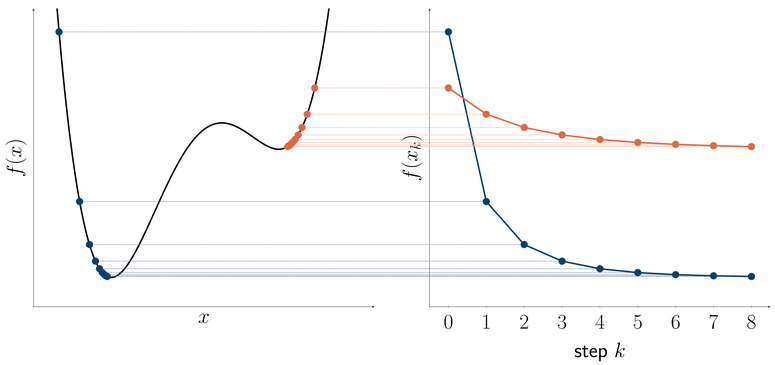
Gradient Span Algorithms Make Predictable Progress in High Dimension

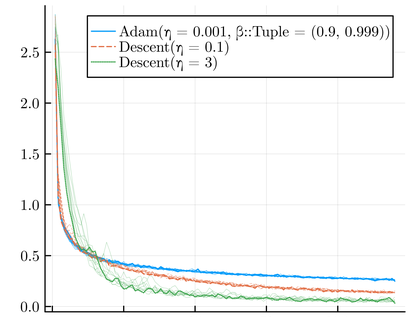
Felix Benning [1,2] Leif Döring [1]

[1]University of Mannheim [2]University of Luxembourg

Low dimensional optimization intuition clashes with observed behavior in machine learning



(a) Low dimensional intuition: Depending on the initialization, gradient descent starts at different heights and converges with different rates to different minima.



(b) Actual training behavior of a 2.3 Mio parameter CNN on the MNIST dataset over 10 random initializations over 120 steps with batch size 1000.

Figure 1. Contrary to the low dimensional intuition, optimization trajectories in high dimensional machine learning optimization problems start at roughly the same height and converge to roughly the same minimum at the same rate.

Gradient span algorithm (simplified)

Gradient span algorithms select the subsequent iterate from the span of previous gradients, that is

$$x_{n+1} \in x_0 + \operatorname{span}\{\nabla f(x_0), \dots, \nabla f(x_n)\}$$

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

More specifically a gradient span algorithm chooses

$$x_{n+1} = x_0 + \sum_{k=0}^{n} h_{n,k} \nabla f(x_k),$$

where the $h_n = (h_{n,k})_{k=0}^n$ are chosen as continuous functions $h_n(I_{n-1}(f))$ of the previous *dimensionless* information $I_{n-1}(f)$ defined as

$$I_n(f) \coloneqq \{f(x_k), k \in [0{:}n]\} \cup \{\langle \nabla f(x_i), \nabla f(x_j)\rangle, i,j \in [0{:}n]\}.$$

Isotropic Gaussian random function

A random function (stochastic process) $(\mathbf{f}(x))_{x\in\mathbb{R}^d}$ is called isotropic if for any orthogonal transformation $U \in O(d)$

$$\mathsf{P}_{\mathbf{f}} = \mathsf{P}_{\mathbf{f}\circ U}.$$

For a Gaussian random function this is equivalent to the mean $\mu_{\mathbf{f}}(x) = \mathsf{E}[\mathbf{f}(x)]$ and covariance $\mathcal{C}_{\mathbf{f}}(x,y) = \mathsf{Cov}(\mathbf{f}(x), \mathbf{f}(y))$ having the form

$$\mu_{\mathbf{f}}(x) = \mu\big(\|x\|^2\big), \qquad \mathcal{C}_{\mathbf{f}}(x,y) = \tfrac{1}{d}\kappa\big(\|x\|^2, \|y\|^2, \langle x,y\rangle\big),$$

Example: A *stationary isotropic* Gaussian random functions has mean and covariance

$$\mu_{\mathbf{f}} = \mu \in \mathbb{R}, \qquad \mathcal{C}_{\mathbf{f}}(x,y) = \tfrac{1}{d}C(\|x-y\|^2),$$

Smoothness and derivatives: The smoothness of a Gaussian random function is intertwined with the smoothness of its covariance function:

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

Essentially: A derivative of $\mathbf{f}$ corresponds to two derivatives of $\mathcal{C}_{\mathbf{f}}$.

