Average Case Analysis and the polynomial view on first order methods

Leonardo Cunha

September 2021

Abstract

The connection between polynomials and first order optimization methods was first explored by [HS+52], and since then several works have built on this connection. In this report, we overview some of these works focusing on the main mathematical ideas and tools they have used. We focus on the average case analysis introduced by [PS20] and present our contributions in the context of acceleration in the average sense for non-strongly convex functions.

These contributions have been made under the supervision of Gauthier Gidel, Fabien Pedregosa, Courtney Paquette and Damien Scieur.

1 Introduction

The analysis of the average complexity of algorithms is an old feature of computer science. Notably, the quicksort algorithm presents worst-case complexity of $O(n^2)$, but average case complexity of $O(n \log n)$ and is empirically faster than HeapSort or MergeSort. Average case complexity also drives much of the decisions made in cryptography [BT06]

The average case analysis of optimization algorithm has stayed an unexplored problem for long because of the ill-defined notion of a distribution over the optimization problems. Recently though, [PS20] has derived a framework to systemically evaluate the complexity of first order methods on distributions of quadratic minimization problems by tying the average of the residuals to the *expected spectral distribution* of the problem, which is a well studied object on Random Matrix Theory.

[Paq+20] has expanded on this work by introducing a new generative model for the problems and deriving the average complexity of the Nesterov Accelerated Method on a particular distribution and showing the strong concentration of the metrics around their expected value.

[SP20] has shown that for a strongly convex problem with eigenvalues supported on a contiguous interval, the optimal average case complexity is asymptotically equal to the one given by the Polyak Heavy Ball method in the worst-case.

In the non-strongly convex case, optimization drastically slows down, as gradient descent presents worst case convergence in $\Theta(\frac{1}{n})$ and Nesterov is $\Theta(\frac{1}{n^2})$, which matches a lower bound on worst case convergence up to a constant factor.

We believe the non strongly convex, worst case scenario is very pessimistic, especially in high dimensions where the problem has enough degrees of freedom to be highly adversarial [citation of the tutorial here+ quotation marks]. Further, as we are dealing with the sublinear convergence of large scale algorithms, accelerated sublinear rates may be the difference between problems that are computationally feasible and those who are not.

A drawback on average case analysis is that it relies, *à priori*, on a much stronger hypothesis, the shape of the *expected spectral distribution* as we'll see, when compared to the worst case analysis, that relies only on the values of the edges of this distribution.

We show that the main aspects of the convergence, which from an optimization point view are the asymptotic rates, are determined by a low dimensional characterization of the distribution. Parametrizing the distributions in terms of the **concentration** of the eigenvalues around the edges, allow us to effectively determine the precise rates for all continuous distributions supported in an interval $]0, L[$,

We can then compare algorithms robustly, allowing an effective choice of algorithms under a more realistic lack of information on the eigenspectra.

In our contributions, we first introduce the Generalized Chebyshev Method, and show the it's rates of GD, Nesterov under the concentration hypothesis, then compare these algorithms on synthetic and real data.

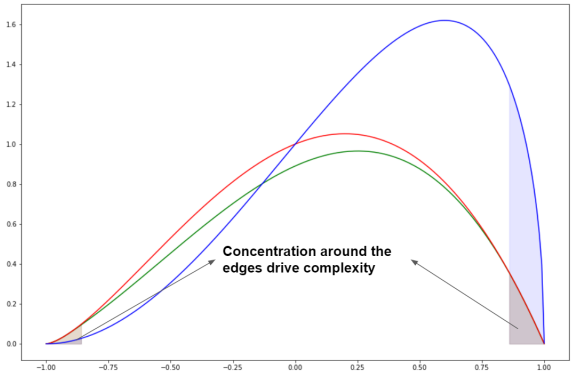

Figure 1: The average case rates for non-strongly problems is determined by the eigenvalue concentration around the edges of the support

Organization In sections 2 we introduce some works built on the connection between first order methods and polynomials. Section 3 overviews [PS20; SP20; DPS20; Paq+20; PP21; Paq+21] which constitute the main body of work on the average case analysis. Sections 4,5 and 6 center on our contributions.

2 Polynomial Abstraction

In this section we introduce the connection between polynomials and first order methods, highlighting how the Chebyshev Semi-Iterative Method [GV61] and the Polyak Heavy Ball method [Pol64], and the average-case analysis framework for random quadratic problems [PS20].

The main result is Theorem 3.1, which relates the expected error with other quantities that will be easier to manipulate, such as the residual polynomial. This is a convenient representation of an optimization method that will allow us in the next section to pose the problem of finding an optimal method as a best approximation problem in the space of polynomials.

Let $\boldsymbol{H} \in \mathbb{R}^{d\times d}$ be a random symmetric positive-definite matrix and $\boldsymbol{x}^\star \in \mathbb{R}^d$ a random vector. These elements

determine the following (random) quadratic minimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^d}\left\{f(\boldsymbol{x})\stackrel{\text{def}}{=}\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}^\star)^\top\boldsymbol{H}(\boldsymbol{x}-\boldsymbol{x}^\star)\right\}. \tag{OPT}$$

Our goal is to quantify the expected error $\mathbb{E}\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|$, where $\boldsymbol{x}_t$ is the t-th update of a first-order method starting from $\boldsymbol{x}_0$ and $\mathbb{E}$ is the expectation over the random $\boldsymbol{H},\boldsymbol{x}_0,\boldsymbol{x}^\star$.

Remark 1. *The expectation in the expected error $\mathbb{E}\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^2$ is over the inputs and not over any randomness of the algorithm, as would be common in the stochastic literature. In this paper we will only consider deterministic algorithms.*

To solve OPT, we will consider *first-order methods*. These are methods in which the sequence of iterates $\boldsymbol{x}_t$ is in the span of previous gradients, i.e.,

$$\boldsymbol{x}_{t+1}\in\boldsymbol{x}_0+\mathbf{span}\{\nabla f(\boldsymbol{x}_0),\dots,\nabla f(\boldsymbol{x}_t)\}. \tag{1}$$

This class of algorithms includes for instance gradient descent and momentum, but not quasi-Newton methods, since the preconditioner could allow the iterates to go outside of the span. Furthermore, we will only consider *oblivious* methods, that is, methods in which the coefficients of the update are known in advance and don't depend on previous updates. This leaves out some methods that are specific to quadratic problems like conjugate gradient.

From First-Order Method to Polynomials. There is an intimate link between first order methods and polynomials that simplifies the analysis on quadratic objectives. Using this link, we will be able to assign to each optimization method a polynomial that determines its convergence. The next Proposition gives a precise statement:

Proposition 2.1. *[HS+52] Let $\boldsymbol{x}_t$ be generated by a first-order method. Then there exists a polynomial P_t of degree t such that $P_t(0)=1$ that verifies*

$$\boldsymbol{x}_t-\boldsymbol{x}^\star=P_t(\boldsymbol{H})(\boldsymbol{x}_0-\boldsymbol{x}^\star)\ . \tag{2}$$

Remark 2. *If the first-order method is further a **momentum method**, i.e.:*

$$\boldsymbol{x}_{t+1}=\boldsymbol{x}_t+m_t(\boldsymbol{x}_t-\boldsymbol{x}_{t-1})+h_t\nabla f(\boldsymbol{x}_t) \tag{3}$$

We can determine the polynomials by the recurrence $P_0=1$ and:

$$P_{t+1}(\lambda)=(1+m_t)P_t(\lambda)+h_t\lambda P_t(\lambda)-m_tP_{t-1}(\lambda) \tag{4}$$

We note that while most popular F.O.M's can be posed a momentum method, Nesterov cannot.

Following [Fis96], we will refer to this polynomial P_t as the *residual polynomial*.

Example 1 (Chebyshev Semi-iterative method)**.** *We can easily derive a worst-case optimal algorithm for the $||\boldsymbol{x}_t-\boldsymbol{x}^\star||$ metric, when H is restricted to be a positive definite matrix. From the proposition above, the task of designing an optimal method is equivalent to solving a polynomial approximation problem:*

Theorem 3 ([MG16]). *The solution to :*

$$P_t^\star = \underset{\deg(P_t)=t, P_t(0)=1}{\operatorname{argmin}} \max_{\lambda\in[\ell,L]} P_t^2(\lambda) \tag{5}$$

Is:

$$P_t^{cheb}(\lambda) := \frac{T_t(\sigma(\lambda))}{T_t(\sigma(0))}$$

Where $\sigma : [\ell, L] \to [-1, 1]$*:*

$$\sigma(\lambda) = \frac{L+\ell}{L-\ell} - \frac{2}{L-\ell}\lambda \tag{6}$$

Is the linear map between the domains and T_t *is the Chebyshev Polynomial of the First Kind.*

Proof. The Chebyshev polynomial T_t are such that $\exists \lambda_1 \leq \ldots \leq \lambda_{t+1}$, $\arg\max_{x\in[-1,1]} |T_t(x)| = |T_t(\lambda_i)| = 1$, and $P_t(\lambda_i) = (-1)^{t+i+1}$, that is the polynomials oscilate between -1 and 1 which are it's maximal points in absolute value.

The same is true for $T_t(\sigma(.))$ and the interval $[\ell, L]$ from the definition of σ.

Suppose R_t is a residual polynomial of same degree as P_t and smaller absolute value, and $Q_t := P_t - R_t$. By this assumption $\operatorname{sign}(Q_t(\lambda)) = (-1)^{t+i+1}$, i.e. it oscilates above and below 0 at λ_i, and thus Q_t has at least $t+1$ zeros, which is an absurd because it has degree at most t. □

We note this method can be derived as a special case of the one we propose in section 4
As $\max_{x\in[-1,1]} T_t(x) = 1, \forall t$*, the convergence rate is determined by* $T_t(\sigma(0))^{-2} = T_t(\frac{L+\ell}{L-\ell})^{-2}$*. If* $\ell > 0$ $\sigma(0) < -1$ *and luckily we have an expression for* T_t *outside* $[-1, 1]$*:*

$$T_t(\lambda) = \frac{1}{2}[(\lambda + \sqrt{\lambda^2 - 1})^t + (\lambda - \sqrt{\lambda^2 - 1})^t] \tag{7}$$

. Thus, considering only the first term on the right:

$$T_t\left(\frac{L+\ell}{L-\ell}\right) \geq \frac{1}{2}\left(\frac{L+\ell}{L-\ell} + \sqrt{\left(\frac{L+\ell}{L-\ell}\right)^2 - 1}\right)^t \tag{8}$$

$$= \frac{1}{2}\left(\frac{\sqrt{L}+\sqrt{\ell}}{\sqrt{L}-\sqrt{\ell}}\right)^t = \left(1 - \Theta\left(\frac{1}{\sqrt{\kappa}}\right)\right)^t \tag{9}$$

Which shows the dependence on the square root of the condition number κ *instead of the condition itself as is the case of gradient descent.*

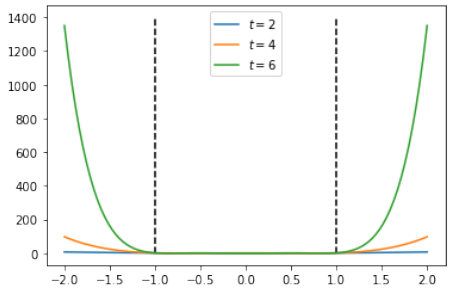

Figure 2: Chebyshev polynomials of the first-kind T_t of different degrees. Note the exponential behaviour, wrt t, outside $[-1, 1]$

The difference between the behaviour outside and inside $[-1, 1]$ *gives an intuitive insight into the contrast of linear and sublinear convergence under strong convexity and without it respectively. This exponential behaviour outside of an interval is not particular to the Chebyshev polynomials, so in general if* $\ell > 0$, $P_t(\sigma(0))$ *behaves like an exponential and like a polynomial otherwise.*

Example 2 (Polyak Heavy Ball)**.** *Consider the method, and the polynomials* P_t*, generated with constant momentum and gradient steps, i.e. equation 4 with* $m_t = m$, $h_t = h$*. Then, polynomials* P_t *can be written [Ped20]:*

$$P_t(\lambda) = m^{t/2}\left(\frac{2m}{1+m}T_t(\sigma(\lambda)) - \frac{m-1}{1+m}U_t(\sigma(\lambda))\right) \tag{10}$$

Where T_t *and* U_t *are the Chebyshev polynomials of the first and second kind respectively and:*

$$\sigma(\lambda) = \frac{1}{2\sqrt{m}}(1 + m - h\lambda) \tag{11}$$

We want to upper bound the expression in eq. 10. We know that for $x \in [-1, 1]$*:*

$$|T_t(x)| \leq 1 \qquad \text{and} \qquad |U_t(x)| \leq t + 1$$

To use this property, we need to have $\max_{\lambda\in[\ell,L]} |\sigma(\lambda)| \leq 1$*, which is equivalent to:*

$$\frac{(1-\sqrt{m})^2}{h} \leq \ell \leq \frac{(1+\sqrt{m})^2}{h} \tag{12}$$

$$0 \leq m \leq 1 \qquad \text{and} \qquad h \leq \frac{2(1+m)}{L} \tag{13}$$

Inside this region we have:

$$\max_{\lambda\in[\ell,L]} P_t(\lambda) \leq m^{t/2}\left(1 + \frac{1-m}{1+m}t\right) =: q_t \tag{14}$$

To minimize the upper bound q_t*, we'll want to minimize* m *while staying inside the region. This will give:*

$$m = \left(\frac{\sqrt{L}-\sqrt{\ell}}{\sqrt{L}+\sqrt{\ell}}\right)^2 \leq 1 \qquad \text{and} \qquad h = \left(\frac{2}{\sqrt{L}+\sqrt{\ell}}\right)^2$$

Which are the parameters of the Polyak Heavy Ball method. This gives:

$$r_t^{Polyak} \leq \left(\frac{\sqrt{L}-\sqrt{\ell}}{\sqrt{L}+\sqrt{\ell}}\right)^2 \left(1+\frac{2\sqrt{L\ell}}{L+\ell}t\right) \tag{15}$$

Which corresponds asymptotically to the square root of the condition number acceleration proved in [Pol64].

