic GRF with covariance

$$\mathsf{Cov}(\hat{\mathbf{f}}_N^\infty(x), \hat{\mathbf{f}}_N^\infty(y)) = \sigma_{\mathbf{A}}^4 \tfrac{R^2}{N}\langle x, y\rangle =: \tfrac{1}{N}\kappa\big(\tfrac{\|x\|^2}{2}, \tfrac{\|y\|^2}{2}, \langle x, y\rangle\big).$$

Random quadratic functions with isotropic features are therefore covered by the (non-stationary) isotropy assumption of the present article. The additional assumption we required over Paquette et al. (2022) was that the mean of $\tilde{\mathbf{x}}$ had to be zero. By a change of coordinate system this essentially implies the origin of the space necessarily has to be the expectation of $\tilde{\mathbf{x}}$.

Remark 37 *The 'correlated feature model' of Paquette et al. (2022) is essentially a change of space. That is, (non-stationary) isotropy with respect to the Hilbertspace $\mathbb{R}^N$ equipped with the inner product*

$$\langle x, y\rangle_\Sigma := \langle x, \Sigma y\rangle$$

results in the correlated feature model. This geometric modification of (non-stationary) isotropy is well known in the stationary isotropic case under the term 'geometric anisotropy' (e.g. Stein, 1999, p. 17).

Appendix B. Conditional Gaussian distributions

Theorem 38 about conditional Gaussian distributions is well known (e.g. Eaton, 2007, Prop. 3.13.), but for your convenience we wrote down our favorite proof of this theorem capturing the intuition behind the statement.

Theorem 38 (Conditional Gaussian distribution) *Let $X \sim \mathcal{N}(\mu, \Sigma)$ be a multivariate Gaussian vector where the covariance matrix is a block matrix of the form*

$$\mu = \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} \quad \textit{and} \quad \Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix},$$

then the conditional distribution of X_2 given X_1 is

$$X_2 \mid X_1 \sim \mathcal{N}(\mu_{2|1}, \Sigma_{2|1}),$$

with conditional mean and variance

$$\begin{aligned}\mu_{2|1} &:= \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1)\\ \Sigma_{2|1} &:= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.\end{aligned}$$

where Σ_{11}^{-1} may be a generalized inverse (cf. Eaton, 2007).

Proof For the general statement we point to the standard literature (e.g. Eaton, 2007). For this proof we will assume for simplicity that Σ is invertible.

Let $\bar{X} := X - \mu$ be the centered version of X. Then there exists a unique lower triangular matrix L such that $\Sigma = LL^T$ (i.e. the Cholesky Decomposition). This results in

the following representation[5]

$$X - \mu =: \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = LY$$

with independent standard normal Y_i, i.e. $Y \sim \mathcal{N}(0, \mathbb{I})$. L_{11} is invertible since Σ_{11} and therefore the map from Y_1 to X_1 is bijective. Conditioning on X_1 is therefore equivalent to conditioning on Y_1. But we have

$$X_2 = \mu_2 + \bar{X}_2 = \underbrace{\mu_2 + L_{21}Y_1}_{\text{conditional expectation}} + \underbrace{L_{22}Y_2}_{\text{conditional distribution}} \tag{55}$$

So it follows that

$$X_2 \mid X_1 \sim \mathcal{N}(\mu_{2|1}, \Sigma_{2|1})$$

with

$$\begin{aligned} \mu_{2|1} &:= \mathbb{E}[X_2 \mid X_1] && = \mu_2 + L_{21}Y_1 \\ \Sigma_{2|1} &:= \mathsf{Cov}[X_2 \mid X_1] = \mathbb{E}\Big[\big(X_2 - \mathbb{E}[X_2 \mid X_1]\big)^2 \mid X_1\Big] && = L_{22}L_{22}^T. \end{aligned}$$

What is left to do, is to find a representation for the L_{ij} using the block matrices of Σ. For this we observe

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \Sigma = LL^T = \begin{bmatrix} L_{11}L_{11}^T & L_{11}L_{21}^T \\ L_{21}L_{11}^T & L_{22}L_{22}^T + L_{21}L_{21}^T \end{bmatrix}. \tag{56}$$

