Only Tails Matter: Average-Case Universality and Robustness in the Convex Regime

Leonardo Cunha [1] **Gauthier Gidel** [1 2] **Fabian Pedregosa** [3] **Damien Scieur** [4] **Courtney Paquette** [5]

Abstract

The recently developed average-case analysis of optimization methods allows a more fine-grained and representative convergence analysis than usual worst-case results. In exchange, this analysis requires a more precise hypothesis over the data generating process, namely assuming knowledge of the expected spectral distribution (ESD) of the random matrix associated with the problem. This work shows that the concentration of eigenvalues near the edges of the ESD determines a problem's asymptotic average complexity. This a priori information on this concentration is a more grounded assumption than complete knowledge of the ESD. This approximate concentration is effectively a middle ground between the coarseness of the worst-case scenario convergence and the restrictive previous average-case analysis. We also introduce the Generalized Chebyshev method, asymptotically optimal under a hypothesis on this concentration and globally optimal when the ESD follows a Beta distribution. We compare its performance to classical optimization algorithms, such as gradient descent or Nesterov's scheme, and we show that, in the average-case context, Nesterov's method is universally nearly optimal asymptotically.

1. Introduction

The analysis of the average complexity of algorithms has a long tradition 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 to distributions of quadratic minimization problems. This derivation is done by relating the average-case convergence rate to the *expected spectral distribution* (ESD) of the objective function's Hessian, which is a well-studied object in random matrix theory. In practice, however, the knowledge of the ESD is a much stronger requirement than the worst-case analysis, which relies only on knowledge of the distribution's support.

Paquette et al. (2022) extended the average-case framework by introducing a noisy generative model for the problems. Among other results, they derived the average complexity of the Nesterov Accelerated Method (Nesterov, 2003) on a particular distribution. Moreover, they showed the concentration of convergence metrics in the infinite sample and dimensional limit.

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.

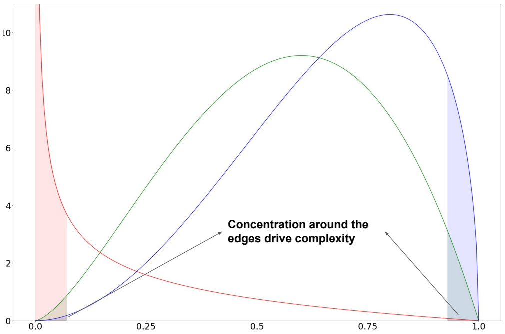

Figure 1. Representation of different spectra with different concentrations of eigenvalues around the support edges. These concentrations determine the average-case rates for non-strongly problems.

[1] MILA and DIRO, Université de Montreal, Montreal, Canada [2] Canada CIFAR AI Chair [3] Google Research [4] Samsung SAIT AI Lab, Montreal, Canada [5] McGill University, Montreal, Canada. Correspondence to: Leonardo Cunha <leonardo-cunha2107@gmail.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

1.1. Current limitations of the average-case analysis

This paper addresses some of the limitations in previous works on average-case analysis. First, little is known about the convergence rate on **convex** but not strongly-convex problems. Also, optimal average-case algorithms require an **exact estimation of the ESD** to guarantee an optimal convergence rate; otherwise, their convergence rate under inexact ESD is unknown. Finally, the **non-smooth** has not yet been discussed in its full generality.

Convex problems. In the smooth and strongly convex case, the optimal worst-case and average-case convergence rates are asymptotically equal (Scieur & Pedregosa, 2020). However, little is known about optimal average-case rates for non-strongly convex problems, as well as the average-case complexity of classical methods such as gradient descent or Nesterov's method, see (Paquette et al., 2022).

Exact estimation of the ESD. Pedregosa & Scieur (2020) developed average-case optimal algorithms that required an exact estimation of the ESD of the problem class. Such an estimation may be complex, if not impossible, to obtain in practice. Despite showing good performance when the ESD is estimated with empirical quantities, there are no theoretical guarantees of the method's performance when the ESD is poorly estimated. Therefore, there is a need to analyze the algorithm's performance under different notions of uncertainty on the spectrum, allowing a practitioner to choose the best algorithm for a practical problem even with imperfect *a priori* information.