Note that while this is one of the tighter bounds we have on Polyak, a strict exponential decay was proven recently in [WLA21].

2.1 Acceleration without momentum

The polynomial abstraction also shows we can achieve "optimal" convergence **without** momentum.

It is easy to see that the polynomial associated to the method:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - h_t \nabla f(\boldsymbol{x}_t) \tag{16}$$

Is:

$$P_t(\lambda) = (1-h_t\lambda)(1-h_{t-1}\lambda)\dots(1-h_0\lambda) \tag{17}$$

So it's roots are $1/h_i$.

We also know that the roots of the t-th degree of the residual chebyshev polynomial P_t^{Cheb} are [Ped21]:

$$\lambda_i = \frac{1}{2}(L+\ell) + \frac{1}{2}(L-\ell)\cos\left(\frac{\pi(i+1/2)}{t}\right) \tag{18}$$

By setting $h_i = \lambda_i$, for $i \leq t$, the polynomial in eq. 17 matches the Chebyshev polynomial at iteration t. This means we can match, for a fixed number iterations, the optimal worst case performance without momentum. This is called the Young Method [You53]

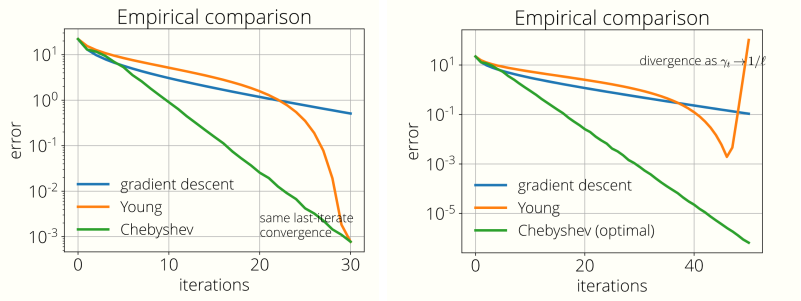

Figure 3: Empirical comparison between Young, Chebyshev method and gradient descent. Note that the vanilla Young method diverges in practice for low values of ℓ Taken from [Ped21]

This is an unstable algorithm in practice, as the number of iterations grow and ℓ goes to 0, because the final steps will converge to $\frac{1}{\ell}$, and small numerical errors in ℓ will lead to divergence.

[LF71], and more recently [AGZ21] have shown that this instability can be reduced with a proper permutation of the Chebyshev steps.

Definition 1. *Let $(\sigma_{2^t})_{t\in\mathbb{N}}$ be a family of tuples of variable length s.t:*

$$\sigma_1 = [1] \tag{19}$$
$$\sigma_{2t} = interlace(\sigma_t, 2t+1-\sigma_t) \tag{20}$$
$$interlace([a_1,\dots,a_n],(b_1,\dots,b_n]) = [a_1,b_1,a_2,b_2,\dots,a_n,b_n] \tag{21}$$

For instance:

$$\sigma_8 = [1,8,4,5,2,7,3,6] \tag{22}$$

We define the fractal Chebyshev schedule as $(h_{\sigma_T(t)})_{t=1,\dots,T}$, where $h_t = \frac{1}{\lambda_t}$, with λ as in eq. 18

If we consider the gradient containing additive noise, may it be from a stochastic or numerical source:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - h_t \nabla f(\boldsymbol{x}_t) + \eta_t \tag{23}$$

The error at each step depends on "partial polynomials":

$$\boldsymbol{x}_t - \boldsymbol{x}^\star = P_{0:t-1}(\boldsymbol{H})(\boldsymbol{x}_0 - \boldsymbol{x}^\star) + \sum_{t'=2}^{t} P_{t':t-1}(\boldsymbol{H})\eta_{t'-1} \tag{24}$$

$$P_{i:j}(\boldsymbol{H}) = \prod_{\tau=i}^{j}(I - h_t\boldsymbol{H}) \tag{25}$$

[AGZ21] shows that for the Fractal Chebyshev Polynomials:

$$\sum_{t'=s}^{t} \sup_{\lambda\in[\ell,L]} |P_{t':t}| = \mathcal{O}(\hat{\kappa}^{1+\frac{1}{\ln 4}\log\hat{\kappa}}) = o(\hat{\kappa}^{1,73}) \tag{26}$$

Where $\hat{\kappa} = \frac{L}{\ell}$. This bound is independent on t, so if we consider a bound like $||\eta_t|| < \epsilon$, we have a time independent bound on the error. Though a lower bound on the regular schedule unstability is not shown, eq. 26 and experiments suggest the much greater stability of the fractal schedule

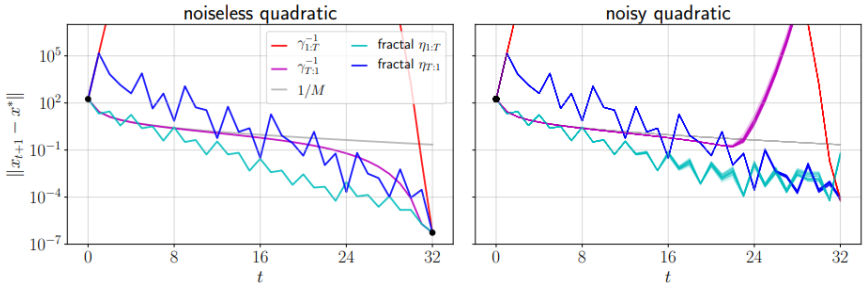

Figure 4: Comparison between stabilized and unstabilized versions of Young's method. Taken from [AGZ21]

Cyclic step sizes [Gou+21; Oym21] showed provable acceleration instead with cyclical gradient steps in the setting where:

$$\Lambda := \mathbf{Sp}(\boldsymbol{H}) \subset [\ell_1, L_1] \cup [\ell_2, L_2] \tag{27}$$

[Gou+21] does so with the concept of *equioscilation*, which is a generalization of the ideas in the proof of theorem 3.

Definition 2. *A polynomial P_t equioscilates on a set Λ if $P_t(0) = 1$ and $\exists \lambda_1 < \ldots < \lambda_{t+1}$ s.t.:*

$$P_t(\lambda_i) = (-1)^i \max_{x \in \Lambda} |P_t(x)| \tag{28}$$

Consider the problem:

$$\min_{P_t \in \mathbb{R}_t[X]} \max_{x \in \Lambda} |P_t(x)| \quad \text{s.t.} \quad P_t(0) = 1 \tag{29}$$

For a general set $\Lambda \subset \mathbb{R}$.We can consider equivalently the solution to:

$$\max_{P_t \in \mathbb{R}_t[X]} P_t(0) \quad \text{s.t.} \quad \sup_{x \in \Lambda} |P_t(x)| = 1 \tag{30}$$

And divide the polynomial by it's value on 0.

Let σ_K be a degree K polynomial s.t $\sup_{x \in \Lambda} |\sigma_K(x)| = 1$. We consider solutions of the type:

$$P_t(\lambda) = \frac{T_n(\sigma_K(\lambda))}{T_n(\sigma_K(0))} \tag{31}$$

With $t = Kn$. Note that we are only able to define these polynomials with degrees in intervals of size K. This is at the origin of the cyclic step sizes.

It's easy to see that the minimum solution in this class happens when σ_K *equioscilates* in Λ, according to def. 2. [Gou+21] further shows this solution is global, in the sense of eq. 29 if $\sigma_K^{-1}([-1, 1]) = \bar{\Lambda}$.

In general such a solution is not analytically described: we can find an approximation to the optimal K degree polynomial, but we can't show that the degree itself is optimal
When Λ is as in eq. 27 and $L_2 - \ell_2 = L_1 - \ell_1$, we can find a globally optimal solution with:

$$\rho = \frac{L_2 + \ell_1}{L_2 - \mu_1}, R = \frac{\ell_2 - L_1}{L_2 - \ell_1} \tag{32}$$

$$m = \left(\frac{\sqrt{\rho^2 - R^2} - \sqrt{\rho^2 - 1}}{\sqrt{1 - R^2}} \right)^2 \tag{33}$$

$$\sigma_2(\lambda) = 2 \left(\frac{1 + m}{2\sqrt{m}} \right) \left(1 - \frac{\lambda}{L_1} \right) \left(1 - \frac{\lambda}{\ell_2} \right) - 1 \tag{34}$$

We now consider the recurrence generating the optimal polynomials:

$$P_{2(n+1)}(\lambda) := \frac{T_{n+1}(\sigma_k(\lambda))}{T_{n+1}(\sigma_k(0))} = 2\sigma_k(\lambda) \frac{T_n(\sigma_k(\lambda))}{T_n(\sigma_k(0))} \underbrace{\frac{T_n(\sigma_k(0))}{T_{n+1}(\sigma_k(0))}}_{a_n} - \frac{T_{n-1}(\sigma_k(\lambda))}{T_{n+1}(\sigma_k(0))} \underbrace{\frac{T_{n-1}(\sigma_k(0))}{T_{n+1}(\sigma_k(0))}}_{b_n} \tag{35}$$

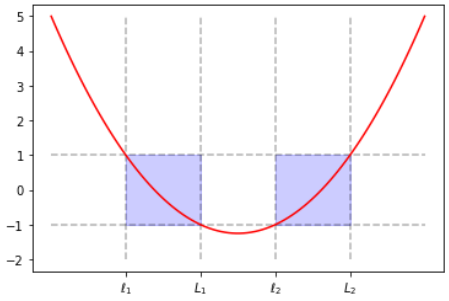

Figure 5: Graph for σ_2. Note the optimality conditions are observed

We can derive a method that tracks these polynomials. [Gou+21] shows that an iteration of the type:

$$c = \sqrt{\frac{\rho^2 - R^2}{1 - R^2}} \tag{36}$$

$$\omega_t = \left(1 - \frac{1}{4c^2}\omega_{t-1}\right)^{-1} \tag{37}$$

$$h_t = \begin{cases} \frac{\omega_t}{L_1} & \text{if } t = 2k \\ \frac{\omega_t}{\ell_2} & \text{if } t = 2k+1 \end{cases} \tag{38}$$

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - h_t \nabla f(\boldsymbol{x}_t) + (\omega_t - 1)(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) \tag{39}$$

With $\omega_0 = 2$, follows $\boldsymbol{x}_t - \boldsymbol{x}^\star = P_t(\boldsymbol{H})(\boldsymbol{x}_0 - \boldsymbol{x}^\star)$. Though we can define an all-time optimal method, [Gou+21] focuses on an asymptotic variant.
Lemma 4 tells us a_n, b_n have a limit a_∞, b_∞, The polynomials who follow recurrence:

$$\tilde{P}_t(\lambda) = 2a_\infty \sigma_K(\lambda)\tilde{P}_{t-K} - b_\infty \tilde{P}_{t-2K}(\lambda) \tag{40}$$

Are then produced by an iteration like eq. 39, when $\omega_t = \omega_\infty = 1 + m$ and have the same (optimal) asymptotic rate. This is called the cyclic heavy ball with (cyclic) steps $h_1 = \frac{1+m}{L_1}, h_2 = \frac{1+m}{\ell_2}$

The optimal asymptotic rate for a given Λ, and the optimal σ_K, is given by:

$$(\sigma_K(0) - \sqrt{\sigma_K(0)^2 - 1})^{\frac{1}{K}} \tag{41}$$

These results allow us to show significant acceleration over the standard Heavy Ball method in some regimes of Λ. Indeed when:

$$\Lambda = [\ell, (1+\gamma)\ell] \cup [L - \gamma\ell, L] \tag{42}$$

The rate is $(1 - \Theta(1))^t$ w.r.t $\kappa = \frac{L}{\ell}$, i.e. it is independent of κ.

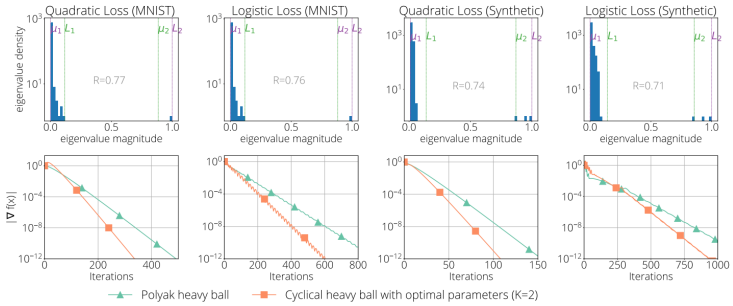

Figure 6: Experimental results contrasting Polyak Heavy Ball with Cyclic heavy Ball. The different slopes in semi log scale indicate different asymptotic convergence rates. Taken from [Gou+21]

3 Average-Case Analysis

The previous section shows the usefulness of the "polynomial abstraction". It turns out this is especially true when considering *distributions of problems*, that can be represented by letting $\boldsymbol{H}$ be a random matrix.

A convenient way to describe a random matrix is through its *empirical spectral distribution*, which we now define.

Definition 3. *(**Weighted/Expected spectral distribution**). Let $\boldsymbol{H}$ be a random matrix with eigenvalues $\{\lambda_1, \dots, \lambda_d\}$. The **empirical spectral distribution** of $\boldsymbol{H}$, called $\mu_{\boldsymbol{H},\,\alpha}$, is the probability measure*

$$\mu_{\boldsymbol{H}} \overset{def}{=} \frac{1}{d}\sum_{i=1}^{d}\delta_{\lambda_i}, \tag{43}$$

where δ_{λ_i} is the Dirac delta, a distribution equal to zero everywhere except at λ_i and whose integral over the entire real line is equal to one.

*Since $\boldsymbol{H}$ is random, the empirical spectral distribution $\mu_{\boldsymbol{H}}$ is a random measure (a random variable in the space of measures). Its expectation over $\boldsymbol{H}$ is called the **expected spectral distribution** (e.s.d.) and we denote it*

$$\mu \overset{def}{=} \mathbb{E}_{\boldsymbol{H}}[\mu_{\boldsymbol{H}}]\,. \tag{44}$$

We can link the e.s.d. of $\boldsymbol{H}$ to the convergence of a first order method on the distribution of $\boldsymbol{H}$ In the following we'll consider $x_0 - x^\star$ and $\boldsymbol{H}$ to be independent, with $x_0 - x^\star$ sampled isotropically. This isotropic hypothesis is not necessary and we could derive a similar analysis for more general distribution of $x_0 - x^\star$.