Using $Y_1 = L_{11}^{-1}\bar{X}_1$ and the insertion of an identity matrix this implies

$$L_{21}Y_1 = L_{21}(L_{11}^T L_{11}^{-T})(L_{11}^{-1}\bar{X}_1) \overset{(56)}{=} \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1).$$

resulting in the desired conditional expectation $\mu_{2|1}$. The conditional variance follows alike

$$\begin{aligned} \Sigma_{2|1} = L_{22}L_{22}^T &\overset{(56)}{=} \Sigma_{22} - L_{21}L_{21}^T \\ &= \Sigma_{22} - \underbrace{L_{21}(L_{11}^T}_{\overset{(56)}{=}\Sigma_{21}} \underbrace{L_{11}^{-T})(L_{11}^{-1}}_{\overset{(56)}{=}\Sigma_{11}^{-1}} \underbrace{L_{11})L_{21}^T}_{\overset{(56)}{=}\Sigma_{12}}. \end{aligned}$$

■

5. It is straightforward to check that $Y := L^{-1}(X - \mu)$ is a multivariate Gaussian vector of centered, uncorrelated entries with variance 1. They are therefore iid standard normal since uncorrelated multivariate Gaussian random variables are independent. Sometimes this representation is even taken as the definition of a multivariate normal distribution. This is the only real place we use the invertibility of Σ in its entirety, the invertibility of Σ_{11}^{-1} is used later on because we do not want to get into the business of generalized inverses here.

Remark 39 (Decomposition) *$X_2 \mid X_1 \sim \mathcal{N}(\mu_{2|1}, \Sigma_{2|1})$ always implies that there exists there exists $Y_2 \sim \mathcal{N}(0, \mathbb{I})$ independent of X_1 such that the following equality is true in distribution*

$$X_2 = \mu_{2|1} + \sqrt{\Sigma_{2|1}} Y_2, \tag{57}$$

where $\sqrt{\Sigma}$ denotes the cholesky decomposition of Σ. If the covariance matrix is moreover invertible, then Y_2 can be determined constructively as in the proof of Theorem 38 such that equation holds always and not just in distribution (cf. (55)). This might be true in the non-invertible case as well but would require deeper insights about singular matrices.

Appendix C. Strict positive definite derivatives

While covariance matrices are always positive definite, they are not always strict positive definite (i.e. invertible). Recall, that a random function $\mathbf{f}$ and its covariance function $\mathcal{C}_\mathbf{f}$ are *strict* positive definite if the matrix $(\mathcal{C}_\mathbf{f}(x_k, x_l))_{k,l=1,\dots,n}$ is strict positive definite for any finite $x_1, \dots, x_n$. It is already known that, whenever $\mathbf{f}$ is stationary isotropic and non-constant, $\mathbf{f}$ is strict positive definite if the dimension satisfies $N \geq 2$ (Sasvári, 2013, Theorem 3.1.5). But since we require the 'jet' $\mathrm{J}^1\mathbf{f} = (\mathbf{f}, \nabla\mathbf{f})$ to be strict positive definite in Theorem 12, we prove a generalization of Theorem 3.1.6 in Sasvári (2013). This generalization will be sufficient to prove that all stationary isotropic covariance functions which are valid in all dimensions (cf. Definition 8) are covered as we will see in Corollary 42. This result will be used to remove the strict positive definiteness assumption in Corollary 9.

If $\mathbf{g}$ is a random function with multivariate output, then $\mathcal{C}_\mathbf{g}(x_k, x_j)$ is already a matrix for fixed k, j and the collection over k, j is really a tensor. To avoid introducing this machinery and explaining positive definiteness for tensors, we will take the following equivalent statement for positive definiteness of random functions as definition.

Definition 40 (Strict positive definite random function) *The covariance $\mathcal{C}_\mathbf{g}$ of a random function $\mathbf{g} : \mathbb{R}^N \to \mathbb{R}^m$ and the random function itself is called* strict positive definite *if for all $w_k \in \mathbb{R}^m$ and distinct $x_1, \dots, x_n \in \mathbb{R}^N$ the equality*

$$0 = \mathsf{Var}\Big[\sum_{k=1}^n w_k^T \mathbf{g}(x_k)\Big] \overset{(58)}{=} \sum_{k,l=1}^n w_k^T \mathcal{C}_\mathbf{g}(x_k, x_l) w_l$$

implies $w_k = 0$ for all k.

