Random Function Descent

Felix Benning
University of Mannheim
felix.benning@uni-mannheim.de

Leif Döring
University of Mannheim
leif.doering@uni-mannheim.de

Abstract

Classical worst-case optimization theory neither explains the success of optimization in machine learning, nor does it help with step size selection. In this paper we demonstrate the viability and advantages of replacing the classical 'convex function' framework with a 'random function' framework. With complexity $\mathcal{O}(n^3d^3)$, where n is the number of steps and d the number of dimensions, Bayesian optimization with gradients has not been viable in large dimension so far. By bridging the gap between Bayesian optimization (i.e. random function optimization theory) and classical optimization we establish viability. Specifically, we use a 'stochastic Taylor approximation' to rediscover gradient descent, which is scalable in high dimension due to $\mathcal{O}(nd)$ complexity. This rediscovery yields a specific step size schedule we call Random Function Descent (RFD). The advantage of this random function framework is that RFD is scale invariant and that it provides a theoretical foundation for common step size heuristics such as gradient clipping and gradual learning rate warmup.

1 Introduction

Cost function minimization is one of the most fundamental mathematical problems in machine learning. Gradient-based methods, popular for this task, require a step size, typically chosen using established heuristics. This article aims to deepen the theoretical understanding of these heuristics and proposes a new algorithm based on this insight.

Classical optimization theory uses L-smoothness, which limits the rate of change of the gradient by L, to provide some convergence guarantees for learning rates smaller than $1/L$ [e.g. 38]. As this theory is based on an upper bound (the worst case), the learning rate $1/L$ is naturally much more conservative than necessary on average. Even if L was known, this learning rate would therefore be impractical. Since line search algorithms typically require access to full cost function evaluations, the field of machine learning (ML) therefore relies heavily on step size heuristics [e.g. 48, 49, 42, 20]. To investigate these heuristics, we introduce new ideas based on a 'random function' perspective.

While automatic step size selection in the convex function framework is possible [11], convexity is generally only satisfied asymptotically and locally. So the understanding of the initial stages of optimization, which includes the warmup heuristic [20], greatly benefits from a framework which also admits non-convex functions. This objective is achieved by the 'random function' framework we investigate.

Many successful algorithms in computer science are significantly slower in the worst case than in the average case based on a probabilistic framework (e.g. Quicksort [23] or the simplex algorithm [e.g. 6]). On random quadratic functions the average case behavior of first order optimizers is already being investigated by the ML community [e.g. 58, 43, 33, 12, 9, 40, 41]. Interested in the landscape of high dimensional random functions as a model for 'spin glasses', the physics community independently started studying the average case of optimization as well [e.g. 4, 15, 37, 51, 24], albeit not geared for ML algorithms.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Average case analysis fundamentally requires a prior distribution over possible cost functions. The evaluations seen so far then result in a posterior over the cost of other parameter inputs. Using this posterior for optimization is called "Bayesian optimization" (BO) [e.g. 32, 47, 16, 2], which is best known in the context of low dimensional optimization (e.g. hyperparameter tuning) in the ML community. BO is treated like a zero order method for low dimensional problems due to the $\mathcal{O}(n^3)$ complexity for the covariance matrix inversion of the n evaluations seen so far, which increases to $\mathcal{O}(n^3d^3)$ when gradient information is included [e.g. 35, 55], where d is the input dimension of our cost function. This limits classic BO to relatively small dimensions even under sparsity considerations [e.g. 45, 39].

While the BO algorithms developed in the 'random function framework' might not have been viable in high dimension so far, due to their computational complexity, this framework is already used to explain the high relative frequency of saddle points in high dimension [10] and to explain the highly predictable progress optimizers make on high dimensional cost functions [5].

In this work we bridge the gap between BO and (computationally viable) gradient based methods, derived from the first Taylor approximation, with the introduction of a stochastic Taylor approximation based on a forgetful BO posterior. The optimization method "Random Function Descent" (RFD), resulting from the minimization of this stochastic Taylor approximation, coincides with a specific form of gradient descent which establishes its viability in high dimension. The advantages of its BO heritage are scale invariance and an explicit step size schedule, which illuminates the inner workings of step size heuristics such as gradient clipping [42] and gradual learning rate warmup [20].

Our contributions and outline The main goal of this paper is to demonstrate the **viability** and **advantages** of replacing the classical "convex function" framework with a "random function" framework. Theorem 4.2 is the main theoretical result establishing **viability** (computatability and scalable complexity) for a given covariance model. Section 6 is concerned with practical estimation of the covariance model and viability is demonstrated with a practical example in the MNIST case study (Section 7). The **advantages** of this approach are scale invariance (Advantage 2.3) and an explicit step size schedule, which does not require expensive tuning and explains existing ML heuristics such as warmup (cf. Section 5.2). This explanation of the initial stage of optimization could never be delivered by the convex framework, because the convexity assumption is not fulfilled initially so it can at best explain asymptotic behavior.

Sec. 2 We motivate a stochastic Taylor approximation and RFD and prove its scale-invariance.

Sec. 3 We briefly motivate and discuss the common distributional assumptions in BO.

Sec. 4 We establish the connection between RFD and gradient descent.

Sec. 5 We investigate the step size schedule suggested by RFD. In particular we

0. calculate explicit formulas for the step size schedules resulting from common covariance models (Table 1, Sec. C),
1. analyze the general asymptotic behavior (Sec. 5.1),
2. discuss how RFD explains gradient clipping and learning rate warmup (Sec. 5.2),

Sec. 6 We develop a non-parametric variance estimation method, which is robust with respect to the choice of covariance kernel. Finally, we present an extension of RFD to mini-batch losses.

Sec. 7 We conduct a case study on the MNIST dataset.

Sec. 8 We discuss extensions (see also Sec. E) and limitations.

2 The random function descent algorithm

The classic derivation of gradient descent [e.g. 38, p. 29], adds an L-smoothness based trust bound to the first Taylor approximation, $T[J(\theta) \mid J(w), \nabla J(w)]$, of the cost function J around w resulting in the gradient step

$$w - \tfrac{1}{L}\nabla J(w) = \operatorname*{argmin}_{\theta} T[J(\theta) \mid J(w), \nabla J(w)] + \tfrac{L}{2}\|\theta - w\|^2.$$

Our unusual notation for the Taylor approximation $T[J(\theta) \mid J(w), \nabla J(w)]$ is meant to highlight the connection to the stochastic Taylor approximation we define below.

Definition 2.1 (Stochastic Taylor approximation)**.** We define the first order stochastic Taylor approximation of a random (cost) function[1] $\mathbf{J}$ around w by the conditional expectation

$$\mathbb{E}[\mathbf{J}(\theta) \mid \mathbf{J}(w), \nabla\mathbf{J}(w)].$$

This is the best L^2 approximation [30, Cor. 8.17] of $\mathbf{J}(\theta)$ provided first order knowledge of $\mathbf{J}$ at w.

We call this the 'stochastic Taylor approximation' because this approximation only makes use of derivatives in a single point. While the standard Taylor approximation is a polynomial approximation, the 'stochastic Taylor approximation' is the best approximation in an L^2 sense and already mean-reverting by itself, i.e. it naturally incorporates covariance-based trust (cf. Figure 1). While L-smoothness-based trust *guarantees* that the gradient still points in the direction we are going (for learning rates smaller $1/L$), covariance based trust tells us whether the derivative is still negative *on average*. Minimizing the stochastic Taylor approximation is therefore optimized for the average case. Since convergence proofs for gradient descent typically rely on an improvement *guarantee*, proving convergence is significantly harder in the average case and we answer this question only partially in Corollary 5.3.

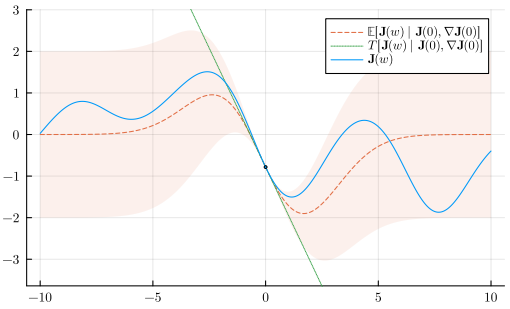

Figure 1: The stochastic Taylor approximation naturally contains a trust bound in contrast to the classical one. Here $\mathbf{J}$ is a Gaussian random function (with covariance as in Equation (11), with length scale $s = 2$ and variance $\sigma^2 = 1$). The ribbon represents two conditional standard deviations around the conditional expectation.

Definition 2.2 (Random Function Descent – RFD)**.** Select w_{n+1} as the minimizer[2] of the first order stochastic Taylor approximation

$$w_{n+1} := \operatorname*{argmin}_w \mathbb{E}[\mathbf{J}(w) \mid \mathbf{J}(w_n), \nabla\mathbf{J}(w_n)].$$

Properties of RFD Before we make RFD more explicit in Section 4, we discuss some properties which are easier to see in the abstract form.

First, observe that RFD is greedy and forgetful in the same way gradient descent is greedy and forgetful when derived as the minimizer of the regularized first Taylor approximation, or the Newton method as the minimizer of the second Taylor approximation. This is because the Taylor approximation only uses derivatives from the last point w_n (forgetful), and we minimize this approximation (greedy). Since momentum methods retain some information about past gradients, they are not as forgetful. We therefore expect a similar improvement could be made for RFD in the future.

Second, it is well known that classical gradient descent with exogenous step sizes (and most other first order methods) lack the scale invariance property of the Newton method [e.g. 21, 13]. Scale invariance means that scaling the input parameters w or the cost itself (e.g. by switching from the mean squared error to the sum squared error) does not change the points selected by the optimization method.

Advantage 2.3 (Scale invariance)**.** *RFD is invariant to additive shifts and positive scaling of the cost* $\mathbf{J}$*. RFD is also invariant with respect to transformations of the parameter input of* $\mathbf{J}$ *by differentiable bijections whose Jacobian is invertible everywhere (e.g. invertible linear maps).*

While invariance to bijections of inputs is much stronger than the affine invariance offered by the Newton method, non-linear bijections will typically break the 'isotropy' assumption of the following

[1] **Remark on terminology**: "stochastic process" [e.g. 53], "random field" [e.g. 1] and "random function" [e.g. 36] are all synonyms. However the latter seems most descriptive of random variables in the set of functions. "Gaussian processes" are naturally Gaussian stochastic processes, i.e. Gaussian random functions. To better distinguish random functions from deterministic functions, we use bold letters to denote random functions (as the usual convention of capitalizing random variables often clashes with other conventions for functions).

[2] we ignore throughout the main body that argmin could be set-valued and that the w_n would be random variables (cf. Section D.1.1 for a formal approach).

section which makes RFD explicit. This invariance should therefore be viewed as an opportunity to look for the bijection of inputs which ensures isotropy (e.g. a whitening transformation). The discussion of geometric anisotropy in Section E.1 is conducive to build an understanding of this.

3 A distribution over cost functions

It is impossible to make average case analysis explicit without a distribution over functions, so we use the canonical distributional assumption of Bayesian optimization [e.g. 16, 55, 44], 'isotropic Gaussian random functions'. This assumption was also used in the high dimensional setting by Dauphin et al. [10] to argue that saddle points are much more common than minima in high dimension, which is often cited to explain why second order methods are uncommon in machine learning.

To motivate isotropy, we note that in average case analysis the uniform distribution is popular, since it weighs all problem instances equally (e.g. all possible permutations in sorting). Isotropy is such a uniformity assumption, which essentially requires "$\mathbb{P}(\mathbf{J} = J) = \mathbb{P}(\mathbf{J} = J \circ \phi)$", for all isometries ϕ. In other words, the probability that our cost function is equal to J is equal to the probability that it is equal to a shifted and turned version of J, given by $J \circ \phi$.

Since the probability of any single realization of a cost function J is zero, the equation we put in quotes is mathematically unsound. The formal definition follows below.

Definition 3.1 (Isotropy)**.** A random function $\mathbf{J}$ is called isotropic if its distribution stays the same under isometric transformations of its input, i.e. for any isometry ϕ we have

$$\mathbb{P}_{\mathbf{J}} = \mathbb{P}_{\mathbf{J} \circ \phi}.$$

If $\mathbf{J}$ is Gaussian, isotropy is well known [e.g. 44, 1] to be equivalent to the condition that there exists $\mu \in \mathbb{R}$ and a function $C : \mathbb{R} \to \mathbb{R}$ such that for all $w, \tilde{w} \in \mathbb{R}^d$ the expectation and covariance are

$$\mathbb{E}[\mathbf{J}(w)] = \mu, \qquad \operatorname{Cov}(\mathbf{J}(w), \mathbf{J}(\tilde{w})) = C\big(\tfrac{\|w-\tilde{w}\|^2}{2}\big).$$

For these isotropic Gaussian random functions we use the notation $\mathbf{J} \sim \mathcal{N}(\mu, C)$.

We discuss generalizations to isotropy in Section F and E.1, but for ease of exposition we retain the (stationary) isotropy assumption throughout the main body. Note that the Gaussian assumption can be statistically tested in practice (cf. Figure 4), but it is also straightforward to reproduce our results with the "best linear unbiased estimator" (BLUE) (Section E.3) in place of the conditional expectation to remove the Gaussian assumption. We finally want to highlight that, in contrast to the uniformity assumption on finite sets, 'isotropic Gaussian random functions' leave us with a family of plausible distributions. It is therefore necessary to estimate μ and C, which is the topic of Section 6.

4 Relation to gradient descent

While we were able to define RFD abstractly without any assumptions on the distribution $\mathbb{P}_{\mathbf{J}}$ of the random cost $\mathbf{J}$, an explicit calculation requires distributional assumptions and we have motivated isotropic Gaussian random functions in Section 3 for this purpose. The assumption of isotropy allows for an explicit version of the stochastic Taylor approximation which then immediately leads to an explicit version of RFD.

Lemma 4.1 (Explicit first order stochastic Taylor approximation)**.** *For $\mathbf{J} \sim \mathcal{N}(\mu, C)$, the first order stochastic Taylor approximation is given by*

$$\mathbb{E}[\mathbf{J}(w - \mathbf{d}) \mid \mathbf{J}(w), \nabla\mathbf{J}(w)] = \mu + \frac{C\big(\frac{\|\mathbf{d}\|^2}{2}\big)}{C(0)}(\mathbf{J}(w) - \mu) - \frac{C'\big(\frac{\|\mathbf{d}\|^2}{2}\big)}{C'(0)}\langle \mathbf{d}, \nabla\mathbf{J}(w)\rangle.$$

The explicit version of RFD follows by fixing the step size $\eta = \|\mathbf{d}\|$ and optimizing over the direction first.

Theorem 4.2 (Explicit RFD)**.** *Let $\mathbf{J} \sim \mathcal{N}(\mu, C)$, then RFD coincides with gradient descent*

$$w_{n+1} = w_n - \eta_n^* \tfrac{\nabla\mathbf{J}(w_n)}{\|\nabla\mathbf{J}(w_n)\|},$$

Table 1: RFD step size (cf. Figure 2 and Eq. (11), (13), (14) for the formal definitions of the models). In particular, s is the length scale in all covariance models.

Model		RFD step size η^* for $\mathbf{J}(w) \leq \mu$		A-RFD
		General case $\left(\text{with } \Theta = \frac{\|\nabla \mathbf{J}(w)\|}{\mu - \mathbf{J}(w)}\right)$	$\mathbf{J}(w) = \mu$	$\Theta \to 0$
Matérn	ν			
	$3/2$	$\frac{s}{\sqrt{3}} \frac{1}{\left(1 + \frac{\sqrt{3}}{s\Theta}\right)}$	$\approx 0.58s$	$\frac{1}{3}s^2\Theta$
	$5/2$	$\frac{s}{\sqrt{5}} \frac{(1-\zeta)+\sqrt{4+(1+\zeta)^2}}{2(1+\zeta)}$ with $\zeta := \frac{\sqrt{5}}{3s\Theta}$.	$\approx 0.72s$	$\frac{3}{5}s^2\Theta$
Squared-exponential	∞	$\frac{s^2}{\sqrt{\left(\frac{\mu - \mathbf{J}(w)}{2}\right)^2 + s^2\|\nabla \mathbf{J}(w)\|^2} + \frac{\mu - \mathbf{J}(w)}{2}} \|\nabla \mathbf{J}(w)\|$	s	$s^2\Theta$
Rational quadratic	β	$s\sqrt{\beta}\, \underset{\eta}{\text{Root}} \left(-1 + \frac{\sqrt{\beta}}{s\Theta}\eta + (1+\beta)\eta^2 + \frac{\sqrt{\beta}}{s\Theta}\eta^3\right)$	$s\sqrt{\frac{\beta}{1+\beta}}$	$s^2\Theta$

where the RFD step sizes are given by

$$\eta_n^* := \underset{\eta \in \mathbb{R}}{\operatorname{argmin}} \frac{C\left(\frac{\eta^2}{2}\right)}{C(0)}(\mathbf{J}(w_n) - \mu) - \eta \frac{C'\left(\frac{\eta^2}{2}\right)}{C'(0)} \|\nabla \mathbf{J}(w_n)\|. \tag{1}$$

While the descent direction is a universal property for all isotropic Gaussian random functions, it follows from (1) that the step sizes depend much more on the specific covariance structure. In particular it depends on the decay rate of the covariance acting as the trust bound.

Remark 4.3 (Scalable complexity). While Bayesian optimization typically has computational complexity $\mathcal{O}(n^3 d^3)$ in number of steps n and dimensions d [55, 45], RFD under the isotropy assumption has the same computational complexity as gradient descent (i.e. $\mathcal{O}(nd)$).

Remark 4.4 (Step until the given information is no longer informative). While L-smoothness-based trust prescribes step sizes that *guarantee* the slope to point downwards over the entire step, RFD prescribes steps which are exactly large enough that the gradient is no longer correlated to the previously observed evaluation. This is because the first order condition demands

$$0 \overset{!}{=} \nabla \mathbb{E}[\mathbf{J}(w) \mid \mathbf{J}(w_n), \nabla \mathbf{J}(w_n)] = \mathbb{E}[\nabla \mathbf{J}(w) \mid \mathbf{J}(w_n), \nabla \mathbf{J}(w_n)].$$

And for measurable functions $\phi : \mathbb{R}^{d+1} \to \mathbb{R}$ such that $\Phi = \phi(\mathbf{J}(w_n), \nabla \mathbf{J}(w_n))$ is sufficiently integrable, Φ is then uncorrelated from $\partial_i \mathbf{J}(w)$ by the first order condition

$$\operatorname{Cov}(\partial_i \mathbf{J}(w), \Phi) = \mathbb{E}\Big[\underbrace{\mathbb{E}[\partial_i \mathbf{J}(w) \mid \mathbf{J}(w_n), \nabla \mathbf{J}(w_n)]}_{=0}(\Phi - \mathbb{E}[\Phi])\Big] = 0.$$

5 The RFD step size schedule

While classical theory leads to 'learning rates', RFD suggests 'step sizes' applied to normalized gradients representing the actual length of the step size. In the following we thus make the distinction

$$w_{n+1} = w_n - \underbrace{h_n}_{\text{'learning rate'}} \nabla \mathbf{J}(w_n) = w_n - \underbrace{\eta_n}_{\text{'step size'}} \frac{\nabla \mathbf{J}(w_n)}{\|\nabla \mathbf{J}(w_n)\|}.$$

To get a better feel for the step sizes suggested by RFD, it is enlightening to divide (1) by $\mu - \mathbf{J}(w_n)$ which results in a minimization problem

$$\eta^* := \eta^*(\Theta) := \underset{\eta}{\operatorname{argmin}}\, q_\Theta(\eta) \qquad \text{for} \qquad q_\Theta(\eta) := -\frac{C\left(\frac{\eta^2}{2}\right)}{C(0)} - \eta \frac{C'\left(\frac{\eta^2}{2}\right)}{C'(0)}\Theta, \tag{2}$$

which is only parametrized by the "gradient cost quotient"

$$\Theta_n = \frac{\|\nabla \mathbf{J}(w_n)\|}{\mu - \mathbf{J}(w_n)},$$

i.e. $\eta_n^* = \eta^*(\Theta_n)$. This minimization problem can be solved explicitly for the most common [44, ch. 4] differentiable isotropic covariance models, see Table 1, Figure 2 and Appendix C for details.

Figure 2 can be interpreted as follows: At the start of optimization, the cost should be roughly equal to the average cost $\mu \approx \mathbf{J}(w)$, so the gradient cost quotient Θ is infinite and the step sizes are therefore given by $\eta^*(\infty)$ (also listed in its own column in Table 1). As we start minimizing, the difference $\mu - \mathbf{J}(w)$ becomes positive. Towards the end of minimization this difference no longer changes as the cost no longer decreases. I.e. towards the end the gradient cost quotient Θ is roughly linear in the gradient $\|\nabla \mathbf{J}(w)\|$. The derivative $\frac{d}{d\Theta}\eta^*(0)$ of $\eta^*(\Theta)$ at zero then effectively results in a constant asymptotic learning rate.

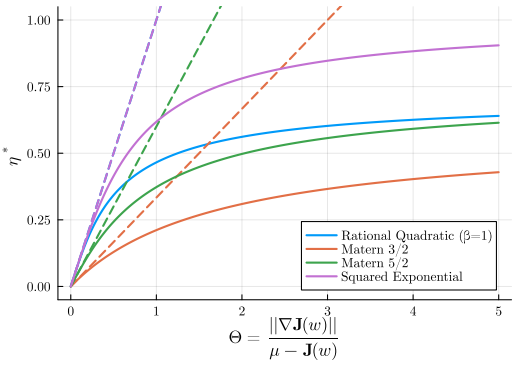

Figure 2: RFD step sizes as a function of $\Theta = \frac{\|\nabla \mathbf{J}(w)\|}{\mu - \mathbf{J}(w)}$ assuming scale $s = 1$ (cf. Table 1). A-RFD (Definition 5.1) is plotted as dashed lines. A-RFD of the rational quadratic coincides with A-RFD of the squared exponential covariance.

5.1 Asymptotic learning rate

To explain the claim above, note that the gradient cost quotient Θ converges to zero towards the end of optimization, because the gradient norm converges to zero. A first order Taylor expansion of η^* would therefore imply

$$\eta^*(\Theta) \approx \eta^*(0) + \tfrac{d}{d\Theta}\eta^*(0)\Theta = \underbrace{\frac{\frac{d}{d\Theta}\eta^*(0)}{\mu - \mathbf{J}(w)}}_{\text{asymptotic learning rate}} \|\nabla \mathbf{J}(w)\|$$

assuming $\eta^*(0) = 0$ and differentiability of η^*, which is a reasonable educated guess based on the the examples in Figure 2. But since the RFD step sizes η^* are abstractly defined as an argmin, it is necessary to formalize this intuition for general covariance models. First, we define asymptotic step sizes as an object towards which we can prove convergence. Then we prove convergence, proving they are well defined. In addition, we obtain a more explicit formula for the asymptotic learning rate.

Definition 5.1 (A-RFD)**.** We define the step sizes of "asymptotic RFD" (A-RFD) to be the minimizer of the second order Taylor approximation $T_2 q_\Theta$ of q_Θ around zero

$$\hat{\eta}(\Theta) := \operatorname*{argmin}_\eta T_2 q_\Theta(\eta) = \frac{C(0)}{-C'(0)}\Theta = \underbrace{\frac{C(0)}{C'(0)(\mathbf{J}(w) - \mu)}}_{\text{asymptotic learning rate}} \|\nabla \mathbf{J}(w)\|.$$

In the following we prove that these are truly asymptotically equal to the step sizes η^* of RFD.

Proposition 5.2 (A-RFD is well defined)**.** *Let $\mathbf{J} \sim \mathcal{N}(\mu, C)$ and assume there exists $\eta_0 > 0$ such that the correlation for larger distances $\eta \geq \eta_0$ are bounded smaller than 1, i.e. $\frac{C(\eta^2/2)}{C(0)} < \rho \in (0, 1)$. Then the step sizes of RFD are asymptotically equal to the step sizes of A-RFD, i.e.*

$$\hat{\eta}(\Theta) \sim \eta^*(\Theta) \quad \textit{as} \quad \Theta \to 0.$$

Note that the assumption is essentially always satisfied, since the Cauchy-Schwarz inequality implies

$$C\left(\tfrac{\|w - \tilde{w}\|^2}{2}\right) = \operatorname{Cov}(\mathbf{J}(w), \mathbf{J}(\tilde{w})) \leq \sqrt{\operatorname{Var}(\mathbf{J}(w)) \operatorname{Var}(\mathbf{J}(\tilde{w}))} = C(0),$$

where equality requires the random variables to be almost surely equal [30]. If the random function is not periodic or constant, this will generally be strict. In the proof, this requirement is only needed to ensure that η^* is not very large. The smallest local minimum of q_Θ is always close to $\hat{\eta}$ even without this assumption (which ensures it is a global minimum).