Intuition of dimensional scaling

Observe that

$$\|\nabla \mathbf{f}(x)\|^2 = d\overbrace{\frac{1}{d}\sum_{i=1}^{d}\big(\partial_{x_i}\mathbf{f}(x)\big)^2}^{\text{Law of large numbers}} \asymp d\,\mathsf{E}[(\partial_{x_1}\mathbf{f}(x))^2], \tag{1}$$

if the partial derivatives $\partial_{x_i}\mathbf{f}(x)$ are identically distributed. They are in the isotropic case! For simplicity assume the stationary case, then

$$\begin{aligned}\partial_{x_i}\partial_{y_j}\mathcal{C}_{\mathbf{f}}(x,y) &= \partial_{x_i} C'(\|y-x\|^2)2\langle y-x, e_j\rangle, \\ &= -2C''(\|y-x\|^2)(y-x)_i(y-x)_j - 2C'(\|x-y\|^2)\delta_{ij},\end{aligned} \tag{2}$$

Consequently, for $x = y$ we have

$$\mathsf{Cov}(\partial_{x_i}\mathbf{f}(x), \partial_{x_j}\mathbf{f}(x)) = -2C'(0)\delta_{ij}.$$

Since $\partial_{x_i}\mathbf{f}(x)$ are Gaussian as limits of Gaussian difference quotients, them being uncorrelated implies they are independent.

Intuitively: In d-dimensional space, $d-1$ directions are orthogonal to the gradient. So the function is flat in almost all directions. Picking any random direction or point leaves us at the mean.

Main result (simplified)

Theorem 12 (Predictable progress). Let $(\mathbf{f}_d(x))_{x\in\mathbb{R}^d}$ be a sequence of sufficiently smooth isotropic Gaussian random functions with isotropic mean function $\mu\colon [0,\infty) \to \mathbb{R}$ and covariance function $\kappa : \mathbb{X} \to \mathbb{R}$ with $\mathbb{X} = \{(r,s,t) \in [0,\infty)^2 \times \mathbb{R} : t^2 \le rs\}$. And assume that a sequence of evaluation points $(x_n)_{n\in\mathbb{N}}$ is chosen by a gradient span algorithm $\mathfrak{G}$. Then, subject to mild technical assumptions,

$$I_n(\mathbf{f}_d) \xrightarrow{p} \mathscr{I}_n(\mathfrak{G}, \mu, \kappa) \qquad (d \to \infty),$$

where $I_n(\mathfrak{G}, \mu, \kappa)$ are deterministic quantities that only depend on μ, κ and the algorithm $\mathfrak{G}$.

Corollary. Considering what is contained in $I_n(\mathbf{f}_d)$ this implies

$$\mathbf{f}_d(x_n) \xrightarrow{p} \mathfrak{f}_n(\mathfrak{G}, \mu, \kappa) \qquad \|\nabla \mathbf{f}_d(x_n)\| \xrightarrow{p} \mathfrak{g}_n(\mathfrak{G}, \mu, \kappa)$$

for deterministic limiting quantities $\mathfrak{f}_n(\mathfrak{G}, \mu, \kappa)$ and $\mathfrak{g}_n(\mathfrak{G}, \mu, \kappa)$.

Intuition of the result

Induction over number of evaluations n:

1. Induction start: At initialization point x_0 the claim follows from the dimensional scaling intuition.
2. Induction step: The main complication is that the evaluation points x_i are dependent on the random function $\mathbf{f}$. To solve this we need to condition on all pervious evaluations. If the x_i were deterministic, this conditional distribution is well known for Gaussian random functions, that is

$$\mathbf{g}(x_{n+1}) \coloneqq \begin{pmatrix} \mathbf{f}_d(x_{n+1}) \\ \nabla \mathbf{f}_d(x_{n+1}) \end{pmatrix} = \mu_{n+1|n} + \sqrt{\Sigma_{n+1|n}} \begin{pmatrix} Y_1 \\ \vdots \\ Y_d \end{pmatrix} \tag{3}$$

for independent $Y_i \sim \mathcal{N}(0,1)$ and

$$\begin{aligned}\mu_{n+1|n} &= \mathsf{E}[\mathbf{g}(x_{n+1}) \mid \mathbf{g}(x_{[1:n]})] \\ \Sigma_{n+1|n} &= \mathsf{Cov}(\mathbf{g}(x_{n+1}) \mid \mathbf{g}(x_{[1:n]}))\end{aligned}$$

Sparsity

Observe that the covariance matrix $\Sigma_{n+1|n}$ in (3) makes it difficult to obtain a similar LLN as in (1). While every entry of $\Sigma_{n+1|n}$ is of order $O(\frac{1}{d})$, it is also of size $(d+1)\times(d+1)$. Specifically,

$$\begin{aligned}\Sigma_{n+1|r} = \mathsf{Cov}(\mathbf{g}(x_{n+1})) \\ - \mathsf{Cov}\big(\mathbf{g}(x_{n+1}), \mathbf{g}(x_{[0:n]})\big)\Sigma_n^{-1}\mathsf{Cov}(\mathbf{g}(x_{[0:n]}), \mathbf{g}(x_{n+1}))\end{aligned}$$

with $\Sigma_n = \mathsf{Cov}(\mathbf{g}(x_{[0:n]}))$. Now the point is, that we would like a similar sparsity as in (2). Observe that the partial derivatives are uncorrelated if $y - x$ is orthogonal to the basis vector e_i or e_j. After the first step this will no longer be the case. However, if we choose the adapted span

$$V_n = \operatorname{span}\{x_0, \dots, x_n, \nabla \mathbf{f}(x_0), \dots, \nabla \mathbf{f}(x_n)\},$$

then the majority of directions in $V_n^\perp$ are orthogonal to all $x_i - x_j$. The directional derivatives in the directions $V_n^\perp$ are again iid like in (2), which yields the LLN we need, as $\dim V_n^\perp = d - 2(n+1) \sim d$ as $d \to \infty$.

