Journal of Machine Learning Research 25 (2024) 1-78 Submitted 5/23; Revised 12/23; Published 3/24

Tight Convergence Rate Bounds for Optimization Under Power Law Spectral Conditions

Maksim Velikanov MAKSIM.VELIKANOV@TII.AE
Technology Innovation Institute, Abu Dhabi, UAE;
CMAP, Ecole Polytechnique, Paris, France

Dmitry Yarotsky D.YAROTSKY@SKOLTECH.RU
Center for Artificial Intelligence Technology
Skolkovo Institute of Science and Technology
Moscow, Russia

Editor: Mehryar Mohri

Abstract

Performance of optimization on quadratic problems sensitively depends on the low-lying part of the spectrum. For large (effectively infinite-dimensional) problems, this part of the spectrum can often be naturally represented or approximated by power law distributions, resulting in power law convergence rates for iterative solutions of these problems by gradient-based algorithms. In this paper, we propose a new spectral condition providing tighter upper bounds for problems with power law optimization trajectories. We use this condition to build a complete picture of upper and lower bounds for a wide range of optimization algorithms – Gradient Descent, Steepest Descent, Heavy Ball, and Conjugate Gradients – with an emphasis on the underlying schedules of learning rate and momentum. In particular, we demonstrate how an optimally accelerated method, its schedule, and convergence upper bound can be obtained in a unified manner for a given shape of the spectrum. Also, we provide first proofs of tight lower bounds for convergence rates of Steepest Descent and Conjugate Gradients under spectral power laws with general exponents. Our experiments show that the obtained convergence bounds and acceleration strategies are not only relevant for exactly quadratic optimization problems, but also fairly accurate when applied to the training of neural networks.

Keywords: Gradient Descent, Steepest Descent, Heavy Ball, Conjugate Gradients, power-law spectrum, convergence rate, tight bounds, non-strongly-convex least squares, acceleration, neural networks

©2024 Maksim Velikanov and Dmitry Yarotsky.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v25/23-0698.html.

Contents

1. Introduction

Modern large-scale optimization problems, such as training of neural networks, are typically solved by some variants of Gradient Descent (GD) or its accelerated versions. Examples of such methods include Stochastic Gradient Descent (SGD), GD with momentum (Polyak, 1964; Qian, 1999), Nesterov's accelerated gradient (Nesterov, 1983), Conjugate Gradients (CG, Hestenes and Stiefel (1952)), ADAM (Kingma and Ba, 2014). These first order methods strike a good balance between universality, efficiency and complexity, which is crucial for high-dimensional applications (where, for example, higher-order Hessian-based methods would be prohibitively expensive).

While real world optimization problems can be characterized by a multitude of different aspects, the key features of first order methods are well captured by examining optimization of quadratic loss functions $L(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T A\mathbf{w} - \mathbf{w}^T\mathbf{b}$, that typically serve as reasonable approximation to the actual objective functions near local or global minima. The main challenge in optimizing such quadratic losses is their ill-conditioning, i.e. some of the eigenvalues of A being much smaller than the others. The convergence rate of the optimization is determined by the condition number of A and notably degrades as this number tends to infinity. The extreme case is when the condition number is effectively infinite, i.e. the eigenvalues of A can be arbitrarily small. In this case, there are well-known classical bounds (see e.g. section 6.1 of Polyak (1987)) for the convergence rate in terms of the initial error $\|\mathbf{w}_0 - \mathbf{w}_*\|$, where $\mathbf{w}_0$ is the starting point and $\mathbf{w}_*$ is the minimizer. Specifically, for the vanilla GD $\mathbf{w}_{n+1} = \mathbf{w}_n - \alpha\nabla L(\mathbf{w}_n)$ with learning rate $\alpha < 2/\lambda_{\max}$, where $\lambda_{\max}$ denotes the largest eigenvalue of A, we have

$$L(\mathbf{w}_n) - L(\mathbf{w}_*) \leq \frac{\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{4\alpha e n}. \tag{1}$$

For optimization by Conjugate Gradients (CG), we have

$$L(\mathbf{w}_n) - L(\mathbf{w}_*) \leq \frac{\lambda_{\max}\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{2(2n+1)^2}. \tag{2}$$

These bounds suggest, in particular, that the convergence is $O(n^{-1})$ for GD, and $O(n^{-2})$ for CG.

However, bounds (1), (2) are crude in that they do not use any information about the distribution of eigenvalues in the segment $[0, \lambda_{\max}]$ and about the expansion coefficients of the initial displacement $\mathbf{w}_0 - \mathbf{w}_*$ over the eigenbasis of A. As a results, actual convergence rates in practical problems can be drastically different from the above $O(n^{-1})$ or $O(n^{-2})$. In fact, the experimentally observed convergence can even be slower than $O(n^{-1})$, seemingly contradicting the theory. In Figure 1 (left) we show the loss trajectory of a neural network in a very basic example – learning the standard MNIST digit classifier (LeCun et al., 2010) by basic GD in a kernel regime (see Section H.1 for details). We see that up to very late iterations, loss evolves as

$$L(\mathbf{w}_n) \propto n^{-\xi}, \quad \xi \approx 0.25. \tag{3}$$

This power law can be explained theoretically by observing that both the eigenvalue distribution and the cumulative distribution of target expansion coefficients in this problem are also approximate power laws, with exponents $\kappa \approx 0.34$ and $\nu \approx 1.35$, respectively (see

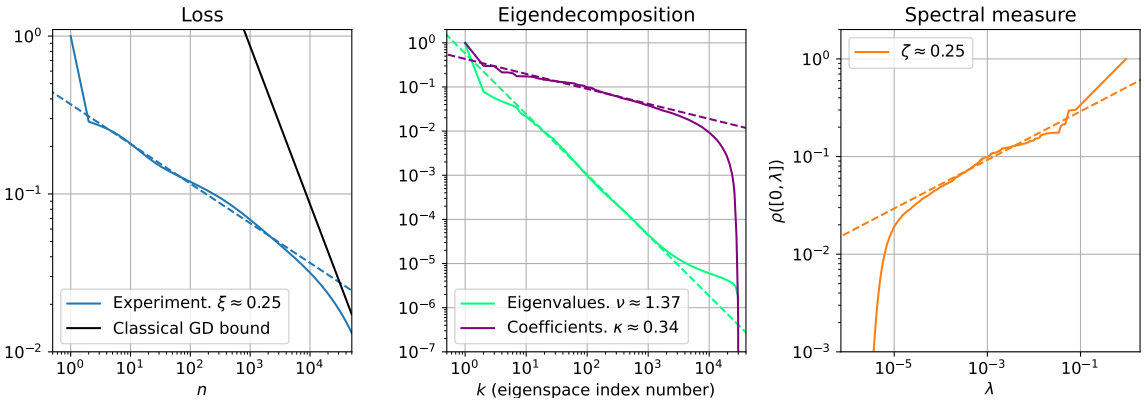

Figure 1: Spectral properties and GD loss of a MNIST classifier learned in a kernel regime. **Left:** The experimental trajectory $L(\mathbf{w}_n)$ of GD loss, the fitted power law (3), and the classical $O(n^{-1})$ bound (1). **Center:** The eigenvalues $\lambda_k \propto k^{-\nu}, \nu \approx 1.37$, and the cumulative distribution of target expansion coefficients, $\sum_{s=k}^{k_{\max}} \lambda_s c_s^2 \propto k^{-\kappa}, \kappa \approx 0.34$. The target expansion coefficients c_k are defined by the spectral expansion $\mathbf{w}_* = \sum_k c_k \mathbf{e}_k$ of the minimizer vector $\mathbf{w}_*$ over the eigenvectors $\mathbf{e}_k$. **Right:** The spectral measure $\rho([0,\lambda]) = \sum_{k:\lambda_k<\lambda} \lambda_k c_k^2 \propto \lambda^\zeta, \xi = \zeta = \frac{\kappa}{\nu} \approx 0.25$. See Section 2.1 for a general definition.

Figure 1 (center) and later sections for details). The power law (3) for the loss can then be derived from these spectral laws with the exponent given by $\xi = \frac{\kappa}{\nu} \approx 0.25$.

The apparent contradiction between the $O(n^{-1})$ theoretical bound (1) and the much slower experimental convergence (3) is explained by the heavy tail of the eigendecomposition of the fitted function. If we attempt to view the problem as effectively infinite-dimensional (which is convenient for abstract theory involving spectral power laws), then under condition $\frac{\kappa}{\nu} \leq 1$ the minimizer $\mathbf{w}_*$ does not exist as a finite-norm vector (is "unattainable"). Accordingly, the norm $\|\mathbf{w}_0 - \mathbf{w}_*\|$ appearing in (3) becomes infinite and bound (3) becomes vacuous. If instead we treat the space as high- but finite-dimensional, then the norm $\|\mathbf{w}_0 - \mathbf{w}_*\|$ is finite but very large, so that bound (1) is still too crude to reflect the actual convergence. Note, at the same time, that in the context of predictive modeling we are primarily interested in the loss $L(\mathbf{w}_n)$ rather than the norm $\|\mathbf{w}_n - \mathbf{w}_*\|$, since the former directly reflects the performance of the models while the latter only characterizes convergence in terms of the internal structure of the model and depends on the model parameterization, the choice of the norm, etc. In the case of MNIST, despite large values of the norm $\|\mathbf{w}_n - \mathbf{w}_*\|$, the model trains well and achieves high accuracy even on the test set (see Section 6). This suggests that a theory describing training realistic machine learning models even as simple as a MNIST classifier need not in general assume existence of a finite-norm solution $\mathbf{w}_*$.

A power-law structure of the spectrum is a common property of many large-scale optimization problems, in particular in machine learning: see e.g. recent works Cui et al. (2021); Bahri et al. (2021); Lee et al. (2020); Canatar et al. (2021); Kopitkov and Indelman (2020); Dou and Liang (2021); Atanasov et al. (2021); Bordelon and Pehlevan (2021); Basri et al. (2020); Bietti (2021). One particularly interesting modern scenario is optimization

of neural networks in the "infinitely wide" NTK regime (Jacot et al., 2018) (or some other "lazy training" setting where the learning problem is linearized (Chizat et al., 2019)). Ill-conditioning here results naturally from overparameterization. In the NTK regime, the neural network effectively becomes a linear model with an explicit kernel (Lee et al., 2019). This can be used to derive explicit power laws for the corresponding spectral distributions and GD convergence rates. For example, when fitting a d-variate indicator function by a ReLU network using the continuous-time GD, the leading term in the loss evolution can be found as $Cn^{-1/(d+1)}$ with some explicit constant C (Velikanov and Yarotsky, 2021). A number of recent works experimentally verify and exploit power law asymptotics of the kernel eigenvalues, e.g. for the analysis of generalization (Bahri et al., 2021; Canatar et al., 2021; Lee et al., 2020; Jin et al., 2021).

This shows that power law spectral conditions are natural assumptions for abstract optimization theory. The power-law structure of the spectrum is commonly described by "source condition" and "capacity condition" (Caponnetto and De Vito, 2007). To the best of our knowledge, the first comprehensive study of several fundamental algorithms such as GD, CG and Heavy Ball (HB) in this setting was performed by Nemirovsky and Polyak who established a number of upper and lower bounds for convergence rates (Nemirovskiy and Polyak, 1984a,b) under the source condition. Their work was later extended in various directions by multiple authors. In particular, Brakhage (1987) introduced HB with a special schedule based on Jacobi polynomials, providing improved convergence bounds. Hanke (1991, 1996) pointed out several important connections between CG and the theory of orthogonal polynomial and proved tight lower bounds for convergence of CG in some special cases. Gilyazov and Gol'dman (2013) established upper bounds for convergence of the method of Steepest Descent (SD). In recent years, capacity and source conditions have been used in the context of kernel methods and Stochastic GD (SGD) to obtain power law convergence rate bounds $O(n^{-\xi})$ with different exponents ξ (Berthier et al., 2020b; Zou et al., 2021; Nitanda and Suzuki, 2021; Varre et al., 2021).

Our contribution. The present work is a comprehensive study of the fundamental first order optimization algorithms GD, CG, HB and SD in problems with a power-law type of the spectrum. On the one hand, our aim is to paint a complete rigorous picture of attainable convergence rates. We consider separately the scenarios with constant, non-constant predefined, and adaptive learning rates. For each algorithm we prove a power-law upper bound and a matching lower bound showing that the upper bound is tight. On the other hand, we introduce a new type of assumption to describe problems with a power-law type of the spectrum. We show that our new assumption provides a more accurate description of convergence rates, and develop a methodology of working with it. We highlight now some particular contributions of our work.

1. We give first general proofs of tight lower bounds for SD and CG, which were previously missing in the literature. This completes the full picture of upper and lower bounds for all considered algorithms.

2. For optimization problems with a power-law loss asymptotic $L(\mathbf{w}_n) \sim Cn^{-\zeta}$, we show that the optimal upper bound under the classical source condition acquires an

additional logarithmic factor $O(n^{-\zeta}\log n)$, while the upper bound under our spectral assumption recovers the correct rate $O(n^{-\zeta})$.

3. Our spectral assumption naturally treats attainable and unattainable problems in a unified way, in particular covering practical scenarios in which loss converges as a power law with an exponent close to 0, like in the above MNIST example.

4. We show that our spectral assumption simplifies the logic of derivation of optimally accelerated gradient descent methods. As a byproduct, we give a new simple expression for an optimal HB schedule.

5. Our experiments show that the considered accelerated gradient descent methods may achieve their theoretically expected convergence rate $O(n^{-2\zeta})$ for practical quadratic problems as well as for non-linear optimization of neural networks.

Paper organization. We describe our assumptions and optimization algorithms in Section 2. In Section 3 we summarize and briefly discuss our results. Detailed statements of the theoretical upper and lower convergence bounds are presented in Section 4. In Section 5 we accurately compare upper bounds obtained for the same problem using either our spectral condition or the classical source condition. In Section 6 we present experiments with all our optimization algorithms, including applications to neural network training. An additional literature review and proof details are deferred to the appendix.

2. The setting

2.1 Problem definition and spectral assumptions

The assumptions. We assume that the optimized quadratic loss function L is defined on a Hilbert space $\mathcal{H}$ by

$$L(\mathbf{w}) = \frac{1}{2}\|J\mathbf{w} - \mathbf{f}_*\|^2 = \frac{1}{2}\langle \mathbf{w}, A\mathbf{w}\rangle - \langle \mathbf{w}, \mathbf{b}\rangle + \frac{1}{2}\|\mathbf{f}_*\|^2, \tag{4}$$

where $J : \mathcal{H} \to \widetilde{\mathcal{H}}$ is a bounded linear operator mapping $\mathcal{H}$ to another Hilbert space $\widetilde{\mathcal{H}}$, $\mathbf{f}_* \in \widetilde{\mathcal{H}}$, and

$$A = J^\dagger J : \mathcal{H} \to \mathcal{H}, \quad \mathbf{b} = J^\dagger \mathbf{f}_* \in \mathcal{H} \tag{5}$$

($J^\dagger$ denotes the adjoint operator). The spaces $\mathcal{H}$ and $\widetilde{\mathcal{H}}$ are, in general, infinite-dimensional. The form (4) of the quadratic function appears naturally in the setting where J represents a linearized model fitting a target vector $\mathbf{f}_*$ (that, e.g., represents a large number of scalar measurements). If J is written as a matrix, its columns correspond to different "features" used to predict the target.

In the sequel, it is convenient to assume that $\ker(J) = \{0\}$ and that the range $\mathrm{Ran}(J)$ is dense in $\widetilde{\mathcal{H}}$.[1] This implies, in particular, that $\inf_{\mathbf{w}\in\mathcal{H}} L(\mathbf{w}) = 0$.

1. The extension to the general case is obtained easily by projecting or restricting all the vectors and operators to $\mathcal{H} \ominus \ker(J)$ in the space $\mathcal{H}$ and to $\mathrm{Ran}(J)$ in the space $\widetilde{\mathcal{H}}$.

Along with A, consider the unitarily equivalent positive definite operator

$$\widetilde{A} = JJ^\dagger : \widetilde{\mathcal{H}} \to \widetilde{\mathcal{H}}. \tag{6}$$

Given $\mathbf{f}_* \in \widetilde{\mathcal{H}}$, there is a (unique, scalar-valued) associated spectral measure $\rho = \rho_{\widetilde{A},\mathbf{f}_*}$ such that

$$\langle p(\widetilde{A})\mathbf{f}_*, \mathbf{f}_* \rangle = \int_{\mathbb{R}} p(\lambda)\rho(d\lambda) \tag{7}$$

for any polynomial p; this relation can then be extended to general Borel functions (see e.g. Birman and Solomjak (2012)). In particular, if $\widetilde{\mathcal{H}}$ is finite-dimensional or $\widetilde{A}$ is compact, then $\rho = \sum_{k=1}^{\dim \widetilde{\mathcal{H}}} c_k^2 \delta_{\lambda_k}$, where λ_k are the eigenvalues of $\widetilde{A}$, and c_k are the respective coefficients in the expansion of $\mathbf{f}_*$ over the orthonormal eigenvectors of $\widetilde{A}$. The measure ρ is finite ($\rho(\mathbb{R}) = \|\mathbf{f}_*\|^2 < \infty$) and supported on the finite interval $[0, \lambda_{\max}]$, where $\lambda_{\max} = \|\widetilde{A}\| = \|A\|$.

Our **main ("target expansion") spectral condition** is a growth condition on the cumulative distribution function of ρ:

$$\rho((0,\lambda]) \le Q\lambda^\zeta, \quad \lambda \in [0, \lambda_{\max}], \tag{8}$$

where Q and ζ are some positive constants. Note that this condition does not require $\widetilde{A}$ to have a discrete spectrum. It is sometimes convenient to fix $\lambda_{\max} = 1$ and $Q = 1$ for brevity:

$$\rho((0,\lambda]) \le \lambda^\zeta, \quad \lambda \in [0, 1]. \tag{9}$$

Results for general $\lambda_{\max}$ and Q can be recovered by rescaling $J \mapsto \lambda_{\max}^{1/2} J$ and $\mathbf{f}_* \mapsto Q^{1/2}\lambda_{\max}^{\zeta/2}\mathbf{f}_*$; in particular, the loss $L(\mathbf{w}_n)$ is simply multiplied by $Q\lambda_{\max}^\zeta$ (see Section C.1). In the rest of the paper we will always assume that $\lambda_{\max} = 1$, but occasionally keep the coefficient Q (e.g., this will be convenient for comparison with the classical source condition).

Our **secondary ("eigenvalue decay") spectral condition** assumes that the operator $\widetilde{A}$ is compact so that its spectrum is discrete, and that the sorted eigenvalues $\lambda_1 \ge \lambda_2 \ge \ldots > 0$ obey

$$\lambda_k \le \Lambda k^{-\nu} \tag{10}$$

with some positive constants Λ, ν[2]. We will see that this condition will only matter for the algorithm CG, but not the other algorithms we consider (GD, SD and HB).

We say that **the solution is attainable** if there exists $\mathbf{w}_* \in \mathcal{H}$ such that $L(\mathbf{w}_*) = 0$. In terms of the measure ρ, attainabilitity means that $\|\mathbf{w}_*\|^2 = \|J^{-1}\mathbf{f}_*\|^2 = \int \lambda^{-1}\rho(d\lambda) < \infty$. This holds if $\zeta > 1$ in Eq. (8), and generally does not hold for $\zeta \le 1$ (see Section C.1). We don't require attainability: in our setting ζ can be any positive number.

2. Bounding eigenvalues from above may seem counter-intuitive as faster eigenvalue decay for fixed coefficients c_k leads to slower loss convergence. However, our main spectral condition (9) forces the coefficients c_k to decrease if we decrease the eigenvalues λ_k. Moreover, for all considered algorithms except CG the eigenvalue decay condition (10) will not actually matter given condition (9). For CG, faster eigenvalue decay leads to faster loss convergence, which justifies the $\le$ sign in (10)

Comparison with a standard "source condition". Our "target expansion" condition (8) is closely related to so-called "source condition" (Nemirovskiy and Polyak, 1984a; Caponnetto and De Vito, 2007; Berthier et al., 2020b; Varre et al., 2021), which is traditionally used to describe the problems with a power-law type of the spectral distributions. It is convenient to write this latter condition in the form

$$\|A^{-(\zeta'-1)/2}\mathbf{w}_*\|^2 \le Q' \tag{11}$$

with some parameters ζ', Q'. This inequality can be written as an integral inequality w.r.t. the spectral measure ρ given by Eq. (7): using the identity $\|A^a\mathbf{w}_*\|^2 = \int_0^\infty \lambda^{2a-1}\rho(d\lambda)$,

$$\int_0^{\lambda_{\max}} \lambda^{-\zeta'}\rho(d\lambda) \le Q'. \tag{12}$$

Accordingly, the difference between our (8) and classical (12) conditions is akin to the difference between L^∞- and L^1-norm bounds.

There is an approximate correspondence between the two conditions under which our exponent ζ matches the exponent ζ' of the classical condition. More precisely, let, as agreed, $\lambda_{\max} = 1$. Denote by $\mathrm{P}(\zeta, Q)$ the set of all spectral measures ρ on $[0, 1]$ satisfying condition (8), and analogously denote by $\mathrm{P}'(\zeta', Q')$ the set of spectral measures ρ on $[0, 1]$ satisfying the classical source condition (12). Then we prove (see section C.1)

Lemma 2.1. *Assuming* $\zeta, \zeta', Q, Q' > 0$,

$$\mathrm{P}(\zeta, Q) \subseteq \mathrm{P}'(\zeta', Q') \iff \begin{cases} \zeta' < \zeta \\ Q' \ge Q\frac{\zeta}{\zeta-\zeta'} \end{cases} \tag{13}$$

$$\mathrm{P}'(\zeta', Q') \subseteq \mathrm{P}(\zeta, Q) \iff \begin{cases} \zeta \le \zeta' \\ Q \ge Q' \end{cases} \tag{14}$$

This lemma shows that our condition with parameters ζ, Q is slightly weaker than the classical source condition with the same parameters. In particular, while the classical condition with some exponent ζ' always implies our condition with the same exponent, the converse is not true: our condition with some ζ implies the classical condition only for $\zeta' < \zeta$, and the allowed constant $Q' \propto \frac{1}{\zeta-\zeta'}$ deteriorates as $\zeta' \nearrow \zeta$. Nevertheless, we will see that our weaker condition implies loss upper bounds analogous to those available with the classical condition.

2.2 Optimization algorithms

We consider several classical iterative optimization algorithms (Polyak, 1987). All of them are first-order in the sense that they use only the values of the loss function and its gradients from current and previous iterations. It will be convenient to assume that the starting point of these algorithms is $\mathbf{w}_0 = 0$.

Gradient Descent (GD) is given by

$$\mathbf{w}_{n+1} = \mathbf{w}_n - \alpha_n \nabla L(\mathbf{w}_n) \tag{15}$$

$$= \mathbf{w}_n - \alpha_n(A\mathbf{w}_n - \mathbf{b}). \tag{16}$$

We consider two scenarios for GD: the learning rate α_n either does not depend on n, or may depend on n, but with a schedule predefined prior to optimization and depending only on the exponent ζ from the main spectral condition (9).

Steepest Descent (SD) is a modification of GD in which learning rate α_n is adaptively chosen at each iteration to optimize the loss:

$$\alpha_n = \arg\min_{\alpha} L(\mathbf{w}_n - \alpha \nabla L(\mathbf{w}_n)). \tag{17}$$

In our quadratic setting α_n can be explicitly written as

$$\alpha_n = \frac{\|\nabla L(\mathbf{w}_n)\|^2}{\langle A\nabla L(\mathbf{w}_n), \nabla L(\mathbf{w}_n)\rangle}. \tag{18}$$

Heavy Ball (HB) is a basic multi-step method (a.k.a. "GD with momentum") given by

$$\mathbf{w}_{n+1} = \mathbf{w}_n - \alpha_n \nabla L(\mathbf{w}_n) + \beta_n(\mathbf{w}_n - \mathbf{w}_{n-1}) \tag{19}$$

$$= \mathbf{w}_n - \alpha_n(A\mathbf{w}_n - \mathbf{b}) + \beta_n(\mathbf{w}_n - \mathbf{w}_{n-1}) \tag{20}$$

(for $n = 0$ the term $\beta_n(\mathbf{w}_n - \mathbf{w}_{n-1})$ is dropped). As with GD, we assume that the learning rates α_n, β_n are either constant or predefined n-dependent. Throughout the paper we assume, as is common, that $0 \le \beta_n < 1$.

Conjugate Gradients (CG) is defined by the same formula as HB, but (as with SD) with adaptively chosen learning rates minimizing the loss at each step:

$$\alpha_n, \beta_n = \arg\min_{\alpha,\beta} L(\mathbf{w}_n - \alpha \nabla L(\mathbf{w}_n) + \beta(\mathbf{w}_n - \mathbf{w}_{n-1})).$$

For a quadratic loss, the optimal α_n, β_n are given by

$$\alpha_n = \frac{\|\mathbf{r}_n\|^2\langle A\mathbf{p}_n, \mathbf{p}_n\rangle - \langle \mathbf{r}_n, \mathbf{p}_n\rangle\langle A\mathbf{r}_n, \mathbf{p}_n\rangle}{\langle A\mathbf{r}_n, \mathbf{r}_n\rangle\langle A\mathbf{p}_n, \mathbf{p}_n\rangle - \langle A\mathbf{r}_n, \mathbf{p}_n\rangle}, \tag{21}$$

$$\beta_n = \frac{\|\mathbf{r}_n\|^2\langle A\mathbf{r}_n, \mathbf{p}_n\rangle - \langle \mathbf{r}_n, \mathbf{p}_n\rangle\langle A\mathbf{r}_n, \mathbf{r}_n\rangle}{\langle A\mathbf{r}_n, \mathbf{r}_n\rangle\langle A\mathbf{p}_n, \mathbf{p}_n\rangle - \langle A\mathbf{r}_n, \mathbf{p}_n\rangle}, \tag{22}$$

$$\mathbf{r}_n = \nabla L(\mathbf{w}_n) = A\mathbf{w}_n - \mathbf{b}, \quad \mathbf{p}_n = \mathbf{w}_n - \mathbf{w}_{n-1} \tag{23}$$

(see Polyak (1987), Section 3.2.2). The fundamental importance of CG lies in the fact that, for quadratic problems, this algorithm is optimal among all first order methods generating new iterates $\mathbf{w}_{n+1}$ by shifting the initial point $\mathbf{w}_0$ along linear subspaces spanned by the previously computed gradients $\nabla L(\mathbf{w}_0), \ldots, \nabla L(\mathbf{w}_n)$.

3. Overview of results

The complete picture of upper and lower bounds. Our main theoretical result is an essentially complete picture of optimal convergence rates, summarized in Table 1: for each of the algorithms GD, SD, HB, CG, for each type of learning rate schedule (constant, predefined step-dependent, adaptive), for each $\zeta > 0$ we establish an upper bound of the

Table 1: Summary of convergence rates of $L(\mathbf{w}_n)$ for different algorithms and learning rate schedules under spectral assumptions (8) (i.e., $\rho((0,\lambda]) = O(\lambda^\zeta)$) and (10) (i.e., $\lambda_k = O(k^{-\nu})$). Assumption (10) only matters in the case of CG: in all other cases the spectrum does not even need to be discrete. For CG, if only assumption (8) holds, then $L(\mathbf{w}_n) = O(n^{-2\zeta})$; if additionally (10) holds, then $L(\mathbf{w}_n) = O(n^{-(2+\nu)\zeta})$. Each of the bounds in the table is tight in the sense that the respective exponents $\zeta, 2\zeta, (2+\nu)\zeta$ cannot be improved. The subscripts indicate the sections where the respective results are presented: roman for upper bounds, *italic* for lower bounds, and **bold** for both.

	Learning rates		
	constant	predefined n-dependent	adaptive
Single-step	**Gradient Descent (GD)**		**Steepest Descent (SD)**
	$O(n^{-\zeta})_{\mathbf{4.2}}$	$O(n^{-2\zeta})_{\textit{4.3, 4.4, 4.5}}$	$O(n^{-\zeta})_{\mathbf{4.6}}$
Multi-step	**Heavy Ball (HB)**		**Conjugate Gradients (CG)**
	$O(n^{-\zeta})_{\mathbf{4.2}}$	$O(n^{-2\zeta})_{\textit{4.3, 4.4, 4.5}}$	$O(n^{-2\zeta}) \mid O(n^{-(2+\nu)\zeta})_{\textit{4.3, 4.4, 4.5}}$

form $L(\mathbf{w}_n) = O(n^{-\xi})$ and a respective lower bound showing that the exponent ξ cannot be improved.

In all cases except CG, only our primary condition (9) matters for the convergence rate: adding the eigenvalue decay condition does not affect the rate. This is confirmed by the lower bounds, which are constructed to satisfy both conditions. CG is an exceptional case where adding the eigenvalue decay condition allows to improve the upper bound from $O(n^{-2\zeta})$ to $O(n^{-(2+\nu)\zeta})$. We prove that both these bounds are tight.

Note that adaptivity of learning rates does not improve convergence rate for single-step methods (GD vs. SD), but does improve it for multi-step methods (HB vs. CG). The exponents 2ζ of faster algorithms are twice as large as the exponents ζ of the basic ones (cf. (1), (2)). In a d-dimensional setting with finite d CG finds the exact solution after d iterations; the analog of this in our setting is the increased exponent $(2+\nu)\zeta$.

Though theoretically CG has the highest convergence rate $O(n^{-(2+\nu)\zeta})$, its practical implementation is not so efficient because of a fast accumulation of numerical errors. The indicated rate requires the polynomials associated with CG (see Section B) to have roots very close to the eigenvalues of A, which imposes strong requirements on the precision of computations. Also, the $O(n^{-2\zeta})$ convergence of GD with predefined schedule is very sensitive to non-quadratic perturbations of the problem. See experiments in Section 6.

In Table 1 we have four instances which enjoy convergence rates accelerated from $O(n^{-\zeta})$ to $O(n^{-2\zeta})$ or $O(n^{-(2+\nu)\zeta})$. In all these cases the stated rates are achieved using constructions based on Jacobi polynomials $P_n^{(a,b)}$ (see Section 4.3).

The classical bounds $O(n^{-1})$ and $O(n^{-2})$ for GD and CG, respectively (cf. Eqs. (1), (2)), are, up to the coefficients, special cases of the bounds $O(n^{-\zeta})$ and $O(n^{-2\zeta})$ when $\|\mathbf{w}_*\| < \infty$, since by Lemma 2.1 (specifically, by Eq. (14)) our main spectral condition (9) holds with $\zeta = 1$ in this case.

In the multi-step predefined step-dependent scenario, our lower bound (Theorem 4.11) applies to any method linearly expressing current step in terms of past gradients. Accordingly, this bound covers not only Heavy Ball, but also its modifications such as Nesterov Accelerated Gradient (NAG, Nesterov (1983)). We discuss NAG in Appendix A.

As already mentioned, most bounds of Table 1 (or some closely related bounds) already appeared in some form in earlier research (Nemirovskiy and Polyak, 1984a,b; Brakhage, 1987; Hanke, 1991, 1996; Gilyazov and Gol'dman, 2013), albeit under the stronger classical source assumption (11). Below we discuss various new elements of our work which were not present in earlier research.

Optimization with unattainable solutions. In almost all previous research, only the case of attainable solutions $\|\mathbf{w}_*\| < \infty$ (i.e., $\zeta > 1$ in Eq. (8)) is considered. However, as already pointed out in Section 1, even simple realistic problems such as MNIST have unattainable solutions. In fact, one can argue that this non-attainability is typical for a wide range of problems. In particular, it is shown in Velikanov and Yarotsky (2021) that in the d-dimensional kernel regression with kernels having homogeneous singularities of degree α, the task of fitting indicator functions corresponds to the exponents $\nu = 1 + \frac{\alpha}{d}, \kappa = \frac{1}{d}$ and $\zeta = \frac{\kappa}{\mu} = \frac{1}{d+\alpha}$ in Eqs. (8), (10). ReLU neural networks in the NTK regime are effectively such kernels models (Jacot et al., 2018) with $\alpha = 1$, so in these scenarios we always have $\zeta = \frac{1}{d+1} < 1$. Our bounds in Table 1 are valid for all $\zeta > 0$ and show that the non-attainability of the solution is not an obstacle for successful optimization.

Even more importantly, our lower bounds show that, regardless of the optimization algorithm, the exponent ζ in the loss power law $L(\mathbf{w}_n) = O(n^{-\zeta})$ will, in general, be close to 0 if ζ is close to 0, i.e. the optimization will inevitably be quite slow. This agrees with experiment and dispels the excessively optimistic theoretical expectations such as $L(\mathbf{w}_n) = O(n^{-1})$ and $L(\mathbf{w}_n) = O(n^{-2})$ that one might get from Eqs. (1), (2).

New bounds. Our significant new technical contributions are the tight lower bounds $\Omega(n^{-\zeta})$ and $\Omega(n^{-(2+\nu)\zeta})$ for SD and CG (see Theorems 4.16, 4.12). While the respective upper bounds were known from Nemirovskiy and Polyak (1984a); Hanke (1991, 1996); Gilyazov and Gol'dman (2013), the tight lower bounds were not available under any kind of power-law assumption.

We consider our lower bound $\Omega(n^{-(2+\nu)\zeta})$ for CG with discrete spectrum to be especially important, because CG can be viewed as an "ultimate" iterative first order algorithm for quadratic objectives: it essentially reconstructs the objective on the nested sequence of whole Krylov subspaces exhausting the space $\mathcal{H}$, and so in a sense optimally exploits all the iteratively available zero- and first-order information about the objective. Our lower bound then shows that even this optimal exploitation will not generally give fast convergence if ζ and ν are small (unless the problem or the algorithm are improved using some additional information about the problem – e.g., by pre-conditioning).