Figure 2 illustrates that $\eta^* \to 0$ should imply $\Theta \to 0$, resulting in a weak convergence guarantee.

Corollary 5.3. *Assume $\eta^* \to 0$ implies $\Theta \to 0$, the cost $\mathbf{J}$ is bounded, has continuous gradients and RFD converges to some point w_∞. Then w_∞ is a critical point and the RFD step sizes η^* are asymptotically equal to $\hat{\eta}$.*

For the squared exponential covariance model we formally prove that η^* is strictly monotonously increasing in Θ and thus $\eta^* \to 0$ implies $\Theta \to 0$ (Prop. C.3). The 'bounded' and 'continuous gradients' assumptions are almost surely satisfied for all sufficiently smooth covariance functions [cf. 1], where three times differentiable is more than enough smoothness.

5.2 RFD step sizes explain common step size heuristics

Asymptotically, RFD suggests constant learning rates, similar to the classical L-smooth setting. We thus define these asymptotic learning rates (as the limit of the learning rates h_n of iteration n) to be

$$h_\infty := \frac{C(0)}{C'(0)(\mathbf{J}(w_\infty) - \mu)}, \tag{3}$$

where $\mathbf{J}(w_\infty)$ is the cost we reach in the limit. If we used these asymptotic learning rates from the start, step sizes would become too large for large gradients, as RFD step sizes exhibit a plateau (cf. Figure 2). To emulate the behavior of RFD with a piecewise linear function, we could introduce a cutoff whenever our step size exceeds the initial step size $\eta^*(\infty)$, i.e.

$$w_{n+1} = w_n - \min\Big\{h_\infty, \frac{\eta^*(\infty)}{\|\nabla \mathbf{J}(w_n)\|}\Big\}\nabla \mathbf{J}(w_n). \qquad \text{(gradient clipping)}$$

At this point we have rediscovered 'gradient clipping' [42]. Since the rational quadratic covariance has the same asymptotic learning rate h_∞ for every β, its parameter β controls the step size bound $\eta^*(\infty)$ of gradient clipping (cf. Table 1, Figure 2).

Pascanu et al. [42] motivated gradient clipping with the geometric interpretation of movement towards a 'wall' placed behind the minimum. This suggests that clipping should happen towards the end of training. This stands in contrast to a more recent step size heuristic, "(linear) warmup" [20], which suggests smaller learning rates at the start (i.e. $h_0 = \frac{\eta^*(\infty)}{\|\nabla \mathbf{J}(w_0)\|}$) and gradual ramp-up to the asymptotic learning rate h_∞. In other words, gradients are not clipped due to some wall next to the minimum, but because the step sizes would be too large at the start otherwise. Goyal et al. [20] further observe that 'constant warmup' (i.e. a step change of learning rates akin to gradient clipping) performs worse than gradual warmup. Since RFD step sizes suggest this gradual increase, we argue that they may have discovered RFD step sizes empirically (also cf. Figure 3).

6 Mini-batch loss and covariance estimation

Since we do not have access to evaluations of the cost $\mathbf{J}$ in practice, we need to prove some results about stochastic losses ℓ_i before we can apply RFD in practice. For this, assume that we have independent identically distributed (iid) data X_i independent of the true relationship $\mathbf{f}$ drawn from $\mathbb{P}_\mathbf{f}$ resulting in labels $Y_i = \mathbf{f}(X_i) + \varsigma_i$, where we have added independent iid noise ς_i, resulting in loss and cost

$$\ell_i(w) := \ell\big(w, (X_i, Y_i)\big) \quad \text{and} \quad \mathbf{J}(w) := \mathbb{E}[\ell_i(w) \mid \mathbf{f}].$$

In this setting we confirm (cf. Lemma D.9), that the stochastic approximation errors

$$\epsilon_i(w) := \ell_i(w) - \mathbf{J}(w)$$

are independent conditional on the true relationship $\mathbf{f}$. In particular they (and all their derivatives) are uncorrelated and also uncorrelated from $\mathbf{J}$. It follows that mini-batch losses

$$\mathcal{L}_b(w) := \frac{1}{b}\sum_{i=1}^{b} \ell_i(w) = \mathbf{J}(w) + \frac{1}{b}\sum_{i=1}^{b} \epsilon_i(w) \tag{4}$$

have variance

$$\operatorname{Var}(\mathcal{L}_b(w)) = \operatorname{Var}(\mathbf{J}(w)) + \tfrac{1}{b}\operatorname{Var}(\epsilon_1(w)) \overset{\text{isotropy}}{=} C(0) + \tfrac{1}{b}C_\epsilon(0), \tag{5}$$

where we assume $\mathbf{J} \sim \mathcal{N}(\mu, C)$ and $\epsilon_i \sim \mathcal{N}(0, C_\epsilon)$ in the last equation for simplicity. But this step did not yet require the distributional Gaussian assumption beyond the mean and variance.

6.1 Variance estimation

Recall that the asymptotic learning rate h_∞ in Equation (3) only depends on $C(0)$ and $C'(0)$. So if we estimate these values, we are certain to get the right RFD step sizes asymptotically without knowing the entire covariance kernel C.

Equation (5) reveals that for $Z_b := (\mathcal{L}_b(w) - \mu)^2$ we have

$$\mathbb{E}[Z_b] = \beta_0 + \tfrac{1}{b}\beta_1 \qquad \text{i.e.} \qquad Z_b = \beta_0 + \tfrac{1}{b}\beta_1 + \text{noise}$$

with bias $\beta_0 = C(0)$ and slope $\beta_1 = C_\epsilon(0)$. So a linear regression on samples $(\frac{1}{b_k}, Z_{b_k})_{k\leq n}$ allows for the estimation of β_0 and β_1. Using the Gaussian assumption from (5), the variance of Z_b is the (centered) fourth moment of $\mathcal{L}_b$, which is given by

$$\sigma_b^2 := \operatorname{Var}(Z_b) = \mathbb{E}[Z_b^4] - \mathbb{E}[Z_b^2]^2 = 2\operatorname{Var}(\mathcal{L}_b(w))^2 = 2(\beta_0 + \tfrac{1}{b}\beta_1)^2.$$

In particular the variance of Z_b depends on the batch size b. The linear regression is therefore heteroskedastic. Weighted least squares (WLS) [e.g. 28, Theorem 4.2] is designed to handle this case, but for its application the variance of Z_b is needed. Since β_0, β_1 are the parameters we wish to estimate, we find ourselves in the paradoxical situation that we need β to obtain β. Our solution to this problem is to start with a guess of β_0, β_1, apply WLS to obtain a better estimate and repeat this bootstrapping procedure until convergence. Since all Z_b have the same underlying cost $\mathbf{J}$, we sample the parameters w randomly to reduce their covariance (details in Sec. B).

The same procedure can be applied to obtain $C'(0)$, where the counterpart of Equation (5) is given by

$$\operatorname{Var}(\partial_i \mathcal{L}_b(w)) = \operatorname{Var}(\partial_i \mathbf{J}(w)) + \tfrac{1}{b}\operatorname{Var}(\partial_i \epsilon_1(w)) \overset{\text{isotropy}}{=} -(C'(0) + \tfrac{1}{b}C'_\epsilon(0)).$$

Remark 6.1. Under the isotropy assumption the partial derivatives are iid, so the expectation of $\|\nabla \mathcal{L}_b(w)\|^2 = \sum_{i=1}^d (\partial_i \mathcal{L}_b(w))^2$ is this variance scaled by d. In particular the variance needs to scale with $\frac{1}{d}$ to keep the gradient norms (and thus the Lipschitz constant of $\mathbf{J}$) stable. This observation is closely related to "isoperimetry" [e.g. 7], for details see [5]. Removing the isotropy assumption and estimating the variance component-wise is most likely how "adaptive" step sizes [e.g. 14, 29], like the ones used by Adam, work (cf. Sec. E.1).

Batch size distribution Before we can apply linear regression to the samples $(\frac{1}{b_k}, Z_{b_k})_{k\leq n}$, it is necessary to choose the batch sizes b_k. As this choice is left to us, we calculate the variance of our estimator $\hat{\beta}_0$ of β_0 explicitly (Lemma B.2), in order to minimize this variance subject to a sample budget α over the selection of batch sizes

$$\min_{n, b_1, \dots, b_n} \operatorname{Var}(\hat{\beta}_0) \quad \text{s.t.} \quad \underbrace{\sum_{k=1}^n b_k}_{\text{samples used}} \leq \alpha. \tag{6}$$

Since this optimization problem is very difficult to solve, we rephrase it in terms of the empirical distribution of batch sizes $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{b_i}$. Optimizing over distributions is still difficult, but we explain in Section B.1 how to heuristically arrive at the parametrization

$$\nu(b) \propto \exp\big(\lambda_1 \tfrac{1}{\sigma_b^2} - \lambda_2 b\big), \qquad b \in \mathbb{N}$$

where the parameters $\lambda_1, \lambda_2 \geq 0$ can then be used to optimize (6). Due to our usage of σ_b^2 this has to be bootstrapped.

Covariance estimation While the variance estimates above ensure correct asymptotic learning rates, we motivated in Section 5.2 that asymptotic learning rates alone would result in too large step sizes at the beginning. We therefore use the estimates of $C(0)$ and $C'(0)$ to fit a covariance model, effectively acting as a gradient clipper while retaining the asymptotic guarantees. Note that covariance models with less than two parameters are generally fully determined by these values.

6.2 Stochastic RFD (S-RFD)

It is reasonable to ask whether there is a 'stochastic gradient descent'-like counterpart to the 'gradient descent'-like RFD. The answer is yes, and we already have all the required machinery.

Extension 6.2 (S-RFD). *For loss* $\mathbf{J} \sim \mathcal{N}(\mu, C)$ *and stochastic errors* $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, C_\epsilon)$ *we have*

$$\operatorname*{argmin}_{\mathbf{d}} \mathbb{E}[\mathbf{J}(w - \mathbf{d}) \mid \mathcal{L}_b(w), \nabla \mathcal{L}_b(w)] = \eta^*(\Theta) \tfrac{\nabla \mathcal{L}_b(w)}{\|\nabla \mathcal{L}_b(w)\|}$$

with the same step size function η^* as for RFD, but modified Θ

$$\Theta = \frac{C'(0)}{C'(0) + \frac{1}{b}C'_\epsilon(0)} \frac{C(0) + \frac{1}{b}C_\epsilon(0)}{C(0)} \frac{\|\nabla \mathcal{L}_b(w)\|}{\mu - \mathcal{L}_b(w)}.$$

Note, that our non-parametric covariance estimation already provides us with estimates of $C_\epsilon(0)$ and $C'_\epsilon(0)$, so no further adaptions are needed. The resulting asymptotic learning rate is given by

$$h_\infty = \frac{C(0) + \frac{1}{b}C_\epsilon(0)}{(C'(0) + \frac{1}{b}C'_\epsilon(0))(\mathcal{L}_b(w_\infty) - \mu)}. \tag{7}$$

7 MNIST case study

For our case study we use the negative log likelihood loss to train a neural network [3, M7] on the MNIST dataset [34]. We choose this model as one of the simplest state-of-the-art models at the time of selection, consisting only of convolutional layers with ReLU activation interspersed by batch normalization layers and a single dense layers at the end with softmax activation. Assuming isotropy, we estimate μ, $C(0)$ and $C'(0)$ as described in Section 6.1 and deduce the parameters σ^2 and s of the respective covariance model (more details in Section B). We then use the step sizes listed in Table 1 for the 'squared exponential' and 'rational quadratic' covariance in our RFD algorithm.

In Figure 3, RFD is benchmarked against step size tuned Adam [29] and stochastic gradient descent (SGD). Even with early stopping, their tuning would typically require more than 1 epoch worth of samples, *in contrast to RFD* (Section A.1.1). We highlight that A-RFD performs significantly worse than either of the RFD versions which effectively implement some form of learning rate warmup. This is despite the RFD learning rates converging to the asymptotic one within one epoch (ca. 30 out of 60 steps per epoch). The step sizes on the other hand are (up to noise) monotonously decreasing. This stands in contrast to the "wall next to the minimum" motivation of gradient clipping.

Code availability: Our implementation of RFD can be found at `https://github.com/FelixBenning/pyrfd` and the package can also be installed from PyPI via '`pip install pyrfd`'.

8 Limitations and extensions

To cover the vast amount of ground that lays between the 'formulation of a general average case optimization problem' and the 'prototype of a working optimizer with theoretical backing',

1. we used the common [16, 52, 55, 10] *isotropic* and *Gaussian* distributional assumption for $\mathbf{J}$,
2. we used very *simple covariance models* for the actual implementation,
3. we used WLS in our variance estimation procedure despite the *violation of independence*.

Since RFD is defined as the minimizer of an average instead of an upper bound – making it more risk affine – it naturally loses the improvement guarantee driving classical convergence proofs. It is therefore impossible to extend classical optimization proofs and new mathematical theory must be developed. This risk-affinity can also be observed in its comparatively large step sizes (cf. Fig. 3 and Sec. A). On CIFAR-100 [31], the step sizes were *too* large and it is an open question whether assumptions were violated or whether RFD is simply too risk-affine. But since the variance of random functions vanishes asymptotic with high dimension [5] we highly suspect the former (cf. Remark E.5).

Future work will therefore have to target these assumptions. Some of the assumptions were already simplifications for the sake of exposition, and we deferred their relaxation to the appendix. The Gaussian assumption can be relaxed with a generalization to the 'BLUE' (Sec. E.3), isotropy can be generalized to 'geometric anisotropies' (Sec. E.1) and the risk-affinity of RFD can be reduced with confidence intervals (Sec. E.2). Since simple random linear models already violate stationary isotropy (Sec. F.1), we believe that stationarity is the most important assumption to attack in future work.

9 Conclusion

In this paper we have demonstrated the **viability** (computability and scalable complexity) and **advantages** (scale invariance, explainable step size schedule which does not require expensive

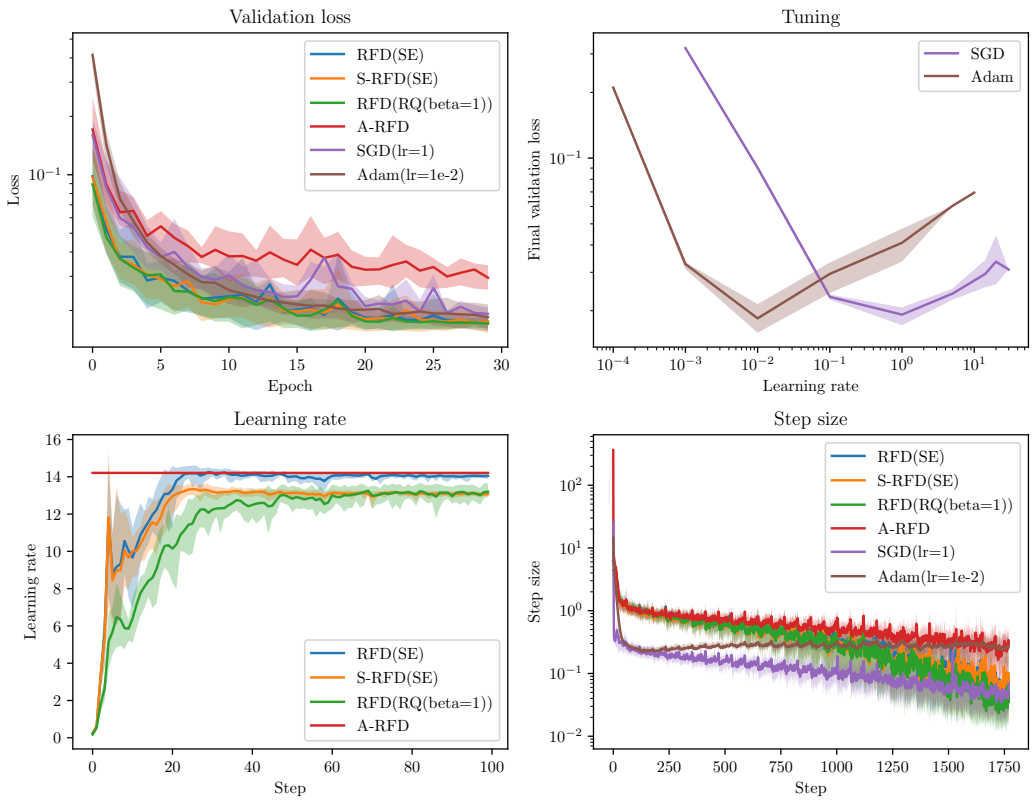

Figure 3: Training on the MNIST dataset (batch size 1024). Ribbons describe the range between the 10% and 90% quantile of 20 repeated experiments while lines represent their mean. SE stands for the squared exponential (11) and RQ for the rational quadratic (13) covariance. The validation loss uses the test data set, which provides a small advantage to Adam and SGD, as we also use it for tuning.

tuning) of replacing the classical "convex function" framework with the "random function" framework. Along the way we bridged the gap between Bayesian optimization (not scalable so far) and classical optimization methods (scalable). This theoretical framework not only sheds light on existing step size heuristics, but can also be used to develop future heuristics.

We envision the following improvements to RFD in the future:

1. The *reliability* of RFD can be improved by generalizing the distributional assumptions to cover more real world scenarios. In particular we are interested in the generalization to non-stationary isotropy because we suspect that regularization such as weight and batch normalization [46, 25] are used to patch violations of stationarity (cf. Section F).
2. The *performance* of RFD can also be improved. Since RFD is forgetful while momentum methods retains some information it is likely fruitful to relax the full forgetfulness. Furthermore, we suspect that adaptive learning rates [e.g. 14, 29], such as those used by Adam, can be incorporated with geometric anisotropies (cf. Sec. E.1). Performance could also be further improved by estimating the covariance (locally) online instead of globally at the start. Finally, the implementation itself can be made more performant.

Acknowledgement

We extend our sincere gratitude to our colleagues at the University of Mannheim, with special thanks to Rainer Gemulla and Julie Naegelen for insightful discussions and invaluable feedback. The Experiments in this work were partially carried out on the compute cluster of the state of Baden-Würtemberg (bwHPC).

References

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

[2] Apoorv Agnihotri and Nipun Batra. "Exploring Bayesian Optimization". In: *Distill* 5.5 (2020-05-05), e26. ISSN: 2476-0757. DOI: 10.23915/distill.00026.

[3] Sanghyeon An et al. *An Ensemble of Simple Convolutional Neural Network Models for MNIST Digit Recognition*. 2020-10-04. DOI: 10.48550/arXiv.2008.10400. Pre-published.

[4] Antonio Auffinger and Qiang Zeng. "Complexity of Gaussian Random Fields with Isotropic Increments". In: *Communications in Mathematical Physics* 402.1 (2023-08-01), pp. 951–993. ISSN: 1432-0916. DOI: 10.1007/s00220-023-04739-0.

[5] Felix Benning and Leif Döring. *Gradient Span Algorithms Make Predictable Progress in High Dimension*. 2024-10-13. DOI: 10.48550/arXiv.2410.09973. Pre-published.

[6] Karl Heinz Borgwardt. *The Simplex Method: A Probabilistic Analysis*. Softcover reprint of the original 1st ed. 1987 edition. Berlin Heidelberg: Springer, 1986-11-01. 282 pp. ISBN: 978-3-540-17096-9.

[7] Sebastien Bubeck and Mark Sellke. "A Universal Law of Robustness via Isoperimetry". In: *Advances in Neural Information Processing Systems*. Vol. 34. Virtual Event: Curran Associates, Inc., 2021, pp. 28811–28822. arXiv: 2105.12806 [cs, stat]. URL: https://proceedings.neurips.cc/paper/2021/hash/f197002b9a0853eca5e046d9ca4663d5-Abstract.html (visited on 2023-09-22).

[8] Youngmin Cho and Lawrence Saul. "Kernel Methods for Deep Learning". In: *Advances in Neural Information Processing Systems*. Vol. 22. Curran Associates, Inc., 2009. URL: https://proceedings.neurips.cc/paper/2009/hash/5751ec3e9a4feab575962e78e006250d-Abstract.html (visited on 2023-04-03).

[9] Leonardo Cunha et al. "Only Tails Matter: Average-Case Universality and Robustness in the Convex Regime". In: *Proceedings of the 39th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, 2022-06-28, pp. 4474–4491. URL: https://proceedings.mlr.press/v162/cunha22a.html (visited on 2023-11-09).

[10] Yann N Dauphin et al. "Identifying and Attacking the Saddle Point Problem in High-Dimensional Non-Convex Optimization". In: *Advances in Neural Information Processing Systems*. Vol. 27. Montréal, Canada: Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper/2014/hash/17e23e50bedc63b4095e3d8204ce063b-Abstract.html (visited on 2022-06-10).

[11] Aaron Defazio and Konstantin Mishchenko. "Learning-Rate-Free Learning by D-Adaptation". In: *Proceedings of the 40th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, 2023-07-03, pp. 7449–7479. arXiv: 2301.07733 [cs.LG]. URL: https://proceedings.mlr.press/v202/defazio23a.html (visited on 2024-03-31).

[12] Percy Deift and Thomas Trogdon. "The Conjugate Gradient Algorithm on Well-Conditioned Wishart Matrices Is Almost Deterministic". In: *Quarterly of Applied Mathematics* 79.1 (2021-03), pp. 125–161. ISSN: 0033-569X, 1552-4485. DOI: 10.1090/qam/1574. arXiv: 1901.09007 [cs, math].

[13] P. Deuflhard and G. Heindl. "Affine Invariant Convergence Theorems for Newton's Method and Extensions to Related Methods". In: *SIAM Journal on Numerical Analysis* 16.1 (1979-02), pp. 1–10. ISSN: 0036-1429. DOI: 10.1137/0716001.

[14] John Duchi, Elad Hazan, and Yoram Singer. "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". In: *The Journal of Machine Learning Research* 12 (2011-07-01), pp. 2121–2159. ISSN: 1532-4435.

[15] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. "Optimization of Mean-Field Spin Glasses". In: *The Annals of Probability* 49.6 (2021-11), pp. 2922–2960. DOI: 10.1214/21-AOP1519.

[16] Peter I. Frazier. "Bayesian Optimization". In: *Recent Advances in Optimization and Modeling of Contemporary Problems*. INFORMS TutORials in Operations Research. Phoenix, Arizona, USA: INFORMS, 2018-10, pp. 255–278. ISBN: 978-0-9906153-2-3. DOI: 10.1287/educ.2018.0188. arXiv: 1807.02811 [cs, math, stat].

[17] Fuchang Gao and Lixing Han. "Implementing the Nelder-Mead Simplex Algorithm with Adaptive Parameters". In: *Computational Optimization and Applications* 51.1 (2012-01-01), pp. 259–277. ISSN: 1573-2894. DOI: 10.1007/s10589-010-9329-3.

[18] 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*. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Sardinia, Italy: JMLR Workshop and Conference Proceedings, 2010-03-31, pp. 249–256. URL: https://proceedings.mlr.press/v9/glorot10a.html (visited on 2023-04-11).

[19] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016-11-10. 801 pp. ISBN: 978-0-262-33737-3. Google Books: omivDQAAQBAJ.

[20] Priya Goyal et al. *Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour*. arXiv, 2018-04-30. arXiv: 1706.02677 [cs]. URL: http://arxiv.org/abs/1706.02677 (visited on 2024-04-02).

[21] Anders Hansson. *Optimization for Learning and Control*. Hoboken, New Jersey: John Wiley & Sons, Inc., 2023. ISBN: 978-1-119-80914-2.

[22] Geoffrey Hinton. "Neural Networks for Machine Learning". Massive Open Online Course (Coursera). 2012. URL: https://www.cs.toronto.edu/~hinton/coursera_lectures.html (visited on 2021-11-16).

[23] C. A. R. Hoare. "Quicksort". In: *The Computer Journal* 5.1 (1962-01-01), pp. 10–16. ISSN: 0010-4620. DOI: 10.1093/comjnl/5.1.10.

[24] 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)*. 2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS). 2022-10, pp. 312–322. DOI: 10.1109/FOCS54457.2022.00037.

[25] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *Proceedings of the 32nd International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, 2015-06-01, pp. 448–456. arXiv: 1502.03167 [cs]. URL: https://proceedings.mlr.press/v37/ioffe15.html (visited on 2021-10-06).

[26] E. T. Jaynes. "Information Theory and Statistical Mechanics". In: *Physical Review* 106.4 (1957-05-15), pp. 620–630. DOI: 10.1103/PhysRev.106.620.