Non-smooth objectives. In this paper we provide an average-case analysis for the generalized Laguerre distribution $\lambda^\alpha e^{-\lambda}$, $\alpha > -1$. This is a distribution with unbounded support, i.e., when the largest eigenvalue (also known as smoothness constant) is bounded. This extends the results of Pedregosa & Scieur (2020) for the Laguerre distribution $e^{-\lambda}$.

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 ESD around the edges of their support. Using this analysis, we also derive a family of average-case **optimal** algorithms, analyze their **robustness**, and finally exhibit a **universality** result for Nesterov's method.

- **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 ESD around the edges of their support. With the proper set of parameters, the GCM method converges at an optimal average-case rate (Theorem 4.2 for smooth problems, Theorem 4.6 for non-smooth problems). These rates are faster than optimal worst-case methods like Nesterov acceleration, and we recover the classical worst-case rates as limits of the average-case (see Table 1). Fig. 6 shows our theoretical analysis to match the practical performance of the algorithms.
- **Robustness.** Developing an optimal algorithm requires knowledge of the exact ESD. However, in practical scenarios, we only have access to an *approximation* of the ESD. In Theorem 4.1 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 4.4, we analyze the asymptotic average-case convergence rate of Nesterov's method. We show that its convergence rate is 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 distribution. This observation contributes to the theoretical understanding of the empirical effectiveness of Nesterov's acceleration.

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

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 $\xi > -1$ is the concentration of eigenvalues around 0 and in the non-smooth case we consider $\mathrm{d}\mu \propto \lambda^\alpha e^{-\lambda} d\lambda$.

2. Average-Case Analysis

This section recalls the average-case analysis framework for random quadratic problems. The main result is Theorem 2.5, 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 the best approximation problem in the space of polynomials.

We define a **random** quadratic problem as follows:

Problem 2.1. Let $\boldsymbol{H} \in \mathbb{R}^{d\times d}$ be a random symmetric positive-definite matrix independent of $\boldsymbol{x}^\star \in \mathbb{R}^d$, which is a random vector and a solution to the problem. We define the random quadratic minimization problem as

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

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 $\boldsymbol{x}^\star$.

Note that the expectations we consider throughout the paper are over problem instances, 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 $\boldsymbol{x}_t$ as

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

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 Methods to Polynomials. There is an intimate link between first-order methods and polynomials that simplifies the analysis of quadratic objectives. The following proposition shows that, with this link, we can assign to each optimization method a polynomial that determines its convergence. We will denote $P_t(\lambda)$ and $P_t(\boldsymbol{H})$ for the polynomials taking arguments in $\mathbb{R}$ and $\mathbb{R}^{n\times n}$ respectively with the same coefficients. Following (Fischer, 1996), we will say a polynomial P_t is *residual* if $P_t(0) = 1$.

Proposition 2.2. *(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*

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

Remark 2.3. If the first-order method is further a **momentum method**, i.e.

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

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

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

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 2.4 (Expected spectral distribution (ESD))**.** . 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

$$\mu_{\boldsymbol{H}} \overset{\text{def}}{=} \frac{1}{d}\textstyle\sum_{i=1}^d \delta_{\lambda_i}, \qquad (3)$$

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 variable in the space of measures. Its expectation over $\boldsymbol{H}$ is called the **expected spectral distribution** (ESD) and we denote it

$$\mu \overset{\text{def}}{=} \mathbb{E}_{\boldsymbol{H}}[\mu_{\boldsymbol{H}}]\,. \qquad (4)$$

The following theorem links the ESD to the average-case convergence of a first-order method when $\boldsymbol{x}_0 - \boldsymbol{x}^\star$ and $\boldsymbol{H}$ are independent.