Theorem 3.1. *Let $\boldsymbol{x}_t$ be generated by a first-order method associated to the polynomial P_t, μ the e.s.d. of H and $\mathbb{E}[(\boldsymbol{x}_0 - \boldsymbol{x}^\star)(\boldsymbol{x}_0 - \boldsymbol{x}^\star)^T] = R^2\boldsymbol{I}$ Then we can write the convergence metrics at time step t as:*

$$\mathbb{E}[\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2] = R^2 \int P_t^2(\lambda)d\mu(\lambda) \tag{45}$$

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] = R^2 \int P_t^2(\lambda)\lambda d\mu(\lambda) \tag{46}$$

$$\mathbb{E}[||\nabla f(\boldsymbol{x}_t)||_2^2] = R^2 \int P_t^2(\lambda)\lambda^2 d\mu(\lambda) \tag{47}$$

This shows that the polynomials are an excellent abstraction: in the following we'll refer directly to the polynomials associated to a given method and omit the R^2 term associated to the initialization. We'll simply refer to metric l as the metric associated to the added λ^l term, i.e. the gradient norm is metric $l = 2$

Theorem 3.1 works as an infinite dimensional limit for the convergence rates. Under some assumptions, we can show it to be a *deterministic* limit

Proposition 3.1. *Let $\boldsymbol{x}_0$ and $\boldsymbol{x}^\star$ in $\mathbb{R}^d$ as s.t:*

$$\mathbb{E}[(\boldsymbol{x}_0 - \boldsymbol{x}^\star)(\boldsymbol{x}_0 - \boldsymbol{x}^\star)^T] = R^2 \boldsymbol{I} \tag{48}$$

$$\mathbb{E}[(\boldsymbol{x}_0 - \boldsymbol{x}^\star)_i^4] = O(\frac{1}{d^2}) \tag{49}$$

And $\boldsymbol{H}$ a random matrix s.t. it's empirical spectral distribution *converges weakly and in probability:*

$$\mu_{\boldsymbol{H}} \xrightarrow[Pr]{d\to\infty} \mu \tag{50}$$

And it's largest eigenvalue is also convergent:

$$\lambda^+_{\boldsymbol{H}} \xrightarrow[Pr]{d\to\infty} \lambda^+ \tag{51}$$

Then we have:

$$||\nabla f(\boldsymbol{x}_k)||^2 \xrightarrow[Pr]{d\to\infty} R^2 \int \lambda^2 P_k^2(\lambda) d\mu \tag{52}$$

This has been stated slightly differently in [Paq+20]. This shows that the if the data generation process scales neatly with the dimension, the gradient, and other metrics with a slight modification to the proof, concentrates around it's expected value.

The hypothesis are relatively weakly, and are supported for instance by the Wishart matrices we present in subsection 4.2

This framework is further linked to the field of **orthogonal polynomials** by the following proposition which gives a construction of an optimal method w.r.t. a given distribution

Proposition 3.2. *[PS20] Let $P_t^\star$ be defined as*

$$P_t^\star := \arg\min_{P_t(0)=1} \int P_t^2(\lambda) \lambda^l d\nu(\lambda) \tag{53}$$

then $(P_t^\star)$ is the family of residual orthogonal polynomials w.r.t. to $\lambda^{l+1} d\nu$

As the theory of orthogonal polynomials is often stated in terms of distributions supported in $[-1, 1]$, we'll often consider σ as in eq. 6 and write our *method* polynomial P_t as $P_t(\lambda) = \tilde{P}_t(\sigma(\lambda))$, where $\tilde{P}_t$ is a classic polynomial, i.e. Chebyshev of the First Kind as in example 1.

This theorem further implies that the optimal first-order method is a momentum method as Favard's theorem [MÁ01] tells us the residual orthogonal polynomials w.r.t. a given distribution can be related through a **three term recurrence**:

$$P_{t+1}(\lambda) = a_t P_t(\lambda) + b_t \lambda P_t(\lambda) + (1 - a_t) P_{t-1}(\lambda) \tag{54}$$

Following remark 2, the optimal method is derived from this recurrence as:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + (a_t - 1)(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) + b_t \nabla f(\boldsymbol{x}_t) \tag{55}$$

3.1 Universality in the Strongly Convex Case

[SP20] showed universal properties in the strongly convex case.

Let's consider a continuous measure μ supported in $[\ell, L]$,with $\ell > 0$, corresponding to $\lambda^{l+1}d\nu$ from proposition 3.2, with P_t it's *residual* orthogonal polynomials, following eq. 54, and p_t the *orthonormal* polynomials to the $[-1, 1]$ shift of μ , i.e. $P_t(\lambda) = \frac{p_t(\sigma(\lambda))}{p_t(\sigma(0))}$, and p_t follows:

$$\alpha_t p_t(\lambda) = (\lambda - \beta_t)p_{t-1}(\lambda) - \alpha_{t-1}p_{t-2}(\lambda) \tag{56}$$

We can rewrite this equation as:

$$P_t(\lambda) = (\sigma(\lambda) - \beta_{t-1})\frac{p_{t-1}(\sigma(\lambda))}{\alpha_t p_t(\sigma(0))} - \alpha_{t-1}\frac{p_{t-2}(\sigma(\lambda))}{\alpha_t p_t(\sigma(0))} \tag{57}$$

We can identify the terms such that:

$$1 - a_t = -\frac{\alpha_{t-1}}{\alpha_t}\frac{p_{t-2}(\sigma(0))}{p_t(\sigma(0))} \tag{58}$$

$$b_t = -\frac{2}{\alpha_t(L-\ell)}\frac{p_{t-1}(\sigma(0))}{p_t(\sigma(0))} \tag{59}$$

Luckily, the *orthonormal polynomials* have useful, and universal, asymptotic properties. In appendix C we use their *weak* asymptotics. Here we'll use their *ratio* asymptotics:

Lemma 4 ([Rah77; MNT85])**.** *Let p_t be the orthonormal polynomial family w.r.t a continuous function supported and strictly positive in $[-1, 1]$.*
For $|\lambda| > 1$:

$$\lim_{t\to\infty}\frac{p_t(\lambda)}{p_{t-1}(\lambda)} = \lambda + \sqrt{\lambda^2 - 1} \tag{60}$$

And letting α_t, β_t as in eq. 56:

$$\lim_{t\to\infty}\alpha_t = \frac{1}{2} \qquad \lim_{t\to\infty}\beta_t = 0 \tag{61}$$

This means that:

$$\lim_{t\to\infty}\frac{p_{t-1}(\sigma(0))}{p_t(\sigma(0))} = \left(\frac{L+\ell}{L-\ell} + \sqrt{\left(\frac{L+\ell}{L-\ell}\right)^2 - 1}\right)^{-1} = \frac{\sqrt{L}-\sqrt{\ell}}{\sqrt{L}+\sqrt{\ell}} \tag{62}$$

$$\lim_{t\to\infty}(1 - a_t) = -\left(\lim_{t\to\infty}\frac{\alpha_{t-1}}{\alpha_t}\right)\left(\lim_{t\to\infty}\frac{p_{t-2}(\sigma(0))}{p_t(\sigma(0))}\right) = -\left(\frac{\sqrt{L}-\sqrt{\ell}}{\sqrt{L}+\sqrt{\ell}}\right)^2 \tag{63}$$

$$\lim_{t\to\infty} b_t = -\frac{2}{L-\ell}\left(\lim_{t\to\infty}\alpha_t^{-1}\right)\left(\lim_{t\to\infty}\frac{p_{t-1}(\sigma(0))}{p_t(\sigma(0))}\right) = -\left(\frac{2}{\sqrt{L}-\sqrt{\ell}}\right)^2 \tag{64}$$

This shows that the momentum and gradient steps converge to those of the Polyak Heavy Ball method. Just compare the above expressions to those in example 2.

We can further show that the asymptotic rates are universal. If P_t is the residual orthogonal w.r.t. $\lambda^{l+1}d\nu$, [PS20] has shown that:

$$\int P_t^2(\lambda)\lambda^l d\nu(\lambda) = \int P_t(\lambda)\lambda^l d\nu(\lambda) =: r_t \tag{65}$$

We can write for r_t:

$$r_t = \int (a_t + \lambda b_t) P_{t-1}(\lambda) + (1 - a_t) P_{t-2}(\lambda) \lambda^l d\nu(\lambda) = a_t \int P_{t-1} \lambda^l d\mu + (1 - a_t) \int P_{t-2} \lambda^l d\mu \tag{66}$$

Because P_{t-1} is orthogonal w.r.t. a constant function and measure $\lambda^{l+1} d\mu$. Thus we have a recurrence:

$$r_t = a_t r_{t-1} + (1 - a_t) r_{t-2}$$
$$r_1 = 1 + b_1 \int \lambda^{l+1} d\mu \qquad r_0 = 1$$

The Poincaré-Perron [Pit02] states that if $\lim_{t\to\infty} a_t = a_\infty$, and $a_t \notin \{0, 1\} \forall t$, then the recurrence has two solution $\{(r_t^1), (r_t^2)\}$ s.t.:

$$\limsup_{t\to\infty} \sqrt[t]{r_t^i} = |\lambda_i| \tag{67}$$

Where λ_i are the roots of the equation $\lambda^2 - a_\infty \lambda - (1 - a_\infty) = 0$, which are 1 and $1 - a_\infty$. Since this method is average case optimal by hypothesis, this rate must be better than any worst case rate, the Polyak rate for instance which we've shown to be $\left(\frac{\sqrt{L}-\sqrt{\ell}}{\sqrt{L}+\sqrt{\ell}}\right)^2 = 1 - a_\infty$.

Thus the asymptotic rate can't be one, and is indeed the same one given by Polyak.

This shows that, at least in an asymptotic sense, in the strongly convex setting the average case is not too different from the worst case and that we cannot do much better than the Polyak method.

3.2 Average Case for Bilinear Games

[DPS20] has extended the average case framework for the solution of an equation:

$$F(\boldsymbol{x}) := \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}^\star) = 0 \tag{68}$$

Where $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ is a real normal matrix, i.e. $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{A}^T\boldsymbol{A}$,. The solution set to this equation is an affine subspace $\mathcal{X}^\star$, s.t. we write:

$$\text{dist}(\boldsymbol{x}, \mathcal{X}^\star) = \min_{\boldsymbol{v}\in\mathcal{X}^\star} ||\boldsymbol{x} - \boldsymbol{v}||^2 \tag{69}$$
$$\mathcal{X}^\star = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}^\star) = 0\} \tag{70}$$

. This generalizes the zero-sum, bilinear game setting, which can be posed as minimax problem:

$$\min_{\boldsymbol{\theta}_1} \max_{\boldsymbol{\theta}_2} l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) := (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_1^\star)^T \boldsymbol{M} (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_2^\star) \tag{71}$$

Where $\boldsymbol{M} \in \mathbb{R}^{d_x \times d_y}$. That's because the vector field to this game [Bal+18] is:

$$F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \begin{bmatrix} \nabla_{\boldsymbol{\theta}_1} \ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ -\nabla_{\boldsymbol{\theta}_2} \ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & \boldsymbol{M} \\ -\boldsymbol{M}^\top & 0 \end{bmatrix}}_{=\boldsymbol{A}} \Big(\underbrace{\begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{bmatrix}}_{=\boldsymbol{x}} - \underbrace{\begin{bmatrix} \boldsymbol{\theta}_1^\star \\ \boldsymbol{\theta}_2^\star \end{bmatrix}}_{=x^\star} \Big) = \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}^\star). \tag{72}$$

In this case the normality of $\boldsymbol{A}$ is ensured. A first order method in this case corresponds to a sequence $(\boldsymbol{x}_t)$ s.t.:

$$\boldsymbol{x}_i \in \boldsymbol{x}_0 + \textbf{span}(\{F(\boldsymbol{x}_j)\}_{j=0}^{i-1}) \tag{73}$$

In this sense, the first order method also has an associated polynomial family P_t, with $P_t(0) = 1$.
Let $\mathbb{E}[(\boldsymbol{x}_0 - \boldsymbol{x}^\star)(\boldsymbol{x}_0 - \boldsymbol{x}^\star)^T] = \frac{R^2}{d}\mathbf{I}$, notice the different scaling, and $\boldsymbol{x}_0, \boldsymbol{x}^\star$ independent to A. Through a similar procedure to theorem 3.1, we can show:

$$\mathbb{E}[\text{dist}(\boldsymbol{x}_t, \mathcal{X}^\star)] = R^2 \int_{\mathbb{C}\setminus\{0\}} |P_t|^2(\lambda) d\mu(\lambda) \tag{74}$$

Where μ is the expected spectral distribution of A.
The main difference with the minimization setting is the domain of the integral. The fact that $\boldsymbol{A}$ is not symmetric, in contrast with $\boldsymbol{H}$ in the minimization setting, adds complex eigenvalues to be considered. Notably, theorem 8 is not valid in this scenario and a complex equivalent is needed:

Lemma 5 ([Van97]). *If μ is a positive Borel measure in $\mathbb{C}$, the minimum of $\int_{\mathbb{C}} |P_t(\lambda)|^2 d\mu(\lambda)$ over the residual polynomials P_t of degree at most t is given by:*

$$P^\star(\lambda) = \frac{\sum_{k=0}^t \phi_k(\lambda)\phi_k(0)^\star}{\sum_{k=0}^t |\phi_k(0)|^2} \tag{75}$$

$$\int_{\mathbb{C}} |P^\star(\lambda)|^2 d\mu(\lambda) = \frac{1}{\sum_{k=0}^t |\phi_k(0)|^2} \tag{76}$$

With ϕ_k the orthonormal polynomials w.r.t. μ

Another difference brought by the complex eigenvalues is the inexistence of an equivalent of Favard's theorem, so in general the optimal solution, as given by the above lemma, cannot be written as in eq. 4.
This issue can be avoided though if we consider μ to have a radial symmetry.