To the best of our knowledge, the only previously available lower bounds for CG in the power-law setting were given in Hanke (1996) and only covered two special cases $\nu = 1, 2$

for which explicit orthogonal polynomials are known. Our approach is completely different: for each $\zeta, \nu > 0$ we give a simple explicit example of the operator J and target $\mathbf{f}_*$ for which spectral conditions (8), (10) hold and $L(\mathbf{w}_n) = \Omega(n^{-(2+\nu)\zeta})$ (assuming $\zeta \notin \mathbb{Z}$; see Theorem 4.12).

Our tight lower bound $\Omega(n^{-\zeta})$ for SD also seems to be new. We give a simple proof based on the limiting periodic behavior of SD (Theorem 4.16).

Finally, our simple construction of the step-dependent schedule ensuring the improved convergence $L(\mathbf{w}_n) = O(n^{-2\zeta})$ for GD (see Theorem 4.9) does not seem to have been described in earlier literature.

Tighter bounds: a weaker spectral assumption. As already mentioned in Section 2, our "target expansion" condition (8) is a weaker version of a more standard "source condition" (11). In Section 5 we show that this difference between conditions can play a significant role. Specifically, we show for the MNIST quadratic optimization problem that the upper bounds based on classical condition (11) poorly describe the actual loss trajectory, while the upper bound based on our condition (8) matches the true trajectory much better. We confirm this empirical observation theoretically for problems with a power-law loss trajectory $L(\mathbf{w}_n) \propto n^{-\zeta}$. We prove that in such problems, the upper bound based on the classical condition acquires an additional logarithmic factor, $O(n^{-\zeta} \log n)$, while the bound based on our condition retains the correct rate $O(n^{-\zeta})$.

Tighter bounds: specifying the constant. A simplest example of an optimization problem exhibiting a $O(n^{-\xi})$ convergence rate is the exact power-law measure $\rho_\zeta([0,\lambda]) = \lambda^\zeta$, a boundary case of our condition (9). We interpret the loss of this problem, $L_n^{(\zeta)}$, as a reference point for convergence rates. It is easy to derive its full loss asymptotic $L_n^{(\zeta)} \overset{n\to\infty}{=} Cn^{-\zeta}(1+o(1))$ with a specific constant C. Then, we are able to provide upper bounds that are quite close to this typical performance. For example, for GD and HB with constant learning rates, the bound asymptotically matches the typical performance: $L(\mathbf{w}_n) \le L_n^{(\zeta)}(1+o(1))$. As for accelerated HB with the rate $O(n^{-2\zeta})$, the bound is just a few times larger than the typical performance, e.g. $L(\mathbf{w}_n) \le 4L_n^{(\zeta)}$ for $\zeta = 1$.

Unified picture of convergence bounds and acceleration. We develop a new, general and transparent approach to simultaneously obtain an upper and a matching lower loss bounds for most of the considered algorithms (except SD) under spectral conditions like (9) (see Sections 4.1 – 4.4). This is done by relating the convergence for general problems satisfying condition $\rho([0,\lambda]) \le G(\lambda)$ to the convergence of a "solvable" problem with "smooth" spectral measure $\rho(d\lambda) = G'(\lambda)d\lambda$. For the solvable problem the optimal learning rate schedule can be found analytically through the 3-term recurrence relation of the related system of polynomials orthogonal with weight $\lambda G'(\lambda)d\lambda$. Then we show that this schedule remains efficient for all problems subject to $\rho([0,\lambda]) \le G(\lambda)$.

4. Upper and lower bounds: detailed results

The structure of our exposition is shown in Figure 2. We start with a block of four sections establishing our methodology of working with spectral condition (9). In Section 4.1, we connect the loss convergence in a general problem described by our condition (9) with the

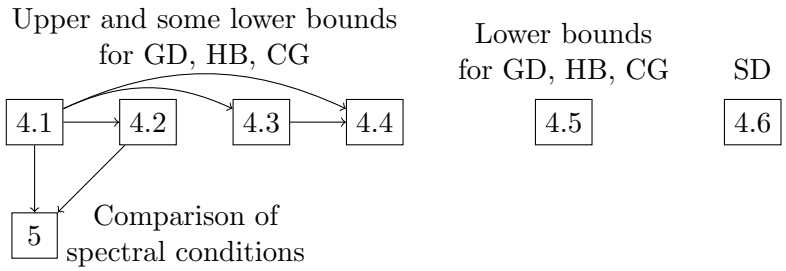

Figure 2: Logical dependencies between sections with main results.

convergence for an exact power-law measure. Then, in Section 4.2, we use this connection to establish upper and lower bounds for constant learning rate algorithms. In Section 4.3, we obtain accelerating strategies for the exact power-law spectral measure. Finally, in Section 4.4, we derive a number of upper bounds based on previously obtained accelerating strategy. In Section 4.5 we derive the lower bounds for general algorithms with predefined schedules and CG applied to a problem with discrete spectrum (10). Lastly, in Section 4.6, we consider the SD algorithm, which requires tools and reasoning different from the other algorithms.

Our proofs rely heavily on the spectral representation of optimization by residual polynomials, which is recalled in Section B. Each of the algorithms of Section 2.2 is represented by a sequence of residual polynomials $p_n(\lambda),\ p_n(0)=1$ so that the n-step solution $\mathbf{f}_n = J\mathbf{w}_n$ satisfies

$$\mathbf{f}_* - \mathbf{f}_n = p_n(\widetilde{A})\mathbf{f}_*. \tag{24}$$

The loss at step n can be expressed through p_n as

$$L(\mathbf{w}_n) = \frac{1}{2}\int p_n(\lambda)^2 \rho(d\lambda). \tag{25}$$

4.1 Worst-case measures under main spectral condition

In non-adaptive GD algorithms, the polynomials $p_n(\lambda)$ are fixed and independent of the problem's measure ρ. In this case, the worst case loss under a condition of the type (9) has a special structure revealed in the following theorem.

Theorem 4.1 (see proof in Section C.2)**.** *Let $G(x)$ be a nondecreasing absolutely continuous function on $[0,1]$ such that $G(0)=0$, and let $q(x)$ be any nonnegative polynomial on $[0,1]$. Consider the integral $\int q(x)\rho(dx)$ as a functional on measures ρ supported on $[0,1]$ and satisfying $\rho([0,x]) \le G(x)$ for all $x \in [0,1]$. Then the maximum of this functional is given by*

$$\sup_{\substack{\rho:\ \mathrm{supp}(\rho)\subset[0,1],\\ \rho([0,x])\le G(x)\forall x}} \int q(x)\rho(dx) = \int \overline{q}(x)G'(x)dx, \qquad \overline{q}(x) = \sup_{y\ge x} q(y). \tag{26}$$

We will refer to $\overline{q}(x)$ as a "flattened polynomial". Considering the case $G(\lambda)=\lambda^\zeta$, we see that this theorem allows to reduce the analysis of upper bounds under condition (9) to

estimating the averages of flattened polynomials $\overline{p_n^2}(x)$ over the exact power-law measure $\rho(d\lambda) = d(\lambda^\zeta)$.

The flattened polynomial $\overline{q}(x)$ can be simply characterized by considering the sequence of largest local maxima $0 \le x_1 < x_2 < \dots x_m \le 1$ of $q(x)$ on $[0,1]$ such that $\{q(x_i)\}_{i=1}^m$ is decreasing. Indeed, take any interval $[x_i, x_{i+1}]$ and denote $y_i \in (x_i, x_{i+1})$ the left most point such that $q(y_i) = q(x_{i+1})$. Then, it is straightforward to see that $\overline{q}(x) = q(x)$ on $[x_i, y_i]$ and $\overline{q}(x) = q(x_{i+1})$ on $[y_i, x_{i+1}]$, hence the name "flattened". See Figure 3 for an illustration.

Let us now outline the structure of convergence rate analysis that is suggested by Theorem 4.1 and will be behind most of our bounds for the algorithms GD and HB. For our main spectral condition (9) we have $G(\lambda) = \lambda^\zeta$, and equation (26) leads to the exact power-law spectral measure $\rho_\zeta(d\lambda) = d(\lambda^\zeta)$ with cumulative distribution function

$$\rho_\zeta((0,\lambda]) = \lambda^\zeta, \quad \lambda \in [0,1]. \tag{27}$$

For this measure, we define the *pair* of the worst-case loss given by Eq. (28) and the exact loss,

$$\overline{L_n^{(\zeta)}} = \frac{1}{2}\int_0^1 \overline{p_n^2}(\lambda) d(\lambda^\zeta), \tag{28}$$

$$L_n^{(\zeta)} = \frac{1}{2}\int_0^1 p_n^2(\lambda) d(\lambda^\zeta). \tag{29}$$

In our results, we will observe the following traits of this pair. First, the worst-case loss $\overline{L_n^{(\zeta)}}$ is not significantly worse than the exact power-law loss $L_n^{(\zeta)}$ and can be tightly bound to it. Then, $L_n^{(\zeta)}$ can be precisely described relying on a simple form of exact power-law measure $d(\lambda^\zeta)$ and properties of the chosen polynomials family $p_n(\lambda)$. Once the pair is characterized, we have the (tightest) upper bound $L(\mathbf{w}_n) \le \overline{L_n^{(\zeta)}}$, and a general (e.g., without discreteness restriction (10)) lower bound $L_n^{(\zeta)}$.

4.2 Constant learning rates (Section D)

Suppose that the learning rate $\alpha_n \equiv \alpha > 0$ and, if present, the momentum parameter $\beta_n \equiv \beta$. The respective residual polynomials for GD and HB are given by (see Section D.2)

$$p_n(\lambda) = (1-\alpha\lambda)^n, \tag{30}$$

$$p_n(\lambda) = \big(\sqrt{\beta}\big)^n \Big(U_n(z) - \sqrt{\beta} U_{n-1}(z)\Big), \quad z(\lambda) = \frac{1+\beta-\alpha\lambda}{2\sqrt{\beta}}, \tag{31}$$

where U_n are the Chebyshev polynomials of the second kind.

Following our strategy described in Section 4.1, we analyze the pair of losses $\overline{L_n^{(\zeta)}}$, $L_n^{(\zeta)}$ for the constant learning rate GD and HB characterized by Eqs. (30) and (31).

Theorem 4.2. *Define residual polynomials $p_n(\lambda)$ with (30) for $\beta = 0$ and with (31) for $0 < \beta < 1$, and assume $\alpha < 2(1+\beta)$. Consider the pair $\overline{L_n^{(\zeta)}}, L_n^{(\zeta)}$ of the worst-case loss (28) and the loss (29) for the exact power-law spectral measure. Then, as $n \to \infty$,*

$$L_n^{(\zeta)} = \frac{\Gamma(\zeta+1)}{2}\Big(\frac{2\alpha n}{1-\beta}\Big)^{-\zeta}(1+o(1)) \tag{32}$$

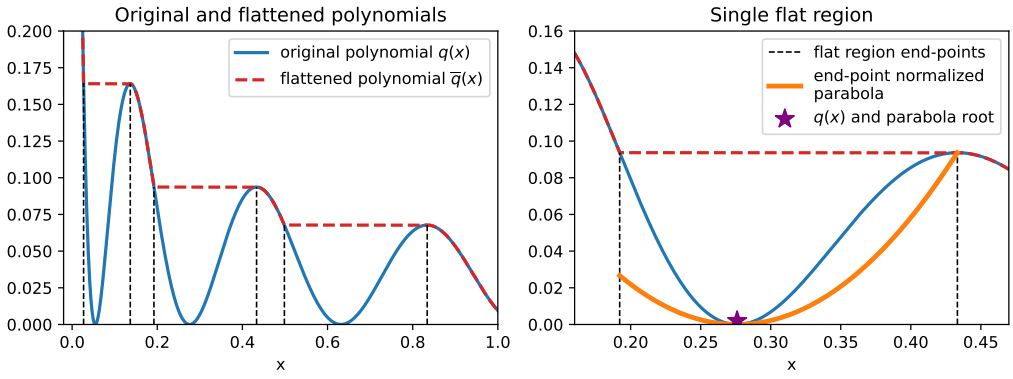

Figure 3: The "flattened polynomial" $\overline{q}(x)$ associated with the worst-case spectral measure (see Theorem 4.1). **Left:** The original polynomial $q(x) = p_n^2(x)$ and respective flattened polynomial $\overline{q}(x)$ for $p_n(x)$ associated with the Jacobi scheduled HB at step $n = 7$ (see Section 4.3). **Right:** Same as the left but zoomed in to a neighborhood of a single flat region of $\overline{q}(x)$. The orange parabola, placed at the respective root of $p_n(x)$ and normalized to match $q(x)$ at the right end of the flat region, is used to estimate the contribution of the flat region to the loss upper bound.

and

$$\overline{L_n^{(\zeta)}} = L_n^{(\zeta)} + \begin{cases} \mathbb{1}_{\alpha>1}\, O\big(u^n\big), & \beta = 0 \\ \mathbb{1}_{z_1<1} O\big(n^2 u^n\big), & \beta > 0 \end{cases} \tag{33}$$

where $z_1 = z(\lambda = 1)$ *and* u *is some value independent of* n *and such that* $0 < u < 1$.

Let's make a few remarks about Theorem 4.2. First, observe from (33) that the worst-case loss $\overline{L_n^{(\zeta)}}$ is just equal to the power-law loss $L_n^{(\zeta)}$ if $\alpha \le 1$ or $z_1 \ge 1$. The reason behind this is that the flattened polynomial from Theorem 4.1 is unchanged: $\overline{p_n^2}(\lambda) = p_n^2(\lambda)$, which holds for $p_n(\lambda)$ monotone decreasing on $[0, 1]$. For vanilla GD ($\beta = 0$) the monotonicity can be seen directly from (30), while for HB ($\beta \neq 0$) it requires more care but intuitively is connected to the localization of the roots of Chebyshev polynomials $U_n(z)$ on $[-1, 1]$.

Next, note that the difference between $\overline{L_n^{(\zeta)}}$ and $L_n^{(\zeta)}$ becomes exponentially small at large steps n. The speed of this exponential decay is given by parameter u, for which we obtain an explicit expression in the proof of the theorem. In particular, in GD without momentum $u = (1 - \alpha)^2$, and we can clearly observe that the convergence condition $\alpha < 2$ is equivalent to the condition $u < 1$ of exponential decay of the correction term.

Finally, we observe from (32) that the constant C in the asymptotic $L_n^{(\zeta)} = Cn^{-\zeta}(1 + o(1))$ can be made arbitrarily small by taking $\beta \nearrow 1$. In other words, higher "inertia" leads to faster convergence. We will see a reflection of this behavior in Section 4.4, where accelerated covergence rate $L_n^{(\zeta)} = O(n^{-2\zeta})$ is reached with the schedule of momentum behaving as $\beta_n \nearrow 1$, $n \to \infty$.

Now, we complete the picture for constant learning rate algorithms by establishing the lower bound in the class of discrete problems characterized by (10).

Theorem 4.3. *Consider the discrete spectral measure* $\rho_{\zeta,\nu} = \sum_{k=1}^{\infty} \left(k^{-\zeta\nu} - (k+1)^{-\zeta\nu}\right)\delta_{k^{-\nu}}$. *Then 1)* $\rho_{\zeta,\nu}$ *satisfies both conditions* (8) *and* (10)*; 2) the loss of constant learning rate GD and HB with* $\alpha < 2(1+\beta)$ *applied to the problem characterized by* $\rho_{\zeta,\nu}$ *is given by the right-hand side of* (32).

Basically, this result indicates that constant learning rate algorithms can not take advantage of discrete power-law spectrum $\lambda_k \le k^{-\nu}$. This fully settles $L(\mathbf{w}_n) = O(n^{-\zeta})$ as the tight bound for GD and HB in the case of constant learning rates.

4.3 A guide to acceleration: exact power-law spectral measure (Section E)

If we want to accelerate GD/HB, in the sense of decreasing the worst-case loss values $\overline{L_n^{(\zeta)}}$, Theorem 4.1 naturally guides us how to do that. Specifically, assume that in search for an accelerated algorithm we end up with a good enough family of polynomials $p_n(\lambda)$ such that $\overline{L_n^{(\zeta)}}$ and the exact power-law loss $L_n^{(\zeta)}$ are not far from each other, e.g. as in Theorem 4.2. Then, instead of minimizing $\overline{L_n^{(\zeta)}}$ we can focus on minimizing $L_n^{(\zeta)}$. The latter problem is well-defined and is given by

$$p_n = \operatorname*{arg\,min}_{q_n:\deg q_n = n, q_n(0)=1} \frac{1}{2}\int_0^1 q_n^2(\lambda)\rho_\zeta(d\lambda). \tag{34}$$

Recall (see Section B and specifically Eq. (117)) that the solution to (34) is exactly the CG algorithm applied to the exact power-law measure ρ_ζ. The corresponding optimal residual polynomial can be found by expressing its variation as $\delta p_n(\lambda) = \lambda r_{n-1}(\lambda)$ with arbitrary degree-$(n-1)$ polynomial r_{n-1} and then equating the variation of loss (29) to zero:

$$\delta L_n^{(\zeta)} = \int_0^1 p_n(\lambda) r_{n-1}(\lambda)\lambda\rho_\zeta(d\lambda) = 0, \tag{35}$$

implying that p_n is an orthogonal polynomial on $[0,1]$ w.r.t. the weight $\lambda\rho_\zeta(d\lambda) = \zeta\lambda^\zeta d\lambda$. Then p_n is a shifted and normalized Jacobi polynomial $P_n^{(\zeta,0)}(x)$:

$$p_n(\lambda) = \frac{P_n^{(\zeta,0)}(1-2\lambda)}{P_n^{(\zeta,0)}(1)}. \tag{36}$$

This leads to the precise convergence rate of Conjugate Gradients under the exact power-law measure ρ_ζ.

Theorem 4.4. *The losses of CG method applied to a problem with measure* (27) *are*

$$L(\mathbf{w}_n) = \frac{\Gamma^2(\zeta+1)n!^2}{2\Gamma^2(\zeta+n+1)} = \frac{\Gamma^2(\zeta+1)}{2} n^{-2\zeta}(1+o(1)) \quad (n\to\infty). \tag{37}$$

This result implies, in particular, that under the main spectral assumption (9) the CG loss $L(\mathbf{w}_n)$ will not, in general, decrease faster than $O(n^{-2\zeta})$. The same is also true for GD and HB, since their losses at any iteration are not less than the respective loss of CG.

The CG solution (36) suggests that other residual polynomials based on Jacobi polynomials might be good candidates for an accelerated GD method under power-law spectral conditions. Moreover, the prospects of applying (36) to practical problems require the robustness of the results with respect to errors in estimating the exponent ζ. To address these questions, we consider a 3-parameter *ansatz* of residual polynomials

$$q_n^{(a,b,r)}(\lambda) = \frac{P_n^{(a,b)}(1-r\lambda)}{P_n^{(a,b)}(1)}, \tag{38}$$

which contains (36) with parameters (a,b,r) set to $a=\zeta,\ b=0,\ r=2$. Then, we have

Proposition 4.5. *Consider residual polynomials $p_n(\lambda)$ given by* (38) *with $a,b>-\frac{1}{2}$ and $r<2$. The respective exact power-law measure loss* (29) *is given by*

$$L_n^{(\zeta)} = \begin{cases} \frac{\zeta\Gamma^2(a+1)B(\zeta,2a-2\zeta+1)}{2^\zeta r^{-\zeta}\Gamma^2(a-\zeta+1)} n^{-2\zeta}\big(1+o(1)\big), & a>\zeta-\frac{1}{2} \\ \frac{2^\zeta\zeta\Gamma^2(a+1)B(\frac{r}{2};\zeta-a-\frac{1}{2},b+\frac{1}{2})}{2\pi r^\zeta} n^{-2a-1}\big(1+o(1)\big), & a<\zeta-\frac{1}{2} \end{cases} \tag{39}$$

Observe that Eqs. (37) and (39) are consistent with each other. But most importantly, the condition $a>\zeta-\frac{1}{2}$ is critical to ensure the optimal convergence rate $O(n^{-2\zeta})$. Once this condition is ensured, the dependence on parameters (a,b,r) becomes *soft*: their variation only smoothly changes the constant without changing the rate $O(n^{-2\zeta})$.

Let us make explicit the connection between ansatz (38) and the associated HB method with n-dependent learning rates $\alpha_n,\ \beta_n$. The connection is enabled by $q_n^{(a,b,r)}$ being obtained from rescaled and n-independently shifted family of orthogonal polynomials. This implies that the sequence $q_n^{(a,b,r)}$ obeys a 3-term recurrence relation, which, due to the residual normalization, has exactly the form of momentum update: $p_{n+1}=p_n-\alpha_n\lambda p_n+\beta_n(p_n-p_{n-1})$. The resulting learning rates for the ansatz (38) are given by

$$\begin{cases} \alpha_n = r\frac{(2n+a+b+1)(2n+a+b+2)}{2(n+a+1)(n+a+b+1)} = 2r+O(n^{-1}), \\ \beta_n = \frac{n(n+b)(2n+a+b+2)}{(n+a+1)(n+a+b+1)(2n+a+b)} = 1-\frac{2a+1}{n}+O(n^{-2}). \end{cases} \tag{40}$$

Special cases of (38) and (40) were previously considered in Brakhage (1987) with parameters $a=\zeta-\frac{1}{2},\ b=-\frac{1}{2},\ r=2$, and in Hanke (1991) with parameters $a=\zeta,\ b=0,\ r=2$. Our general formula (40) allows to give an example of parameters (a,b) different from the cases considered by these authors and having a much simpler expression for learning rates. Specifically, with $a=b$ the Jacobi polynomials in (38) reduce to the ultraspherical polynomials $q_n^{(a,a,r)}=C_n^{a-\frac{1}{2}}(1-r\lambda)/C_n^{a-\frac{1}{2}}(1)$, and the respective learning rates are

$$\begin{cases} \alpha_n = 2r-\frac{2a+1}{n+2a+1}, \\ \beta_n = 1-\frac{2a+1}{n+2a+1}. \end{cases} \tag{41}$$

Our experiments (see Section 6) suggest that the $O(n^{-2\zeta})$ performance is retained even if we simplify the learning rate expressions even further, to the leading terms $\alpha_n=2r,\ \beta_n=1-\frac{2a+1}{n}$ in Eq. (40), but we do not have a proof of optimality in this case.

Importantly, Brakhage (1987) and Hanke (1991) used relatively indirect reasoning to arrive at their accelerated methods based on Jacobi polynomials. In contrast, our approach is straightforward – given a spectral condition $\rho([0,\lambda]) \le G(\lambda)$, one simply needs to take the system of polynomials orthogonal w.r.t. the weight function $\lambda G'(\lambda)$. In particular, we expect that our approach can be generalized to spectral conditions specified by functions $G(\lambda)$ other than power-laws.

4.4 General upper bounds

Jacobi ansatz (Section F.1). The key intuition employed in the previous sections was that the GD method efficiently minimizing $L_n^{(\zeta)}$ would also work for all problems specified by (9). We quantify this intuition in the following way:

Theorem 4.6. *Consider residual polynomials $p_n(\lambda)$ given by (38) with $a, b > -\frac{1}{2}$ and $r \le \frac{2a+1}{a+b+1}$. Then, the worst-case loss (28) is bounded in terms of exact power-law loss (29):*

$$\overline{L_n^{(\zeta)}} \le C_\zeta L_n^{(\zeta)}, \tag{42}$$

where the constant $C_\zeta = C[\rho_\zeta]$ is given by the functional $C[\rho]$ of measure ρ defined as

$$\frac{1}{C[\rho]} = \inf_{c, x_l, x_r \in \operatorname{supp}\rho} \left[\int_{x_l}^{x_r} \frac{(\lambda - c)^2}{\max\left((x_l - c)^2, (x_r - c)^2\right)} \rho(d\lambda) \Big/ \int_{x_l}^{x_r} \rho(d\lambda) \right] \tag{43}$$

The functional $C[\rho]$ has a simple geometric interpretation: the expression minimized in (43) is a ρ-weighted average of a parabola with center at c and normalized by its value at one of the edges x_l, x_r. The origin of this parabola is illustrated in Figure 3 (right): if a flat region of polynomial $\overline{p_n^2}(\lambda)$ contains only a single root, the true polynomials $p_n^2(\lambda)$ can be lower-bounded by a such normalized parabola. Looking at the contribution to the losses (28),(29) from this flat region reveals that their ratio is not worse than the ratio of the ρ-weighted averages of the constant and the normalized parabola over the flat region. Interestingly, the geometric picture depicted on Figure 3 (right) requires only basic properties of polynomials $p_n(\lambda)$: non-degeneracy of the roots and monotonicity of local extrema. We explicitly calculate the functional $C[\rho]$ for the exact power-law measure.

Proposition 4.7. *Let ρ_ζ be defined as in (27). Then*

$$C[\rho_\zeta] = \begin{cases} (\zeta+1)^2, & \zeta \ge 1 \\ 2 + 2/\zeta, & \zeta \le 1 \end{cases} \tag{44}$$

Now, we denote the coefficient in the $a > \zeta - \frac{1}{2}$ case of Eq. (39) by $R(a, r, \zeta)$, and summarize Theorem 4.6 and Propositions 4.5, 4.7 as

Corollary 4.8. *Let $a > \zeta - \frac{1}{2}$, $b > -\frac{1}{2}$, $r \le \frac{2a+1}{a+b+1}$. Then the loss of HB method with the schedule (40) applied to a problem described by condition (9) is*

$$L(\mathbf{w}_n) \le C_\zeta R(a, r, \zeta) n^{-2\zeta}(1 + o(1)) \tag{45}$$

The same bound obviously remains valid for CG, since its loss is dominated by the HB loss.

GD with predefined schedule (Section F.2). The above result ensures an $O(n^{-2\zeta})$ convergence of HB with a suitable problem-independent learning rate schedule. We show that, theoretically, such a rate can also be achieved for GD (i.e., without using momentum):

Theorem 4.9. *Given $\zeta > 0$, there exists a sequence α_n such that for any problem subject to spectral condition* (9)*, GD with this schedule α_n satisfies*

$$L(\mathbf{w}_n) \leq C_\zeta R(a, r, \zeta) 4^{2\zeta} n^{-2\zeta} \big(1 + o(1)\big), \tag{46}$$

where the parameter a and the constants $C_\zeta, R(a, r, \zeta)$ are as in Corollary 4.8.

The idea of the proof is to consider a subsequence of polynomials (38) with growing degrees 2^l, and choose the learning rates α_n as inverse roots of these polynomials.

We remark, however, that this construction requires very large learning rates α_n, which makes this algorithm, in contrast to HB with schedule (40), fairly unstable for non-linear models (see experiments in Section 6).

Conjugate Gradients: discrete spectrum (Section F.3). If the main spectral condition (9) is supplemented by eigenvalue decay condition (10), CG acquires quite different convergence rate $O(n^{-(2+\nu)\zeta})$:

Theorem 4.10. *Assuming spectral conditions* (9) *and* (10)*, the losses of CG satisfy*

$$L(\mathbf{w}_n) \leq C_\zeta R(a, r, \zeta) \Lambda^\zeta (n/2)^{-(2+\nu)\zeta} (1 + o(1)), \tag{47}$$

where the parameter a and the constants $C_\zeta, R(a, r, \zeta)$ are as in Corollary 4.8.

The proof is based on assigning half of the roots of trial polynomials q_n to the largest atoms λ_k of the spectral measure ρ, and then adjusting the remaining roots on the segment $[0, \Lambda(n/2)^{-\nu}]$ by rescaling and invoking Corollary 4.8.

4.5 Further lower bounds

Non-adaptive schedules (Section G.1). If an optimization algorithm has a predefined (non-adaptive) learning rate schedule (as in our GD or HB), then it cannot in general improve the exponent 2ζ in the convergence rate $O(n^{-2\zeta})$, even if we additionally assume the discreteness of the spectrum with a particular power law decay:

Theorem 4.11. *Consider any optimization algorithm of the form*

$$\mathbf{w}_{n+1} = \mathbf{w}_0 + \sum_{j=0}^{n} \alpha_{nj} \nabla L(\mathbf{w}_j) \tag{48}$$

with fixed (problem-independent) α_{nj}. Then for any $\zeta, \nu, \epsilon > 0$ there exists a problem with a compact A and $\mathbf{b}$ subject to

$$\lambda_n = n^{-\nu}(1 + o(1)), \quad n \to \infty, \tag{49}$$

$$\rho((0, \lambda]) = \lambda^\zeta (1 + o(1)), \quad \lambda \to 0+, \tag{50}$$

such that there is an infinite sequence $n_1 < n_2 < \dots$ for which

$$L(\mathbf{w}_{n_s}) > n_s^{-2\zeta - \epsilon}. \tag{51}$$

CG with discrete spectrum (Section G.2). We give an explicit example showing that the bound $L(\mathbf{w}_n) = O(n^{-(2+\nu)\zeta})$ established in Theorem 4.10 for CG under two spectral conditions (9), (10) cannot generally be improved. For any constants $\nu > 0$ and $\zeta > 0$, consider the operator J defined on the space l^2 of square-summable sequences $\mathbf{w} = (w_1, w_2, \dots)$ by

$$(J\mathbf{w})_n = \begin{cases} w_1, & n = 1, \\ n^{-\frac{\nu}{2}} w_n - (\frac{n}{n-1})^{\frac{1-(2+\nu)\zeta}{2}} (n-1)^{-\frac{\nu}{2}} w_{n-1}, & n = 2, 3, \dots \end{cases} \tag{52}$$

Next, let $\mathbf{f}_* = \mathbf{e}_1 = (1, 0, \dots)$. We will show that the quadratic problem (4) defined by these J and $\mathbf{f}_*$ is a desired example.

Let us clarify the idea behind this choice of the operator J. Its two-diagonal form implies that the respective Krylov subspaces are just the standard coordinate subspaces, which allows to easily compute the exact loss trajectory $L(\mathbf{w}_n)$ (statement 1 of the following theorem). On the other hand, the coefficients in Eq. (52) are adjusted to ensure the desired asymptotics of the eigenvalues λ_k and the spectral measure ρ (statements 2 and 3).

Theorem 4.12.

1. *The loss values of CG for the problem defined by the above J and $\mathbf{f}_*$ are*

$$L(\mathbf{w}_n) = \Big(2 \sum_{m=1}^{n+1} m^{(2+\nu)\zeta - 1}\Big)^{-1} = (1 + o(1)) \frac{(2+\nu)\zeta}{2} n^{-(2+\nu)\zeta}, \quad n \to \infty.$$

2. *For any $\nu > 0$ and $\zeta > 0$, $\widetilde{A} = JJ^\dagger$ is a compact operator with eigenvalues $\lambda_k = O(k^{-\nu})$.*

3. *For any non-integer $\zeta > 0$, the spectral measure ρ associated with $\widetilde{A}$ and $\mathbf{f}_*$ satisfies $\rho((0, \lambda]) = O(\lambda^\zeta)$ as $\lambda \to 0+$.*

The restriction to non-integer ζ in Statement 3 is due to our proof technique; it can probably be lifted using a more careful analysis. If ζ is non-integer, then Theorem 4.12 gives precisely an example of a problem satisfying spectral conditions (9), (10) and a lower bound $L(\mathbf{w}_n) = \Omega(n^{-(2+\nu)\zeta})$. If ζ is an integer, then we can still use the theorem for a slightly weaker conclusion: considering operator (52) with ζ replaced by $\zeta + \epsilon$ with an arbitrary $0 < \epsilon < 1$, we get an example satisfying spectral conditions (9), (10) and a lower bound $L(\mathbf{w}_n) = \Omega(n^{-(2+\nu)(\zeta+\epsilon)})$.

Proof [of Theorem 4.12] As a preliminary observation, note that $J^\dagger$ is given by

$$(J^\dagger \mathbf{x})_n = n^{-\nu/2} (x_n - (\tfrac{n+1}{n})^{(1-(2+\nu)\zeta)/2} x_{n+1}), \quad n = 1, 2, \dots \tag{53}$$

Statement 1. In the case of CG, $L(\mathbf{w}_n)$ is obtained by optimizing $L(\mathbf{w})$ over the Krylov subspace spanned by $\{(J^\dagger J)^m J^\dagger \mathbf{e}_1\}_{m=0}^{n-1}$. Note that $(J^\dagger J)^m J^\dagger = J^\dagger (JJ^\dagger)^m = J^\dagger \widetilde{A}^m$ and that $\widetilde{A}$ is three-diagonal, so that the vectors $\{\widetilde{A}^m \mathbf{e}_1\}_{m=0}^{n-1}$ span the coordinate subspace $\mathcal{H}_n$ spanned by $\mathbf{e}_1, \dots, \mathbf{e}_n$. Therefore,

$$L(\mathbf{w}_n) = \min_{\mathbf{x} \in \mathcal{H}_n} \tfrac{1}{2} \|JJ^\dagger \mathbf{x} - \mathbf{e}_1\|^2 = \min_{\mathbf{x} \in \mathcal{H}_n} \tfrac{1}{2} \|\widetilde{A}\mathbf{x} - \mathbf{e}_1\|^2. \tag{54}$$

Consider the vector $\mathbf{v} = (v_1, v_2, \ldots)$ defined by

$$v_m = \begin{cases} m^{((2+\nu)\zeta-1)/2}, & m = 1, \ldots, n+1, \\ 0, & m > n+1. \end{cases} \tag{55}$$

Then, using Eq. (53), $(J^\dagger \mathbf{v})_m = 0$ for $m = 1, \ldots, n$. Accordingly, $\langle \widetilde{A}\mathbf{x}, \mathbf{v}\rangle = \langle J^\dagger \mathbf{x}, J^\dagger \mathbf{v}\rangle = 0$ for any $\mathbf{x} \in \mathcal{H}_n$. On the other hand, it is easy to see that if a vector $\mathbf{u}$ in the coordinate subspace $\mathcal{H}_{n+1}$ is orthogonal to this $\mathbf{v}$, then $\mathbf{u} = \widetilde{A}\mathbf{x}$ for some $\mathbf{x} \in \mathcal{H}_n$. It follows that

$$L(\mathbf{w}_n) = \min_{\mathbf{x} \in \mathcal{H}_{n+1} \ominus \mathbf{v}} \tfrac{1}{2}\|\mathbf{x} - \mathbf{e}_1\|^2 = \frac{\langle \mathbf{e}_1, \mathbf{v}\rangle^2}{2\|\mathbf{v}\|^2} = \Big(2 \sum_{m=1}^{n+1} m^{(2+\nu)\zeta-1}\Big)^{-1}, \tag{56}$$

as desired.