Theorem 2.5. *Let $\boldsymbol{x}_t$ be generated by a first-order method with associated residual polynomial P_t, μ be the ESD, and $\mathbb{E}[(\boldsymbol{x}_0 - \boldsymbol{x}^\star)(\boldsymbol{x}_0 - \boldsymbol{x}^\star)^\top] = R^2\boldsymbol{I}$ for some constant R. Then we have the following identities for different convergence metrics:*

$$\mathbb{E}[\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2] = R^2 \int P_t^2(\lambda)\,\mathrm{d}\mu(\lambda)\,, \qquad (5)$$

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] = \frac{R^2}{2} \int P_t^2(\lambda)\lambda\,\mathrm{d}\mu(\lambda)\,, \qquad (6)$$

$$\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|_2^2] = R^2 \int P_t^2(\lambda)\lambda^2\,\mathrm{d}\mu(\lambda)\,. \qquad (7)$$

This theorem states that polynomials are a powerful abstraction, allowing us to write all our convergence metrics within the same framework. Therefore, in the rest of the paper, we will refer directly to the polynomials associated with a given method. For simplicity, we set $R^2 = 1$.

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

Proposition 2.6. *Let ν be a distribution with continuous support and P_t^l be defined as*

$$P_t^l \overset{def}{=} \arg\min{}_{P_t(0)=1} \int P_t^2(\lambda)\lambda^l d\nu(\lambda)\,. \qquad (8)$$

Then $(P_t^l)_{t=0,1,\ldots}$ is the family of residual orthogonal polynomials w.r.t. to $\lambda^{l+1}d\nu$.

We refer to objective l as the one associated to the added λ^l term, i.e. the function-value is objective $l = 1$. This proposition further implies that the optimal first-order method is a momentum method because 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**,

$$P_{t+1}(\lambda) = (a_t + b_t\lambda)P_t(\lambda) + (1 - a_t)P_{t-1}(\lambda)\,. \tag{9}$$

Following Remark 2.3, the optimal method is derived from this recurrence as

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + (a_t - 1)(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) + b_t \nabla f(\boldsymbol{x}_t)\,. \tag{10}$$

3. Methods

Writing the rates in terms of the *expected spectral distribution* ties the average-case framework to the field of *random matrix theory*. A classical result of this field is that the same spectral distribution arises from different input distributions in the data. For example, the **Marchenko-Pastur** distribution arises in the large sample and dimension limit when we take the Gram matrix of a matrix where the entries are generated i.i.d. from any distribution with mean zero, variance σ^2 and with a bounded fourth moment.

Definition 3.1. The Marchenko-Pastur distribution associated with the parameter r and with scale σ^2 is given by

$$\mathrm{d}\mu_{MP}(\lambda) = \frac{1}{2\pi\sigma^2}\frac{\sqrt{(\lambda^+ - \lambda)(\lambda - \lambda^-)}}{r\lambda}\,\mathrm{d}\lambda\,, \tag{11}$$

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

r in the Gram matrix case, is determined as n/d. The Marchenko-Pastur distribution μ_{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 $\mathrm{d}\mu_{MP} \propto \lambda^{-1/2}\sqrt{\lambda^+ - \lambda}\,\mathrm{d}\lambda$.

Pedregosa & Scieur (2020) first derived the optimal method w.r.t. μ_{MP}, and Paquette et al. (2022) derived Nesterov's rates under the distribution. In this paper we take a more general view and consider Beta distribution.

Definition 3.2. The (generalized) Beta distribution with parameters τ, ξ and scale L are given by the (non-normalized) probability density function

$$\mathrm{d}\mu_{\tau,\xi}(\lambda) \propto \lambda^{\xi}(L - \lambda)^{\tau}\,\mathrm{d}\lambda\,, \tag{12}$$

where $\mathrm{d}\lambda$ denotes the standard Lebesgues measure on $\mathbf{R}$. This family of distributions generalizes the Marchenko-Pastur distribution, and both have similar concentrations near 0 when $\xi \approx -1/2$.