Proposition 3.3. *Let μ be supported on the disk of center C and radius R $D_{C,R}$. $R < C$ s.t. :*

$$d\mu(C + re^{i\theta}) = \frac{1}{2\pi} d\theta d\mu_R(r)$$

Where μ_R is a real measure supported in $[0, R]$. then the optimal method is given by:

$$\boldsymbol{y}_t = \boldsymbol{y}_{t-1} - \frac{1}{C}F(\boldsymbol{y}_{t-1}), \quad \beta_t = \frac{C^{2t}}{K_{t,R}^2} \quad B_t = B_{t-1} + \beta_{t-1} \tag{77}$$

Where $K_{t,r} = \sqrt{2\pi \int_0^R r^{2t} d\mu_R(r)}$. Moreover:

$$\mathbb{E}[\text{dist}(\boldsymbol{x}_t, \mathcal{X}^\star)] = \frac{1}{B_t}$$

[DPS20] further shows this method outperforms gradient descent by a constant factor, in average. It also shows the method converges to an asymptotic version depending only on R and C, much like [SP20] found Polyak as an universal optimal asymptotic.

In the bilinear game setting, given by eq. 72, we can reduce the problem of finding the optimal method to the "real case" of minimizing the function $\frac{1}{2}||F(\boldsymbol{x})||^2$.
If $d_x \leq d_y$ we can directly relate the e.s.d's of A and $\boldsymbol{M}\boldsymbol{M}^T$.

$$\mu_{\boldsymbol{A}}(i\lambda) = \left(1 - \frac{2}{1 + r^{-1}}\right)\delta_0(\lambda) + \frac{2|\lambda|}{1 + r^{-1}} \mu_{\boldsymbol{M}\boldsymbol{M}^T}(\lambda)^2 \tag{78}$$

Where $r = \frac{d_x}{d_y}$. Notably this setting is simpler than the general normal matrix one because the eigenvalues are purely imaginary.
As $\nabla\left(\frac{1}{2}||F(\boldsymbol{x})||^2\right) = -\boldsymbol{A}^2(x - x^\star)$, the optimal method w.r.t. to $\frac{1}{2}||F(\boldsymbol{x})||^2$ is given by:

$$Q_t^\star := \underset{Q_t(0)=1}{\operatorname{argmin}} \int Q_t(\lambda)^2 d\mu_{-\boldsymbol{A}^2(\lambda)} \tag{79}$$

. The optimal polynomial w.r.t. $\mu_{\boldsymbol{A}}$ is given by $P_{2t}^\star(\lambda) = Q_t^\star(-\lambda^2)$ [DPS20]. This leads to the method:

$$\boldsymbol{g}_t = F(\boldsymbol{x}_t - F(\boldsymbol{x}_t)) - F(\boldsymbol{x}_t) \qquad \left(= \frac{1}{2}||\nabla F(\boldsymbol{x}_t)||^2\right) \tag{80}$$

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - h_{t+1}\boldsymbol{g}_t + m_{t+1}(\boldsymbol{x}_{t-1} - \boldsymbol{x}_t) \tag{81}$$

Where h_t, m_t are the optimal gradient step and momentum for $\boldsymbol{M}\boldsymbol{M}^T$.
The parity of the optimal polynomials evidenced above is what gives origin to the intermediary term $\boldsymbol{g}_t$ and the extragradient-like update.

3.3 Stochastic Average Case

[Paq+21; Paq+21] has shown that under a similar framework to [PS20], we can derive the limiting dynamics of *Stochastic* Gradient descent and stochastic momentum.
Following [Paq+20], these works suppose a slightly different generative process for the problems:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} ||\boldsymbol{A}x - b||^2 \tag{82}$$

$$b = \boldsymbol{A}\tilde{\boldsymbol{x}} + \eta \tag{83}$$

With η some noise independent to $\boldsymbol{x}_0 - \tilde{\boldsymbol{x}}$ and $\boldsymbol{A}$, $\mathbb{E}[\eta\eta^T] = \tilde{R}^2\boldsymbol{I}$. It can represent some 'natural' noise or the case $b \notin \mathbf{span}(\boldsymbol{A})$. The SGD iteration corresponds to:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{\gamma}{n}\boldsymbol{A}^T\boldsymbol{P}_k(\boldsymbol{A}\boldsymbol{x}_k - b), \qquad \boldsymbol{P}_k = \sum_{i\in B_k} e_i e_i^T \tag{84}$$

Where B_k is the batch at iteration k and γ is the step-size. [Paq+21] embeds this iteration into a continuous time process, the stochacity due to the incomplete gradient becomes brownian noise in this embedding. Letting $\hat{nu}_{i,t}$ be the i-th spectral component at iteration t, i.e.:

$$\boldsymbol{A} = \boldsymbol{V}\Sigma\boldsymbol{V}^T \tag{85}$$

$$\hat{\nu}_{i,k} = (V^T(\boldsymbol{x}_t - \tilde{x}))_i \tag{86}$$

Then the Doob-Meyer decomposition [Bas96] gives us:

$$\hat{\nu}_{i,t}^2 = \hat{\nu}_{i,0}^2 + \int_0^t \mathcal{A}_{i,s}\,\mathrm{d}\mathbf{s} + \mathcal{M}_{j,t} \tag{87}$$

$$\mathcal{A}_{i,s} := \lim_{\epsilon\to 0} \frac{\mathbb{E}(\hat{\nu}_{i,t+\epsilon} - \hat{\nu}_{i,t}|\mathcal{F}_t)}{\epsilon} \tag{88}$$

Where $\mathbf{s}$ is the Brownian Noise, $\mathcal{M}_{j,\cdot}$ is a martingale and $\mathcal{A}_{i,\cdot}$ is an a.s. negative random process.

If the relative batch size $\frac{|B_k|}{n}$ goes to 0 when n and the dimension go to infinity, the expression for $\mathcal{A}$ can be greatly simplified and we can show, in a strong enough sense, that if $\frac{d}{n} \to r$, then $f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) \underset{n\to\infty}{\to} \psi_0(t)$, where ψ_0 is the solution to:

$$\psi_0(t) = \frac{R}{2}h_1(t) + \underbrace{\frac{\tilde{R}}{2}(rh_0(t) + (1-r))}_{\text{Noise } \eta} + \underbrace{\gamma^2 r \int_0^t h_2(t-s)\psi_0(s)ds}_{\text{incomplete gradient}} \tag{89}$$

$$h_k(t) = \int_0^\infty \lambda^k e^{-2\gamma t\lambda} d\mu(\lambda) \tag{90}$$

[PP21] generalizes this by proposing *dimension adjusted* versions of Nesterov, sDANA, and Polyak's methods, sDAHB, s.t. $\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] \to \psi(t)$ and:

$$\psi(t) = F(T) + \int_0^t \mathcal{I}(t-s)\psi(t)\,\mathrm{d}s \tag{91}$$

When we let the parameters of sDANA and sDAHB be independent of the problem's dimension we retrieve the original methods, and the same equation as for SGD. This shows that in the high dimensional stochastic regime, the changes in dynamics provided by the momentum, with fixed coefficients, are too shy and don't improve convergence.

Equations 89 and 91 are of Volterra type, more specifically of renewal type, and have well studied properties. Indeed for a renewal equation of the form:

$$Z = z + F \star Z \tag{92}$$

We consider the Laguerre transform of F:

$$\hat{F}(\beta) = \int_0^\infty e^{\beta x} F(dx) = 1 \tag{93}$$

The Maltusian Exponent [Asm08] of eq. 92 is the solution to $\hat{F}(\beta^\star) = 1$. $\beta^\star$, if it is different than 0, determines the exponential behaviour of the solution Z.
In the case of eq. 89, we have $F = \gamma^2 r h_2(t)$. To have a convergent solution, we need the Malthusian exponent to be greater than 0, or differently stated, $\hat{F}(0) < 1$:

$$\gamma^2 r \int_0^\infty \int_0^\infty \lambda^2 e^{-2\gamma t\lambda}\,\mathrm{d}\mu(\lambda)\,\mathrm{d}t = \frac{\gamma r}{2}\int_0^\infty \lambda\,\mathrm{d}\mu(\lambda) < 1 \tag{94}$$

[Paq+21] further shows that $\beta^\star \leq \ell$, where ℓ is the left edge of μ, so we can define the optimal step size to be s.t. $\beta^\star = \ell$. This $\gamma^\star$ acts further as phase transition on the dynamics of SGD, i.e $\psi(t;\gamma)$ behaves differently whether γ is greater or lower than $\gamma^\star$
We further note that ψ_0 can be derived in close form for the Marchenko-Pastur measure because of the connections between the Laplace and the Stieltjes transforms.

4 Optimization methods

4.1 Nesterov Accelerated Gradient

The Nesterov acceleraion [Nes03] is considered one of the most important methods in optimization. In the convex case it provably matches a lower bound of $\mathcal{O}(\frac{1}{t^2})$ complexity.

The nesterov iteration, when tuned for eigenvalues in $[0, L]$ is [Paq+20]:

$$\boldsymbol{x}_{t+1} = \boldsymbol{y}_t - \frac{1}{L}\nabla f(\boldsymbol{y}_t) \tag{95}$$

$$\boldsymbol{y}_{t+1} = \boldsymbol{x}_{t+1} + \beta_t(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t) \qquad \beta_t = \frac{t}{t+3} \tag{96}$$

It is a first order method and thus has an associated polynomials, but cannot be written as momentum method as in eq. 4. Indeed, letting $\alpha = 1/L$, we can show the Nesterov polynomials P_t follow the recurrence:

$$P_t(\lambda) = (1+\beta_{t-1})(1-\alpha\lambda)P_t(\lambda) - \beta_{t-1}(1-\alpha\lambda)P_{t-1}(\lambda) \tag{97}$$

By considering the generating function of the R polynomials defined as $R_t((1-u)^{\frac{1}{2}} = P_t(u)(1-u)^{\frac{-t}{2}}$:

$$\mathrm{G}(v, x) = \sum_{k=0}^{\infty} R_k(v)x^k \tag{98}$$

[Paq+20] show that the polynomials corresponding to recurrence 97 can be written as:

$$P_t(\lambda) = \frac{2(1-\alpha\lambda)^{\frac{t+1}{2}}}{t\alpha\lambda}\left(\sqrt{1-\alpha\lambda}L_t(\sqrt{1-\alpha\lambda}) - L_{t+1}(\sqrt{1-\alpha\lambda})\right) \tag{99}$$

4.2 Generalized Chebyshev and Laguerre methods

Being able to write the rates in terms of the *expected spectral distribution* ties the average case framework to the field of *random matrix theory*. Indeed, because of results from this field, certain e.s.d's are considered more natural than others. We illustrate this and motivate following considerations with:

Proposition 4.1 (Marchenko Pastur Theorem)**.** *Let X_n be a $m \times n$ random matrix with X_{ij} i.i.d with variance σ^2 and $Y_n = \frac{1}{n}X_nX_n^T$. Let μ_n be the expected spectrum of Y_n, then, as $n \to \infty$, $\frac{m}{n} \to r$:*

$$\mu_n \xrightarrow{weakly} \max(0, 1-\frac{1}{r})\delta_0 + \mu_{MP}$$

$$d\mu_{MP}(\lambda) = \frac{1}{2\pi\sigma^2}\frac{\sqrt{(\lambda^+ - \lambda)(\lambda - \lambda^-)}}{r\lambda}$$

With $\lambda^+ = \sigma^2(1+\sqrt{r})^2$, $\lambda^- = \max(0, \sigma^2(1-\sqrt{r})^2)$

The Marchenko Pastur distribution μ_{MP} can be considered a natural first model for e.s.d's as it arises universally from i.i.d. noise in the matrix entries. Note there's no specific distribution of X_{ij} considered.
When $r = 1 d\mu_{MP} \propto \lambda^{-1/2}\sqrt{\lambda^+ - \lambda}$, Though practical e.s.d's diverge from it, the concentration near 0 is often verified. [PS20] first derived the optimal method wrt. μ_{MP}, and [Paq+20] derived Nesterov's rates under it. As we are mainly concerned with being robust, a natural step is to consider the Beta weights $d\mu(\lambda) \propto \lambda^{\xi}(L-\lambda)^{\tau}$ but we are mainly interested in distributions with similar concentrations near 0, i.e. $\xi \approx -1/2$.
The optimal method w.r.t. μ and metric l is associated to a shifted Jacobi polynomial $\tilde{P}_t^{\alpha,\beta}$ with $\beta = \xi + l + 1, \alpha = \tau$. When $\alpha = \beta = -1/2$ we retrieve the *Chebyshev Method* [HS+52], so we call this the Generalized Chebyshev Method.
We'll also consider the Laguerre method, which is optimal w.r.t. $d\mu(x) = \frac{x^{\alpha}e^{-x}}{\Gamma(\alpha+1)}$, taking α as a parameter. This method is proposed to optimize non-smooth functions
Both these methods are generalizations of one that have been proposed in [PS20]. We derive the coefficients associated to these methods in appendix B.

Remark 6. *The Generalized Chebyshev also takes the largest eigenvalue L as a parameter, but the rates we'll show are robust to an overestimation of L*

5 Robust Average Case Rates

We will establish our assumption over the spectral distributions. It effectively allows us to parametrize all of our distributions of interest in a way that characterize the asymptotic convergence, diving them into equivalence classes .

Assumption 1. *We will write $\nu_{\tau,\xi}$ for a continuous distribution supported in $(0, L]$ s.t. $\nu'_{\tau,\xi}(x) > 0$ for $x \in [0, L]$, $d\nu_{\tau,\xi} = \Theta(\lambda^{\xi})$ near 0 and $d\nu_{\tau,\xi} = \Theta((L-\lambda)^{\tau})$ near L.*

We argue this is a much more mild assumption to be made than the exact spectral distribution and it covers any distribution modeling a smooth convex problem.
The ξ works as a measure of how close we are to the worst case scenario, as it approaches -1. Samples in finite dimension, of distributions with high values of ξ, will work as strongly convex functions in practice.

We show that $\nu_{\tau,\xi}$ indeed behaves like an equivalence class when considering the asymptotics of the convergence of a Jacobi method: only the concentrations near the edge matter. We do this by singling out from each of these classes the beta distributions for which we can compute the rates, then show the rates to be the same inside $\nu_{\tau,\xi}$.