Note, that the second equality marked with (58) is always true and only represents an equivalent formulation, because after centering $\mathbf{g}$ (without loss of generality, using $\tilde{\mathbf{g}} = \mathbf{g} - \mathbb{E}[\mathbf{g}]$), we have

$$\sum_{k,l=1}^n w_k^T \mathcal{C}_\mathbf{g}(x_k, x_l) w_l = \mathbb{E}\Big[\sum_{k,l=1}^n w_k^T \mathbf{g}(x_k) \mathbf{g}(x_l)^T w_l\Big] = \mathsf{Var}\Big[\sum_{k=1}^n w_k^T \mathbf{g}(x_k)\Big]. \tag{58}$$

Theorem 41 (Strict positive definite derivatives) *Let $\mathbf{f} \colon \mathbb{R}^N \to \mathbb{R}$ be a stationary random function, i.e. its covariance has the form $\mathcal{C}_\mathbf{f}(x, y) = C(x - y)$. Assume the positive definite function C is continuous such that the support of its spectral measure contains*

a non-empty open set. Then up to any order k up to which $\mathbf{f}$ is almost surely differentiable, the jet

$$\mathrm{J}^k\mathbf{f} = (\mathbf{f}, \nabla\mathbf{f}, (\partial_{ij}\mathbf{f})_{i\leq j}, \ldots, (\partial_{i_1\cdots i_k}\mathbf{f})_{i_1\leq\cdots\leq i_k})$$

is strict positive definite.

Theorem 41 is not more general than Sasvári (2013, Theorem 3.1.6) in the sense that conditions are weakend, but it is more general in its implication. That is, $\mathbf{g}$ is strict positive definite and not just $\mathbf{f}$.

The following corollary shows that this fully covers the stationary isotropic covariance functions valid in all dimensions (cf. Definition 8). We use this result about stationary isotropic random functions in Corollary 9 to omit the assumption that $(\mathbf{f}, \nabla\mathbf{f})$ has to be strict positive definite, which is needed for our general result (Theorem 12).

Corollary 42 *Assume that C is a continuous stationary isotropic covariance kernel valid in all dimensions, i.e. defined on the space of square summable sequences ℓ^2 by Lemma 2.3 in Benning and Schölpple (2025). Let $\mathbf{f} \sim \mathcal{N}(\mu, C)$ be a Gaussian random function, which is not almost surely constant. Then the jet*

$$\mathrm{J}^k\mathbf{f} = (\mathbf{f}, \nabla\mathbf{f}, (\partial_{ij}\mathbf{f})_{i\leq j}, \ldots, (\partial_{i_1\cdots i_k}\mathbf{f})_{i_1\leq\cdots\leq i_k})$$

is strict positive definite for any $k \in \mathbb{N}$.

Proof [Proof of Theorem 41] For any n finite and distinct $x_1, \ldots, x_n \in \mathbb{R}^N$ we need that

$$0 = \mathsf{Var}\Big[\sum_{j=1}^n w_j^T \mathrm{J}^k\mathbf{f}(x_j)\Big]$$

implies $w_i = 0$ for all $i = 1, \ldots, n$. Without loss of generality we assume $\mathbf{f}$ (and thus $\mathrm{J}^k\mathbf{f}$) to be centered. We can then rewrite this as

$$\mathsf{Var}\Big[\sum_{j=1}^n w_j^T \mathrm{J}^k\mathbf{f}(x_j)\Big] = \mathsf{Var}\Big[\sum_{j=1}^n \sum_{l=0}^k \sum_{i_1\leq\cdots\leq i_l} w_j^{(l,i_1,\ldots,i_l)} \partial_{i_1\ldots i_l}\mathbf{f}(x_j)\Big],$$

with appropriate indexing of the w_j. Note that l ranges over the order of differentiation contained in $\mathbf{g}$ before we sum over all the partial derivatives in the inner sum. For $k = 1$ this is for example

$$\mathsf{Var}\Big[\sum_{j=1}^n w_j^{(0)}\mathbf{f}(x_j) + \sum_{i=1}^N w_j^{(1,i)}\partial_i\mathbf{f}(x_j)\Big] = \mathsf{Var}\Big[\sum_{j=1}^n w_j^{(0)}\mathbf{f}(x_j) + D_{w_j^{(1,\cdot)}}\mathbf{f}(x_j)\Big] = 0.$$

We now consider the linear differential operator

$$T := \sum_{j=1}^n \sum_{l=0}^k \sum_{i_1\leq\cdots\leq i_l} (-1)^l w_j^{(l,i_1,\ldots,i_l)} \partial_{i_1\ldots i_l}\delta_{x_j},$$

where $\delta_x(f) = f(x)$ is the dirac delta function and $\partial_i \delta_x(f) = -\partial_i f(x)$ is its derivative in the sense of distributions. Recall this means for test functions ϕ, that we have

$$\langle \partial_i \delta_x, \phi \rangle = -\langle \delta_x, \partial_i \phi \rangle = -\partial_i \phi(x).$$

