Only tails matter: Average-Case Universality and Robustness in the Convex Regime

Anonymous authors
Paper under double-blind review

Abstract

Recent works have studied the average convergence properties of first-order optimization methods on distributions of quadratic problems.The average-case framework allows a more fine-grained and representative analysis of convergence than usual worst-case results, in exchange for a more precise hypothesis over the data generating process, namely assuming knowledge of the expected spectral distribution (e.s.d) of the random matrix associated to the problem. In this work, we show that a problem's asymptotic average complexity is determined by the concentration of eigenvalues near the edges of the e.s.d. We argue that having à priori information on this concentration is a more grounded assumption than complete knowledge of the e.s.d., and that basing our analysis on this is effectively a middle ground between the coarseness of the worst-case convergence and this more unrealistic hypothesis. We introduce the Generalized Chebyshev method, asymptotically optimal under a hypothesis on this concentration, and globally optimal when the e.s.d. follows a Beta distribution. We compare its performance to classical optimization algorithms, such as gradient descent or Nesterov's scheme, and we show that, asymptotically, Nesterov's method is universally nearly-optimal in the average-case.

1 Introduction

The analysis of the average complexity of algorithms has a long story in computer science. Notably, the quicksort algorithm presents a worst-case complexity of $O(n^2)$, but it's average case complexity is $O(n \log n)$ which makes it empirically very competitive despite it's worst-case complexity. Average case complexity also drives much of the decisions made in cryptography (Bogdanov & Trevisan, 2006).

Despite it's relevance, average case analyses are difficult to extend to other algorithms, mainly because of the intrinsic issue of defining what would be a typical a distribution over the problems. Recently though, Pedregosa & Scieur (2020) derived a framework to systemically evaluate the complexity of first-order methods when applied on distributions of quadratic minimization problems. This is done by tying the average of the convergence metrics to the *expected spectral distribution* (e.s.d) of the problem, which is a well studied object on random matrix theory . Having access to this quantity in practice, though, is a much stronger hypothesis when compared to the worst-case analysis that relies only on the values of the edges of this distribution.

Paquette et al. (2020) extended the average-case framework by introducing a new noisy generative model for the problems . They further derived the average complexity of the Nesterov Accelerated Method Nesterov (2003) on a particular distribution and showed the strong concentration of the metrics around a limiting value as dimensions go to infinite.

Scieur & Pedregosa (2020) showed that for a strongly convex problem with eigenvalues supported on a contiguous interval, the optimal average case complexity converges asymptotically to the one given by the Polyak Heavy Ball method in the worst-case.

In this work, we approach the non-strongly convex case, with a focus on the smooth subcase. In this scenario optimization drastically slows down, as gradient descent presents worst-case

convergence in $\Theta(\frac{1}{t})$ and Nesterov is $\Theta(\frac{1}{t^2})$, where t is the number of iterations, which matches a lower bound on worst-case convergence up to a constant factor (Nesterov, 2003).

In this paper, we show that there is a gap between the worst-case and average-case complexity on convex problems. This is in contrast with the strongly convex case, where both the worst-case and average case are asymptotically equal.

We also adress the gap between the coarseness of the eigenvalue support assumption commonly found on worst-case analysis and the the too precise assumption of the exact e.s.d. of the problem. We introduce a middle ground by looking at the concentration of eigenvalues around the edges of the support, allowing for a *robust* analysis of the average case convergence. That is, we compare the algorithm's performance under different notions of uncertainty on the spectrum. This allows us to choose the best algorithm for a practical problem based in realistic, i.e. imperfect, *à priori* information.

Contributions From a practical standpoint, we propose the Generalized Chebyshev method which outperforms Nesterov and is competitive with Conjugate Gradient on regression tasks in MNIST and Cifar-10 features,.

From a theoretical standpoint, we show that 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 acceleration is such a good method in practice.

We show in section 4 that the average case asymptotic convergence of Gradient Descent, Nesterov and Generalized Chebyshev depend only on the concentrations of the problem's eigenvalue near the edges. We show these rates to be representative of the practical performance of the algorithms in fig. 5, and retrieve the classical worst-case rates as limits of the average case

Finally, we characterize the optimum average case convergence under the scenarios analogous to the three usual hypothesis in worst-case convergence analysis which we resume in table 1.

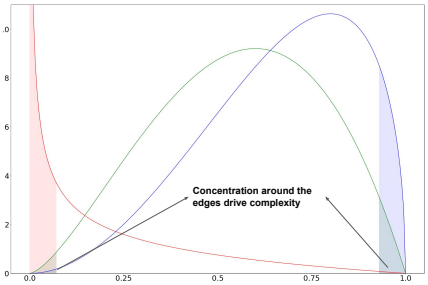

Figure 1: Representation of different spectrums with different concentrations of eigenvalues around the edges of the support. The average-case rates for non-strongly problems are determined by these concentrations

Regime	Worst-case	Average-Case
Strongly conv.	$\left(1-\Theta(\frac{1}{\sqrt{\kappa}})\right)^t$	$\left(1-\Theta(\frac{1}{\sqrt{\kappa}})\right)^t$
Smooth conv.	$\frac{1}{t^2}$	$\frac{1}{t^{2\xi+4}}$
Convex	$\frac{1}{\sqrt{t}}$	$\frac{1}{t^{\alpha+2}}$

Table 1: 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 $d\mu \propto \lambda^\alpha e^{-x}$

2 Average-Case Analysis

In this section, we recall the average-case analysis framework for random quadratic problems. The main result is Theorem 2.1, which relates the expected error to the *expected spectral distribution* and the *residual polynomial*. The one-to-one correspondence between the residual polynomials and first-order methods applied to quadratics will allow us to pose the problem of finding an optimal method as a best approximation problem in the space of polynomials.

We define a **random** quadratic problem:

Problem 1. *Let $\boldsymbol{H} \in \mathbb{R}^{d\times d}$ be a random symmetric positive-definite matrix independent to $\boldsymbol{x}^\star \in \mathbb{R}^d$, a random vector that is the solution to the problem. We define the random quadratic minimization problem as:*

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

We are interested on minimizing the expected errors $\mathbb{E}\|f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)\|$, the expected function-value gap, and $\mathbb{E}||\nabla f(\boldsymbol{x}_t)||^2$, the expected gradient norm, 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 variables $\boldsymbol{H}, \boldsymbol{x}_0$ and $\boldsymbol{x}^\star$.

The expectation we consider is over the problem and not over any randomness of the algorithm.

In this paper, we consider the class of *first-order methods* to minimize (OPT). Methods in this class construct the iterates $\boldsymbol{x}_t$ as:

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

That is, $\boldsymbol{x}_t$ belongs to the span of previous gradients. 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 do not depend on previous updates. This leaves out some methods such as conjugate gradient or methods with line-search.