The optimal method w.r.t. $\mu_{\tau,\xi}$ and objective l is associated to a shifted Jacobi polynomial $\tilde{P}_t^{\alpha,\beta}$ with $\beta = \xi + l + 1, \alpha = \tau$. This is a direct implication of the definition of the Jacobi polynomials as the orthogonal family w.r.t. the weights $(1 - \lambda)^{\alpha}(1 + \lambda)^{\beta}$. When $\alpha = \beta = -1/2$, we retrieve the *Chebyshev Method* (Flanders & Shortley, 1950). It is also related to the "ν-method" of Brakhage (1987), who considers the conjugate gradient algorithm applied to a matrix with a Gamma spectral distribution ($\beta = -1/2$).

We name the following method the *Generalized Chebyshev Method* (GCM).

Algorithm 1 Generalized Chebyshev Method GCM(α, β)

```
**Inputs**: Initial vector $\boldsymbol{x}_0$, distribution parameters $\alpha, \beta$, $L$ =largest eigenvalue of $\boldsymbol{H}$ .
$\boldsymbol{x}_{-1} \leftarrow \boldsymbol{x}_0, \delta_0 \leftarrow 0$
**for** $t = 1, \dots, T$ **do**
    $a_t \leftarrow -\frac{2\left(\beta^2+\alpha\beta+(2t+1)(\alpha+\beta)+2t^2+2\right)(2t+\alpha+\beta+1)}{2(t+1)(t+\alpha+\beta+1)(2t+\alpha+\beta)}$
    $b_t \leftarrow \frac{(2t+\alpha+\beta+1)(2t+\alpha+\beta+2)}{L(t+1)(t+\alpha+\beta+1)}$
    $\gamma_t \leftarrow -\frac{(t+\alpha)(t+\beta)(2t+\alpha+\beta+2)}{(t+1)(t+\alpha+\beta+1)(2t+\alpha+\beta)}$
    $\delta_t \leftarrow \frac{1}{a_t+\gamma_t\delta_{t-1}}$
    $\boldsymbol{x}_t \leftarrow \boldsymbol{x}_{t-1}+(\delta_t a_t-1)(\boldsymbol{x}_{t-1}-\boldsymbol{x}_{t-2})+\delta_t b_t \nabla f(\boldsymbol{x}_{t-1})$
**end for**
```

We consider the Nesterov's method used in (Paquette et al., 2022), which is defined by the iterations:

$$\boldsymbol{x}_{t+1} = \boldsymbol{y}_t - \frac{1}{L}\nabla f(\boldsymbol{y}_t), \tag{13}$$

$$\boldsymbol{y}_{t+1} = \boldsymbol{x}_{t+1} + \frac{t}{t+3}(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t)\,. \tag{14}$$

We also consider the Laguerre method (Pedregosa & Scieur, 2020), which is optimal w.r.t. $\mathrm{d}\mu(\lambda) = \frac{\lambda^{\alpha}e^{-\lambda}}{\Gamma(\alpha+1)}\,\mathrm{d}\lambda$, taking α as a parameter. This method is proposed to optimize non-smooth functions.

Both these methods are generalizations of the ones in (Pedregosa & Scieur, 2020). We show that Algorithm 1 corresponded to polynomials $\tilde{P}_t^{\alpha,\beta}$ and derive the Laguerre method in Appendix B.

Remark 3.3. 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

Throughout the rest of the paper, we make the following assumption about the spectral distributions:

Assumption. Let $\nu_{\tau,\xi}$ be a 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 characterize our distributions of interest only in terms of the behavior at the edges. This assumption is sufficient to determine the asymptotic convergence of algorithms. This (mild) assumption excludes only distributions that decay exponentially near their edges, in which case they effectively behave as strongly convex problems. Moreover, this assumption allows considering a broader class of problems than previous work that assumed complete knowledge of the spectrum of the Hessian.

The ξ parameter measures 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, as samples of low eigenvalues are rarer and rarer with increasing ξ.

We show that the coefficients ξ, τ determine 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 for all distributions with the same concentrations.