Statement 2 is implied by the following (more detailed) characterization of the spectrum of $\widetilde{A}$.

Lemma 4.13. *The operator $\widetilde{A}$ is compact, and the sorted positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots > 0$ satisfy*

$$(2k)^{-\nu} \leq \lambda_k \leq 5k^{-\nu}. \tag{57}$$

Proof The compactness follows since J is approximated in norm by the finite-dimensional operators obtained by truncating the assignment (52). As a result of compactness, the spectrum of $\widetilde{A}$ is discrete and consists of nonnegative eigenvalues; the positive eigenvalues can be sorted in decreasing order. To lower bound the eigenvalues, use the minimax principle:

$$\lambda_k = \max_{\substack{\mathcal{H}_k \subset l^2: \\ \dim \mathcal{H}_k = k}} \min_{\mathbf{x} \in \mathcal{H}_k} \frac{\langle \widetilde{A}\mathbf{x}, \mathbf{x}\rangle}{\|\mathbf{x}\|^2}. \tag{58}$$

Choosing the subspace $\mathcal{H}_k$ spanned by $\mathbf{e}_2, \mathbf{e}_4, \ldots, \mathbf{e}_{2k}$, we get

$$\lambda_k \geq \min_{\mathbf{x} \in \mathcal{H}_k} \frac{\langle \widetilde{A}\mathbf{x}, \mathbf{x}\rangle}{\|\mathbf{x}\|^2} = \min_{\mathbf{x} \in \mathcal{H}_k} \frac{\|J^\dagger \mathbf{x}\|^2}{\|\mathbf{x}\|^2} \tag{59}$$

$$= \min_{\mathbf{x} \in \mathcal{H}_k} \frac{1}{\|\mathbf{x}\|^2} \sum_{m=1}^{k} [(\tfrac{2m}{2m-1})^{1-(2+\nu)\zeta}(2m-1)^{-\nu} + (2m)^{-\nu}] x_{2m}^2 \tag{60}$$

$$\geq (2k)^{-\nu}. \tag{61}$$

To upper bound λ_k use the minimax principle in a different form:

$$\lambda_k = \min_{\substack{\mathcal{G}_k \in l^2: \\ \dim(l^2 \ominus \mathcal{G}_k) = k-1}} \max_{\mathbf{x} \in \mathcal{G}_k} \frac{\langle \widetilde{A}\mathbf{x}, \mathbf{x}\rangle}{\|\mathbf{x}\|^2}. \tag{62}$$

Choosing $\mathcal{G}_k$ spanned by $\mathbf{e}_k, \mathbf{e}_{k+1}, \ldots$, we get

$$\begin{aligned}
\lambda_k &\le \max_{\mathbf{x}\in\mathcal{G}_k} \frac{\langle \widetilde{A}\mathbf{x}, \mathbf{x}\rangle}{\|\mathbf{x}\|^2} = \max_{\mathbf{w}\in\mathcal{G}_k} \frac{\|J^\dagger \mathbf{x}\|^2}{\|\mathbf{x}\|^2} && (63)\\
&= \max_{\mathbf{x}\in\mathcal{G}_k} \frac{1}{\|\mathbf{x}\|^2}\Big[k^{-\nu}x_k^2 + \sum_{m=k+1}^{\infty} (m^{-\nu/2}x_m - (\tfrac{m}{m-1})^{(1-(2+\nu)\zeta)/2}(m-1)^{-\nu/2}x_{m-1})^2\Big] \\
&\le \max_{\mathbf{x}\in\mathcal{G}_k} \frac{1}{\|\mathbf{x}\|^2}\Big[k^{-\nu}x_k^2 + \sum_{m=k+1}^{\infty} \big(2m^{-\nu}x_m^2 + 4(m-1)^{-\nu}x_{m-1}^2\big)\Big] && (64)\\
&\le \max_{\mathbf{x}\in\mathcal{G}_k} \frac{1}{\|\mathbf{x}\|^2}\Big[5k^{-\nu}\sum_{m=k}^{\infty} x_m^2\Big] && (65)\\
&= 5k^{-\nu}. && (66)
\end{aligned}$$

■

Statement 3 relies on the following resolvent bounds.

Proposition 4.14.

1. *Assuming* $2m < \zeta < 2m+1$ *for some integer* $m \ge 0$, *the vectors* $\widetilde{A}^{-m}\mathbf{e}_1$ *and* $\widetilde{A}^{-m}(\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1$ *exist as elements of* l^2 *and*
$$\langle \widetilde{A}^{-m}\mathbf{e}_1, \widetilde{A}^{-m}(\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1\rangle = O(\epsilon^{\zeta-2m-1}), \quad \epsilon \to 0+. \tag{67}$$

2. *Assuming* $2m+1 < \zeta < 2m+2$ *for some integer* $m \ge 0$, *the vectors* $J^{-1}\widetilde{A}^{-m}\mathbf{e}_1$ *and* $J^{-1}\widetilde{A}^{-m}(\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1$ *exist as elements of* l^2 *and*
$$\langle J^{-1}\widetilde{A}^{-m}\mathbf{e}_1, J^{-1}\widetilde{A}^{-m}(\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1\rangle = O(\epsilon^{\zeta-2m-2}), \quad \epsilon \to 0+. \tag{68}$$

The proof of this proposition is quite lengthy, and we defer it to Sections G.2.1 and G.2.2. Let us show how it implies the desired spectral bound.

Assume first that $2m < \zeta < 2m+1$ for some integer $m \ge 0$. By definition of the spectral measure,

$$\begin{aligned}
\langle \widetilde{A}^{-m}\mathbf{e}_1, \widetilde{A}^{-m}(\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1\rangle &= \int_0^\infty \lambda^{-m}\cdot\lambda^{-m}(\lambda+\epsilon)^{-1}\rho(d\lambda) && (69)\\
&\ge \int_0^\epsilon \epsilon^{-2m}(2\epsilon)^{-1}\rho(d\lambda) && (70)\\
&= \tfrac{1}{2}\epsilon^{-1-2m}\rho((0,\epsilon]). && (71)
\end{aligned}$$

It follows then by Statement 1 of Proposition 4.14 that

$$\rho((0,\lambda]) \le 2\epsilon^{1+2m}O(\epsilon^{\zeta-2m-1}) = O(\epsilon^\zeta), \tag{72}$$

as desired.

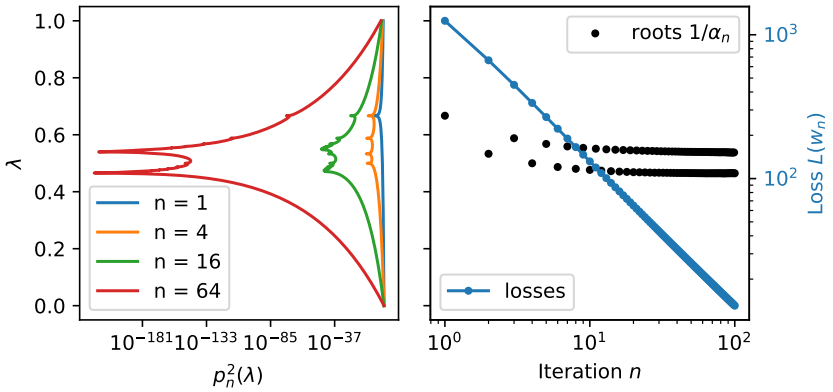

Figure 4: SD applied to the uniform spectral distribution ($\rho((0,\lambda]) = Q\lambda$) on $[0,1]$ converges to a period-2 oscillatory regime.

The case $2m+1 < \zeta < 2m+2$ is analyzed similarly, using part 2 of the proposition and the observation

$$\langle J^{-1}\widetilde{A}^{-m}\mathbf{e}_1, J^{-1}\widetilde{A}^{-m}(\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1\rangle = \int_0^\infty \lambda^{-m-1/2}\cdot\lambda^{-m-1/2}(\lambda+\epsilon)^{-1}\rho(d\lambda). \tag{73}$$

■

4.6 Steepest descent

Our analysis of SD is based on the remarkable asymptotic periodicity of this algorithm: as n increases, the adaptive learning rates α_n start to perform approximate period-2 oscillations, and the subsequences α_{2n} and α_{2n+1} converge (see Figure 4). This effect was first established, for finite-dimensional problems, in Akaike (1959). We will use a generalization to infinite-dimensional spaces proved in Pronzato et al. (2001).

Denote by $\lambda_{\min}$ and $\lambda_{\max}$ the left and right ends of the support of spectral measure ρ:

$$\lambda_{\min} = \sup\{\lambda : \rho((-\infty,\lambda)) = 0\}, \qquad \lambda_{\max} = \inf\{\lambda : \rho((\lambda,\infty)) = 0\}. \tag{74}$$

We will assume that $\lambda_{\min} \neq \lambda_{\max}$ (excluding the trivial case of a Dirac delta), so

$$0 \leq \lambda_{\min} < \lambda_{\max} < \infty. \tag{75}$$

It is convenient to introduce the inverses b_n of the learning rates α_n:

$$b_n = 1/\alpha_n. \tag{76}$$

The values b_n are the roots of the residual polynomials p_n associated with the iterates of SD (see Section B):

$$p_n(\lambda) = \prod_{s=0}^{n-1}(1-\lambda/b_s). \tag{77}$$

By definition of SD, α_n is obtained by optimizing

$$\int_{\lambda_{\min}}^{\lambda_{\max}} (1-\alpha\lambda)^2 p_n^2(\lambda)\rho(d\lambda) \to \min_{\alpha}. \tag{78}$$

This gives

$$\alpha_n = \frac{\int_{\lambda_{\min}}^{\lambda_{\max}} \lambda p_n^2(\lambda)\rho(d\lambda)}{\int_{\lambda_{\min}}^{\lambda_{\max}} \lambda^2 p_n^2(\lambda)\rho(d\lambda)}, \qquad b_n = \frac{\int_{\lambda_{\min}}^{\lambda_{\max}} \lambda^2 p_n^2(\lambda)\rho(d\lambda)}{\int_{\lambda_{\min}}^{\lambda_{\max}} \lambda p_n^2(\lambda)\rho(d\lambda)}. \tag{79}$$

Let us introduce the probability measure σ_n by

$$\sigma_n(d\lambda) = Z_n^{-1}\lambda p_n^2(\lambda)\rho(d\lambda), \tag{80}$$

where $Z_n = \int_{\lambda_{\min}}^{\lambda_{\max}} \lambda p_n^2(\lambda)\rho(d\lambda)$ is the normalizing factor. Eq. (79) shows that b_n is the mean of σ_n:

$$b_n = \int_{\lambda_{\min}}^{\lambda_{\max}} \lambda\sigma_n(d\lambda). \tag{81}$$

Moreover, using Eq. (77), the evolution of the measures σ_n with SD iterations is given simply by

$$\sigma_{n+1}(d\lambda) = D_n^{-1}(\lambda - b_n)^2\sigma_n(d\lambda), \tag{82}$$

where $D_n = \int_{\lambda_{\min}}^{\lambda_{\max}} (\lambda - b_n)^2\sigma_n(d\lambda)$ is the variance of σ_n.

By our assumptions, 0 is not an eigenvalue of $\widetilde{A}$ and so is not an isolated atom of the measure ρ. It follows that the measure $\sigma_0(d\lambda) = Z_0^{-1}\lambda\rho(d\lambda)$ has the same end points $\lambda_{\min}, \lambda_{\max}$ of its support as the measure ρ.

Evolution (82) admits a simple family of special period-2 solutions parameterized by $q \in (0,1)$:

$$\sigma_{2n} = q\delta_{\lambda_{\min}} + (1-q)\delta_{\lambda_{\max}}, \qquad \sigma_{2n+1} = (1-q)\delta_{\lambda_{\min}} + q\delta_{\lambda_{\max}}. \tag{83}$$

The following result shows that any sequence of iterates σ_n is attracted to one of these special solutions.

Theorem 4.15 (Theorem 2 in Pronzato et al. (2001))**.** *Consider iterations* (82) *starting from some compactly supported Borel probability measure σ_0 with end points $\lambda_{\min} < \lambda_{\max}$ of its support.*[3] *Then there exists $q \in (0,1)$ such that for any $\lambda \in (\lambda_{\min}, \lambda_{\max})$*

$$\sigma_{2n}([\lambda_{\min}, \lambda]) \overset{n\to\infty}{\longrightarrow} q, \qquad \sigma_{2n+1}([\lambda_{\min}, \lambda]) \overset{n\to\infty}{\longrightarrow} 1-q. \tag{84}$$

This result implies, in particular, that

$$b_{2n} \overset{n\to\infty}{\longrightarrow} q\lambda_{\min} + (1-q)\lambda_{\max}, \qquad b_{2n+1} \overset{n\to\infty}{\longrightarrow} (1-q)\lambda_{\min} + q\lambda_{\max}. \tag{85}$$

Using Theorem 4.15 and asymptotics (85), it is easy to connect the convergence rates of the SD evolution to those of GD with constant rates. The case $\lambda_{\min} > 0$ is discussed in Section

3. The statement of this theorem in Pronzato et al. (2001) also includes the condition $\lambda_{\min} > 0$, but it is clear that this condition can be dropped since evolution (82) is translation invariant.

5 of Pronzato et al. (2001); it is shown there that in this case the convergence of SD is (like that of GD) exponentially fast:

$$L(\mathbf{w}_n) = O\Big(\Big(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} + \epsilon\Big)^{2n}\Big), \quad n \to \infty, \tag{86}$$

for any $\epsilon > 0$. Consider now the case $\lambda_{\min} = 0$. The loss $L(\mathbf{w}_n)$ can be written in terms of σ_n as

$$L(\mathbf{w}_n) = \frac{Z_n}{2} \int_0^{\lambda_{\max}} \lambda^{-1} \sigma_n(d\lambda). \tag{87}$$

Applying Theorem 4.15, the leading contribution to this integral comes from small neighborhoods of $\lambda = 0$: for any $\widetilde{\lambda} \in (0, \lambda_{\max})$

$$\int_0^{\lambda_{\max}} \lambda^{-1} \sigma_n(d\lambda) = (1 + o(1)) \int_0^{\widetilde{\lambda}} \lambda^{-1} \sigma_n(d\lambda), \quad n \to \infty, \tag{88}$$

and accordingly

$$L(\mathbf{w}_n) = (1 + o(1)) \frac{1}{2} \int_0^{\widetilde{\lambda}} p_n^2(\lambda) \rho(d\lambda), \quad n \to \infty. \tag{89}$$

Now choose $\widetilde{\lambda} = \frac{1}{2} \inf_n b_n$. Using convergence (85) of the values b_n, we have $\widetilde{\lambda} > 0$. Recalling that the values b_n are the roots of the residual polynomials p_n, we can find constants $c_1, c_2 > 0$ such that for any $\lambda \in [0, \widetilde{\lambda}]$ and n

$$e^{-nc_1\lambda} \le p_n^2(\lambda) \le e^{-nc_2\lambda} \tag{90}$$

and so

$$(1 + o(1)) \frac{1}{2} \int_0^{\widetilde{\lambda}} e^{-nc_1\lambda} \rho(d\lambda) \le L(\mathbf{w}_n) \le (1 + o(1)) \frac{1}{2} \int_0^{\widetilde{\lambda}} e^{-nc_2\lambda} \rho(d\lambda), \quad n \to \infty. \tag{91}$$

Integrating by parts and making the change of variable $nc\lambda = t$,

$$\int_0^{\widetilde{\lambda}} e^{-nc\lambda} \rho(d\lambda) = e^{-nc\widetilde{\lambda}} \rho((0, \widetilde{\lambda}]) + \int_0^{nc\widetilde{\lambda}} e^{-t} \rho((0, \tfrac{t}{cn}])dt. \tag{92}$$

The first term falls off exponentially, while in the case of the power law measure $\rho((0, \lambda]) = \min(\lambda^\zeta, \lambda_{\max}^\zeta)$ the second term equals $\Gamma(\zeta + 1)(cn)^{-\zeta}(1 + o(1))$. Combined with Eq. (91), this immediately implies the desired upper and lower loss bounds:

Theorem 4.16. *Assuming the main spectral condition* (9)*, the SD loss obeys* $L(\mathbf{w}_n) = O(n^{-\zeta})$*. On the other hand, if we assume a lower bound* $\rho((0, \lambda]) = \Omega(\lambda^\zeta)$*, then* $L(\mathbf{w}_n) = \Omega(n^{-\zeta})$*.*

Recall the discrete measure $\rho_{\zeta,\nu} = \sum_{k=1}^{\infty} (k^{-\zeta\nu} - (k+1)^{-\zeta\nu}) \delta_{k^{-\nu}}$ that appeared in Theorem 4.3 and satisfies both main spectral condition (9) and eigenvalue decay condition (10). It is easy to see that $\rho((0, \lambda]) \ge 2^{-\zeta\nu} \lambda^{-\zeta}$ for $0 < \lambda \le 1$, so both statements of Theorem 4.16 are applicable to $\rho_{\zeta,\nu}$. It follows that the loss convergence bound $O(n^{-\zeta})$ is tight even if the main spectral condition (9) is supplemented by the eigenvalue decay condition (10).

We remark that a $O(n^{-\zeta})$ upper bound for the loss was obtained previously by a different method, based on moment inequalities, in Gilyazov and Gol'dman (2013) (see their Theorem 2.2.5). However, that method seems to require the stronger source condition (11) and does not produce tight lower bounds.

5. Comparison of spectral conditions

As discussed in Section 2.1, our target expansion condition (8) is a variant of the more standard source condition (12). In this section we compare the two versions and argue that our condition (8) can be more convenient and natural in applications. We have already shown in Lemma 2.1 and Section 4 that our condition (8) with a particular exponent ζ is slightly weaker than the respective source condition (12), but leads to similar power-law loss bounds $O(n^{-\zeta}), O(n^{-2\zeta}), O(n^{-(2+\nu)\zeta})$. We will argue now that, moreover, our condition generally better fits practical power-law spectra and produces tighter bounds when optimized over spectral parameters.

Upper bounds for classical source condition. We briefly recap the classical technique used for obtaining loss upper bounds under the classical source condition (12) (see, e.g. Polyak (1987); Nemirovskiy and Polyak (1984a); Brakhage (1987)). Recall that the loss is given by $L_n = \frac{1}{2}\int_0^1 p_n^2(\lambda)\rho(d\lambda)$ with a residual polynomial p_n associated with a particular optimization algorithm. Consider p_n as fixed and the loss $L_n = L_n(\rho)$ as a function of measure ρ. Under the classical source condition with parameters ζ', Q', the largest value of L_n is

$$\sup_{\rho\in \mathrm{P}'(\zeta',Q')} L_n(\rho) = \sup_{\rho\in \mathrm{P}'(\zeta',Q')} \frac{1}{2}\int_0^1 [\lambda^{\zeta'} p_n^2(\lambda)]\lambda^{-\zeta'}\rho(d\lambda) = \frac{Q'}{2}\sup_{0\le\lambda\le 1}\left[\lambda^{\zeta'} p_n^2(\lambda)\right]. \tag{93}$$

The value $\omega(\zeta', p_n) \equiv \sup_{0\le\lambda\le 1}\left[\lambda^{\zeta'} p_n^2(\lambda)\right]$ is the main object studied in Polyak (1987); Nemirovskiy and Polyak (1984a); Brakhage (1987) and other related works to characterize convergence rates. Note that the loss in (93) is maximized at the rescaled Dirac delta $\rho^* = Q'(\lambda^*)^{\zeta'}\delta_{\lambda^*}$, where $\lambda^* = \arg\max_{0\le\lambda\le 1}[\lambda^{\zeta'} p_n^2(\lambda)]$. This shows that the tightest upper bound under the source condition is

$$L_n^{UB}(\zeta', Q') = \sup_{\rho\in \mathrm{P}'(\zeta',Q')} L_n(\rho) = \frac{Q'}{2}\sup_{0\le\lambda\le 1}\left[\lambda^{\zeta'} p_n^2(\lambda)\right], \tag{94}$$

and the bound is especially accurate for measures close to the delta measure ρ^*. The value λ^* is n-dependent, so for any fixed measure $\rho \in \mathrm{P}(\zeta', Q')$ the bound (94) is necessarily suboptimal for all steps n except for a finite number of them.

This result is in stark contrast to its counterpart for our condition (9) described by Theorem 4.1. Specifically, if $p_n^2(\lambda)$ is monotone decreasing, the loss L_n is maximized by the exact power-law measure $\rho(d\lambda) = Qd\lambda^{\zeta}$. In the more general case of non-monotone $p_n^2(\lambda)$, the mass of the worst-case measure becomes partially redistributed towards the local maxima of $p_n^2(\lambda)$ while still being rather well-distributed overall (see proof of Theorem 4.1 for details). For problems with approximately power-law spectral measures, such well-distributed character of the worst-case measure results in accurate upper bounds for all steps n.

As an example of application of Eq. (94), consider vanilla GD with learning rate $\alpha < 2$. The respective polynomial is $p_n(\lambda) = (1-\alpha\lambda)^n$. The position of the Dirac delta can be found exactly by differentiating $\lambda^{\zeta'}(1-\alpha\lambda)^{2n}$ and is given by $\lambda^* = \alpha^{-1}\frac{\zeta'}{2n+\zeta'}$. Substituting

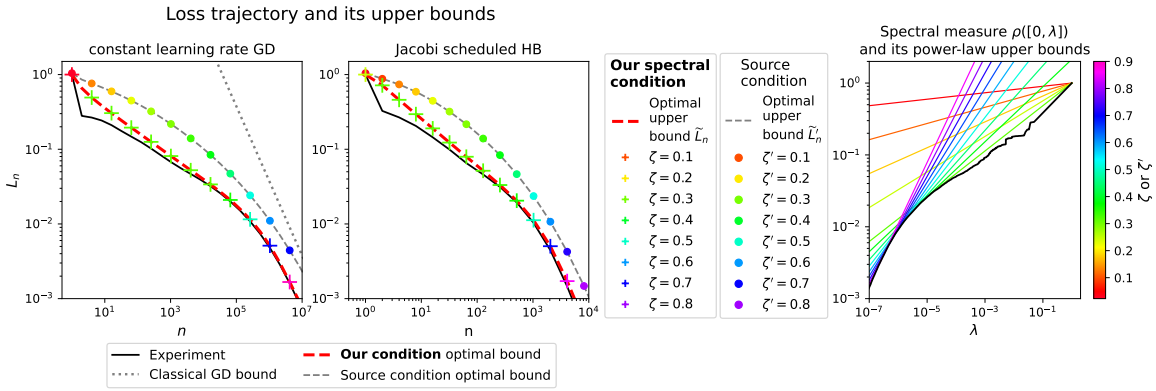

Figure 5: Comparison of the experimental loss and different upper bounds for a kernel regression on the MNIST dataset. **Left:** Loss trajectories and respective bounds for constant learning rate GD (left subfigure) and Jacobi scheduled HB (right subfigure). For both GD and HB, the two upper bound curves are given by the functions $\widetilde{L}_n, \widetilde{L}'_n$ defined in Eqs. (96),(97); in the GD case we additionally show the crude classical bound (1) corresponding to $\zeta = 1$ and requiring a very large constant C. The colors of the dots reflect the optimized values of ζ, ζ' in Eqs. (96), (97). **Right:** The actual spectral distribution ρ and different spectral bounds $\rho((0, \lambda]) \leq Q\lambda^{\zeta}$ with varying ζ and respective optimal Q.

this into (94) gives

$$L_n^{UB}(\zeta', Q') = \frac{Q'}{2}\left(1 - \frac{\zeta'}{2n+\zeta'}\right)^{2n}\left(\frac{\zeta'}{\alpha(2n+\zeta')}\right)^{\zeta'} \overset{n\to\infty}{=\!=} \frac{Q'}{2}\left(\frac{\zeta'}{2\alpha e}\right)^{\zeta'} n^{-\zeta'}(1+o(1)). \quad (95)$$

This $O(n^{-\zeta'})$ bound seems reasonable, but we will see later that it is suboptimal: it can only hold when the true loss does not have a power-law behavior with the same exponent ζ'.

A practical example. The above arguments suggest that our spectral condition and respective bounds should be more efficient than the classical source condition and respective bounds for problems with approximate power-law spectra. In Figure 5 we verify this conclusion experimentally on a kernel regression problem for the MNIST dataset, optimized either with constant learning rate GD or HB with Jacobi-based schedule (40) (see Section H.1 for further details).

For each step n and a given distribution ρ, we compute the respective optimal bounds $\widetilde{L}_n(\rho), \widetilde{L}'_n(\rho)$ obtained with our and classical source condition by

$$\widetilde{L}_n(\rho) = \inf_{\zeta, Q:\rho\in \mathrm{P}(\zeta,Q)} \sup_{\widetilde{\rho}\in \mathrm{P}(\zeta,Q)} L_n(\widetilde{\rho}), \quad (96)$$

$$\widetilde{L}'_n(\rho) = \inf_{\zeta', Q':\rho\in \mathrm{P}'(\zeta',Q')} \sup_{\widetilde{\rho}\in \mathrm{P}'(\zeta',Q')} L_n(\widetilde{\rho}). \quad (97)$$

In either case, in the inner supremum we choose the tightest upper bound available for given parameters ζ, Q or ζ', Q', and then in the outer infimum optimize it over all admissible parameters.

We observe in Figure 5 that the curves $\widetilde{L}_n$ corresponding to our spectral condition lie much closer to the actual loss trajectory than the curves $\widetilde{L}'_n$ corresponding to the classical source condition, in agreement with our prediction. Accordingly, when using our spectral condition, the optimal ζ stays the same until the late stages of training ($n \sim 10^5$ for GD and $n \sim 10^3$ for HB), meaning that a single spectral condition with fixed ζ, Q can efficiently describe the loss evolution. In contrast, for the classical source condition (12), the optimal parameters ζ', Q' are constantly changing along the whole optimization trajectory.

Theoretical suboptimality of the classical source condition. We state now the theoretical suboptimality result announced earlier and corroborating theoretical expectations and the experimental observations.

Theorem 5.1. *Assume that, for a certain spectral measure ρ, the sequence of the loss values under GD with constant learning rate $\alpha < 1$ is $L_n = Cn^{-\xi}(1 + o(1))$. Then, the respective optimal upper bounds $\widetilde{L}'_n, \widetilde{L}_n$ defined in Eqs.* (96), (97) *are given by*

$$\widetilde{L}_n = \left[\frac{Q}{2}\Gamma(\xi+1)(2\alpha)^{-\xi}\right] n^{-\xi}(1+o(1)), \quad Q = \sup_{\lambda\in(0,1]} \rho([0,\lambda])/\lambda^{\xi} < \infty, \tag{98}$$

$$\widetilde{L}'_n = \left[C\frac{\xi^{\xi+1}}{\Gamma(\xi+1)e^{\xi-1}}\right] \log(n) n^{-\xi}(1+o(1)). \tag{99}$$

This result shows that if the actual loss decreases as a power law, then the optimal upper bound (98) based on our spectral condition will agree with the actual loss up to a constant factor, while the optimal bound (99) based on the classical source condition will be off by at least a factor of $\log n$, even when we optimize the bound over the parameters Q', ζ'.

In the remainder of this section, let us outline the proof of Theorem 5.1 (see Section C.3 for details). First, we show by tauberian-type arguments that the loss asymptotic $L_n = Cn^{-\xi}(1 + o(1))$ implies a respective power-law asymptotic of the spectral measure: $\rho([0, \lambda]) = Q_\rho \lambda^{\xi}(1 + o(1))$, with $Q_\rho = 2C\frac{(2\alpha)^{\xi}}{\Gamma(\xi+1)}$. One can think of this as a partial converse (for $\beta = 0$) of theorem 4.2, hence the value of the constant Q_ρ.

Next, consider the exact power-law measure $\rho_\xi([0, \lambda]) = Q_\rho \lambda^{\xi}$. While the full proof needs to carefully take into account the correction $\rho - \rho_\xi$ at finite λ (in particular, leading to $Q > Q_\rho$ in (98)), the exact power-law measure captures the essence of the optimal bounds (98), (99). The optimal bound (98) for our condition is basically given by $L_n^{(\xi)}$ from theorem 4.2, since for the exact power-law measure ρ_ξ we have $\widetilde{L}_n(\rho_\xi) = Q_\rho \overline{L_n^{(\xi)}}$.

Turning to the second result (99), we note that the inner supremum in (97) is already derived in (95). As for the outer infimum in (97), the smallest possible Q' at a given ζ' can be inferred from lemma 2.1: $Q'(\zeta') = Q_\rho \frac{\xi}{\xi-\zeta'}$. From this point, we only need to estimate the optimal ζ' at a given iteration n:

$$\widetilde{L}'_n(\rho_\xi) = \inf_{0<\zeta'<\xi} \frac{Q_\rho}{2}\frac{\xi}{\xi-\zeta'}\left(\frac{\zeta'}{2\alpha e}\right)^{\zeta'} n^{-\zeta'}(1+o(1)) \overset{(*)}{=} \frac{Q_\rho \xi^{\xi+1}}{2(2\alpha e)^{\xi}} n^{-\xi}(1+o(1)) \inf_{0<\varepsilon<\xi} \frac{n^{\varepsilon}}{\varepsilon}. \tag{100}$$

Here in $(*)$, we took out all the factors that behave regularly at $\zeta' = \xi$, while the last infimum over $\varepsilon = \xi - \zeta'$ captures the essential tradeoff within the classical source condition:

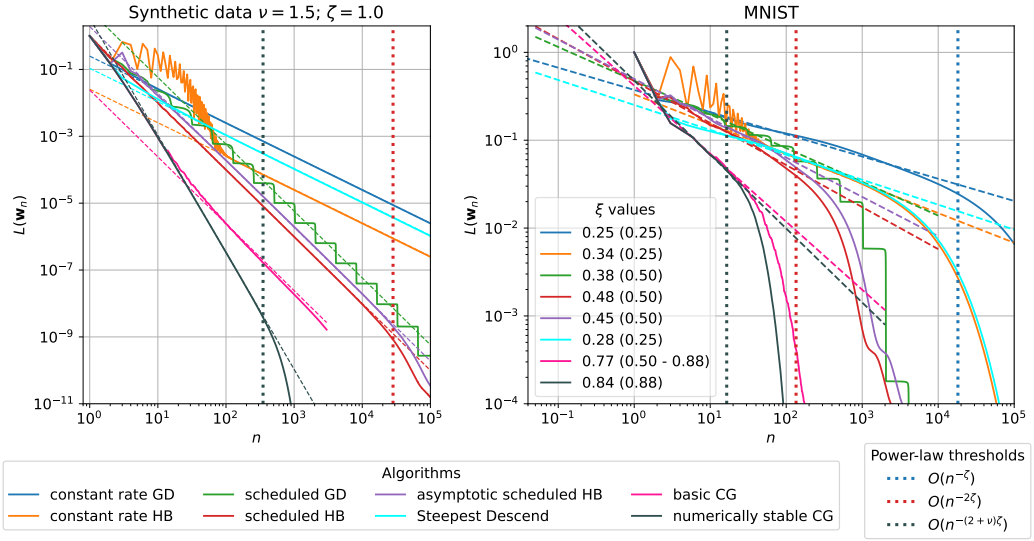

Figure 6: Loss trajectories of different optimization algorithms for the artificial diagonal problem with $\nu = 1.5$ and $\zeta = 1$ **(Left)**, and for (the $\{0, \dots, 9\}$-valued version of) MNIST learned by the NTK kernel of shallow ReLU network **(Right)**. The dashed lines are the fitted power-laws; the fitted (and calculated theoretically) exponents are shown in the legend. Dotted vertical lines correspond to the estimated threshold of validity of loss power-law (102).

higher values of ζ' are more favorable on the level of the rate $O(n^{-\zeta'})$ but they come at a price of a large constant $\propto \frac{1}{\xi - \zeta'}$. The logarithm in (99) appears as a result of this tradeoff:

$$\varepsilon_n^* \equiv \arg\min_{\varepsilon > 0} \frac{n^\varepsilon}{\varepsilon} = \frac{1}{\log n}, \qquad \inf_{\varepsilon > 0} \frac{n^\varepsilon}{\varepsilon} = \frac{n^{\varepsilon_n^*}}{\varepsilon_n^*} = e \log n. \tag{101}$$

6. Experiments

Diagonal matrices. We start with an artificial quadratic problem in which we can directly control the exponents ζ and ν: $\widetilde{A}$ is diagonal with eigenvalues $\lambda_k = k^{-\nu}$, and the respective coefficients of $\mathbf{f}_*$ are $c_k = k^{-\frac{\zeta\nu - 1}{2}}$. The size of $\widetilde{A} \in \mathbb{R}^{M \times M}$ is $M = 10^6$. The optimization results are shown in Figure 6 (Left). For all considered algorithms except CG, the losses have power-law rates with exponents ξ in accordance with Table 1 (shown by dashed lines). In Figure 6 the asymptotic scheduled HB algorithms are defined using the simplified versions of learning rate and momenta, obtained by discarding the $O(\dots)$ terms in Eq. (40). While we do not have a theoretical convergence rate for this method, we see that it has the same rate $O(n^{-2\zeta})$ as the full scheduled HB. This suggests that the correct asymptotic of α_n, β_n at $n \to \infty$ is a deeper reason for acceleration.