From First-Order Method to Polynomials. There is an intimate link between first-order methods and polynomials that simplifies the analysis on quadratic objectives. The next proposition shows that, with this link, we can assign to each optimization method a polynomial that determines its convergence. Following Fischer (1996), we will say a polynomial P_t is *residual* if $P_t(0) = 1$.

Proposition 1. *(Hestenes et al., 1952) Let $\boldsymbol{x}_t$ be generated by a first-order method. Then there exists a residual polynomial P_t of degree t, that verifies*

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

Remark 1. *If the first-order method is further a* ***momentum method****, i.e.*

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

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

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

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

A convenient way to collect statistics on the spectrum of a matrix is through its *empirical spectral distribution.*

Definition 1. ***(Expected spectral distribution [e.s.d])****. Let $\boldsymbol{H}$ be a random matrix with eigenvalues $\{\lambda_1, \ldots, \lambda_d\}$. The* ***empirical spectral distribution*** *of $\boldsymbol{H}$, called $\mu_{\boldsymbol{H}}$, is the probability measure*

$$\mu_{\boldsymbol{H}} \stackrel{def}{=} \frac{1}{d}\textstyle\sum_{i=1}^d \delta_{\lambda_i}, \tag{3}$$

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*** *and we denote it*

$$\mu \stackrel{def}{=} \mathbb{E}_{\boldsymbol{H}}[\mu_{\boldsymbol{H}}]. \tag{4}$$

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 will consider $\boldsymbol{x}_0 - \boldsymbol{x}^\star$ and $\boldsymbol{H}$ to be independent, with $\boldsymbol{x}_0 - \boldsymbol{x}^\star$ sampled isotropically.

Theorem 2.1. *Let $\boldsymbol{x}_t$ be generated by a first-order method associated to the polynomial P_t, the measure μ the e.s.d. of H, and $\mathbb{E}[(\boldsymbol{x}_0 - \boldsymbol{x}^\star)(\boldsymbol{x}_0 - \boldsymbol{x}^\star)^T] = R^2\boldsymbol{I}$ for some constant R. Then we can write the convergence metrics at time step t as*

$$\mathbb{E}[\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2] = R^2 \int P_t^2(\lambda) d\mu(\lambda), \qquad \mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] = R^2 \int P_t^2(\lambda)\lambda d\mu(\lambda)$$
$$\text{and} \qquad \mathbb{E}[||\nabla f(\boldsymbol{x}_t)||_2^2] = R^2 \int P_t^2(\lambda)\lambda^2 d\mu(\lambda). \tag{5}$$

This shows that polynomials are a powerful abstraction as they allows us to write all of our convergence metrics in terms of it . For simplicity, we set $R^2 = 1$ and we will refer directly to the polynomials associated to a given method. We will refer to objective l as the metric associated to the added λ^l term, i.e. the function-value is objective $l = 1$.

This framework is linked to the field of **orthogonal polynomials** by the next proposition. We construct an optimal method w.r.t. a given distribution through a family of orthogonal polynomials associated to it.

Proposition 2.1. *(Pedregosa & Scieur, 2020) Let P_t^l be defined as*

$$P_t^l := \arg\min_{P_t(0)=1} \int P_t^2(\lambda)\lambda^l d\nu(\lambda). \tag{6}$$

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

This theorem further implies that the optimal first-order method is a momentum method as Favard's theorem Marcellán & Álvarez-Nodarse (2001) tells us the orthogonal polynomials w.r.t. a given distribution are related through a **three term recurrence**,

$$P_{t+1}(\lambda) = a_t P_t(\lambda) + b_t \lambda P_t(\lambda) + (1 - a_t)P_{t-1}(\lambda). \tag{7}$$

Following Remark 1, the optimal method is derived from this recurrence as

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + (a_t - 1)(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) + b_t \nabla f(\boldsymbol{x}_t)\,. \tag{8}$$

3 Generalized Chebyshev Method

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. Indeed, the following proposition shows that the same distribution arises when we take the gram matrix of random i.i.d. features. We first recall the definition of weak convergence.

Definition 2. *We say a sequence of measures (μ_n) converges weakly to a measure μ if for any bounded and continuous function f we have*

$$\lim_{n\to\infty} \int f(x)\,\mathrm{d}\mu_n(x) = \int f(x)\,\mathrm{d}\mu(x)\,. \tag{9}$$

Proposition 3.1 (Marchenko Pastur Theorem (Bai & Silverstein, 2010))**.** *Let $\boldsymbol{X}_n$ be a $m \times n$ random matrix with X_{ij} i.i.d with variance σ^2 and $\boldsymbol{Y}_n = \frac{1}{n}\boldsymbol{X}_n\boldsymbol{X}_n^T$. Let μ_n be the expected spectrum of $\boldsymbol{Y}_n$, then, as $n \to \infty$ and $\frac{m}{n} \to r$*

$$\mu_n \xrightarrow{weakly} \max(0, 1 - \frac{1}{r})\delta_0 + \mu_{MP},$$
$$d\mu_{MP}(\lambda) = \frac{1}{2\pi\sigma^2}\frac{\sqrt{(\lambda^+ - \lambda)(\lambda - \lambda^-)}}{r\lambda},$$

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 is no specific distribution of X_{ij} considered.

When $r = 1$, i.e. $n = d$, we have $d\mu_{MP} \propto \lambda^{-1/2}\sqrt{\lambda^+ - \lambda}$. Though practical e.s.d's do not take the exact shape of the MP distribution, the same concentration near 0 is often verified.

Pedregosa & Scieur (2020) first derived the optimal method w.r.t. μ_{MP}, and Paquette et al. (2020) derived Nesterov's rates under it. As we are concerned with being robust, a natural step is to consider the Beta weights $d\mu(\lambda) \propto \lambda^{\xi}(L-\lambda)^{\tau}$, which can be seen as a generalization of the MP distribution. We are mainly interested in distributions similar to MP in the sense that they have similar concentrations near 0, i.e. $\xi \approx -1/2$. This is because this approximate concentration is often found in real data.

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* (Flanders & Shortley, 1950). As such, we call methods of this form the *Generalized Chebyshev Method* (GCM).