[27] Richard Arnold Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. 6th ed. Upper Saddle River, N.J: Pearson College Div, 2007. 767 pp. ISBN: 978-0-13-187715-3.

[28] Steven M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. USA: Prentice-Hall, Inc., 1993-02. 595 pp. ISBN: 978-0-13-345711-7.

[29] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *Proceedings of the 3rd International Conference on Learning Representations*. ICLR. San Diego, 2015. arXiv: 1412.6980.

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

[31] Alex Krizhevsky. *Learning Multiple Layers of Features from Tiny Images*. 2009. URL: https://www.cs.toronto.edu/%20kriz/learning-features-2009-TR.pdf (visited on 2024-05-21).

[32] H. J. Kushner. "A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise". In: *Journal of Basic Engineering* 86.1 (1964-03-01), pp. 97–106. ISSN: 0021-9223. DOI: 10.1115/1.3653121.

[33] Jonathan Lacotte and Mert Pilanci. "Optimal Randomized First-Order Methods for Least-Squares Problems". In: *Proceedings of the 37th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, 2020-11-21, pp. 5587–5597. URL: https://proceedings.mlr.press/v119/lacotte20a.html (visited on 2023-11-09).

[34] Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. *THE MNIST DATABASE of Handwritten Digits*. 2010. URL: http://yann.lecun.com/exdb/mnist/ (visited on 2024-01-24).

[35] Daniel James Lizotte. “Practical Bayesian Optimization”. PhD thesis. Alberta, Canada: University of Alberta, 2008. 168 pp.

[36] G. Matheron. “The Intrinsic Random Functions and Their Applications”. In: *Advances in Applied Probability* 5.3 (1973-12), pp. 439–468. ISSN: 0001-8678, 1475-6064. DOI: 10.2307/1425829.

[37] Andrea Montanari. “Optimization of the Sherrington–Kirkpatrick Hamiltonian”. In: *SIAM Journal on Computing* (2021-01-07), FOCS19–1. ISSN: 0097-5397. DOI: 10.1137/20M132016X.

[38] Yurii Evgen’evič Nesterov. *Lectures on Convex Optimization*. Second edition. Springer Optimization and Its Applications; Volume 137. Cham: Springer, 2018. ISBN: 978-3-319-91578-4. DOI: 10.1007/978-3-319-91578-4.

[39] Misha Padidar et al. “Scaling Gaussian Processes with Derivative Information Using Variational Inference”. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 6442–6453. URL: https://proceedings.neurips.cc/paper/2021/hash/32bbf7b2bc4ed14eb1e9c2580056a989-Abstract.html (visited on 2023-05-17).

[40] Courtney Paquette et al. “Halting Time Is Predictable for Large Models: A Universality Property and Average-Case Analysis”. In: *Foundations of Computational Mathematics* 23.2 (2022-02), pp. 597–673. ISSN: 1615-3383. DOI: 10.1007/s10208-022-09554-y. arXiv: 2006.04299 [math, stat].

[41] Elliot Paquette and Thomas Trogdon. “Universality for the Conjugate Gradient and MINRES Algorithms on Sample Covariance Matrices”. In: *Communications on Pure and Applied Mathematics* 76.5 (2022-09-01), pp. 1085–1136. ISSN: 1097-0312. DOI: 10.1002/cpa.22081.

[42] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. “On the Difficulty of Training Recurrent Neural Networks”. In: *Proceedings of the 30th International Conference on Machine Learning*. International Conference on Machine Learning. Atlanta: PMLR, 2013-05-26, pp. 1310–1318. URL: https://proceedings.mlr.press/v28/pascanu13.html (visited on 2024-04-02).

[43] Fabian Pedregosa and Damien Scieur. “Acceleration through Spectral Density Estimation”. In: *Proceedings of the 37th International Conference on Machine Learning*. International Conference on Machine Learning. Virtual Event (formerly Vienna): PMLR, 2020-11-21, pp. 7553–7562. URL: https://proceedings.mlr.press/v119/pedregosa20a.html (visited on 2023-11-09).

[44] Carl Edward Rasmussen and Christopher K.I. Williams. *Gaussian Processes for Machine Learning*. 2nd ed. Adaptive Computation and Machine Learning 3. Cambridge, Massachusetts: MIT Press, 2006. 248 pp. ISBN: 0-262-18253-X. URL: http://gaussianprocess.org/gpml/chapters/RW.pdf (visited on 2023-04-06).

[45] Filip de Roos, Alexandra Gessner, and Philipp Hennig. “High-Dimensional Gaussian Process Inference with Derivatives”. In: *Proceedings of the 38th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, 2021-07-01, pp. 2535–2545. URL: https://proceedings.mlr.press/v139/de-roos21a.html (visited on 2023-05-15).

[46] 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*. Vol. 29. Barcelona, Spain: Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper/2016/hash/ed265bc903a5a097f61d3ec064d96d2e-Abstract.html (visited on 2023-10-16).

[47] Bobak Shahriari et al. “Taking the Human Out of the Loop: A Review of Bayesian Optimization”. In: *Proceedings of the IEEE* 104.1 (2016-01), pp. 148–175. ISSN: 1558-2256. DOI: 10.1109/JPROC.2015.2494218.

[48] Leslie N. Smith. *A Disciplined Approach to Neural Network Hyper-Parameters: Part 1 – Learning Rate, Batch Size, Momentum, and Weight Decay*. 2018-04-24. DOI: 10.48550/arXiv.1803.09820. arXiv: 1803.09820 [cs, stat]. Pre-published.

[49] Leslie N. Smith. “Cyclical Learning Rates for Training Neural Networks”. In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). 2017-03, pp. 464–472. DOI: 10.1109/WACV.2017.58.

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

[51] Eliran Subag. "Following the Ground States of Full-RSB Spherical Spin Glasses". In: *Communications on Pure and Applied Mathematics* 74.5 (2021), pp. 1021–1044. ISSN: 1097-0312. DOI: 10.1002/cpa.21922.

[52] Ziyu Wang et al. "Bayesian Optimization in a Billion Dimensions via Random Embeddings". In: *Journal of Artificial Intelligence Research* 55 (2016-02-19), pp. 361–387. ISSN: 1076-9757. DOI: 10.1613/jair.4806.

[53] C.K.I. Williams and D. Barber. "Bayesian Classification with Gaussian Processes". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.12 (1998-12), pp. 1342–1351. ISSN: 1939-3539. DOI: 10.1109/34.735807.

[54] Christopher K. I. Williams. "Computation with Infinite Neural Networks". In: *Neural Computation* 10.5 (1998-07-01), pp. 1203–1216. ISSN: 0899-7667. DOI: 10.1162/089976698300017412.

[55] Jian Wu et al. "Bayesian Optimization with Gradients". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/64a08e5f1e6c39faeb90108c430eb120-Abstract.html (visited on 2022-06-02).

[56] Han Xiao, Kashif Rasul, and Roland Vollgraf. *Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms*. 2017-09-15. DOI: 10.48550/arXiv.1708.07747. arXiv: 1708.07747 [cs, stat]. Pre-published.

[57] Matthew D. Zeiler. "ADADELTA: An Adaptive Learning Rate Method". 2012-12-22. arXiv: 1212.5701 [cs].

[58] Guodong Zhang et al. "Which Algorithmic Choices Matter at Which Batch Sizes? Insights From a Noisy Quadratic Model". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019. arXiv: 1907.04164 [cs, stat]. URL: https://proceedings.neurips.cc/paper/2019/hash/e0eacd983971634327ae1819ea8b6214-Abstract.html (visited on 2023-11-09).

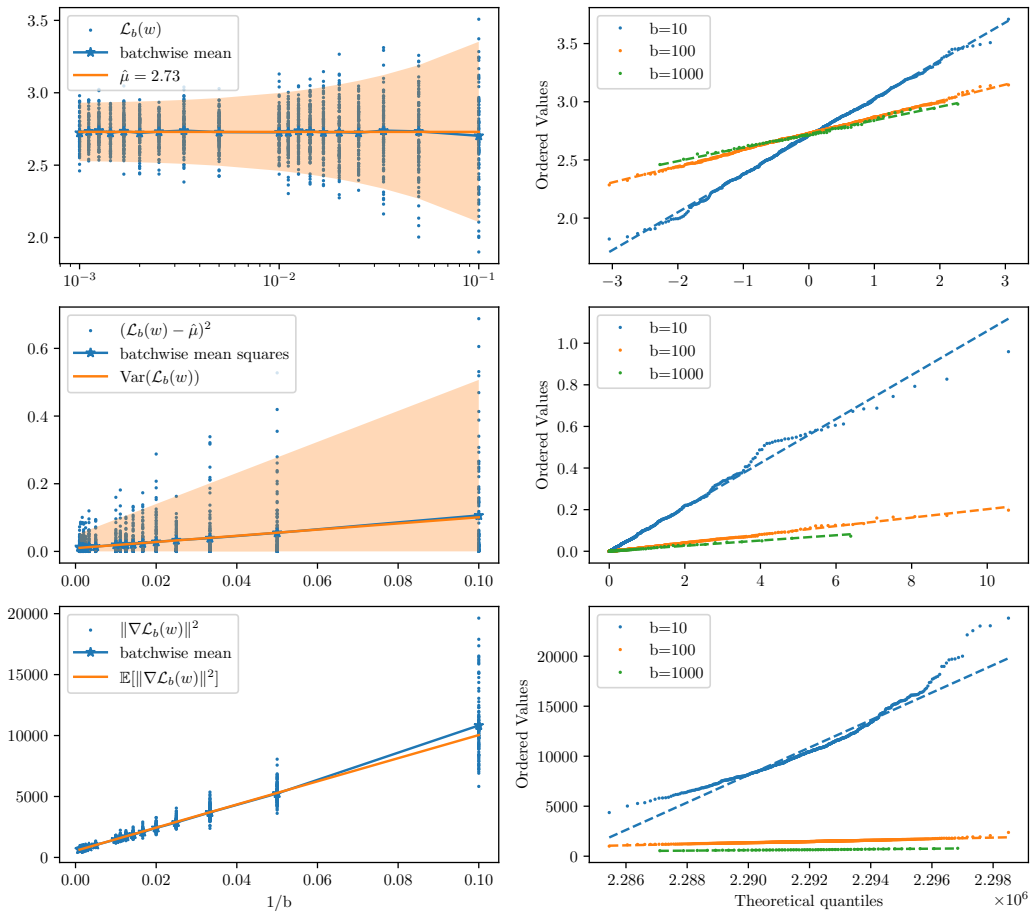

Figure 4: Visualization of the variance estimation (Section 6.1) with 95%-confidence intervals based on the assumed distribution. Quantile-quantile (QQ) plots of the losses (against a normal distribution), squared losses (against a $\chi^2(1)$ distribution) and squared gradient norms (against a $\chi^2(d)$-distribution) are displayed on the right for a selection of batch sizes.

Appendix: Random Function Descent

A Experiments

A.1 Covariance estimation

In Figure 4 we visualize weighted least squares (WLS) regression of the covariance estimation from Section 6.1. Note, that we sampled much more samples per batch size for these plots than RFD would typically require by itself in order to be able to plot batch-wise means and batch-wise QQ-plots. The batch size distribution we described in Section B.1 would avoid sampling the same batch size multiple times to ensure better stability of the regression and generally requires much fewer samples than were used for this visualization (cf. A.1.1)

We can observe from the QQ-plots on the right, that the Gaussian assumption is essentially justified for the losses, resulting in a $\chi^2(1)$ distribution for the squared losses and a $\chi^2(d)$ distribution for the gradient norms squared. The confidence interval estimate for the squared norms appears to be much too small (it is plotted, but too small to be visible). Perhaps this signifies a violation of the isotropy

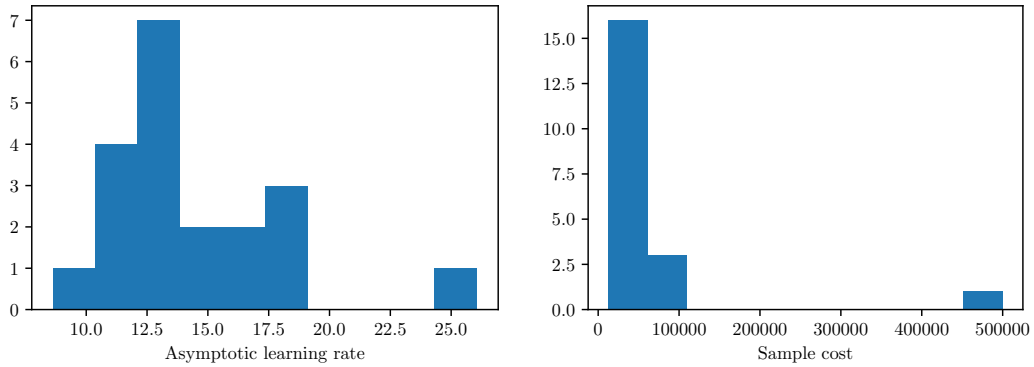

Figure 5: 20 repeated covariance estimations of model M7 [3] applied to the MNIST dataset. On the left are the resulting asymptotic learning rates (assuming a final loss of zero) and on the right are the samples used until the stopping criterion interrupted sampling.

assumption as the variance of

$$\|\nabla \mathcal{L}_b(w)\|^2 = \sum_{i=1}^{d} (\partial_i \mathcal{L}_b(w))^2$$

does not appear to be the variance of independent $\chi^2(d)$ Gaussian random variables, and the independence only follows from the isotropy assumption.

A.1.1 Sampling efficiency and stability

To evaluate the sampling efficiency and stability of our variance estimation process, we repeated the covariance estimation of the model model M7 [3] applied to the MNIST dataset 20 times (Figure 5). We used a tolerance of $\mathrm{tol} = 0.3$ as a stopping criterion for the estimated relative standard deviation (10).

At this tolerance, the asymptotic learning rate already seems relatively stable (in the same order of magnitude) and the sample cost is quite cheap. The majority of runs ($16/20$ runs or 80%) required less than $60\,000$ samples (1 epoch). There was one large outlier which used $500\,589$ samples. A closer inspection revealed, that after the initial sample to estimate the optimal batch size distribution, it sampled almost exclusively at batch sizes 20 (which was the minimal cutoff to avoid instabilities caused by batch normalization) and batch sizes between 1700 and 1900. It therefore seems like the initial batch of samples caused a very unfavorable batch size distribution which then required a lot of samples to recover from. Our selection of an initial sample size of 6000 might therefore have been too small.

A more extensive empirical study is needed to tune this estimation process, but the process promises to be very sample efficient. Classical step size tuning would train models for a short duration in order to evaluate the performance of a particular learning rate [e.g. 48], but a single epoch worth of samples is very hard to beat.

Our implementation of this process on the other hand is very inefficient as of writing. Piping data of differing batch sizes into a model is not a standard use case. We implement this by repeatedly initializing data loaders, which is anything but performance friendly.

A.2 Other models and datasets

To estimate the effect of the batch size on RFD, we trained the same model (M7 [3]) on MNIST with batch size 128 (Figure 6). We can see that the asymptotic learning rate of S-RFD is reduced at a smaller batch size (cf. Equation 7) but the performance is barely different. Overall, RFD seems to be slightly too risk-affine, selecting larger step sizes than the tuned SGD models.

We also trained a different model (M5 [3]) on the Fashion MNIST dataset [56] with batch size 128 (Figure 7). Since the validation loss increases after epoch 5, early stopping would have been

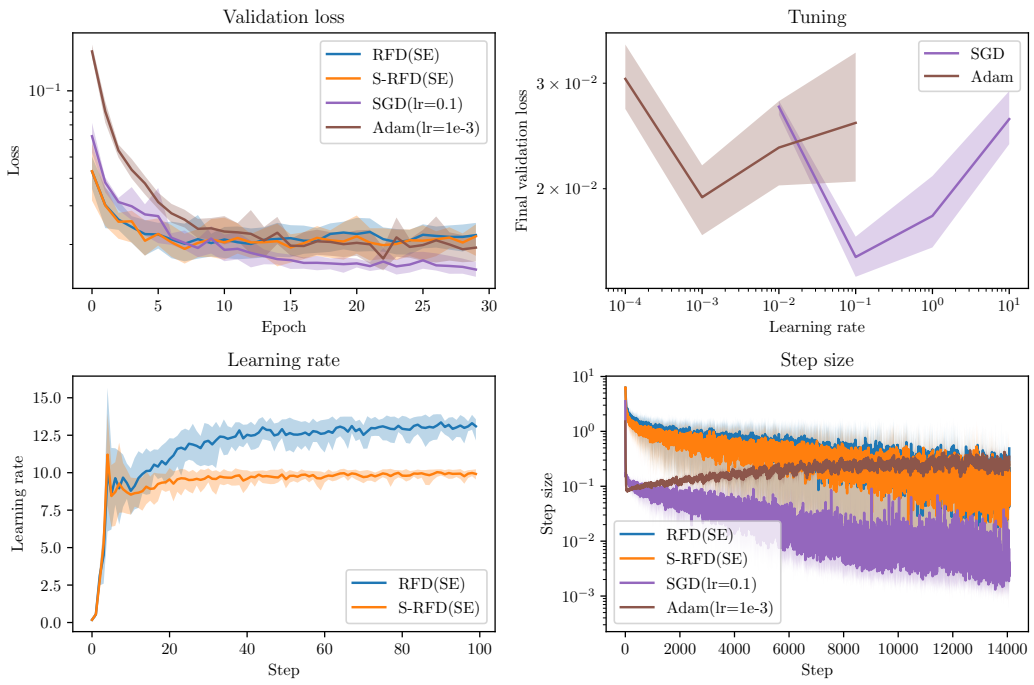

Figure 6: Training model M7 [3] with batch size 128 on MNIST [34].

appropriate. We therefore include Adam with learning rate 10^{-3}, despite Adam with learning rate 10^{-4} technically performing better at the end of training. We can generally see, that RFD comes very close to tuned performance at the time early stopping would have been appropriate. Again, learning rates seem to be slightly too large (risk-affine) in comparison to tuned SGD.

B Variance estimation in detail

Recall that we are interested in the regression

$$Z_b(w) = (\mathcal{L}_b(w) - \mu)^2 \sim \beta_0 + \frac{1}{b}\beta_1$$

where the variance of Z_b is given by

$$\sigma_b^2 = 2(\beta_0 + \tfrac{1}{b}\beta_1)^2.$$

under the Gaussian assumption on $\mathcal{L}_b$.

More specifically we for minibatch sizes $(b_k)_{k\leq n}$ and parameter vectors $(w_k)_{k\leq n}$ we want to sample mini batch losses

$$\mathcal{L}^{(k)} := \mathcal{L}_{b_k}(w_k) = \mathbf{J}(w_k) + \frac{1}{b_k}\sum_{i=1}^{b_k} \epsilon_{k,i}(w_k)$$

As the $\epsilon_{k,i}$ are all conditionally independent and therefore uncorrelated, we have

$$\mathrm{Cov}(\mathcal{L}^{(k)}, \mathcal{L}^{(l)}) = \mathrm{Cov}(\mathbf{J}(w_k), \mathbf{J}(w_l)) = C\big(\tfrac{\|w_k - w_l\|^2}{2}\big)$$

Since the covariance kernel C is typically monotonously falling in the distance of parameters $\|w_k - w_l\|^2$, we want to select them as spaced out as possible to minimize the covariance of $\mathcal{L}^{(k)}$ (which is the next best thing to iid samples). Randomly selecting w_i with Glorot initialization [18] will ensure a good spread.

Note that Glorot initialization places all parameters approximately on the same sphere. This is because Glorot initialization initializes all parameters independently, therefore their norm is the

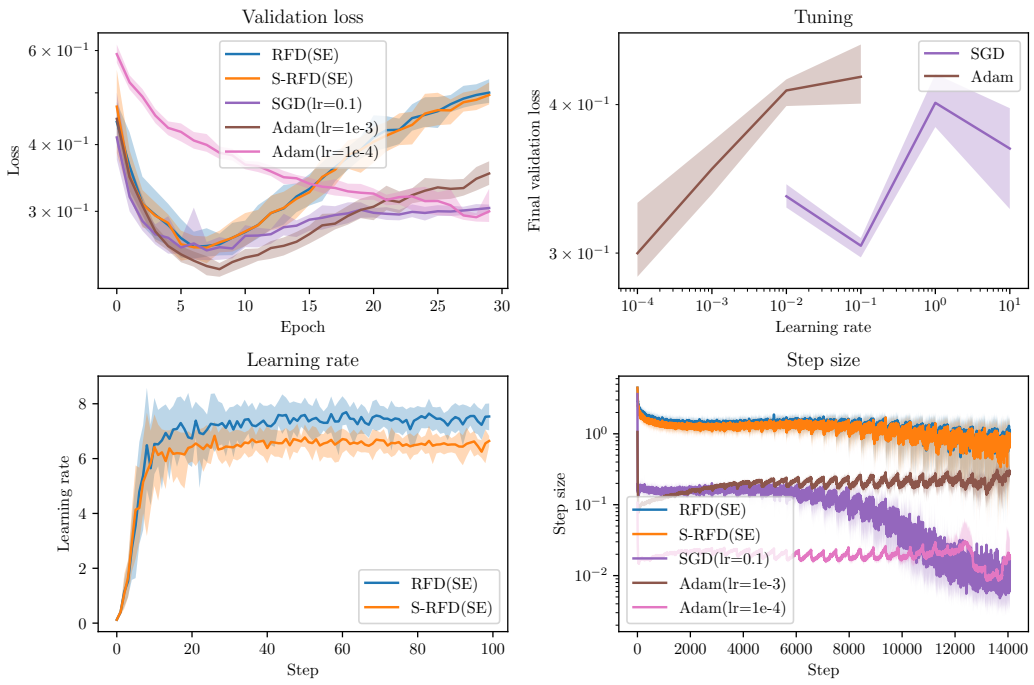

Figure 7: Model M5 [3] trained on Fashion MNIST [56] with batch size 128.

sum of independent squares, which converges by the law of large numbers due to the normalization Glorot uses. Since stationary isotropy and non-stationary isotropy coincides on the sphere, this is an important effect to consider (cf. Section F).

What is left, is the selection of the batch sizes b_k.

B.1 Batch size distribution

Since we plan to use the data set $(\frac{1}{b_k}, \mathcal{L}^{(k)})_{k \leq n}$ for weighted least squares (WLS) regression and do not have a selection process for the batch sizes b_k yet, it might be appropriate to select the batch sizes b_k in such a way, that the variance of our estimator $\hat{\beta}_0$ of β_0 is minimized. Here we choose $\mathrm{Var}(\hat{\beta}_0)$ and not $\mathrm{Var}(\hat{\beta}_1)$ as our optimization target, since $\beta_0 = C(0)$ is used to fit the covariance model, while $\beta_1 = C_\epsilon(0)$ is only required for S-RFD. Without deeper analysis β_0 therefore seems to be more important.

Optimization over n parameters b_k is quite difficult, but we can simplify this optimization problem by considering the empirical batch size distribution

$$\nu_n = \frac{1}{n} \sum_{k=1}^{n} \delta_{b_k}.$$

Using a random variable B distributed according to ν_n, the total number of sample losses can then be expressed as

$$\sum_{k=1}^{n} b_k = n\mathbb{E}[B] = \text{samples used}.$$

Under an (unrealistic) independence assumption, the variance $\mathrm{Var}(\hat{\beta}_0)$ also has a simple representation in terms of ν_n (Lemma B.2). We now want to minimize this variance subject to compute

constraint α limiting the number of sample losses we can use resulting in the optimization problem

$$\mathrm{Var}(\hat{\beta}_0) = \frac{1}{n} \underbrace{\frac{1}{\mathbb{E}[\frac{1}{\sigma_B^2}]}}_{\text{'variance of } Z_B\text{'}} \underbrace{\frac{\mathbb{E}[\frac{1}{\sigma_B^2 B^2}]}{\mathbb{E}\Big[\frac{1}{\sigma_B^2}\big(\frac{1}{B} - \mathbb{E}[\frac{1}{B\sigma_B^2\mathbb{E}[1/\sigma_B^2]}]\big)^2\Big]}}_{\text{inverse of the 'spread' of } \frac{1}{B}} \quad \text{s.t.} \quad n\mathbb{E}[B] \leq \alpha. \tag{8}$$

where we recall that σ_B^2 is the variance of Z_B. So the inverse of the expectation of its inverse is roughly the average variance of Z_B. The second half is the fraction of a weighted second moment divided by the weighted variance. Unless the mean is at zero, the former will be larger. In particular we want a spread of data otherwise the variance would be zero. This is in some conflict with the variance of Z_B.

