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 with the problem. This work shows 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 a priori information on this concentration is a more grounded assumption than complete knowledge of the e.s.d. basing our analysis on the approximate concentration is effectively a middle ground between the coarseness of the worst-case scenario convergence and the restrictive previous average-case analysis. 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. Average-case complexity, for instance, drives much of the decisions made in cryptography (Bogdanov & Trevisan, 2006).

Despite their relevance, average-case analyses are difficult to extend to other algorithms, partly because of the intrinsic issue of defining a typical distribution over problem instances. 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 relating the average-case convergence rate to the *expected spectral distribution* (e.s.d) of the objective function's Hessian, which is a well-studied object on random matrix theory. Having access to this object in practice 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 noisy generative model for the problems. They further derived the average complexity of the Nesterov Accelerated Method (Nesterov, 2003) on a particular distribution. They showed the strong concentration of the metrics around a limiting value as dimensions go to infinity.

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 (Polyak, 1964) in the worst-case.

1.1 CURRENT LIMITATIONS OF THE AVERAGE-CASE ANALYSIS

When analyzing the state of the art of average-case methods on quadratics problems, we observe significant limitations that we address in this paper. First, little is known

about the convergence rate on **convex problems**. Also, optimal average-case algorithms require an **exact estimation of the e.s.d** to guarantee an optimal convergence rate, their convergence rate under inexact e.s.d. is not known. Finally, the **non-smooth** is also discussed in (Pedregosa & Scieur, 2020), but with little details.

Convex problems. The minimization of non-strongly convex problems is drastically slower than their strongly convex counterpart, as Gradient Descent presents worst-case convergence in $(\frac{1}{t})$ and Nesterov is $(\frac{1}{t^2})$. In the strongly convex case, both the worst-case and average-case are asymptotically equal. However, little is known on optimal average-case rates for convex problems, as well as the average-case complexity of classical methods such as gradient descent or Nesterov's method, see (Paquette et al., 2020).

Exact estimation of the e.s.d. In (Pedregosa & Scieur, 2020), the theoretical study of optimal algorithms in the average-case requires an exact estimation of the e.s.d. of the problem class. Such estimation may be hard, nor impossible to obtain in practical scenarios. Despite showing good performance when the e.s.d. is estimated with empirical quantities, there are no theoretical guarantees on the performance of the method when the e.s.d. is poorly estimated. There is therefore a need to analyze the algorithm's performance under different notions of uncertainty on the spectrum. This allows a practitioner to choose the best algorithm for a practical problem, even with imperfect *a priori* information.

Non-smooth. Pedregosa & Scieur (2020) briefly introduce average-case optimal rates on non-smooth problems, when the e.s.d. is the Laguerre distribution e . In this paper, we extend the analysis to the generalized Laguerre distribution $e\ ;\ >\ 1$.

1.2 Contributions

Our main contribution is a fine-grained analysis of the average-case complexity on convex quadratic problems: we show that a problem's complexity depends on the concentration of the eigenvalues of e.s.d. around the edges of their support. From this perspective, we propose a family of **optimal algorithms** in the average-case, analyze their **robustness**, and finally exhibit a **universality** result for Nesterov's method. More precisely,

- **(Optimal algorithms).** In Section 3, we propose the Generalized Chebyshev Method (GCM, Algorithm 1), a family of algorithms whose parameters depend on the concentration of the e.s.d. around the edges of their support. If the parameters of the GCM method are set properly, the algorithm converges at an optimal average-case rate (Theorem 3 for smooth problems, Theorem 6 for non-smooth problems), a rate that we show is faster than worst-case optimal methods like Nesterov acceleration. We show these rates to be representative of the practical performance of the algorithms in Fig. 6, and retrieve the classical worst-case rates as limits of the average-case (see Table 1).
- **(Robustness).** Developing an optimal algorithm requires the knowledge of the exact e.s.d. However, in practical scenarios, we only have access to an *approximation* of the e.s.d. In Theorem 2 in Section 4 we analyze the rate of GCM in the presence of such a mismatch. We also analyze the optimal average-case rates of distributions representing the smooth convex, non-smooth convex, and strongly convex settings and compare them with the worst-case rates (Table 1).
- **(Universality).** Finally, in Theorem 4, we analyze the asymptotic average-case convergence rate of Nesterov's method. We show that its convergence rate is nearly optimal (up to a logarithmic factor) under some natural assumptions over the data, namely a concentration of eigenvalues around 0 similar to the Marchenko-Pastur measure. This contributes to the theoretical understanding of the numerical efficiency of Nesterov's acceleration.

2 Average-Case Analysis

In this section, we recall the average-case analysis framework for random quadratic problems. The main result is Theorem 1, which relates the expected error to the *expected spectral*

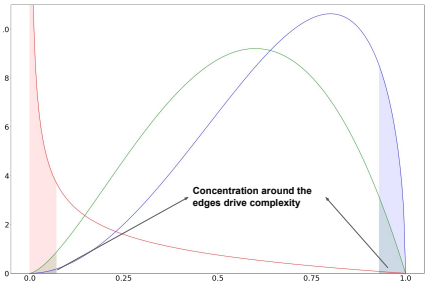


Figure 1: Representation of di□erent spectra with di□erent concentrations of eigenvalues around the edges of the support. The average-case rates for non-strongly problems are determined by these concentrations

Regime	Worst-case	Average-Case
Strongly conv.	$(1 \quad (1=\sqrt{\ }))^t$	$(1 \quad (1=\sqrt{\ }))^t$
Smooth conv.	$1=t^2$	$1=t^{2\ +4}$
Convex	$1=\sqrt{t}$	$1=t^{\ +2}$

Table 1: Comparison between function value worst-case and average-case convergence. is the condition number in the smooth strongly convex case. In the smooth convex case $>$ 1 is the concentration of eigenvalues around 0 (see Assumption 1) and in the non-smooth case we consider d / e

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 the best approximation problem in the space of polynomials.

We define a **random** quadratic problem:

Problem 1. *Let* H 2 $\mathbb{R}^{d\ \ d}$ *be a random symmetric positive-definite matrix independent to* $x^?$ 2 $\mathbb{R}^d$*, a random vector that is the solution to the problem. We define the random quadratic minimization problem as*

$$\min_{x2\mathbb{R}^d} \left\{ f(x) := \frac{1}{2}(x \quad x^?)^{>} H (x \quad x^?) \right\} : \tag{OPT}$$

We are interested on minimizing the expected errors $\mathbb{E}kf(x_t) \quad f(x^?)k$*, the expected function-value gap, and* $\mathbb{E}jj\, r\, f(x_t)jj^2$*, the expected gradient norm, where* x_t *is the* t*-th update of a first-order method starting from* x_0 *and* $\mathbb{E}$ *is the expectation over the random variables* $H; x_0$ *and* $x^?$*.*

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* (F.O.M's) to minimize (OPT). Methods in this class construct the iterates x_t as

$$x_t \ 2 \ x_0 + \mathbf{span}\{r\, f(x_0); : : : ; r\, f(x_{t\ \ 1})\} : \tag{1}$$