Algorithm 1: GCM(α, β)

Inputs: Initial vector $\boldsymbol{x}_0$, function f, smoothness parameter estimate L

$\boldsymbol{x}_{-1} \leftarrow \mathbf{0}, \delta_0 \leftarrow 0$

for $t = 1, \dots, T$ **do**

$a_t \leftarrow -\frac{2\left(\beta^2+\alpha\beta+(2n+1)(\alpha+\beta)+2n^2+2\right)(2n+\alpha+\beta+1)}{2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)}$

$b_t \leftarrow \frac{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)}{L(n+1)(n+\alpha+\beta+1)}$

$\gamma_t \leftarrow -\frac{(n+\alpha)(n+\beta)(2n+\alpha+\beta+2)}{(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)}$

$\delta_t \leftarrow \frac{1}{a_t+\gamma_t\delta_{t-1}}$

$\boldsymbol{x}_t \leftarrow \boldsymbol{x}_{t-1} + (\delta_t a_t - 1)(\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}) + \delta_t b_t \nabla f(\boldsymbol{x}_{t-1})$

We 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 Pedregosa & Scieur (2020). We show that Algorithm 1 corresponded to polynomials $\tilde{P}_t^{\alpha,\beta}$ derive the Laguerre method in appendix B.

Remark 2. *The Generalized Chebyshev takes the largest eigenvalue L as a parameter, but the rates we will show are robust to an overestimation of L.*

4 Robust Average Case Rates

We will state our assumption over the spectral distributions. It effectively allows us to parametrize all of our distributions of interest in a way that characterizes 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 milder 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 3. *A Generalized Chebyshev Method with parameters (α, β) applied to a problem with e.s.d. as in assumption 1 has rates*

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] \sim L \cdot C_{1,\nu}^{\alpha,\beta} \begin{cases} t^{-1-2\beta} & \text{if } \alpha < \tau + 1/2 \text{ and } \beta < \xi + 3/2 \\ t^{-2(\xi+2)} \log t & \text{if } \alpha = \tau + 1/2 \text{ and } \beta = \xi + 3/2 \\ t^{2(\max\{\alpha-\beta-\tau,-\xi-1\}-1)} & \text{if } \alpha > \tau + 1/2 \text{ or } \beta > \xi + 3/2 \end{cases}, \tag{10}$$

$$\mathbb{E}[||\nabla f(\boldsymbol{x}_t)||_2^2] \sim L^2 \cdot C_{2,\nu}^{\alpha,\beta} \begin{cases} t^{-1-2\beta} & \text{if } \alpha < \tau + 1/2 \text{ and } \beta < \xi + 5/2 \\ t^{-2(\xi+3)} \log t & \text{if } \alpha = \tau + 1/2 \text{ and } \beta = \xi + 5/2 \\ t^{2(\max\{\alpha-\beta-\tau,-\xi-2\}-1)} & \text{if } \alpha > \tau + 1/2 \text{ or } \beta > \xi + 5/2 \end{cases}, \tag{11}$$

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

Theorem 4 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 4. *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. We remark that the above theorems imply that, at least asymptotically, the Jacobi method is robust for a suboptimal choice of parameter β up to 1/2 below the optimal choice and infinitely above.

For completeness, we also derive worst-case rates for the Jacobi method:

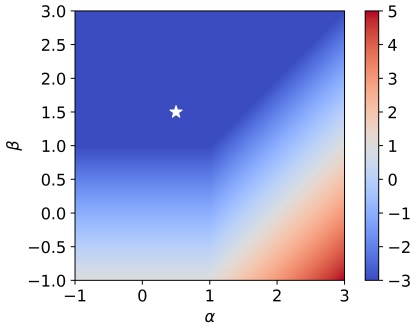

Figure 2: Colormap for the function value average rates for the Marchenko Pastur distribution and different tunings of the Generalized Chebyshev Method, α and. The white represents the optimal tuning, and the color represents the exponent, e.g. 'cooler' is better.

Proposition 4.1. *Let f be a convex, L-smooth quadratic function. Then, for the Generalized Chebyshev Method with parameters (α, β), we have:*

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) \leq C_1 L \begin{cases} t^{2(\alpha-\beta)} & \text{if} \quad \alpha > \beta - 1 \\ t^{-1-2\beta}, & \text{if} \quad \alpha \leq \beta - 1 \quad \beta \leq \frac{1}{2} \\ t^{-2}, & \text{if} \quad \alpha \leq \beta - 1 \quad \beta \geq \frac{1}{2} \end{cases}, \tag{12}$$

$$||\nabla f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)|| \leq C_2 L^2 \begin{cases} t^{2(\alpha-\beta)} & \text{if} \quad \alpha > \beta - 2 \\ t^{-1-2\beta}, & \text{if} \quad \alpha \leq \beta - 2 \quad \beta \leq 3/2 \\ t^{-4}, & \text{if} \quad \alpha \leq \beta - 2 \quad \beta \geq 3/2 \end{cases}. \tag{13}$$

For a reasonable choice of α, β, i.e. $\beta \geq \frac{1}{2}$, $\alpha \leq \beta - 1$. the function value achieves the theoretical lower bound of t^{-2}.

We now analyse the convergence of the Nesterov method. Nesterov (2003) has shown that it matches up to a a constant factor a lower bound on the worst-case complexity of non

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

strongly convex problems. A natural question is if this performance would translate to good average case rates. To do so, we will extend Paquette et al. (2020) proof for the Nesterov method under the MP distribution.

Theorem 5. *Let ν as in Assumption 1. Then for the Nesterov method*

$$\mathbb{E}[f(\boldsymbol{x}_t)-f(\boldsymbol{x}^\star)] \sim C'_{1,\nu}\begin{cases} t^{-2(\xi+2)} & \text{if } \xi<-1/2 \\ t^{-3}\log t & \text{if } \xi=-1/2 \\ t^{-(\xi+7/2)} & \text{if } \xi>-1/2 \end{cases}, \qquad \mathbb{E}[||\nabla f(\boldsymbol{x}_t)||_2^2] \sim C'_{2,\nu} t^{-(\xi+9/2)}. \tag{14}$$

The difference between the assymp. rates of Nesterov and the optimal ones are $t\xi+l-1/2$, 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 6. *Let ν as in Assumption 1. Then for gradient descent*

$$\mathbb{E}[f(\boldsymbol{x}_t)-f(\boldsymbol{x}^\star)] = \Theta(t^{-(\xi+2)}), \qquad \mathbb{E}[||\nabla f(\boldsymbol{x}_t)||_2^2] = \Theta(t^{-(\xi+3)}). \tag{15}$$

Observe for the function value that the rate for Nesterov is t^{-2} and the rate for Gradient Descent is t^{-1} when $\xi\to-1$.

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

Theorem 7. *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*

$$\mathbb{E}[f(\boldsymbol{x}_t)-f(\boldsymbol{x}^\star)] = \Theta(t^{-(\alpha+2)}). \tag{16}$$

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 especially pessimistic in this scenario.