CG has the expected $\sim n^{-(2+\nu)\zeta}$ asymptotic only up to iteration $n_e \approx 20$, around which the asymptotic switches to $\sim n^{-2\zeta}$. This happens because of numerical errors (see further

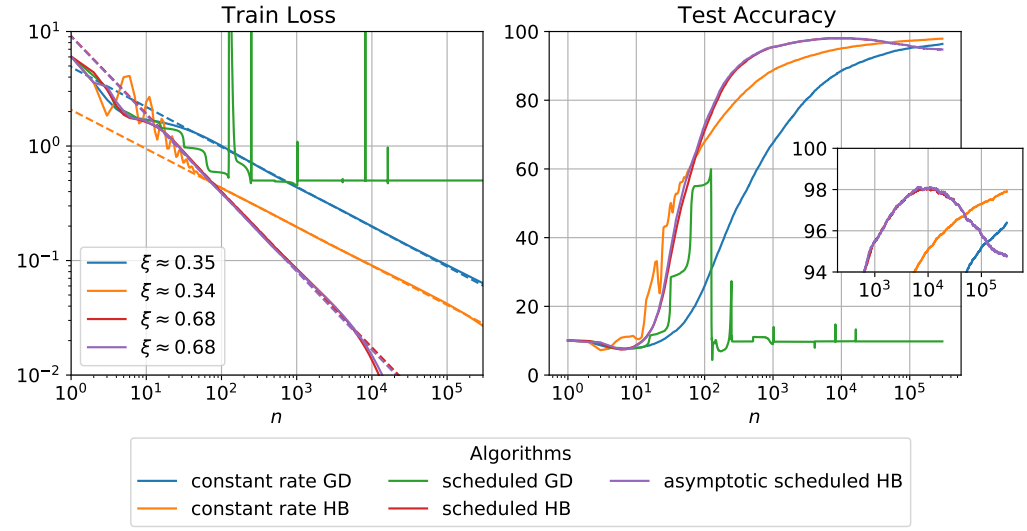

Figure 7: Train loss and test accuracy of different algorithms on MNIST learned by a shallow width-1000 neural network.

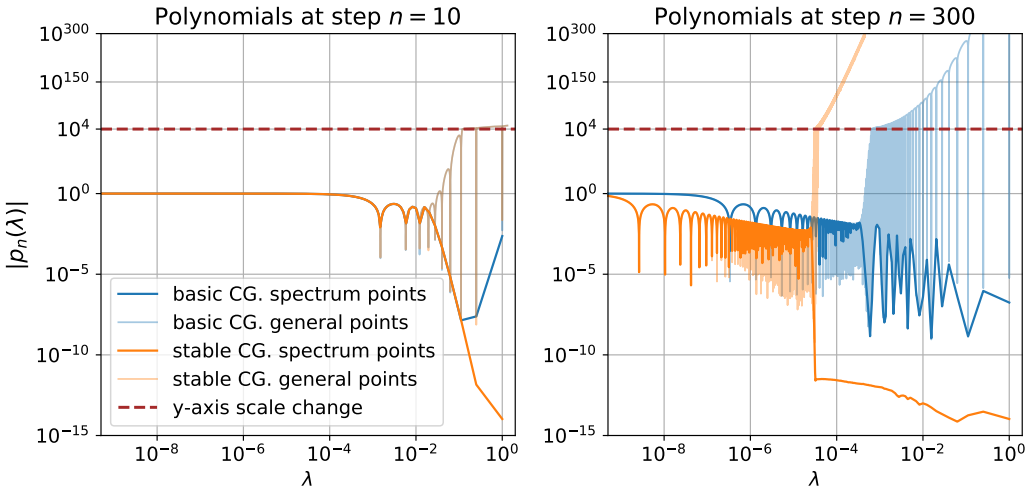

Figure 8: Residual polynomials for stable and unstable CG.

discussion in paragraph "CG polynomials" below). A version of CG modified to ensure stability exhibits the $\sim n^{-(2+\nu)\zeta}$ convergence to the very end.

Scheduled GD has a "staircase" shape because the schedule consists of size-2^l chunks (see Section F.2).

Note that for faster algorithms, such as CG or Jacobi scheduled HB, the power-law behavior of the loss breaks down at sufficiently large iteration n. This iteration can be estimated theoretically, as we explain below, and is depicted by vertical dotted lines in Figure 6.

Intervals of validity of loss power laws. When applied to real-life problems with approximately power-law spectra, the respective power-law behavior of optimization loss trajectories can be expected to hold only for moderately large iterations. In a real-life finite dimensional problem, the infinite-dimensional approximation $\rho([0,\lambda]) \sim \lambda^\zeta$ breaks down for $\lambda \lesssim \lambda_{\text{low}}$ with some characteristic value λ_{low} (e.g., the minimal positive eigenvalue). Under optimization, the loss is given by $L(\mathbf{w}_n) = \frac{1}{2}\int_0^{\lambda_{\max}} p_n^2(\lambda)\rho(d\lambda)$ with a suitable residual polynomial p_n. At large n, under the assumption of a power-law measure ρ, the leading contribution to this integral comes from the spectral interval $(0, \lambda_{\text{low}})$. Accordingly, the loss power law $L(\mathbf{w}_n) \propto n^{-s\zeta}$, where $s = 1, 2$ or $2+\nu$, breaks down for $n \gtrsim n_{\text{th}}$ with some characteristic iteration number n_{th}. In Section H.2 we derive a (non-rigorous) estimate of n_{th} :

$$n_{\text{th}} \propto \begin{cases} \frac{1-\beta}{\alpha\lambda_{\text{low}}}, & \text{for constant rate algorithms } (s=1) \\ \frac{1}{\sqrt{\lambda_{\text{low}}}}, & \text{for algorithms based on Jacobi polynomials } (s=2) \\ \lambda_{\text{low}}^{-\frac{1}{\nu+2}}, & \text{for (numerically stable) Conjugate Gradients } (s=2+\nu) \end{cases} \tag{102}$$

In the experiments, we choose λ_{low} either as the minimum eigenvalue (in the artificial power-law problems) or as a value at which we experimentally observe the breakdown of the spectral power-law (for MNIST).

Realistic quadratic problems. As an example of a realistic quadratic problem we take a subset of MNIST (of size $M = 30000$) and consider the scalar regression problem with targets given by the numerical values of corresponding digits $y \in \{0, 1, \ldots, 9\}$. The matrix $\widetilde{A}$ is the NTK of an infinitely wide, single-hidden layer network. The results are depicted in Figure 6 (right). Again, we observe power-law dependencies up to the estimated thresholds. The numerical entries in the legend have the form $\xi_{\text{exp}}(\xi_{\text{theor}})$, where ξ_{exp} is the "experimental" exponent estimated directly from the loss trajectory, and ξ_{theor} is the respective "theoretical" exponent given by $\zeta, 2\zeta$, or $(2+\nu)\zeta$. Here the values $\nu \approx 1.37$ and $\zeta \approx 0.25$ are in turn estimated from the empirically found ρ (Figure 1 (right)) and the eigenvalues λ_k and partial sums of target expansion coefficients (Figure 1 (center)). We see a reasonable agreement between ξ_{exp} and ξ_{theor}. Like with synthetic data, the asymptotic scheduled HB performs similarly to its full counterpart.

Neural networks. We consider a shallow fully-connected ReLU network with 1000 hidden neurons and train it on the full MNIST with MSE loss calculated on one-hot encoded classes. Note that this is no longer a quadratic problem. We restrict ourselves to optimization algorithms with predefined schedules due to their computational efficiency compared to adaptive algorithms (in which the 1D nonlinear problem of step optimization has to be solved in each iteration). Also, we use full-batch gradient descent in accordance with the rest of the paper. The results are shown in Figure 7.

For all algorithms except scheduled GD we observe behavior similar to the quadratic case, and in particular asymptotic HB is very close to its full counterpart. The relation between the fitted exponents ξ holds true: they are twice as large for scheduled methods as for constant learning rate methods.

The unstable behavior of scheduled GD is explained by large step-sizes α_n present in the schedule (at steps $n \approx 2^l$). When the problem is quadratic, large α_n are compensated by smaller ones chosen at other steps n, but non-quadratic perturbations break this compensation mechanism.

CG polynomials. In Figure 8 we plot CG polynomials $p_n(\lambda)$ for the basic and the numerically stable algorithms, calculated either at the spectral points λ_k, or also between them. At step $n = 10$ the two polynomials mostly coincide except for big λ. At step $n = 300$ the polynomials are different, and for either of them we observe two λ–regions with a sharp transition point $\widetilde{\lambda}$. For $\lambda > \widetilde{\lambda}$, the values of $p_n(\lambda)$ vanish at the spectral points but are extremely large in between, meaning that the roots of $p_n(\lambda)$ are located exactly at the spectral points λ_k. The rest of the roots are located at $\lambda < \widetilde{\lambda}$ and seem to optimize the overall envelope of $p_n(\lambda)$ instead of only root positions. This agrees with construction used in upper bound (47). As, due to numerical errors, the polynomial of basic CG places its roots in the region $\lambda \geq \widetilde{\lambda}$ with lower precision, the value of $\widetilde{\lambda}$ is higher in this case and hence convergence on $[0, \widetilde{\lambda}]$ is worse.

7. Conclusion

We have considered a wide range of first-order optimization methods including Gradient Descent, Steepest Descent, Heavy Ball, and Conjugate Gradients, with constant, non-constant predefined, and adaptive learning rates. Under power-law spectral assumptions with target exponent ζ and eigenvalue exponent ν the convergence rates of these methods are given by $O(n^{-\xi})$, where $\xi = \zeta, 2\zeta$ or $(2+\nu)\zeta$, depending on the method. The basic rate with $\xi = \zeta$ applies to Gradient Descent with constant learning rates and also to Steepest Descent. To reliably achieve the first accelerated rate 2ζ with Heavy Ball, a specific Jacobi-based schedule of learning rate and momenta is required, with β_n approaching 1 so that $1 - \beta_n \propto n^{-1}$. Finally, the fastest rate $(2+\nu)\zeta$ is achieved by Conjugate Gradients – the only method out of those we have considered that can take advantage of the discreteness of the problem spectrum by exactly fitting the target function in certain eigenspaces.

We prove that all our upper bounds are tight. For each upper bound we provide an example problem whose convergence rate matches that of the upper bound, and in some cases also has a very close coefficient. An important aspect of our approach is a power-law spectral assumption that is somewhat different from the classical source condition. We show, both experimentally and theoretically, that our condition much better describes problems whose actual loss trajectory is well approximated by a power-law. Specifically, for a problem with power-law loss asymptotic $L(\mathbf{w}_n) \sim n^{-\xi}$ our condition provides the matching bound $L(\mathbf{w}_n) \leq const \cdot n^{-\xi}$ while the best usage of the classical source condition can only provide a bound with additional logarithmic factor, $L(\mathbf{w}_n) \leq const \cdot n^{-\xi} \log n$.

Our theoretical results are confirmed by experiments with both simulated and real problems, including classifying MNIST by a neural network (which is only an approximately quadratic problem). In all experiments we observe a clear power law dependence of the loss on the optimization step n for steps that are neither too large nor too small, i.e. whenever both the infinite-dimensional approximation and asymptotic formulas are applicable. The respective exponents and their mutual relations agree well with theoretical predictions (un-

less the method is affected strongly by noise, as with CG, or by non-quadratic corrections, as with the optimally scheduled GD applied to a neural network).

Finally, let us outline a few natural topics for future research. First, as discussed in Section 4.3, Heavy Ball with various Jacobi-based schedules with the asymptotic form $\alpha_n \sim const,\ 1-\beta_n \propto n^{-1}$ can ensure the same $O(-2\zeta)$ convergence rate. We hypothesize that under the general spectral condition $\rho([0,\lambda]) \leq G(\lambda)$, the asymptotic of $G(\lambda)$ at small eigenvalues $\lambda \to 0$ can be translated into a certain asymptotic of $1-\beta_n$ at large iterations n for optimal HB. Second, it would be interesting to investigate whether weak non-quadratic perturbations of quadratic problems allow to retain the accelerated rate $O(n^{-2\zeta})$. Our experiments with a neural network on MNIST confirm this possibility. Third, it would be interesting to include stochasticity into consideration, as mini-batch stochastic gradient descent is a necessary requirement for any GD method to be used in modern deep learning applications.

Acknowledgments

We acknowledge support from the Russian Ministry of Science and Higher Education, grant No. 075-10-2021-068.

Appendix A. Related work

Optimization by GD, SD, HB and CG under power law spectral assumptions. The first study of GD, HB and CG under power-law spectral assumptions (in a form somewhat different from ours; see discussion at the end of Section 2.1) was performed in Nemirovskiy and Polyak (1984a) (upper bounds) and Nemirovskiy and Polyak (1984b) (lower bounds). These two works proved or conjectured some of the bounds appearing in our Table 1. While these two papers only considered scheduled HB based on Chebyshev polynomials, Brakhage (1987) generalized it to a "ν-method" based on general Jacobi polynomials, which allowed him to obtain the tight $O(n^{-2\nu})$ upper bound analogous to our Corollary 4.8 for HB with predefined schedules. SD was analyzed in Gilyazov and Gol'dman (2013) who proved a $O(n^{-\zeta})$ upper bound (their Theorem 2.2.5). However, the proof of its tightness (supplemented in our Theorem 4.16) does not seem to have been known prior to our work. Various aspects of optimization by HB and CG were discussed in Hanke (1991) and Hanke (1996). In particular, the latter paper gave a proof of the lower bound for CG in the special case of exponents $\nu = 1, 2$. All of these works relied on the classical source condition and only considered problems with attainable solutions.

The recent work Berthier et al. (2020a), although focusing on a specific application domain of gossip problem, uses a spectral condition (see their Proposition 5.5 or Definition I.2) which is different from the classical source condition and much closer to our condition, and also considers a Jacobi-based optimization algorithm. However, both Berthier et al. (2020a) and earlier works Brakhage (1987); Hanke (1991) rely on classical asymptotic properties of Jacobi polynomials for the proofs of upper bounds, e.g. Theorem 7.32.2 of Szego (1939). This approach quickly provides the desired $O(n^{-2\zeta})$ rate but does not specify the constant. In contrast, our flattened polynomial construction of Theorem 4.1 followed by accurate estimations in Theorem 4.6 and Proposition 4.7 lead to an explicit and tight constant in the convergence bound (e.g. overestimation by at most a factor of $C_\zeta = 4$ for $\zeta = 1$).

SGD. Analogs of our power law spectral conditions (8) and (10) are well-known in literature on kernel methods, regularized regression and SGD (Caponnetto and De Vito, 2007; Steinwart et al., 2009; Varre et al., 2021). Convergence of SGD under these or similar conditions has been studied in Berthier et al. (2020b); Zou et al. (2021); Varre et al. (2021); Velikanov et al. (2022). SGD subsumes GD as a special case of noiseless gradient evaluation, but is in a sense more complex than all the algorithms we discuss in this paper because even for linear models the loss evolution under SGD is not generally expressible in terms of only spectral data. The most common version of SGD is mini-batch SGD in which the stochasticity is due to random sampling of the underlying data. In contrast to GD, SD and HB (cf. Table 1), convergence rates of SGD do depend directly, in general, on the eigenvalue decay exponent ν. In particular, for mini-batch SGD with constant learning rates the respective exponent equals $\min(\zeta, 2 - 1/\nu)$; moreover, optimization diverges if $\nu < 1$.

Kernel methods and NTK. Power law eigenvalue decay bounds are known to generally hold for integral operators with kernels satisfying suitable regularity assumptions (Widom, 1963; Kühn, 1987; Ritter et al., 1995; Ferreira and Menegatto, 2009; Birman and Solomjak, 1970; Williams and Rasmussen, 2006).

In the NTK regime of training wide neural networks the network model essentially becomes a kernel model (Neal, 2012; Jacot et al., 2018) with explicit kernels (Cho and Saul, 2009; Lee et al., 2019). Several recent studies empirically verify and exploit power law assumptions for the NTK spectrum (Bahri et al., 2021; Canatar et al., 2021; Lee et al., 2020; Nitanda and Suzuki, 2021; Jin et al., 2021). Specific powers of eigenvalue decay and eigenfunction expansion coefficients for ReLU networks and some classes of target functions are derived in Velikanov and Yarotsky (2021).

Steepest Descent. See Kantorovich and Akilov (1964) for a general introduction to Steepest Descent. In a general non-strongly convex case, convergence of the iterates to a solution (if it exists) was proved in Fridman (1962). In Kammerer and Nashed (1971) an explicit $\|\mathbf{w}_n - \mathbf{w}_*\|^2 = O(n^{-1})$ bound was proved in the non-strongly convex case under assumption $\|\widetilde{A}^{-1}\mathbf{f}_*\| < \infty$. The $O(n^{-\zeta})$ convergence upper bound under a power-law spectral condition was proved in Gilyazov and Gol'dman (2013) using moment inequalities from Krasnoselskii et al. (1972). Our approach in Section is 4.6 is rather different from these works and relies on the observation that SD converges to a period-2 oscillatory regime. This effect was established by Akaike (1959) in the finite-dimensional setting and by Pronzato et al. (2001) in the infinite-dimensional setting. Compared to Gilyazov and Gol'dman (2013), our approach is applicable under our slightly weaker spectral assumption (9) and additionally proves the tightness of the loss upper bound.

Heavy Ball. Multi-step methods have long been used in numerical linear algebra. As a method of optimization for general (non-quadratic) problems, Heavy Ball was proposed in Polyak (1964). HB can be interpreted as a simplest method with the momentum term (Qian, 1999). Flammarion and Bach (2015) introduced a general family of methods that includes HB with $\beta_n = 1 - 2/n$ as well as averaged GD (Polyak and Juditsky, 1992). Some variants of GD with momentum are optimal with respect to averaged case optimization scenarios (Pedregosa and Scieur, 2020; Lacotte and Pilanci, 2020).

Conjugate Gradients. Method of Conjugate Gradients was proposed in Hestenes and Stiefel (1952) and extensively studied afterwards (Daniel, 1971; Hestenes, 2012). The extension of the method to non-quadratic problems was first proposed in Fletcher and Reeves (1964). Stability of CG is a complex issue that has also been analyzed extensively (Hestenes and Stiefel, 1952; Björck et al., 1998; Meurant and Strakoš, 2006; Fischer, 2011). A $\|\mathbf{w}_n - \mathbf{w}_*\|^2 = O(n^{-1})$ convergence bound for CG in a gapless infinite-dimensional setting was proved in Kammerer and Nashed (1972). A version of the bound $L(\mathbf{w}_n) = O(n^{-2\zeta})$ was proved in Nemirovskiy and Polyak (1984a), and in the same paper it was observed that this rate can be improved if the spectrum is discrete. Hanke (1991, 1996) gave a version of the $L(\mathbf{w}_n) = O(n^{-(2+\nu)\zeta})$ bound and proved its tightness in the cases $\nu = 1, 2$, for which a classical system of orthogonal polynomials is available. Our general proof of the tightness of the $O(n^{-(2+\nu)\zeta})$ bound for CG under the power law eigenvalue decay assumption (Section 4.5) is inspired by Theorem 2.1.7 in Nesterov (2003) which proves the tightness of the bound $L(\mathbf{w}_n) = O(n^{-2})$ in a setting of finite norm solution $\|\mathbf{w}_*\| < \infty$. However, the proof of our bound is significantly more difficult.

Nesterov Accelerated Gradient (NAG). NAG (Nesterov, 1983) is a modification of Heavy Ball (19) in which the gradient is computed after applying the momentum term

rather than before:

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \beta_n(\mathbf{w}_n - \mathbf{w}_{n-1}) - \alpha_n \nabla L(\mathbf{w}_n + \beta_n(\mathbf{w}_n - \mathbf{w}_{n-1})). \tag{103}$$

For quadratic problems, the analog of Eq. (20) then reads

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \beta_n(\mathbf{w}_n - \mathbf{w}_{n-1}) - \alpha_n[(A\mathbf{w}_n - \mathbf{b}) + \beta_n A(\mathbf{w}_n - \mathbf{w}_{n-1})]. \tag{104}$$

NAG is a practically widely used method and it is known to provide improved upper bounds for general convex problems (Nesterov, 1983). However, it does not seem to improve on Heavy Ball in the purely quadratic case considered in the present paper, at least in terms of the optimal convergence exponent. Specifically, assuming that the coefficients α_n, β_n are non-adaptive (predefined), both NAG and Heavy Ball are subject to our Theorem 4.11 showing that they cannot generally have a rate $L(\mathbf{w}_n) = O(n^{-\xi})$ with $\xi > 2\zeta$, while the rate $L(\mathbf{w}_n) = O(n^{-2\zeta})$ is attained by Heavy Ball by Corollary 4.8.

Appendix B. Background on polynomials for optimization

The polynomial representation of optimization updates. The optimization algorithms of Section 2.2 and their properties can be conveniently expressed in terms of polynomials of the operator A (or $\widetilde{A}$). Suppose first for simplicity that our optimization problem $L(\mathbf{w}) = \frac{1}{2}\langle \mathbf{w}, A\mathbf{w}\rangle - \langle \mathbf{w}, \mathbf{b}\rangle + \frac{1}{2}\|\mathbf{f}_*\|^2 \to \min_{\mathbf{w}}$ has a finite-norm optimizer $\mathbf{w}_*$ such that $A\mathbf{w}_* = \mathbf{b}$. Consider the deviations $\delta\mathbf{w} = \mathbf{w} - \mathbf{w}_*$ of the points $\mathbf{w}$ from the solution $\mathbf{w}_*$. For the basic GD or SD, we have

$$\delta\mathbf{w}_n = \mathbf{w}_{n-1} - \alpha_{n-1}(A\mathbf{w}_{n-1} - \mathbf{b}) - \mathbf{w}_* \tag{105}$$

$$= (1 - \alpha_{n-1}A)\delta\mathbf{w}_{n-1}, \tag{106}$$

and so, by iterating,

$$\delta\mathbf{w}_n = p_n(A)\delta\mathbf{w}_0, \tag{107}$$

where p_n is the degree-n polynomial

$$p_n(\lambda) = \prod_{s=1}^{n}(1 - \alpha_{s-1}\lambda). \tag{108}$$

The respective loss is

$$L(\mathbf{w}_n) = \frac{1}{2}\langle A\delta\mathbf{w}_n, \delta\mathbf{w}_n\rangle \tag{109}$$

$$= \frac{1}{2}\int \lambda p_n^2(\lambda)\rho_{A,\mathbf{w}_*}(d\lambda) \tag{110}$$

$$= \frac{1}{2}\int p_n^2(\lambda)\rho_{\widetilde{A},\mathbf{f}_*}(d\lambda), \tag{111}$$

where $\rho_{A,\mathbf{w}_*}$ and $\rho_{\widetilde{A},\mathbf{f}_*}$ are the spectral measures associated (as in Eq. (7)) with $A, \mathbf{w}_*$ and $\widetilde{A}, \mathbf{f}_*$, respectively.

Representation (111) (with $\rho_{\widetilde{A},\mathbf{f}_*}$) can alternatively be reached without assuming the existence of the solution $\mathbf{w}_*$, by considering the deviations $\delta\mathbf{f} = \mathbf{f} - \mathbf{f}_*$ in the target space and similarly observing that

$$\delta\mathbf{f}_n = p_n(A)\delta\mathbf{f}_0, \tag{112}$$

with the same polynomial p_n.

In the case of HB and CG, the iterations have the more general form

$$\delta\mathbf{w}_{n+1} = (1 - \alpha_n A)\delta\mathbf{w}_n + \beta_n(\delta\mathbf{w}_n - \delta\mathbf{w}_{n-1}). \tag{113}$$

This again yields the polynomial representation $\delta\mathbf{w}_n = p_n(A)\delta\mathbf{w}_0$, but with a degree-$n$ polynomial p_n depending on $\{\alpha_s, \beta_s\}_{s=0}^{n-1}$ in a more complicated way:

$$p_0 = 1, \tag{114}$$
$$p_1 = 1 - \alpha_0\lambda, \tag{115}$$
$$p_{n+1} = (1 - \alpha_n\lambda)p_n + \beta_n(p_n - p_{n-1}). \tag{116}$$

Note that p_n is necessarily a *residual* polynomial, in the sense that $p_n(0) = 1$.

As mentioned in Section 2.2, CG has the important property of being optimal among all first order methods generating new iterates $\mathbf{w}_{n+1}$ by shifting the initial point $\mathbf{w}_0$ along linear subspaces spanned by the previously computed gradients $\nabla L(\mathbf{w}_0), \ldots, \nabla L(\mathbf{w}_n)$. In terms of the respective residual polynomials p_n, this means that they minimize the loss functional over all residual polynomials of given degree:

$$p_n = \operatorname*{arg\,min}_{q_n:\deg q_n = n, q_n(0)=1} \frac{1}{2}\int q_n^2(\lambda)\rho_{\widetilde{A},\mathbf{f}_*}(d\lambda). \tag{117}$$

See the book Fischer (2011) for more details on the polynomial representation of optimization methods.

Jacobi polynomials. As shown in Section 4.3, Jacobi polynomials $P_n^{(a,b)}$ arise as an optimal choice for power-law spectral measure. We heavily use these polynomials in many of our results.

The appearance of Jacobi polynomials in our setting is related to their orthogonality w.r.t. power-law weight function:

$$\int_{-1}^{1}(1-x)^a(1+x)^b P_n^{(a,b)}(x)P_m^{(a,b)}(x)dx = C_n\delta_{nm}. \tag{118}$$

Here δ_{nm} is Kronecker delta function and C_n are the constants depending on normalization of the polynomials. We adopt the standard normalization of Jacobi polynomials by their value at $x = 1$:

$$P_n^{(a,b)}(1) = \binom{n+a}{n}. \tag{119}$$

Jacobi polynomials, like any system of orthogonal polynomials, enjoy three-term recurrence relations. Specifically,

$$\begin{aligned}2(n+1)(n+a+b+1)(2n+a+b)P_{n+1}^{(a,b)}(x) = \\ (2n+a+b)(2n+a+b+1)(2n+a+b+2)xP_n^{(a,b)}(x) \\ +(2n+a+b+1)(a^2-b^2)P_n^{(a,b)}(x) \\ -2(n+a)(n+b)(2n+a+b+2)P_{n-1}^{(a,b)}(x).\end{aligned} \tag{120}$$

Appendix C. Main spectral condition

In this section, we collect the proofs of the results concerning either general properties of our spectral condition (9) or its relation to the classical source condition (12).

C.1 Basic properties

Proof of Lemma 2.1.
Inclusion $\mathrm{P}(\zeta,Q)\subseteq\mathrm{P}'(\zeta',Q')$ (Eq. (13)). To test this inclusion for a certain pair of ζ,Q and ζ',Q', we need to check

$$\sup_{\rho\in\mathrm{P}(\zeta,Q)}\int_0^1\lambda^{-\zeta'}\rho(d\lambda)\le Q'. \tag{121}$$

First, consider $\zeta'\ge\zeta$ and the exact power-law measure $\rho(d\lambda)=Qd(\lambda^\zeta)\in\mathrm{P}(\zeta,Q)$. Then, the integral in (121) diverges as

$$\lim_{\varepsilon\to0}\int_\varepsilon^1\lambda^{-\zeta'}Qd(\lambda^\zeta)=\begin{cases}\lim\limits_{\varepsilon\to0}\frac{Q\zeta}{\zeta'-\zeta}\left(\varepsilon^{\zeta-\zeta'}-1\right)=\infty, & \zeta'>\zeta\\ \lim\limits_{\varepsilon\to0}Q\zeta\log(\varepsilon^{-1})=\infty, & \zeta'=\zeta\end{cases} \tag{122}$$

which makes $\zeta'<\zeta$ a necessary condition for inclusion. Assuming this condition, the supremum in (121) can be evaluated using integration by parts:

$$\int_0^1\lambda^{-\zeta'}\rho(d\lambda)=\lambda^{-\zeta'}\rho([0,\lambda])\Big|_0^1+\zeta'\int_0^1\lambda^{-\zeta'-1}\rho([0,\lambda])d\lambda. \tag{123}$$

Note that both terms in (123) are well defined thanks to the constraint $\rho([0,\lambda])\le Q\lambda^\zeta$. Importantly, the right-hand side of (123) is a pointwise positive linear functional of the cumulative distribution function $\rho([0,\lambda])$, which implies that the supremum in (121) is reached at the exact power-law measure $\rho(d\lambda)=Qd(\lambda^\zeta)$, and its value is

$$\sup_{\rho\in\mathrm{P}(\zeta,Q)}\int_0^1\lambda^{-\zeta'}\rho(d\lambda)=Q+Q\frac{\zeta'}{\zeta-\zeta'}=Q\frac{\zeta}{\zeta-\zeta'}. \tag{124}$$

This computation implies that $Q'\ge Q\frac{\zeta}{\zeta-\zeta'}$ is equivalent to the desired inclusion for $\zeta'<\zeta$, which completes the proof of (13).

Inclusion $\mathrm{P}'(\zeta', Q') \subseteq \mathrm{P}(\zeta, Q)$ (Eq. (14)). First note that this inclusion cannot hold if $\zeta' < \zeta$. Indeed, in that case the equivalence (13) would imply $\mathrm{P}(\widetilde{\zeta}, \widetilde{Q}) \subseteq \mathrm{P}(\zeta, Q)$ for any $\widetilde{\zeta} \in (\zeta', \zeta)$ and some $\widetilde{Q}$, which contradicts $\rho(d\lambda) = \widetilde{Q} d(\lambda^{\widetilde{\zeta}}) \notin \mathrm{P}(\zeta, Q)$.

For $\zeta' \geq \zeta$ the inclusion can be tested with

$$\sup_{\rho \in \mathrm{P}'(\zeta', Q')} \left[\sup_{\lambda \in (0,1]} \rho([0,\lambda])/\lambda^{\zeta} \right] \leq Q, \tag{125}$$

where we used that $\rho(\{0\}) = 0$ in our setting (see section 2) to account for $\lambda = 0$ case of (9). Note that the expression $\rho([0,\lambda])/\lambda^{\zeta}$ is bounded for $\rho \in \mathrm{P}'(\zeta', Q'),\ \lambda \in (0,1]$ as

$$\lambda^{-\zeta} \rho([0,\lambda]) \leq \lambda^{\zeta'-\zeta} \int_0^{\lambda} \lambda_1^{-\zeta'} \rho(d\lambda_1) \leq Q' \tag{126}$$

Actually, this bound is tight, as can be shown by taking $\rho = Q'\delta_1 \in \mathrm{P}'(\zeta', Q')$ and $\lambda = 1$. This makes the value of the supremum in (125) equal to Q', thus establishing equivalence (14). This completes the proof of Lemma 2.1.

Attainability. Let ρ be the spectral measure supported on $[0,1]$ and satisfying our main spectral condition $\rho((0,\lambda]) \leq Q\lambda^{\zeta}$ with some $Q, \zeta > 0$. Recall that the attainability condition reads $\|\mathbf{w}_*\|^2 = \|J^{-1}\mathbf{f}_*\|^2 = \int_0^1 \lambda^{-1}\rho(d\lambda) < \infty$. If $\zeta \leq 1$, then, in general, the solution is not attainable, as can be seen by considering the exact power law $\rho((0,\lambda]) = \lambda^{\zeta}$. On the other hand, if $\zeta > 1$ then, by Lemma 2.1, $\mathrm{P}(\zeta, Q) \subseteq \mathrm{P}'(1, Q\frac{\zeta}{\zeta-\zeta'})$, implying that the solution is attainable.

Scaling properties. An important property of our quadratic optimization problem is its transformation under rescaling of the input data by $J \mapsto cJ$ or by $\mathbf{f}_* \mapsto c\mathbf{f}_*$. Under these rescalings, all the optimization algorithms of Section 2.2 and the spectral conditions (8) and (10) retain their structure, but the quantities appearing in their description get rescaled by $u \mapsto c^a u$ with various scaling exponents a. In Table 2 we list these scaling exponents.

As an application of this observation, if we have a result for a special case when two scalar parameters are fixed, we can derive the corresponding general result by rescaling $J \mapsto cJ$ and $\mathbf{f}_* \mapsto c'\mathbf{f}_*$ with suitable c and c'. In particular, suppose that we have a bound for $L(\mathbf{w}_n)$ when $\lambda_{\max} = 1$ and $Q = 1$. Then the corresponding bound for general $\lambda_{\max}$ and Q can be obtained by taking $c = \lambda_{\max}^{1/2}$ and $c' = Q^{1/2}\lambda_{\max}^{\zeta/2}$: we see that the loss will be rescaled by $L(\mathbf{w}_n) \mapsto (c')^2 L(\mathbf{w}_n) = Q\lambda_{\max}^{\zeta} L(\mathbf{w}_n)$.

C.2 Proof of Theorem 4.1

First, lets us examine the structure of the function $\overline{q}(x) = \sup_{y \geq x} q(y)$ introduced in Section 4.1. Since $q(x)$ is a polynomial, it has a finite number of local maxima on $[0,1]$, from which we choose a maximal length sequence $0 \leq x_1 < x_2 < \ldots < x_m \leq 1$ such that the values at subsequent local maxima are decreasing: $q(x_i) > q(x_{i+1})$. Then, picking m points y_i such that $y_i,\ i = 1 \ldots m-1$, is the leftmost point in (x_i, x_{i+1}) satisfying $q(y_i) = q(x_{i+1})$ and $y_m = 1$, allows to characterize $\overline{q}(x)$ as

$$\overline{q}(x) = \begin{cases} q(x_1), & 0 \leq x < x_1 \\ q(x), & x_i \leq x \leq y_i, \quad i = 1 \ldots m \\ q(x_{i+1}), & y_i < x < x_{i+1}, \quad i = 1 \ldots m-1 \end{cases} \tag{127}$$

Table 2: Scaling exponents a in the transformations $u \mapsto c^a u$ of various quantities u appearing in the descriptions of optimization algorithms (Section 2.2) and spectral conditions (8) and (10) under the transformations $J \mapsto cJ$ and $\mathbf{f}_* \mapsto c\mathbf{f}_*$.

	J	$\mathbf{f}_*$	A	$\mathbf{b}$	$\mathbf{w}_n$	α_n	β_n	$L(\mathbf{w}_n)$	Q	Λ	$\lambda_{\max}$	ζ	ν
$J \mapsto cJ$	1	0	2	1	-1	-2	0	0	-2ζ	2	2	0	0
$\mathbf{f}_* \mapsto c\mathbf{f}_*$	0	1	0	1	1	0	0	2	2	0	0	0	0

The representation (127) can be verified by direct comparison with the definition $\overline{q}(x) = \sup_{y \geq x} q(y)$ in each of the three cases.