That is, 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 of 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* x_t *be generated by a first-order method. Then there exists a residual polynomial* P_t *of degree* t*, that verifies*

$$x_t \quad x^? = P_t(H)(x_0 \quad x^?) : \tag{2}$$

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

$$x_{t+1} = x_t + h_t\, r\, f(x_t) + m_t(x_t \quad x_{t\ 1}).$$

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

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

We note that while most popular F.O.M's can be posed as a momentum method, the Nesterov method 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* H *be a random matrix with eigenvalues* $\{\ _1; \ldots; \ _d\}$. *The* empirical spectral distribution *of* H*, called* $\ _H$*, is the probability measure*

$$\ _H := \frac{1}{d} \sum_{i=1}^{d} \ _i; \tag{3}$$

where $\ _i$ *is the Dirac delta, a distribution equal to zero everywhere except at* $\ _i$ *and whose integral over the entire real line is equal to one.*

Since H *is random, the empirical spectral distribution* $\ _H$ *is a random variable in the space of measures. Its expectation over* H *is called the* expected spectral distribution *and we denote it*

$$:= \mathbb{E}_H[\ _H]. \tag{4}$$

We can link the e.s.d. of H to the convergence of a first-order method on the distribution of H. In the following we will consider $x_0 \quad x^?$ and H to be independent, with $x_0 \quad x^?$ sampled isotropically.

Theorem 1. *Let* x_t *be generated by a first-order method associated to the polynomial* P_t*, the measure the e.s.d. of* H*, and* $\mathbb{E}[(x_0 \quad x^?)(x_0 \quad x^?)^T] = R^2 I$ *for some constant* R*. Then we can write the convergence metrics at time step* t *as*

$$\mathbb{E}[\|x_t \quad x^?\|^2] = R^2 \int \quad P_t^2(\ )\, d\ (\ ); \qquad \mathbb{E}[f(x_t) \quad f(x^?)] = R^2 \int \quad P_t^2(\ )\ d\ (\ )$$
$$\text{and} \qquad \mathbb{E}[\|r\, f(x_t)\|_2^2] = R^2 \int \quad P_t^2(\ )\ ^2 d\ (\ ). \tag{5}$$

This shows that polynomials are a powerful abstraction as they allow us to write all of our convergence metrics within the same framework . 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 one associated to the added $\ ^l$ 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 ((Pedregosa & Scieur, 2020)). *Let* P_t^l *be defined as*

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

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

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**,

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

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

$$x_{t+1} = x_t + (a_t \quad 1)(x_t \quad x_{t\ 1}) + b_t\, r\, f(x_t). \tag{8}$$

3 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. Indeed, it can be shown that the same distribution arises when we take the gram matrix of random centered i.i.d. features with variance ^{2}: the **Marchenko Pastur** distribution.

Definition 2 (MP distribution)**.** *The Marchenko Pastur distribution associated with the parameter* r *and with scale* ^{2} *is given by*

$$d_{MP}( ) = \frac{1}{2 \ ^2} \frac{p}{r} \frac{(\ ^+ \quad)(\quad)}{}; \tag{9}$$

with $^+ = \ ^2(1 + p\overline{r})^2, \quad = \ ^2 \max(0; (1 \quad p\overline{r})^2).$

The Marchenko Pastur distribution $_{MP}$ can be considered a natural first model for e.s.d's as it arises universally from matrices with i.i.d. entries,under mild low moment assumptions, there is no specific distribution of the matrix to be considered. It can be seen as a model for the white-noise in the data. When $r = 1$, i.e. $n = d$, we have $d_{MP} / \ ^{1=2} p\frac{}{^+}$.

Pedregosa & Scieur (2020) first derived the optimal method w.r.t. $_{MP}$, and Paquette et al. (2020) derived Nesterov's rates under the distribution. As we are concerned with being robust, a natural step is to consider the Beta weights.

Definition 3. *The (generalized) Beta weights with parameters* $\ ;$ *and scale* L *are given by the (non-normalized) pdf*

$$d( ) = \quad (L \quad) : \tag{10}$$

This family of distribution generalizes the MP distribution, and both have similar concentrations near 0 when $1=2$.

The optimal method w.r.t. and objective l is associated to a shifted Jacobi polynomial $P_t^{\ ;}$ with $= \ + l + 1; \ = \ .$ When $= \ = \ 1=2$, we retrieve the *Chebyshev Method* (Flanders & Shortley, 1950). As such, we name our proposed methods the *Generalized Chebyshev Method* (GCM).

Algorithm 1: GCM($\ ;$)

Inputs: Initial vector x_0, function f, smoothness parameter estimate L

$x_{\ 1} \quad 0; \ _0 \quad 0$

for $t = 1; : : : ; T$ **do**

$$a_t \quad \frac{2(\ ^2 + \ + (2t+1)(\ + \ ) + 2t^2 + 2)(2t + \ + \ + 1)}{2(t+1)(t + \ + \ + 1)(2t + \ + \ )}$$

$$b_t \quad \frac{(2t + \ + \ + 1)(2t + \ + \ + 2)}{L(t+1)(t + \ + \ + 1)}$$

$$_t \quad \frac{(t + \ )(t + \ )(2t + \ + \ + 2)}{(t+1)(t + \ + \ + 1)(2t + \ + \ )}$$

$$_t \quad \frac{1}{a_t + \ _t \ _{t \ 1}}$$

$$x_t \quad x_{t \ 1} + (\ _t a_t \quad 1)(x_{t \ 1} \quad x_{t \ 2}) + \ _t b_t \ r \ f(x_{t \ 1})$$

We'll consider the Nesterov's method used in Paquette et al. (2020), which is defined by the iterations:

$$x_{t+1} = y_t \quad \frac{1}{L} r \ f(y_t) \tag{11}$$

$$y_{t+1} = x_{t+1} + \frac{t}{t+3}(x_{t+1} \quad x_t) \tag{12}$$

We also consider the Laguerre method, which is optimal w.r.t. $d \ (x) = \frac{x \ e^{\ x}}{(\ +1)}$, taking as a parameter. This method is proposed to optimize non-smooth functions.
Both these methods are generalizations of ones that have been proposed in Pedregosa &

Scieur (2020). We show that Algorithm 1 corresponded to polynomials $P_t^{;}$ and 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 $;$ for a continuous distribution supported in $(0; L]$ s.t. $^{0}_{;}(x) > 0$ for $x \in [0; L]$, $d_{;} = (\ )$ near 0 and $d_{;} = ((L\ )\ )$ near L.*

Assumption 1 is quite nonrestrictive, in that, the spectral distribution of the Hessian for any smooth convex problem can be identified with some $;$ in this class. It is a milder assumption than (1) assuming complete knowledge of the spectrum of the Hessian or (2) the specific distribution on the entries of your data. Moreover the assumption encompasses the frequently used MP (e.g. Martin & Mahoney (2021); Pennington & Bahri (2017)) and Uniform distributions We note there's no need to consider eigenvalues situated at 0 as they do not contribute to the optimization process.

The 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 will work as strongly convex functions in practice.

We show that $;$ indeed behaves like an equivalence class when considering the asymptotics of the convergence of the methods: 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 $;$.