Theorem 7. *A Generalized Chebyshev Method with parameters (α, β) applied to a problem with e.s.d. as in assumption 1 has rates:*

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] \sim L \cdot C_{1,\nu}^{\alpha,\beta} \begin{cases} t^{-1-2\beta} & \text{if } \alpha < \tau + 1/2 \text{ and } \beta < \xi + 3/2 \\ t^{-2(\xi+2)} \log t & \text{if } \alpha = \tau + 1/2 \text{ and } \beta = \xi + 3/2 \\ t^{2(\max\{\alpha-\beta-\tau,-\xi-1\}-1)} & \text{if } \alpha > \tau + 1/2 \text{ or } \beta > \xi + 3/2 \end{cases} \tag{100}$$

$$\mathbb{E}[||\nabla f(\boldsymbol{x}_t)||_2^2] \sim L^2 \cdot C_{2,\nu}^{\alpha,\beta} \begin{cases} t^{-1-2\beta} & \text{if } \alpha < \tau + 1/2 \text{ and } \beta < \xi + 5/2 \\ t^{-2(\xi+3)} \log t & \text{if } \alpha = \tau + 1/2 \text{ and } \beta = \xi + 5/2 \\ t^{2(\max\{\alpha-\beta-\tau,-\xi-2\}-1)} & \text{if } \alpha > \tau + 1/2 \text{ or } \beta > \xi + 5/2 \end{cases} \tag{101}$$

Where $C_{\nu}^{\alpha,\beta}$ is a distribution dependent constant.

Theorem 8 shows that a proper choice of α, β can indeed make the Jacobi polynomial asymptotically optimal w.r.t. to any $\nu_{\tau,\xi}$.

Theorem 8. *Let ν follow assumption 1. The optimal asymptotic rate for $\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)]$ is $t^{-2(\xi+2)}$ and is attained by the Chebyshev Method with parameters $(\tau, \xi + 2)$.*
The optimal rate for $\mathbb{E}[||\nabla f(\boldsymbol{x}_t)||_2^2]$ is $t^{-2(\xi+3)}$ and is attained by the Chebyshev Method with parameters $(\tau, \xi + 3)$.

For the function value ($l = 1$), we find rates that approach t^{-2} as $\xi \to -1$, showing the worst case as a limit (over the considered distribution) on the average case.

Remark 9. *We can contextualize our results on the field of orthogonal polynomials as asymptotics on the values of the Christoffel Functions at 0 [Tot05]:*

$$\lambda_t(\mu, x) = \inf_{P_t(x):=1, \deg(P_t) \le t} \int P_t^2 d\mu = \Big(\sum_{k=0}^{t} p_k(x; \mu)^2 \Big)^{-1} \tag{102}$$

For the equivalence classes $\nu_{\tau,\xi}$

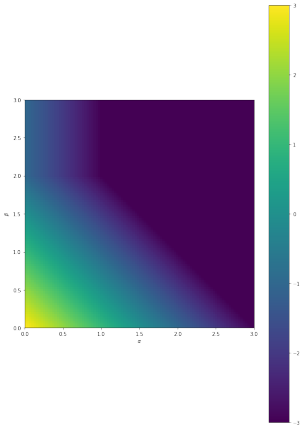

Figure 7: Colormap for the function value rates for the Marchenko Pastur distribution. The color represents the negative exponent, e.g. lower is better.

We remark that the above theorems imply that, at least asymptotically, the Jacobi method is robust for a suboptimal choice of parameters up to 1/2 below the optimal choice of β and infinitely above.
For completeness, we also derive worst case rates for the Jacobi method:

Proposition 5.1. *Let f be a convex, L-smooth quadratic function. Then, For the Generalized Chebyshev Method with parameters (α, β) we have:*

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) \leq C_1 L \begin{cases} t^{2(\alpha-\beta)} & \text{if} \quad \alpha > \beta - 1 \\ t^{-1-2\beta}, & \text{if} \quad \alpha \leq \beta - 1 \quad \beta \leq \frac{1}{2} \\ t^{-2}, & \text{if} \quad \alpha \leq \beta - 1 \quad \beta \geq \frac{1}{2} \end{cases} \tag{103}$$

$$||\nabla f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)|| \leq C_2 L^2 \begin{cases} t^{2(\alpha-\beta)} & \text{if} \quad \alpha > \beta - 2 \\ t^{-1-2\beta}, & \text{if} \quad \alpha \leq \beta - 2 \quad \beta \leq 3/2 \\ t^{-4}, & \text{if} \quad \alpha \leq \beta - 2 \quad \beta \geq 3/2 \end{cases} \tag{104}$$

For the function value ($l = 1$) and a reasonable choice of α, β this means effectively that the worst case rates are t^{-2}

[Nes03] has shown that the Nesterov matches up to a a constant factor a lower bound on the worst case complexity of non strongly convex problems. A natural question is if this performance would translate to good average case rates.
We'll extend [Paq+20] proof for the Nesterov method under the MP distribution

Theorem 10. *Let ν as in assumption 1, then, for the Nesterov method:*

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] \sim C'_{1,\nu} \begin{cases} t^{-2(\xi+2)} & \text{if } \xi < -1/2 \\ t^{-3} \log t & \text{if } \xi = -1/2 \\ t^{-(\xi+7/2)} & \text{if } \xi > -1/2 \end{cases} \qquad \mathbb{E}[||\nabla f(\boldsymbol{x}_t)||_2^2] \sim C'_{2,\nu} t^{-(\xi+9/2)} \tag{105}$$

The optimality gap of nesterov is $t\xi + l - 1/2$, when $\xi + l > 1/2$, $\log t$ when $\xi + l = 1/2$ and 0 otherwise. This shows that Nesterov is almost optimal when the concentrations near 0 are relatively high, i.e. low ξ

Theorem 11. *Let ν as in assumption 1, then for gradient descent:*

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] = \Theta(t^{-(\xi+2)}) \qquad \mathbb{E}[||\nabla f(\boldsymbol{x}_t)||_2^2] = \Theta(t^{-(\xi+3)}) \tag{106}$$

Observe, for the function value, that we find the t^{-2} rates for Nesterov and t^{-1} for Gradient Descent when $\xi \to -1$

(τ, ξ)/Method	Chebyshev $(\frac{1}{2}, \frac{5}{2})$	Chebyshev $(\frac{1}{2}, \frac{3}{2})$	Nesterov	G.D.
$(\frac{1}{2}, \frac{1}{2})$	t^{-5}	t^{-4}	t^{-4}	$t^{-\frac{5}{2}}$
$(\frac{1}{2}, \frac{-1}{2})$	t^{-3}	t^{-3}	$t^{-3}\log t$	$t^{-\frac{3}{2}}$

Table 1: Comparison of asymptotic rates for the function-value for different methods an (τ, ξ) values

Lastly, we consider the optimal rates for a Gamma distribution.

Theorem 12. *Let $\alpha > -1$ and μ_α be a Gamma distribution, i.e. $d\mu_\alpha(x) = \frac{x^\alpha e^{-x}}{\Gamma(\alpha+1)}$. The optimal rates are given by the Laguerre method of appropriate tuning and:*

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] = \Theta(t^{-(\alpha+2)}) \tag{107}$$

Note that this result does not have the same universality of the others because of the non-compacity of the distribution's support.
These rates are contrasted to the worst case lower bound on the optimization of non-smooth functions by first order methods, which gives:

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \geq \frac{C}{\sqrt{t}}$$

These rates are not found when $\alpha \to -1$, indicating that the worst case is specially pessimistic in this scenario.

6 Experiments

We synthesize the e.s.d's either by means of the Marchenko Pastur that enables us to simulate (τ, ξ) values of $(1/2, 3/2)$ and $(1/2, 5/2)$. Other values are simulated by sampling from the corresponding Beta distribution and applying a random rotation to the diagonal matrix comprising those samples.
Because we work with the asymptotic rates on the infinite dimensional case, i.e. we take $d \to \infty$ and then $t \to \infty$, they are representative of the experiments on the regime $t < d$.
Our theoretical estimates for the Nesterov method and Gradient Descent are precise under the approximate range $-1 < \xi < 0$, distributions with higher ξ need many samples otherwise they behave as strongly convex functions.

The same isn't true for the GCM. If $\beta < \beta^\star$ or ξ is low the empirical findings diverge from the theoretical. We believe this is due to numerical instability under these regimes as the metrics also have much larger variance than in the other regimes. This shown in appendix D.

The GCM with $\beta > \beta^\star$ perform corresponding to the theory, and it's non-asymptotically very close to the performance of $\beta^\star$. High values of β also perform very well on non-synthetic data, suggesting in practice we should use these values.

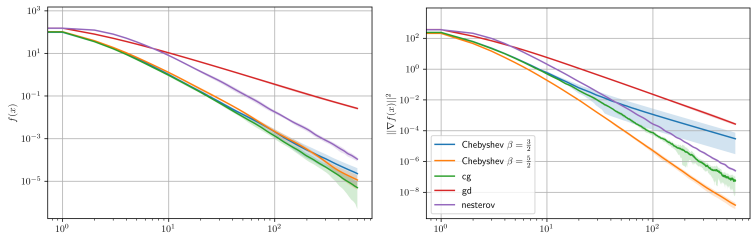

Figure 8: Rates for a synthetic problem, simulating the Marchenko Pastur distribution. *Left:* function value. Note that both tunings of the GCM achieve performance in function value very close to the one of Conjugate Gradient, which is optimal for every draw of the problem. *Right*: gradient norm

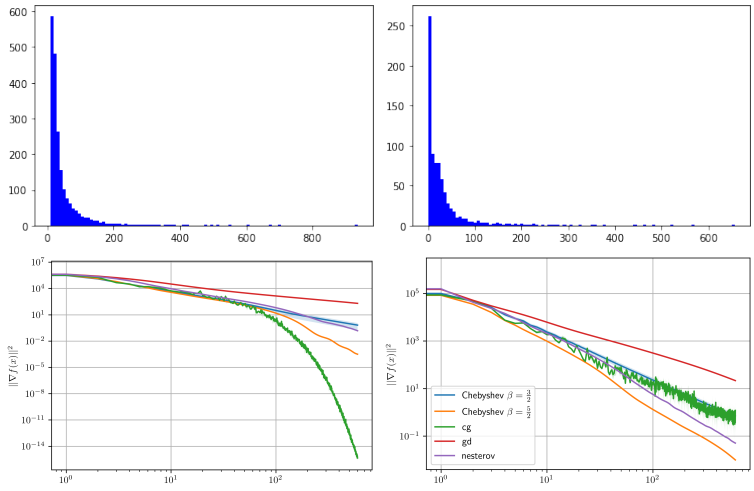

Figure 9: Spectrum and gradient norm rates for regression problems. *Left:* CIFAR-10 Inception features *Right*: MNIST features. Here we choose to compare gradient norms as the minimum function value is not known. The properly tuned GCM achieves remarkable performance under these non-synthetic spectrum's.

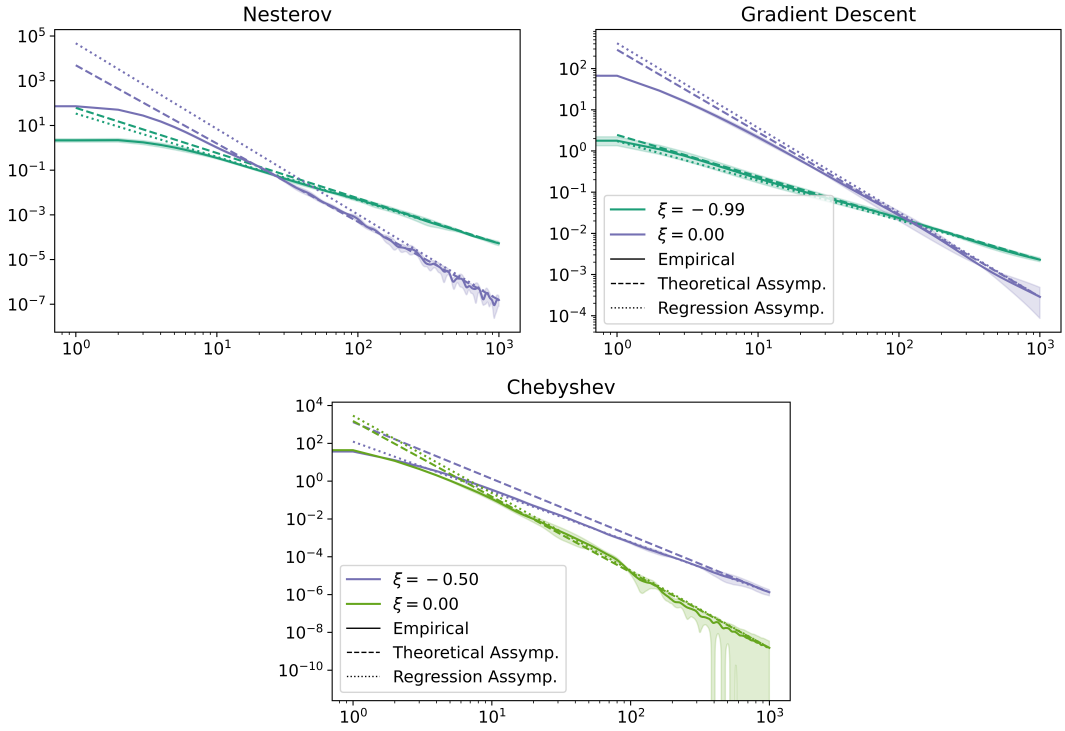

Figure 10: Comparison between experiments run on synthetic Beta distribution and theoretical asymptotic. Y-axis is function value

7 Conclusion

From a practical standpoint, we have proposed f.o.m., the Generalized Chebyshev method which shows remarkable performance on real world datasets.
From a theoretical standpoint, we've shown the Nesterov method, which is easier to implement and more numerically stable, to be very close to the asymptotic optimum under some natural assumptions over the data, namely a concentration of eigenvalues around 0 similar to the Marchenko-Pastur measure. This adds to the theoretical understanding of why Nesterov acceleratio is such a good method in pratice.

We have shown that the average case asymptotic convergence of the better known f.o.m. depend only on the concentrations of the problem's eigenvalue near the edges. We have shown these rates to be representative of the practical performance of the algorithms, and retrieve the classical worst case rates as limits of the average case

Finally, we have characterized the optimum average case convergence under the scenario analogous to the three usual hypothesis in worst-case convergence analysis.