But first, let us get rid of n. Note that we would always increase n until our compute budget is used up, since this always reduces variance. So we approximately have $n\mathbb{E}[B] = \alpha$. Thus

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\mathbb{E}[B]}{\alpha} \frac{1}{\mathbb{E}[\frac{1}{\sigma_B^2}]} \frac{\mathbb{E}[\frac{1}{\sigma_B^2 B^2}]}{\mathbb{E}\Big[\frac{1}{\sigma_B^2}\big(\frac{1}{B} - \mathbb{E}[\frac{1}{B\sigma_B^2\mathbb{E}[1/\sigma_B^2]}]\big)^2\Big]}$$

Since α is now just resulting in a constant factor, it can be assumed to be 1 without loss of generality. Over batch size distributions ν we therefore want to solve the minimization problem

$$\min_{\nu} \underbrace{\frac{\mathbb{E}[B]}{\mathbb{E}[\frac{1}{\sigma_B^2}]}}_{\text{moments}} \underbrace{\frac{\mathbb{E}[\frac{1}{\sigma_B^2 B^2}]}{\mathbb{E}\Big[\frac{1}{\sigma_B^2}\big(\frac{1}{B} - \mathbb{E}[\frac{1}{B\sigma_B^2\mathbb{E}[1/\sigma_B^2]}]\big)^2\Big]}}_{\text{spread}} \tag{9}$$

Example B.1 (If we did not require spread)**.** If we were not concerned with the variance of batch sizes, we could select a constant $B = b$. Then it is straightforward to minimize the moments factor manually

$$\min_b \frac{\mathbb{E}[b]}{\mathbb{E}[\frac{1}{\sigma_b^2}]} = b\sigma_b^2 = 2b(\beta_0 + \frac{1}{b}\beta_1)^2,$$

resulting in $\frac{1}{b} = \frac{\beta_0}{\beta_1}$. In other words: If we did not have to be concerned with the spread of B there is one optimal selection to minimize the first factor. But in reality we have to trade-off this target with the spread of B.

To ensure a good spread of data, we use the maximum entropy distribution for B, with the moment constraints

$$\mathbb{E}[-B] \geq -\frac{\alpha}{n} \qquad \text{average sample usage}$$
$$\mathbb{E}[\frac{1}{\sigma_B^2}] \geq \theta \qquad Z_B \text{ variance}$$

which capture the first factor. Maximizing entropy under moment constraints is known [26] to result in the Boltzmann (a.k.a. Gibbs) distribution

$$\nu(b) = \mathbb{P}(B = b) \propto \exp\Big(\lambda_1 \frac{1}{\sigma_b^2} - \lambda_2 b\Big),$$

where λ_1, λ_2 depend on the momentum constraints. We can now forget the origin of this distribution and use λ_1, λ_2 as parameters for the distribution ν in Equation (9) to get close to its minimum. In practice we use a zero order black box optimizer (Nelder-Mead [17]). One could calculate the expectations of (9) under this distribution explicitly and take manual derivatives with respect to λ_i to investigate this further, but we wanted to avoid getting too distracted by this tangent.

We also use the estimated relative standard deviation

$$\text{rel\_std} = \frac{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_0)}}{\hat{\beta}_0} \tag{10}$$

as a stopping criterion for sampling. Without extensive testing we found a tolerance of $\text{rel\_std} < \text{tol} = 0.3$ to be reasonable, cf. Section A.1.1.

Lemma B.2 (Variance of $\hat{\beta}_0$ in terms of the empirical batch size distribution)**.** *Assuming independence of the samples* $((\frac{1}{b_k}), Z_{b_k})_{k\leq n}$*, the variance of* $\hat{\beta}_0$ *is given by*

$$\operatorname{Var}(\hat{\beta}_0) = \frac{1}{n}\frac{1}{\mathbb{E}[\frac{1}{\sigma_B^2}]}\frac{\mathbb{E}[\frac{1}{\sigma_B^2 B^2}]}{\mathbb{E}\Big[\frac{1}{\sigma_B^2}\big(\frac{1}{B} - \mathbb{E}[\frac{1}{B\sigma_B^2\mathbb{E}[1/\sigma_B^2]}]\big)^2\Big]}$$

where B *is distributed according to the empirical batch size distribution* $\nu_n = \frac{1}{n}\sum_{k=1}^n \delta_{b_k}$*.*

Proof. With the notation $\sigma_k^2 = \sigma_{b_k}^2$ to describe the variance of Z_{b_k} it follows from [cf. 28, Thm. 4.2] that the variance of the estimator $\hat{\beta}$ of β using n samples is given by

$$\begin{aligned}\operatorname{Var}(\hat{\beta}) &= (H^T C^{-1} H)^{-1} \\ &= \frac{1}{\big(\sum_k \frac{1}{\sigma_k^2}\big)\big(\sum_k \frac{1}{(\sigma_k b_k)^2}\big) - \big(\sum_k \frac{1}{\sigma_k^2 b_k}\big)^2}\begin{pmatrix} \sum_k \frac{1}{(\sigma_k b_k)^2} & -\sum_k \frac{1}{\sigma_k^2 b_k} \\ -\sum_k \frac{1}{\sigma_k^2 b_k} & \sum_k \frac{1}{\sigma_k^2}\end{pmatrix}\end{aligned}$$

where

$$C := \begin{pmatrix}\sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2\end{pmatrix} \qquad H := \begin{pmatrix}1 & \frac{1}{b_1} \\ \vdots & \\ 1 & \frac{1}{b_n}\end{pmatrix}.$$

In particular we have

$$\operatorname{Var}(\hat{\beta}_0) = \frac{\sum_k \frac{1}{\sigma_k^2 b_k^2}}{\big(\sum_k \frac{1}{\sigma_k^2}\big)\big(\sum_k \frac{1}{\sigma_k^2 b_k^2}\big) - \big(\sum_k \frac{1}{\sigma_k^2 b_k}\big)^2}.$$

With the help of $\theta := \sum_j \frac{1}{\sigma_j^2}$ and $\lambda_k := \frac{1}{\sigma_k^2\theta}$, we can reorder the divisor. For this note that since the λ_k sum to 1 we have

$$\begin{aligned}\sum_k \lambda_k\Big(\frac{1}{b_k} - \sum_j \lambda_j \frac{1}{b_j}\Big)^2 &= \sum_k \lambda_k\Big(\frac{1}{b_k^2} - 2\frac{1}{b_k}\sum_j \lambda_j \frac{1}{b_j} + \Big(\sum_j \lambda_j \frac{1}{b_j}\Big)^2\Big) \\ &= \sum_k \lambda_k \frac{1}{b_k^2} - 2\Big(\sum_k \lambda_k \frac{1}{b_k}\Big) + \Big(\sum_k \lambda_k \frac{1}{b_j}\Big)^2 \\ &= \sum_k \lambda_k \frac{1}{b_k^2} - \Big(\sum_k \lambda_k \frac{1}{b_k}\Big)^2\end{aligned}$$

Where the above is essentially the well known statement $\mathbb{E}[(Y - \mathbb{E}[Y])^2] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$ for an appropriate selection of Y. This implies that our divisor is given by a weighted variance

$$\theta^2 \sum_k \lambda_k\Big(\frac{1}{b_k} - \sum_j \lambda_j \frac{1}{b_j}\Big)^2 = \theta \sum_k \frac{1}{\sigma_k^2 b_k^2} - \Big(\sum_k \frac{1}{\sigma_k^2 b_k}\Big)^2,$$

where it is only necessary to plug in the definition of θ to see the right term is exactly our divisor. Expanding both the enumerator as well as the divisor by $\frac{1}{n}$, we obtain

$$\operatorname{Var}(\hat{\beta}_0) = \frac{1}{\theta}\frac{\frac{1}{n}\sum_k \frac{1}{\sigma_k^2 b_k^2}}{\frac{1}{n}\sum_k \frac{1}{\sigma_k^2}\big(\frac{1}{b_k} - \sum_j \lambda_j \frac{1}{b_j}\big)^2}$$

Since $\theta = n\mathbb{E}[1/\sigma_B^2]$ for $B \sim \frac{1}{n}\sum_{k=1}^n \delta_{b_k}$ and $\lambda_k = \frac{1}{n\sigma_k^2\mathbb{E}[1/\sigma_B^2]}$, the above can thus be written as

$$\operatorname{Var}(\hat{\beta}_0) = \frac{1}{n}\frac{1}{\mathbb{E}[1/\sigma_B^2]}\frac{\mathbb{E}[\frac{1}{\sigma_B^2 B^2}]}{\mathbb{E}\Big[\frac{1}{\sigma_B^2}\big(\frac{1}{B} - \mathbb{E}[\frac{1}{B\sigma_B^2\mathbb{E}[1/\sigma_B^2]}]\big)^2\Big]},$$

which proves our claim. □

C Covariance models

In this section we calculate the step sizes of the covariance models listed in Table 1 and plotted in Figure 2. Additionally we calculate the asymptotic learning rate of A-RFD and prove an Assumption of Corollary 5.3 for the squared exponential covariance (Prop. C.3).

C.1 Squared exponential

The squared exponential covariance function is given by

$$C\big(\tfrac{\|x-y\|^2}{2}\big) = \sigma^2 \exp\big(-\tfrac{\|x-y\|^2}{2s^2}\big). \tag{11}$$

Note that σ^2 will play no role in the step sizes of RFD due to its scale invariance (cf. Advantage 2.3).

Theorem C.1. *Let* $\mathbf{J} \sim \mathcal{N}(\mu, C)$ *where* C *is the* ***squared exponential*** *covariance function* (11)*, then we have*

$$\eta^* \frac{\nabla \mathbf{J}(w)}{\|\nabla \mathbf{J}(w)\|} = \operatorname*{argmin}_{\mathbf{d}} \mathbb{E}[\mathbf{J}(w-\mathbf{d}) \mid \mathbf{J}(w), \nabla \mathbf{J}(w)]$$

with RFD step size

$$\eta^* = \frac{s^2 \|\nabla \mathbf{J}(w)\|}{\sqrt{\big(\frac{\mu - \mathbf{J}(w)}{2}\big)^2 + s^2 \|\nabla \mathbf{J}(w)\|^2} + \frac{\mu - \mathbf{J}(w)}{2}}.$$

Proof. The covariance function C is of the form

$$C(h) = \sigma^2 e^{-\frac{h}{s^2}}.$$

By Equation (2)

$$\eta^* = -\operatorname*{argmin}_{\eta} \tfrac{C(\frac{\eta^2}{2})}{C(0)} - \eta \tfrac{C'(\frac{\eta^2}{2})}{C'(0)} \Theta.$$

where $\Theta = \frac{\|\nabla \mathbf{J}(w)\|}{\mu - \mathbf{J}(w)}$. We calculate

$$-\frac{C\big(\frac{\eta^2}{2}\big)}{C(0)} - \eta \frac{C'\big(\frac{\eta^2}{2}\big)}{C'(0)} \Theta = -e^{-\frac{\eta^2}{2s^2}} (1 + \eta\Theta).$$

This results in the first order condition

$$0 \overset{!}{=} \frac{\eta}{s^2} e^{-\frac{\eta^2}{2s^2}} (1 + \eta\Theta) - e^{-\frac{\eta^2}{2s^2}} \Theta = \frac{e^{-\frac{\eta^2}{2s^2}}}{s^2} (\eta^2 \Theta + \eta - s^2 \Theta).$$

Since the exponential can never be zero, we have to solve a quadratic equation. Its solution results in

$$\eta^*(\Theta) = \sqrt{\big(\tfrac{1}{2\Theta}\big)^2 + s^2} - \tfrac{1}{2\Theta}. \tag{12}$$

At this point we could stop, but the result is numerically unstable as it suffers from catastrophic cancellation. To solve this issue we set $x = \frac{1}{2\Theta}$ and reorder

$$\eta^* = \sqrt{x^2 + s^2} - x = (\sqrt{x^2 + s^2} - x) \frac{\sqrt{x^2 + s^2} + x}{\sqrt{x^2 + s^2} + x} = \frac{\cancel{x^2} + s^2 - \cancel{x^2}}{\sqrt{x^2 + s^2} + x}.$$

Re-substituting $x = \frac{1}{2\Theta} = \frac{\mu - \mathbf{J}(w)}{2\|\nabla \mathbf{J}(w)\|}$, we finally get

$$\eta^* = \frac{s^2}{\sqrt{\big(\frac{\mu - \mathbf{J}(w)}{2\|\nabla \mathbf{J}(w)\|}\big)^2 + s^2} + \frac{\mu - \mathbf{J}(w)}{2\|\nabla \mathbf{J}(w)\|}} = \frac{s^2 \|\nabla \mathbf{J}(w)\|}{\sqrt{\big(\frac{\mu - \mathbf{J}(w)}{2}\big)^2 + s^2 \|\nabla \mathbf{J}(w)\|^2} + \frac{\mu - \mathbf{J}(w)}{2}}. \qquad \square$$

Proposition C.2 (A-RFD for the Squared Exponential Covariance)**.** *If* $\mathbf{J}$ *is isotropic with squared exponential covariance* (11)*, then the step size of A-RFD is given by*

$$\hat{\eta} = \frac{s^2}{\mu - \mathbf{J}(w)} \|\nabla \mathbf{J}(w)\|,$$

Proof. By Definition 5.1 of A-RFD and $\Theta = \frac{\|\nabla \mathbf{J}(w)\|}{\mu - \mathbf{J}(w)}$ we have

$$\hat{\eta}(\Theta) = \frac{C(0)}{-C'(0)}\Theta = \frac{\sigma^2 \exp(0)}{\frac{\sigma^2}{s^2}\exp(0)} \frac{\|\nabla \mathbf{J}(w)\|}{\mu - \mathbf{J}(w)} = s^2 \frac{\|\nabla \mathbf{J}(w)\|}{\mu - \mathbf{J}(w)}.$$ □

Proposition C.3. *If* $\mathbf{J}$ *is isotropic with squared exponential covariance* (11)*, then the RFD step sizes are strictly monotonously increasing in* Θ*.*

Proof. Since we know that $\Theta \to 0$ implies $\eta^* \sim \hat{\eta} \to 0$ strict monotonicity of η^* in Θ is sufficient to show that $\eta^* \to 0$ also implies $\Theta \to 0$. So we take the derivative of (12) resulting in

$$\frac{d}{d\Theta}\eta^* = \frac{1 - \frac{1}{\sqrt{1+s^2(2\Theta)^2}}}{2\Theta^2},$$

which is greater zero for all $\Theta > 0$. □

C.2 Rational quadratic

The **rational quadratic** covariance function is given by

$$C\big(\tfrac{\|x-y\|}{2}\big) = \sigma^2 \left(1 + \frac{\|x-y\|^2}{\beta s^2}\right)^{-\beta/2} \qquad \beta > 0. \tag{13}$$

It can be viewed as a scale mixture of the squared exponential and converges to the squared exponential in the limit $\beta \to \infty$ [44, p. 87].

Theorem C.4 (Rational Quadratic)**.** *For* $\mathbf{J} \sim \mathcal{N}(\mu, C)$ *where* C *is the* ***rational quadratic covariance*** *we have for* $\Theta = \frac{\|\nabla \mathbf{J}(w)\|}{\mu - \mathbf{J}(w)} \geq 0$ *that the RFD step size is given by*

$$\eta^* = s\sqrt{\beta}\operatorname*{Root}_{\eta}\left(-1 + \tfrac{\sqrt{\beta}}{s\Theta}\eta + (1+\beta)\eta^2 + \tfrac{\sqrt{\beta}}{s\Theta}\eta^3\right).$$

The unique root of the polynomial in η *can be found either directly with a formula for polynomials of third degree (e.g. using Cardano's method) or by bisection as it is contained in* $[0, 1/\sqrt{1+\beta}]$*.*

Proof. By Theorem 4.2 we have

$$\eta^* = \operatorname*{argmin}_{\eta} -\frac{C\big(\frac{\eta^2}{2}\big)}{C(0)} - \eta\frac{C'\big(\frac{\eta^2}{2}\big)}{C'(0)}\Theta$$

for $C(x) = \sigma^2(1 + \frac{2x}{\beta s^2})^{-\beta/2}$. We therefore need to minimize

$$f\big(\frac{\eta}{\sqrt{\beta}s}\big) := -\left(1 + \frac{\eta^2}{\beta s^2}\right)^{-\beta/2} - \eta\left(1 + \frac{\eta^2}{\beta s^2}\right)^{-\beta/2-1}\Theta.$$

Substitute in $\tilde{\eta} := \frac{\eta}{\sqrt{\beta}s}$, then the first order condition is

$$0 \overset{!}{=} f'(\tilde{\eta}) = -\frac{d}{d\tilde{\eta}}\left(1+\tilde{\eta}^2\right)^{-\beta/2} + \sqrt{\beta}s\tilde{\eta}\left(1+\tilde{\eta}^2\right)^{-\beta/2-1}\Theta$$

Dividing both sides by $\sqrt{\beta}s\Theta$ we get

$$\begin{aligned} 0 = \frac{f'(\tilde{\eta})}{\sqrt{\beta}s\Theta} &= \tfrac{\beta}{2}(1+\tilde{\eta}^2)^{-\frac{\beta}{2}-1}2\tilde{\eta}\tfrac{1}{\sqrt{\beta}s\Theta} + (1+\tilde{\eta}^2)^{-\frac{\beta}{2}-2}\left[1 + \tilde{\eta}^2 - (\tfrac{\beta}{2}+1)2\tilde{\eta}^2\right] \\ &= (1+\tilde{\eta}^2)^{-\frac{\beta}{2}-2}\underbrace{\left[\beta\tilde{\eta}\tfrac{1}{\sqrt{\beta}s\Theta}(1+\tilde{\eta}^2) - [1 - \tilde{\eta}^2(1+\beta)]\right]}_{=-1+\frac{\sqrt{\beta}}{s\Theta}\tilde{\eta}+(1+\beta)\tilde{\eta}^2+\frac{\sqrt{\beta}}{s\Theta}\tilde{\eta}^3} \end{aligned}$$

Since $\Theta \geq 0$ and $\beta > 0$ all coefficients of the polynomial are positive except for the shift. The polynomial thus starts out at -1 in zero and only increases from there. Therefore there exists a unique positive critical point which is a minimum.

At the point $\tilde{\eta} = \sqrt{1+\beta}$ the quadratic term is already larger than 1 so the polynomial is positive and we have passed the root. The minimum is therefore contained in the interval $[0, \sqrt{1+\beta}]$.

After finding the minimum in $\tilde{\eta}$ we return to η by multiplication with $\sqrt{\beta}s$. □

Proposition C.5 (A-RFD for the Rational Quadratic Covariance)**.** *If* $\mathbf{J}$ *is isotropic with rational quadratic covariance* (13)*, then the step size of A-RFD is given by*

$$\hat{\eta} = \frac{s^2}{\mu - \mathbf{J}(w)} \|\nabla \mathbf{J}(w)\|.$$

Proof. $C(x) = \sigma^2(1 + \frac{2x}{\beta s^2})^{-\beta/2}$ implies by Definition 5.1 of A-RFD and $\Theta = \frac{\|\nabla \mathbf{J}(w)\|}{\mu - \mathbf{J}(w)}$

$$\hat{\eta}(\Theta) = \frac{C_{\mathbf{J}}(0)}{-C'_{\mathbf{J}}(0)}\Theta = \frac{\sigma^2(1+0)^{-\beta/2}}{\frac{\sigma^2}{s^2}(1+0)^{-\beta/2-1}} \frac{\|\nabla \mathbf{J}(x)\|}{\mu - \mathbf{J}(x)} = s^2 \frac{\|\nabla \mathbf{J}(x)\|}{\mu - \mathbf{J}(x)}.$$ □

C.3 Matérn

Definition C.6. The Matérn model parametrized by $s > 0, \nu \geq 0, \sigma^2 \geq 0$ is given by

$$C\big(\tfrac{\|x-y\|^2}{2}\big) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\|x-y\|}{s}\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}\|x-y\|}{s}\right) \tag{14}$$

where K_ν is the modified Bessel function.

For $\nu = p + \frac{1}{2}$ with $p \in \mathbb{N}_0$, it can be simplified [cf. 44, sec. 4.2.1] to

$$C\big(\tfrac{\|x-y\|^2}{2}\big) = \sigma^2 e^{-\frac{\sqrt{2\nu}\|x-y\|}{s}} \frac{p!}{(2p)!} \sum_{k=0}^{p} \frac{(2p-k)!}{(p-k)!k!} \left(\frac{2\sqrt{2\nu}}{s}\|x-y\|\right)^k$$

The Matérn model encompasses Rasmussen and Williams [44]

- the **nugget effect** for $\nu = 0$ (independent randomness)
- the **exponential model** for $\nu = \frac{1}{2}$ (Ornstein-Uhlenbeck process)
- the **squared exponential model** for $\nu \to \infty$ with the same scale s and variance σ^2.

The random functions induced by the Matérn model are a.s. $\lfloor \nu \rfloor$-times differentiable Rasmussen and Williams [44], i.e. the smoothness of the model increases with increasing ν. While the exponential covariance model with $\nu = \frac{1}{2}$ results in a random function which is not yet differentiable, larger ν result in increasing differentiability. As differentiability starts with $\nu = \frac{3}{2}$ and we have a more explicit formula for $\nu = p + \frac{1}{2}$ the cases $\nu = \frac{3}{2}$ and $\nu = \frac{5}{2}$ are of particular interest.

> "[F]or $\nu \geq 7/2$, in the absence of explicit prior knowledge about the existence of higher order derivatives, it is probably very hard from finite noisy training examples to distinguish between values of $\nu \geq 7/2$ (or even to distinguish between finite values of ν and $\nu \to \infty$, the smooth squared exponential, in this case)" [44, p. 85].

Theorem C.7. *Assuming* $\mathbf{J} \sim \mathcal{N}(\mu, C)$ *is a random function where* C *is the Matérn covariance such that* $\nu = p + \frac{1}{2}$ *with* $p \in \{1, 2\}$*. Then the RFD step is given for* $\Theta := \frac{\|\nabla \mathbf{J}(w)\|}{\mu - \mathbf{J}(w)} \geq 0$ *by*

- $p = 1$
$$\eta^* = \frac{s}{\sqrt{3}} \frac{1}{\left(1 + \frac{\sqrt{3}}{s\Theta}\right)}$$
- $p = 2$
$$\eta^* = \frac{s}{\sqrt{5}} \frac{(1-\zeta) + \sqrt{4 + (1+\zeta)^2}}{2(1+\zeta)} \qquad \zeta := \frac{\sqrt{5}}{3s\Theta}.$$

Proof. We define $\mathcal{C}(\eta) := C(\frac{\eta^2}{2})$, which implies

$$\mathcal{C}'(\eta) = C'\big(\tfrac{\eta^2}{2}\big)\eta$$

or conversely

$$C'\big(\tfrac{\eta^2}{2}\big) = \frac{1}{\eta}\mathcal{C}'(\eta). \tag{15}$$

By Theorem 4.2, we need to calculate

$$\eta^* = \operatorname*{argmin}_{\eta} -\tfrac{C(\frac{\eta^2}{2})}{C(0)} - \eta\tfrac{C'(\frac{\eta^2}{2})}{C'(0)}\Theta. \tag{16}$$

Discarding σ w.l.o.g. due to scale invariance (Advantage 2.3), we have in the case $p = 1$

$$\mathcal{C}(\eta) = \Big(1 + \frac{\sqrt{3}}{s}\eta\Big)\exp\Big(-\frac{\sqrt{3}}{s}\eta\Big).$$

The derivative is then given by

$$\mathcal{C}'(\eta) = -\big(\tfrac{\sqrt{3}}{s}\big)^2\eta\exp\big(-\tfrac{\sqrt{3}}{s}\eta\big)$$

which implies using (15)

$$C'\big(\tfrac{\eta^2}{2}\big) = -\big(\tfrac{\sqrt{3}}{s}\big)^2\exp\big(-\tfrac{\sqrt{3}}{s}\eta\big) \tag{17}$$

We therefore need to minimize (16) which is given by

$$\operatorname*{argmin}_{\eta} -\big(1 + \tfrac{\sqrt{3}}{s}\eta\big)\exp\big(-\tfrac{\sqrt{3}}{s}\eta\big) - \eta\exp\big(-\tfrac{\sqrt{3}}{s}\eta\big)\Theta = \operatorname*{argmin}_{\eta} -\big(1 + (\tfrac{\sqrt{3}}{s} + \Theta)\eta\big)\exp\big(-\tfrac{\sqrt{3}}{s}\eta\big).$$

The first order condition is

$$0 \overset{!}{=} \Big(\tfrac{\sqrt{3}}{s}\big(\cancel{1} + (\tfrac{\sqrt{3}}{s} + \Theta)\eta\big) - (\cancel{\tfrac{\sqrt{3}}{s}} + \Theta)\Big)$$

which (divided by Θ and noting that the exponential can never be zero) is equivalent to

$$0 \overset{!}{=} \tfrac{\sqrt{3}}{s}(\tfrac{\sqrt{3}}{s\Theta} + 1)\eta - 1$$

reordering for η implies

$$\eta \overset{!}{=} \frac{s}{\sqrt{3}}\frac{1}{\big(1 + \frac{\sqrt{3}}{s\Theta}\big)}.$$

It is also not difficult to see that this is the point where the derivative switches from negative to positive (i.e. a minimum).