5 Experiments

We simulate the e.s.d's in two ways. The Marchenko Pastur distribution, which we sample by taking $H=\boldsymbol{X}\boldsymbol{X}^T$ where $\boldsymbol{X}$ has i.i.d samples, that enables us to simulate (τ,ξ) values of $(1/2,3/2)$ and $(1/2,5/2)$.

Other values of (τ,ξ) are simulated by sampling $\Lambda\in\mathbb{R}^d$ from the corresponding Beta distribution and taking $\boldsymbol{H}=\boldsymbol{U}\,\mathbf{diag}(\Lambda)\boldsymbol{U}^T$, where $\boldsymbol{U}$ is an independently sampled orthonormal matrix.

We let $\boldsymbol{x}^\star=0$ and sample $\boldsymbol{x}_0$ from a centered gaussian distribution, the dynamics are the same as in the general case

Our theoretical rates in theorems 6 5 for the Nesterov method and Gradient Descent are precise under the approximate range $-1<\xi<0$ as we show in figure 5. Distributions with higher ξ need many samples otherwise they behave as strongly convex functions.

The same is not true for the Generalized Chebyshev Method. 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.

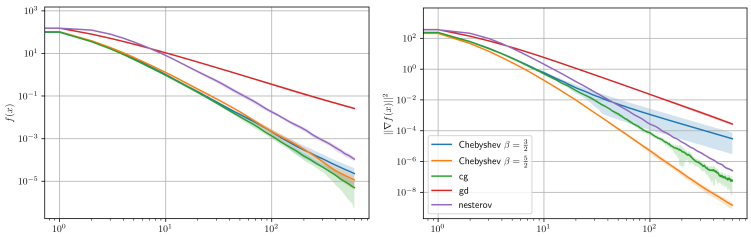

Figure 3: Rates for a synthetic problem, simulating the Marchenko Pastur distribution. Note that both tunings of the GCM achieve performance in function value very close to the one of Conjugate Gradient, which is optimal for every draw of the problem.

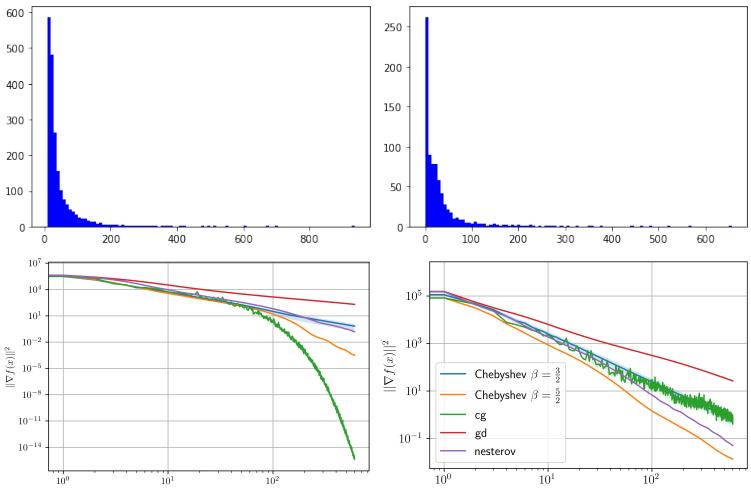

Figure 4: *Above:* Empirical spectrum for the covariance matrix of the features. *Below:* Gradient norms throughout iterations. *Left:* CIFAR-10 Inception features *Right*: MNIST features. Here we choose to compare gradient norms as the minimum function value is not known. The properly tuned GCM achieves remarkable performance under these non-synthetic spectrum's.

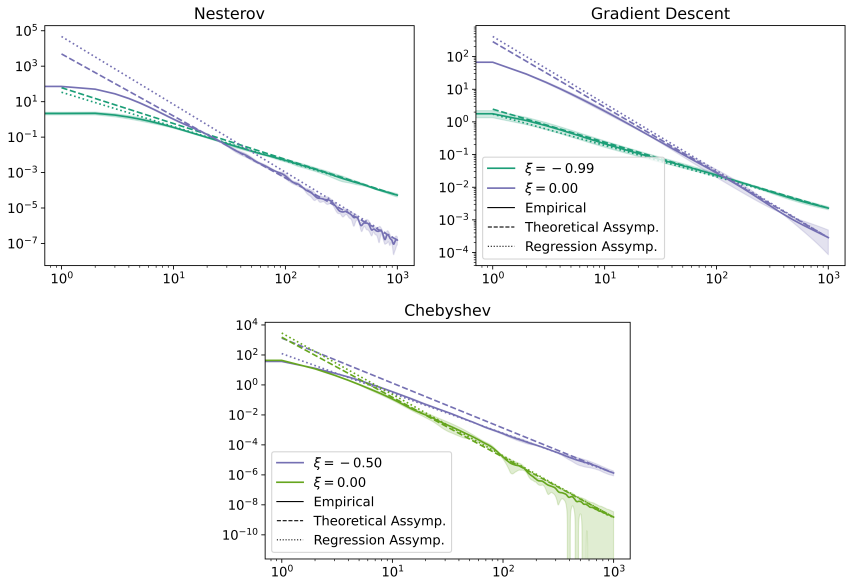

Figure 5: Comparison between experiments run on synthetic Beta distribution and theoretical asymptotic. Y-axis is function value

6 Further work

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 Wang et al. (2021), which has shown the size of the convergent neighborhood for the Polyak method for general strongly convex smooth functions.

Bibliography

Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Andrej Bogdanov and Luca Trevisan. Average-case complexity. *arXiv preprint cs/0606037*, 2006.

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

Donald A Flanders and George Shortley. Numerical determination of fundamental modes. *Journal of Applied Physics*, 21(12):1326–1332, 1950.

Magnus Rudolph Hestenes, Eduard Stiefel, et al. *Methods of conjugate gradients for solving linear systems*, volume 49. NBS Washington, DC, 1952.

Francisco Marcellán and Renato Álvarez-Nodarse. On the "favard theorem" and its extensions. *Journal of computational and applied mathematics*, 127(1-2):231–254, 2001.

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

Courtney Paquette, Bart van Merriënboer, Elliot Paquette, and Fabian Pedregosa. Halting time is predictable for large models: A universality property and average-case analysis. *arXiv preprint arXiv:2006.04299*, 2020.

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

Damien Scieur and Fabian Pedregosa. Universal average-case optimality of polyak momentum. *arXiv preprint arXiv:2002.04664*, 2020.

Gabor Szegö. Orthogonal polynomials, vol. 23. In *American Mathematical Society Colloquium Publications*, 1975.

Walter Van Assche. Weak convergence of orthogonal polynomials. *Indagationes Mathematicae*, 6(1):7–23, 1995.

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*, pp. 10816–10827. PMLR, 2021.