Theorem 2 (GCM average-case rates)**.** *A Generalized Chebyshev Method with parameters $(\ ;\ )$ applied to a problem with e.s.d. as in Assumption 1 has average-case rates*

$$\mathrm{E}[f(x_t)\quad f(x^?)]\quad L\quad C_{1;}^{;}\ \begin{cases} t^{\ 1\ 2} & \text{if } <\ +1=2 \text{ and } <\ +3=2 \\ t^{\ 2(\ +2)}\log t & \text{if } =\ +1=2 \text{ and } =\ +3=2 \\ t^{2(\max f\quad ;\quad 1g\ 1)} & \text{if } >\ +1=2 \text{ or } >\ +3=2 \end{cases} ; \tag{13}$$

$$\mathrm{E}[jj\ r\ f(x_t)jj_2^2]\quad L^2\quad C_{2;}^{;}\ \begin{cases} t^{\ 1\ 2} & \text{if } <\ +1=2 \text{ and } <\ +5=2 \\ t^{\ 2(\ +3)}\log t & \text{if } =\ +1=2 \text{ and } =\ +5=2 \\ t^{2(\max f\quad ;\quad 2g\ 1)} & \text{if } >\ +1=2 \text{ or } >\ +5=2 \end{cases} ; \tag{14}$$

where $C_{;}^{;}$ is a distribution dependent constant.

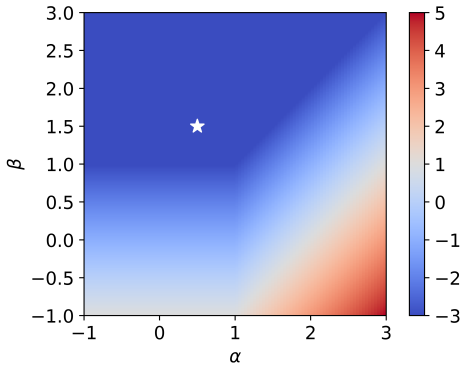
Theorem 2, which is illustrated by fig. 2 shows that overestimating , and underestimating will still leave us with the optimal asymptotic rates, so a good rule of thumb for calibrating the algorithm is to use high and low .

Theorem 3 shows that a proper choice of $;$ can indeed make the Jacobi polynomial asymptotically optimal w.r.t. to any $;$.

Theorem 3 (Optimal Rates)**.** *Let follow Assumption 1. The optimal asymptotic average-case rates for $\mathrm{E}[f(x_t)\quad f(x^?)]$ and $\mathrm{E}[jj\ r\ f(x_t)jj_2^2]$ are attained by the GCM with parameters $(\ ;\ +2)$ and $(\ ;\ +3)$, respectively, and read*

$$\mathrm{E}[f(x_t)\quad f(x^?)] = \ (t^{\ 2(\ +2)}); \qquad \mathrm{E}[jj\ r\ f(x_t)jj_2^2] = \ (t^{\ 2(\ +3)}):$$

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



Method	Parameters (;)	
	$(\frac{1}{2};\frac{1}{2})$	$(\frac{1}{2};\ \frac{1}{2})$
GCM(= $\frac{1}{2}$; = $\frac{5}{2}$)	$t^{\ 5}$	$t^{\ 3}$
GCM(= $\frac{1}{2}$; = $\frac{3}{2}$)	$t^{\ 4}$	$t^{\ 3}$
Nesterov	$t^{\ 4}$	$t^{\ 3}\log t$
Gradient Descent	$t^{-\frac{5}{2}}$	$t^{-\frac{3}{2}}$

Figure 3 & Table 1: The figure illustrates the robustness of the Generalized Chebyshev Method with parameters (;) for a *fixed problem* corresponding to the Marchenko-Pastur distribution (= $\frac{1}{2}$; = $\frac{1}{2}$). The color represents the exponent a of the average-case rate $O(t^a)$ of the method for different values of and . The white star represents the optimal tuning and the blue area is the set of parameters for which the method converges. Note we have a large of region that guarantee the same optimal asymptotic rate. The table compares the asymptotic average-case rates for the function-value for different methods with different (;) values.

We remark that the above theorems imply that, at least asymptotically, the GCM is robust for a suboptimal choice of parameter up to 1=2 below the optimal choice and infinitely above.

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

Proposition 3 (GCM worst-case rates)**.** *Let f be a convex, L-smooth quadratic function. Then, for the Generalized Chebyshev Method with parameters (;), we have worst-case rates*

$$f(x_t) \quad f(x^?) \quad C_1 L \ \begin{cases} t^{2(\ \ )} & \text{if} \quad > \quad 1 \\ t^{\ 1\ 2}; & \text{if} \quad 1 \quad \frac{1}{2} \\ t^{\ 2}; & \text{if} \quad 1 \quad \frac{1}{2} \end{cases}; \tag{15}$$

$$jj\, r\, f(x_t) \quad f(x^?)jj \quad C_2 L^2 \ \begin{cases} t^{2(\ \ )} & \text{if} \quad > \quad 2 \\ t^{\ 1\ 2}; & \text{if} \quad 2 \quad 3{=}2 \\ t^{\ 4}; & \text{if} \quad 2 \quad 3{=}2 \end{cases}; \tag{16}$$

For a reasonable choice of ; , i.e. $\frac{1}{2}$, 1. the function value achieves the theoretical lower bound of $t^{\ 2}$.

We now analyze the convergence of the Nesterov method. Nesterov (2003) has shown that it matches up to 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. To do so, we will extend Paquette et al. (2020) proof for the Nesterov method rates under the MP distribution.

Theorem 4 (Nesterov average-case rates)**.** *Let as in Assumption 1. Then for the Nesterov method, we have average-case rates*

$$\mathrm{E}[f(x_t) \quad f(x^?)] \quad C_{1;}^{0} \ \begin{cases} t^{\ 2(\ +2)} & \text{if} \quad < \quad 1{=}2 \\ t^{\ 3}\log t & \text{if} \quad = \quad 1{=}2 \\ t^{\ (\ +7{=}2)} & \text{if} \quad > \quad 1{=}2 \end{cases}; \quad \mathrm{E}[jj\, r\, f(x_t)jj_2^2] \quad C_{2;}^{0}\ t^{\ (\ +9{=}2)}; \tag{17}$$

The difference between the asymptotic average-case rates of Nesterov and the optimal ones are $t^{\ +/\ 1{=}2}$, when + / > 1=2, $\log t$ when + / = 1=2 and 0 otherwise. This shows that Nesterov is almost optimal when the concentrations near 0 are relatively high, i.e. low .