Now, assume that the original polynomial q is upper bounded, $q(x) \leq g(x)$, by some absolutely continuous and non-increasing $g(x)$. Then, integrating by parts, the respective "loss" integral can be upper-bounded as

$$\begin{aligned}\int_0^1 q(x)\rho(dx) &\leq \int_0^1 g(x)\rho(dx) = \int_0^1 g(x)d\rho([0,x]) \\ &= g(1)\rho([0,1]) + \int_0^1 (-g'(x))\rho([0,x])dx \\ &\overset{(1)}{\leq} g(1)G(1) + \int_0^1 (-g'(x))G(x)dx = \int_0^1 g(x)G'(x)dx\end{aligned} \tag{128}$$

where in (1) we used that $-g'(x) \geq 0$ due to $g(x)$ being non-decreasing, and that $g(1) \geq 0$ since the polynomial q is by assumption nonnegative. Note that $\overline{q}(x)$ given by (127) is absolutely continuous and non-decreasing. Thus, the bound (128) applies with $g(x) = \overline{q}(x)$ which sets the r.h.s of (26) as an upper bound for the loss integral.

Next, we show that the obtained upper bond is reached with a specific spectral measure

$$\rho^* = G(x_1)\delta_{x_1} + \sum_{i=1}^{m-1} \big(G(x_{i+1}) - G(x_i)\big)\delta_{x_{i+1}} + \sum_{i=1}^{m} \rho_i^*, \quad \rho_i^*(dx) = \mathbb{1}_{[x_i,y_i]}G'(x)dx \tag{129}$$

which is a mix of Dirac delta measures x_i and "smooth" measures with density ρ_i^*, supported on $[x_i, y_i]$. Note that ρ^* satisfies the required condition $\rho^*([0,x]) \leq G(x)$. Direct substitution of ρ^* into the loss integral gives

$$\begin{aligned}\int_0^1 q(x)\rho^*(dx) &= q(x_1)G(x_1) + \sum_{i=1}^{m-1} q(x_{i+1}) \int_{y_i}^{x_{i+1}} G'(x)dx + \sum_{i=1}^{m} \int_{x_i}^{y_i} q(x)G'(x)dx \\ &= \int_0^1 \overline{q}(x)G'(x)dx\end{aligned} \tag{130}$$

C.3 Proof of Theorem 5.1

Our proof consists of three steps. In Step 1 we will show that a power-law asymptotic of the loss implies a power-law asymptotic of the spectral measure. Then, in Step 2 we derive the asymptotic of the bound $\widetilde{L}'_n$, and in Step 3 the asymptotic of the bound $\widetilde{L}_n$.

Step 1. We will use the following general lemma.

Lemma C.1. *Suppose that ρ is a Borel measure on the segment $[0,1]$, and $a > 0$ is a constant. Assume that $\int_0^1 (1-a\lambda)^{2n}\rho(d\lambda) = n^{-\xi}(1+o(1))$ as $n \to \infty$, with some constant $\xi > 0$. Then $\rho([0,\lambda]) = (\Gamma(\xi+1))^{-1}(2a\lambda)^{\xi}(1+o(1))$ as $\lambda \searrow 0$.*

Proof This lemma can be derived from the general theory of abelian–tauberian power-law relations (Feller (1991), Section XIII.5), but we find it simpler to just give a direct proof mimicking original Karamata's arguments (Karamata, 1930).

We argue that, under the hypotheses of the lemma, for all sufficiently regular functions $g : [0,1] \to \mathbb{R}$ holds

$$\lim_{n\to\infty} n^{\xi} \int_0^1 (1-a\lambda)^{2n} g((1-a\lambda)^{2n})\rho(d\lambda) = I(g), \tag{131}$$

where

$$I(g) = (\Gamma(\xi+1))^{-1} \int_0^{\infty} e^{-y} g(e^{-y}) dy^{\xi}. \tag{132}$$

Indeed, for monomials $g(x) = x^k$ both sides of Eq. (131) equal $(k+1)^{-\xi}$. By linearity, Eq. (131) then holds for all polynomials.

Now observe that the integral on the l.h.s. of Eq. (131) is monotone in g – in the sense that if $g_1(x) \le g_2(x)$ for all $x \in [0,1]$, then the same inequality holds for the respective integrals.

Suppose next that a function g is such that for any $\epsilon > 0$ one can find polynomials $g_{\pm}$ for which $g_-(x) \le g(x) \le g_+(x)$ on $[0,1]$ and $I(g_+) - I(g_-) < \epsilon$. Then, using the above mentioned monotonicity, Eq. (131) holds for the function g, too.

Clearly, this condition holds for the function

$$g(x) = \begin{cases} 1/x, & x \in [e^{-1}, 1], \\ 0, & \text{otherwise.} \end{cases} \tag{133}$$

Substituting in Eq. (131), we find

$$\lim_{n\to\infty} n^{\xi}\rho([0, a^{-1}(1-e^{-1/(2n)})]) = (\Gamma(\xi+1))^{-1}, \tag{134}$$

implying the claim of the lemma. ■

Recalling that the loss of Gradient Decent with constant learning rate α is given by $L_n = \frac{1}{2}\int_0^1 (1-\alpha\lambda)^{2n}\rho(d\lambda)$, the asymptotic $L_n = Cn^{-\xi}(1+o(1))$ and lemma C.1 imply

$$\rho([0,\lambda]) = Q_\rho \lambda^{\xi}(1+o(1)), \quad Q_\rho = 2C\tfrac{(2\alpha)^{\xi}}{\Gamma(\xi+1)}. \tag{135}$$

Step 2. We use spectral asymptotic (135) derived above to calculate the optimal upper bound $\widetilde{L}'_n$ as defined in Eq. (97):

$$\widetilde{L}'_n(\rho) = \inf_{\zeta',Q':\rho\in \mathrm{P}'(\zeta',Q')} \sup_{\widetilde{\rho}\in \mathrm{P}'(\zeta',Q')} L_n(\widetilde{\rho}). \tag{136}$$

Note that Eq. (95) already gives the supremum $L_n^{UB}(\zeta', Q') = \sup_{\widetilde{\rho}\in \mathrm{P}'(\zeta', Q')} L_n(\widetilde{\rho})$, and we only need to optimize it over Q' and $\zeta' < \xi$. At a given ζ', the minimal possible Q' is simply $Q'(\zeta') = \int_0^1 \lambda^{-\zeta'}\rho(d\lambda)$, so optimization reduces to that over ζ' with this $Q'(\zeta')$. Expecting the need to take $\zeta' \nearrow \xi$ at large n, we denote $\varepsilon = \xi - \zeta'$ and calculate

$$\begin{aligned}\lim_{\varepsilon\searrow 0} \varepsilon Q'(\zeta') &= \lim_{\varepsilon\searrow 0} \varepsilon \left[\rho([0,1]) + \zeta' \int_0^1 \lambda^{-\zeta'-1}\rho([0,\lambda])d\lambda\right] \\ &= Q_\rho \lim_{\varepsilon\searrow 0} \varepsilon\zeta' \int_0^1 \lambda^{\varepsilon-1}(1+o(1))d\lambda \\ &= Q_\rho \lim_{\varepsilon\searrow 0} \zeta' \int_0^1 (1+o(1))d\lambda^{\varepsilon} \\ &= Q_\rho \xi,\end{aligned} \tag{137}$$

where in the first line we integrated by parts and in the second used Eq. (135). It follows that $Q'(\zeta')$ asymptotically behaves as

$$Q'(\zeta') = Q_\rho \frac{\xi}{\xi-\zeta'}(1+o(1)), \quad \zeta' \nearrow \xi. \tag{138}$$

Recalling the form of the upper bound (95), we calculate $\widetilde{L}_n'$ as

$$\begin{aligned}\widetilde{L}_n' &= \inf_{0<\zeta'<\xi} L_n^{UB}(\zeta', Q'(\zeta')) \\ &= \inf_{0<\zeta'<\xi} \frac{Q'(\zeta')}{2}\left(\frac{\zeta'}{2\alpha e}\right)^{\zeta'} n^{-\zeta'}(1+o_n(1)) \\ &= \inf_{0<\zeta'<\xi} \frac{Q_\rho\xi}{2}\left(\frac{\zeta'}{2\alpha e}\right)^{\zeta'} \frac{n^{-\zeta'}}{\xi-\zeta'}(1+o_n(1))(1+o_{\zeta'}(1)) \\ &\overset{\zeta'=\xi-\varepsilon}{=\!=} \frac{Q_\rho\xi}{2}\left(\frac{\xi}{2\alpha e}\right)^{\xi} n^{-\xi}(1+o_n(1)) \inf_{0<\varepsilon<\xi} \frac{n^\varepsilon}{\varepsilon}(1+o_\varepsilon(1)).\end{aligned} \tag{139}$$

Here we added subscripts to distinguish different $o(1)$ corrections, and used that the $o_n(1)$ correction from (95) is in fact uniform for $\zeta' \in [0, c_1]$ with any finite c_1.

Recall the optimal bound $\widetilde{L}_n'(\rho_\xi)$ for exact power-law measure given in (100). Substitution of the infimum $\inf_{0<\varepsilon<\xi} \frac{n^\varepsilon}{\varepsilon} = e\log n,\ n > e^{\frac{1}{\xi}}$ and the expression for Q_ρ into (100) gives the desired statement (99) of the theorem.

However, we still need to argue that this result is not affected by the factor $1+o_\varepsilon(1)$ appearing in $\inf_{0<\varepsilon<\xi} \frac{n^\varepsilon}{\varepsilon}(1+o_\varepsilon(1))$. To this end, it clearly suffices to show that the optimal $\varepsilon \to 0$ as $n\to\infty$. By tracing back our expression $1+o_\varepsilon(1)$ to formula (138), this expression is bounded away from 0 on the interval $[0,\xi]$. Then, on any interval $[c,\xi]$ with $c>0$ we get a power-law lower bound

$$\inf_{c\le\varepsilon<\xi} \frac{n^\varepsilon}{\varepsilon}(1+o_\varepsilon(1)) = \Omega(n^c), \quad n\to\infty. \tag{140}$$

This shows by comparison with the logarithmic expression $\inf_{0<\varepsilon<\xi} \frac{n^\varepsilon}{\varepsilon} = e\log n$ that the values ε bounded away from 0 are indeed asymptotically suboptimal. This completes the computation of $\widetilde{L}_n'$.

Step 3. Finally, we calculate the optimal bound $\widetilde{L}_n$ under our source condition (9), as defined in Eq. (96):

$$\widetilde{L}_n(\rho) = \inf_{\zeta,Q:\rho\in \mathrm{P}(\zeta,Q)} \sup_{\widetilde{\rho}\in \mathrm{P}(\zeta,Q)} L_n(\widetilde{\rho}). \tag{141}$$

First, recall that the inner supremum here is given by theorem 4.1, where for GD with $\alpha \leq 1$ the flattened polynomial $\overline{(1-\alpha\lambda)^{2n}} = (1-\alpha\lambda)^{2n}$. Thus, we have

$$\begin{aligned}\sup_{\widetilde{\rho}\in P(\zeta,Q)} L_n(\widetilde{\rho}) &= \frac{Q}{2}\int_0^1 (1-\alpha\lambda)^{2n} d(\lambda^\zeta) = \frac{Q}{2}\int_0^{\alpha^{-1}} (1-\alpha\lambda)^{2n} d(\lambda^\zeta) + O((1-\alpha)^{2n}) \\ &= \frac{Q}{2}\zeta\alpha^{-\zeta}\frac{\Gamma(2n+1)\Gamma(\zeta)}{\Gamma(2n+\zeta+1)} + O((1-\alpha)^{2n}) \overset{n\to\infty}{=\!=} \frac{Q}{2}\Gamma(\zeta+1)(2\alpha n)^{-\zeta}(1+o(1)),\end{aligned} \tag{142}$$

where we recognized the integral $\int_0^1 (1-z)^{2n} z^{\zeta-1} dz$ as a Beta function and substituted its expression in terms of Gamma functions.[4]

To optimize this expression over Q and ζ, note that we can take any $\zeta \leq \xi$, and at the given ζ the minimal constant Q is

$$Q(\zeta) = \sup_{\lambda\in(0,1]} \rho([0,\lambda])/\lambda^\zeta = \sup_{\lambda\in(0,1]} Q_\rho \lambda^{\xi-\zeta}(1+o(1)). \tag{143}$$

We note a couple of properties of $Q(\zeta)$:

1. $Q(\zeta) \nearrow Q(\xi)$ as $\zeta \nearrow \xi$, because for any $\lambda \in (0,1]$ the function $\zeta \mapsto \rho([0,\lambda])/\lambda^\zeta$ is monotone non-decreasing and converging to $\rho([0,\lambda])/\lambda^\xi$ as $\zeta \nearrow \xi$.

2. $Q(\zeta)$ is bounded away from 0 on the interval $0 \leq \zeta \leq \xi$, since $Q(\zeta) \geq \rho([0,1]) > 0$.

Property 2) and representation (143) imply that the infimum of $\sup_{\widetilde{\rho}\in P(\zeta,Q)} L_n(\widetilde{\rho})$ over ζ and $Q(\zeta)$ is attained at a ζ deviating from ξ by at most $O(1/\log n)$; in particular the optimal ζ converges to ξ as $n \to \infty$. But then, using property 1) we get the desired asymptotic (98):

$$\widetilde{L}_n(\rho) = \frac{Q(\xi)}{2}\Gamma(\xi+1)(2\alpha n)^{-\xi}(1+o(1)). \tag{144}$$

This completes the proof of the theorem.

Appendix D. Constant learning rates

D.1 Proof of Theorem 4.2: the case of GD ($\beta = 0$)

First, we express worst-case loss (28) through exact power-law loss (29).

4. Actually, the same computation is performed in the proof of the theorem 4.2, see eq. (147). We repeat it here simply for convenience.

If $\alpha \le 1$, the polynomial $p_n^2(\lambda) = (1-\alpha\lambda)^{2n}$ is monotone decreasing and therefore $\overline{p_n^2}(\lambda) = p_n^2(\lambda)$. This implies that $\overline{L_n^{(\zeta)}} = L_n^{(\zeta)}$. If $1 < \alpha < 2$, the flattened polynomials $\overline{p_n^2}(\lambda)$ differ from $p_n^2(\lambda)$ on a single flat region and are given by

$$\overline{p_n^2}(\lambda) = \begin{cases} p_n^2(\lambda), & \lambda < \frac{2-\alpha}{\alpha} \\ (\alpha-1)^{2n}, & \frac{2-\alpha}{\alpha} \le \lambda \le 1 \end{cases} \tag{145}$$

The associated worst-case loss is

$$\begin{aligned} \overline{L_n^{(\zeta)}} &= \frac{1}{2}\int_0^{\frac{2-\alpha}{\alpha}} p_n^2(\lambda) d(\lambda^\zeta) + \frac{1}{2}\int_{\frac{2-\alpha}{\alpha}}^1 (\alpha-1)^{2n} d(\lambda^\zeta) \\ &= \frac{1}{2}\int_0^{\frac{2-\alpha}{\alpha}} p_n^2(\lambda) d(\lambda^\zeta) + O\big((\alpha-1)^{2n}\big) \\ &= \frac{1}{2}\int_0^{1} p_n^2(\lambda) d(\lambda^\zeta) + O\big((\alpha-1)^{2n}\big), \end{aligned} \tag{146}$$

which is exactly the $\beta = 0$ part of (33) with $u = (1-\alpha)^2$. Finally, we calculate the loss under exact power-law measure as

$$\begin{aligned} L_n^{(\zeta)} &= \frac{1}{2}\int_0^1 (1-\alpha\lambda)^{2n} d(\lambda^\zeta) = \frac{1}{2}\zeta\int_0^{\frac{1}{\alpha}} (1-\alpha\lambda)^{2n}\lambda^{\zeta-1} d\lambda + O\big((\alpha-1)^{2n}\big) \\ &= \frac{1}{2}\zeta\alpha^{-\zeta}\frac{\Gamma(2n+1)\Gamma(\zeta)}{\Gamma(2n+1+\zeta)} + O\big((\alpha-1)^{2n}\big) \\ &= \frac{1}{2}\Gamma(\zeta+1)(2n\alpha)^{-\zeta}(1+o(1)) + O\big((\alpha-1)^{2n}\big) \end{aligned} \tag{147}$$

Here in the second line, we recognized the integral representation of the Beta function $B(a,b) = \int_0^1 (1-z)^{a-1} z^{b-1} dz$ and expressed it through the Gamma functions. In the last line, we used $x\Gamma(x) = \Gamma(x+1)$ and and asymptotic of Gamma function $\Gamma(x+a) = \Gamma(x)x^a(1+o(1))$.

D.2 Proof of Theorem 4.2: the case of HB ($\beta \neq 0$)

Structure of HB residual polynomials. We start with deriving expression for residual polynomial corresponding to HB method with step-size α and momentum β. These residual polynomials satisfy recurrence relation with constant coefficients

$$p_{n+1}(\lambda) = p_n(\lambda) - \alpha\lambda p_n(\lambda) + \beta(p_n(\lambda) - p_{n-1}(\lambda)), \qquad p_0(\lambda) = p_{-1}(\lambda) = 1. \tag{148}$$

Linear transformations of the polynomials $p_n(\lambda) = c^n q_n(z), z = ax + b$ lead to new polynomials q_n with different constants in their recurrence relations, which we choose to be that of Chebyshev polynomials.

$$p_n(\lambda) = (\sqrt{\beta})^n q_n(z(\lambda)), \quad z(\lambda) = \frac{1-\alpha\lambda+\beta}{2\sqrt{\beta}} \tag{149}$$

$$q_{n+1}(z) = 2zq_n(z) - q_{n-1}(z), \quad q_0(z) = 1, \quad q_{-1}(z) = \sqrt{\beta} \tag{150}$$

The initial conditions in (150) are satisfied with $q_n(z) = U_n(z) - \sqrt{\beta}U_{n-1}(z)$, where $U_n(z)$ are Chebyshev polynomials of second kind

$$U_n(z) = \begin{cases} \frac{\sin((n+1)\varphi)}{\sin\varphi}, & |z| \le 1 \\ \frac{\left((z+\sqrt{z^2-1})^{n+1} - (z-\sqrt{z^2-1})^{n+1}\right)}{2\sqrt{z^2-1}}, & |z| \ge 1 \end{cases} \tag{151}$$

Here $\cos\varphi = z$. Thus, we derived representation (31) for HB residual polynomials.

Let's list properties of $q_n(z)$ which will be useful in the subsequent parts of the proof.

1. *Monotonocity w.r.t. z*:
   $q_n(z)^2$ is monotone decreasing for $z \in (-\infty, -1]$ and monotone increasing for $z \in [1, \infty)$.

2. *Monotonocity w.r.t. n*:

$$(\sqrt{\beta})^{n+1} q_{n+1}(z) \le (\sqrt{\beta})^n q_n(z) \quad \text{for } z \in [1, \frac{1+\beta}{2\sqrt{\beta}}] \tag{152}$$

The first property follows from the fact that all $n-1$ zeros of the derivative $\frac{d}{dz}q_n(z)$ are located between n roots of $q_n(z)$, which in turn are located on $(-1, 1)$. To get the latter, note that the zero of $q_n(z) = U_n(z) - \sqrt{\beta}U_{n-1}(z) = \Big(\sin((n+1)\varphi) - \sqrt{\beta}\sin n\varphi\Big)/\sin\varphi$ is equivalent to

$$\begin{cases} \tan n\varphi = -\frac{\sin\varphi}{\cos\varphi - \sqrt{\beta}}, & \cos\varphi \ne \sqrt{\beta} \\ \cos n\varphi = 0, & \cos\varphi = \sqrt{\beta} \end{cases} \tag{153}$$

Here the first equation has at least $n-1$ solutions: a single solution on each interval $\frac{\pi}{2} + \pi k < n\varphi < \frac{\pi}{2} + \pi(k+1), \quad k = 0, \dots, n-2$. The remaining solution can be found in the interval containing $\cos\varphi = \sqrt{\beta}$, or exactly on the boundary if the second equation in (153) is satisfied.

The obtain the second property, note that it is equivalent to $r_n(z) \le 1$ where $r_{n+1}(z) = \frac{\sqrt{\beta}q_{n+1}(z)}{q_n(z)}$ and satisfies $r_{n+1}(z) = 2\sqrt{\beta}z - \frac{\beta}{r_n(z)}$ due to (150). Observing that $r_0(z) = 1$ we proceed by induction and assume that $r_n(z) \le 1$ for $z \in [1, \frac{1+\beta}{2\sqrt{\beta}}]$. Then, using that all $q_n(z)$ are positive for $z \ge 1$ and therefore $r_n(z) > 0$, we get $r_{n+1} \le 2\sqrt{\beta}z - \beta \le 1$ for $z \in [1, \frac{1+\beta}{2\sqrt{\beta}}]$.

Bounding the worst-case loss. First, let's bound $q_n(z)$ inside the oscillatory region $z \in [-1, 1]$. Since $|U_n(z)| \le n+1$ for $z \in [-1, 1]$, we get $|q_n(z)| = |U_n(z) - \sqrt{\beta}U_{n-1}(z)| \le 2n+1$.

Next, we bound $q_n(z)$ to the left of oscillatory region: $z < -1$. For convenience, we denote $z_\pm = z \pm \sqrt{z^2-1}$, and write

$$\begin{aligned} q_n(z) &= \frac{1}{2\sqrt{z^2-1}}\left[z_+^n\Big(z + \sqrt{z^2-1} - \sqrt{\beta}\Big) + z_-^n\Big(-z + \sqrt{z^2-1} + \sqrt{\beta}\Big)\right] \\ &= \frac{z_+^n + z_-^n}{2} + (z - \sqrt{\beta})\sum_{k=0}^{n-1} z_+^k z_-^{n-1-k} \end{aligned} \tag{154}$$

using the representation above and the fact that $|z_-| \geq |z_+|$ for $z < -1$, we get

$$\begin{aligned} |q_n(z)| &\leq \frac{|z_+|^n + |z_-|^n}{2} + (\sqrt{\beta} - z)\sum_{k=0}^{n-1} |z_+|^k |z_-|^{n-1-k} \\ &\leq |z_-|^n \left(1 + \frac{n(\sqrt{\beta} - z)}{|z_-|}\right) \leq (2n+1)|z_-|^n \end{aligned} \tag{155}$$

Now, we are ready to bound the flattened HB polynomial

$$\overline{p_n^2}(\lambda) = (\sqrt{\beta})^n \overline{q_n^2}\big(z(\lambda)\big), \quad \overline{q_n^2}(z) = \sup_{z_1 \leq y \leq z} q_n^2(y) \tag{156}$$

where $z_1 = z(\lambda = 1) = \frac{1-\alpha+\beta}{2\sqrt{\beta}}$. Now, recall the monotonicity properties of $q_n^2(z)$ on $(-\infty, -1]$ and $[1, \infty)$. Then, for $z_1 \geq 1$ we immediately get $\overline{q_n^2}(z) = z_n^2(z)$, while for $z_1 < 1$ we first get a single bound on $[z_1, 1]$ as

$$|q_n^2(z)| \leq \max\Big\{2n+1,\ \mathbb{1}_{z_1<-1}(2n+1)|z_-(z_1)|^n\Big\} = (2n+1)\left|z_1 - \sqrt{z_1^2-1}\right|^n \tag{157}$$

where for $z_1 \in (-1, 1)$ the square root $\sqrt{z_1^2 - 1}$ is understood in the complex sense.

Combining the obtained bounds, we can compactly characterize the flattened polynomial as

$$\overline{q_n^2}(z) = q_n^2(z) + \mathbb{1}_{z_1<1} O\Big(n^2 \left|z_1 - \sqrt{z_1^2-1}\right|^{2n}\Big), \tag{158}$$

implying for the worst-case loss

$$\overline{L_n^{(\zeta)}} = \frac{1}{2}\int_0^1 p_n^2(\lambda) d(\lambda^\zeta) + \int_0^1 \mathbb{1}_{z_1<1} O\Big(n^2\beta^n \left|z_1 - \sqrt{z_1^2-1}\right|^{2n}\Big) d(\lambda^\zeta) \tag{159}$$

which is exactly the momentum case of (33) with $u = \beta\left|z_1 - \sqrt{z_1^2-1}\right|^2$.

Calculating the loss under the exact power-law measure. While this can be done in a number of ways, we choose the approach based on the generating functions of $p_n^2(\lambda)$ and L_n. The approach is based on the connection between the asymptotic of the loss $L_n^{(\zeta)}$ and the singularity of its generating function

$$G_L(t) = \sum_{n=0}^{\infty} t^n L_n^{(\zeta)} \tag{160}$$

at $t = 1$. The two are connected by Tauberian theorem (Feller (1991), p. 445) which states that if generating function $G(t) = \sum_n t^n a_n$ of a sequence a_n has asymptotic $G(1-\varepsilon) = C\varepsilon^{-\rho}(1 + o(1)),\ \rho > 0$, then

$$\sum_{k=1}^{n} a_k = \frac{C}{\Gamma(\rho+1)} n^\rho (1 + o(1)), \quad n \to \infty. \tag{161}$$

We will apply this theorem to the sequence $a_n = n^m L_n^{(\zeta)}$, where $m = \lfloor \zeta \rfloor$ is required to get a divergent behavior of the partial sums.

First, recall that thanks to (149), (154) we can write HB residual polynomials in the form $p_n(\lambda) = f_+(z(\lambda))\big(\sqrt{\beta}z_+(\lambda)\big)^n + f_-(z(\lambda))\big(\sqrt{\beta}z_-(\lambda)\big)^n$. Then, generating function of $p_n^2(\lambda)$ can be immediately written as

$$\begin{aligned} G_p(t,\lambda) \equiv \sum_{n=0}^{\infty} t^n p_n^2(\lambda) &= \sum_{n=0}^{\infty} \Big[f_+^2 (t\beta z_+^2)^n + f_-^2 (t\beta z_-^2)^n + 2 f_+ f_- (t\beta z_+ z_-)^n \Big] \\ &= \frac{f_+^2}{1-t\beta z_+^2} + \frac{f_-^2}{1-t\beta z_-^2} + \frac{2f_+f_-}{1-t\beta z_+ z_-} \end{aligned} \tag{162}$$

Substituting $z(\lambda)$ into $f_+(z), f_-(z), z_+(z), z_-(z)$ and straightforwardly simplifying the expression (e.g., using symbolic computer algebra software) reveals that $G_p(t,\lambda)$ is a rational function of its arguments equal to

$$G_p(t,\lambda) = \frac{(1-\beta t)(1-\beta^2 t) + 2\alpha\beta\lambda t}{(1-\beta t)\Big((1-t)(1-\beta^2 t) + \alpha\lambda t(2+2\beta-\alpha\lambda t)\Big)} \tag{163}$$

$$\overset{t=1-\varepsilon}{=} \frac{1}{\varepsilon + \frac{2\alpha}{1-\beta}\lambda}\Big(1 + O(\varepsilon) + O(\lambda)\Big), \quad \text{as} \quad \varepsilon \searrow 0 \quad \text{and} \quad \lambda \searrow 0 \tag{164}$$

Here we observed from (163) that when stability condition $\alpha < 2(1+\beta)$ is satisfied, $G_p(t,\lambda)$ on $[0,1]^2$ is regular everywhere except the singularity at $t=1$, $\lambda = 0$.

Focusing on the contribution to the loss $L_{a,n}^{(\zeta)} = \frac{1}{2}\int_0^a p_n^2(\lambda) d(\lambda^\zeta)$ from $[0,a]$, $a \le 1$ (to be specified later), we write m-th derivative of its generating function $G_{L,a}(t) = \sum_{n=0}^{\infty} t^n L_{a,n}^{(\zeta)}$ as

$$\begin{aligned} \Big(t\frac{d}{dt}\Big)^m G_{L,a}(t) &= \frac{1}{2}\int_0^a \Big(t\frac{\partial}{\partial t}\Big)^m G_p(t,\lambda) d(\lambda^\zeta) \\ &= \frac{m!}{2}\int_0^a \frac{\zeta\lambda^{\zeta-1}}{\Big(\varepsilon + \frac{2\alpha}{1-\beta}\lambda\Big)^{m+1}}\big(1 + O(\varepsilon) + O(\lambda)\big) d\lambda \\ &= \frac{m!\zeta}{2}\left(\frac{2\alpha}{1-\beta}\right)^{-\zeta} \varepsilon^{\zeta-m-1} \int_0^{\frac{2\alpha a}{\varepsilon(1-\beta)}} \frac{x^{\zeta-1}}{(1+x)^{m+1}}\big(1 + (1+x)O(\varepsilon)\big) dx \\ &= \frac{m!\zeta}{2}\left(\frac{2\alpha}{1-\beta}\right)^{-\zeta} \varepsilon^{\zeta-m-1}(1+O(\varepsilon)) \int_0^{\infty} \frac{x^{\zeta-1}dx}{(1+x)^{m+1}} \\ &= \frac{\Gamma(\zeta+1)\Gamma(m+1-\zeta)}{2}\left(\frac{2\alpha}{1-\beta}\right)^{-\zeta} \varepsilon^{\zeta-m-1}(1+O(\varepsilon)) \end{aligned} \tag{165}$$

where in the second-to-last line, we recognized the integral representation of Beta function $B(\zeta, m+1-\zeta)$ and subsequently expressed it in terms of Gamma functions. Observing that $\big(t\frac{d}{dt}\big)^m G_{L,a}(t)$ is the generating function of the sequence $n^m L_{a,n}^{(\zeta)}$, we apply Tauberian theorem to get asymptotic of the partial sums

$$\sum_{k=0}^{n} k^m L_{a,k}^{(\zeta)} = \frac{\Gamma(\zeta+1)}{2(m+1-\zeta)}\left(\frac{2\alpha}{1-\beta}\right)^{-\zeta} n^{m+1-\zeta}(1+o(1)) \tag{166}$$

Now, we choose $a = \min\{1, (1-\sqrt{\beta})^2/\alpha\}$ where the second option corresponds to the border of the oscillating region $z(a) = 1$ of polynomials $p_n(\lambda)$. Then, the monotonicity property (152) imply monotonicity of $p_n^2(\lambda)$ on $[0, a]$, and therefore monotonicity of $L_{a,k}^{(\zeta)}$. This enables to use Lemma D.1 below on partial sums (166) and get $L_{a,n}^{(\zeta)} = \frac{\Gamma(\zeta+1)}{2}(\frac{2\alpha n}{1-\beta})^{-\zeta}(1 + o(1))$, which is the same as (32) thanks to exponentially suppressed (see eq. (157)) contribution to the loss from $\lambda \in [a, 1]$.

Lemma D.1. *Assume a sequence a_n is monotonically decreasing, and there is $m \geq 0$ such that $\sum_{k=1}^{n} k^m a_k = n^\xi(1 + o(1))$ with some $\xi > 0$. Then, $a_n = \xi n^{\xi-m-1}(1 + o(1))$.*

Proof Take a fixed $r > 0$ and consider the partial sums $S_n = \sum_{k=n}^{n'-1} k^m a_k$ in the chunks $[n, n')$, $n' = \lfloor n(1 + r)\rfloor$. In the limit $n \to \infty$ we have

$$S_n = n^\xi\big[(1 + r)^\xi - 1\big](1 + o(1)) \tag{167}$$

$$S_n \leq a_n \sum_{k=n}^{n'-1} k^m = a_n n^{m+1}\frac{(1 + r)^{m+1} - 1}{m + 1}(1 + o(1)) \tag{168}$$

Combining these two estimates yields the bound

$$a_n \geq n^{\xi-m-1}\frac{(m + 1)[(1 + r)^\xi - 1]}{(1 + r)^{m+1} - 1}(1 + o(1)) \tag{169}$$

As r was arbitrary, we take $r \searrow 0$ in (169) and get $a_n \geq \xi n^{\xi-m-1}(1 + o(1))$. Next, we take a fixed $0 < r < 1$ and consider the partial sums in the chunks $[n', n)$, $n' = \lfloor n(1-r)\rfloor$. Then, similar reasoning gives $a_n \leq \xi n^{\xi-m-1}(1 + o(1))$, thus completing the proof. ■

D.3 Proof of Theorem 4.3

First, observe that the measure $\rho_{\zeta,\nu}$ trivially satisfies the condition (10) since the eigenvalues corresponding to $\rho_{\zeta,\nu}$ are $\lambda_k = k^{-\nu}$. Next, we take $\lambda \in [\lambda_k, \lambda_{k-1})$ and evaluate the respective cumulative distribution function $\rho_{\zeta,\nu}([0, \lambda])$ as

$$\rho_{\zeta,\nu}([0, \lambda]) = \sum_{l\geq k}^{\infty}\Big(l^{-\zeta\nu} - (l + 1)^{-\zeta\nu}\Big) = k^{-\zeta\nu} = \lambda_k^\zeta \leq \lambda^\zeta, \tag{170}$$

which confirms that $\rho_{\zeta,\nu}$ satisfies the main condition (8).