A Proofs of Section 2

Theorem 2.1. *Let $\boldsymbol{x}_t$ be generated by a first-order method associated to the polynomial P_t, the measure μ the e.s.d. of H, and $\mathbb{E}[(\boldsymbol{x}_0 - \boldsymbol{x}^\star)(\boldsymbol{x}_0 - \boldsymbol{x}^\star)^T] = R^2\boldsymbol{I}$ for some constant R. Then we can write the convergence metrics at time step t as*

$$\mathbb{E}[\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2] = R^2 \int P_t^2(\lambda) d\mu(\lambda), \qquad \mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] = R^2 \int P_t^2(\lambda)\lambda d\mu(\lambda)$$
$$\textit{and} \qquad \mathbb{E}[||\nabla f(\boldsymbol{x}_t)||_2^2] = R^2 \int P_t^2(\lambda)\lambda^2 d\mu(\lambda). \tag{5}$$

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

$$\mathbb{E}_H[g(tr(\boldsymbol{H}))] = \int g(\lambda)\,\mathrm{d}\mu(\lambda) \tag{17}$$

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.:

$$\begin{aligned}
\mathbb{E}||\boldsymbol{x}_t - \boldsymbol{x}^\star||^2 &= \mathbb{E}[tr((\boldsymbol{x}_0 - \boldsymbol{x}^\star)^T P_t(\boldsymbol{H})^2(\boldsymbol{x}_0 - \boldsymbol{x}^\star))] && (18)\\
&= \mathbb{E}_{\boldsymbol{H},\boldsymbol{x}_0-\boldsymbol{x}^\star}[tr(P_t(\boldsymbol{H})^2(\boldsymbol{x}_0 - \boldsymbol{x}^\star)(\boldsymbol{x}_0 - \boldsymbol{x}^\star)^T] && (19)\\
&= \mathbb{E}_{\boldsymbol{H}}\left[P_t(\boldsymbol{H})^2\mathbb{E}_{\boldsymbol{x}_0-\boldsymbol{x}^\star}[(\boldsymbol{x}_0 - \boldsymbol{x}^\star)(\boldsymbol{x}_0 - \boldsymbol{x}^\star)^T])\right] && (20)\\
&= R^2\mathbb{E}_{\boldsymbol{H}}[P_t(tr(\boldsymbol{H}))^2] = R^2\int P_t(\lambda)^2\,\mathrm{d}\mu(\lambda) && (21)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] &= \mathbb{E}[tr((\boldsymbol{x}_0 - \boldsymbol{x}^\star)^T P_t(\boldsymbol{H})\boldsymbol{H}P_t(\boldsymbol{H})(\boldsymbol{x}_0 - \boldsymbol{x}^\star))] && (22)\\
&= \mathbb{E}_{\boldsymbol{H}}[(\lambda P_t)(\mathrm{tr}(H))^2], && (23)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}||\nabla f(\boldsymbol{x}_t))||^2 &= \mathbb{E}[tr((\boldsymbol{x}_0 - \boldsymbol{x}^\star)^T P_t(\boldsymbol{H})\boldsymbol{H}^2P_t(\boldsymbol{H})(\boldsymbol{x}_0 - \boldsymbol{x}^\star))] && (24)\\
&= \mathbb{E}_{\boldsymbol{H}}[(\lambda^2 P_t)(\mathrm{tr}(H))^2] && (25)
\end{aligned}$$

□

Proposition 2.1. *(Pedregosa & Scieur, 2020) Let P_t^l be defined as*

$$P_t^l := \arg\min_{P_t(0)=1} \int P_t^2(\lambda)\lambda^l d\nu(\lambda). \tag{6}$$

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 GCM and Laguerre method derivation

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

Lemma 8. *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

$$P_t(\lambda) = (a_t + b_t\lambda)P_{t-1}(\lambda) + (1 - a_t)P_{t-2}(\lambda) \tag{26}$$

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

$$a_t = \delta_t\alpha_t \tag{27}$$
$$b_t = \delta_t\beta_t \tag{28}$$
$$\delta_t = (\alpha_t + \gamma_t\delta_{t-1}) \qquad (\delta_0 = 0) \tag{29}$$

The proof of this is presented in Pedregosa & Scieur (2020). Further, we know how to compute the recurrence for the polynomials of a shifted distribution:

Lemma 9. *Let $(\tilde{P}_t)$ be a family polynomials orthogonal w.r.t following*

$$\tilde{P}_t(\lambda) = (\alpha_t + \beta_t\lambda)\tilde{P}_{t-1}(\lambda) + \gamma_t\tilde{P}_{t-2}(\lambda), \tag{30}$$

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. (30), with:

$$\alpha'_t = \alpha_t + b\beta_t \tag{31}$$
$$\beta'_t = a\beta_t \tag{32}$$
$$\gamma'_t = \gamma_t \tag{33}$$

The lemma is self-evident by considering eq. (30) with argument $m^{-1}(\lambda)$

These results are enough 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 $\eta(x)$, and then transform to the residual ones. We slightly simplify these computations and use remark 1 to get Algorithm 1.

We know (Szegö, 1975) that the Laguerre polynomials L_t^α, with usual normalization, follow the recurrence

$$L_t^\alpha(\lambda) = \left(\frac{2t+\alpha-1}{t} - \frac{1}{t}\lambda\right)L_{t-1}^\alpha(\lambda) + \frac{t+\alpha-1}{t}L_{t-2}^\alpha(\lambda) \tag{34}$$

As we don't have to shift the domains, we have only to apply lemma 8 to get the Laguerre method.Further, we can get a explicit expression for $\delta_t = \frac{t}{t+\alpha}$, simplifying the expression.

Algorithm 2: Laguerre(α)

Inputs: Initial vector $\boldsymbol{x}_0$, function f
$\boldsymbol{x}_{-1} \leftarrow \mathbf{0}$
for $t = 1, \ldots, T$ **do**
  $\boldsymbol{x}_t \leftarrow \boldsymbol{x}_{t-1} + \frac{t-1}{t+\alpha}(\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}) - \frac{1}{t+\alpha}\nabla f(\boldsymbol{x}_{t-1})$

C Proofs of section 3

In the following we will 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 $\mu'(x) = \tilde{\mu}'(m(x))$:

$$\int P_t^2(x)\mu'(x)\,\mathrm{d}x \propto \int \tilde{P}_t^2(x)\tilde{\mu}'(x)\,\mathrm{d}x \tag{35}$$

So all the asymptotics are the same. The Jacobi polynomials $P_t^{\alpha,\beta}$ are those orthogonal w.r.t $d\mu(x) = (1-x)^\alpha(1+x)^\beta$. Most works use the normalization $\tilde{P}_t^{\alpha,\beta}(-1) = (-1)^t\binom{t+\beta}{t}$. We will write $\tilde{P}_t^{\alpha,\beta}$ for this normalization and $P_t^{\alpha,\beta}$ for the residual polynomials