Theorem 5 (Gradient Descent average-case rates)**.** *Let as in Assumption 1. Then for gradient descent*

$$\mathrm{E}[f(x_t) \quad f(x^?)] = \quad (t^{\ (\ +2)}); \qquad \mathrm{E}[jj\, r\, f(x_t)jj_2^2] = \quad (t^{\ (\ +3)}); \tag{18}$$

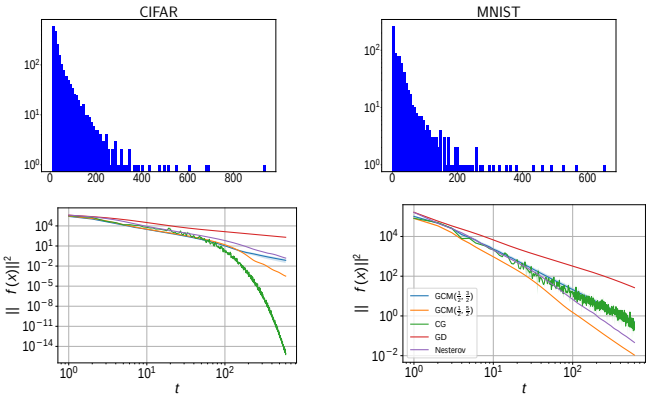


Figure 4: *Above:* Empirical spectrum for the covariance matrix of the features. *Below:* Gradient norms throughout iterations. *Left:* CIFAR-10 Inception features *Right*: MNIST features. Here we choose to compare gradient norms as the minimum function value is not known. The properly tuned GCM achieves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

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

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

Theorem 6 (Laguerre method rates)**.** *Let* $>\ 1$ *and* $;\ >\ 1$ *be a Gamma distribution, i.e.* $d\ (x) = \frac{x\ e^{\ x}}{(\ +1)}$*. The optimal rates are given by the Laguerre method of appropriate tuning and*

$$\mathsf{E}[f(x_t)\ \ f(x^?)] = \ (t^{\ (\ +2)}): \tag{19}$$

Note that this result does not have the same universality as 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

$$f(x_k)\ \ f(x^?)\ \ \frac{C}{t}:$$

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

Remark 3. *All of the expected rates we state are almost deterministic on the high dimensional setting as per the concentration results shown in Paquette et al. (2020)*

5 Experiments

We simulate the e.s.d's in two ways. The Marchenko Pastur distribution, which we sample by taking $H\ =\ X\,X^{T}$ where X has i.i.d. gaussian samples. This enables us to simulate $(\ ;\ )$ values of $(1{=}2;\ \ 1{=}2)$. Other values of $(\ ;\ )$ are simulated by sampling $\ 2\ \mathsf{R}^{d}$ from the corresponding Beta distribution and taking $H\ =\ U\,\mathbf{diag}(\ )U^{T}$, where U is an independently sampled orthonormal matrix.

We let $x^{?}\ =\ 0$ and sample x_0 from a centered Gaussian distribution, the dynamics are the same as in the general case. In all experiments we use the problem's instance largest eigenvalue to calibrate each method (e.g. Gradient Descent's stepsize is $1{=}L$). Our theoretical rates in Theorem 5 and Theorem 4 respectively for the Nesterov method and Gradient Descent are precise under the approximate range $\ 1 <\ \ < 0$ as we show in Figure 6. Distributions with higher $\ $ need many samples otherwise they behave as strongly convex functions.

The same is not true for the Generalized Chebyshev Method. If $\ <\ ^{?}$ or $\ $ is low the empirical findings diverge from the theoretical. We believe this is due to numerical instability

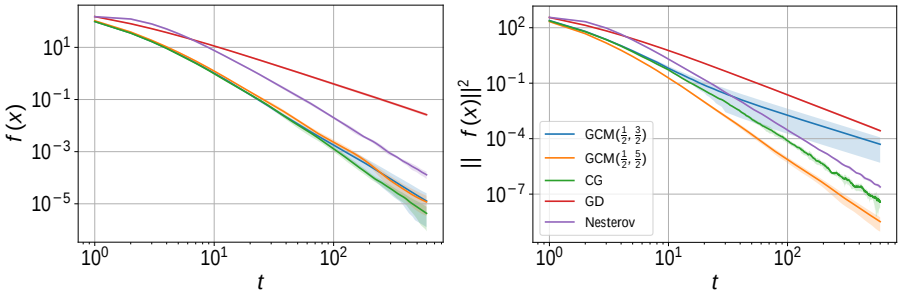


Figure 5: Rates for a synthetic problem, simulating the Marchenko Pastur distribution. Note that both tunings of the GCM achieve performance in function value very close to the one of Conjugate Gradient, which is optimal for every draw of the problem.

under these regimes as the metrics also have much larger variance than in the other regimes. We've not been able though to pinpoint the exact source of this supposed instability. This is shown in appendix D.

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

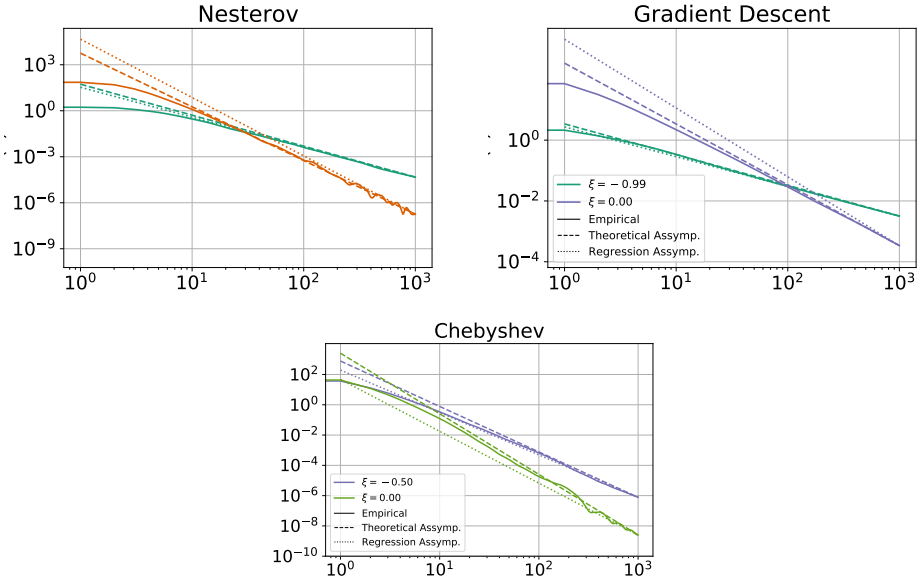


Figure 6: Comparison between experiments run on synthetic Beta distribution and theoretical asymptotic. Y-axis is the function value

6 Conclusion and further work

In this paper, we've established that the assymptotic convergence of first order methods on quadratic problems in the convex regime depend on the concentration of the Hessian's eigenvalue near the edges of the spectrum's support. We further contributed to the theoretic understanding of the Nesterov's method performance and established the contrast between the worst-case and average-case in the main regimes considered in Optimization.

Bibliography

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.

Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

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.

Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *International Conference on Machine Learning*, pp. 2798–2806. PMLR, 2017.

B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 04, 1964.

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.