Let us now consider the case $p = 2$, i.e.

$$\mathcal{C}(\eta) = \big(1 + \tfrac{\sqrt{5}}{s}\eta + \tfrac{5}{3s^2}\eta^2\big)\exp\big(-\tfrac{\sqrt{5}}{s}\eta\big),$$

which results in

$$\mathcal{C}'(\eta) = -\tfrac{5}{3s^2}\big(\eta + \tfrac{\sqrt{5}}{s}\eta^2\big)\exp\big(-\tfrac{\sqrt{5}}{s}\eta\big),$$

i.e. by (15)

$$C'\big(\tfrac{\eta^2}{2}\big) = -\tfrac{5}{3s^2}\big(1 + \tfrac{\sqrt{5}}{s}\eta\big)\exp\big(-\tfrac{\sqrt{5}}{s}\eta\big). \tag{18}$$

We therefore need to minimize (16) which is given by

$$\underbrace{\Big(-\big(1 + \tfrac{\sqrt{5}}{s}\eta + \tfrac{5}{3s^2}\eta^2\big) - \eta\big(1 + \tfrac{\sqrt{5}}{s}\eta\big)\Theta\Big)}_{=-\left(1+\left(\frac{\sqrt{5}}{s}+\Theta\right)\eta+\left(\frac{5}{3s^2}+\frac{\sqrt{5}}{s}\Theta\right)\eta^2\right)}\exp\big(-\tfrac{\sqrt{5}}{s}\eta\big).$$

The first order condition results in

$$\begin{aligned}0 &\overset{!}{=} \tfrac{\sqrt{5}}{s}\Big(\cancel{1} + \big(\tfrac{\sqrt{5}}{s} + \Theta\big)\eta + \big(\tfrac{5}{3s^2} + \tfrac{\sqrt{5}}{s}\Theta\big)\eta^2\Big) - \Big(\big(\cancel{\tfrac{\sqrt{5}}{s}} + \Theta\big) + 2\big(\tfrac{5}{3s^2} + \tfrac{\sqrt{5}}{s}\Theta\big)\eta\Big)\\ &= -\Theta + \big(\tfrac{5}{3s^2} - \tfrac{\sqrt{5}}{s}\Theta\big)\eta + \tfrac{\sqrt{5}}{s}\big(\tfrac{5}{3s^2} + \tfrac{\sqrt{5}}{s}\Theta\big)\eta^2\end{aligned}$$

Dividing everything by Θ and using $\zeta := \frac{\sqrt{5}}{3s\Theta}$ we get

$$0 \overset{!}{=} -1 - \big(\zeta - 1\big)\big(\tfrac{\sqrt{5}}{s}\eta\big) + \big(\zeta + 1\big)\big(\tfrac{\sqrt{5}}{s}\eta\big)^2$$

Taking a closer look at the sign changes of the derivative it becomes obvious, that the positive root is the minimum, i.e.

$$\frac{\sqrt{5}}{s}\eta \overset{!}{=} \frac{(1-\zeta) + \sqrt{(1-\zeta)^2 + 4(1+\zeta)}}{2(1+\zeta)} = \frac{(1-\zeta) + \sqrt{4 + (1+\zeta)^2}}{2(1+\zeta)}. \qquad \square$$

Proposition C.8 (A-RFD for the Matérn Covariance)**.** *If* $\mathbf{J}$ *is isotropic with Matérn covariance* (14) *such that* $\nu = p + \frac{1}{2}$*, then the step size of A-RFD for* $p \in \{1, 2\}$ *is given by*

- $p = 1$
$$\hat{\eta} = \frac{s^2}{3} \frac{\|\nabla \mathbf{J}(x)\|}{\mu - \mathbf{J}(x)}$$
- $p = 2$
$$\hat{\eta} = \frac{3s^2}{5} \frac{\|\nabla \mathbf{J}(x)\|}{\mu - \mathbf{J}(x)}$$

Proof. Noting $\Theta = \frac{\|\nabla \mathbf{J}(x)\|}{\mu - \mathbf{J}(x)}$, we have by Definition 5.1 of A-RFD for $p = 1$

$$\hat{\eta} = \frac{C(0)}{-C'(0)} \Theta \overset{(17)}{=} \frac{s^2}{3} \Theta,$$

and in the case $p = 2$

$$\hat{\eta} = \frac{C(0)}{-C'(0)} \Theta \overset{(18)}{=} \frac{3s^2}{5} \Theta. \qquad \square$$

D Proofs

In this section we prove all the claims made in the main body.

D.1 Section 2: Random function descent

D.1.1 Formal RFD

As we mentioned in a footnote at the definition of RFD, the fact that the parameters become random variables as they are selected by random gradients poses some mathematical challenges which would have been distracting to address in the main body. In following paragraphs leading up to Definition D.1 we introduce and discuss the probability theory required to provide a mathematically sound definition.

For a fixed cost distribution $\mathbb{P}_{\mathbf{J}}$ and any weight vectors w and $\tilde{w}$ the conditional distribution

$$\mathbb{E}[\mathbf{J}(\tilde{w}) \mid \mathbf{J}(w), \nabla \mathbf{J}(w)]$$

is by its axiomatic definition a $(\mathbf{J}(w), \nabla \mathbf{J}(w))$-measurable random variable. By the factorization lemma [30, Cor. 1.9.7], there therefore exists a measurable function $(j, g) \mapsto \varphi_{w,\tilde{w}}(j, g)$ such that the following equation holds almost surely

$$\varphi_{w,\theta}(\mathbf{J}(w), \nabla \mathbf{J}(w)) = \mathbb{E}[\mathbf{J}(\tilde{w}) \mid \mathbf{J}(w), \nabla \mathbf{J}(w)]. \tag{19}$$

Since it is possible to calculate $\varphi_{w,\tilde{w}}$ explicitly in the Gaussian case (cf. G.1), the function

$$\Phi_{\mathbb{P}_{\mathbf{J}}} : \begin{cases} (\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d) \to \mathbb{R} \\ (w, j, g) \mapsto \operatorname{argmin}_{\tilde{w}} \varphi_{w,\tilde{w}}(j, g), \end{cases}$$

which implements some tie-breaker rules for set valued argmin is measurable when $\mathbf{J}$ is Gaussian and its covariance function is sufficiently smooth. To prove measurability in the general case is a difficult problem of its own, which we do not attempt to solve here, since we would not utilize the conditional expectation outside of the Gaussian case anyway (cf. Section E.3). For deterministic w, we therefore have

$$\begin{aligned} \Phi_{\mathbb{P}_{\mathbf{J}}}(w, \mathbf{J}(w), \nabla \mathbf{J}(w)) &= \operatorname*{argmin}_{\tilde{w}} \varphi_{w,\tilde{w}}(\mathbf{J}(w), \nabla \mathbf{J}(w)) \\ &\overset{(19)}{=} \operatorname*{argmin}_{\tilde{w}} \mathbb{E}[\mathbf{J}(\tilde{w}) \mid \mathbf{J}(w), \nabla \mathbf{J}(w)]. \end{aligned}$$

So if the parameter vectors w_n were deterministic, our formal definition of RFD and our initial definition would coincide. But for random weights W (19) stops to hold in general[3], i.e.

$$\varphi_{W,\tilde{w}}(\mathbf{J}(W), \nabla\mathbf{J}(W)) \neq \mathbb{E}[\mathbf{J}(\tilde{w}) \mid \mathbf{J}(W), \nabla\mathbf{J}(W)].$$

If this equation does not need to hold, we similarly have in general

$$\Phi_{\mathbb{P}_\mathbf{J}}(W, \mathbf{J}(W), \nabla\mathbf{J}(W)) \neq \operatorname*{argmin}_{\tilde{w}} \mathbb{E}[\mathbf{J}(\tilde{w}) \mid \mathbf{J}(W), \nabla\mathbf{J}(W)].$$

So the following definition is not just a restatement of the original definition of RFD.

Definition D.1 (Formal RFD)**.** For a Gaussian random cost function $\mathbf{J}$, we define the RFD algorithm with starting point $W_0 = w_0 \in \mathbb{R}^d$ by

$$W_{n+1} := \Phi_{\mathbb{P}_\mathbf{J}}(W_n, \mathbf{J}(W_n), \nabla\mathbf{J}(W_n))$$

This is what we effectively do in Theorem 4.2 under the additional isotropy assumption, where we calculate the argmin under the assumption that w is deterministic (i.e. we determine $\Phi_{\mathbb{P}_\mathbf{J}}$), before we plug-in the random variables W_n to obtain W_{n+1}. Similarly this is how the step size prescriptions of RFD actually work. We first assume deterministic weights and later plug the random variables into our formulas. For this reason, we avoided large letters indicating random variables for parameters w in the main body.

D.1.2 Scale invariance

Advantage 2.3 (Scale invariance)**.** *RFD is invariant to additive shifts and positive scaling of the cost* $\mathbf{J}$*. RFD is also invariant with respect to transformations of the parameter input of* $\mathbf{J}$ *by differentiable bijections whose Jacobian is invertible everywhere (e.g. invertible linear maps).*

Before we get to the proof, let us quickly formulate the statement in mathematical terms. Let w_n be the parameters selected optimizing $\mathbf{J}$ starting in w_0 and $\tilde{w}_n$ the parameters selected by the same optimizer optimizing $\tilde{\mathbf{J}}$ starting in $\tilde{w}_0$.

If we apply affine linear scaling to cost $\mathbf{J}$ such that $\tilde{\mathbf{J}}(w) = a\mathbf{J}(w) + b$ and start optimization in the same point, i.e. $w_0 = \tilde{w}_0$, then we expect a scale invariant optimizer to select

$$w_n = \tilde{w}_n.$$

If we scale inputs on the other hand (or more generally map them with a bijection ϕ), then we expect for $\tilde{\mathbf{J}} := \mathbf{J} \circ \phi$ and starting point $\tilde{w}_0 = \phi^{-1}(w_0)$, that this relationship is retained by a scale invariant optimizer, i.e.

$$\tilde{w}_n = \phi^{-1}(w_n).$$

Why do we use a different starting point? As an illustrating example, assume that ϕ maps miles into kilometers. Then $\tilde{\mathbf{J}}$ accepts miles, while $\mathbf{J}$ accepts kilometers. Then we have to map the initial starting point w_0 of $\mathbf{J}$ measured in kilometers into miles $\tilde{w}_0$. ϕ^{-1} is precisely this transformation from kilometers into miles. A scale invariant optimizer should retain this relation, i.e. no matter if the input is measured in miles or kilometers the same points are selected.

Proof. The following proof will be split into three parts. The first two parts of the proof will address a more general audience and ignore the mathematical subtleties we discussed in Section D.1.1. In the third part we explain to the interested probabilists how to resolve these issues.

1. **Invariance with regard to affine linear scaling**

Let $\tilde{\mathbf{J}}(w) := a\mathbf{J}(w) + b$ where $a > 0$ and $b \in \mathbb{R}$ and assume $\tilde{w}_0 = w_0$. With the induction start given, we only require the induction step to prove $\tilde{w}_n = w_n$.

[3] E.g. consider the random variable

$$W = \big(\operatorname{argmin} \mathbf{J}\big)\mathbf{1}_{\mathbf{J}(\tilde{w})>0} + \big(\operatorname{argmax} \mathbf{J}\big)\mathbf{1}_{\mathbf{J}(\tilde{w})<0}.$$

In this case, $\mathbf{J}(W)$ is much more informative of $\mathbf{J}(\tilde{w})$ than $\mathbf{J}(w)$ at some deterministic w.

For the induction step, we assume this equation holds up to n. Since $\phi(x) = ax + b$ is a measurable bijection, the sigma algebra[4] generated by

$$(\tilde{\mathbf{J}}(w_n), \nabla\tilde{\mathbf{J}}(w_n)) = (\phi \circ \mathbf{J}(w_n), a\nabla\mathbf{J}(w_n))$$

is therefore equal to the sigma algebra generated by $(\mathbf{J}(w_n), \nabla\mathbf{J}(w_n))$. This implies

$$\begin{aligned}
\tilde{w}_{n+1} &= \operatorname*{argmin}_w \mathbb{E}[\tilde{\mathbf{J}}(w) \mid \tilde{\mathbf{J}}(\tilde{w}_n), \nabla\tilde{\mathbf{J}}(\tilde{w}_n)] \\
&\overset{\text{induction}}{=} \operatorname*{argmin}_w \mathbb{E}[\tilde{\mathbf{J}}(w) \mid \tilde{\mathbf{J}}(w_n), \nabla\tilde{\mathbf{J}}(w_n)] \\
&\overset{\text{sigma alg.}}{=} \operatorname*{argmin}_w \mathbb{E}[\tilde{\mathbf{J}}(w) \mid \mathbf{J}(w_n), \nabla\mathbf{J}(w_n)] \\
&\overset{\text{linearity}}{=} \operatorname*{argmin}_w a\mathbb{E}[\mathbf{J}(w) \mid \mathbf{J}(w_n), \nabla\mathbf{J}(w_n)] + b \\
&\overset{\text{monotonicity}}{=} \operatorname*{argmin}_w \mathbb{E}[\mathbf{J}(w) \mid \mathbf{J}(w_n), \nabla\mathbf{J}(w_n)] \\
&\overset{\text{def.}}{=} w_{n+1}
\end{aligned} \tag{20}$$

Where we have used the linearity of the conditional expectation and the strict monotonicity of $\phi(x) = ax + b$.

2. **Invariance with regard to certain input bijections**

Let ϕ be a differentiable bijection whose jacobian is invertible everywhere and assume $\tilde{\mathbf{J}} := \mathbf{J} \circ \phi$. Since ϕ is a bijection, $\phi(M)$ is the domain of $\mathbf{J}$ whenever M is the domain of $\tilde{\mathbf{J}}$.

For a starting point $w_0 \in \phi(M)$ we now assume $\tilde{w}_0 = \phi^{-1}(w_0) \in M$ and are again going to prove the claim

$$\tilde{w}_n = \phi^{-1}(w_n).$$

by induction. Assume that we have this claim up to n. Then we have by induction

$$\tilde{\mathbf{J}}(\tilde{w}_n) = \mathbf{J} \circ \phi(\phi^{-1}(w_n)) = \mathbf{J}(w_n) \tag{21}$$

and

$$\nabla\tilde{\mathbf{J}}(\tilde{w}_n) = \nabla_{\tilde{w}_n}(\mathbf{J} \circ \phi(\tilde{w}_n)) = \phi'(\tilde{w}_n)(\nabla\mathbf{J})(\phi(\tilde{w}_n)) = \phi'(\tilde{w}_n)\nabla\mathbf{J}(w_n).$$

Since $\phi'(\tilde{w}_n)$ is invertible by assumption, the sigma algebras generated by $(\tilde{\mathbf{J}}(\tilde{w}_n), \nabla\tilde{\mathbf{J}}(\tilde{w}_n))$ and $\mathbf{J}(w_n), \nabla\mathbf{J}(w_n)$ are identical. But this results in the induction step

$$\begin{aligned}
\tilde{w}_{n+1} &= \operatorname*{argmin}_{w \in M} \mathbb{E}[\tilde{\mathbf{J}}(w) \mid \tilde{\mathbf{J}}(\tilde{w}_n), \nabla\tilde{\mathbf{J}}(\tilde{w}_n)] \\
&\overset{\text{sigma alg.}}{=} \operatorname*{argmin}_{w \in M} \mathbb{E}[\tilde{\mathbf{J}}(w) \mid \mathbf{J}(w_n), \nabla\mathbf{J}(w_n)] \\
&\overset{\text{def.}}{=} \operatorname*{argmin}_{w \in M} \mathbb{E}[\mathbf{J} \circ \phi(w) \mid \mathbf{J}(w_n), \nabla\mathbf{J}(w_n)] \\
&= \phi^{-1}\Big(\underbrace{\operatorname*{argmin}_{\theta \in \phi(M)} \mathbb{E}[\mathbf{J}(\theta) \mid \mathbf{J}(w_n), \nabla\mathbf{J}(w_n)]}_{\overset{\text{def.}}{=} w_{n+1}}\Big).
\end{aligned} \tag{22}$$

where we simply optimize over $\theta = \phi(w)$ instead of w and correct the argmin at the end.

3. **Addressing the subtleties**

In equation (20) we have really proven for deterministic w

$$\Phi_{\mathbb{P}_{\tilde{\mathbf{J}}}}(w, \tilde{\mathbf{J}}(w), \nabla\tilde{\mathbf{J}}(w)) = \Phi_{\mathbb{P}_{\mathbf{J}}}(w, \mathbf{J}(w), \nabla\mathbf{J}(w)).$$

But this implies with the induction assumption $W_n = \tilde{W}_n$

$$\tilde{W}_{n+1} = \Phi_{\mathbb{P}_{\tilde{\mathbf{J}}}}(\tilde{W}_n, \tilde{\mathbf{J}}(\tilde{W}_n), \nabla\tilde{\mathbf{J}}(\tilde{W}_n)) \overset{\text{ind.}}{=} \Phi_{\mathbb{P}_{\mathbf{J}}}(W_n, \mathbf{J}(W_n), \nabla\mathbf{J}(W_n)) = W_{n+1}.$$

[4] if you are unfamiliar with sigma algebras read them as "information".

Similarly we have proven in (22) that

$$\Phi_{\mathbb{P}_{\tilde{\mathbf{J}}}}\big(\phi^{-1}(w), \tilde{\mathbf{J}}(\phi^{-1}(w)), \nabla\tilde{\mathbf{J}}(\phi^{-1}(w))\big) = \phi^{-1}\big(\Phi_{\mathbb{P}_{\mathbf{J}}}(w, \mathbf{J}(w), \nabla\mathbf{J}(w))\big).$$

By the induction assumption $\tilde{W} = \phi^{-1}(W_n)$, this implies

$$\begin{aligned}
\tilde{W}_{n+1} &= \Phi_{\mathbb{P}_{\tilde{\mathbf{J}}}}(\tilde{W}_n, \tilde{\mathbf{J}}(\tilde{W}_n), \nabla\tilde{\mathbf{J}}(\tilde{W}_n)) \\
&\overset{\text{ind.}}{=} \Phi_{\mathbb{P}_{\tilde{\mathbf{J}}}}\big(\phi^{-1}(W_n), \tilde{\mathbf{J}}(\phi^{-1}(W_n)), \nabla\tilde{\mathbf{J}}(\phi^{-1}(W_n))\big) \\
&= \phi^{-1}\big(\Phi_{\mathbb{P}_{\mathbf{J}}}(W_n, \mathbf{J}(W_n), \nabla\mathbf{J}(W_n))\big) \\
&= \phi^{-1}(W_{n+1}).
\end{aligned}$$

□

D.2 Section 4: Relation to gradient descent

Lemma 4.1 (Explicit first order stochastic Taylor approximation)**.** *For $\mathbf{J} \sim \mathcal{N}(\mu, C)$, the first order stochastic Taylor approximation is given by*

$$\mathbb{E}[\mathbf{J}(w-\mathbf{d}) \mid \mathbf{J}(w), \nabla\mathbf{J}(w)] = \mu + \frac{C\big(\frac{\|\mathbf{d}\|^2}{2}\big)}{C(0)}(\mathbf{J}(w)-\mu) - \frac{C'\big(\frac{\|\mathbf{d}\|^2}{2}\big)}{C'(0)}\langle\mathbf{d}, \nabla\mathbf{J}(w)\rangle.$$

Proof. $(\mathbf{J}(w), \nabla\mathbf{J}(w), \mathbf{J}(w-\mathbf{d}))$ is a Gaussian vector for which the conditional distribution is well known. It is only necessary to calculate the covariance matrix. The key ingredient here is to observe that $\mathbf{J}(w), \partial_1\mathbf{J}(w), \ldots, \partial_d\mathbf{J}(w)$ are all independent, trivializing matrix inversion.

More formally, by Lemma G.2 we have

$$\operatorname{Cov}\left(\begin{pmatrix}\mathbf{J}(w)\\ \nabla\mathbf{J}(w)\end{pmatrix}\right) = \begin{pmatrix}C(0) & \\ & -C'(0)\mathbb{I}_{d\times d}\end{pmatrix}$$

and

$$\operatorname{Cov}\left(\mathbf{J}(w-\mathbf{d}), \begin{pmatrix}\mathbf{J}(w)\\ \nabla\mathbf{J}(w)\end{pmatrix}\right) = \begin{pmatrix}C\big(\frac{\|\mathbf{d}\|^2}{2}\big)\\ C'\big(\frac{\|\mathbf{d}\|^2}{2}\big)\mathbf{d}\end{pmatrix}.$$

By Theorem G.1 we therefore know that

$$\mathbb{E}[\mathbf{J}(w-\mathbf{d}) \mid \mathbf{J}(w), \nabla\mathbf{J}(w)] = \mu + \begin{pmatrix}C\big(\frac{\|\mathbf{d}\|^2}{2}\big)\\ C'\big(\frac{\|\mathbf{d}\|^2}{2}\big)\mathbf{d}\end{pmatrix}^T \begin{pmatrix}C(0) & \\ & -C'(0)\mathbb{I}_{d\times d}\end{pmatrix}^{-1} \begin{pmatrix}\mathbf{J}(w)-\mu\\ \nabla\mathbf{J}(w)\end{pmatrix},$$

which immediately yields the claim. □

Theorem 4.2 (Explicit RFD)**.** *Let $\mathbf{J} \sim \mathcal{N}(\mu, C)$, then RFD coincides with gradient descent*

$$w_{n+1} = w_n - \eta_n^* \tfrac{\nabla\mathbf{J}(w_n)}{\|\nabla\mathbf{J}(w_n)\|},$$

where the RFD step sizes are given by

$$\eta_n^* := \operatorname*{argmin}_{\eta\in\mathbb{R}} \frac{C\big(\frac{\eta^2}{2}\big)}{C(0)}(\mathbf{J}(w_n)-\mu) - \eta\frac{C'\big(\frac{\eta^2}{2}\big)}{C'(0)}\|\nabla\mathbf{J}(w_n)\|. \tag{1}$$

Proof. The explicit version of RFD follows essentially by fixing the step size $\eta = \|\mathbf{d}\|$ and optimizing over the direction first. With Lemma 4.1 we have

$$\begin{aligned}
&\min_{\mathbf{d}} \mathbb{E}[\mathbf{J}(w-\mathbf{d}) \mid \mathbf{J}(w), \nabla\mathbf{J}(w)] \\
&= \min_{\eta\geq 0}\min_{\mathbf{d}:\|\mathbf{d}\|=\eta} \mu + \frac{C\big(\frac{\eta^2}{2}\big)}{C(0)}(\mathbf{J}(w)-\mu) - \frac{C'\big(\frac{\eta^2}{2}\big)}{C'(0)}\langle\mathbf{d}, \nabla\mathbf{J}(w)\rangle \\
&= \min_{\eta\geq 0} \mu + \frac{C\big(\frac{\eta^2}{2}\big)}{C(0)}(\mathbf{J}(w)-\mu) - \frac{C'\big(\frac{\eta^2}{2}\big)}{C'(0)}\begin{cases}\max\limits_{\mathbf{d}:\|\mathbf{d}\|=\eta}\langle\mathbf{d}, \nabla\mathbf{J}(w)\rangle & \frac{C'(\frac{\eta^2}{2})}{C'(0)} \geq 0\\ \min\limits_{\mathbf{d}:\|\mathbf{d}\|=\eta}\langle\mathbf{d}, \nabla\mathbf{J}(w)\rangle & \frac{C'(\frac{\eta^2}{2})}{C'(0)} < 0.\end{cases}
\end{aligned}$$

By Lemma G.3 and Corollary G.4 the maximizing or minimizing step direction is then given by

$$\mathbf{d}(\eta) = \pm\eta\frac{\nabla\mathbf{J}(w)}{\|\nabla\mathbf{J}(w)\|}.$$

Where it is typically to be expected, that we have a positive sign. Since that depends on the covariance though, we avoid this problem with the following argument: Since η only appears as η^2 in the remaining equation, we can optimize over $\eta \in \mathbb{R}$ in the outer minimization instead of over $\eta \geq 0$ to move the sign into the step size η and set without loss of generality

$$\mathbf{d}(\eta) = \eta\frac{\nabla\mathbf{J}(w)}{\|\nabla\mathbf{J}(w)\|}.$$

Since $\langle\mathbf{d}(\eta), \nabla\mathbf{J}(w)\rangle = \eta\|\nabla\mathbf{J}(w)\|$ the remaining outer minimization problem over the step size is then given by

$$\min_{\eta\in\mathbb{R}} \frac{C\big(\frac{\eta^2}{2}\big)}{C(0)}(\mathbf{J}(w) - \mu) - \eta\frac{C'\big(\frac{\eta^2}{2}\big)}{C'(0)}\|\nabla\mathbf{J}(w)\|,$$

Its minimizer is by definition the RFD step size as given in the Theorem. □

D.3 Section 5: RFD-step sizes

Proposition D.2 (Tayloring the step size optimization problem)**.** *The second order Taylor approximation of the step size optimization problem*

$$q_\Theta(\eta) = -\frac{C\big(\frac{\eta^2}{2}\big)}{C(0)} - \eta\frac{C'\big(\frac{\eta^2}{2}\big)}{C'(0)}\Theta$$

around zero is given by

$$T_2q_\Theta(\eta) = -1 - \eta\Theta + \eta^2\frac{-C'(0)}{2C(0)} \quad \textit{minimized by} \quad \hat{\eta} := \operatorname*{argmin}_\eta T_2q_\Theta(\eta) = \tfrac{C(0)}{-C'(0)}\Theta.$$

Furthermore, the Taylor residual is bounded by

$$\big|q(\eta) - T_2q(\eta)\big| \leq \eta^3c_0\big(\tfrac{\eta}{4} + \Theta\big)$$

with $c_0 = \frac{1}{2}\max\{\sup_{\theta\in[0,1]}|C''(\theta)|, |C'(0)|\}(\frac{1}{C(0)} + \frac{1}{|C'(0)|}) < \infty$.