The higher order derivatives in the sense of distributions are similarly defined. Using this operator we now obtain more succinctly

$$0 = \mathsf{Var}\Big[\sum_{j=1}^{n} \sum_{l=0}^{k} \sum_{i_1 \le \cdots \le i_l} w_j^{(l,i_1,\ldots,i_l)} \partial_{i_1 \ldots i_l} \mathbf{f}(x_j)\Big] = \mathsf{Var}[T\mathbf{f}].$$

While $T\mathbf{f}$ is well defined, we now want to move this operator outside. For this we want to use the bilinearity of the covariance. But the covariance has two inputs and $T\mathcal{C}_\mathbf{f}$ is then not well defined because the differential operator T might be applied to the first or second input of $\mathcal{C}_\mathbf{f}$. To avoid this issue, we define with some abuse of notation $T^t f(t) := Tf$ such that we can write $T^t \mathcal{C}_\mathbf{f}(t,s)$ when we mean $T\mathcal{C}_\mathbf{f}(\cdot, s)$. Note that T is a linear combination of basis elements $\partial_{i_1 \ldots i_l} \delta_x$. So by bilinearity of the covariance it is sufficient to check, that we can move these basis elements out of the covariance, i.e.

$$\begin{aligned}\mathsf{Cov}(\partial_{i_1 \ldots i_l} \delta_x \mathbf{f}, \partial_{j_1 \ldots j_{l'}} \delta_y \mathbf{f}) &= \mathsf{Cov}(\partial_{i_1 \ldots i_l} \mathbf{f}(x), \partial_{j_1 \ldots j_{l'}} \mathbf{f}(y)) \\ &= (\partial_{i_1 \ldots i_l} \delta_x)^t \, (\partial_{j_1 \ldots j_{l'}} \delta_y)^s \mathcal{C}_\mathbf{f}(t,s).\end{aligned}$$

Since T is a linear combinations of these, we get by the bilinearity of the covariance

$$0 = \mathsf{Var}[T\mathbf{f}] = T^x T^y \mathcal{C}_\mathbf{f}(x,y).$$

In the remainder of the proof we will essentially show, that this variance can be represented as an integral over the absolute value of the fourier transform of T with respect to the spectral measure of $\mathcal{C}_\mathbf{f}$. This forces the fourier transform of T and therefore T to be zero.

Using the spectral representation of $\mathcal{C}_\mathbf{f}$ (e.g. Sasvári, 2013, Theorem 1.7.4) given by

$$\mathcal{C}_\mathbf{f}(x,y) = \int e^{i\langle x-y,t\rangle} \sigma(dt),$$

we move the operator into the integral (by moving sums and derivatives into the integral)

$$0 = T^x T^y \mathcal{C}_\mathbf{f}(x,y) = \int T^x e^{i\langle x,t\rangle} T^y e^{-i\langle y,t\rangle} \sigma(dt) = \int |T e^{i\langle \cdot,t\rangle}| \sigma(dt),$$

where we use $T\bar{f} = \overline{Tf}$ for the last equation, which follows from

- $\delta_x(\overline{f}) = \overline{\delta_x(f)}$
- $D_v \delta_x(\overline{f}) = D_v \overline{f(x)} = \overline{D_v f(x)} = \overline{D_v \delta_x(f)}$ because from $|z| = |\overline{z}|$ follows

$$\lim_{t \to 0} \frac{|\overline{f(x+tv)} - \overline{f(x)} - \overline{D_v f(x)} t|}{t} = 0,$$

- induction for higher order derivatives.

So we have that $P(t) := Te^{i\langle \cdot, t\rangle}$ must be zero σ-almost everywhere. Since

$$P(t) = \sum_{j=1}^{n} \sum_{l=0}^{k} \sum_{i_1 \leq \cdots \leq i_l} w_j^{(l,i_1,\ldots,i_l)} \partial_{i_1 \ldots i_l} e^{i\langle x_j, t\rangle}$$

is continuous in t, it can only be zero σ-almost everywhere if it is zero on the support of σ. Since the support of the spectral measure σ contains an open subset by assumption of the theorem, P must be zero on this open subset. But then P must be zero everywhere as an analytic function. As $P(t)$ is the fourier transform of T in the sense of distributions, i.e.