A Proofs of Section 2

Theorem 1. *Let x_t be generated by a first-order method associated to the polynomial P_t, the measure the e.s.d. of H, and $\mathbb{E}[(x_0 - x^?)(x_0 - x^?)^T] = R^2 I$ for some constant R. Then we can write the convergence metrics at time step t as*

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

Proof. We remark that by the definition of the expected spectral distribution of H, we have for continuous g

$$\mathbb{E}_H[g(tr(H))] = \int g(\lambda)\, \mathrm{d}\mu(\lambda) \tag{20}$$

We know that $x_t - x^? = P_t(H)(x_0 - x^?)$. We can write $\|x_t - x^?\|^2$ in terms of a trace and use the independence of H and $x_0 - x^?$ to connect it to the e.s.d.:

$$\begin{aligned}
\mathbb{E}\|x_t - x^?\|^2 &= \mathbb{E}[tr((x_0 - x^?)^T P_t(H)^2 (x_0 - x^?))] && (21)\\
&= \mathbb{E}_{H, x_0 - x^?}[tr(P_t(H)^2 (x_0 - x^?)(x_0 - x^?)^T] && (22)\\
&= \mathbb{E}_H[P_t(H)^2 \mathbb{E}_{x_0 - x^?}[(x_0 - x^?)(x_0 - x^?)^T])] && (23)\\
&= R^2 \mathbb{E}_H[P_t(tr(H))^2] = R^2 \int P_t(\lambda)^2\, \mathrm{d}\mu(\lambda) && (24)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[f(x_t) - f(x^?)] &= \mathbb{E}[tr((x_0 - x^?)^T P_t(H) H P_t(H)(x_0 - x^?))] && (25)\\
&= \mathbb{E}_H[(\lambda P_t)(\mathrm{tr}(H))^2]; && (26)
\end{aligned}$$

where λP_t is also a polynomial. As $\nabla f(x_t) = H(x_t - x^?)$.

$$\begin{aligned}
\mathbb{E}\|\nabla f(x_t))\|^2 &= \mathbb{E}[tr((x_0 - x^?)^T P_t(H) H^2 P_t(H)(x_0 - x^?))] && (27)\\
&= \mathbb{E}_H[(\lambda^2 P_t)(\mathrm{tr}(H))^2] && (28)
\end{aligned}$$

□

Proposition 2 ((Pedregosa & Scieur, 2020)). *Let P_t^l be defined as*

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

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

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

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

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

□

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 7. *Let* (P_t) *be a family polynomials following*

$$P_t(\ ) = (\ _t + \ _t\ )P_{t\ 1})\ ) + \ _tP_{t\ 2}(\ );$$

with P_0 *a constant polynomial and* $P_t \neq 0; 8t$. *Then*

$$P_t(\ ) = (a_t + b_t\ )P_{t\ 1}(\ ) + (1 \quad a_t)P_{t\ 2}(\ ) \tag{29}$$

is the recurrence for $P_t(\ ) = P_t(\ )=P_t(0)$. *With:*

$$a_t = \ _t\ _t \tag{30}$$
$$b_t = \ _t\ _t \tag{31}$$
$$\ _t = (\ _t + \ _t\ _t\ _1) \quad (\ _0 = 0) \tag{32}$$

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 8. *Let* (P_t) *be a family polynomials orthogonal w.r.t following*

$$P_t(\ ) = (\ _t + \ _t\ )P_{t\ 1}(\ ) + \ _tP_{t\ 2}(\ ); \tag{33}$$

and define polynomials P_t *s.t. :*

$$P_t(m(\ )) = P_t(\ );$$

with $m(\ ) = a\ \ + b$ *a non singular affine transform. Then* P_t *follows a recurrence like in eq.* (33)*, with:*

$$\ ^0_t = \ _t + b\ _t \tag{34}$$
$$\ ^0_t = a\ _t \tag{35}$$
$$\ ^0_t = \ _t \tag{36}$$

The lemma is self-evident by considering eq. (33) with argument $m\ ^1(\ )$

These results are enough to get the recurrence relation for the residual polynomial w.r.t $x\ (L\ \ x)$. We begin by the standard Jacobi polynomials, which are orthogonal w.r.t $(1\ \ x)\ (1 + x)$ and follow a recurrence according to $\ _t; \ _t; \ _t$ below Szegö (1975):

$$\ _t = \frac{(2n + \ \ + \ )(2n + \ \ + \ \ 1)}{2n(n + \ \ + \ )} \tag{37}$$
$$\ _t = \frac{(\ ^2 \ \ ^2)(2n + \ \ + \ \ 1)}{2n(n + \ \ + \ )(2n + \ \ + \ \ 2)} \tag{38}$$
$$\ _t = \frac{2(n + \ \ 1)(n + \ \ 1)(2n + \ \ + \ )}{2n(n + \ \ + \ )(2n + \ \ + \ \ 2)} \tag{39}$$

We then shift the distribution according to $\ (x)$, 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 , with usual normalization, follow the recurrence

$$L_t(\ ) = \ \frac{2t + \ \ 1}{t} \ \ \frac{1}{t} \ \ L_{t\ 1}(\ ) + \frac{t + \ \ 1}{t}L_{t\ 2}(\ ) \tag{40}$$

As we don't have to shift the domains, we have only to apply lemma 7 to get the Laguerre method.Further, we can get a explicit expression for $\ _t = \frac{t}{t+}$, simplifying the expression.

Algorithm 2: Laguerre( )

Inputs: Initial vector x_0, function f

$x_{\ 1} \quad \mathbf{0}$

for $t = 1;:::;T$ **do**

$\quad x_t \quad x_{t\ 1} + \frac{t\ 1}{t+}(x_{t\ 1} \quad x_{t\ 2}) \quad \frac{1}{t+} r\, f(x_{t\ 1})$

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 $m(\ ) : [0; L] \ ! \ [\ 1; 1]$ 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 $P_t(x) = \bar{P}_t(m(x))$ and $^{0}(x) = \tilde{}^{0}(m(x))$:

$$\int P_t^2(x)\ ^{0}(x)\,\mathrm{d}x \ / \quad \int \bar{P}_t^2(x)\tilde{}^{0}(x)\,\mathrm{d}x \tag{41}$$

So all the asymptotics are the same. The Jacobi polynomials $P_{t\ ;\ }$ are those orthogonal w.r.t $d\ (x) = (1 \quad x)\ (1 + x)$. Most works use the normalization $\bar{P}_{t\ ;\ }(\ \ 1) = (\ \ 1)^t\ _t^{t+}$. We will write $P_{t\ ;\ }$ for this normalization and $P_{t\ ;\ }$ for the residual polynomials