Proof. Using the Taylor approximation with the mean value reminder for C, we get

$$\begin{aligned} C\big(\tfrac{\eta^2}{2}\big) &= C(0) + C'(0)\tfrac{\eta^2}{2} + C''(\theta_2)\frac{\big(\frac{\eta^2}{2}\big)^2}{2!} \\ C'\big(\tfrac{\eta^2}{2}\big) &= C'(0) + C''(\theta_1)\tfrac{\eta^2}{2} \end{aligned}$$

for some $\theta_1, \theta_2 \in [0, \frac{\eta^2}{2}]$. This implies

$$q(\eta) - \underbrace{\Big(-\big(1 + \tfrac{C'(0)}{C(0)}\tfrac{\eta^2}{2}\big) - \eta\Theta\Big)}_{=:T_2q_\Theta(\eta)} = -\frac{C''(\theta_2)}{C(0)}\frac{\eta^4}{2^3} - \frac{C''(\theta_1)}{C'(0)}\frac{\eta^3}{2}\Theta$$

By the following error the optimistically defined $T_2q_\Theta(\eta)$ is really the second Taylor approximation (which can be confirmed manually, but we deduce it by arguing that its residual is in $\mathcal{O}(\eta^3)$). More specifically,

$$\big|q(\eta) - T_2q(\eta)\big| \leq \eta^3\Big(\frac{\sup_{\theta\in[0,\frac{\eta^2}{2}]}|C''(\theta)|}{2C(0)}\frac{\eta}{4} + \frac{\sup_{\theta\in[0,\frac{\eta^2}{2}]}|C''(\theta)|}{2|C'(0)|}\Theta\Big) \overset{\text{Lem. D.8}}{\leq} \eta^3c_0\big(\tfrac{\eta}{4} + \Theta\big)$$

It is easy to see for $\mathbf{J}(w) < \mu$ that $T_2q(\eta)$ is a convex parabola due to $C'(0) < 0$. We thus have

$$\hat{\eta} := \operatorname*{argmin}_\eta T_2q_\Theta(\eta) = \tfrac{C(0)}{-C'(0)}\Theta.$$ □

Theorem D.3 (Details of Proposition 5.2)**.** *Let* $\mathbf{J} \sim \mathcal{N}(\mu, C)$ *and assume there exists* $\eta_0 > 0$ *such that the correlation for larger distances* $\eta \geq \eta_0$ *are bounded smaller than* 1*, i.e.* $\frac{C(\eta^2/2)}{C(0)} < \rho \in (0,1)$*. Then there exists* $K, \Theta_0 > 0$ *such that for all* $\Theta < \Theta_0$

$$1 - K\Theta \leq \frac{\eta^*(\Theta)}{\hat{\eta}(\Theta)} \leq 1 + K\Theta.$$

In particular we have $\eta^*(\Theta) \sim \hat{\eta}(\Theta)$ *as* $\Theta \to 0$ *or equivalently as* $\hat{\eta} \to 0$*.*

Proof. This follows immediately from Lemma D.4, Lemma D.5 and Lemma D.6. □

Corollary 5.3. *Assume* $\eta^* \to 0$ *implies* $\Theta \to 0$*, the cost* $\mathbf{J}$ *is bounded, has continuous gradients and RFD converges to some point* w_∞*. Then* w_∞ *is a critical point and the RFD step sizes* η^* *are asymptotically equal to* $\hat{\eta}$*.*

Proof. Assuming RFD converges, its step sizes η^* converge to zero. But this implies $\Theta \to 0$ by assumption, i.e.

$$\Theta = \frac{\|\nabla \mathbf{J}(w)\|}{\mu - \mathbf{J}(w)} \to 0$$

Since $\mathbf{J}(w)$ is bounded, this implies $\|\nabla \mathbf{J}(w)\| \to 0$ and by continuity the of the gradient, it is zero in its limit. Thus we converge to a stationary point. The asymptotic equality follows by Lemma D.4 and Lemma D.5, as we know η^* converges so we do not require the assumptions of Lemma D.6. □

D.3.1 Locating the Minimizer

In the following we want to rule out locations for the RFD step size η^* by proving $q_\Theta(\eta) > q_\Theta(\hat{\eta})$ for a wide range of η. For this endeavour the relative position of the step size η relative to $\hat{\eta}$ is a useful re-parametrization

$$\eta := \eta(\lambda) = \lambda\hat{\eta}.$$

Due to $\hat{\eta} = \frac{C(0)}{-C'(0)}\Theta$ we obtain

$$T_2 q_\Theta(\eta) = -1 - \eta\Theta + \tfrac{\eta^2}{2}\tfrac{-C'(0)}{C(0)} = -1 + \lambda(\tfrac{\lambda}{2} - 1)\hat{\eta}\Theta$$

On the other hand we have for the bound

$$|q_\Theta(\eta) - T_2 q_\Theta(\eta)| \leq \lambda^3 \hat{\eta}^3 c_0 \Big(\lambda \tfrac{C(0)}{4|C'(0)|} + 1\Big)\Theta$$

Since $\hat{\eta} = \eta(1)$ we thus obtain

$$\begin{aligned}
\frac{q_\Theta(\eta) - q_\Theta(\hat{\eta})}{\hat{\eta}\Theta} &\geq \frac{T_2 q_\Theta(\eta) - |q_\Theta(\eta) - T_2 q_\Theta(\eta)| - T_2 q_\Theta(\hat{\eta}) - |q_\Theta(\hat{\eta}) - T_2 q_\Theta(\hat{\eta})|}{\hat{\eta}\Theta} \\
&\geq \underbrace{\big(\lambda(\tfrac{\lambda}{2} - 1) - (-\tfrac{1}{2})\big)}_{=\frac{1}{2} - \lambda + \frac{\lambda^2}{2}} - \hat{\eta}^2 c_0 \Big[\lambda^3 \Big(\lambda \tfrac{C(0)}{4|C'(0)|} + 1\Big) + \Big(\tfrac{C(0)}{4|C'(0)|} + 1\Big)\Big] \\
&= \tfrac{1}{2}(1 - \lambda)^2 - \hat{\eta}^2 c_0 \Big[\lambda^3 \Big(\lambda \tfrac{C(0)}{4|C'(0)|} + 1\Big) + \Big(\tfrac{C(0)}{4|C'(0)|} + 1\Big)\Big].
\end{aligned} \tag{23}$$

This equation will be the basis of a number of lemmas ruling out various step sizes as minimizers.

Lemma D.4 (Ruling out small step sizes)**.** *If the step size is (much) smaller than the asymptotic step size* $\hat{\eta} = \hat{\eta}(\Theta)$*, then it can not be a minimizer. More specifically*

$$\frac{\eta}{\hat{\eta}} \in [0, 1 - c_1\Theta) \implies q_\Theta(\eta) > q_\Theta(\hat{\eta})$$

where $c_1 := 2\frac{C(0)}{|C'(0)|}\sqrt{c_0\big(\frac{C(0)}{4|C'(0)|} + 1\big)} < \infty$*.*

Proof. Here we consider the case $\eta \leq \hat{\eta}$, i.e. $\lambda \in [0,1]$. By (23) we have

$$\begin{aligned}\frac{q_\Theta(\eta) - q_\Theta(\hat{\eta})}{\hat{\eta}\Theta} &\geq \tfrac{1}{2}(1-\lambda)^2 - \hat{\eta}^2 c_0\Big[\lambda^3\Big(\lambda\tfrac{C(0)}{4|C'(0)|}+1\Big) + \Big(\tfrac{C(0)}{4|C'(0)|}+1\Big)\Big] \\ &\geq \tfrac{1}{2}(1-\lambda)^2 - 2\hat{\eta}^2 c_0\Big(\tfrac{C(0)}{4|C'(0)|}+1\Big) \\ &\overset{!}{>} 0\end{aligned}$$

for which

$$(1-\lambda)^2 > 4\hat{\eta}^2 c_0\Big(\tfrac{C(0)}{4|C'(0)|}+1\Big)$$

is sufficient or equivalently

$$\lambda < 1 - 2\hat{\eta}\sqrt{c_0\Big(\tfrac{C(0)}{4|C'(0)|}+1\Big)} = 1 - \Theta\underbrace{2\tfrac{C(0)}{|C'(0)|}\sqrt{c_0\Big(\tfrac{C(0)}{4|C'(0)|}+1\Big)}}_{=:c_1}$$

So for $\lambda \in [0, 1-\Theta c_1)$ we have $q_\Theta(\eta) > q_\Theta(\hat{\eta})$. □

Lemma D.5 (Ruling out medium sized step sizes as minimizer)**.** *For $c_2 = 2c_1$ and $\Theta \leq \Theta_0 := \frac{1}{5c_1}$, we have*

$$\frac{\eta}{\hat{\eta}} \in \big(1 + c_2\Theta, \tfrac{1}{c_2\Theta}\big) \implies q_\Theta(\eta) > q_\Theta(\hat{\eta})$$

Proof. Here we consider the case $\lambda \geq 1$, i.e. $\eta > \hat{\eta}$. Again starting with (23) we get

$$\begin{aligned}\frac{q(\eta) - q(\hat{\eta})}{\hat{\eta}\Theta} &\geq \tfrac{1}{2}(1-\lambda)^2 - \hat{\eta}^2 c_0\Big[\lambda^3\Big(\lambda\tfrac{C(0)}{4|C'(0)|}+1\Big) + \Big(\tfrac{C(0)}{4|C'(0)|}+1\Big)\Big] \\ &\geq \tfrac{1}{2}(\lambda-1)^2 - 2\lambda^4\hat{\eta}^2 c_0\Big(\tfrac{C(0)}{4|C'(0)|}+1\Big) \\ &\overset{!}{>} 0,\end{aligned}$$

for which

$$\lambda - 1 > 2\lambda^2\hat{\eta}\sqrt{c_0\Big(\tfrac{C(0)}{4|C'(0)|}+1\Big)} = c_1\Theta\lambda^2$$

or equivalently

$$\lambda - 1 - c_1\Theta\lambda^2 > 0$$

is sufficient. Note that this is a concave parabola in λ. So it is positive between its zeros which are characterized by

$$c_1\Theta\lambda^2 - \lambda + 1 = 0.$$

They are thus given by

$$\lambda_{1/2} = \frac{1 \pm \sqrt{1 - 4c_1\Theta}}{2c_1\Theta}.$$

So whenever $\lambda \in (\lambda_1, \lambda_2)$ we have that $q_\Theta(\eta) > q_\Theta(\hat{\eta})$. In particular for $4c_1\Theta \leq 1$ or equivalently $\Theta \leq \frac{1}{4c_1}$ we have

$$\lambda_2 \geq \frac{1}{2c_1\Theta} = \frac{1}{c_2\Theta}$$

To get a bound on λ_1 note that the original equation was essentially

$$\lambda \geq 1 + c_1\Theta\lambda^2$$

with equality for $\lambda = \lambda_1$, if Θ is reduced, the inequality remains, which implies that λ_1 is decreasing with Θ. So assuming the inequality is satisfied for a particular λ e.g. $\lambda = \sqrt{2}$ which requires

$$\sqrt{2} \geq 1 + 2c_1\Theta \iff \Theta \leq \tfrac{\sqrt{2}-1}{2c_1},$$

then we know that $\lambda_1 \leq \sqrt{2}$ for all smaller Θ. This implies for $\Theta \leq \Theta_0 = \frac{1}{5c_1} \leq \frac{\sqrt{2}-1}{2c_1}$

$$\lambda_1 = 1 + c_1\Theta\lambda_1^2 \leq 1 + \underbrace{2c_1}_{c_2}\Theta.$$

□

Lemma D.6 (Ruling out large step sizes as minimizer)**.** *If there exists step size $\eta_0 > 0$ such that the correlation is bounded by some $\rho < 1$, i.e.*

$$\frac{C\left(\frac{\eta^2}{2}\right)}{C(0)} \leq \rho \in (0,1),$$

for larger step sizes $\eta \geq \eta_0$, then there exist $\Theta_0 > 0$ such that for all $\Theta < \Theta_0$

$$\frac{\eta}{\hat{\eta}} \in \left(1 + c_2\Theta, \infty\right) \implies q(\eta) > q(\hat{\eta}),$$

where c_2 is the constant from Lemma D.5.

Proof. The upper bound $\frac{1}{c_2\Theta}$ in Lemma D.5 is only due to the loss of precision of the Taylor approximation. To remove it, we take a closer look at the actual q_Θ itself. We have the following bound for our asymptotic minimum

$$\begin{aligned}\frac{q_\Theta(\hat{\eta})}{\Theta} &\leq \frac{T_2 q_\Theta(\hat{\eta}) + |q_\Theta(\hat{\eta}) - T_2 q_\Theta(\hat{\eta})|}{\Theta} = -\frac{1}{\Theta} - \frac{1}{2}\hat{\eta} + \hat{\eta}^3 \underbrace{c_0\Big(\tfrac{C(0)}{4|C'(0)|} + 1\Big)}_{=:c_3} \\ &\leq -\frac{1}{\Theta} + \hat{\eta}^3 c_3\end{aligned}$$

Which means we have for

$$\begin{aligned}\frac{q_\Theta(\eta) - q_\Theta(\hat{\eta})}{\Theta} &\geq \Big(1 - \frac{C\left(\frac{\eta^2}{2}\right)}{C(0)}\Big)\frac{1}{\Theta} - \eta\frac{C'\left(\frac{\eta^2}{2}\right)}{C'(0)} - \hat{\eta}^3 c_3 \\ &\overset{\text{Lemma D.7}}{\geq} \Big(1 - \frac{C\left(\frac{\eta^2}{2}\right)}{C(0)}\Big)\frac{1}{\Theta} - \frac{\sqrt{C(0)}}{\sqrt{-C'(0)}} - \hat{\eta}^3 c_3 \\ &\geq \big(1 - M\big)\frac{1}{\Theta} - \frac{\sqrt{C(0)}}{\sqrt{-C'(0)}} - \hat{\eta}^3 c_3 \\ &\overset{!}{>} 0,\end{aligned}$$

where we use the assumption that there exists $\rho \in (0,1)$ such that $\rho \geq \frac{C\left(\frac{\eta^2}{2}\right)}{C(0)}$ for all $\eta \geq \eta_0$ and the fact that we only need to consider $\eta \geq \frac{1}{c_2\Theta}$ (due to Lemma D.5) which allows a translation of η_0 into some maximal Θ_0. Note that $\hat{\eta} \sim \Theta$ vanishes as $\Theta \to 0$, so eventually the term $(1-M)\frac{1}{\Theta}$ dominates. Selecting Θ_0 small small enough is thus sufficient to cover everything that is not already covered by Lemma D.5. □

D.3.2 Technical bounds

Lemma D.7 (Bound on the first derivative of the covariance)**.**

$$\sup_{\eta \geq 0} |C'\left(\tfrac{\eta^2}{2}\right)\eta| \leq \sqrt{-C'(0)C(0)}$$

Proof. Since we have

$$\operatorname{Cov}(D_v\mathbf{J}(x), \mathbf{J}(y)) = C'\left(\tfrac{\|x-y\|^2}{2}\right)\langle x - y, v\rangle$$

we have for a standardized vector $\|v\| = 1$ and $x - y = \eta v$ by Cauchy-Schwarz

$$|C'\left(\tfrac{\eta^2}{2}\right)\eta| = |\operatorname{Cov}(D_v\mathbf{J}(x), \mathbf{J}(y))| \overset{\text{C.S.}}{\leq} \sqrt{\operatorname{Var}(D_v\mathbf{J}(x))\operatorname{Var}(\mathbf{J}(y))} = \sqrt{-C'(0)C(0)}.$$

As the bound is independent of η this yields the claim. □

Lemma D.8 (Bound on the second derivative of the covariance)**.**

$$\sup_{\theta \geq 0} |C''(\theta)| \leq \max\Big\{\sup_{\theta \in [0,1]} |C''(\theta)|, |C'(0)|\Big\}$$

Proof. Note that

$$\operatorname{Cov}(D_v\mathbf{J}(x), D_w\mathbf{J}(y)) = -C''\big(\tfrac{\|x-y\|^2}{2}\big)\langle x-y, v\rangle\langle x-y, w\rangle - C'\big(\tfrac{\|x-y\|^2}{2}\big)\langle v, w\rangle$$

Selecting v, w as orthonormal vectors (e.g. $v = e_1, w = e_2$) and $x - y := \eta(v + w)$ for some $\eta > 0$ results in $\|x - y\|^2 = 2\eta^2$ and thus by the Cauchy-Schwarz inequality

$$\big|-C''(\eta^2)\eta^2\big| = \big|\operatorname{Cov}(D_v\mathbf{J}(x), D_w\mathbf{J}(y))\big| \overset{\text{C.S.}}{\leq} \sqrt{\operatorname{Var}(D_v\mathbf{J}(x))\operatorname{Var}(D_w\mathbf{J}(y))} = \sqrt{(-C'(0))^2}$$

This implies the claim. □

D.4 Section 6: Stochastic loss

Lemma D.9. *The stochastic approximation errors*

$$\epsilon_i(w) := \ell_i(w) - \mathbf{J}(w)$$

are identically distributed, centered random functions, which are independent conditional on $\mathbf{f}$*. In particular,*

$$\mathbb{E}[\epsilon_i(w)\epsilon_j(\tilde{w})] = \mathbb{E}[\epsilon_i(w)\epsilon_j(\tilde{w}) \mid \mathbf{f}] = 0 \quad \forall j \neq i.$$

Proof. The ϵ_i are independent random functions conditional on $\mathbf{f}$, since for any $n \in \mathbb{N}$, any bounded measurable functions h and g

$$\begin{aligned}
&\mathbb{E}\Big[h\big(\epsilon_i(w_1), \dots, \epsilon_i(w_n)\big)g\big(\epsilon_j(w_1), \dots, \epsilon_j(w_n)\big) \mid \mathbf{f}\Big] \\
&= \Big[h\big(\epsilon_i(w_1), \dots, \epsilon_i(w_n)\big)\underbrace{\mathbb{E}\Big[g\big(\epsilon_j(w_1), \dots, \epsilon_j(w_n)\big) \mid \mathbf{f}, X_i, \varsigma_i\Big]}_{\overset{(*)}{=}\mathbb{E}\left[g(\epsilon_j(w_1),\dots,\epsilon_j(w_n))|\mathbf{f}\right]}\Big| \mathbf{f}\Big] \\
&= \mathbb{E}\Big[h\big(\epsilon_i(w_1), \dots, \epsilon_i(w_n)\big) \mid \mathbf{f}\Big]\mathbb{E}\Big[g\big(\epsilon_j(w_1), \dots, \epsilon_j(w_n)\big) \mid \mathbf{f}\Big],
\end{aligned}$$

where $(*)$ uses the fact that ϵ_j does not depend on the independent X_i, ς_i. Since almost by definition

$$\mathbb{E}[\epsilon_i \mid \mathbf{f}] = \mathbb{E}[\ell(\cdot, X_i, Y_i) \mid \mathbf{f}] - \mathbf{J}(\cdot) = 0,$$

the stochastic approximation errors are thus uncorrelated

$$\mathbb{E}[\epsilon_i\epsilon_j] = \mathbb{E}\Big[\mathbb{E}[\epsilon_i\epsilon_j \mid \mathbf{f}]\Big] = \mathbb{E}\Big[\mathbb{E}[\epsilon_i \mid \mathbf{f}]\mathbb{E}[\epsilon_j \mid \mathbf{f}]\Big] = 0.$$

□

Extension 6.2 (S-RFD)**.** *For loss* $\mathbf{J} \sim \mathcal{N}(\mu, C)$ *and stochastic errors* $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, C_\epsilon)$ *we have*

$$\operatorname*{argmin}_{\mathbf{d}} \mathbb{E}[\mathbf{J}(w - \mathbf{d}) \mid \mathcal{L}_b(w), \nabla\mathcal{L}_b(w)] = \eta^*(\Theta)\tfrac{\nabla\mathcal{L}_b(w)}{\|\nabla\mathcal{L}_b(w)\|}$$

with the same step size function η^* *as for RFD, but modified* Θ

$$\Theta = \frac{C'(0)}{C'(0) + \frac{1}{b}C'_\epsilon(0)}\frac{C(0) + \frac{1}{b}C_\epsilon(0)}{C(0)}\frac{\|\nabla\mathcal{L}_b(w)\|}{\mu - \mathcal{L}_b(w)}.$$

Proof. Since ϵ_i are conditionally independent between each other and to $\mathbf{J}$, as entire functions, the same holds true for $\nabla\epsilon_i$. As all the mixed covariances disappear we have

$$\begin{aligned}
\operatorname{Cov}\Big(\begin{pmatrix}\mathcal{L}_b(w)\\ \nabla\mathcal{L}_b(w)\end{pmatrix}\Big) &= \operatorname{Cov}\Big(\begin{pmatrix}\mathbf{J}(w)\\ \nabla\mathbf{J}(w)\end{pmatrix}\Big) + \frac{1}{b^2}\sum_{i=1}^{b}\operatorname{Cov}\Big(\begin{pmatrix}\epsilon_i(w)\\ \nabla\epsilon_i(w)\end{pmatrix}\Big) \\
&= \begin{pmatrix}C(0) & \\ & -C'(0)\mathbb{I}_{d\times d}\end{pmatrix} + \frac{1}{b^2}\sum_{i=1}^{b}\begin{pmatrix}C_\epsilon(0) & \\ & -C'_\epsilon(0)\mathbb{I}_{d\times d}\end{pmatrix} \\
&= \begin{pmatrix}C(0) + \frac{1}{b}C_\epsilon(0) & \\ & -\Big(C'(0) + \frac{1}{b}C'_\epsilon(0)\Big)\mathbb{I}_{d\times d}.\end{pmatrix}
\end{aligned}$$

by Lemma G.2. If you want to break up the first step we recommend considering individual entries of the covariance matrix to convince yourself that all the mixed covariances disappear. Together with the fact

$$\begin{aligned}
&\operatorname{Cov}\Big(\mathbf{J}(w-\mathbf{d}), \begin{pmatrix}\mathcal{L}_b(w)\\ \nabla\mathcal{L}_b(w)\end{pmatrix}\Big)\\
&= \operatorname{Cov}\Big(\mathbf{J}(w-\mathbf{d}), \begin{pmatrix}\mathbf{J}(w)\\ \nabla\mathbf{J}(w)\end{pmatrix}\Big) + \frac{1}{b^2}\sum_{i=1}^{b} \underbrace{\operatorname{Cov}\Big(\mathbf{J}(w-\mathbf{d}), \begin{pmatrix}\epsilon_i(w)\\ \nabla\epsilon_i(w)\end{pmatrix}\Big)}_{=0}\\
&= \begin{pmatrix} C(\frac{\|\mathbf{d}\|^2}{2}) \\ C'(\frac{\|\mathbf{d}\|^2}{2})\mathbf{d}\end{pmatrix}.
\end{aligned}$$

The rest is analogous to Lemma 4.1 and Theorem 4.2, so we only sketch the remaining steps.