Theorem 4.1 (GCM average-case rates)**.** *The Generalized Chebyshev Method with parameters (α, β) applied to a problem with expected spectral distribution $\nu_{\tau,\xi}$ has average-case rates $\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] \sim L \cdot C^{\alpha,\beta}_{1,\nu} t^{e_1}$ and $\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|_2^2] \sim L^2 \cdot C^{\alpha,\beta}_{2,\nu} t^{e_2}$, with the exponents*

$$e_1 = \begin{cases} -1-2\beta & \textit{if} \quad \alpha < \tau + \frac{1}{2} \quad \wedge \quad \beta < \xi + \frac{3}{2}, \\ -2(\xi+2)\log t & \textit{if} \quad \alpha = \tau + \frac{1}{2} \quad \wedge \quad \beta = \xi + \frac{3}{2}, \\ 2(\max\{\alpha - \beta - \tau, -\xi - 1\} - 1) & \textit{otherwise}, \end{cases}$$

$$e_2 = \begin{cases} -1-2\beta & \textit{if} \quad \alpha < \tau + \frac{1}{2} \quad \wedge \quad \beta < \xi + \frac{5}{2}, \\ -2(\xi+3)\log t & \textit{if} \quad \alpha = \tau + \frac{1}{2} \quad \wedge \quad \beta = \xi + \frac{5}{2}, \\ 2(\max\{\alpha - \beta - \tau, -\xi - 2\} - 1) & \textit{otherwise}, \end{cases}$$

*where $\wedge$ denotes the logical "*and*" operator and $C^{\alpha,\beta}_{\cdot,\nu}$ are distribution dependent constants.*

Proof Sketch. We first compute the convergence rates assuming that the ESD is a generalized Beta distribution with parameters τ, ξ. For this we use Theorem 2.5 and Lemma C.1. For a general $\nu_{\tau,\xi}$, with the aid of Lemma C.2, we then show that the mass away from the edges is negligible, i.e.,

$$\int_{\epsilon}^{L-\epsilon} P_t^{\alpha,\beta}(\lambda)^2 \lambda^l d\nu_{\tau,\xi}\,, \tag{15}$$

is $O(t^{-1-2\beta}) = O(t^{e_\cdot}), \forall \epsilon > 0$. As the mass near the edges must be similar to the one for the Beta weights, the result follows. □

Theorem 4.1, illustrated by Fig. 2, shows that overestimating β and underestimating α will still leave us with the optimal asymptotic rates. So a good rule of thumb for calibrating the algorithm is to use high β and low α.

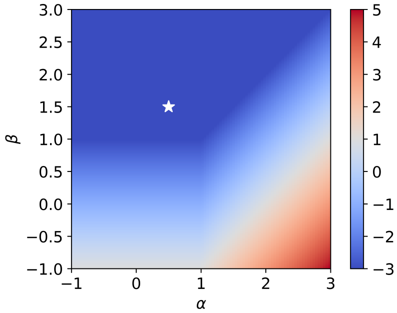

Figure 2. The figure illustrates the robustness of the Generalized Chebyshev Method with parameters (α, β) for a *fixed problem* corresponding to the Marchenko-Pastur distribution $(\tau = \frac{1}{2}, \xi = -\frac{1}{2})$. The color represents the exponent a of the average-case rate $O(t^a)$ of the method for different values of α and β. The white star represents the optimal tuning and the blue area is the set of parameters for which the method converges. Note that a large region guarantees the same optimal asymptotic rate.

Method	**Parameters** (τ, ξ)	
	$(\frac{1}{2}, \frac{1}{2})$	$(\frac{1}{2}, -\frac{1}{2})$
GCM$(\alpha = \frac{1}{2}, \beta = \frac{5}{2})$	t^{-5}	t^{-3}
GCM$(\alpha = \frac{1}{2}, \beta = \frac{3}{2})$	t^{-4}	t^{-3}
Nesterov	t^{-4}	$t^{-3}\log t$
Gradient descent	$t^{\frac{-5}{2}}$	$t^{\frac{-3}{2}}$

Table 2. The table compares the asymptotic average-case rates for the function-value for different methods and pairs (τ, ξ).