Theorem 2 (GCM average-case rates)**.** *A Generalized Chebyshev Method with parameters* $(\ ;\ )$ *applied to a problem with e.s.d. as in Assumption 1 has average-case rates*

$$\mathsf{E}[f(x_t) \quad f(x^?)] \quad L \quad C_{1;}^{;} \begin{cases} t^{\ 1\ 2} & \text{if} \quad < \ + 1{=}2 \ \text{and} \quad < \ + 3{=}2 \\ t^{\ 2(\ +2)} \log t & \text{if} \quad = \ + 1{=}2 \ \text{and} \quad = \ + 3{=}2 \\ t^{2(\max f \quad ;\quad 1g\ 1)} & \text{if} \quad > \ + 1{=}2 \ \text{or} \quad > \ + 3{=}2 \end{cases} ; \tag{13}$$

$$\mathsf{E}[\|r\, f(x_t)\|_2^2] \quad L^2 \quad C_{2;}^{;} \begin{cases} t^{\ 1\ 2} & \text{if} \quad < \ + 1{=}2 \ \text{and} \quad < \ + 5{=}2 \\ t^{\ 2(\ +3)} \log t & \text{if} \quad = \ + 1{=}2 \ \text{and} \quad = \ + 5{=}2 \\ t^{2(\max f \quad ;\quad 2g\ 1)} & \text{if} \quad > \ + 1{=}2 \ \text{or} \quad > \ + 5{=}2 \end{cases} ; \tag{14}$$

where $C_{\ ;\ }$ *is a distribution dependent constant.*

Proof. We will prove that for any and , $;\ > \ 1$, $l > 0$ and following Assumption 1, we have

$$\int P_{t\ ;\ }(x)^2 x^l d\ _{\ ;\ l}(x) \quad L^l C_{\ ;\ } \begin{cases} t^{\ 1\ 2} & \text{if} \quad < \ + 1{=}2 \ \text{and} \quad < \ + 1{=}2 \\ t^{\ 2(\ +1)} \log t & \text{if} \quad = \ + 1{=}2 \ \text{and} \quad = \ + 1{=}2 \\ t^{2(\max f \quad ;\quad g\ 1)} & \text{if} \quad > \ + 1{=}2 \ \text{or} \quad > \ + 1{=}2 \end{cases}$$

We will first show this result for the Beta weights, then show that distributions with the same concentration behave similarly.
The normalization of $P_{t\ ;\ }$ is s.t.[Szegö (1975) (4.3.3)]:

$$\int_{\ 1}^{1} \bar{P}_{t\ ;\ }(x)(1 \quad x)^{\ }(1{+}x)^{\ }\,\mathrm{d}x = \frac{2^{\ +\ +1}}{2n + \quad + \quad + 1} \frac{(n + \quad + 1)\ (n + \quad + 1)}{(n + 1)\ (n + \quad + \quad + 1)} = \quad (t^{\ 1}) \tag{42}$$

Further, the residual polynomials are s.t. $\|P_{t\ ;\ }\|_{\ ;\ } = \quad (t \quad)\|\bar{P}_{t\ ;\ }\|_{\ ;\ }$, from the definition of the classical normalization.
We state the result (Exercise 91, Generalisation of 7.34.1) from Szegö (1975):

Lemma 9. *We have*

$$\int_0^1 (1-x)\, P_t^{\alpha,\beta}(x)^2 dx \qquad (h\ ) \tag{43}$$

$$h := \begin{cases} t^{2(\ \ 1)} & \text{if} \quad > \quad + 1=2 \\ t^{\ 1} \log t & \text{if} \quad = \quad + 1=2 \\ t^{\ 1} & \text{if} \quad < \quad + 1=2 \end{cases} \tag{44}$$

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

$$\int_{-1}^1 P_t(x)^2 (1-x)^{\ } (1+x)^{\ } dx = \int_0^1 (1-x)^{\ } |P_t^{\alpha,\beta}(x)|^2 dx + \int_0^1 (1-x)^{\ } |P_t^{\beta,\alpha}(x)|^2 dx \tag{45}$$

We can then show our result for $\mathrm{d}^{\ \ ;\ \ }{}_{l}(x) = x^{\ \ l}(L\ \ x)$ by carefully considering each of the cases on Lemma 9 and the maximum of each term in eq. 45, and an added $t^{\ \ 2}$ from the different normalization. With this, we have the wanted result for the Beta weights
It remains to show:

$$\int_0^1 P_t^{\alpha,\beta}(x)^2\, \mathrm{d}^{\ \ ;\ \ }{}_{l}(x) = \int_0^1 (1-x)^{\ } P_t^{\alpha,\beta}(x)^2 dx \tag{46}$$

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 10. *Let* $\quad$ *be a measure and* (p_t) *it's family of orthonormal polynomials s.t.* p_t *follow the recurrence:*

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

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

$$\int f(x) p_t^2(x) d^{\ \ }(x)\ \ ! \ \frac{1}{\ } \int_{-1}^1 p\frac{f(x)}{1\ \ x^2} dx \tag{47}$$

Let $\quad$ s.t.

$$x \quad 1 \qquad) \ j\mathrm{d}^{\ \ ;\ \ } A(1\ \ x)^{\ } j \quad B(1\ \ x) \tag{48}$$

We observe that for $0 < x < 1$ $\quad$, $f(x) = \frac{d^{\ \ ;\ \ }}{(1\ \ x)^{\ } (1+x)}$ is bounded.

We get from an application of 10, and the observation that $P_t^{\ \ ;\ \ } = N_t p_t^{\ \ ;\ \ }$, with $N_t = (t^{\ \ 1=2})$:

$$\int_{-1}^1 (1-x)^{\ } P_t^{\alpha,\beta}(x)^2\, \mathrm{d}x = \underbrace{\int_0^1 (1-x)^{\ } P_t^{\alpha,\beta}(x)^2\, \mathrm{d}x}_{(h\ \ )} + \underbrace{\int_{-1}^1 (1-x)^{\ } P_t^{\alpha,\beta}(x)^2\, \mathrm{d}x}_{(t^{\ \ 1})}\ \ ) \tag{49}$$

$$\int_{-1}^1 (1-x)^{\ } P_t^{\alpha,\beta}(x)^2\, \mathrm{d}x = \quad (h\ \ ) \tag{50}$$

$$\int_{-1}^1 P_t^{\alpha,\beta}(x)^2\, \mathrm{d}^{\ \ ;\ \ }{}_{l}(x) = \underbrace{\int_0^1 P_t^{\alpha,\beta}(x)^2 f(x)(1-x)^{\ } (1+x)^{\ } \mathrm{d}x)}_{(t^{\ \ 1})} + \begin{pmatrix} 0 \\ \int_{-1}^1 (1-x)^{\ } P_t^{\alpha,\beta}(x)^2\, \mathrm{d}x \\ 1 \end{pmatrix}_{(h\ \ )} \tag{51}$$

□

Theorem 3 (Optimal Rates)**.** *Let follow Assumption 1. The optimal asymptotic average-case rates for* $\mathsf{E}[f(x_t) \quad f(x^?)]$ *and* $\mathsf{E}[jj\, r\, f(x_t)jj_2^2]$ *are attained by the GCM with parameters* $(\ ; \ + 2)$ *and* $(\ ; \ + 3)$*, respectively, and read*

$$\mathsf{E}[f(x_t) \quad f(x^?)] = \quad (t^{\ 2(\ +2)}); \qquad \mathsf{E}[jj\, r\, f(x_t)jj_2^2] = \quad (t^{\ 2(\ +3)}):$$

Proof. We will prove that for $\ ; \ > \ 1$ If $\ = \ $ and $\ = \ + l + 1$ (i.e., are optimal), the rate of convergence reads

$$\min_{P_t(0)=1} \int P_t^2(\ )^{\ l} d \ (\ ) = \int_0^{\ l} P_t^{\ ; \ }(\ )^2 (L \quad)^{\ +l} d \quad = \quad (t^{\ 2(\ +l+1)}) \tag{52}$$

Showing the second equality is easy by considering theorem 2, and that is further the minimum asymptotic rate for the Beta distribution $\ ; \ $.