To bound the loss under the measure $\rho_{\zeta.\nu}$, we first do so for $\rho_{\zeta,\nu}([0, \lambda])$. Take a $k_0 \geq 0$ such that $G_{\zeta,\nu}(\lambda) = \lambda^\zeta - \zeta\nu\lambda^{\zeta+1/\nu}$ is increasing on $[0, \lambda_{k_0}]$ and consider again $\lambda \in (\lambda_{k+1}, \lambda_k]$, $k \geq k_0$:

$$G_{\zeta,\nu}(\lambda) \leq G_{\zeta,\nu}(\lambda_k) = k^{-\zeta\nu} - \zeta\nu k^{-\zeta\nu-1} \leq (k + 1)^{-\zeta\nu} = \lambda_{k+1}^\zeta \leq \rho([0, \lambda]) \tag{171}$$

Thus, we established that $\rho_{\zeta,\nu}([0, \lambda]) \geq G_{\zeta,\nu}(\lambda)$ for $\lambda \in [0, \lambda_{k_0}]$. Now, let $p_n(\lambda)$ be the residual polynomial of the considered GD algorithm and λ_0 be it's left-most zero. Since

$p_n(\lambda)$ is monotone decreasing on $[0, \lambda_0]$ (see the proof of Theorem 4.2), the contribution to the loss from $[0, \lambda^*]$, $\lambda^* = \min(\lambda_{k_0}, \lambda_0)$ is given by

$$\begin{aligned}\int_0^{\lambda^*} p_n^2(\lambda)\rho_{\zeta,\nu}(d\lambda) &= p_n^2(\lambda^*)\rho_{\zeta,\nu}([0,\lambda^*]) - \int_0^{\lambda^*} \Big(\frac{d}{d\lambda}p_n^2(\lambda)\Big)\rho_{\zeta,\nu}([0,\lambda])d\lambda \\ &\geq p_n^2(\lambda^*)G_{\zeta,\nu}(\lambda^*) - \int_0^{\lambda^*} \Big(\frac{d}{d\lambda}p_n^2(\lambda)\Big)G_{\zeta,\nu}(\lambda)d\lambda = \int_0^{\lambda^*} p_n^2(\lambda)G'_{\zeta,\nu}(\lambda)d\lambda\end{aligned} \tag{172}$$

Referring to the proof of Theorem (4.2) and eq. (154) we observe that on any $[a, 1]$, $a > 0$ the residual polynomials decay uniformly as $p_n^2(\lambda) = O(r^n)$, $r < 1$. Using this and (32) we bound the loss as

$$\begin{aligned}L_n &= \frac{1}{2}\int_0^1 p_n^2(\lambda)\rho_{\zeta,\nu}(d\lambda) = \frac{1}{2}\int_0^{\lambda^*} p_n^2(\lambda)\rho_{\zeta,\nu}(d\lambda) + O(r^n) \\ &\geq \frac{1}{2}\int_0^{\lambda^*} p_n^2(\lambda)G'_{\zeta,\nu}(\lambda)d\lambda + O(r^n) = \frac{1}{2}\int_0^1 p_n^2(\lambda)\Big[d(\lambda^\zeta) - \zeta\nu d(\lambda^{\zeta+1/\nu})\Big] + O(r^n) \\ &= \frac{1}{2}\int_0^1 p_n^2(\lambda)d(\lambda^\zeta) + O(n^{-\zeta-1/\nu}) + O(r^n) = \frac{\Gamma(\zeta+1)}{2}\left(\frac{2\alpha n}{1-\beta}\right)^{-\zeta}(1+o(1))\end{aligned} \tag{173}$$

Appendix E. Accelerated methods for exact power-law spectral measure

Proof of Theorem 4.4. We substitute CG residual polynomial given by (36) into the loss (25)

$$p_n(\lambda) = \frac{P_n^{(\zeta,0)}(1-2\lambda)}{P_n^{(\zeta,0)}(1)}. \tag{174}$$

Then, by a change of variables,

$$L(\mathbf{w}_n) = \frac{1}{2}\int_0^1 p_n^2(\lambda)d\lambda^\zeta = \frac{\zeta}{2^{\zeta+1}\big(P_n^{(\zeta,0)}(1)\big)^2}\int_{-1}^1 (1-x)^{\zeta-1}(P_n^{(\zeta,0)}(x))^2 dx. \tag{175}$$

We will use Rodrigues' formula for $P_n^{(a,b)}$:

$$P_n^{(a,b)}(x) = \frac{(-1)^n}{2^n n!}(1-x)^{-a}(1+x)^{-b}\frac{d^n}{dx^n}[(1-x)^{a+n}(1+x)^{b+n}]. \tag{176}$$

It gives (with $a = \zeta, b = 0$)

$$\int_{-1}^1 (1-x)^{\zeta-1}(P_n^{(\zeta,0)}(x))^2 dx = \frac{(-1)^n}{2^n n!}\int_{-1}^1 (1-x)^{-1}\frac{d^n}{dx^n}[(1-x)^{\zeta+n}(1+x)^n]P_n^{(\zeta,0)}(x)dx. \tag{177}$$

Observe that we can write

$$(1-x)^{-1}P_n^{(\zeta,0)}(x) = P_n^{(\zeta,0)}(1)(1-x)^{-1} + q_{n-1}(x) \tag{178}$$

with some polynomial q_{n-1} of degree $n-1$. Suppose that we perform repeated integration by parts in the r.h.s. of (177), moving all the derivatives $\frac{d^n}{dx^n}$ from $(1-x)^{\zeta+n}(1+x)^n$ to

$(1-x)^{-1}P_n^{(\zeta,0)}(x)$. Thanks to the condition $\zeta > 0$, all the boundary terms will vanish. Moreover, since $d^n q_{n-1}/dx^n = 0$, only the first term in the r.h.s. of Eq. (178) will give a nonvanishing contribution to the resulting integral, specifically

$$\begin{aligned}\int_{-1}^{1}(1-x)^{-1}\frac{d^n}{dx^n}[(1-x)^{\zeta+n}(1+x)^n]P_n^{(\zeta,0)}(x)dx \\ &= (-1)^n n! P_n^{(\zeta,0)}(1)\int_{-1}^{1}(1-x)^{-n-1}[(1-x)^{\zeta+n}(1+x)^n]dx \\ &= (-1)^n n! P_n^{(\zeta,0)}(1)2^{\zeta+n}\int_0^1 t^{\zeta-1}(1-t)^n dt \\ &= (-1)^n n! P_n^{(\zeta,0)}(1)2^{\zeta+n}B(\zeta,n+1).\end{aligned} \tag{179}$$

Using the fact that

$$P_n^{(\zeta,0)}(1) = \frac{\Gamma(\zeta+n+1)}{n!\Gamma(\zeta+1)}, \tag{180}$$

we finally obtain

$$L(\mathbf{w}_n) = \frac{\zeta}{2^{\zeta+1}\big(P_n^{(\zeta,0)}(1)\big)^2}\cdot\frac{(-1)^n}{2^n n!}\cdot(-1)^n n! P_n^{(\zeta,0)}(1)2^{\zeta+n}\cdot\frac{\Gamma(\zeta)\Gamma(n+1)}{\Gamma(\zeta+n+1)} \tag{181}$$

$$= \frac{\Gamma^2(\zeta+1)n!^2}{2\Gamma^2(\zeta+n+1)} \tag{182}$$

$$= \frac{\Gamma^2(\zeta+1)}{2}n^{-2\zeta}(1+o(1)) \quad (\lambda\to 0+). \tag{183}$$

Proof of Proposition 4.5. The principal difference between $a > \zeta - \frac{1}{2}$ and $a < \zeta - \frac{1}{2}$ is that in the former case the dominating contribution to the integral comes from $\lambda \sim n^{-2}$ while in the latter case the dominant contribution comes from $\lambda \sim 1$.

Let's start with $a > \zeta - \frac{1}{2}$. The classical asymptotic expansion of Jacobi polynomials $P_n^{(a,b)}(\cos\theta)$ at small θ (Szegö (1959), Theorem 8.21.12.) states for a fixed c, ε and $N = n + \frac{1}{2}(a+b+1)$

$$\begin{aligned}\sin\left(\frac{\theta}{2}\right)^a\cos\left(\frac{\theta}{2}\right)^b P_n^{(a,b)}(\cos\theta) =& N^{-a}\frac{\Gamma(n+a+1)}{n!}\left(\frac{\theta}{\sin\theta}\right)^{\frac{1}{2}}J_a(N\theta) \\ &+\begin{cases}\theta^{a+2}O(n^a), & \theta < cn^{-1}\\ \theta^{\frac{1}{2}}O(n^{-\frac{3}{2}}), & cn^{-1}<\theta<\pi-\varepsilon\end{cases}\end{aligned} \tag{184}$$

Using that $z^{-a}J_a(z)$ is bounded and also $|J_a(z)| = O(z^{-\frac{1}{2}})$ uniformly, we adopt (184) to our needs and write an asymptotic form

$$\begin{aligned}&\Big(P_n^{(a,b)}(\cos\theta)\Big/P_n^{(a,b)}(1)\Big)^2 = \\ &=\frac{\Gamma^2(a+1)}{2N^{-2a-1}}\frac{N\theta J_a^2(N\theta)}{\big(\sin\frac{\theta}{2}\big)^{2a+1}\big(\cos\frac{\theta}{2}\big)^{2b+1}}+\begin{cases}\theta^2 O(1), & \theta<cn^{-1}\\ \theta^{-2a}O(n^{-2a-2}), & cn^{-1}<\theta<\pi-\varepsilon\end{cases} \\ &=\Gamma^2(a+1)\left(\frac{N\theta}{2}\right)^{-2a}J_a^2(N\theta)+\begin{cases}\theta O(1), & \theta<cn^{-1}\\ \theta^{-2a}O(n^{-2a-1}), & cn^{-1}<\theta<\pi-\varepsilon\end{cases}\end{aligned} \tag{185}$$

Next, we use coordinate transformation $\cos\theta = 1 - r\lambda$, $d\lambda^\zeta = r^{-\zeta}2^{1-\zeta}\zeta\theta^{2\zeta-1}(1+O(\theta))d\theta$ and obtained asymptotic form to calculate the integral in the left-hand side of (39)

$$\begin{aligned}
\int_0^1 (q_n^{(a,b,r)}(\lambda))^2 d\lambda^\zeta &\overset{(1)}{=} \Gamma^2(a+1)r^{-\zeta}2^{2a+1-\zeta}\zeta\int_0^{\theta_r}\theta^{2\zeta-1-2a}N^{-2a}J_a^2(N\theta)d\theta \\
&\quad + O(1)\int_0^{cn^{-1}}\theta^{2\zeta}d\theta + O(n^{-2a-1})\int_{cn^{-1}}^{\theta_r}\theta^{2\zeta-1-2a}d\theta \\
&\overset{(2)}{=}\Gamma^2(a+1)r^{-\zeta}2^{2a+1-\zeta}\zeta N^{-2\zeta}\int_0^{N\theta_r} z^{2\zeta-1-2a}J_a^2(z)dz + n^{-2\zeta}O(\delta_{\zeta,a}(n)) \\
&=\Gamma^2(a+1)r^{-\zeta}2^{2a+1-\zeta}\zeta N^{-2\zeta}\int_0^{\infty} z^{2\zeta-1-2a}J_a^2(z)dz + n^{-2\zeta}O(\delta_{\zeta,a}(n)) \\
&\overset{(3)}{=}\Big(\frac{r}{2}\Big)^{-\zeta}\frac{\zeta\Gamma^2(a+1)B(\zeta,2a-2\zeta+1)}{\Gamma^2(a-\zeta+1)}n^{-2\zeta} + n^{-2\zeta}O(\delta_{\zeta,a}(n))
\end{aligned} \tag{186}$$

where in (1) $\cos\theta_r = 1 - r$. In (2), error term $\delta_{\zeta,a}(n)$ comes from estimation of the last two integrals in (1) and is given by

$$\delta_{\zeta,a}(n) = \begin{cases} n^{-1}, & a > \zeta \\ n^{-1}\log n, & a = \zeta \\ n^{2\zeta-2a-1}, & \zeta - \frac{1}{2} < a < \zeta \end{cases} \tag{187}$$

This error term gives more fine-grained characterization of the correction than $o(1)$ term in (39), where it was omitted for brevity. Finally, in (3) we used known integral for Bessel function, which can be found e.g. in DLMF (§10.22).

Now we proceed with the second case $a < \zeta - \frac{1}{2}$. Using the first asymptotic in (185) and analyzing the error terms similarly to (186) we get

$$\begin{aligned}
\int_0^1 (q_n^{(a,b,r)}(\lambda))^2 d\lambda^\zeta &= \frac{2^\zeta\zeta\Gamma^2(a+1)}{r^\zeta N^{2a+1}}\int_0^{\theta_r}\frac{N\theta J_a^2(N\theta)d(\sin\frac{\theta}{2})}{\big(\sin\frac{\theta}{2}\big)^{2a-2\zeta+2}\big(\cos\frac{\theta}{2}\big)^{2b+1}} + O(n^{-2a-2}) \\
&\overset{(1)}{=}\frac{2^\zeta\zeta\Gamma^2(a+1)}{\pi r^\zeta N^{2a+1}}\int_0^{\theta_r}\big(\sin\tfrac{\theta}{2}\big)^{2\zeta-2a-2}\big(\cos\tfrac{\theta}{2}\big)^{-2b-1}d\sin\tfrac{\theta}{2} + o(n^{-2a-1}) \\
&\overset{(2)}{=}\frac{2^\zeta\zeta\Gamma^2(a+1)}{2\pi r^\zeta n^{2a+1}}\int_0^{\frac{r}{2}} x^{\zeta-a-\frac{3}{2}}(1-x)^{-b-\frac{1}{2}}dx + o(n^{-2a-1}) \\
&=\frac{2^\zeta\zeta\Gamma^2(a+1)B(\frac{r}{2};\zeta-a-\frac{1}{2},b+\frac{1}{2})}{2\pi r^\zeta}n^{-2a-1}\big(1+o(1)\big)
\end{aligned} \tag{188}$$

Here in (1) we used the property that $\lim_{n\to\infty}\int_0^1 f(x)[nxJ_a^2(nx)]dx = \pi^{-1}\int_0^1 f(x)dx$ for functions $f(x)$ integrable on $(0,1)$ and Lipschitz on any $(\varepsilon,1)$. This property follows from $zJ_a(z)^2$ being bounded, and asymptotic of Bessel function $zJ_a^2(z) = 2\pi^{-1}\cos^2(z-\alpha) + O(z^{-1})$. In (2) we changed integration coordinate to $x = \sin^2\frac{\theta}{2}$.

Learning rate schedule associated with Jacobi ansatz (38)**.** In this section, we obtain the learning rate schedule (40). Note that we can set $r = 1$ in derivation but receiver it in the end since it always comes in combination $r\lambda$, therefore multiplicative modifying learning rate.

Now, we start with standard recurrence relations (120) and first substitute $x = 1 - \lambda$:

$$\begin{aligned}
&2(n+1)(n+a+b+1)(2n+a+b)P_{n+1}^{(a,b)}(1-\lambda) = \\
&\quad -(2n+a+b)(2n+a+b+1)(2n+a+b+2)\lambda P_n^{(a,b)}(1-\lambda) \\
&\quad +\Big[(2n+a+b+1)(a^2-b^2)+(2n+a+b)(2n+a+b+1)(2n+a+b+2)\Big]P_n^{(a,b)}(1-\lambda) \\
&\quad -2(n+a)(n+b)(2n+a+b+2)P_{n-1}^{(a,b)}(1-\lambda).
\end{aligned} \tag{189}$$

Next step is to add normalization $P_n^{(a,b)}(1-\lambda) = P_n^{(a,b)}(1)p_n^{(a,b)}(\lambda)$, where according to (119) $P_n^{(a,b)}(1) = \frac{\Gamma(n+a+1)}{\Gamma(n+1)\Gamma(a+1)}$. We get

$$\begin{aligned}
&p_{n+1}^{(a,b)}(\lambda) = \\
&\quad -\frac{(2n+a+b+1)(2n+a+b+2)}{2(n+a+1)(n+a+b+1)}\lambda p_n^{(a,b)}(\lambda) \\
&\quad +\Big[\frac{(2n+a+b+1)(2n+a+b+2)}{2(n+a+1)(n+a+b+1)}+\frac{(2n+a+b+1)(a^2-b^2)}{2(n+a+1)(n+a+b+1)(2n+a+b)}\Big]p_n^{(a,b)}(\lambda) \\
&\quad -\frac{n(n+b)(2n+a+b+2)}{(n+a+1)(n+a+b+1)(2n+a+b)}p_{n-1}^{(a,b)}(\lambda)
\end{aligned} \tag{190}$$

Comparing with (116), this gives exactly (40) with $r = 1$. Then, r is recovered by setting $\alpha_n \to r\alpha_n$. Finally, the asymptotic form in (40) is obtained by a simple Taylor expansion with respect to $\frac{1}{n}$.

Appendix F. Non-constant learning rates: upper bounds

F.1 Accelerated Heavy Ball rates

Proof of Theorem 4.6 From the properties of polynomials $q_n^{(a,b,r)}$, we will take only non-degeneracy of zeros and monotonicity of local maxima. The former follows directly from the same property of Jacobi polynomials. The monotonicity property is also inherited from Jacobi polynomials and the respective argument is implicitly given in Section 7.32 of Szegö (1959). For completeness, we formulate and prove the monotonicity property here.

Lemma F.1. *Assume $a, b > -\frac{1}{2}$ and let $x_0 = \frac{b-a}{a+b+1}$. Next, denote $\{x_i\}_{i=1}^m$ the positions of local maxima of $|P_n^{(a,b)}(x)|$ on $(x_0, 1)$ sorted in increasing order: $x_0 < x_1 < \ldots < x_m < 1$. Then, the values at local maxima and at the endpoints form an increasing sequence*

$$|P_n^{(a,b)}(x_0)| < |P_n^{(a,b)}(x_1)| < \ldots < |P_n^{(a,b)}(x_m)| < |P_n^{(a,b)}(1)| \tag{191}$$

Proof Recall that $y(x) = P_n^{(a,b)}(x)$ satisfy differential equation

$$(1-x^2)y'' + \Big(b-a-(a+b+2)x\Big)y' + n(n+a+b+1)y = 0 \tag{192}$$

Then, to characterize $y(x)$ at local extrema we introduce function $f(x)$ and calculate its derivative taking into account differential equation for $y(x)$.

$$f(x) = \Big(y(x)\Big)^2 + \frac{1-x^2}{n(n+a+b+1)}\Big(y'(x)\Big)^2 \tag{193}$$

$$f'(x) = \frac{2(a+b+1)}{n(n+a+b+1)}(x-x_0)\Big(y'(x)\Big)^2 \tag{194}$$

From the derivative expression we see that $f(x)$ is monotonously increasing on $[x_0, x]$. Now observe that $f(x) = y^2(x)$ at local minima x_i and at endpoint $x = 1$, which implies monotonicity of maxima $|P_n^{(a,b)}(x_1)| < \ldots < |P_n^{(a,b)}(x_m)| < |P_n^{(a,b)}(1)|$. For the left endpoint x_0 we notice that $y^2(x_0) \le f(x_0) < f(x_1) = y^2(x_1)$ which completes the proof. ■

Note that according to (38), restriction on r means $\lambda \in [0,1]$ maps to $[1-r, 1] \subseteq [x_0, 1]$ in the argument of $P_n^{(a,b)}(x)$ with $x_0 = \frac{b-a}{a+b+1}$. Then, according the lemma F.1, for the local maxima $\{\lambda_i\}_{i=1}^m$ of $q_n^2(\lambda)$ on $(0,1)$ we have

$$|q_n(0)| > |q_n(\lambda_1)| > \ldots > |q_n(\lambda_m)| > |q_n(1)| \tag{195}$$

From this point, we will not require any additional properties of $q_n^{(a,b,r)}$, and therefore denote $p_n(\lambda) \equiv q_n^{(a,b,r)}(\lambda)$. From the proof of Theorem 4.1, we recall the structure of flattened polynomial $\overline{q_n^2}(\lambda)$ given by (127). Monotonicity of local maxima of $q_n(\lambda)$ means that x_i in (127) are simply local maxima of $q_n(\lambda)$, and, in particular, $x_1 = 0$.

Now, we focus on the contribution to the losses (28) and (29) from a single flat region $[y_i, x_{i+1}]$. Let c be the root of $p_n(\lambda)$ on $[y_i, x_{i+1}]$, and denote $\widetilde{p}_n(\lambda) = p_n(\lambda)/(\lambda - c)$. As $\widetilde{p}_n(\lambda)$ has all its roots outside of $[y_i, x_{i+1}]$, on this segment $\widetilde{p}_n^2(\lambda)$ is either 1) monotonically increasing and then decreasing 2) monotonically decreasing 3) monotonically increasing. Therefore, the minima of $\widetilde{p}_n^2(\lambda)$ on $[y_i, x_{i+1}]$ is attained at one of the ends of the segments. Taking into account that $\int_{y_i}^{x_{i+1}} \overline{p_n^2}(\lambda) d(\lambda^\zeta) = p_n^2(x_{i+1})(x_{i+1}^\zeta - y_i^\zeta)$, we have

$$\begin{aligned}
&\int_{y_i}^{x_{i+1}} p_n^2(\lambda) d(\lambda^\zeta) \Big/ \int_{y_i}^{x_{i+1}} \overline{p_n^2}(\lambda) d(\lambda^\zeta) \\
&= \int_{y_i}^{x_{i+1}} \widetilde{p}_n^2(\lambda)(\lambda - c)^2 d(\lambda^\zeta) \Big/ \left(p_n^2(x_{i+1}) \int_{y_i}^{x_{i+1}} d(\lambda^\zeta)\right) \\
&\ge \int_{y_i}^{x_{i+1}} \min\big(\widetilde{p}_n^2(y_i), \widetilde{p}_n^2(x_{i+1})\big)(\lambda - c)^2 d(\lambda^\zeta) \Big/ \left(p_n^2(x_{i+1}) \int_{y_i}^{x_{i+1}} d(\lambda^\zeta)\right) \\
&= \int_{y_i}^{x_{i+1}} \frac{(\lambda - c)^2}{\max\big((y_i - c)^2, (x_{i+1} - c)^2\big)} d(\lambda^\zeta) \Big/ \int_{y_i}^{x_{i+1}} d(\lambda^\zeta) \ge \frac{1}{C[\rho_\zeta]}
\end{aligned} \tag{196}$$

Here, we observed that the expression in the last line is a single realization of the expression minimized in (43), and therefore can be bounded with respective infimum $\frac{1}{C[\rho_\zeta]}$. Thus, we have bounded the ratio of integrals $\int \overline{p_n^2}(\lambda) d(\lambda^\zeta) \Big/ \int p_n^2(\lambda) d(\lambda^\zeta)$ on $[y_i, x_{i+1}]$ with $C[\rho_\zeta]$. As the same bound trivially holds on $[x_i, y_i]$ (flattened and original polynomials are equal), and the respective segments cover the whole $[0,1]$, we get (42).

Proof of Proposition 4.7. Let's denote the ratio of integrals under the infimum in (43) as $C[\rho](c, x_l, x_r)$. Then, for the exact power law measure $\rho_\zeta([0,\lambda]) = \lambda^\zeta$ the ratio becomes invariant under scaling transformations: $C[\rho_\zeta](\eta c, \eta x_l, \eta x_r) = C[\rho_\zeta](c, x_l, x_r), \forall \eta > 0$. This scale invariance implies that it is sufficient only to consider the case $x_r = 1$. Now, we can simply denote $x_l = x$.

We reduce the space of (x, c) potentially containing the infimum by noting that for $c \notin [x, 1]$, it is always beneficial to move c to the nearest endpoint of $[x, 1]$. Next, we take advantage of monotonicity of the density $p_\zeta(\lambda) = \frac{d\rho_\zeta([0,\lambda])}{d\lambda} = \zeta\lambda^{\zeta-1}$ to further narrow down the search space: for any $c \in [x, 1]$ we compare it with its reflection $c' = x + 1 - c$ w.r.t. window center $c_0 = (x + 1)/2$

$$\begin{aligned} C[\rho](c', x, 1) - C[\rho](c, x, 1) &\propto \int_x^1 \big[(\lambda - c')^2 - (\lambda - c)^2\big] p(\lambda) d\lambda \\ &\propto (c - c_0) \int_x^1 \big[\lambda - c_0\big] p(\lambda) d\lambda \propto (c - c_0) \int_0^{\frac{1-x}{2}} z\Big[p\big(c_0 - z\big) - p\big(c_0 + z\big)\Big] dz \end{aligned} \tag{197}$$

Here and in the remaining parts of the proof, the proportionality sign $\propto$ denotes equality up to a positive multiplicative factor. From the last line, we see that for increasing density $p(\lambda)$ it is always more beneficial to be in the right half of the window $c > c_0$, and vice versa for decreasing $p(\lambda)$. In the case of constant $p(\lambda)$, as for $\zeta = 1$, both halves of the window $[x, 1]$ are equivalent.

The right (left) position of c w.r.t. window center c_0 implies that parabola in (43) is normalized by its left(right) endpoint. Slightly abusing the fact that after fixing the normalization endpoint, the positions of c away from the intended half of the window are always suboptimal, we may write

$$C_\zeta^{-1} = \inf_{c,\, 0 \le x < 1} C_\zeta(x, c) \tag{198}$$

$$C_\zeta(x, c) \equiv \begin{cases} \langle (\lambda - c)^2 \rangle_x / \langle (1 - c)^2 \rangle_x, & 0 < \zeta \le 1 \\ \langle (\lambda - c)^2 \rangle_x / \langle (x - c)^2 \rangle_x, & \zeta > 1 \end{cases} \tag{199}$$

where angle brackets denote the integral $\langle f(\lambda) \rangle_x \equiv \int_x^1 f(\lambda) p_\zeta(\lambda) d\lambda$. Now we proceed with finding the optimal point $(x^*, c^*) = \operatorname{argmin} C_\zeta(x, c)$ and respective value C_ζ separately for the cases $\zeta \le 1$ and $\zeta > 0$. In both cases, it turns out that at the optimum $x^* = 0$, which makes it easy to find respective c^*. However, showing that $x^* = 0$ is technically challenging, and we had to use symbolic computation, e.g. Wolfram Mathematica Inc..

Decreasing density ($\zeta \le 1$). First, let's find optimal $c = c^*(x)$ at a given x. Since $C_\zeta(x, c)$ is a rational function in c, the optimum is given by a zero of the derivative

$$\frac{\partial C_\zeta(x, c)}{\partial c} = 2\frac{\big\langle (\lambda - c)^2 \big\rangle_x}{(1 - c)^3 \langle 1 \rangle_x} - 2\frac{\langle \lambda - c \rangle_x}{(1 - c)^2 \langle 1 \rangle_x} \propto \operatorname{sgn}(1 - c) \Big(c \langle 1 - \lambda \rangle_x - \langle \lambda(1 - \lambda) \rangle_x \Big) \tag{200}$$

From this expression, we see that the minimum is indeed unique and achieved at

$$c^*(x) = \frac{\langle \lambda(1 - \lambda) \rangle_x}{\langle 1 - \lambda \rangle_x} \tag{201}$$

Next, as the global minimum of $C_\zeta(x,c)$ is located on the curve $c=c^*(x)$, we may analyze the derivative along the curve $\frac{d}{dx}C_\zeta(x,c^*(x)) = \frac{\partial}{\partial x}C_\zeta(x,c^*(x))$

$$\begin{aligned}\frac{\partial}{\partial x}C_\zeta(x,c^*) &\propto \frac{\partial}{\partial x}\frac{\left\langle(\lambda-c^*)^2\right\rangle_x}{\langle 1\rangle_x} = -\frac{(x-c^*)^2 p_\zeta(x)}{\langle 1\rangle_x} + \frac{\left\langle(\lambda-c^*)^2\right\rangle_x}{(\langle 1\rangle_x)^2}p_\zeta(x)\\ &\propto \left\langle(\lambda-c^*)^2-(x-c^*)^2\right\rangle_x = \left\langle\lambda^2-x^2\right\rangle_x - 2c^*\left\langle\lambda-x\right\rangle_x\\ &\propto \left\langle\lambda^2-x^2\right\rangle_x\left\langle 1-\lambda\right\rangle_x - 2\left\langle\lambda-x\right\rangle_x\left\langle\lambda(1-\lambda)\right\rangle_x \equiv g_1(x)\end{aligned} \tag{202}$$

Now we will show that $g_1(x)$, and therefore the derivative $\frac{d}{dx}C_\zeta(x,c^*(x))$, is non-negative for $x\in(0,1)$ implying that the global minimum is achieved at $x=0$. First, observe that $g_1(x)$ can be written as an explicit function of x by substituting moments $\langle\lambda^n\rangle_x = \frac{\zeta}{\zeta+n}(1-x^{\zeta+n})$. Next, we perform a *top-down* step: use symbolic computations to evaluate several derivatives of $g_1(x)$ in the form of the following statements

1. $g_1(x) = \frac{d}{dx}g_1(x) = \left(\frac{d}{dx}\right)^2 g_1(x) = \left(\frac{d}{dx}\right)^3 g_1(x) = 0$ at $x=1$.
2. Denote $g_2(x) = x^{3-\zeta}\left(\frac{d}{dx}\right)^3 g_1(x)$. Then $g_2(x) = \frac{d}{dx}g_2(x) = \left(\frac{d}{dx}\right)^2 g_2(x) = 0$ at $x=1$.
3. $g_2(x=0) = -\zeta(4+\zeta(\zeta^2-4\zeta+7))/(\zeta+1) < 0$ for $0<\zeta\le 1$.
4. $\left(\frac{d}{dx}\right)^3 g_2(x) = 10(1-\zeta)\zeta^2$ at $x=1$, and $\left(\frac{d}{dx}\right)^4 g_2(x) = 8(1-\zeta)\zeta^2(1+2\zeta)x^{\zeta-2}$.

Now we proceed with a *bottom-up* step: use simple expressions of lowest derivatives to reconstruct the positivity of $g_1(x)$. It will be convenient to call *sign signature* of a function the sequence of its signs on a given interval, e.g. $f(x)=(2x-1)^2-0.5$ has sign signature $(+-+)$ on interval $(0,1)$. Then

1. $\left(\frac{d}{dx}\right)^4 g_2(x) > 0$ and $\left(\frac{d}{dx}\right)^3 g_2(x=1) > 0$ implies that $\left(\frac{d}{dx}\right)^3 g_2(x)$ has sign signature either $(-+)$ or $(+)$ on $(0,1)$.
2. Sign signature of $\left(\frac{d}{dx}\right)^3 g_2(x)$ and $\left(\frac{d}{dx}\right)^2 g_2(x=1)=0$ implies that $\left(\frac{d}{dx}\right)^2 g_2(x)$ has sign signature either $(+-)$ or $(-)$ on $(0,1)$.
3. Sign signature of $\left(\frac{d}{dx}\right)^2 g_2(x)$ and $\frac{d}{dx}g_2(x=1)=0$ implies that $\frac{d}{dx}g_2(x)$ has sign signature either $(-+)$ or $(+)$ on $(0,1)$.
4. Sign signature of $\frac{d}{dx}g_2(x)$ implies that maximum of $g_2(x)$ on $[0,1]$ is reached either at $x=0$ or $x=1$. Since $g_2(1)=0$ and $g_2(0)<0$, we have $g_2(x)\le 0$ and therefore $\left(\frac{d}{dx}\right)^3 g_1(x)\le 0$ on $(0,1)$.
5. $g_1(x) = \frac{d}{dx}g_1(x) = \left(\frac{d}{dx}\right)^2 g_1(x) = \left(\frac{d}{dx}\right)^3 g_1(x) = 0$ at $x=1$ and $\left(\frac{d}{dx}\right)^3 g_1(x)\le 0$ on $(0,1)$ implies that $g_1(x)\ge 0$ on $(0,1)$, which completes the argument.

Finally, we can proceed with calculating the value at the global minimum $C_\zeta(x=0,c^*(x=0))$. When $x=0$, the moments are $\langle\lambda^n\rangle_0 = \zeta/(\zeta+n)$, which after substitution into (201) gives $c^*=\zeta/(\zeta+2)$. Then we again substitute the moments into $C_\zeta(0,c^*)$ and get

$$C_\zeta^{-1} = C_\zeta(0,c^*(0)) = \left(\frac{\zeta+2}{2}\right)^2\left(\frac{\zeta}{\zeta+2} - 2\frac{\zeta}{\zeta+1}\frac{\zeta}{\zeta+2} + \left(\frac{\zeta}{\zeta+2}\right)^2\right) = \frac{\zeta}{2(\zeta+1)} \tag{203}$$

Increasing density ($\zeta > 1$). Similarly to $\zeta \leq 1$ case, we start with obtaining optimal c at fixed x by calculating the derivative

$$\frac{\partial C_\zeta(x,c)}{\partial c} = 2\frac{\left\langle (\lambda-c)^2\right\rangle_x}{(x-c)^3\left\langle 1\right\rangle_x} - 2\frac{\left\langle \lambda - c\right\rangle_x}{(x-c)^2\left\langle 1\right\rangle_x} \propto \operatorname{sgn}(x-c)\left(\left\langle \lambda(\lambda-x)\right\rangle_x - c\left\langle \lambda - x\right\rangle_x\right) \quad (204)$$

which gives the optimal position of the parabola root

$$c^*(x) = \frac{\left\langle \lambda(\lambda-x)\right\rangle_x}{\left\langle \lambda - x\right\rangle_x}. \quad (205)$$

Next, we again search for the global minimum of $C_\zeta(x,c)$ on the curve $c = c^*(x)$, by analyzing the derivative along the curve $\frac{d}{dx}C_\zeta(x,c^*(x)) = \frac{\partial}{\partial x}C_\zeta(x,c^*(x))$

$$\begin{aligned}
\frac{\partial}{\partial x}C_\zeta(x,c^*) &= \frac{\partial}{\partial x}\frac{\left\langle (\lambda-c^*)^2\right\rangle_x}{\left\langle (x-c^*)^2\right\rangle_x} = p_\zeta(x)\frac{\left\langle (\lambda-c^*)^2 - (x-c^*)^2\right\rangle_x}{(x-c^*)^2\left\langle 1\right\rangle_x^2} - 2\frac{\left\langle (\lambda-c^*)^2\right\rangle_x}{(x-c^*)^3\left\langle 1\right\rangle_x} \\
&\propto -p_\zeta(x)(x-c^*)\left\langle \lambda^2 - x^2 - 2(\lambda-x)c^*\right\rangle_x + 2\left\langle (\lambda-c^*)^2\right\rangle_x\left\langle 1\right\rangle_x \\
&= \frac{\left\langle (\lambda-x)^2\right\rangle_x}{\left\langle \lambda-x\right\rangle_x^2}\Big[2\left\langle 1\right\rangle_x\left(\left\langle \lambda^2\right\rangle_x\left\langle 1\right\rangle_x - \left\langle \lambda\right\rangle_x^2\right) - p_\zeta(x)\left\langle \lambda-x\right\rangle_x\left\langle (\lambda-x)^2\right\rangle_x\Big] \\
&\propto 2\left\langle 1\right\rangle_x\left(\left\langle \lambda^2\right\rangle_x\left\langle 1\right\rangle_x - \left\langle \lambda\right\rangle_x^2\right) - p_\zeta(x)\left\langle \lambda-x\right\rangle_x\left\langle (\lambda-x)^2\right\rangle_x \equiv g_1(x)
\end{aligned} \quad (206)$$

Continuing the same strategy as for the case $\zeta \leq 1$, we will show $g_1(x) > 0$ on $(0,1)$ by exploiting the explicit form of $g_1(x)$ and symbolic computations. *top-down* step:

1. $\frac{\frac{d}{dx}g_1(x)}{(1-x)x^{\zeta-2}} \equiv g_2(x)$ is a polynomial in variables (x, x^ζ).
2. $g_1(x) = g_2(x) = \frac{d}{dx}g_2(x) = \left(\frac{d}{dx}\right)^2 g_2(x) = 0$ at $x = 1$.
3. $\left(\frac{d}{dx}\right)^3 g_2(x) = 2(\zeta-1)\zeta^3(1-x)x^{\zeta-2} > 0$ on $(0,1)$.