Simulations on isotropic Gaussian random functions

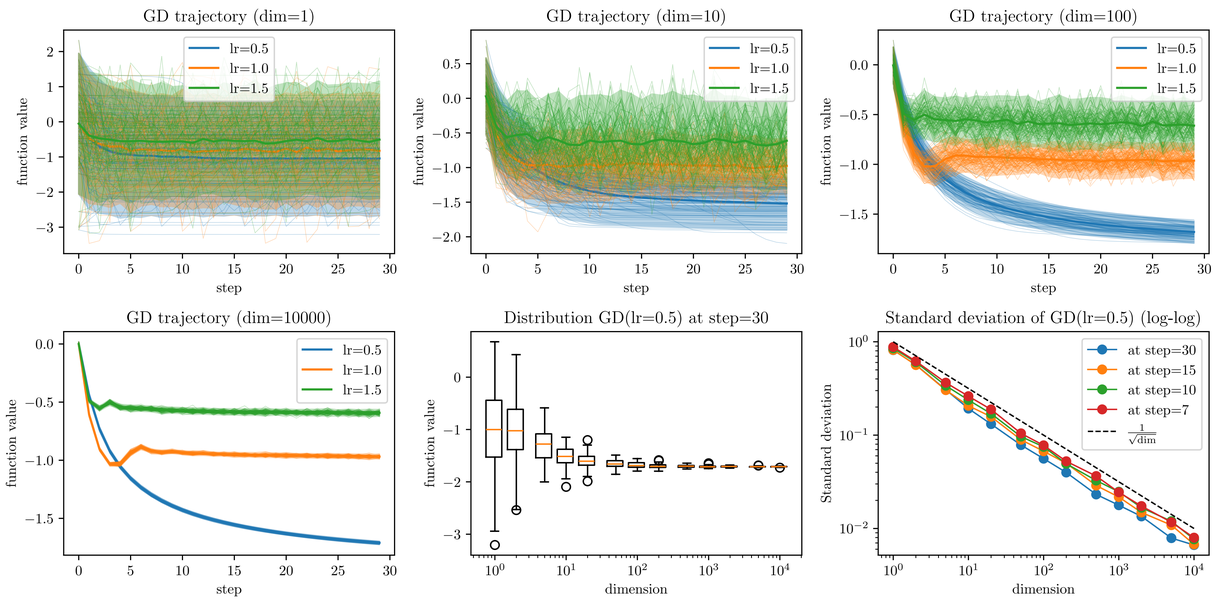


Figure 2. For each learning rate we simulated 100 gradient descent trajectories on a centered isotropic Gaussian random function with covariance $\mathcal{C}_{\mathbf{f}}(x,y) = \frac{1}{d}\exp(-\frac{\|x-y\|^2}{2})$.

Some of our related work

1. If the optimization surface is a random function, can we use this to obtain a better optimizer? Probably: See 'Random function descent' [1] (Algorithm assumes **stationary** isotropy, often false)
2. How does the abstract isotropic covariance function κ look like? A series characterization is given in [3].
3. How to deal with random variables plugged into random functions? This is a non-trivial measure theory problem, see [4]. For example let $Y \sim \mathcal{N}(\mu, 1)$ independent of $U \sim \mathcal{U}(0,1)$ and $\mathbf{f}(x) \coloneqq Y$, then

$$\mathsf{E}[\mathbf{f}(x) \mid U] \overset{\text{a.s.}}{=} \mu \mathbf{1}_{U \neq x} =: g(x) \quad \text{but} \quad \mathsf{E}[\mathbf{f}(U) \mid U] \neq g(U).$$

References

[1] 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.

[2] Felix Benning and Leif Döring. Gradient Span Algorithms Make Predictable Progress in High Dimension. *Journal of Machine Learning Research*, 27(121):1–62, 2026.

[3] Felix Benning and Max David Schölpple. Schoenberg characterization of continuous non-stationary isotropic positive definite kernels. *Accepted at: Journal of Fourier Analysis and Applications*, 2026.

[4] Felix Benning and Grégoire Szymanski. Measure Theory of Conditionally Independent Random Function Evaluation. 2026.

felix.benning@gmail.com