Applying Theorem G.1 as in Lemma 4.1 we obtain a stochastic version of the stochastic Taylor approximation ("stochastic^2 Taylor approximation" perhaps?)

$$\mathbb{E}[\mathbf{J}(w-\mathbf{d}) \mid \mathcal{L}_b(w), \nabla\mathcal{L}_b(w)] = \mu + \frac{C(\frac{\|\mathbf{d}\|^2}{2})}{C(0)+\frac{1}{b}C_\epsilon(0)}(\mathcal{L}_b(w)-\mu) - \frac{C'(\frac{\|\mathbf{d}\|^2}{2})}{C'(0)+\frac{1}{b}C'_\epsilon(0)}\langle \mathbf{d}, \mathcal{L}_b(w)\rangle.$$

Minimizing this subject to a constant step size as in Theorem 4.2 results in

$$\begin{aligned}
\eta^* &= \operatorname*{argmin}_{\eta\in\mathbb{R}} \frac{C(\frac{\|\mathbf{d}\|^2}{2})}{C(0)+\frac{1}{b}C_\epsilon(0)}(\mathcal{L}_b(w)-\mu) - \eta\frac{C'(\frac{\|\mathbf{d}\|^2}{2})}{C'(0)+\frac{1}{b}C'_\epsilon(0)}\|\mathcal{L}_b(w)\|\\
&= \operatorname*{argmin}_{\eta\in\mathbb{R}} -\frac{C(\frac{\|\mathbf{d}\|^2}{2})}{C(0)} - \eta\frac{C'(\frac{\|\mathbf{d}\|^2}{2})}{C'(0)+\frac{1}{b}C'_\epsilon(0)}\frac{C(0)}{C(0)+\frac{1}{b}C_\epsilon(0)}\frac{\|\mathcal{L}_b(w)\|}{\mu-\mathcal{L}_b(w)},
\end{aligned}$$

where we divided the term by $\frac{C(0)}{C(0)+\frac{1}{b}C_\epsilon(0)}\frac{1}{\mu-\mathcal{L}_b(w)} \geq 0$ to obtain the last equation. The claim follows by definition of $\eta^*(\Theta)$ and our redefinition of Θ. □

E Extensions

In this section we present a few possible extensions to Theorem 4.2, which are all composable, i.e. it is possible to combine these extensions without any major problems (including S-RFD, i.e. Extension 6.2).

E.1 Geometric anisotropy/Adaptive step sizes

In this section, we discuss the generalization of isotropy to "geometric anisotropies" [50, p. 17], which provide good insights into the inner workings of adaptive learning rates (e.g. AdaGrad [14] and Adam [29]).

Definition E.1 (Geometric Anisotropy)**.** We say a random function $\mathbf{J}$ exhibits a "geometric anisotropy", if there exists an invertible matrix A such that $\mathbf{J}(x) = \mathbf{g}(Ax)$ for some isotropic random function $\mathbf{g}$.

This implies that the expectation of $\mathbf{J}$ is still constant ($\mathbb{E}[\mathbf{J}(x)] = \mathbb{E}[\mathbf{g}(Ax)] = \mu$) and the covariance function of $\mathbf{J}$ is given by

$$\operatorname{Cov}(\mathbf{J}(x), \mathbf{J}(y)) = \operatorname{Cov}(\mathbf{g}(Ax), \mathbf{g}(Ay)) = C\Big(\frac{\|A(x-y)\|^2}{2}\Big) = C\Big(\frac{\|x-y\|^2_{A^TA}}{2}\Big) \tag{24}$$

where $\|\cdot\|_\Sigma$ is the norm induced by $\langle x, y\rangle_\Sigma := \langle x, \Sigma y\rangle$ for some strictly positive definite matrix $\Sigma = A^TA$. Here (24) characterizes the set of random functions with a geometric anisotropy in the Gaussian case, because for an $\mathbf{J}$ with such a covariance we can always obtain an isotropic $\mathbf{g}$ by $\mathbf{g}(x) := \mathbf{J}(A^{-1}x)$. This is the whitening transformation we suggest looking for in order to ensure isotropy in the context of scale invariance (Section 2).

An important observation is, that Theorem F.2 implies that $\mathbf{J}$ is still stationary, so the distribution of $\mathbf{J}$ is still invariant to translations. If stationarity is a problem, this is therefore not the solution. But geometric anisotropies are a beautiful model to explain preconditioning and adaptive step sizes. For this, we first determine the RFD steps.

Extension E.2 (RFD steps under geometric anisotropy)**.** *Let* $\mathbf{J}$ *be a Gaussian random function which exhibits a "geometric anisotropy"* A *and is based on an isotropic random function* $\mathbf{g} \sim \mathcal{N}(\mu, C)$*. Then the RFD steps are given by*

$$\eta^* \frac{\Sigma^{-1}\nabla\mathbf{J}(w)}{\|\Sigma^{-1}\nabla\mathbf{J}(w)\|_\Sigma} = \operatorname*{argmin}_{\mathbf{d}} \mathbb{E}[\mathbf{J}(w-\mathbf{d}) \mid \mathbf{J}(w), \nabla\mathbf{J}(w)]$$

with

$$\eta^* = \operatorname*{argmin}_{\eta} q_\Theta(\eta) \quad \textit{where} \quad \Theta = \frac{\|\Sigma^{-1}\nabla\mathbf{J}(w)\|_\Sigma}{\mu - \mathbf{J}(w)}.$$

Proofsketch. There are two ways to see this. Either we apply scale invariance (Advantage 2.3) directly to translate the steps on $\mathbf{g}$ into steps on $\mathbf{J}$. Alternatively one can manually retrace the steps of the proof. Details in Subsection E.1.1 □

The step direction is therefore

$$\Sigma^{-1}\nabla\mathbf{J}(x)$$

and Σ^{-1} acts as a preconditioner. So how would one obtain Σ? As it turns out the following holds true (by Lemma G.2)

$$\mathbb{E}[\nabla\mathbf{J}(w)\nabla\mathbf{J}(w)^T] = A^T\mathbb{E}[\nabla\mathbf{g}(w)\nabla\mathbf{g}(w)^T]A = A^T(-C'(0)\mathbb{I})A = -C'(0)\Sigma$$

In their proposal of the first "adaptive" method, AdaGrad, Duchi et al. [14] suggest to use the matrix

$$G_t = \sum_{k=1}^{t} \nabla\mathbf{J}(w_k)\nabla\mathbf{J}(w_k)^T,$$

which is basically already looking like an estimation method of Σ. They then restrict themselves to $\operatorname{diag}(G_t)$ due to the computational costs of a full matrix inversion. This results in entry-wise ("adaptive") learning rates. Later adaptive methods like RMSProp [22], AdaDelta [57] and in particular Adam [29] replace this sum with an exponential mean estimate, i.e. in the case of Adam the decay rate β_2 is used to get an exponential moving average

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)\operatorname{diag}(\nabla\mathbf{J}(w_t)\nabla\mathbf{J}(w_t)^T) = \beta_2 v_{t-1} + (1-\beta_2)(\nabla\mathbf{J}(w_t))^2.$$

They then take the expectation

$$\mathbb{E}[v_t] = \mathbb{E}\Big[(1-\beta_2)\sum_{k=1}^{t}\beta_2^{t-k}\nabla\mathbf{J}(w_k)^2\Big] = \mathbb{E}\Big[(1-\beta_2)\sum_{k=1}^{t}\beta_2^{t-k}\nabla\mathbf{J}(w_k)^2\Big] = (1-\beta_2^t)\underbrace{\mathbb{E}[\nabla\mathbf{J}(x_t)^2]}_{\propto\operatorname{diag}(\Sigma)}$$

So $\hat{v}_t = v_t/(1-\beta_2^t)$ in the Adam optimizer is essentially an estimator for $\operatorname{diag}(\Sigma)$. It is noteworthy, that Kingma and Ba [29] already used the expectation symbol. This is despite the fact, that they did not yet model the optimization objective $\mathbf{J}$ as a random function.

We can not yet explain why they then use the square root of their estimate $\operatorname{diag}(\Sigma)^{-1/2}$ instead of $\operatorname{diag}(\Sigma)^{-1}$ itself. This might have something to do with the fact that the estimation of G_t happens online and the $\mathbf{J}(w_k)$ are therefore highly correlated. Another reason might be that the inverse of an estimator has different properties than the estimator itself. Finally, the fact that only the diagonal is used might also be the reason, if the preconditioner $\operatorname{diag}(\Sigma)^{-1/2}$ is simply better when we restrict ourselves to diagonal matrices.

E.1.1 Proof of Extension E.2

Since the application of scale invariance provides no intuition, we provide a proof which retraces some of the steps of the original proof.

Recall, that for an isotropic random function $\mathbf{g}$ we have the stochastic Taylor approximation

$$\mathbb{E}[\mathbf{g}(w-\mathbf{d}) \mid \mathbf{g}(x), \nabla\mathbf{g}(x)] = \mu + \frac{C\left(\frac{\|\mathbf{d}\|^2}{2}\right)}{C(0)}(\mathbf{g}(w)-\mu) + \frac{C'\left(\frac{\|\mathbf{d}\|^2}{2}\right)}{C'(0)}\langle \mathbf{d}, \nabla\mathbf{g}(w)\rangle$$

This implies for a random function with geometric anisotropy $\mathbf{J}(w) = \mathbf{g}(Aw)$ that

$$\begin{aligned}
\mathbb{E}[\mathbf{J}(w-\mathbf{d}) \mid \mathbf{J}(w), \nabla\mathbf{J}(w)] &= \mathbb{E}[\mathbf{g}(A(w-\mathbf{d})) \mid \mathbf{g}(Aw), \nabla\mathbf{g}(Aw)] \\
&= \mu + \frac{C\left(\frac{\|A\mathbf{d}\|^2}{2}\right)}{C(0)}(\mathbf{g}(Aw)-\mu) - \frac{C'\left(\frac{\|A\mathbf{d}\|^2}{2}\right)}{C'(0)}\langle A\mathbf{d}, \nabla\mathbf{g}(Aw)\rangle \\
&= \mu + \frac{C\left(\frac{\|\mathbf{d}\|_\Sigma^2}{2}\right)}{C(0)}(\mathbf{J}(w)-\mu) - \frac{C'\left(\frac{\|\mathbf{d}\|_\Sigma^2}{2}\right)}{C'(0)}\langle \mathbf{d}, \underbrace{A^T\nabla\mathbf{g}(Aw)}_{=\nabla\mathbf{J}(w)}\rangle
\end{aligned}$$

with $\Sigma := A^TA$. As in the original proof, we now optimize over the direction first, while keeping the step size constant, although we now fix the step size with regard to the norm $\|\cdot\|_\Sigma$ (which basically means that we still do the optimization in the isotropic space). Note that

$$\max_{\mathbf{d}} \langle \mathbf{d}, \nabla\mathbf{J}(x)\rangle \quad \text{s.t.} \quad \|\mathbf{d}\|_\Sigma = \eta$$

is equivalent to

$$\max_{\mathbf{d}} \langle \mathbf{d}, \Sigma^{-1}\nabla\mathbf{J}(x)\rangle_\Sigma \quad \text{s.t.} \quad \|\mathbf{d}\|_\Sigma = \eta$$

which is solved by

$$\pm\eta\frac{\Sigma^{-1}\nabla\mathbf{J}(x)}{\|\Sigma^{-1}\nabla\mathbf{J}(x)\|_\Sigma}$$

The remainder of the proof is exactly the same as in the original.

E.2 Conservative RFD

In the first paragraph of Section 2 we motivated the relation between RFD and classical optimization with the observation, that gradient descent is the minimizer of a regularized first order Taylor approximation

$$\tfrac{1}{L}\nabla\mathbf{J}(w) = \operatorname*{argmin}_{\mathbf{d}} T[\mathbf{J}(w-\mathbf{d}) \mid \mathbf{J}(w), \nabla\mathbf{J}(w)] + \tfrac{L}{2}\|w\|^2.$$

This regularized Taylor approximation is in fact an upper bound on our function under the L-smoothness assumption [38], i.e.

$$\mathbf{J}(w-\mathbf{d}) \leq T[\mathbf{J}(w-\mathbf{d}) \mid \mathbf{J}(w), \nabla\mathbf{J}(w)] + \tfrac{L}{2}\|\mathbf{d}\|^2$$

An improvement on of this upper bound compared to $\mathbf{J}(x)$ therefore guarantees an improvement of the loss. This guarantee was lost with the conditional expectation (on purpose, as we wanted to consider the average case). Losing this guarantee also makes convergence proofs more difficult as they typically make use of this improvement. In view of the confidence intervals of Figure 1, it is natural to ask for a similar upper bound in the random setting, where this can only be the top of an confidence interval. This is provided in the following theorem

Lemma E.3 (An γ-upper bound)**.** *We have*

$$\mathbb{P}\Big(\mathbf{J}(w-\mathbf{d}) \leq \mathbb{E}[\mathbf{J}(w-\mathbf{d}) \mid \mathbf{J}(w), \nabla\mathbf{J}(w)] + \rho_\gamma(\|\mathbf{d}\|^2)\Big) \geq \gamma$$

for $\rho_\gamma(\eta^2) := \Phi^{-1}(\gamma)\sigma(\eta^2)$ *with*

$$\sigma^2(\eta^2) := C(0) - \frac{C\left(\frac{\eta^2}{2}\right)^2}{C(0)} - \frac{C'\left(\frac{\eta^2}{2}\right)^2}{-C'(0)}\eta^2$$

where Φ *is the cumulative distribution function (cdf) of the standard normal distribution.*

Proof. Note that the conditional variance is with the usual argument about the covariance matrices (cf. the proof of Thoerem 4.2) using Lemma G.2 and an application of Theorem G.1 given by

$$\sigma^2(\|w\|^2) := \operatorname{Var}[\mathbf{J}(w-\mathbf{d}) \mid \mathbf{J}(w), \nabla\mathbf{J}(w)] = C(0) - \frac{C\left(\frac{\|\mathbf{d}\|^2}{2}\right)^2}{C(0)} - \frac{C'\left(\frac{\|\mathbf{d}\|^2}{2}\right)^2}{-C'(0)}\|\mathbf{d}\|^2.$$

Since the conditional distribution is normal (by Theorem G.1), we have

$$\frac{\mathbf{J}(w-\mathbf{d}) - \mathbb{E}[\mathbf{J}(w-\mathbf{d}) \mid \mathbf{J}(w), \nabla\mathbf{J}(w)]}{\sigma(\|w\|^2)} \sim \mathcal{N}(0,1).$$

But this implies the claim

$$\begin{aligned}
&\mathbb{P}\Big(\mathbf{J}(w-\mathbf{d}) \le \mathbb{E}[\mathbf{J}(w-\mathbf{d}) \mid \mathbf{J}(w), \nabla\mathbf{J}(w)] + \rho_\gamma(\|\mathbf{d}\|^2)\Big) \\
&= \mathbb{P}\Big(\frac{\mathbf{J}(w-\mathbf{d}) - \mathbb{E}[\mathbf{J}(w-\mathbf{d}) \mid \mathbf{J}(w), \nabla\mathbf{J}(w)]}{\sigma(\|w\|^2)} \le \Phi^{-1}(\gamma)\Big) \\
&= \Phi(\Phi^{-1}(\gamma)) = \gamma.
\end{aligned}$$

To avoid the Gaussian assumption, one could apply the Markov inequality instead, or another applicable concentration inequality. □

Using this upper bound, we obtain a natural conservative extension of RFD

Extension E.4 (γ-conservative RFD)**.** *Let $\mathbf{J} \sim \mathcal{N}(\mu, C)$ and $\rho_\gamma(\eta^2) = \Phi^{-1}(\gamma)\sigma(\eta^2)$, where σ is the conditional standard deviation as defined in Lemma E.3. Then the conservative RFD step direction is given by*

$$\eta^* \frac{\nabla\mathbf{J}(w)}{\|\nabla\mathbf{J}(w)\|} = \operatorname*{argmin}_{\mathbf{d}} \mathbb{E}[\mathbf{J}(w-\mathbf{d}) \mid \mathbf{J}(w), \nabla\mathbf{J}(w)] + \rho_\gamma(\|\mathbf{d}\|^2)$$

and the γ-conservative RFD step size is given by

$$\eta^* = \operatorname*{argmin}_{\eta} \frac{C\left(\frac{\eta^2}{2}\right)}{C(0)}(\mathbf{J}(w)-\mu) - \eta\frac{C'\left(\frac{\eta^2}{2}\right)}{C'(0)}\|\nabla\mathbf{J}(w)\| + \rho_\gamma(\eta^2).$$

Proof. The proof is the same as in Theorem 4.2 with Lemma 4.1 replaced by Lemma E.3. □

Taking multiple steps should generally have an averaging effect, so we expect faster convergence for almost risk neutral minimization of the conditional expectation (i.e. $\gamma \approx \frac{1}{2}$). Here γ is a natural parameter to vary conservatism. In a software implementation it might be a good idea to call this parameter 'conservatism' and rescale it to be in $[0,1]$ instead of $[\frac{1}{2},1]$. But formulas look cleaner with γ.

In Bayesian optimization it is much more common to reverse this approach and minimize a lower confidence bound ('conservatism' < 0 or $\gamma < \frac{1}{2}$) in order to encourage exploration. But since RFD is forgetful, this is not a good idea for RFD.

Remark E.5 (Conservative RFD coincides asymptotically with RFD in high dimension). While conservative RFD might seem like a good approach to fix the instability of RFD under the isotropy assumption on some optimization problems, the variance generally vanishes in high dimension [see 5] and conservative RFD coincides asymptotically with RFD. We therefore believe that the underlying issue is not an overly risk-affine algorithm but rather that distributional assumptions, in particular the stationarity assumption, are violated when instabilities occur (cf. Section F).

Nevertheless, conservative RFD might be a good approach for lower dimensional, risk-sensitive applications.

E.3 Beyond the Gaussian assumption

In this section we sketch how the extension beyond the Gaussian case using the "best linear unbiased estimator" BLUE [e.g. 27, ch. 7] works.

For this we recapitulate what a BLUE is. A **linear estimator** $\hat{Y}$ of Y using $X_1, \dots, X_n$ is of the form

$$\hat{Y} \in \operatorname{span}\{X_1, \dots, X_n\} + \mathbb{R}.$$

The set of **unbiased linear estimators** is defined as

$$\begin{aligned} \mathrm{LUE} = \mathrm{LUE}\left[Y \mid X_1, \dots, X_n\right] &= \{\hat{Y} \in \operatorname{span}\{X_1, \dots, X_n\} + \mathbb{R} : \mathbb{E}[\hat{Y}] = \mathbb{E}[Y]\} \\ &= \{\hat{Y} + \mathbb{E}[Y] : \hat{Y} \in \operatorname{span}\{X_1 - \mathbb{E}[X_1], \dots, X_n - \mathbb{E}[X_n]\}\}. \end{aligned} \tag{25}$$

And the BLUE is the **best linear unbiased estimator**, i.e.

$$\mathrm{BLUE}[Y \mid X_1, \dots, X_n] := \operatorname*{argmin}_{\hat{Y} \in \mathrm{LUE}} \mathbb{E}[\|\hat{Y} - Y\|^2]. \tag{26}$$

Other risk functions to minimize are possible, but this is the usual one.

Lemma E.6. *If $X, Y_1, \dots, Y_n$ are multivariate normal distributed, then we have*

$$\begin{aligned} \mathrm{BLUE}[Y \mid X_1, \dots, X_n] &= \mathbb{E}[Y \mid X_1, \dots, X_n] \\ &\left(= \operatorname*{argmin}_{\hat{Y} \in \{f(X_1, \dots, X_n) : f \text{ meas.}\}} \mathbb{E}[\|Y - \hat{Y}\|^2] \right). \end{aligned}$$

Proof. We observe that the conditional expectation of Gaussian random variables is linear (Theorem G.1). So as a linear function its L^2 risk must be larger or equal to that of the BLUE. And as an L^2 projection [30, Cor. 8.17] the conditional expectation was already optimal. □

If we now replace the conditional expectation with the BLUE, then all our theory remains the same because the result in Theorem G.1 remains the BLUE for general distributions [27]. Instead of minimizing

$$\min_{\mathbf{d}} \mathbb{E}[\mathbf{J}(w - \mathbf{d}) \mid \mathbf{J}(w), \nabla \mathbf{J}(w)]$$

we can therefore always minimize

$$\min_{\mathbf{d}} \mathrm{BLUE}[\mathbf{J}(w - \mathbf{d}) \mid \mathbf{J}(w), \nabla \mathbf{J}(w)]$$

without the Gaussian assumption and all our results can be translated to this case. The reader only needs to replace all mentions of Theorem G.1 with the BLUE equivalent and replace all "idependence" claims with "uncorrelated".

F Input invariance

In this section we generalize the notion of isotropy to non-stationary isotropy and discuss why we believe this generalization is necessary. Recall that we motivated isotropy as an invariant distribution with regard to isometric transformations of the input. In particular its distribution stays invariant with regard to translations (also known as stationarity), which we do not believe plausible for cost functions, because the cost at zero $\mathbf{J}(0)$ behaves fundamentally different from the cost of any other parameter vector.

In the following we will therefore generalize this notion to general input invariant distributions. And we will discuss their applicability to machine learning after we characterize the named categories.

Definition F.1 (Input Invariance)**.** A random function $\mathbf{f}$ is Φ-input invariant, if [5]

$$\mathbb{P}_{\mathbf{f}} = \mathbb{P}_{\mathbf{f} \circ \phi} \qquad \forall \phi \in \Phi.$$

For certain sets of Φ we give these Φ-input invariant distributions names

[5] The input to a random function is somewhat ambiguous since it is a random variable, i.e. function from the probability space Ω into function space, so its first input should be $\omega \in \Omega$. Formally, the definition should therefore be: For all measurable sets of functions A

$$\mathbb{P}_{\mathbf{f}}(\phi_*^{-1}(A)) = \mathbb{P}_{\mathbf{f}}(A) \quad \forall \phi \in \Phi$$

where $\phi_* : f \mapsto f \circ \phi$ denotes the pullback. But this is less helpful for an intuitive understanding.

- If Φ is the set of *isometries*, we call $\mathbf{f}$ (stationary) **isotropic**.
- If Φ is the set of *translations*, we call $\mathbf{f}$ **stationary**.
- If Φ is the set of *linear isometries*, we call $\mathbf{f}$ **non-stationary isotropic**.

We further say a random function $\mathbf{f}$ is n-weakly Φ-input invariant, if for all $\phi \in \Phi$, all $k \leq n$ and all x_i

$$\mathbb{E}[\mathbf{f}(\phi(x_1)) \cdot \cdots \cdot \mathbf{f}(\phi(x_k))] = \mathbb{E}[\mathbf{f}(x_1) \cdot \cdots \cdot \mathbf{f}(x_k)].$$

Since second moments fully determine Gaussian distributions, 2-weakly input invariance is special, because it is equivalent to full input invariance in the Gaussian case. So an omitted n equals 2. "Weakly isometry invariant" naturally becomes "weakly isotropic", etc.

While stationary and stationary isotropic random functions are well known [e.g. 44, 1], we are not aware of research on non-stationary isotropy although we doubt the concept is new. It turns out that the different notions of input isometry have simple characterizations in terms of the covariance functions. We present these in Theorem F.2 of which the stationary isotropic and stationary case are already well known.