Regime	Worst case	Average case
Smooth Strongly convex	$\left(1-\Theta(\frac{1}{\sqrt{\kappa}})\right)^t$	$\left(1-\Theta(\frac{1}{\sqrt{\kappa}})\right)^t$
Smooth convex	$\frac{1}{t^2}$	$\frac{1}{t^{2\xi+4}}$
Non-smoooth convex	$\frac{1}{\sqrt{t}}$	$\frac{1}{t^{\alpha+2}}$

Table 2: Comparison between function value worst-case and average case convergence. κ is the condition number in the smooth strongly convex case. In the smooth convex case ξ is the concentration of eigenvalues around 0 and in the non-smooth case we consider $d\mu \propto \lambda^\alpha e^{-x}$

We consider analysing the Generalized Chebyshev methods and the average convergence in problems beyond quadratics minimization the main follow-up direction to our work. We conjecture one can show local convergence, and local average case rates, for general convex functions by a similar method to [WLA21], which has shown the size of the cnvergent neighborhood for the Polyak method for general strongly convex smooth functions.

References

[MG16] Wladimir Markoff and J Grossmann. "Über Polynome, die in einem gegebenen Intervalle möglichst wenig von Null abweichen". In: *Mathematische Annalen* 77.2 (1916), pp. 213–258.

[HS+52] Magnus Rudolph Hestenes, Eduard Stiefel, et al. *Methods of conjugate gradients for solving linear systems*. Vol. 49. 1. NBS Washington, DC, 1952.

[You53] David Young. "On Richardson's method for solving linear systems with positive definite matrices". In: *Journal of Mathematics and Physics* 32.1-4 (1953), pp. 243–255.

[GV61] Gene H Golub and Richard S Varga. "Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods". In: *Numerische Mathematik* 3.1 (1961), pp. 157–168.

[Pol64] Boris T Polyak. "Some methods of speeding up the convergence of iteration methods". In: *Ussr computational mathematics and mathematical physics* 4.5 (1964), pp. 1–17.

[LF71] Vyacheslav Ivanovich Lebedev and SA Finogenov. "The order of choice of the iteration parameters in the cyclic Čebyšev iteration method". In: *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki* 11.2 (1971), pp. 425–438.

[Sze75] Gabor Szegö. "Orthogonal polynomials, vol. 23". In: *American Mathematical Society Colloquium Publications*. 1975.

[Rah77] EA Rahmanov. "On the asymptotics of the ratio of orthogonal polynomials". In: *Mathematics of the USSR-Sbornik* 32.2 (1977), p. 199.

[MNT85] Attila Máté, Paul Nevai, and Vilmos Totik. "Asymptotics for the ratio of leading coefficients of orthonormal polynomials on the unit circle". In: *Constructive Approximation* 1.1 (1985), pp. 63–69.

[Van95] Walter Van Assche. "Weak convergence of orthogonal polynomials". In: *Indagationes Mathematicae* 6.1 (1995), pp. 7–23.

[Bas96] Richard F Bass. "The Doob-Meyer decomposition revisited". In: *Canadian Mathematical Bulletin* 39.2 (1996), pp. 138–150.

[Fis96] Bernd Fischer. *Polynomial based iteration methods for symmetric linear systems*. SIAM, 1996.

[Van97] Walter Van Assche. "Orthogonal polynomials in the complex plane and on the real line". In: *Field Institute Communications* 14 (1997), pp. 211–245.

[MÁ01] Francisco Marcellán and Renato Álvarez-Nodarse. "On the "Favard theorem" and its extensions". In: *Journal of computational and applied mathematics* 127.1-2 (2001), pp. 231–254.

[Pit02] Mihály Pituk. "More on Poincaré's and Perron's theorems for difference equations". In: *The Journal of Difference Equations and Applications* 8.3 (2002), pp. 201–216.

[Nes03] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2003.

[Tot05] Vilmos Totik. "Orthogonal polynomials". In: *arXiv preprint math/0512424* (2005).

[BT06] Andrej Bogdanov and Luca Trevisan. "Average-case complexity". In: *arXiv preprint cs/0606037* (2006).

[Asm08] Søren Asmussen. *Applied probability and queues*. Vol. 51. Springer Science & Business Media, 2008.

[Bal+18] David Balduzzi et al. "The mechanics of n-player differentiable games". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 354–363.

[DPS20] Carles Domingo-Enrich, Fabian Pedregosa, and Damien Scieur. "Average-case acceleration for bilinear games and normal matrices". In: *arXiv preprint arXiv:2010.02076* (2020).

[Paq+20] Courtney Paquette et al. "Halting time is predictable for large models: A universality property and average-case analysis". In: *arXiv preprint arXiv:2006.04299* (2020).

[Ped20] Fabian Pedregosa. *Momentum: when Chebyshev meets Chebyshev*. Dec. 2020. URL: http://fa.bianp.net/blog/2020/momentum/.

[PS20] Fabian Pedregosa and Damien Scieur. "Acceleration through spectral density estimation". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 7553–7562.

[SP20] Damien Scieur and Fabian Pedregosa. "Universal average-case optimality of polyak momentum". In: *arXiv preprint arXiv:2002.04664* (2020).

[AGZ21] Naman Agarwal, Surbhi Goel, and Cyril Zhang. "Acceleration via Fractal Learning Rate Schedules". In: *arXiv preprint arXiv:2103.01338* (2021).

[Gou+21] Baptiste Goujaud et al. "Super-Acceleration with Cyclical Step-sizes". In: *arXiv preprint arXiv:2106.09687* (2021).

[Oym21] Samet Oymak. "Provable Super-Convergence with a Large Cyclical Learning Rate". In: *IEEE Signal Processing Letters* (2021).

[PP21] Courtney Paquette and Elliot Paquette. "Dynamics of Stochastic Momentum Methods on Large-scale, Quadratic Models". In: *arXiv preprint arXiv:2106.03696* (2021).

[Paq+21] Courtney Paquette et al. "SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality". In: *arXiv preprint arXiv:2102.04396* (2021).

[Ped21] Fabian Pedregosa. *Acceleration without Momentum*. Apr. 2021. URL: http://fa.bianp.net/blog/2021/no-momentum/.

[WLA21] Jun-Kun Wang, Chi-Heng Lin, and Jacob D Abernethy. "A Modular Analysis of Provable Acceleration via Polyak's Momentum: Training a Wide ReLU Network and a Deep Linear Network". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 10816–10827.

A Proofs of Section 2

Theorem 3.1. *Let $\boldsymbol{x}_t$ be generated by a first-order method associated to the polynomial P_t, μ the e.s.d. of H and $\mathbb{E}[(\boldsymbol{x}_0 - \boldsymbol{x}^\star)(\boldsymbol{x}_0 - \boldsymbol{x}^\star)^T] = R^2\boldsymbol{I}$ Then we can write the convergence metrics at time step t as:*

$$\mathbb{E}[\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2] = R^2 \int P_t^2(\lambda) d\mu(\lambda) \tag{45}$$

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] = R^2 \int P_t^2(\lambda)\lambda d\mu(\lambda) \tag{46}$$

$$\mathbb{E}[||\nabla f(\boldsymbol{x}_t)||_2^2] = R^2 \int P_t^2(\lambda)\lambda^2 d\mu(\lambda) \tag{47}$$

Proof. We remark that by the definition of the expected spectral distribution μ of $\boldsymbol{H}$, we have for continuous g:

$$\mathbb{E}_H[g(tr(\boldsymbol{H}))] = \int g(\lambda)\,\mathrm{d}mu(\lambda) \tag{108}$$

We know that $\boldsymbol{x}_t - \boldsymbol{x}^\star = P_t(\boldsymbol{H})(\boldsymbol{x}_0 - \boldsymbol{x}^\star)$. We can write $||\boldsymbol{x}_t - \boldsymbol{x}^\star||^2$ in terms of a trace and use the independence of $\boldsymbol{H}$ and $\boldsymbol{x}_0 - \boldsymbol{x}^\star$ to connect it to the e.s.d.:

$$\begin{aligned}
\mathbb{E}||\boldsymbol{x}_t - \boldsymbol{x}^\star||^2 &= \mathbb{E}[tr((\boldsymbol{x}_0 - \boldsymbol{x}^\star)^T P_t(\boldsymbol{H})^2(\boldsymbol{x}_0 - \boldsymbol{x}^\star))] && (109)\\
&= \mathbb{E}_{\boldsymbol{H},\boldsymbol{x}_0-\boldsymbol{x}^\star}[tr(P_t(\boldsymbol{H})^2(\boldsymbol{x}_0 - \boldsymbol{x}^\star)(\boldsymbol{x}_0 - \boldsymbol{x}^\star)^T] && (110)\\
&= \mathbb{E}_{\boldsymbol{H}}\left[P_t(\boldsymbol{H})^2\mathbb{E}_{\boldsymbol{x}_0-\boldsymbol{x}^\star}[(\boldsymbol{x}_0 - \boldsymbol{x}^\star)(\boldsymbol{x}_0 - \boldsymbol{x}^\star)^T])\right] && (111)\\
&= R^2\mathbb{E}_{\boldsymbol{H}}[P_t(tr(\boldsymbol{H}))^2] = R^2 \int P_t(\lambda)^2\,\mathrm{d}\mu(\lambda) && (112)
\end{aligned}$$

For the gradient and function value the reasoning is the same by noticing:

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] &= \mathbb{E}[tr((\boldsymbol{x}_0 - \boldsymbol{x}^\star)^T P_t(\boldsymbol{H})\boldsymbol{H}P_t(\boldsymbol{H})(\boldsymbol{x}_0 - \boldsymbol{x}^\star))] && (113)\\
&= \mathbb{E}_{\boldsymbol{H}}[(\lambda P_t)(\mathrm{tr}(H))^2] && (114)
\end{aligned}$$

Where λP_t is also a polynomial. As $\nabla f(\boldsymbol{x}_t) = \boldsymbol{H}(\boldsymbol{x}_t - \boldsymbol{x}^\star)$

$$\begin{aligned}
\mathbb{E}||\nabla f(\boldsymbol{x}_t))||^2 &= \mathbb{E}[tr((\boldsymbol{x}_0 - \boldsymbol{x}^\star)^T P_t(\boldsymbol{H})\boldsymbol{H}^2P_t(\boldsymbol{H})(\boldsymbol{x}_0 - \boldsymbol{x}^\star))] && (115)\\
&= \mathbb{E}_{\boldsymbol{H}}[(\lambda^2 P_t)(\mathrm{tr}(H))^2] && (116)
\end{aligned}$$

□

Proposition 3.2. *[PS20] Let $P_t^\star$ be defined as*

$$P_t^\star := \arg\min_{P_t(0)=1} \int P_t^2(\lambda)\lambda^l d\nu(\lambda) \tag{53}$$

then $(P_t^\star)$ is the family of residual orthogonal polynomials w.r.t. to $\lambda^{l+1}d\nu$

Proof. We differentiate the expression for the metrics w.r.t. to the coefficients of the polynomials:

$$\begin{aligned}
\frac{d}{da_k}\left(\int \lambda^l P_t^2(\lambda) d\mu(\lambda)\right) &= \int \lambda^l \frac{\mathrm{d}}{\mathrm{d}a_k}\left(\sum_{k=0}^{t} a_k\lambda^k P_t(\lambda)\right)\mathrm{d}\mu(\lambda) =\\
&= 2\cdot\left(\int \lambda^{l+k} P_t(\lambda) d\mu(\lambda)\right) = 0
\end{aligned}$$

This means that $P_t(\lambda)$ is orthogonal to any polynomial of degree $t-1$ w.r.t to the intern product $\langle ., . \rangle_{\lambda^{l+1} d\mu}$

□

B Coefficients of the Jacobi Method

We'll first state two lemmas that allow us to construct the optimal polynomials. With them in hand the procedure is trivial.

Lemma 13. *Let $(\tilde{P}_t)$ be a family polynomials following*

$$\tilde{P}_t(\lambda) = (\alpha_t + \beta_t \lambda)\tilde{P}_{t-1})\lambda) + \gamma_t \tilde{P}_{t-2}(\lambda)$$

With $\tilde{P}_0$ a constant polynomial and $\tilde{P}_t \neq 0, \forall t$. Then:

$$P_t(\lambda) = (a_t + b_t \lambda)P_{t-1}(\lambda) + (1 - a_t)P_{t-2}(\lambda) \tag{117}$$

Is the recurrence for $P_t(\lambda) = \tilde{P}_t(\lambda)/\tilde{P}_t(0)$. With:

$$a_t = \delta_t \alpha_t \tag{118}$$
$$b_t = \delta_t \beta_t \tag{119}$$
$$\delta_t = (\alpha_t + \gamma_t \delta_{t-1}) \qquad (\delta_0 = 0) \tag{120}$$

The proof of this is presented in [PS20]. Further, we know how to compute the recurrence for the polynomials of a shifted distribution:

Lemma 14. *Let $(\tilde{P}_t)$ be a family polynomials orthogonal w.r.t following*

$$\tilde{P}_t(\lambda) = (\alpha_t + \beta_t \lambda)\tilde{P}_{t-1}(\lambda) + \gamma_t \tilde{P}_{t-2}(\lambda \tag{121}$$

And define polynomials P_t s.t. :

$$P_t(m(\lambda)) = \tilde{P}_t(\lambda)$$

With $m(\lambda) = a\lambda + b$ a non singular affine transform. Then P_t follows a recurrence like in eq. (121), with:

$$\alpha'_t = \alpha_t + b\beta_t \tag{122}$$
$$\beta'_t = a\beta_t \tag{123}$$
$$\gamma'_t = \gamma_t \tag{124}$$

The lemma is self-evident by considering eq. (121) with argument $m^{-1}(\lambda)$

Then to get the recurrence relation for the residual polynomial w.r.t $x^\beta(L-x)^\alpha$, we begin by the standard jacobi polynomials, which are orthogonal w.r.t $(1-x)^\alpha(1+x)^\beta$ and follow a recurrence according to $\alpha_t, \beta_t, \gamma_t$ below, shift the distribution according to $\eta(x)\ldots$, and then transform to the residual ones:

Proposition B.1. *The residual polynomials w.r.t. $d\mu(\lambda) = \lambda^\beta(L-\lambda)^\alpha$, follow the recurrence:*

$$P_t(\lambda) = (a_t + b_t \lambda)P_{t-1}(\lambda) + (1 - a_t)P_{t-2}(\lambda)$$

With:

$$\alpha_t = \frac{(\alpha^2-\beta^2)(2n+\alpha+\beta+1)}{2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)} \tag{125}$$

$$\beta_t = \frac{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)}{(2(n+1)(n+\alpha+\beta+1)} \tag{126}$$

$$\gamma_t = -\frac{(n+\alpha)(n+\beta)(2n+\alpha+\beta+2)}{(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)} \tag{127}$$

$$\tilde{a}_t = \alpha_t - \beta_t \tag{128}$$

$$\tilde{b}_t = \frac{2}{L}\beta_t \tag{129}$$

$$\delta_t = \frac{1}{\tilde{a}_t + c_t\delta_{t-1}} \qquad (\delta_0 = 0) \tag{130}$$

$$a_t, b_t = \delta_t\tilde{a}_t, \delta_t\tilde{b}_t \tag{131}$$

C Proofs of section 3

In the following we'll consider shifted versions of the spectral distributions. This shift is written as an affine transform $m(\lambda) : [0, L] \to [-1, 1]$ because most results in the theory of orthogonal polynomials are stated in terms of distributions supported in $[-1, 1]$. This can be seen as an additional layer of abstraction because the quantities evaluated with the shifted distributions and polynomials are proportional, i.e. if $P_t(x) = \tilde{P}_t(m(x))$ and $\mu'(x) = \tilde{\mu}'(m(x))$:

$$\int P_t^2(x)\mu'(x)\,\mathrm{d}x \propto \int \tilde{P}_t^2(x)\tilde{\mu}'(x)\,\mathrm{d}x \tag{132}$$

So all the asymptotics are the same and we consider ν restricted to $[-1, 1]$. The Jacobi polynomials $P_t^{\alpha,\beta}$ are orthogonal w.r.t; $d\mu(x) = (1-x)^\alpha(1+x)^\beta$, we note that most results give them in terms of normalization $\tilde{P}_t^{\alpha,\beta}(-1) = (-1)^t\binom{t+\beta}{t}$. We'l write $\tilde{P}_t^{\alpha,\beta}$ for this normalization and $P_t^{\alpha,\beta}$ for the residual polynomials

Theorem 7. *A Generalized Chebyshev Method with parameters (α, β) applied to a problem with e.s.d. as in assumption 1 has rates:*

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] \sim L \cdot C_{1,\nu}^{\alpha,\beta} \begin{cases} t^{-1-2\beta} & \text{if } \alpha < \tau + 1/2 \text{ and } \beta < \xi + 3/2 \\ t^{-2(\xi+2)}\log t & \text{if } \alpha = \tau + 1/2 \text{ and } \beta = \xi + 3/2 \\ t^{2(\max\{\alpha-\beta-\tau,-\xi-1\}-1)} & \text{if } \alpha > \tau + 1/2 \text{ or } \beta > \xi + 3/2 \end{cases} \tag{100}$$

$$\mathbb{E}[||\nabla f(\boldsymbol{x}_t)||_2^2] \sim L^2 \cdot C_{2,\nu}^{\alpha,\beta} \begin{cases} t^{-1-2\beta} & \text{if } \alpha < \tau + 1/2 \text{ and } \beta < \xi + 5/2 \\ t^{-2(\xi+3)}\log t & \text{if } \alpha = \tau + 1/2 \text{ and } \beta = \xi + 5/2 \\ t^{2(\max\{\alpha-\beta-\tau,-\xi-2\}-1)} & \text{if } \alpha > \tau + 1/2 \text{ or } \beta > \xi + 5/2 \end{cases} \tag{101}$$

Where $C_\nu^{\alpha,\beta}$ is a distribution dependent constant.

Proof. We'll prove that for any α and β, $\xi, \tau > -1$, $l > 0$ and ν following Assumption 1, we have:

$$\int P_t^{\alpha,\beta}(x)^2 x^l d\nu_{\tau,\xi-l}(x) \sim L^l C_\nu^{\alpha,\beta} \begin{cases} t^{-1-2\beta} & \text{if } \alpha < \tau + 1/2 \text{ and } \beta < \xi + 1/2 \\ t^{-2(\xi+1)}\log t & \text{if } \alpha = \tau + 1/2 \text{ and } \beta = \xi + 1/2 \\ t^{2(\max\{\alpha-\beta-\tau,-\xi\}-1)} & \text{if } \alpha > \tau + 1/2 \text{ or } \beta > \xi + 1/2 \end{cases}$$

We'll first show this result for the Beta weights, then show that distributions with the same concentration behave similarly.
The normalization of $\tilde{P}_t^{\alpha,\beta}$ is s.t.[[Sze75] (4.3.3)]:

$$\int_{-1}^{1} \tilde{P}_t^{\alpha,\beta}(x)(1-x)^{\alpha}(1+x)^{\beta}\,\mathrm{d}x = \frac{2^{\alpha+\beta+1}}{2n+\alpha+\beta+1}\frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{\Gamma(n+1)\Gamma(n+\alpha+\beta+1)} = \Theta(t^{-1}) \tag{133}$$

The residual polynomials then are s.t. $|P_t^{\alpha,\beta}| = \Theta(t^{-\beta})\tilde{P}_t^{\alpha,\beta}$.
The following result (Exercise 91, Generalisation of 7.34.1) from [Sze75]:

Lemma 15. *We have:*

$$\int_0^1 (1-x)^{\tau} P_t^{\alpha,\beta}(x)^2 dx \sim \Theta(h_{\tau}^{\alpha}) \tag{134}$$

$$h_{\tau}^{\alpha} := \begin{cases} t^{2(\alpha-\tau-1)} & \text{if } \alpha > \tau + 1/2 \\ t^{-1}\log t & \text{if } \alpha = \tau + 1/2 \\ t^{-1} & \text{if } \alpha < \tau + 1/2 \end{cases} \tag{135}$$

Noting that $\tilde{P}_t^{\alpha,\beta}(x) = (-1)^t \tilde{P}_t^{\beta,\alpha}(-x)$, we can write:

$$\int_{-1}^{1} \tilde{P}_t(x)^2(1-x)^{\tau}(1+x)^{\xi}dx = \Theta\left(\int_0^1 (1-x)^{\tau}|\tilde{P}_t^{\alpha,\beta}(x)|^2 dx\right) + \Theta\left(\int_0^1 (1-x)^{\xi}|\tilde{P}_t^{\beta,\alpha}(x)|^2 dx\right) \tag{136}$$

We can then show our result for $\mathrm{d}\nu_{\tau,\xi-l}(x) = x^{\xi-l}(L-x)^{\alpha}$ by carefully considering each of the cases on Lemma 15 and the maximum of each term in eq. 136, and an added $t^{-2\beta}$ from the different normalization. With this, we have the wanted result for the Beta weights
It remains to show:

$$\int_0^1 \tilde{P}_t^{\alpha,\beta}(x)^2\,\mathrm{d}\nu_{\tau,\xi}(x) = \Theta\left(\int_0^1 (1-x)^{\tau}\tilde{P}_t^{\alpha,\beta}(x)^2 dx\right) \tag{137}$$

And the rest follows from the same arguments. We do this with the help of this lemma shown in [Van95] relating to the weak convergence of the orthogonal polynomials:

Lemma 16. *Let μ be a measure and (p_t) it's family of orthonormal polynomials s.t. p_t follow the recurrence:*

$$xp_t(x) = a_t p_{t+1}(x) + b_t p_t(x) + a_{t-1}p_{t-1}(x)$$

and a_t, b_t converge respectively to a, b. Then for any f continuous and bounded:

$$\int f(x)p_t^2(x)d\mu(x) \to \frac{1}{\pi}\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}dx \tag{138}$$

Let ϵ s.t.

$$x \geq 1-\epsilon \Rightarrow |\,\mathrm{d}\nu_{\tau,\xi} - A(1-x)^{\tau}| \leq B(1-x)^{\tau} \tag{139}$$

We observe that for $0 < x < 1 - \epsilon$, $f(x) = \frac{d\nu_{\tau,\xi}}{(1-x)^\alpha(1+x)^\beta}$ is bounded.
We get from an application of 16, and the observation that $\tilde{P}_t^{\alpha,\beta} = \mathcal{N}_t p_t^{\alpha,\beta}$, with $\mathcal{N}_t = \Theta(t^{-1/2})$:

$$\underbrace{\int_0^1 (1-x)^\tau \tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}x}_{\Theta(h_\tau^\alpha)} = \underbrace{\int_0^{1-\epsilon} (1-x)^\tau \tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}x}_{\Theta(t^{-1})} + \int_{1-\epsilon}^1 (1-x)^\tau \tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}x \Rightarrow \tag{140}$$

$$\int_{1-\epsilon}^1 (1-x)^\tau \tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}x = \Theta(h_\tau^\alpha) \tag{141}$$

$$\int_0^1 \tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}\nu_{\tau,\xi}(x) = \underbrace{\int_0^{1-\epsilon} \tilde{P}_t^{\alpha,\beta}(x)^2 f(x)(1-x)^\alpha(1+x)^\beta \,\mathrm{d}x)}_{\Theta(t^{-1})} + \Theta\left(\underbrace{\int_{1-\epsilon}^1 (1-x)^\tau \tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}x}_{\Theta(h_\tau^\alpha)}\right) \tag{142}$$

□

Theorem 8. *Let ν follow assumption 1. The optimal asymptotic rate for $\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)]$ is $t^{-2(\xi+2)}$ and is attained by the Chebyshev Method with parameters $(\tau, \xi+2)$.*
The optimal rate for $\mathbb{E}[||\nabla f(\boldsymbol{x}_t)||_2^2]$ is $t^{-2(\xi+3)}$ and is attained by the Chebyshev Method with parameters $(\tau, \xi+3)$.

Proof. We'll prove that for $\tau, \xi > -1$ If $\alpha = \tau$ and $\beta = \xi + l + 1$ (i.e., are optimal), the rate of convergence reads

$$\min_{P_t(0)=1} \int P_t^2(\lambda)\lambda^l d\nu(\lambda) = \Theta\left(\int_0^l \tilde{P}_t^{\alpha,\beta}(\lambda)^2 (L-\lambda)^\tau \lambda^{\xi+l} \,\mathrm{d}\lambda\right) = \Theta(t^{-2(\xi+l+1)}) \tag{143}$$

Showing the second equality is easy by considering theorem 7, and that is further the minimum asymptotic rate for the Beta distribution.
As $\tilde{P}_t^{\alpha,\beta}$ has the same rate on ν as on the Beta distribution , the minimum rate for ν is lower bounded by the r.h.s. We argue that, setting $P_t^\nu = \frac{p^\nu}{p_t^\nu(0)}$ the optimal residual and orthonormal polynomials and $\mu_{\tau,\xi}$ the Beta distribution w.r.t, P_t^ν must have the same rate on ν as it does on ν. Indeed, setting ϵ_1, ϵ_2 as in eq. 139, we argue:

$$\int_{1-\epsilon_2}^1 P_t^\nu(x)^2 d\nu(x) = \left(\Theta \int_{1-\epsilon_2}^1 P_t^\nu(x)^2 d\mu(x)\right) \tag{144}$$

$$\int_{-1}^{-1+\epsilon_1} P_t^\nu(x)^2 d\nu(x) = \Theta\left(\int_{-1}^{-1+\epsilon_1} P_t^\nu(x)^2 d\mu(x)\right) \tag{145}$$

$$\int_{-1+\epsilon_1}^{1-\epsilon_2} P_t^\nu(x)^2 d\nu(x) = \Theta\left(\int_{-1+\epsilon_1}^{1-\epsilon_2} P_t^\nu(x)^2 d\mu(x)\right) = \Theta\left(\frac{1}{p_t^\nu(-1)^2}\right) \tag{146}$$

$$\tag{147}$$

Where the first two equations come from the fact that $\nu = \Theta(\mu)$ near -1 and 1 and the third from lemma 16.
This effectively upper bounds the rates on ν because the rates of P_t^ν on $\mu_{\tau,\xi}$ can't be lower than $-2(\xi+1)$. □

Proposition 5.1. *Let f be a convex, L-smooth quadratic function. Then, For the Generalized Chebyshev Method with parameters (α, β) we have:*

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) \leq C_1 L \begin{cases} t^{2(\alpha-\beta)} & \text{if} \quad \alpha > \beta - 1 \\ t^{-1-2\beta}, & \text{if} \quad \alpha \leq \beta - 1 \quad \beta \leq \frac{1}{2} \\ t^{-2}, & \text{if} \quad \alpha \leq \beta - 1 \quad \beta \geq \frac{1}{2} \end{cases} \tag{103}$$

$$||\nabla f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)|| \leq C_2 L^2 \begin{cases} t^{2(\alpha-\beta)} & \text{if} \quad \alpha > \beta - 2 \\ t^{-1-2\beta}, & \text{if} \quad \alpha \leq \beta - 2 \quad \beta \leq 3/2 \\ t^{-4}, & \text{if} \quad \alpha \leq \beta - 2 \quad \beta \geq 3/2 \end{cases} \tag{104}$$

Proof. rates] We'll prove that: $\sup_{x\in[0,L]} x^l P_t^{\alpha,\beta}(x)^2 = O(L^l t^{v(\alpha,\beta,l)})$. Where:

$$v(\alpha, \beta, l) = \begin{cases} 2(\alpha-\beta) & \text{if} \quad \alpha > \beta - l \\ -1-2\beta, & \text{if} \quad \alpha \leq \beta - l \quad \beta \leq l - \frac{1}{2} \\ -2l, & \text{if} \quad \alpha \leq \beta - l \quad \beta \geq l - \frac{1}{2} \end{cases} \tag{148}$$