By setting p_t and $P_t \ = \frac{p_t}{p_t(0)}$ the optimal orthonormal and residual and polynomials w.r.t. we show that P_t must have the same rate on $\ ; \ $ as it does on , thus the optimal rate of cannot be lower than the optimal rate of $\ ; \ $. Indeed, setting $_1; \ _2$ as in eq. 48:

$$\int_{\ 1}^{\ 1} P_t \ (x)^2 d \ (x) = \int_{\ 1}^{\ 1} P_t \ (x)^2 d \ _{\ ; \ }(x) \tag{53}$$

$$\int_{\ 1+\ _1}^{\ 1 \quad \ _2} P_t \ (x)^2 d \ (x) = \int_{\ 1+\ _1}^{\ 1 \quad \ _2} P_t \ (x)^2 d \ _{\ ; \ }(x) \tag{54}$$

$$\int_{\ 1+\ _1}^{\ 1 \quad \ _2} P_t \ (x)^2 d \ (x) = \int_{\ 1+\ _1}^{\ 1 \quad \ _2} P_t \ (x)^2 d \ _{\ ; \ }(x) = \quad \frac{1}{p_t(\ 1)^2} \tag{55}$$

$$\tag{56}$$

Where the first two equations come from the fact that $\ = \ (\ ; \ )$ near $\ 1$ and 1 and the third from lemma 10.

This effectively upper bounds the rates on because the rates of P_t on $\ ; \ $ can't be lower than $\ 2(\ + l + 1)$. □

Proposition 3 (GCM worst-case rates)**.** *Let* f *be a convex, L-smooth quadratic function. Then, for the Generalized Chebyshev Method with parameters* $(\ ; \ )$*, we have worst-case rates*

$$f(x_t) \quad f(x^?) \quad C_1 L \begin{cases} t^{2(\ \ )} & \text{if} \quad > \quad 1 \\ t^{\ 1 \ 2}; & \text{if} \quad 1 \quad \frac{1}{2} \\ t^{\ 2}; & \text{if} \quad 1 \quad \frac{1}{2} \end{cases}; \tag{15}$$

$$jj\, r\, f(x_t) \quad f(x^?)jj \quad C_2 L^2 \begin{cases} t^{2(\ \ )} & \text{if} \quad > \quad 2 \\ t^{\ 1 \ 2}; & \text{if} \quad 2 \quad 3=2 \\ t^{\ 4}; & \text{if} \quad 2 \quad 3=2 \end{cases}: \tag{16}$$

Proof. rates] We will prove that: $\sup_{x2[0;L]} x^l P_t^{\ ; \ }(x)^2 = O(L^{\ l} t^{v(\ ; \ ;l)})$. Where:

$$v(\ ; \ ; l) = \begin{cases} 2(\ \ ) & \text{if} \quad > \quad l \\ 1 \quad 2 \ ; & \text{if} \quad l \quad l \quad \frac{1}{2} \\ 2l; & \text{if} \quad l \quad l \quad \frac{1}{2} \end{cases} \tag{57}$$

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

$$P_t^{\ ; \ }(\cos \ ) = \begin{cases} O(t^{\ 1=2}) & \text{if} \quad < \quad \frac{1}{2} \\ O(t^{\ \ }) & \text{if} \quad \frac{1}{2}; 0 \quad ct^{\ 1} \\ {}^{1=2}O(t^{\ 1=2}) & \text{if} \quad \frac{1}{2}; \ > ct^{\ 1} \end{cases} \tag{58}$$

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

$$\sup_{x2[0;L]} x^l P_t^{\ ; \ }(x)^2 = \quad \max \left(\sup_{x2[0;1]} P_t^{\ ; \ }(x)^2; \sup_{x2[0;1]} (1 \quad x)^l P_t^{\ ; \ }(x)^2 \right) ! \tag{59}$$

The $(1 \quad x)^l$ term, corresponds to $(2\sin(\frac{}{2}))^l$ in the variable , which is $O(\ ^{2l})$. The rest follows from carefully considering the expressions given by eq. 58. □

Theorem 4 (Nesterov average-case rates)**.** *Let as in Assumption 1. Then for the Nesterov method, we have average-case rates*

$$\mathsf{E}[f(x_t) - f(x^?)] \quad C^{\emptyset}_{1;} \quad \begin{cases} t^{2(+2)} & \text{if } < 1=2 \\ t^{3}\log t & \text{if } = 1=2 \\ t^{(+7=2)} & \text{if } > 1=2 \end{cases} ; \quad \mathsf{E}[jj\, r\, f(x_t)jj_2^2] \quad C^{\emptyset}_{2;}\, t^{(+9=2)}: \tag{17}$$

Proof. We will prove:

$$\int_0^1 P_t^{\mathrm{Nes}}(\ )^2\ ^{l}\mathrm{d} \quad ;\quad {}^{l} \quad C^{\emptyset} \quad \begin{cases} t^{2(+1)} & \text{if } 0 < < 1=2 \\ t^{3}\log t & \text{if } = 1=2 \\ t^{(+5=2)} & \text{if } > 1=2 \end{cases} \tag{60}$$

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

$$P_t(\ ) \quad \frac{2J_1(t\sqrt{\ })}{t\sqrt{\ }}e^{\ t=2} \tag{61}$$

In the sense that:

$$\int_0^1 u^{l}\ P_t^2(u) \quad \frac{4J_1^2(t\sqrt{u})}{t^2 u}e^{\ ut}\ 4\,\mathrm{d}\ _{MP}(u) = O(t^{\ (l+25=12))}) \tag{62}$$

The arguments can be easily used to show that such an integral is $O(t^{\ (+l+31=12)})$ when evaluated wrt a general d s.t $^{\emptyset} = \ (\ )$ 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. $[0;1] = [0;\bar{t}] \cup [\bar{t};\sqrt{\bar{t}}] \cup [\sqrt{\bar{t}};1]$ 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 > 0:

$$\int_{\bar{t}}^{\sqrt{\bar{t}}} u\ \frac{4J_1^2(t\sqrt{u})}{t^2 u}e^{\ ut}\,\mathrm{d}u \tag{63}$$

We note the asymptotic for J_1^2:

$$J_1^2(\sqrt{tv}) \quad \frac{1}{\sqrt{tv}}(1+\cos(2\sqrt{tv}+2\ )) \tag{64}$$

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

$$\int_{\bar{t}}^{\sqrt{\bar{t}}} u\ \frac{4J_1^2(t\sqrt{u})}{t^2 u}e^{\ ut}\,\mathrm{d}u = \quad t^{\ 2}\int_1^{\ } v^{\ 1}J_1^2(\sqrt{tv})e^{\ v}\,\mathrm{d}v \tag{65}$$

$$= \quad t^{\ 2}\int_0^1 v^{\ 1}\frac{1}{\sqrt{tv}}e^{\ v}\,\mathrm{d}v \tag{66}$$

$$= \quad t^{\ \frac{5}{2}}\Big(\underbrace{\int_0^1 v^{\ \frac{3}{2}}\frac{1}{\sqrt{tv}}e^{\ v}\,\mathrm{d}v}_{(\ \frac{1}{2};\ )}\Big) \tag{67}$$