Theorem 3. *A Generalized Chebyshev Method with parameters (α, β) applied to a problem with e.s.d. as in assumption 1 has rates*

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] \sim L \cdot C_{1,\nu}^{\alpha,\beta} \begin{cases} t^{-1-2\beta} & \text{if } \alpha < \tau + 1/2 \text{ and } \beta < \xi + 3/2 \\ t^{-2(\xi+2)}\log t & \text{if } \alpha = \tau + 1/2 \text{ and } \beta = \xi + 3/2 \\ t^{2(\max\{\alpha-\beta-\tau,-\xi-1\}-1)} & \text{if } \alpha > \tau + 1/2 \text{ or } \beta > \xi + 3/2 \end{cases}, \tag{10}$$

$$\mathbb{E}[||\nabla f(\boldsymbol{x}_t)||_2^2] \sim L^2 \cdot C_{2,\nu}^{\alpha,\beta} \begin{cases} t^{-1-2\beta} & \text{if } \alpha < \tau + 1/2 \text{ and } \beta < \xi + 5/2 \\ t^{-2(\xi+3)}\log t & \text{if } \alpha = \tau + 1/2 \text{ and } \beta = \xi + 5/2 \\ t^{2(\max\{\alpha-\beta-\tau,-\xi-2\}-1)} & \text{if } \alpha > \tau + 1/2 \text{ or } \beta > \xi + 5/2 \end{cases}, \tag{11}$$

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

Proof. We will 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 will 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.[Szegö (1975) (4.3.3)]:

$$\int_{-1}^{1} \tilde{P}_t^{\alpha,\beta}(x)(1-x)^\alpha(1+x)^\beta\,\mathrm{d}x = \frac{2^{\alpha+\beta+1}}{2n+\alpha+\beta+1}\frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{\Gamma(n+1)\Gamma(n+\alpha+\beta+1)} = \Theta(t^{-1}) \tag{36}$$

Further, the residual polynomials are s.t. $|P_t^{\alpha,\beta}| = \Theta(t^{-\beta})|\tilde{P}_t^{\alpha,\beta}|$, from the definition of the classical normalization.
We state the result (Exercise 91, Generalisation of 7.34.1) from Szegö (1975):

Lemma 10. *We have:*

$$\int_0^1 (1-x)^\tau P_t^{\alpha,\beta}(x)^2 dx \sim \Theta(h_\tau^\alpha) \tag{37}$$

$$h_\tau^\alpha := \begin{cases} t^{2(\alpha-\tau-1)} & \text{if } \alpha > \tau + 1/2 \\ t^{-1}\log t & \text{if } \alpha = \tau + 1/2 \\ t^{-1} & \text{if } \alpha < \tau + 1/2 \end{cases} \tag{38}$$

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

$$\int_{-1}^{1} \tilde{P}_t(x)^2(1-x)^\tau(1+x)^\xi dx = \Theta\left(\int_0^1 (1-x)^\tau|\tilde{P}_t^{\alpha,\beta}(x)|^2 dx\right) + \Theta\left(\int_0^1 (1-x)^\xi|\tilde{P}_t^{\beta,\alpha}(x)|^2 dx\right) \tag{39}$$

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 10 and the maximum of each term in eq. 39, 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:

$$\int_0^1 \tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}\nu_{\tau,\xi}(x) = \Theta\left(\int_0^1 (1-x)^{\tau} \tilde{P}_t^{\alpha,\beta}(x)^2 dx\right) \tag{40}$$

And the rest follows from the same arguments. We do this with the help of this lemma shown in Van Assche (1995) relating to the weak convergence of the orthogonal polynomials:

Lemma 11. *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:

$$\int f(x) p_t^2(x) d\mu(x) \to \frac{1}{\pi}\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}} dx \tag{41}$$

Let ϵ s.t.

$$x \geq 1-\epsilon \Rightarrow |\,\mathrm{d}\nu_{\tau,\xi} - A(1-x)^{\tau}| \leq B(1-x)^{\tau} \tag{42}$$

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 11, and the observation that $\tilde{P}_t^{\alpha,\beta} = \mathcal{N}_t p_t^{\alpha,\beta}$, with $\mathcal{N}_t = \Theta(t^{-1/2})$:

$$\underbrace{\int_0^1 (1-x)^{\tau}\tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}x}_{\Theta(h_{\tau}^{\alpha})} = \underbrace{\int_0^{1-\epsilon} (1-x)^{\tau}\tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}x}_{\Theta(t^{-1})} + \int_{1-\epsilon}^1 (1-x)^{\tau}\tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}x \Rightarrow \tag{43}$$

$$\int_{1-\epsilon}^1 (1-x)^{\tau}\tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}x = \Theta(h_{\tau}^{\alpha}) \tag{44}$$

$$\begin{aligned}\int_0^1 \tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}\nu_{\tau,\xi}(x) = &\underbrace{\int_0^{1-\epsilon} \tilde{P}_t^{\alpha,\beta}(x)^2 f(x)(1-x)^{\alpha}(1+x)^{\beta} \,\mathrm{d}x)}_{\Theta(t^{-1})} \\ &+ \Theta\left(\underbrace{\int_{1-\epsilon}^1 (1-x)^{\tau}\tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}x}_{\Theta(h_{\tau}^{\alpha})}\right)\end{aligned} \tag{45}$$

□

Theorem 4. *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 will prove that for $\tau, \xi > -1$ If $\alpha = \tau$ and $\beta = \xi + l + 1$ (i.e., are optimal), the rate of convergence reads

$$\min_{P_t(0)=1} \int P_t^2(\lambda)\lambda^l d\nu(\lambda) = \Theta\left(\int_0^l \tilde{P}_t^{\alpha,\beta}(\lambda)^2 (L-\lambda)^{\tau}\lambda^{\xi+l} \,\mathrm{d}\lambda\right) = \Theta(t^{-2(\xi+l+1)}) \tag{46}$$

Showing the second equality is easy by considering theorem 3, 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. 42, we argue:

$$\int_{1-\epsilon_2}^{1} P_t^\nu(x)^2 d\nu(x) = \left(\Theta \int_{1-\epsilon_2}^{1} P_t^\nu(x)^2 d\mu(x)\right) \tag{47}$$

$$\int_{-1}^{-1+\epsilon_1} P_t^\nu(x)^2 d\nu(x) = \Theta\left(\int_{-1}^{-1+\epsilon_1} P_t^\nu(x)^2 d\mu(x)\right) \tag{48}$$

$$\int_{-1+\epsilon_1}^{1-\epsilon_2} P_t^\nu(x)^2 d\nu(x) = \Theta\left(\int_{-1+\epsilon_1}^{1-\epsilon_2} P_t^\nu(x)^2 d\mu(x)\right) = \Theta\left(\frac{1}{p_t^\nu(-1)^2}\right) \tag{49}$$

$$\tag{50}$$

Where the first two equations come from the fact that $\nu = \Theta(\mu)$ near -1 and 1 and the third from lemma 11.
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 4.1. *Let f be a convex, L-smooth quadratic function. Then, for the Generalized Chebyshev Method with parameters (α, β), we have:*