$$P(t) = \int \Big(\sum_{j=1}^{n} \sum_{l=0}^{k} \sum_{i_1 \leq \cdots \leq i_l} w_j^{(l,i_1,\ldots,i_l)} \partial_{i_1 \ldots i_l} \delta_{x_j} \Big)(x) e^{i\langle x, t\rangle} dx = \mathcal{F}[T](t),$$

this requires T to be zero by linearity and invertibility of the Fourier transform. But since the $\partial_{i_1 \ldots i_l} \delta_{x_j}$ are linear independent[6] for distinct x_j the only way T can be zero is, if all the w_j are zero. This finishes the proof. ■

The general requirement, the existence of a non-empty open sets in the support of the spectral measure, is satisfied for the stationary isotropic random functions valid in all dimensions (Corollary 42). To prove this result we require the following lemma. It proves that stationary isotropic random function in ℓ^2 is either almost surely constant or the 'Schoenberg measure' has positive measure on $(0, \infty)$. Where the Schoenberg measure ν refers to the measure in Schoenberg's characterization of stationary isotropic covariance kernels on ℓ^2 given by

$$\mathcal{C}_{\mathbf{f}}(x, y) = C(\tfrac{\|x-y\|^2}{2}) = \int_{[0,\infty)} \exp(-t^2 \tfrac{\|x-y\|^2}{2}) \nu(dt) \tag{59}$$

Lemma 43 (Constant random functions) *Let C be a stationary isotropic covariance kernel valid in all dimensions and let $\mathbf{f} \sim \mathcal{N}(\mu, C)$ be a continuous stationary isotropic random function.*

If the Schoenberg measure ν of C has no mass on $(0, \infty)$, i.e. $\nu((0, \infty)) = 0$, then $\mathbf{f}$ is almost surely constant.

6. To see the linear independence of $\partial_{i_1 \ldots i_l} \delta_{x_j}$, consider that the finite x_k are distinct, so they have a minimal distance. Rescale a bump function with zero slope at the top, e.g.

$$\phi(x) = \begin{cases} \exp\Big(-\frac{1}{1-\|x\|^2}\Big) & \|x\| < 1 \\ 0 & \|x\| \geq 1 \end{cases}$$

such that it is centered on some x_j and it is zero at all other x_k (and zero at all derivatives). This implies

$$0 = \langle T, \phi \rangle = w_j^{(0)} \phi(x_j)$$

and thus $w_j^{(0)} = 0$. Then construct similar test functions to ensure the other prefactors have to be zero, by placing a non-zero slope at x_j while ensuring it is zero at all other x_k.

Proof By Schoenberg's characterization (59) we have

$$C(r) = \int_{[0,\infty)} \exp(-t^2 r)\nu(dt) \overset{\nu((0,\infty)=0)}{=} \nu(\{0\}).$$

This implies a constant covariance

$$\mathsf{Cov}(\mathbf{f}(x), \mathbf{f}(y)) = \nu(\{0\}).$$

Assuming $\mathbf{f}$ to be centered (without loss of generality) by switching to $\tilde{\mathbf{f}}_N = \mathbf{f} - \mu$, this implies

$$\mathbb{E}[(\mathbf{f}(x) - \mathbf{f}(y))^2] = \mathbb{E}[\mathbf{f}(x)^2] - 2\mathbb{E}[\mathbf{f}(x)\mathbf{f}(y)] + \mathbb{E}[\mathbf{f}(y)^2] = 0.$$

Thus $\mathbf{f}(x) = \mathbf{f}(y)$ almost surely for all x, y. Via the union over a dense countable subset of $\mathbb{R}^N$ and a.s. continuity of $\mathbf{f}$ we get

$$\mathbb{P}(\mathbf{f} \text{ a.s. constant}) = \mathbb{P}\Big(\mathbf{f}(x) = \mathbf{f}(y) \quad \forall x, y \in \mathbb{R}^N\Big) = 1.$$

■

Using this lemma, we can prove that all stationary isotropic covariance kernels on ℓ^2 have strictly positive definite derivatives.

Corollary 44 *Assume that C is a continuous stationary isotropic covariance kernel valid in all dimensions, i.e. defined on the space of square summable sequences ℓ^2 by Lemma 2.3 in Benning and Schölpple (2025). Let $\mathbf{f} \sim \mathcal{N}(\mu, C)$ be a Gaussian random function, which is not almost surely constant. Then the jet*

$$\mathrm{J}^k\mathbf{f} = (\mathbf{f}, \nabla\mathbf{f}, (\partial_{ij}\mathbf{f})_{i\leq j}, \ldots, (\partial_{i_1\cdots i_k}\mathbf{f})_{i_1\leq\cdots\leq i_k})$$

is strict positive definite for any $k \in \mathbb{N}$.