Then, the *bottom-up* argumentation is the following

1. $\left(\frac{d}{dx}\right)^2 g_2(x=1) = 0$ and $\left(\frac{d}{dx}\right)^3 g_2(x) > 0$ on $(0,1)$ implies $\left(\frac{d}{dx}\right)^2 g_2(x) < 0$ on $(0,1)$.
2. $\frac{d}{dx}g_2(x=1) = 0$ and $\left(\frac{d}{dx}\right)^2 g_2(x) < 0$ on $(0,1)$ implies $\frac{d}{dx}g_2(x) > 0$ on $(0,1)$.
3. $g_2(x=1) = 0$ and $\frac{d}{dx}g_2(x) > 0$ on $(0,1)$ implies $g_2(x) < 0$ on $(0,1)$, and, therefore, $\frac{d}{dx}g_1(x) < 0$ on $(0,1)$.
4. $g_1(x=1) = 0$ and $\frac{d}{dx}g_1(x) < 0$ on $(0,1)$ implies $g_1(x) > 0$ on $(0,1)$.

Having shown that at the minimum $x^* = 0$, we find the optimal position of the parabola root to be $c^* = c^*(x=0) = \frac{\zeta+1}{\zeta+2}$. Plugging c^* into $C_\zeta(0,c)$ gives

$$C_\zeta^{-1} = C_\zeta(0,c^*) = \left(\frac{\zeta+2}{\zeta+1}\right)^2\left(\frac{\zeta}{\zeta+2} - 2\frac{\zeta(\zeta+1)}{(\zeta+1)(\zeta+2)} + \left(\frac{\zeta+1}{\zeta+2}\right)^2\right) = \frac{1}{(\zeta+1)^2} \quad (207)$$

F.2 Gradient Descent with predefined schedule

Preliminaries: "reduced" polynomials. We will use a construction based on "reduced" polynomials $p_{n,m}(x), 0 \le m \le n$. Given a residual (equal to 1 at $x = 0$) polynomial $p_n(x)$ of degree n we define the corresponding reduced polynomials by

$$p_{n,m}(x) \equiv \prod_{i=1}^{m} \Big(1 - \frac{x}{x_i}\Big), \tag{208}$$

where x_i are the roots of $p_n(x)$ sorted in the decreasing order $x_1 \ge x_2 \ge \ldots$. In particular, $p_{n,n}(x) = p_n(x)$. We will need the following technical lemma about residual polynomials

Lemma F.2. *Let $p_n(x)$ be a residual polynomial of degree n such that $|p_n(x)| \le 1$ if $x \in [0, a]$. Then the same bound also holds for the corresponding reduced polynomials:*

$$|p_{n,m}(x)| \le 1 \;\; \textit{for all} \;\; x \in [0, a] \;\; \textit{and} \;\; 0 \le m \le n. \tag{209}$$

Proof Let's fix m and divide the segment $[0, a]$ into two parts: $[0, 2x_m]$ and $[2x_m, a]$ We will prove bound (208) separately for each part. (If $x_m < 0$ or $2x_m > a$, then there is only one nontrivial part that covers $[0, a]$, and we consider only the respective single case.) Recall that the initial polynomial $p_n(x)$ and reduced polynomial $p_{n,m}(x)$ can be written as

$$p_n(x) = \prod_{i=1}^{n} \Big(1 - \frac{x}{x_i}\Big), \quad p_{n,m}(x) = \prod_{i=1}^{m} \Big(1 - \frac{x}{x_i}\Big). \tag{210}$$

1. *Case $x \in [0, 2x_m]$.* In this case we have $|1 - \frac{x}{x_i}| \le 1,\ i \le m$, and thus $|p_{n,m}(x)| \le 1$.

2. *Case $x \in [2x_m, a]$.* In this case we write

$$|p_{n,m}(x)| = \frac{|p_n(x)|}{\prod_{i=m+1}^{n} |1 - \frac{x}{x_i}|}. \tag{211}$$

   Then for $x \in [2x_m, a]$ and $i > m$, if $x_i > 0$, then

$$\Big|1 - \frac{x}{x_i}\Big| = \frac{x}{x_i} - 1 \ge \frac{x}{x_m} - 1 \ge 1. \tag{212}$$

   The same inequality $|1 - \frac{x}{x_i}| \ge 1$ clearly also holds if $x_i < 0$. Thus, in any case $|1 - \frac{x}{x_i}| \ge 1$. It follows then from (211) that $p_{n,m}(x) \le p_n(x) \le 1$.

■

Construction of learning rates α_n. Given $\zeta > 0$, fix some $a > b > -\frac{1}{2}$ and $r \le 2$ and consider the residual polynomials p_n obtained by shifting and normalizing the Jacobi polynomials as in Eq. (38):

$$p_n(x) = \frac{P_n^{(a,b)}(1 - r\lambda)}{P_n^{(a,b)}(1)}. \tag{213}$$

A well-known result from Szegö (1959) states that if $a > b \geq -\frac{1}{2}$, then the largest value of the Jacobi polynomial $P_n^{(a,b)}$ on the segment $[-1,1]$ is reached at $z=1$:

$$\max_{|z|\leq 1}\left|P_n^{(a,b)}(z)\right| = P_n^{(a,b)}(1). \tag{214}$$

It follows that our polynomials p_n satisfy the condition $\max_{0\leq x\leq 1}|p_n(x)| = 1$ of Lemma F.2.

Now we describe a construction of schedule $\{\alpha_i\}$ which gives the convergence rate $O(n^{-2\zeta})$ for GD. Informally, we will build our GD polynomial $\widetilde{q}_k(x)$ by sequentially taking the roots of $p_1(x), p_2(x), p_4(x), p_8(x), \ldots$. More precisely, to determine α_i we first find the largest l such that $i \geq 2^l$, and denote $l_i \equiv l$, $n_i \equiv 2^l$, $m_i \equiv i - 2^l + 1$. Then we set

$$\alpha_i = \frac{1}{x_{m_i}^{(n_i)}}, \tag{215}$$

where $x_m^{(n)}$ is the m'th root of $p_n(x)$ (as usual, taken in decreasing order). In this way the polynomial $\widetilde{q}_k(x)$ corresponding to our scheduled GD is

$$\widetilde{q}_k(x) = p_{n_k,m_k}(x)\prod_{l=0}^{l_k-1} p_{2^l}(x). \tag{216}$$

We can now prove the main result.

Proof of Theorem 4.9. As already mentioned, the polynomials p_n satisfy the hypothesis of Lemma F.2 and so, by this lemma, $|p_{n,m}(x)| \leq 1$ on $[0,1]$. We apply this bound to $\widetilde{q}_k(x)$:

$$|\widetilde{q}_k(x)| = |p_{n_k,m_k}(x)||p_{n_k/2}(x)|\prod_{l=0}^{l_k-2}|p_{2^l}(x)| \leq |p_{n_k/2}(x)| \tag{217}$$

Using Corollary 4.8, we then get

$$L(\mathbf{w}_k) \leq C_\zeta R(a,r,\zeta)\left(\frac{n_k}{2}\right)^{-2\zeta}\left(1+O(\frac{1}{n_k})\right) \leq C_\zeta R(a,r,\zeta)4^{2\zeta}k^{-2\zeta}\Big(1+O(\frac{1}{k})\Big), \tag{218}$$

where in the last inequality we used $\frac{k}{2} < n_k \leq k$. This completes the proof of Theorem 4.9.

F.3 Conjugate Gradients: discrete spectrum

Proof of Theorem 4.10. Consider the degree-n residual polynomial q_n of the form

$$q_n(\lambda) = \prod_{s=1}^{n}(1-\lambda/a_s) = \Big(\prod_{s=1}^{\lfloor n/2\rfloor}(1-\lambda/\lambda_s)\Big)r_n(x) \tag{219}$$

where $\lambda_1 \geq \ldots \geq \lambda_{\lfloor n/2 \rfloor}$ are the $\lfloor n/2 \rfloor$ largest eigenvalues (atoms of the measure ρ) and r_n is some degree-$(n - \lfloor n/2 \rfloor)$ residual polynomial. Then,

$$\int_0^1 q_n^2(\lambda)\rho(d\lambda) = \int_0^{\lambda_{\lfloor n/2 \rfloor}} q_n^2(\lambda)\rho(d\lambda) \tag{220}$$

$$\leq \int_0^{\lambda_{\lfloor n/2 \rfloor}} r_n^2(\lambda)\rho(d\lambda) \tag{221}$$

$$\leq \int_0^{\Lambda(n/2)^{-\nu}} r_n^2(\lambda)\rho(d\lambda) \tag{222}$$

$$= \int_0^1 r_n^2(\Lambda(n/2)^{-\nu}t)\rho(d\Lambda(n/2)^{-\nu}t) \tag{223}$$

$$= \Lambda^{\zeta}(n/2)^{-\nu\zeta} \int_0^1 r_n^2(\Lambda(n/2)^{-\nu}t)\rho_n(dt), \tag{224}$$

where the measure ρ_n is defined for Borel subsets $X \subset \mathbb{R}$ by rescaling

$$\rho_n(X) = \Lambda^{-\zeta}(n/2)^{\nu\zeta}\rho(\Lambda(n/2)^{-\nu}X). \tag{225}$$

The measure ρ_n satisfies the same power law bound (9) as ρ:

$$\rho_n((0,\lambda]) = \Lambda^{-\zeta}(n/2)^{\nu\zeta}\rho((0,\Lambda(n/2)^{-\nu}\lambda]) \tag{226}$$

$$\leq \Lambda^{-\zeta}(n/2)^{\nu\zeta}(\Lambda(n/2)^{-\nu}\lambda)^{\zeta} \tag{227}$$

$$= \lambda^{\zeta}. \tag{228}$$

It follows that we can apply Corollary 4.8 and find r_n such that

$$\int_0^1 r_n^2(\Lambda(n/2)^{-\nu}t)\rho_n(dt) \leq 2QC_{\zeta}R(a,r,\zeta)(n/2)^{-2\zeta}\big(1+o(1)\big). \tag{229}$$

Combining with (224), this gives the desired bound (47):

$$L(\mathbf{w}_n) \leq \frac{1}{2}\int_0^1 q_n^2(\lambda)\rho(d\lambda) \leq QC_{\zeta}R(a,r,\zeta)\Lambda^{-\zeta}(n/2)^{-(\nu+2)\zeta}\big(1+o(1))\big). \tag{230}$$

Appendix G. Non-constant learning rates: lower bounds

G.1 Non-adaptive schedules

Proof of Theorem 4.11. Consider the power law distribution $\rho_{\zeta}((0,\lambda]) = \lambda^{\zeta}$ with $\lambda_{\max} = 1$. Let us define discrete distributions $(\rho_{\zeta,r})_{r\in[0,1]}$ subject to the spectral conditions (49), (50) of the theorem and such that

$$\rho_{\zeta} = \int_0^1 \rho_{\zeta,r}dr. \tag{231}$$

To this end, we set

$$\rho_{\zeta,r} = \sum_{k=1}^{\infty} \rho([(k+1)^{-\nu}, k^{-\nu}])\delta_{a_{k,r}} \tag{232}$$

with some $a_{k,r} \in [(k+1)^{-\nu}, k^{-\nu}]$. It is clear that thus defined $\rho_{\zeta,r}$ satisfies Eqs. (49), (50), and one can also satisfy Eq. (231) by suitably adjusting $a_{k,r}$.

We will construct the distribution ρ corresponding to the desired A and $\mathbf{b}$ by joining a sequence of segments of the distributions $\rho_{\zeta,r}$:

$$\rho = \sum_{k=1}^{\infty} \rho_{\zeta,r_k}|_{[h_k,h_{k-1})}, \quad h_0 = 1. \tag{233}$$

It is easy to see that if $h_k \to 0+$ sufficiently fast, say $h_k \le h_{k-1}/2$ for all k, then such ρ also satisfies the required conditions (49), (50).

Consider the first step of the construction of ρ. Arguing as in Section B, the loss $L(\mathbf{w}_n)$ of a general multistep method (48) can be written as

$$L(\mathbf{w}_n) = \frac{1}{2}\int_0^1 q_n^2(\lambda)\rho(d\lambda), \tag{234}$$

where q_n is some residual polynomial of degree n. We know from the exact solution of the minimization problem

$$\min_{\widetilde{q}_{n_1}:\deg \widetilde{q}_{n_1}=n_1, \widetilde{q}_{n_1}(0)=1} \frac{1}{2}\int_0^1 \widetilde{q}_{n_1}^2 \rho_\zeta(d\lambda) \tag{235}$$

by a rescaled Jacobi polynomial (see Theorem 4.4) that

$$\frac{1}{2}\int_0^1 q_{n_1}^2 \rho_\zeta(d\lambda) > C n_1^{-2\zeta}, \tag{236}$$

where q_{n_1} is the residual polynomial corresponding to the given optimization algorithm and C is an absolute constant. Choose n_1 sufficiently large so that

$$\frac{1}{2}\int_0^1 q_{n_1}^2 \rho_\zeta(d\lambda) > n_1^{-2\zeta-\epsilon}. \tag{237}$$

It follows from the decomposition (231) that there exists r_1 such that this inequality remains valid if we replace ρ_ζ by ρ_{ζ,r_1} :

$$\frac{1}{2}\int_0^1 q_{n_1}^2 \rho_{\zeta,r_1}(d\lambda) > n_1^{-2\zeta-\epsilon}. \tag{238}$$

We can then choose h_1 sufficiently small so that

$$\frac{1}{2}\int_{h_1}^1 q_{n_1}^2 \rho_{\zeta,r_1}(d\lambda) > n_1^{-2\zeta-\epsilon}. \tag{239}$$

Consider now the second step of the construction of ρ. Using the homogeneity of the distribution ρ_ζ, the lower bound (236) extends to the segment $[0, h_1]$ with the additional factor h_1^ζ :

$$\frac{1}{2}\int_0^{h_1} q_{n_2}^2 \rho_\zeta(d\lambda) > C h_1^\zeta n_2^{-2\zeta}. \tag{240}$$

Arguing as before, we then choose a sufficiently large n_2, a suitable r_2, and a sufficiently small h_2 such that

$$\frac{1}{2}\int_{h_2}^{h_1} q_{n_2}^2 \rho_{\zeta,r_2}(d\lambda) > n_2^{-2\zeta-\epsilon}. \tag{241}$$

Continuing this process, we obtain the full desired expansion (233).

G.2 CG with discrete spectrum

G.2.1 Proof of Proposition 4.14 for $0 < \zeta < 1$

In this section we prove Proposition 4.14 for $0 < \zeta < 0$, i.e. we prove only Statement 1 with $m = 0$. The remaining cases will be considered in Section G.2.2.

Denote $\mathbf{x} = (\widetilde{A} + \epsilon)^{-1}\mathbf{e}_1$. In coordinates, the equation $(\widetilde{A} + \epsilon)\mathbf{x} = \mathbf{e}_1$ is a system of finite difference equations

$$-(n-1)^{-\nu}(\tfrac{n}{n-1})^g x_{n-1} \tag{242}$$
$$+((n-1)^{-\nu}(\tfrac{n}{n-1})^{2g} + n^{-\nu})x_n \tag{243}$$
$$-n^{-\nu}(\tfrac{n+1}{n})^g x_{n+1} = -\epsilon x_n, \quad n = 2, 3, \dots \tag{244}$$
$$x_1 - 2^g x_2 = 1 - \epsilon x_1, \tag{245}$$

where we introduced the constant

$$g = \frac{1 - (2+\nu)\zeta}{2}. \tag{246}$$

Let us make the substitution

$$y_n = n^g x_n. \tag{247}$$

Then the finite difference equations become

$$-(n-1)^{-(\nu+2g)} n^g y_{n-1} \tag{248}$$
$$+((n-1)^{-(\nu+2g)} n^{2g} + n^{-\nu}) n^{-g} y_n \tag{249}$$
$$-n^{-(\nu+g)} y_{n+1} = -\epsilon n^{-g} y_n, \quad n = 2, 3, \dots \tag{250}$$
$$y_1 - y_2 = 1 - \epsilon y_1. \tag{251}$$

We further introduce the variable h by

$$h = \epsilon^{1/(2+\nu)} \tag{252}$$

and the variable θ_n by

$$1 - h\theta_n = \frac{y_n}{y_{n+1}}. \tag{253}$$

By multiplying the difference equation by $n^{\nu+g}/y_n$, we can then rewrite it as

$$-(\tfrac{n-1}{n})^{-(\nu+2g)}(1 - h\theta_{n-1}) \tag{254}$$
$$+(\tfrac{n-1}{n})^{-(\nu+2g)} + 1 \tag{255}$$
$$-(1 - h\theta_n)^{-1} = -h^2 (hn)^\nu, \quad n = 2, 3, \dots. \tag{256}$$

Introducing the variable s by

$$s = nh, \tag{257}$$

we then get

$$\theta_{n-1} = \Big(\frac{\theta_n}{1 - h\theta_n} - hs^\nu\Big)\Big(\frac{s-h}{s}\Big)^{\nu+2g}, \quad n = 2, 3, \dots \tag{258}$$

This system of finite difference equations has a one-parameter family of solutions that can be specified by one value θ_{n_0} at a particular n_0. We will now identify a special solution (θ^*_n) for which $\mathbf{x} \in l^2$. We expect the components x_n of this $\mathbf{x}$ to have the same sign and decay to 0 sufficiently fast as $n \to +\infty$. By Eq. (253), these conditions will be satisfied if we ensure that $\theta^*_n < 0$ for all n and $\theta^*_n \to -\infty$ sufficiently fast as $n \to \infty$ (note that this need not be the case for a generic solution (θ_n) since it may diverge at a finite n or start increasing at some n). Importantly, we will establish growth bounds for the solution (θ^*_n) that hold uniformly in h.

Lemma G.1. *Let constants a, b be such that $a > \nu$ and $0 < b < \nu/2$. Then there exists a unique solution (θ^*_n) of Eq. (258) such that we have*

$$-s^a \le \theta^*_n \le -s^b \text{ provided } h < h_0 \text{ and } s = nh > s_0 \tag{259}$$

with some constants $h_0, s_0 > 0$.

Proof Let $G_{s,h} : \mathbb{R} \to \mathbb{R}$ denote the transformation in the iteration law (258):

$$\theta_{n-1} = G_{nh,h}(\theta_n). \tag{260}$$

Consider the intervals

$$I_s = [-s^a, -s^b]. \tag{261}$$

We show now that under our iteration law the intervals I_s are ordered by inclusion.

Lemma G.2. *There exist constants $h_0, s_0 > 0$ such that for all $h < h_0$ and $s = hn > s_0$ we have*

$$G_{nh,h}(I_{hn}) \subset I_{h(n-1)}. \tag{262}$$

Proof By monotonicity of $G_{h,s}$, Eq. (262) will be established if we show

$$-(s-h)^b \ge G_{s,h}(-s^b), \tag{263}$$
$$-(s-h)^a \le G_{s,h}(-s^a). \tag{264}$$

Fulfilling condition (263). This inequality is equivalent to

$$-(1-h/s)^{b-(\nu+2g)} \ge -\frac{1}{1+hs^b} - hs^{\nu-b}. \tag{265}$$

Since we assume that h is sufficiently small and s sufficiently large, we can write $-(1-h/s)^{b-(\nu+2g)} \ge -1 - Ch/s$ with some absolute constant C. Therefore, it is sufficient to establish

$$-Ch/s \ge \frac{hs^b}{1+hs^b} - hs^{\nu-b}. \tag{266}$$

Dividing by h and bounding $1 + hs^b \ge 1$, this in turn reduces to

$$-C/s \ge s^b - s^{\nu-b}. \tag{267}$$

Clearly, this inequality holds for sufficiently large s if $b < \nu/2$.

Fulfilling condition (264). By a similar argument, it suffices to fulfill

$$C/s \le \frac{s^a}{1+h_0 s^a} - s^{\nu-a}. \tag{268}$$

This holds for all sufficiently large s if we choose any $a > \nu$ and h_0 small enough. ■

Lemma G.2 yields a nested sequence of compact intervals

$$I_{hn_0} \supset G_{h(n_0+1),h}(I_{h(n_0+1)}) \supset G_{h(n_0+1),h}(G_{h(n_0+2),h}(I_{h(n_0+2)})) \supset \dots, \tag{269}$$

where $n_0 = \lceil s_0/h \rceil$. This sequence has a non-empty intersection I. Then, a sequence θ_n^* such that $\theta_{n_0}^* \in I$ satisfies the desired bounds (259).

We argue now that such a sequence θ_n^* is unique. It is easy to see that if a solution θ_n^* satisfies the upper bound in (259), then the respective sequence x_n belongs to l^2. Different sequences θ_n^* would correspond to different l^2 sequences x_n. However, the equation $(\widetilde{A} + \epsilon)\mathbf{x} = \mathbf{e}_1$ has a unique l^2 solution $\mathbf{x}$. ■

We study now the behavior of θ_n^* at small n. It is convenient to introduce the new variables ω_n by

$$\theta_n = s^{\nu+2g}\omega_n. \tag{270}$$

Then the difference equation (258) becomes

$$\omega_{n-1} = \frac{\omega_n}{1 - hs^{\nu+2g}\omega_n} - hs^{-2g}, \quad n = 2, 3, \dots \tag{271}$$

Let ω_n^* be the sequence ω_n corresponding to the sequence θ_n^* found in Lemma G.1, and s_0 be as in this lemma.

Lemma G.3. *Let $0 < \zeta < 1$. Then there exist constants $c < d < 0$ such that*

$$c \le \omega_n^* \le d \text{ provided } h < h_0 \text{ and } s = nh < s_0 \tag{272}$$

with some constant $h_0 > 0$.

Proof *Lower bound.* By Lemma G.1 we have $\omega_n^* < 0$ for $n \ge n_0 = \lceil s_0/h \rceil$, and Eq. (271) then implies that $\omega_n^* < 0$ for all n; moreover,

$$\omega_{n-1}^* \ge \omega_n^* - hs^{-2g}, \quad s = nh, \quad n = 2, 3, \dots, n_0. \tag{273}$$

Note that by the definition of g in Eq. (246) and the inequality $\zeta > 0$ we have

$$2g < 1. \tag{274}$$

It follows that for any $n = 1, 2, \dots, n_0$

$$\omega_n^* \ge \omega_{n_0}^* - \int_{nh}^{n_0 h} s^{-2g} ds + O(h + h^{1-2g}) \tag{275}$$

$$\ge -s_0^{-(\nu+2g)} s_0^a - (1-2g)^{-1} s_0^{1-2g} + O(1) \ge c, \quad (h \to 0) \tag{276}$$

for a suitable constant c.

Upper bound. On the other hand, Eq. (271) implies that

$$\omega^*_{n-1} \le \frac{\omega^*_n}{1 - hs^{\nu+2g}\omega^*_n}, \quad n = 2, 3, \dots \tag{277}$$

This is equivalent to

$$(\omega^*_{n-1})^{-1} \ge (\omega^*_n)^{-1} - hs^{\nu+2g}, \quad n = 2, 3, \dots \tag{278}$$

Note that by the definition (246) of g we have

$$\nu + 2g + 1 = (2 + \nu)(1 - \zeta), \tag{279}$$

so that the inequality $\zeta < 1$ yields

$$\nu + 2g > -1. \tag{280}$$

It follows that for any $n = 1, 2, \dots, n_0$

$$(\omega^*_n)^{-1} \ge (\omega^*_{n_0})^{-1} - \int_{nh}^{n_0 h} s^{\nu+2g} ds + O(h + h^{\nu+2g+1}) \tag{281}$$

$$\ge -s_0^{\nu+2g} s_0^{-b} - (\nu + 2g + 1)^{-1} s_0^{\nu+2g+1} + O(1) \ge d^{-1}, \quad (h \to 0) \tag{282}$$

for a suitable constant $d < 0$, which implies the desired bound. □ ■

Lemmas G.1 and G.3 allow us to control the initial element x_1 of the sequence $\mathbf{x}$. From Eqs. (247), (251), and (253) we have

$$x_1 = y_1 = \Big(\epsilon - \frac{h\theta^*_1}{1 - h\theta^*_1}\Big)^{-1}. \tag{283}$$

By Eqs. (252), (270), (279), and Lemma G.3 we have

$$h\theta^*_1 = h^{\nu+2g+1}\omega^*_1 = \epsilon^{1-\zeta}\omega^*_1 = O(\epsilon^{1-\zeta}), \quad (\epsilon \to 0). \tag{284}$$

Since $\omega^*_1 < d < 0$ for all ϵ, it follows that if $\zeta < 1$, then

$$x_1 = O(\epsilon^{\zeta-1}), \quad (\epsilon \to 0). \tag{285}$$

This is the desired bound, since $x_1 = \langle \mathbf{e}_1, (\widetilde{\mathcal{A}} + \epsilon)^{-1}\mathbf{e}_1 \rangle$.

G.2.2 Proof of Proposition 4.14 for $\zeta > 1$

We retain the notation introduced in the previous section. Throughout this section, we write $a_n = O(b_n)$ meaning that $|a_n| \le Cb_n$ for all n with some constant $C > 0$ that might depend on ν and ζ but not n or ϵ.

We start with a technical lemma that describes the special solution θ^*_n for $\zeta > 1$ (thus complementing Lemma G.3 that covers $\zeta < 1$).

Lemma G.4. *If $\zeta > 1$, then for sufficiently small ϵ the special sequence θ_n^* satisfies*

$$\theta_n^* \leq \frac{(\nu+2)(1-\zeta)}{nh}. \tag{286}$$

Proof By Lemma G.1, if h is small enough then for sufficiently large n we have $\theta_n^* \leq -s^{\nu/2}$ and hence bound (286) is satisfied if n is large enough. We prove now that if it is satisfied for some $n \geq 2$, then it is also satisfied for $n-1$. Consider Eq. (258) for θ_{n-1}^*:

$$\theta_{n-1}^* = \Big(\frac{\theta_n^*}{1-h\theta_n^*} - hs^\nu\Big)\Big(\frac{s-h}{s}\Big)^{\nu+2g}. \tag{287}$$

Recall that $\nu + 2g + 1 = (\nu+2)(1-\zeta)$. Denote $a = (\nu+2)(1-\zeta)$. Using the fact that the function $x \mapsto x/(1-hx)$ is increasing on $(-\infty, 0)$ and the assumption $\theta_n^* \leq a/s$, we get

$$\begin{aligned}
\theta_{n-1}^* &\leq \frac{\theta_n^*}{1-h\theta_n^*}\Big(\frac{s-h}{s}\Big)^{\nu+2g} && (288)\\
&\leq \frac{a/s}{1-(h/s)a}\Big(\frac{s-h}{s}\Big)^{\nu+2g} && (289)\\
&= \frac{1}{s-h}\frac{a}{1-(h/s)a}\Big(\frac{s-h}{s}\Big)^{a} && (290)\\
&\leq \frac{a}{s-h}, && (291)
\end{aligned}$$

where in the last step we used the inequality

$$\frac{(1-x)^a}{1-xa} \geq 1, \quad 0 < x < 1, a < 0 \tag{292}$$

with $x = h/s$. ■

Our proof of Proposition 4.14 is based on the following extended version of this proposition that contains bounds on the growth of the involved sequences.

Proposition G.5. *Let $n_0 = \lfloor s_0/h \rfloor$ with the constant s_0 appearing in Lemma G.1.*

1. *Assuming $2m < \zeta$ for some integer $m \geq 1$, the vectors $\widetilde{A}^{-m}\mathbf{e}_1$ and $\widetilde{A}^{-m}(\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1$ exist as elements of l^2 and*

$$(\widetilde{A}^{-m}\mathbf{e}_1)_n = O(n^{\frac{-1-(2+\nu)(\zeta-2m)}{2}}), \tag{293}$$

$$(\widetilde{A}^{-m}(\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1)_n = \begin{cases} O(n^{\frac{-1-(2+\nu)(\zeta-2m-2)}{2}}), & n \leq n_0 \\ O(\epsilon^{-1}n^{\frac{-1-(2+\nu)(\zeta-2m)}{2}}), & n > n_0 \end{cases} \tag{294}$$

2. *Assuming $2m+1 < \zeta$ for some integer $m \geq 0$, the vectors $J^{-1}\widetilde{A}^{-m}\mathbf{e}_1$ and $J^{-1}\widetilde{A}^{-m}(\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1$ exist as elements of l^2 and*

$$(J^{-1}\widetilde{A}^{-m}\mathbf{e}_1)_n = O(n^{\frac{-1-(2+\nu)(\zeta-2m-1)}{2}}), \tag{295}$$

$$(J^{-1}\widetilde{A}^{-m}(\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1)_n = \begin{cases} O(n^{\frac{-1-(2+\nu)(\zeta-2m-3)}{2}}), & n \leq n_0 \\ O(\epsilon^{-1}n^{\frac{-1-(2+\nu)(\zeta-2m-1)}{2}}), & n > n_0 \end{cases} \tag{296}$$

Let us first show that this proposition implies desired Proposition 4.14 in all cases except $0 < \zeta < 1$ (covered in the previous section). Let $2m < \zeta < 2m+1$ for some integer $m \geq 1$, then, using Eqs. (293), (294),

$$\langle \widetilde{A}^{-m}\mathbf{e}_1, \widetilde{A}^{-m}(\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1\rangle = \sum_{n=1}^{\infty}(\widetilde{A}^{-m}\mathbf{e}_1)_n(\widetilde{A}^{-m}(\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1)_n \tag{297}$$

$$= \sum_{n=1}^{n_0} + \sum_{n=n_0+1}^{\infty} \tag{298}$$

$$= \sum_{n=1}^{n_0} O(n^{\frac{-1-(2+\nu)(\zeta-2m)}{2}})O(n^{\frac{-1-(2+\nu)(\zeta-2m-2)}{2}}) \tag{299}$$

$$+ \sum_{n=n_0+1}^{\infty} O(n^{\frac{-1-(2+\nu)(\zeta-2m)}{2}})O(\epsilon^{-1}n^{\frac{-1-(2+\nu)(\zeta-2m)}{2}}) \tag{300}$$

$$= \sum_{n=1}^{n_0} O(n^{-1-(2+\nu)(\zeta-2m-1)}) + \sum_{n=n_0+1}^{\infty} O(\epsilon^{-1}n^{-1-(2+\nu)(\zeta-2m)}) \tag{301}$$

$$= O(n_0^{-(2+\nu)(\zeta-2m-1)}) \tag{302}$$

$$= O(\epsilon^{\zeta-2m-1}), \tag{303}$$

which is the desired bound (67). Note that here we used both inequalities $2m < \zeta < 2m+1$ and the identity $\epsilon = h^{2+\nu}$ to get Eq. (302).

By a similar reasoning, if $2m+1 < \zeta < 2m+2$ with some $m \geq 0$, then Eqs. (295), (296) imply desired Eq. (68) of Proposition 4.14. We have thus fully proved Proposition 4.14 assuming Proposition G.5, and it remains to prove the latter.

Proof We prove Proposition G.5 by induction. The base of induction is Statement 2 with $m = 0$ (corresponding to $\zeta > 1$). In the induction step, we either derive Statement 1 for m from Statement 2 for $m-1$, or derive Statement 2 for m from Statement 1 with the same m.

Base of induction: Statement 2 for $m = 0$. Given any $\mathbf{u} \in l^2$, denote $\mathbf{w} = J^{-1}\mathbf{u}$. If $\mathbf{w} \in l^2$, its components satisfy the equations

$$w_1 = u_1, \tag{304}$$

$$n^{-\frac{\nu}{2}}w_n - (\tfrac{n}{n-1})^{\frac{1-(2+\nu)\zeta}{2}}(n-1)^{-\frac{\nu}{2}}w_{n-1} = u_n, \quad n = 2,3,\dots \tag{305}$$

The system can be solved iteratively, starting from w_1 and computing w_n from w_{n-1} using Eq.(305):

$$w_n = (\tfrac{n}{n-1})^{\frac{1+\nu-(2+\nu)\zeta)}{2}}w_{n-1} + n^{\frac{\nu}{2}}u_n \tag{306}$$

$$= (\tfrac{n}{n-1})^{\frac{1+\nu-(2+\nu)\zeta)}{2}}\Big((\tfrac{n-1}{n-2})^{\frac{1+\nu-(2+\nu)\zeta)}{2}}w_{n-2} + (n-1)^{\frac{\nu}{2}}u_{n-1}\Big) + n^{\frac{\nu}{2}}u_n, \tag{307}$$

$$= n^{\frac{-1-(2+\nu)(\zeta-1)}{2}}\sum_{k=1}^{n} k^{\frac{(2+\nu)\zeta-1}{2}}u_k, \qquad n = 2,3,\dots \tag{308}$$

In the special case $\mathbf{u} = \mathbf{e}_1$ we get the explicit solution

$$w_n = n^{\frac{-1-(2+\nu)(\zeta-1)}{2}}, \tag{309}$$

proving desired Eq. (295) for $m = 0$. It is also clear that this $\mathbf{w} \in l^2$ as long as $\zeta > 1$.