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) \leq C_1 L \begin{cases} t^{2(\alpha-\beta)} & \text{if} \quad \alpha > \beta - 1 \\ t^{-1-2\beta}, & \text{if} \quad \alpha \leq \beta - 1 \quad \beta \leq \frac{1}{2} \\ t^{-2}, & \text{if} \quad \alpha \leq \beta - 1 \quad \beta \geq \frac{1}{2} \end{cases}, \tag{12}$$

$$||\nabla f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)|| \leq C_2 L^2 \begin{cases} t^{2(\alpha-\beta)} & \text{if} \quad \alpha > \beta - 2 \\ t^{-1-2\beta}, & \text{if} \quad \alpha \leq \beta - 2 \quad \beta \leq 3/2 \\ t^{-4}, & \text{if} \quad \alpha \leq \beta - 2 \quad \beta \geq 3/2 \end{cases}. \tag{13}$$

Proof. rates] We will prove that: $\sup_{x\in[0,L]} x^l P_t^{\alpha,\beta}(x)^2 = O(L^l t^{v(\alpha,\beta,l)})$. Where:

$$v(\alpha,\beta,l) = \begin{cases} 2(\alpha-\beta) & \text{if} \quad \alpha > \beta - l \\ -1-2\beta, & \text{if} \quad \alpha \leq \beta - l \quad \beta \leq l - \frac{1}{2} \\ -2l, & \text{if} \quad \alpha \leq \beta - l \quad \beta \geq l - \frac{1}{2} \end{cases} \tag{51}$$

From Szegö (1975), Theorem 7.32.2, if $\theta < \frac{\pi}{2}$:

$$\tilde{P}_t^{\alpha,\beta}(\cos\theta) = \begin{cases} O(t^{-1/2}) & \text{if} \quad \alpha < -\frac{1}{2} \\ O(t^\alpha) & \text{if} \quad \alpha \geq -\frac{1}{2}, 0 \leq \theta \leq ct^{-1} \\ \theta^{-\alpha-1/2} O(t^{-1/2}) & \text{if} \quad \alpha \geq -\frac{1}{2}, \theta > ct^{-1} \end{cases} \tag{52}$$

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

$$\sup_{x\in[0,L]} x^l P_t^{\alpha,\beta}(x)^2 = \Theta\left(\max\left\{\sup_{x\in[0,1]} P_t^{\alpha,\beta}(x)^2, \sup_{x\in[0,1]} (1-x)^l P_t^{\beta,\alpha}(x)^2\right\}\right) \tag{53}$$

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. 52. □

Theorem 5. *Let ν as in Assumption 1. Then for the Nesterov method*

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] \sim C'_{1,\nu} \begin{cases} t^{-2(\xi+2)} & \text{if } \xi < -1/2 \\ t^{-3}\log t & \text{if } \xi = -1/2 \\ t^{-(\xi+7/2)} & \text{if } \xi > -1/2 \end{cases}, \qquad \mathbb{E}[||\nabla f(\boldsymbol{x}_t)||_2^2] \sim C'_{2,\nu} t^{-(\xi+9/2)}. \tag{14}$$

Proof. We will prove:

$$\int_0^1 P_t^{\text{Nes}}(\lambda)^2 \lambda^l \, \mathrm{d}\nu_{\tau,\xi-l} \sim C'_\nu \begin{cases} t^{-2(\xi+1)} & \text{if } 0 < \xi < 1/2 \\ t^{-3} \log t & \text{if } \xi = 1/2 \\ t^{-(\xi+5/2)} & \text{if } \xi > 1/2 \end{cases} \tag{54}$$

Paquette et al. (2020) has shown that the nesterov polynomials P_t are asymptotically, in t:

$$P_t(\lambda) \sim \frac{2J_1(t\sqrt{\alpha\lambda})}{t\sqrt{\alpha\lambda}} e^{-\alpha\lambda t/2} \tag{55}$$

In the sense that:

$$\int_0^1 u^l \left[\tilde{P_t^2}(u) - \frac{4J_1^2(t\sqrt{u})}{t^2 u} e^{-ut} \right] 4 \, \mathrm{d}\mu_{MP}(u) = O(t^{-(l+25/12))} \tag{56}$$

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 will bound.
We consider first, for some $\epsilon > 0$:

$$\int_{\frac{\epsilon}{t}}^{\frac{\epsilon}{\sqrt{t}}} u^\xi \frac{4J_1^2(t\sqrt{u})}{t^2 u} e^{-ut} \, \mathrm{d}u \tag{57}$$

We note the asymptotic for J_1^2:

$$J_1^2(\sqrt{tv}) \sim \frac{1}{\pi\sqrt{tv}}(1 + \cos(2\sqrt{tv} + 2\gamma)) \tag{58}$$

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

$$\int_{\frac{\epsilon}{t}}^{\frac{\epsilon}{\sqrt{t}}} u^\xi \frac{4J_1^2(t\sqrt{u})}{t^2 u} e^{-ut} \, \mathrm{d}u = \Theta\left(t^{-2-\xi} \int_\epsilon^\infty v^{\xi-1} J_1^2(\sqrt{tv}) e^{-v} \, \mathrm{d}v \right) \tag{59}$$

$$= \Theta\left(t^{-2-\xi} \int_\epsilon^\infty v^{\xi-1} \frac{1}{\pi\sqrt{tv}} e^{-v} \, \mathrm{d}v \right) \tag{60}$$

$$= \Theta\left(t^{-\frac{5}{2}-\xi} \underbrace{\int_\epsilon^\infty v^{\xi-\frac{3}{2}} \frac{1}{\pi\sqrt{tv}} e^{-v} \, \mathrm{d}v}_{\Gamma(\xi-\frac{1}{2},\epsilon)} \right) \tag{61}$$

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 $v = t^2 u$:

$$\int_0^{\frac{\epsilon}{t}} u^\xi \frac{4J_1^2(t\sqrt{u})}{t^2 u} e^{-ut} \, \mathrm{d}u = \Theta\left(t^{-2(\xi+1)} \int_0^{t\epsilon} v^\xi \frac{J_1^2(\sqrt{v})}{v} e^{-\frac{v}{t}} \, \mathrm{d}v \right) \tag{62}$$