Where the cosinus term goes to 0 from the Riemann-Lebesgue lemma and is the incomplete Gamma function.
The term corresponding to the interval $[\ t^{\ 1=2};1]$ is exponentially small. Indeed, because of the exponential $e^{\ ut}$ it is $O(e^{\ \sqrt{t}})$. 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 $[0;\bar{t}]$ region, doing the change of variables $v = t^2u$:

$$\int_0^{\bar{t}} u\ \frac{4J_1^2(t\sqrt{u})}{t^2 u}e^{\ ut}\,\mathrm{d}u = \quad t^{\ 2(+1)}\int_0^t v\ \frac{J_1^2(\sqrt{v})}{v}e^{\ \frac{v}{t}}\,\mathrm{d}v \tag{68}$$

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

$$\frac{J_1^2(\sqrt{v})}{v} \sim \frac{1}{4}; \qquad v \to 0 \tag{69}$$

$$\frac{J_1^2(\sqrt{v})}{v} = O(v^{-3/2}); \qquad v \to \infty \tag{70}$$

So we divide this integral aswell:

$$t^{-2(\xi+1)} \int_1^t v^{\xi} \frac{J_1^2(\sqrt{v})}{v} e^{-\frac{v}{t}} dv = \Theta\left(t^{-2(\xi+1)} \int_1^t v^{\xi-\frac{3}{2}} dv\right) = \Theta\left(I_\xi(t) t^{-\xi-\frac{5}{2}}\right) \tag{71}$$

$$t^{-2(\xi+1)} \int_0^1 v^{\xi} \frac{J_1^2(\sqrt{v})}{v} e^{-\frac{v}{t}} dv = \Theta\left(t^{-2(\xi+1)} \int_0^1 v^{\xi} dv\right) = \Theta\left(t^{-2(\xi+1)}\right) \tag{72}$$

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 5 (Gradient Descent average-case rates)**.** *Let μ as in Assumption 1. Then for gradient descent*

$$\mathbb{E}[f(x_t) - f(x^\star)] = \Theta(t^{-(\xi+2)}); \qquad \mathbb{E}[\|\nabla f(x_t)\|_2^2] = \Theta(t^{-(\xi+3)}). \tag{18}$$

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

$$\int_0^1 (1-\lambda)^{2t} \lambda^l d\mu_{\xi;l} = \Theta(t^{-(\xi+l+1)}) \tag{73}$$

We know, for the Beta weights, that:

$$\int_0^1 (1-\lambda)^{2t+\beta} \lambda^{\xi+l} d\lambda = \frac{\Gamma(l+\xi+1)\Gamma(2t+\beta+1)}{\Gamma(2t+l+\xi+\beta+2)} = \Theta(t^{-(\xi+l+1)}) \tag{74}$$

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

$$\int_\epsilon^1 (1-\lambda)^{2t+\beta} \lambda^{\xi+l} d\lambda = O((1-\epsilon)^{2t}) \tag{75}$$

Then:

$$\int_\epsilon^1 (1-\lambda)^{2t} \lambda^l d\mu_{\xi;l} = O((1-\epsilon)^{2t}) \tag{76}$$

$$\int_0^\epsilon (1-\lambda)^{2t} \lambda^l d\mu_{\xi;l} = \int_0^\epsilon (1-\lambda)^{2t+\beta} \lambda^{\xi+l} d\lambda = \Theta(t^{-(\xi+l+1)}) \tag{77}$$

☐

Theorem 6 (Laguerre method rates)**.** *Let $\xi > -1$ and μ_ξ; $\xi > -1$ be a Gamma distribution, i.e. $d\mu_\xi(x) = \frac{x^\xi e^{-x}}{\Gamma(\xi+1)}$. The optimal rates are given by the Laguerre method of appropriate tuning and*

$$\mathbb{E}[f(x_t) - f(x^\star)] = \Theta(t^{-(\xi+2)}). \tag{19}$$

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

$$\int L_t^\xi(x)^2 d\mu_\xi(x) = L_t^\xi(0) = \binom{n+\xi}{n} \tag{78}$$

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

$$\sum_{k=0}^{t} L_t^\xi(x) = L_t^{\xi+1}(x) \tag{79}$$

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

$$\begin{aligned}
\mathrm{E}[f(x_t) \quad f(x^?)] &= \int P_t{}^{+2}(\ )^2 d\ {}_{+1}(\ ) = \frac{t+\ +2}{t}{}^2 \int L_t{}^{+2} d\ {}_{+1}(\ ) \\
&= \frac{t+\ +2}{t}{}^2 \sum_{k=0}^{\infty} \int L_k{}^{+1}(\ ) d\ {}_{+1}(\ ) \\
&= \frac{t+\ +2}{t}{}^2 \sum_{k=0}^{\infty} \frac{k+\ +1}{k} = \frac{t+\ +2}{t}{}^2 \frac{t+\ +2}{t} \\
&= \frac{t+\ +2}{t}{}^1 = (t^{\ (\ +2)})
\end{aligned} \tag{80}$$

□

D Additional Experiments

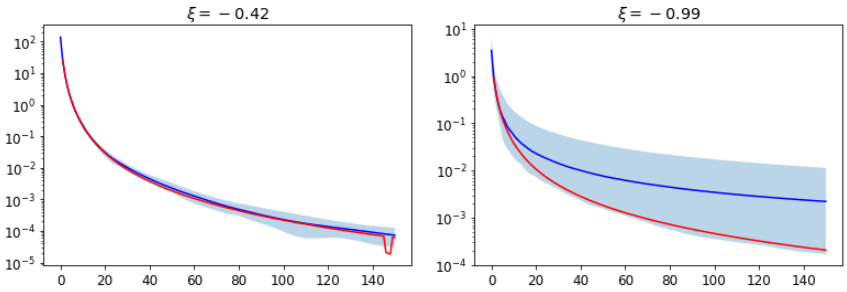


Figure 7: Empirical vs Theoretical function-value performance for GCM($^?$; $^?$) . Red lines are given by numerical integration, shades are minimum and maximum values under 10 runs and the blue line is the mean

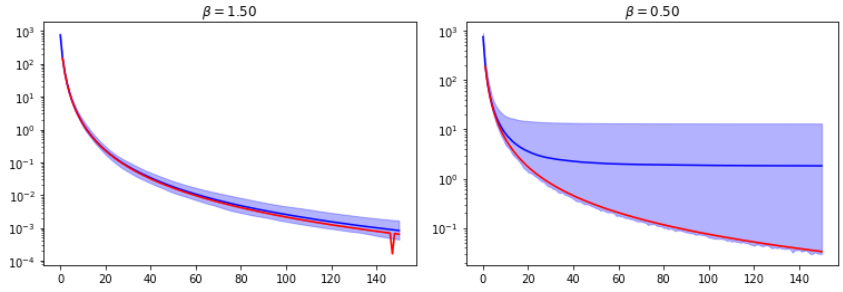


Figure 8: Empirical vs Theoretical function-value performance under Marchenko Pastur distribution. Red lines are given by numerical integration,shades are minimum and maximum values under 10 runs and the blue line is the mean

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.