Proof [Proof of Corollary 42] Since $\psi(x) = e^{-\frac{\|x\|^2 t^2}{2}}$ is the characteristic function of $Y_t \sim \mathcal{N}(0, t\mathbb{I}_{N\times N})$, Schoenbergs characterization (59) of the covariance implies up to scaling

$$\begin{aligned}
\mathcal{C}_{\mathbf{f}}(x, \tilde{x}) = C\big(\tfrac{\|x-\tilde{x}\|^2}{2}\big) &= \int_{[0,\infty)} e^{-\frac{t^2\|x-\tilde{x}\|^2}{2}}\nu(dt) \\
&= \int_{[0,\infty)} \mathbb{E}[e^{i\langle x-\tilde{x}, Y_t\rangle}]\nu(dt) \\
&= \int e^{i\langle x-\tilde{x}, y\rangle} \underbrace{\int_{(0,\infty)} e^{-\frac{\|y\|^2}{2t}}\nu(dt)}_{=:\varphi_\sigma(y)} dy + \nu(\{0\}) \\
&= \int e^{i\langle x-\tilde{x}, y\rangle}\sigma(dy)
\end{aligned}$$

with spectral measure

$$\sigma(A) := \int_A \varphi_\sigma(y)dy + \tfrac{1}{N}\nu\{0\}\delta_0(A).$$

Since $\nu((0,\infty)) \neq 0$ because $\mathbf{f}$ is not almost surely constant (Corollary 43), its density φ_σ is continuous and positive $\varphi_\sigma(y) > 0$ for all $y \in \mathbb{R}^N$ and thus $\operatorname{supp}(\sigma) = \mathbb{R}^N$. In particular Theorem 41 is applicable and finishes the proof. ■

Appendix D. Technical results

Recall that $\kappa_3 > 0$ was used in (34) to prove that the gradient has positive length. This fact follows from strict positive definiteness, which we are now going to prove.

Lemma 45 *Let κ be a (non-stationary) isotropic covariance kernel valid in all dimensions. Let $\mathbf{f} \sim \mathcal{N}(\mu, \kappa)$ and assume $(\mathbf{f}, \nabla\mathbf{f})$ to be strict positive definite, then for any $x \in \mathbb{R}^N$*

$$\kappa_3\big(\tfrac{\|x\|^2}{2}, \tfrac{\|x\|^2}{2}, \|x\|^2\big) > 0.$$

Proof Since the norm $\|\cdot\|^2$ is rotation invariant, we can assume without loss of generality $x = \lambda e_1$ for the standard basis vector $e_1 \in \mathbb{R}^N$. Since $\partial_2\mathbf{f}$ is strict positive definite, we know that

$$0 < \mathsf{Var}(\partial_2\mathbf{f}(x)) = \mathsf{Cov}(\partial_2\mathbf{f}(\lambda e_1), \partial_2\mathbf{f}(\lambda e_1)) = \tfrac{1}{N}\kappa_3\big(\tfrac{\|x\|^2}{2}, \tfrac{\|x\|^2}{2}, \|x\|^2\big),$$

where we have used (18) of Lemma 20 and $\|\lambda e_1\| = \|x\|$ in the last equation. ■

In the stationary isotropic case, we can say more. We do not only have $\kappa_3 = -C'(0) > 0$, but we have $C'(r) < 0$ for all $r \geq 0$.

Lemma 46 *Let C be a stationary isotropic covariance kernel valid in all dimensions. If $\mathbf{f} \sim \mathcal{N}(\mu, C)$ is* not *almost surely constant, then $C'(r) < 0$ for all $r \geq 0$.*

Proof By Corollary 43 we have $\nu((0,\infty)) > 0$ for the Schoenberg measure ν in Schoenbergs's characterization (59). Thus by the continuity of measures there exists $a, b > 0$ such that $\nu([a,b]) > 0$. Then by (59) we have

$$\begin{aligned} -C'(r) &= \int_{[0,\infty)} t^2 \exp(-t^2 r)\nu(dt) \\ &\geq \int_{[a,b]} t^2 \exp(-t^2 r)\nu(dt) \\ &\geq \nu([a,b]) \inf_{s\in[a,b]} s^2 \exp(-s^2 r) > 0. \end{aligned}$$

■