Now let $\mathbf{u} = (\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1$. Let us first bound the components u_n, using results of Section G.2.1 with $\mathbf{x} = \mathbf{u}$ and Lemma G.4. First we observe that u_1 is uniformly bounded for all suficiently small ϵ: by Eq. (283) and Lemma G.4, as long as $\zeta > 1$,

$$|u_1| = (\epsilon - \tfrac{h\theta_1^*}{1-h\theta_1^*})^{-1} \tag{310}$$

$$= (\epsilon + 1 - \tfrac{1}{1-h\theta_1^*})^{-1} \tag{311}$$

$$\leq (1 - \tfrac{1}{1+(\nu+2)(\zeta-1)})^{-1} < \infty. \tag{312}$$

Next we obtain a bound on u_n for $n \leq n_0$. Using the definition of θ^* and Lemma G.4,

$$|x_n| = n^{-g}|y_n| \tag{313}$$

$$= n^{-g}|y_1| \prod_{k=1}^{n-1}(1 - h\theta_k^*)^{-1} \tag{314}$$

$$\leq n^{\frac{-1+(2+\nu)\zeta}{2}}|x_1| \prod_{k=1}^{n-1}\Big(1 - \tfrac{(\nu+2)(1-\zeta)}{k}\Big)^{-1} \tag{315}$$

$$= n^{\frac{-1+(2+\nu)\zeta}{2}} O(n^{(\nu+2)(1-\zeta)}) \tag{316}$$

$$= O(n^{\frac{-1-(\nu+2)(\zeta-2)}{2}}). \tag{317}$$

Now using Eq. (308), we get for $n \leq n_0$

$$|((\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1)_n| = n^{\frac{-1-(2+\nu)(\zeta-1)}{2}} \sum_{k=1}^{n} k^{\frac{(2+\nu)\zeta-1}{2}} O(k^{\frac{-1-(\nu+2)(\zeta-2)}{2}}) \tag{318}$$

$$= n^{\frac{-1-(2+\nu)(\zeta-1)}{2}} \sum_{k=1}^{n} O(k^{-1+\nu+2}) \tag{319}$$

$$= n^{\frac{-1-(2+\nu)(\zeta-1)}{2}} O(n^{\nu+2}) \tag{320}$$

$$= O(n^{\frac{-1-(2+\nu)(\zeta-3)}{2}}), \tag{321}$$

which is the desired Eq. (295).

Now consider $n > n_0$. By Lemma G.1, we can assume w.l.o.g. (if necessary, increasing s_0) that $\theta_n^* < -1$ for $n \geq n_0$. Then, with $\mathbf{u} = (\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1$ and h small enough,

$$|u_n| = n^{-g}|u_{n_0}|n_0^g \prod_{k=n_0}^{n-1}(1 - h\theta_k^*)^{-1} \tag{322}$$

$$= O\Big(n^{-g} n_0^{(\nu+2)(1-\zeta)}(1+h)^{n_0-n}\Big) \tag{323}$$

$$= O\Big(h^{(\nu+2)(\zeta-1)+g}(nh)^{-g}(1+h)^{-n}\Big) \tag{324}$$

$$= O\Big(h^{(\nu+2)(\zeta-1)+g} e^{-nh/2}\Big). \tag{325}$$

It follows that for $n > n_0$

$$|(J^{-1}(\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1)_n| \le n^{\frac{-1-(2+\nu)(\zeta-1)}{2}} \sum_{k=1}^{\infty} k^{\frac{(2+\nu)\zeta-1}{2}} |u_k| \tag{326}$$

$$= n^{\frac{-1-(2+\nu)(\zeta-1)}{2}} O\Big(n_0^{\nu+2} + h^{(\nu+2)(\zeta-1)+g} \sum_{k=n_0}^{\infty} k^{\frac{(2+\nu)\zeta-1}{2}} e^{-kh/2}\Big) \tag{327}$$

$$= n^{\frac{-1-(2+\nu)(\zeta-1)}{2}} O\Big(h^{-(\nu+2)} + h^{(\nu+2)(\zeta-1)+2g} \sum_{k=n_0}^{\infty} (kh)^{\frac{(2+\nu)\zeta-1}{2}} e^{-kh/2}\Big) \tag{328}$$

$$= n^{\frac{-1-(2+\nu)(\zeta-1)}{2}} O\Big(h^{-(\nu+2)} + h^{(\nu+2)(\zeta-1)+2g-1}\Big) \tag{329}$$

$$= n^{\frac{-1-(2+\nu)(\zeta-1)}{2}} O(\epsilon^{-1}), \tag{330}$$

which is the desired bound (296).

Induction step: Statement 1 for m from Statement 2 for $m-1$. Note that $(\widetilde{A})^{-1} = (J^\dagger)^{-1}J^{-1}$ and so we can represent

$$\widetilde{A}^{-m}\mathbf{e}_1 = (J^\dagger)^{-1}(J^{-1}\widetilde{A}^{-(m-1)}\mathbf{e}_1), \tag{331}$$

$$\widetilde{A}^{-m}(\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1 = (J^\dagger)^{-1}(J^{-1}\widetilde{A}^{-(m-1)}(\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1). \tag{332}$$

Let us examine the operator $(J^\dagger)^{-1}$. Given any $\mathbf{u} \in l^2$, denote $\mathbf{w} = (J^\dagger)^{-1}\mathbf{u}$. If $\mathbf{w} \in l^2$, its components satisfy the equations

$$n^{-\frac{\nu}{2}}(w_n - (\tfrac{n+1}{n})^{\frac{1-(2+\nu)\zeta}{2}} w_{n+1}) = u_n, \quad n = 1, 2, \dots \tag{333}$$

These equations can be solved iteratively, with w_n expressed via w_{n+1} :

$$w_n = n^{\frac{-1+(2+\nu)\zeta}{2}} \Big(n^{\frac{1+\nu-(2+\nu)\zeta}{2}} u_n + (n+1)^{\frac{1-(2+\nu)\zeta}{2}} w_{n+1}\Big) \tag{334}$$

$$= n^{\frac{-1+(2+\nu)\zeta}{2}} \Big(\sum_{m=n}^{q-1} k^{\frac{-1-(2+\nu)(\zeta-1)}{2}} u_k + q^{\frac{1-(2+\nu)\zeta}{2}} w_q\Big) \tag{335}$$

for any $q > n$. It is convenient to take the limit $q \to \infty$. If we assume that

$$\sum_{k=1}^{\infty} k^{\frac{-1-(2+\nu)(\zeta-1)}{2}} |u_k| < \infty \tag{336}$$

and

$$w_q = o(q^{\frac{-1+(2+\nu)\zeta}{2}}), \tag{337}$$

then we can take this limit, obtaining

$$w_n = n^{\frac{-1+(2+\nu)\zeta}{2}} \sum_{k=n}^{\infty} k^{\frac{-1-(2+\nu)(\zeta-1)}{2}} u_k. \tag{338}$$

In fact, if we just assume condition (336) and define w_n by Eq. (338), these w_n clearly satisfy equations (333) and condition (337). Accordingly, it suffices to only check condition (336).

We now apply this expansion to $\mathbf{u} = J^{-1}\widetilde{A}^{-(m-1)}\mathbf{e}_1$ and $\mathbf{u} = J^{-1}\widetilde{A}^{-(m-1)}(\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1$. Observe first that in both cases condition (336) is fulfilled thanks to induction hypotheses (295), (296) for $m-1$, since they imply

$$k^{\frac{-1-(2+\nu)(\zeta-1)}{2}}|u_k| = k^{\frac{-1-(2+\nu)(\zeta-1)}{2}}O(k^{\frac{-1-(2+\nu)(\zeta-2m+1)}{2}}) \tag{339}$$

$$= O(k^{-1-(2+\nu)(\zeta-m)}) \tag{340}$$

and, by assumption, $\zeta > 2m \geq m$. Taking $\mathbf{u} = J^{-1}\widetilde{A}^{-(m-1)}\mathbf{e}_1$, we obtain desired Eq. (293):

$$(\widetilde{A}^{-m}\mathbf{e}_1)_n = n^{\frac{-1+(2+\nu)\zeta}{2}}\sum_{k=n}^{\infty} k^{\frac{-1-(2+\nu)(\zeta-1)}{2}}(J^{-1}\widetilde{A}^{-(m-1)}\mathbf{e}_1)_k \tag{341}$$

$$= n^{\frac{-1+(2+\nu)\zeta}{2}}\sum_{k=n}^{\infty} O(k^{-1-(2+\nu)(\zeta-m)}) \tag{342}$$

$$= n^{\frac{-1+(2+\nu)\zeta}{2}}O(n^{-(2+\nu)(\zeta-m)}) \tag{343}$$

$$= O(n^{\frac{-1-(2+\nu)(\zeta-2m)}{2}}). \tag{344}$$

Taking $\mathbf{u} = J^{-1}\widetilde{A}^{-(m-1)}(\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1$, in the case $n > n_0$ we obtain desired Eq. (294) by a completely similar argument. In the case $n \leq n_0$ we obtain the desired bound by

$$(\widetilde{A}^{-m}(\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1)_n = n^{\frac{-1+(2+\nu)\zeta}{2}}\sum_{k=n}^{\infty} k^{\frac{-1-(2+\nu)(\zeta-1)}{2}}(J^{-1}\widetilde{A}^{-(m-1)}(\widetilde{A}+\epsilon)^{-1}\mathbf{e}_1)_k \tag{345}$$

$$= n^{\frac{-1+(2+\nu)\zeta}{2}}\Big(\sum_{k=n}^{n_0} k^{\frac{-1-(2+\nu)(\zeta-1)}{2}}O(k^{\frac{-1-(2+\nu)(\zeta-2m-1)}{2}}) \tag{346}$$

$$+ \sum_{k=n_0+1}^{\infty} k^{\frac{-1-(2+\nu)(\zeta-1)}{2}}O(\epsilon^{-1}k^{\frac{-1-(2+\nu)(\zeta-2m+1)}{2}})\Big) \tag{347}$$

$$= n^{\frac{-1+(2+\nu)\zeta}{2}}\Big(\sum_{k=n}^{n_0} O(k^{-1-(2+\nu)(\zeta-m-1)}) + O(\epsilon^{-1}n_0^{-(2+\nu)(\zeta-m)})\Big) \tag{348}$$

$$= n^{\frac{-1+(2+\nu)\zeta}{2}}\Big(O(n^{-(2+\nu)(\zeta-m-1)} - n_0^{-(2+\nu)(\zeta-m-1)}) \tag{349}$$

$$+ O(n_0^{-(2+\nu)(\zeta-m-1)})\Big) \tag{350}$$

$$= n^{\frac{-1+(2+\nu)\zeta}{2}}O(n^{-(2+\nu)(\zeta-m-1)}) \tag{351}$$

$$= O(n^{\frac{-1-(2+\nu)(\zeta-2m-2)}{2}}), \tag{352}$$

where we used the fact that $m \geq 1$ and $\zeta > 2m \geq m+1$.

Induction step: Statement 2 for m from Statement 1 for the same m. Applying again Eq. (308) with $\mathbf{u} = \widetilde{\mathcal{A}}^{-m}\mathbf{e}_1$, we get for $\zeta > 2m+1$ with $m \geq 1$

$$|(J^{-1}\widetilde{A}^{-m}\mathbf{e}_1)_n| \leq n^{\frac{-1-(2+\nu)(\zeta-1)}{2}} \sum_{k=1}^{n} k^{\frac{(2+\nu)\zeta-1}{2}} |u_k| \tag{353}$$

$$= n^{\frac{-1-(2+\nu)(\zeta-1)}{2}} \sum_{k=1}^{n} k^{\frac{(2+\nu)\zeta-1}{2}} O(k^{\frac{-1-(2+\nu)(\zeta-2m)}{2}}) \tag{354}$$

$$= n^{\frac{-1-(2+\nu)(\zeta-1)}{2}} \sum_{k=1}^{n} O(k^{-1+(2+\nu)m}) \tag{355}$$

$$= n^{\frac{-1-(2+\nu)(\zeta-1)}{2}} O(n^{(2+\nu)m}) \tag{356}$$

$$= O(n^{\frac{-1-(2+\nu)(\zeta-2m-1)}{2}}), \tag{357}$$

which is the desired bound (295). The $n \leq n_0$ case of bound (296) is obtained similarly. In the case $n > n_0$ we have

$$|(J^{-1}\widetilde{A}^{-m}\mathbf{e}_1)_n| \leq n^{\frac{-1-(2+\nu)(\zeta-1)}{2}} \Big(\sum_{k=1}^{n_0} k^{\frac{(2+\nu)\zeta-1}{2}} |u_k| + \sum_{k=n_0+1}^{n} k^{\frac{(2+\nu)\zeta-1}{2}} |u_k| \Big) \tag{358}$$

$$= n^{\frac{-1-(2+\nu)(\zeta-1)}{2}} \Big(\sum_{k=1}^{n_0} k^{\frac{(2+\nu)\zeta-1}{2}} O(k^{\frac{-1-(2+\nu)(\zeta-2m-2)}{2}}) \tag{359}$$

$$+ \sum_{k=n_0+1}^{n} k^{\frac{(2+\nu)\zeta-1}{2}} O(\epsilon^{-1} k^{\frac{-1-(2+\nu)(\zeta-2m)}{2}}) \Big) \tag{360}$$

$$= n^{\frac{-1-(2+\nu)(\zeta-1)}{2}} \Big(O(n_0^{(2+\nu)(m+1)}) + O(\epsilon^{-1} \sum_{k=n_0+1}^{n} k^{-1+(2+\nu)m}) \Big) \tag{361}$$

$$= n^{\frac{-1-(2+\nu)(\zeta-1)}{2}} \Big(O(\epsilon^{-1} n_0^{(2+\nu)m}) + O(\epsilon^{-1}(n^{(2+\nu)m} - n_0^{(2+\nu)m})) \Big) \tag{362}$$

$$= n^{\frac{-1-(2+\nu)(\zeta-1)}{2}} O(\epsilon^{-1} n^{(2+\nu)m}) \tag{363}$$

$$= O(\epsilon^{-1} n^{\frac{-1-(2+\nu)(\zeta-2m-1)}{2}}), \tag{364}$$

which is the desired bound (296) for $n > n_0$. This completes the proof of the proposition. ■

Appendix H. Experiments

H.1 Details of experiments

Algorithms. Let us describe details of each of the eight algorithms present in our experiments (see the legend of Figure 6).

For "constant rate GD" and "constant rate HB" we used parameters $\alpha = 1$ and $\beta = 0.9$.

The algorithm "scheduled HB" uses schedule (40) for α_n, β_n with parameters $a = \zeta, b = 0, r = 1$. The "asymptotic scheduled HB" uses asymptotic $n \to \infty$ version of Jacobi schedule

given by the rightmost part of (40), and additionally set $\alpha_n = 1$ (it is rather unnecessary artifact of our experimentation). The difference between limiting values of α_n in "scheduled HB" ($\alpha_n \to 2$) and "asymptotic scheduled HB" ($\alpha_n = 1$) explains the slight advantage of the former in Figure 6.

The algorithm "scheduled GD" uses the schedule based on the roots of the same residual polynomials we used for "scheduled HB", as described in the proof of theorem 4.9 in section F.2. As this schedule continuously "fills" roots of Jacobi polynomials of degrees 2^l, we see respective stair-like structure in figure 6 and spikes in figure 7.

The adaptive algorithms "Steepest Descent" and "basic CG" in our experiments are given by formulas (18) and (21),(22) since we apply them only to quadratic problems.

The "numerically stable CG" algorithm is meant to fix the problems of "basic CG" as we expect the convergence rate for CG to be $\sim n^{-(2+\nu)\zeta}$. In particular, numerical errors accumulate during the run of the algorithm, leading to non-exact placement of the roots of respective residual polynomial $p_n(\lambda)$ at spectral points λ_k. We resolve this issue by introducing some kind of checking procedure on each step. As CG is known to produce a system of orthogonal steps $\Delta\mathbf{w}_n = \mathbf{w}_{n+1} - \mathbf{w}_n$ for quadratic problems, we directly check this orthogonality on each step. Specifically, before making new step $\Delta\mathbf{w}_{n+1}$, we first eliminate all its components along previously made steps $\Delta\mathbf{w}_l, l \leq n$ (which we store during the run of the algorithm). Then, after $\Delta\mathbf{w}_{n+1}$ is made orthogonal to all $\Delta\mathbf{w}_l, l \leq n$, we correct the magnitude of the step to fully eliminate the component of $\mathbf{f}_{n+1} - \mathbf{f}_*$ in this direction. The described procedure is equivalent to formulas (21), (22) in exact arithmetic, but is required for actual implementation of CG to reach convergence rate $\sim n^{-(2+\nu)\zeta}$, as can be seen from experiments in Figure 6.

The MNIST-based quadratic problem. In figures 1 and 6 we took the first $M = 30000$ MNIST images from the usual train subset, and flattened them into $d = 784$–dimensional vectors $\{\mathbf{x}_i\}_{i=1}^M = \mathcal{D}$. Then we normalize each vector using the dataset mean $\mathbf{m} = \frac{1}{M}\sum_i \mathbf{x}_i$ and variance $r^2 = \frac{1}{M}\sum_i \|\mathbf{x_i} - \mathbf{m}\|^2$ by $\mathbf{x}_i \to (\mathbf{x}_i - \mathbf{m})/r$. Then the scalar targets y_i were obtained simply as numerical values $\{0, 1, \ldots, 8, 9\}$ of the digits corresponding to the images $\mathbf{x}_i$. Then, instead of formulating the quadratic problem in parameter space $\mathbf{w}$, where we would need to specify matrix A and vector $\mathbf{b}$, we consider the problem in output space where we need matrix $\widetilde{A}$ and vector $\mathbf{f}_*$. Components of the latter are given simply by our targets y_i, and the matrix $\widetilde{A}$ is obtained by evaluating a kernel $K(\mathbf{x}, \mathbf{x}')$ on our data with $\widetilde{A}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$. For the kernel we take the NTK of infinitely wide shallow ReLU network given by (see e.g. Lee et al. (2019))

$$K(\mathbf{x}, \mathbf{x}') = \frac{\|\mathbf{x}\|\|\mathbf{x}'\|\big(\sin\varphi + 2\cos\varphi(\pi - \varphi)\big)}{2\pi}, \quad \cos\varphi = \frac{\mathbf{x}^T\mathbf{x}'}{\|\mathbf{x}\|\|\mathbf{x}'\|}. \tag{365}$$

For figure 5 we repeat the same procedure but on a full training set of MNIST ($M = 5000$), which was possible due to the availability of additional computational resources at the later times of our work on this paper. Also, in figure 5 we changed the kernel from NTK to sigmoid kernel $K(\mathbf{x}, \mathbf{x}') = \tanh(\mathbf{x}^T\mathbf{x}' + 1)$, which seem to better illustrate the described phenomenon.

The neural network experiment. We consider the standard MNIST classification problem with one-hot encoding of the 10 classes. We consider a simple shallow ReLU network

of width $N = 1000$ with the NTK parametrization. Its function $f(\mathbf{x})$ can be written as

$$f(\mathbf{x}) = \frac{1}{\sqrt{N}} \sum_{l=1}^{N} \mathbf{c}_l \, \mathrm{ReLU}(\mathbf{w}_l^T \mathbf{x} + b_l) \tag{366}$$

where $\mathbf{w}_l \in \mathbb{R}^{784}$, $b_k \in \mathbb{R}$, $\mathbf{c}_l \in \mathbb{R}^{10}$ are the parameters of the neuron l.

Then we train this network on the full MNIST dataset with standard train-test split. Importantly, we don't use mini-batches during training steps, but process the whole train dataset of size $M = 50000$ during optimization. Thus, considering full-batch GD allows us to stay close to our main setting, with the only difference being non-linearity of the model.

H.2 Finding the end of the loss power law region

Let us formulate a general principle allowing to estimate the transition point n_{th} on the loss curve where the power-law region ends. The end of the power-law region in the loss is due to the end of the power law region in the spectral measure asymptotic. In particular, we assume that the power-law asymptotic $\rho[(\lambda_1, \lambda_2)] \sim \lambda_2^\zeta - \lambda_1^\zeta$ holds in the region $\lambda_1, \lambda_2 \gtrsim \lambda_{\mathrm{low}}$ with λ_{low} being (an estimated) end of this power-law region. For synthetic data from figure 6 (a) this would be simply the lowest eigenvalue $\lambda_{\mathrm{low}} = M^{-\nu}$; for the MNIST-based quadratic problem from figure 6 (b-d) we visually set $\lambda_{\mathrm{low}} = 5 \times 10^{-5}$. Finally, n_{th} is simply a step when for a chosen optimization algorithm the region $[0, \lambda_{\mathrm{low}}]$ can no longer be ignored.

Next, define (approximately) a point $\lambda(n)$ as the point where the residual polynomial $p_n(\lambda)$ of considered optimization algorithm starts to significantly deviate from its value at the origin $p_n(\lambda = 0) = 1$, and then is expected to rapidly converge to zero $p_n(\lambda) \to 0$ as $\lambda \gg \lambda(n)$. Then the loss of the algorithm at step n can be estimated as $L_n \sim \rho([0, \lambda(n)])$. Suppose that only the $\rho([\lambda_{\mathrm{low}}, \lambda(n)])$ part of this loss is defined by asymptotic spectral power law, while the $\rho([0, \lambda_{\mathrm{low}}])$ part is unknown. Then the fraction h of "controlled" loss on step n can be estimated as

$$h(n) = \frac{\lambda(n)^\zeta - \lambda_{\mathrm{low}}^\zeta}{\lambda(n)^\zeta} = 1 - \left(\frac{\lambda(n)}{\lambda_{\mathrm{low}}} \right)^{-\zeta} \tag{367}$$

As this fraction reaches some predefined tolerance threshold h_0 (e.g. $h_0 = 0.5$) we can say that the loss no longer follows its power-law and therefore we are at threshold step n_{th}. Formally, n_{th} is defined by the equation

$$h_0 = 1 - \left(\frac{\lambda(n_{\mathrm{th}})}{\lambda_{\mathrm{low}}} \right)^{-\zeta} \tag{368}$$

To actually apply this principle we need to know $\lambda(n)$. Let us find it for the algorithms considered in this work. For algorithms with constant rates α, β the residual polynomial at small λ has large n asymptotic $p_n(\lambda) \sim \exp(-\frac{n\alpha\lambda}{1-\beta})$. Indeed, for $\beta = 0$ we have $p_n(\lambda) = (1 - \alpha\lambda)^n \approx \exp(-\alpha n)$, while for the case with momentum, one needs to use representation (31) together with (154) near $\lambda = 0$ ($z = \frac{1+\beta}{2\sqrt{\beta}}$). Thus we define $\lambda(n) = \frac{1-\beta}{\alpha n}$. Next, for the algorithms with predefined schedules based on Jacobi polynomials we recall asymptotic (185), which says that the polynomial start to deviate from 1 at $\lambda n^2 \sim 1$, therefore $\lambda(n) =$

n^{-2}. Finally, for (stable) Conjugate Gradient method we may assume that $p_n(\lambda)$ is simply $q_{n/2}^{(a,b)}(\lambda/\lambda_{n/2})$ as in the proof of 4.10. Then we again apply Jacobi polynomials asymptotic to find that $\lambda(n) = n^{-\nu-2}$. To summarize, we have established

$$\lambda(n) = \begin{cases} \frac{1-\beta}{\alpha n}, & \text{for constant rate algorithms} \\ \frac{1}{n^2}, & \text{for algorithms based on Jacobi polynomials} \\ n^{-\nu-2}, & \text{for (numerically stable) Conjugate Gradients} \end{cases} \tag{369}$$

Solving (368) with (369) gives

$$n_{\text{th}} = \begin{cases} (1-h_0)^{\frac{1}{\zeta}} \frac{1-\beta}{\alpha \lambda_{\text{low}}}, & \text{for constant rate algorithms} \\ (1-h_0)^{\frac{1}{\zeta}} \frac{1}{\sqrt{\lambda_{\text{low}}}}, & \text{for algorithms based on Jacobi polynomials} \\ (1-h_0)^{\frac{1}{\zeta}} \lambda_{\text{low}}^{-\frac{1}{\nu+2}}, & \text{for (numerically stable) Conjugate Gradients} \end{cases} \tag{370}$$

This result agrees with Table 1 and also admits the following interpretation: the critical values n_{th} approximately correspond to the step numbers at which the order of the loss magnitude approximately matches the value $\lambda_{\text{low}}^{\zeta}$ associated with the measure $\rho((0, \lambda_{\text{low}}])$ under the power-law spectral assumption.

References

Hirotugu Akaike. On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method. *Annals of the Institute of Statistical Mathematics*, 11(1):1–16, 1959.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. 2021. doi: 10.48550/ARXIV.2111.00034. URL `https://arxiv.org/abs/2111.00034`.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021.

Ronen Basri, Meirav Galun, Amnon Geifman, David Jacobs, Yoni Kasten, and Shira Kritchman. Frequency bias in neural networks for input of non-uniform density, 2020. URL `https://arxiv.org/abs/2003.04560`.

Raphaël Berthier, Francis Bach, and Pierre Gaillard. Accelerated gossip in networks of given dimension using jacobi polynomial iterations. *SIAM Journal on Mathematics of Data Science*, 2(1):24–47, 2020a. doi: 10.1137/19M1244822. URL `https://doi.org/10.1137/19M1244822`.

Raphaël Berthier, Francis Bach, and Pierre Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. *arXiv preprint arXiv:2006.08212*, 2020b.

Alberto Bietti. Approximation and learning with deep convolutional models: a kernel perspective, 2021. URL https://arxiv.org/abs/2102.10032.

M Š Birman and M Z Solomjak. Asymptotic behavior of the spectrum of weakly polar integral operators. *Mathematics of the USSR-Izvestiya*, 4(5):1151–1168, oct 1970. doi: 10.1070/im1970v004n05abeh000948. URL https://doi.org/10.1070/im1970v004n05abeh000948.

M.S. Birman and M.Z. Solomjak. *Spectral Theory of Self-Adjoint Operators in Hilbert Space*. Mathematics and its Applications. Springer Netherlands, 2012. ISBN 9789400945869. URL https://books.google.ru/books?id=unPrCAAAQBAJ.

Åke Björck, Tommy Elfving, and Zdenek Strakos. Stability of conjugate gradient and lanczos methods for linear least squares problems. *SIAM Journal on Matrix Analysis and Applications*, 19(3):720–736, 1998.

Blake Bordelon and Cengiz Pehlevan. Learning curves for sgd on structured features, 2021. URL https://arxiv.org/abs/2106.02713.

Helmut Brakhage. On ill-posed problems and the method of conjugate gradients. In *Inverse and ill-posed Problems*, pages 165–175. Elsevier, 1987.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *arXiv preprint arXiv:2006.13198*, 2021.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32:2937–2947, 2019.

Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22, pages 342–350. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2009/file/5751ec3e9a4feab575962e78e006250d-Paper.pdf.

Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime, 2021. URL https://arxiv.org/abs/2105.15004.

J.W. Daniel. *The Approximate Minimization of Functionals*. Prentice-Hall series in automatic computation. Prentice-Hall, 1971. ISBN 9780130438775. URL https://books.google.ru/books?id=kEUZAQAAIAAJ.

Xialiang Dou and Tengyuan Liang. Training neural networks as learning data-adaptive kernels: Provable representation and approximation benefits. *Journal of the American Statistical Association*, 116(535):1507–1520, 2021.

William Feller. *An introduction to probability theory and its applications, Volume 2*, volume 81. John Wiley & Sons, 1991.

JC Ferreira and VA Menegatto. Eigenvalues of integral operators defined by smooth positive definite kernels. *Integral Equations and Operator Theory*, 64(1):61–81, 2009.

Bernd Fischer. *Polynomial based iteration methods for symmetric linear systems*. SIAM, 2011.

Nicolas Flammarion and Francis Bach. From averaging to acceleration, there is only a step-size. In *Conference on Learning Theory*, pages 658–695. PMLR, 2015.

R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *The Computer Journal*, 7(2):149–154, 01 1964. ISSN 0010-4620. doi: 10.1093/comjnl/7.2.149. URL `https://doi.org/10.1093/comjnl/7.2.149`.

V. M. Fridman. On the convergence of methods of steepest descent type. *Usp. Mat. Nauk*, 17(3(105)):201–204, 1962. ISSN 0042-1316.

Sergei Farshatovich Gilyazov and Nataliâ L'vovna Gol'dman. *Regularization of ill-posed problems by iteration methods*, volume 499. Springer Science & Business Media, 2013.

Martin Hanke. Accelerated landweber iterations for the solution of ill-posed equations. *Numerische mathematik*, 60(1):341–373, 1991.

Martin Hanke. Asymptotics of orthogonal polynomials and the numerical solution of ill-posed problems. *Numerical Algorithms*, 11(1):203–213, 1996.

M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 49:409–436, 1952.

M.R. Hestenes. *Conjugate Direction Methods in Optimization*. Stochastic Modelling and Applied Probability. Springer New York, 2012. ISBN 9781461260486. URL `https://books.google.ru/books?id=nc3cBwAAQBAJ`.

Wolfram Research, Inc. Mathematica, Version 13.2. URL `https://www.wolfram.com/mathematica`. Champaign, IL, 2022.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

Hui Jin, Pradeep Kr Banerjee, and Guido Montúfar. Learning curves for gaussian process regression with power-law priors and targets. *arXiv preprint arXiv:2110.12231*, 2021.

William J Kammerer and M Zuhair Nashed. Steepest descent for singular linear operators with nonclosed range. *Applicable Analysis*, 1(2):143–159, 1971.

William J Kammerer and M Zuhair Nashed. On the convergence of the conjugate gradient method for singular linear operator equations. *SIAM Journal on Numerical Analysis*, 9 (1):165–181, 1972.

Leonid Kantorovich and Gleb Akilov. *Functional analysis in normed spaces*. Number 46. Pergamon Press;[distributed in the Western Hemisphere by Macmillan, New York], 1964.

Jovan Karamata. Sur certains 'Tauberian theorems' de M.M. Hardy et Littlewood. 1930.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Dmitry Kopitkov and Vadim Indelman. Neural spectrum alignment: Empirical study. In *International Conference on Artificial Neural Networks*, pages 168–179. Springer, 2020.

MA Krasnoselskii, GM Vainikko, PP Zabreiko, Ya B Rutitskii, and V Ya Stetsenko. Approximate solutions of operator equations, noordhoff, groningen, 1972. *MR*, 52:6515, 1972.

Thomas Kühn. Eigenvalues of integral operators with smooth positive definite kernels. *Archiv der Mathematik*, 49(6):525–534, 1987.

Jonathan Lacotte and Mert Pilanci. Optimal randomized first-order methods for least-squares problems. In *International Conference on Machine Learning*, pages 5587–5597. PMLR, 2020.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In H. Wallach, H. Larochelle, A. Beygelzimer, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/0d1a9651497a38d8b1c3871c84528bd4-Paper.pdf`.

Jaehoon Lee, Samuel S Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, and Jascha Sohl-Dickstein. Finite versus infinite neural networks: an empirical study. *arXiv preprint arXiv:2007.15801*, 2020.

Gérard Meurant and Zdeněk Strakoš. The lanczos and conjugate gradient algorithms in finite precision arithmetic. *Acta Numerica*, 15:471–542, 2006.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Arkadi S Nemirovskiy and Boris T Polyak. Iterative methods for solving linear ill-posed problems under precise information. I. *Izv. Akad. Nauk SSSR. Tekhn. Kibernet.*, (2), 1984a. [In Russian].

Arkadi S Nemirovskiy and Boris T Polyak. Iterative methods of solving linear ill-posed problems with precise information. II.". *Izv. Akad. Nauk SSSR. Tekhn. Kibernet.*, (3), 1984b. [In Russian].

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. In *Dokl. Akad. Nauk SSSR*, volume 269, pages 543–547, 1983.

Atsushi Nitanda and Taiji Suzuki. Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=PULSD5qI2N1`.

Fabian Pedregosa and Damien Scieur. Acceleration through spectral density estimation. In *International Conference on Machine Learning*, pages 7553–7562. PMLR, 2020.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

Boris T. Polyak. *Introduction to Optimization*. Optimization Software, New York, 1987.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Luc Pronzato, Henry P Wynn, and Anatoly A Zhigljavsky. Renormalised steepest descent in hilbert space converges to a two-point attractor. *Acta Applicandae Mathematica*, 67 (1):1–18, 2001.

Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

Klaus Ritter, Grzegorz W Wasilkowski, and Henryk Woźniakowski. Multivariate integration and approximation for random fields satisfying sacks-ylvisaker conditions. *The Annals of Applied Probability*, pages 518–540, 1995.

Ingo Steinwart, Don R Hush, Clint Scovel, et al. Optimal rates for regularized least squares regression. In *COLT*, pages 79–93, 2009.

Gabor Szego. *Orthogonal polynomials*. American Mathematical Society Providence, 4th ed. edition, 1939. ISBN 0821810235.

Gabor Szegö. *Orthogonal Polynomials*. Number v. 23 in American Mathematical Society colloquium publications. American Mathematical Society, 1959. ISBN 9780821889527.

Aditya Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of sgd for least-squares in the interpolation regime. *arXiv preprint arXiv:2102.03183*, 2021.

Maksim Velikanov and Dmitry Yarotsky. Explicit loss asymptotics in the gradient descent training of neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Maksim Velikanov, Denis Kuznedelev, and Dmitry Yarotsky. A view of mini-batch SGD via generating functions: conditions of convergence, phase transitions, benefit from negative momenta. In *The Eleventh International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=bzaPGEllsjE`.

Harold Widom. Asymptotic behavior of the eigenvalues of certain integral equations. *Transactions of the American Mathematical Society*, 109(2):278–295, 1963.

Christopher K Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. Benign overfitting of constant-stepsize sgd for linear regression. *arXiv preprint arXiv:2103.12692*, 2021.