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

$$\frac{J_1^2(\sqrt{v})}{v} \sim \frac{1}{4}, \qquad v \to 0 \tag{63}$$

$$\frac{J_1^2(\sqrt{v})}{v} = O(v^{-3/2}), \qquad v \to \infty \tag{64}$$

So we divide this integral aswell:

$$t^{-2(\xi+1)} \int_1^{t\epsilon} v^{\xi} \frac{J_1^2(\sqrt{v})}{v} e^{-\frac{v}{t}} \, \mathrm{d}v = \Theta\left(t^{-2(\xi+1)} \int_{\epsilon}^{t\epsilon} v^{\xi-\frac{3}{2}} \, \mathrm{d}v\right) = \Theta\left(I_{\xi}(t) t^{-\xi-\frac{5}{2}}\right) \tag{65}$$

$$t^{-2(\xi+1)} \int_0^1 v^{\xi} \frac{J_1^2(\sqrt{v})}{v} e^{-\frac{v}{t}} \, \mathrm{d}v = \Theta\left(t^{-2(\xi+1)} \int_0 \epsilon^1 v^{\xi} \, \mathrm{d}v\right) = \Theta\left(t^{-2(\xi+1)}\right) \tag{66}$$

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 6. *Let ν as in Assumption 1. Then for gradient descent*

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^{\star})] = \Theta(t^{-(\xi+2)}), \qquad \mathbb{E}[||\nabla f(\boldsymbol{x}_t)||_2^2] = \Theta(t^{-(\xi+3)}). \tag{15}$$

Proof. Considering that $P_t^{\mathrm{GD}}(\lambda) = (1 - \frac{\lambda}{L})^t$ we will prove :

$$\int_0^1 (1-\lambda)^{2t} \lambda^l \, \mathrm{d}\nu_{\tau,\xi-l} = \Theta(t^{-(\xi+l+1)} \tag{67}$$

We know, for the Beta weights, that:

$$\int_0^1 (1-\lambda)^{2t+\tau} \lambda^{\xi+l} \, \mathrm{d}\lambda = \frac{\Gamma(l+\xi+1)\Gamma(2t+\tau+1)}{\Gamma(2t+l+\xi+\tau+2)} = \Theta(t^{-(\xi+l+1)}) \tag{68}$$

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

$$\int_{\epsilon}^1 (1-\lambda)^{2t+\tau} \lambda^{\xi+l} \, \mathrm{d}\lambda = \mathcal{O}((1-\epsilon)^{2t}) \tag{69}$$

Then:

$$\int_{\epsilon}^1 (1-\lambda)^{2t} \lambda^l \, \mathrm{d}\nu_{\tau,\xi-l} = \mathcal{O}((1-\epsilon)^{2t}) \tag{70}$$

$$\int 0^{\epsilon} (1-\lambda)^{2t} \lambda^l \, \mathrm{d}\nu_{\tau,\xi-l} = \Theta\left(\int_0^{\epsilon} (1-\lambda)^{2t+\tau} \lambda^{\xi+l} \, \mathrm{d}\lambda\right) = \Theta(t^{-(\xi+l+1)}) \tag{71}$$

□

Theorem 7. *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*

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^{\star})] = \Theta(t^{-(\alpha+2)}). \tag{16}$$

Proof. Let L_t^{α} be the Laguerre polynomials with the usual normalization Szegö (1975):

$$\int L_t^{\alpha}(x)^2 d\mu_{\alpha}(x) = L_t^{\alpha}(0) = \binom{n+\alpha}{n} \tag{72}$$

We further now [Szegö (1975) (5.1.13)]]:

$$\sum_{k=0}^{t} L_t^{\alpha}(x) = L_t^{\alpha+1}(x) \tag{73}$$

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

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^{\star})] &= \int P_t^{\alpha+2}(\lambda)^2 d\mu_{\alpha+1}(\lambda) = \binom{t+\alpha+2}{t}^{-2} \int L_t^{\alpha+2} d\mu_{\alpha+1}(\lambda) \\
&= \binom{t+\alpha+2}{t}^{-2} \sum_{k=0}^{t} \left[\int L_k^{\alpha+1}(\lambda) d\mu_{\alpha+1}(\lambda)\right] \\
&= \binom{t+\alpha+2}{t}^{-2} \sum_{k=0}^{t} \binom{k+\alpha+1}{k} = \binom{t+\alpha+2}{t}^{-2} \binom{t+\alpha+2}{t} \\
&= \binom{t+\alpha+2}{t}^{-1} = \Theta(t^{-(\alpha+2)})
\end{aligned} \tag{74}$$

□

D Additional Experiments

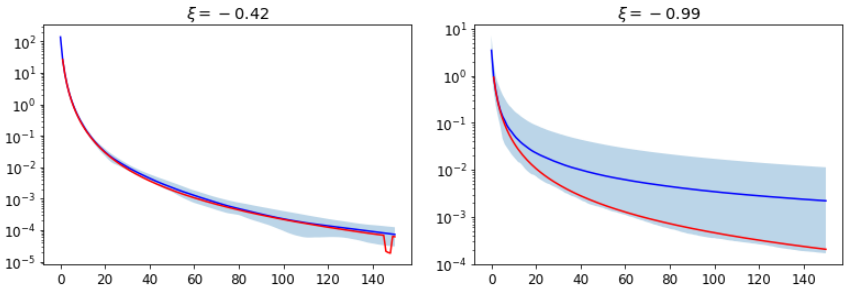

Figure 6: Empirical vs Theoretical function-value performance for GCM($\alpha^\star, \beta^\star$) . Red lines are given by numerical integration, shades are minimum and maximum values under 10 runs

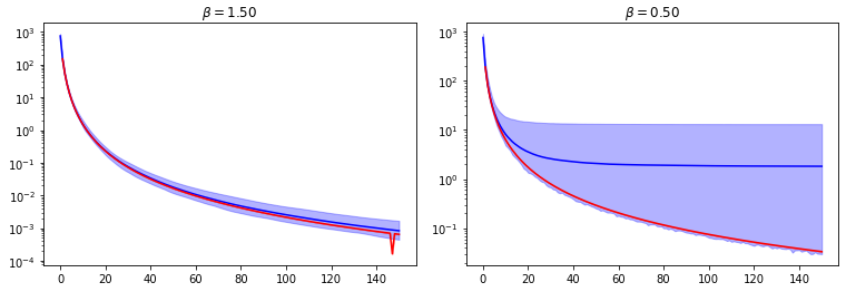

Figure 7: Empirical vs Theoretical function-value performance under Marchenko Pastur distribution. Red lines are given by numerical integration,shades are minimum and maximum values under 10 runs

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.