From [Sze75], Theorem 7.32.2, if $\theta < \frac{\pi}{2}$:

$$\tilde{P}_t^{\alpha,\beta}(\cos\theta) = \begin{cases} O(t^{-1/2}) & \text{if} \quad \alpha < -\frac{1}{2} \\ O(t^\alpha) & \text{if} \quad \alpha \geq -\frac{1}{2}, 0 \leq \theta \leq ct^{-1} \\ \theta^{-\alpha-1/2} O(t^{-1/2}) & \text{if} \quad \alpha \geq -\frac{1}{2}, \theta > ct^{-1} \end{cases} \tag{149}$$

We observe that, from the symmetry of the jacobi polynomials:

$$\sup_{x\in[0,L]} x^l P_t^{\alpha,\beta}(x)^2 = \Theta\left(\max\left\{\sup_{x\in[0,1]} P_t^{\alpha,\beta}(x)^2, \sup_{x\in[0,1]} (1-x)^l P_t^{\beta,\alpha}(x)^2\right\}\right) \tag{150}$$

The $(1-x)^l$ term, corresponds to $(2\sin(\frac{\theta}{2}))^l$ in the variable θ, which is $O(\theta^{2l})$. The rest follows from carefully considering the expressions given by eq. 149. □

Theorem 10. *Let ν as in assumption 1, then, for the Nesterov method:*

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] \sim C'_{1,\nu} \begin{cases} t^{-2(\xi+2)} & \text{if } \xi < -1/2 \\ t^{-3}\log t & \text{if } \xi = -1/2 \\ t^{-(\xi+7/2)} & \text{if } \xi > -1/2 \end{cases} \qquad \mathbb{E}[||\nabla f(\boldsymbol{x}_t)||_2^2] \sim C'_{2,\nu} t^{-(\xi+9/2)} \tag{105}$$

Proof. We'll prove:

$$\int_0^1 P_t^{\text{Nes}}(\lambda)^2 \lambda^l \,\mathrm{d}\nu_{\tau,\xi-l} \sim C'_\nu \begin{cases} t^{-2(\xi+1)} & \text{if } 0 < \xi < 1/2 \\ t^{-3}\log t & \text{if } \xi = 1/2 \\ t^{-(\xi+5/2)} & \text{if } \xi > 1/2 \end{cases} \tag{151}$$

[Paq+20] has shown that the nesterov polynomials P_t are asymptotically, in t:

$$P_t(\lambda) \sim \frac{2J_1(t\sqrt{\alpha\lambda})}{t\sqrt{\alpha\lambda}} e^{-\alpha\lambda t/2} \tag{152}$$

In the sense that:

$$\int_0^1 u^l \left[\tilde{P_t^2}(u) - \frac{4J_1^2(t\sqrt{u})}{t^2 u} e^{-ut}\right] 4\,\mathrm{d}\mu_{MP}(u) = O(t^{-(l+25/12)}) \tag{153}$$

The arguments can be easily used to show that such an integral is $O(t^{-(\alpha+l+31/12)})$ when evaluated wrt a general $\mathrm{d}\mu$ s.t $\mu' = \Theta(\lambda^\alpha)$ near 0.
We can thus consider our integral of interest substituting P_t^{Nes} by it's Bessel asymptotic and dividing it into three regions, i.e. $[0,1] = [0, \frac{\epsilon}{t}] \cup [\frac{\epsilon}{t}, \frac{\epsilon}{\sqrt{t}}] \cup [\frac{\epsilon}{\sqrt{t}}, 1]$ corresponding to two different regimes for the Bessel function. The first region will give us the asymptotic and the others we'll bound.
We consider first, for some $\epsilon > 0$:

$$\int_{\frac{\epsilon}{t}}^{\frac{\epsilon}{\sqrt{t}}} u^\xi \frac{4J_1^2(t\sqrt{u})}{t^2 u} e^{-ut}\,\mathrm{d}u \tag{154}$$

We note the asymptotic for J_1^2:

$$J_1^2(\sqrt{tv}) \sim \frac{1}{\pi\sqrt{tv}}(1 + \cos(2\sqrt{tv} + 2\gamma)) \tag{155}$$

Doing the change of variable $v = tu$, and identifying the upper limit of the interval, which is $\epsilon t^{1/2}$ to ∞:

$$\int_{\frac{\epsilon}{t}}^{\frac{\epsilon}{\sqrt{t}}} u^\xi \frac{4J_1^2(t\sqrt{u})}{t^2 u} e^{-ut}\,\mathrm{d}u = \Theta\left(t^{-2-\xi}\int_\epsilon^\infty v^{\xi-1} J_1^2(\sqrt{tv}) e^{-v}\,\mathrm{d}v\right) \tag{156}$$

$$= \Theta\left(t^{-2-\xi}\int_\epsilon^\infty v^{\xi-1}\frac{1}{\pi\sqrt{tv}} e^{-v}\,\mathrm{d}v\right) \tag{157}$$

$$= \Theta\left(t^{-\frac{5}{2}-\xi}\underbrace{\int_\epsilon^\infty v^{\xi-\frac{3}{2}}\frac{1}{\pi\sqrt{tv}} e^{-v}\,\mathrm{d}v}_{\Gamma(\xi-\frac{1}{2},\epsilon)}\right) \tag{158}$$

Where the cosinus term goes to 0 from the Riemann-Lebesgue lemma and Γ is the incomplete Gamma function.
The term corresponding to the interval $[\epsilon t^{-1/2}, 1]$ is exponentially small. Indeed, because of the exponential e^{-ut} it is $O(e^{-\epsilon\sqrt{t}})$. This shows that the integral concentrates in a region that is closer and closer to 0 and that only the behaviour of the distribution near 0 matters.We have for the $[0, \frac{\epsilon}{t}]$ region, doing the change of variables $v = t^2 u$:

$$\int_0^{\frac{\epsilon}{t}} u^\xi \frac{4J_1^2(t\sqrt{u})}{t^2 u} e^{-ut}\,\mathrm{d}u = \Theta\left(t^{-2(\xi+1)}\int_0^{t\epsilon} v^\xi \frac{J_1^2(\sqrt{v})}{v} e^{-\frac{v}{t}}\,\mathrm{d}v\right) \tag{159}$$

And the $e^{\frac{-v}{t}}$ is $\Theta(1)$. We have the following Bessel asymptotics:

$$\frac{J_1^2(\sqrt{v})}{v} \sim \frac{1}{4}, \qquad v \to 0 \tag{160}$$

$$\frac{J_1^2(\sqrt{v})}{v} = O(v^{-3/2}), \qquad v \to \infty \tag{161}$$

So we divide this integral aswell:

$$t^{-2(\xi+1)}\int_1^{t\epsilon} v^\xi \frac{J_1^2(\sqrt{v})}{v} e^{-\frac{v}{t}}\,\mathrm{d}v = \Theta\left(t^{-2(\xi+1)}\int_\epsilon^{t\epsilon} v^{\xi-\frac{3}{2}}\,\mathrm{d}v\right) = \Theta\left(I_\xi(t) t^{-\xi-\frac{5}{2}}\right) \tag{162}$$

$$t^{-2(\xi+1)}\int_0^1 v^\xi \frac{J_1^2(\sqrt{v})}{v} e^{-\frac{v}{t}}\,\mathrm{d}v = \Theta\left(t^{-2(\xi+1)}\int_0^1 \epsilon^1 v^\xi\,\mathrm{d}v\right) = \Theta\left(t^{-2(\xi+1)}\right) \tag{163}$$

Where $I_\xi(t) = \log t$ if $\xi = \frac{1}{2}$ and 1 otherwise.The nesterov rate is then $I_\xi(t) t^{-\xi-\frac{5}{2}}$ if $\xi \geq \frac{1}{2}$ and $t^{-2(\xi+1)}$ if $0 < \xi < \frac{1}{2}$ □

Theorem 11. *Let ν as in assumption 1, then for gradient descent:*

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] = \Theta(t^{-(\xi+2)}) \qquad \mathbb{E}[||\nabla f(\boldsymbol{x}_t)||_2^2] = \Theta(t^{-(\xi+3)}) \tag{106}$$

Proof. Considering that $P_t^{\mathrm{GD}}(\lambda) = (1 - \frac{\lambda}{L})^t$ we'll prove :

$$\int_0^1 (1-\lambda)^{2t}\lambda^l \,\mathrm{d}\nu_{\tau,\xi-l} = \Theta(t^{-(\xi+l+1)} \tag{164}$$

We know, for the Beta weights, that:

$$\int_0^1 (1-\lambda)^{2t+\tau}\lambda^{\xi+l} \,\mathrm{d}\lambda = \frac{\Gamma(l+\xi+1)\Gamma(2t+\tau+1)}{\Gamma(2t+l+\xi+\tau+2)} = \Theta(t^{-(\xi+l+1)}) \tag{165}$$

We can indentify this asymptotic to the interval $\int_0^\epsilon$ for any ϵ because:

$$\int_\epsilon^1 (1-\lambda)^{2t+\tau}\lambda^{\xi+l} \,\mathrm{d}\lambda = \mathcal{O}((1-\epsilon)^{2t}) \tag{166}$$

Then:

$$\int_\epsilon^1 (1-\lambda)^{2t}\lambda^l \,\mathrm{d}\nu_{\tau,\xi-l} = \mathcal{O}((1-\epsilon)^{2t}) \tag{167}$$

$$\int 0^\epsilon (1-\lambda)^{2t}\lambda^l \,\mathrm{d}\nu_{\tau,\xi-l} = \Theta\left(\int_0^\epsilon (1-\lambda)^{2t+\tau}\lambda^{\xi+l} \,\mathrm{d}\lambda\right) = \Theta(t^{-(\xi+l+1)}) \tag{168}$$

□

Theorem 12. *Let $\alpha > -1$ and μ_α be a Gamma distribution, i.e. $d\mu_\alpha(x) = \frac{x^\alpha e^{-x}}{\Gamma(\alpha+1)}$. The optimal rates are given by the Laguerre method of appropriate tuning and:*

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] = \Theta(t^{-(\alpha+2)}) \tag{107}$$

Proof. Let L_t^α be the Laguerre polynomials with the usual normalization [Sze75]:

$$\int L_t^\alpha(x)^2 d\mu_\alpha(x) = L_t^\alpha(0) = \binom{n+\alpha}{n} \tag{169}$$

We further know [[Sze75] (5.1.13)]:

$$\sum_{k=0}^t L_t^\alpha(x) = L_t^{\alpha+1}(x) \tag{170}$$

Thus, letting P_t^α be the residual laguerre polynomial, we consider:

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] &= \int P_t^{\alpha+2}(\lambda)^2 d\mu_{\alpha+1}(\lambda) = \binom{t+\alpha+2}{t}^{-2} \int L_t^{\alpha+2} d\mu_{\alpha+1}(\lambda) \\
&= \binom{t+\alpha+2}{t}^{-2} \sum_{k=0}^t \left[\int L_k^{\alpha+1}(\lambda) d\mu_{\alpha+1}(\lambda)\right] = \binom{t+\alpha+2}{t}^{-2} \sum_{k=0}^t \binom{k+\alpha+1}{k} \\
&= \binom{t+\alpha+2}{t}^{-2} \binom{t+\alpha+2}{t} = \binom{t+\alpha+2}{t}^{-1} = \Theta(t^{-(\alpha+2)})
\end{aligned} \tag{171}$$

□

D Additional Experiments

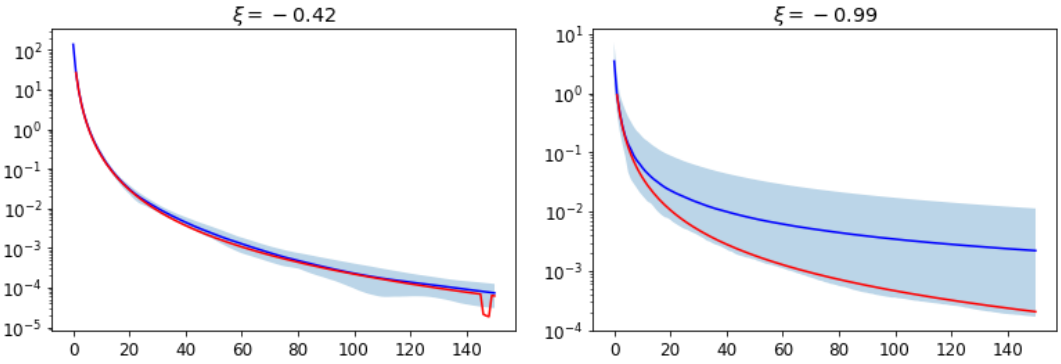

Figure 11: Empirical vs Theoretical function-value performance for GCM($\alpha^\star, \beta^\star$) . Red lines are given by numerical integration, shades are minimum and maximum values under 10 runs

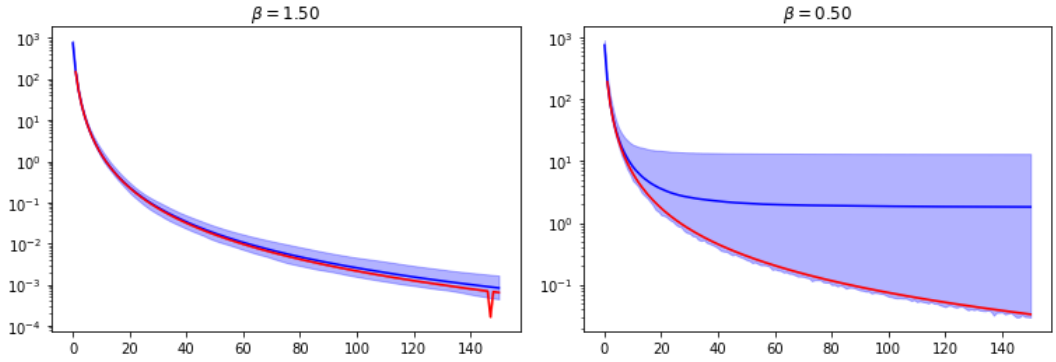

Figure 12: Empirical vs Theoretical function-value performance under Marchenko Pastur distribution. Red lines are given by numerical integration,shades are minimum and maximum values under 10 runs

We note that in the regimes where the empirical average performance doesn't match the theoretical one, we can still find samples of problems who do match. This and the much larger variance on the function-value, this discrepancy is due to numerical unstability in these regimes.