Theorem F.2 (Characterization of Weak Input Invariances)**.** *Let $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}$ be a random function, then $\mathbf{f}$ is*

1. *weakly stationary, if and only if there exists $\mu \in \mathbb{R}$ and function $C : \mathbb{R}^d \to \mathbb{R}$ such that for all x, y*
$$\mu_{\mathbf{f}}(x) = \mu, \qquad \mathcal{C}_{\mathbf{f}}(x, y) = C(x - y).$$
2. *weakly non-stationary isotropic, if and only if there exist functions $\mu : \mathbb{R}_{\geq 0} \to \mathbb{R}$ and $\kappa : D \to \mathbb{R}$ with $D = \{\lambda \in \mathbb{R}^2_{\geq 0} \times \mathbb{R} : |\lambda_3| \leq 2\sqrt{\lambda_1 \lambda_2}\} \subseteq \mathbb{R}^3$. such that for all x, y*
$$\mu_{\mathbf{f}}(x) = \mu\big(\tfrac{\|x\|^2}{2}\big)$$
$$\mathcal{C}_{\mathbf{f}}(x, y) = \kappa\big(\tfrac{\|x\|^2}{2}, \tfrac{\|y\|^2}{2}, \langle x, y\rangle\big)$$
3. *weakly stationary isotropic, if and only if there exists $\mu \in \mathbb{R}$ and a function $C : \mathbb{R}_{\geq 0} \to \mathbb{R}$ such that for all x, y*
$$\mu_{\mathbf{f}}(x) = \mu, \qquad \mathcal{C}_{\mathbf{f}}(x, y) = C\big(\tfrac{\|x-y\|^2}{2}\big)$$

Proof. The proof essentially follow as a corollary from a characterization of isometries (Proposition F.4). For details see Subsec F.2. □

Non-stationary isotropy is therefore a generalization of stationary isotropy. It allows the zero parameter vector to have special meaning because the distribution is only invariant to linear isometries (i.e. rotations and reflections) which keep the zero in place.

It is important to highlight, that a geometric anisotropy (Section E.1) retains stationarity, while breaking non-stationary isotropy. A similar geometric generalization could also be applied to non-stationary isotropy.

Another important observation is the fact, that non-stationary isotropy coincides with stationary isotropy on the sphere. I.e. when $\|x\|$ and $\|y\|$ are constant, the function

$$\kappa\big(\tfrac{\|x\|^2}{2}, \tfrac{\|y\|^2}{2}, \tfrac{\|x-y\|^2}{2}\big)$$

only depends on $\|x - y\|$ and the mean is also constant. In other words, we have stationary isotropy on the sphere.

Isotropy might therefore 'get by' as an assumption in machine learning, as parameters are typically initialized on the sphere. This is because Glorot intialization [18] samples parameter entries independently, so their squared norm

$$\|w\|^2 = \sum_{i=1}^{d} (w^{(i)})^2$$

is a sum of independent random variables which are normalized such that a law of large numbers applies. Up to small variance their lengths are therefore all the same, and are placed on a sphere at this radius.

If we leave this sphere, this equivalence stops being true. Weight normalization [46], batch normalization [25], weight decay [e.g. 19] or equivalently L^2 regularization, etc. might all contribute to keep this assumption intact.

But in the following section we will see, that even simple linear regressions considered by researches investigating the average case behavior on quadratic functions [e.g. 58, 43, 33, 12, 9, 40, 41], require non-stationary isotropy. Moreover the covariance kernels suggested by investigations into random neuronal networks [e.g. 54, 8] are also non-stationary isotropic but not stationary isotropic.

F.1 Random linear regression

In this section, we determine the distribution of the cost function induced by a simple linear regression. For this we define the mean squared sample loss

$$\ell_i(w) = (Y - f_w(X))^2,$$

where the random data X is mapped by the true relationship $\mathbf{f}$ to labels $Y = \mathbf{f}(X)$ and

$$f_w(x) = \langle x, w\rangle$$

is a linear model. If the true relationship $\mathbf{f}$ is also a random linear function $\mathbf{f}(x) = \langle \theta, x\rangle$ with random signal $\theta \sim \mathcal{N}(0, \mathbb{I})$ independent of input $X \sim \mathcal{N}(0, \mathbb{I})$, then the cost function is given by

$$\begin{aligned}\mathbf{J}(w) &= \mathbb{E}[\ell_i(w) \mid \mathbf{f}] = \mathbb{E}[\langle \theta - w, X\rangle^2 \mid \theta] \\ &= (\theta - w)^T \mathbb{E}[XX^T](\theta - w) \\ &= \|\theta - w\|^2\end{aligned}$$

Lemma F.3. *The expectation and covariance of* $\mathbf{J}$ *are given by*

$$\begin{aligned}\mathbb{E}[\mathbf{J}(w)] &= \textit{const.} + \|w\|^2 \\ \operatorname{Cov}(\mathbf{J}(w), \mathbf{J}(\tilde{w})) &= \textit{const.} + 4\langle w, \tilde{w}\rangle\end{aligned}$$

In particular, the cost $\mathbf{J}$ *is non-stationary isotropic, but not stationary isotropic.*

Proof. Its expectation is given by

$$\begin{aligned}\mathbb{E}[\mathbf{J}(w)] = \mathbb{E}[\|\theta - w\|^2] &= \mathbb{E}[\|\theta\|^2] - 2\langle \underbrace{\mathbb{E}[\theta]}_{=0}, w\rangle + \|w\|^2 \\ &= \text{const.} + \|w\|^2\end{aligned}$$

In particular it is not constant, but dependent on $\|w\|^2$, which means that we do not have stationary isotropy. But there is still hope for non-stationary isotropy, and this is essentially true as can be seen by calculating

$$\begin{aligned}\operatorname{Cov}(\mathbf{J}(w), \mathbf{J}(\tilde{w})) &= \mathbb{E}\Big[(\mathbf{J}(w) - \mathbb{E}[\mathbf{J}(w)])(\mathbf{J}(\tilde{w}) - \mathbb{E}[\mathbf{J}(\tilde{w})])\Big] \\ &= \mathbb{E}\Big[(\|\theta\|^2 - \mathbb{E}\|\theta\|^2 - 2\langle\theta, w\rangle)(\|\theta\|^2 - \mathbb{E}\|\theta\|^2 - 2\langle\theta, \tilde{w}\rangle)\Big] \\ &= \operatorname{Var}(\|\theta\|^2 - \mathbb{E}\|\theta\|^2) \\ &\quad - 2\mathbb{E}[(\|\theta\|^2 - \mathbb{E}\|\theta\|^2)\langle\theta, w\rangle] \\ &\quad - 2\mathbb{E}[(\|\theta\|^2 - \mathbb{E}\|\theta\|^2)\langle\theta, \tilde{w}\rangle] \\ &\quad + 4w^T\mathbb{E}[\theta\theta^T]\tilde{w} \\ &= \operatorname{Var}(\|\theta\|^2 - \mathbb{E}\|\theta\|^2) + 4\langle w, \tilde{w}\rangle \\ &= \text{const.} + 4\langle w, \tilde{w}\rangle\end{aligned}$$

because the terms in the middle are zero, e.g.

$$\mathbb{E}[(\|\theta\|^2 - \mathbb{E}\|\theta\|^2)\langle\theta, w\rangle] = \langle \underbrace{\mathbb{E}[\|\theta\|^2\theta]}_{=0}, w\rangle - \mathbb{E}\|\theta\|^2\langle \underbrace{\mathbb{E}[\theta]}_{=0}, w\rangle$$

where the entries of $\mathbb{E}[\|\theta\|^2\theta]$ are zero, because of independence and first moments being zero and third moments being zero. □

F.2 Proof of Theorem F.2

Proposition F.4 (Characterizing isometries)**.** *Let $\mathcal{X}$ be a vectorspace and $x_i, y_i \in \mathcal{X}$ for $i = 1, \dots, n$, then the following pairs of statements are equivalent*

1. *(a) $x_i - x_j = y_i - y_j$ for all i, j*
   *(b) there exists a **translation** ϕ with $\phi(x_i) = y_i$ for all i.*

In the remainder we further assume $\mathcal{X}$ to be a Hilbertspace,

2. *(a) $\|x_i\| = \|y_i\|$ and $\|x_i - x_j\| = \|y_i - y_j\|$ for all i, j*
   *(b) there exists a **linear isometry** ϕ with $\phi(x_i) = y_i$ for all i.*
3. *(a) $\|x_i - x_j\| = \|y_i - y_j\|$ for all i, j*
   *(b) there exists an **(affine) isometry** ϕ with $\phi(x_i) = y_i$ for all i.*

Proof. (1a) $\Rightarrow$ (1b)**:** we define

$$\phi(x) := x + (y_0 - x_0),$$

which implies

$$\phi(x_i) = x_i - x_0 + y_0 = (y_i - y_0) + y_0 = y_i.$$

(1b) $\Rightarrow$ (1a)**:** Let $\phi(x) = x + c$ for some c. Then we immediately have

$$y_i - y_j = \phi(x_i) - \phi(x_j) = x_i + c - (x_j + c) = x_i - x_j.$$

(2a) $\Rightarrow$ (2b)**:** By the polarization formula, for all i, j

$$\langle x_i, x_j \rangle = \frac{\|x_i\|^2 + \|y_i\|^2 - \|x_i - x_j\|^2}{2} = \langle y_i, y_j \rangle.$$

We apply the Gram-Schmidt orthonormalization procedure to both x_i and y_i such that

$$U_{k_n} = \operatorname{span}(u_1, \dots, u_{k_n}) = \operatorname{span}(x_1, \dots, x_n)$$

for orthonormal u_i where we skip x_m if it is already in $U_{k_{m-1}}$ (resulting in $k_m = k_{m-1}$), and similarly

$$V_{k_n} = \operatorname{span}(v_1, \dots, v_{k_n}) = \operatorname{span}(y_1, \dots, y_n).$$

Since this procedure only uses scalar products, we inductively get

$$\langle x_k, u_j \rangle = \langle y_k, v_j \rangle \quad \forall k, j$$

We now extend u_i and v_i to orthonormal basis of $\mathcal{X}$ and define the linear mapping by its behavior on the basis elements $\phi : u_i \mapsto v_i$. Mapping an orthonormal basis to an orthonormal basis is an isometry and we have

$$\begin{aligned}
\phi(x_k) &= \phi\Big(\sum_{j=1}^{k} \langle x_k, u_j \rangle u_j\Big) \\
&= \sum_{j=1}^{k} \langle x_k, u_j \rangle \phi(u_j) = \sum_{j=1}^{k} \langle y_k, v_j \rangle v_j \\
&= y_k.
\end{aligned}$$

(2b) $\Rightarrow$ (2a)**:** Isometries preserve distances by definition. This implies $\|x_i - x_j\| = \|y_i - y_j\|$. And linear functions map 0 to 0, so we have

$$\|x_i\| = \|x_i - 0\| = \|\phi(x_i) - \phi(0)\| = \|y_i\|.$$

(3a) $\Rightarrow$ (3b)**:** We define

$$\tilde{x}_i = x_i - x_0$$

and similarly for y. In particular, $\tilde{x}_0 = \tilde{y}_0 = 0$. Since $\tilde{x}_i$ and $\tilde{y}_i$ satisfy the requirements of 2, there exists a linear isometry $\tilde{\phi}$ with $\tilde{\phi}(\tilde{x}_i) = \tilde{y}_i$. Then the isometry

$$\phi : x \mapsto \tilde{\phi}(x - x_0) + y_0$$

does the job.

(3b) $\Rightarrow$ (3a)**:** This is precisely the distance preserving property of Isometries. $\square$

Theorem F.2 (Characterization of Weak Input Invariances)**.** *Let* $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}$ *be a random function, then* $\mathbf{f}$ *is*

1. *weakly stationary, if and only if there exists* $\mu \in \mathbb{R}$ *and function* $C : \mathbb{R}^d \to \mathbb{R}$ *such that for all* x, y
$$\mu_{\mathbf{f}}(x) = \mu, \qquad \mathcal{C}_{\mathbf{f}}(x, y) = C(x - y).$$
2. *weakly non-stationary isotropic, if and only if there exist functions* $\mu : \mathbb{R}_{\geq 0} \to \mathbb{R}$ *and* $\kappa : D \to \mathbb{R}$ *with* $D = \{\lambda \in \mathbb{R}^2_{\geq 0} \times \mathbb{R} : |\lambda_3| \leq 2\sqrt{\lambda_1 \lambda_2}\} \subseteq \mathbb{R}^3$*. such that for all* x, y
$$\mu_{\mathbf{f}}(x) = \mu\big(\tfrac{\|x\|^2}{2}\big)$$
$$\mathcal{C}_{\mathbf{f}}(x, y) = \kappa\big(\tfrac{\|x\|^2}{2}, \tfrac{\|y\|^2}{2}, \langle x, y\rangle\big)$$
3. *weakly stationary isotropic, if and only if there exists* $\mu \in \mathbb{R}$ *and a function* $C : \mathbb{R}_{\geq 0} \to \mathbb{R}$ *such that for all* x, y
$$\mu_{\mathbf{f}}(x) = \mu, \qquad \mathcal{C}_{\mathbf{f}}(x, y) = C\big(\tfrac{\|x-y\|^2}{2}\big)$$

Proof. Starting from the mean and covariance function it is easy to check 2-weak non-stationary isotropy. So we only need to check the other direction.

The proof is essentially an application of Prop. F.4. For brevity (and since the other two results are well known), we will only prove the weakly non-stationary isotropic case (the other two cases can be proven with minor adjustments to the proof).

Without loss of generality, we will find the slightly different representation
$$\mathbb{E}[\mathbf{f}_d(x)] = \tilde{\mu}(\|x\|) \quad \text{and} \quad \mathcal{C}_{\mathbf{f}_d}(x, y) = \tilde{\kappa}(\|x\|, \|y\|, \langle x, y\rangle),$$
where the domain of $\tilde{\kappa}$ is given by $\tilde{D} = \{\lambda \in \mathbb{R}^2_{\geq 0} \times \mathbb{R} : |\lambda_3| \leq \lambda_1 \lambda_2\}$. The representation of the theorem is then equivalent by a change to
$$\mu(\lambda) := \tilde{\mu}\big(\tfrac{\lambda^2}{2}\big) \quad \text{and} \quad \kappa(\lambda_1, \lambda_2, \lambda_2) := \tilde{\kappa}\big(\tfrac{\lambda_1^2}{2}, \tfrac{\lambda_2^2}{2}, \lambda_3\big).$$

First we want to find μ. Let v be some vector (w.l.o.g. $\|v\| = 1$). Then we define
$$\mu(r) := \mathbb{E}[\mathbf{f}_d(rv)]$$
Now we need to show that this definition of μ is an appropriate mean function. For this choose any $x \in \mathcal{X}$. Then for $r = \|x\|$ there exists by Prop. F.4 (2.) a non-stationary isometry ϕ such that $\phi(x) = rv$ (we use $n = 1$). With 1-weak non-stationary isotropy of $\mathbf{f}_d$ this implies
$$\mathbb{E}[\mathbf{f}_d(x)] = \mathbb{E}[\mathbf{f}_d(rv)] = \mu(r) = \mu(\|x\|).$$
Next we need to define $\kappa(r_x, r_y, r_{xy})$. For this, choose two orthonormal vectors v, w. For every $r = (r_x, r_y, r_{xy}) \in \tilde{D}$ we define
$$x^*(r) = r_x v$$
$$y^*(r) = \frac{r_{xy}}{r_x} v + \sqrt{r_y^2 - \tfrac{r_{xy}^2}{r_x^2}}\, w.$$
Where $r \in \tilde{D}$ ensures $|r_{xy}| \leq r_x r_y$ and thus $r_y^2 - \frac{r_{xy}^2}{r_x^2} \geq 0$. Then we have
$$\|x^*(r)\| = r_x, \quad \|y^*(r)\| = r_y, \quad \text{and} \quad \langle x^*(r), x^*(y)\rangle = r_{xy}, \tag{27}$$
and define
$$\kappa(r_x, r_y, r_{xy}) := \mathcal{C}_{\mathbf{f}_d}(x^*(r), y^*(r)).$$
Again, we need to show that this kernel does the job. For this, choose any $x, y \in \mathcal{X}$. For
$$r := (\|x\|, \|y\|, \langle x, y\rangle),$$
which is in $\tilde{D}$ by the Cauchy-Schwarz inequality, the induced $x^*(r)$ and $y^*(r)$ satisfy by (27)
$$\|x^*(r)\| = \|x\|, \quad \|y^*(r)\| = \|y\| \quad \text{and} \quad \|x^*(r) - y^*(r)\| = \|x - y\|.$$
By Prop. F.4 (2.) there therefore exists an isometry ϕ such that $\phi(x) = x^*(r)$ and $\phi(y) = y^*(r)$. By 2-weak input isotropy of $\mathbf{f}_d$ we conclude
$$\mathcal{C}_{\mathbf{f}_d}(x, y) \overset{\text{isotrop.}}{=} \mathcal{C}_{\mathbf{f}_d}(x^*(r), y^*(r)) \overset{\text{def.}}{=} \kappa\big(\|x\|, \|y\|, \langle x, y\rangle\big). \qquad \square$$

G Technical

G.1 Conditional Gaussian distribution

For the following well known result we found a tidy proof giving insight into the reason it is true, so we wrote it down for your convenience but do not even expect this particular proof to be new.

Theorem G.1 (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 assuming Σ_{11} is invertible, 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

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

Proof. Let $\bar{X} := X - \mu$ be the centered version of X. There exists some lower triangular matrix L (even if Σ is only positive semidefinite only not uniquely) such that $\Sigma = LL^T$ (i.e. the Cholesky Decomposition). We can then write without loss of generality

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

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

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

So it follows that

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

with

$$\begin{aligned} \mu_{2|1} &:= \mu_2 + L_{21}Y_1 \\ \Sigma_{2|1} &:= L_{22}L_{22}^T. \end{aligned}$$

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

$$\Sigma = LL^T = \begin{bmatrix} L_{11}L_{11}^T & L_{11}L_{21}^T \\ L_{21}L_{11}^T & L_{22}L_{22}^T + L_{21}L_{21}^T \end{bmatrix}$$

This implies

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

so we have the desired conditional expectation, and finally

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

□

G.2 Covariance of derivatives

By Swapping integration and differentiation we have for a centered random function $\mathbf{f}$

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

So the covariance of a derivative of $\mathbf{f}$ with $\mathbf{f}$ is equal to a partial derivative of the covariance function [more details in 1]. Similarly other covariances can be calculated, e.g.

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

For this reason the derivatives of the covariance function are interesting as they represent the covariance of derivatives.

Applying this observation to isotropic covariance functions

$$\mathrm{Cov}(\mathbf{f}(x), \mathbf{f}(y)) = C\big(\tfrac{\|x-y\|^2}{2}\big)$$

we obtain.

Lemma G.2 (Covariance of derivatives)**.** *Let $\mathbf{f} \sim \mathcal{N}(\mu, C)$ and $\mathbf{d} = x - y$, then*

Cov	$\mathbf{f}(y)$	$\partial_j \mathbf{f}(y)$
$\mathbf{f}(x)$	$C(\frac{\|\mathbf{d}\|^2}{2})$	$-C'(\frac{\|\mathbf{d}\|^2}{2})\langle \mathbf{d}, e_j\rangle$
$\partial_i \mathbf{f}(x)$	$C'(\frac{\|\mathbf{d}\|^2}{2})\langle \mathbf{d}, e_i\rangle$	$-\Big[C''(\frac{\|\mathbf{d}\|^2}{2})\langle \mathbf{d}, e_j\rangle\langle \mathbf{d}, e_i\rangle + C'(\frac{\|\mathbf{d}\|^2}{2})\langle e_j, e_i\rangle\Big]$

G.3 Constrained linear optimization

Let U be a vectorspace. We define the projection of a vector w onto U by

$$P_U(w) := \operatorname*{argmin}_{v \in U} \|v - w\|^2$$

Lemma G.3 (Constrained maximiziation of scalar products)**.** *For a linear subspace $U \subseteq \mathbb{R}^d$, we have*

$$\max_{\substack{v \in U \\ \|v\| = \lambda}} \langle v, w\rangle = \lambda \|P_U(w)\| \tag{28}$$

$$\operatorname*{argmax}_{\substack{v \in U \\ \|v\| = \lambda}} \langle v, w\rangle = \lambda \frac{P_U(w)}{\|P_U(w)\|} \tag{29}$$

Before we get to the proof let us note that this immediately results in the following corollary about minimization.

Corollary G.4 (Constrained minimization of scalar products)**.**

$$\min_{\substack{v \in U \\ \|v\| = \lambda}} \langle v, w\rangle = -\lambda \|P_U(w)\| \tag{30}$$

$$\operatorname*{argmin}_{\substack{v \in U \\ \|v\| = \lambda}} \langle v, w\rangle = -\lambda \frac{P_U(w)}{\|P_U(w)\|} \tag{31}$$

Proof of Corollary G.4. The trick is to move one '$-$' outside from $w = -(-w)$

$$\min_{\substack{v \in U \\ \|v\| = \lambda}} \langle v, w\rangle = -\max_{\substack{v \in U \\ \|v\| = \lambda}} \langle v, -w\rangle = -\lambda \|P_U(w)\|$$

where we have used in the last equation that the projection is linear (we can move the minus sign out) and the norm removes the inner minus sign. The argmin argument is similar. □

Proof of Lemma G.3. **Step 1:** We claim that

$$v^* = \lambda \frac{P_U(w)}{\|P_U(w)\|}$$

results in the value $\langle v^*, w\rangle = \lambda \|P_U(w)\|$.

For this we consider

$$P_U(w) = \operatorname*{argmin}_{v\in U} \underbrace{\|v-w\|^2}_{=\|v\|^2-2\langle v,w\rangle+\|w\|^2} \tag{32}$$
$$= \operatorname*{argmin}_{v\in U} \underbrace{\|v\|^2 - 2\langle v, w\rangle}_{=:f(v)}$$

we know that $t \mapsto f(t\langle w\rangle_U)$ is minimized at $t = 1$ by the definition of $\langle w\rangle_U$. The first order condition implies

$$0 \overset{!}{=} \frac{d}{dt} = 2t\|P_U(w)\|^2 - 2\langle P_U(w), w\rangle$$

and thus

$$1 = t^* = \frac{\langle P_U(w), w\rangle}{\|P_U(w)\|^2}$$

Multiplying both sides by $\lambda\|P_U(w)\|$ finishes this step

$$\lambda\|P_U(w)\| = \Big\langle \underbrace{\lambda\frac{P_U(w)}{\|P_U(w)\|}}_{=v^*}, w\Big\rangle. \tag{33}$$

Step 2: By (33), we know that we can achieve the value we claim to be the maximum (and know the location v^* to do so). So if we prove that we can not exceed this value, then it is a maximum and v^* is the argmax. This would finish the proof. What remains to be shown is therefore

$$\langle v, w\rangle \le \lambda\|P_U(w)\| \qquad \forall v \in U : \|v\| = \lambda.$$

Let $v \in U$ with $\|v\| = \lambda$. Then for any $\mu \in \mathbb{R}$ we can plug μv into f from (32) to get

$$\begin{aligned}\mu^2\lambda^2 - 2\mu\langle v, w\rangle = f(\mu v)\\ &\overset{(32)}{\ge} f(P_U w) = \|P_U(w)\|^2 - 2\langle P_U w, w\rangle\\ &= -\langle P_U w, w\rangle\end{aligned}$$

where the last equation follows from (33) with $\lambda = \|P_U w\|$. Reordering we get for all μ

$$\langle P_U w, w\rangle + \mu^2\lambda^2 \ge 2\mu\langle v, w\rangle$$

We now select $\mu = \frac{\|P_U w\|}{\lambda} > 0$ and divide both sides by μ to get

$$2\langle v, w\rangle \le \underbrace{\Big\langle \underbrace{\frac{P_U(w)}{\mu}}_{=v^*}, w\Big\rangle + \lambda\|P_U(w)\|}_{\overset{(33)}{=}\lambda\|P_U w\|} = 2\lambda\|P_U w\|$$

Dividing both sides by 2 yields the claim. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 8 recapitulates all the assumptions made and highlights possible generalizations.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Section D provides all proofs of statements made in the main body and follows an identical structure for easier cross reference.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: While we do not have the necessary space to discuss all implementation details, we believe that we have discussed all relevant insights necessary to reproduce our results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code of our algorithm is fairly well commented and passes pylint and flake8 linting. The code to perform the benchmarks is less polished and we have not seeded the covariance estimation sampling process, but, since we obtained similar results over multiple runs (Section A.1.1), we are confident that our results are reproducible.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Training on the MNIST data set is fairly standard, so we feel like our brief outline is sufficient.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Quantiles are plotted in Figure 3 and the figures of Section A and we provided a histogram of asymptotic learning rates resulting from multiple covariance estimation runs (Section A.1.1).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: We have not kept careful track of resources used, as MNIST is a fairly small dataset for machine learning standards. We believe the compute was comparatively negligible, although the use of multiple GPUs was helpful in repeating experiments in parallel.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: See broader impacts.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Since we focus on optimization theory our work has no societal impact beyond the advancement of the the field of Machine Learning, which may have many societal consequences, but none we feel necessary to address.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite datasets and models used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.