Theorem 4.2 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 4.2 (Optimal Rates)**.** *Consider the distribution $\nu_{\tau,\xi}$. The optimal asymptotic average-case rates for $\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)]$ and $\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|_2^2]$ are attained by the GCM with parameters $(\tau, \xi + 2)$ and $(\tau, \xi + 3)$, respectively, and read*

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

$$\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|_2^2] = \Theta(t^{-2(\xi+3)})\,. \tag{17}$$

Proof Sketch. When $\nu_{\tau,\xi}$ is a Beta distribution, the result follows from Theorem 4.1.

For a general $\nu_{\tau,\xi}$, we consider the optimal method for this expected spectral distribution. We show that its performance on the corresponding Beta weight $\mu_{\tau,\xi}$ is similar to that of $\nu_{\tau,\xi}$, as the integrals concentrate on the edges. It follows that the optimal performance for $\nu_{\tau,\xi}$ is asymptotically the same as for $\mu_{\tau,\xi}$. □

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 GCM is robust for a suboptimal choice of parameter β up to $1/2$ below the optimal choice and infinitely above.

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

Proposition 4.3 (GCM worst-case rates)**.** *Let f be a convex, L-smooth quadratic function. Then, for the Generalized Chebyshev Method with parameters (α, β), we have worst-case rates $f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) \leq C_1 L t^{e_3}$ and $\|\nabla f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)\| \leq C_2 L^2 t^{e_4}$, with the exponents*

$$e_3 = \begin{cases} 2(\alpha-\beta) & \textit{if} \quad \alpha > \beta - 1, \\ -1-2\beta & \textit{if} \quad \alpha \leq \beta - 1 \wedge \beta \leq 1/2, \\ -2 & \textit{otherwise}, \end{cases} \tag{18}$$

$$e_4 = \begin{cases} 2(\alpha-\beta) & \textit{if} \quad \alpha > \beta - 2, \\ -1-2\beta & \textit{if} \quad \alpha \leq \beta - 2 \wedge \beta \leq 3/2, \\ -4 & \textit{otherwise}. \end{cases} \tag{19}$$

The logical and is noted $\wedge$.

For common choice of α, β, i.e. $\beta \geq \frac{1}{2}$, $\alpha \leq \beta - 1$, the rate of decay of the function suboptimality 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 extend the proof of Paquette et al. (2022, Lemma B.2) for the Nesterov method under the Marchenko-Pastur distribution.

Theorem 4.4 (Nesterov average-case rates)**.** *Consider the distribution $\nu_{\tau,\xi}$. Then for the Nesterov method, we have average-case rates*

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] \sim C'_{1,\nu} \begin{cases} t^{-2(\xi+2)} & \textit{if}\ \xi < -1/2, \\ t^{-3}\log t & \textit{if}\ \xi = -1/2, \\ t^{-(\xi+7/2)} & \textit{if}\ \xi > -1/2, \end{cases} \tag{20}$$

$$\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|_2^2] \sim C'_{2,\nu} t^{-(\xi+9/2)}. \tag{21}$$

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

Theorem 4.5 (Gradient descent average-case rates)**.** *Consider the distribution $\nu_{\tau,\xi}$. Then for gradient descent*

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] = \Theta(t^{-(\xi+2)}), \tag{22}$$

$$\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|_2^2] = \Theta(t^{-(\xi+3)}). \tag{23}$$

Proof Sketch. From the simple form of the polynomial associated with gradient descent, we have a closed-form (Eq. (78)) expression for the performance of gradient descent on the Beta weights. For a general $\nu_{\tau,\xi}$, the mass that is ϵ away from 0 is $O((1-\epsilon)^{2t})$; thus, the performance of gradient descent is asymptotically the same as on the Beta weights. □

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

We now consider the optimal rates for a Gamma distribution.

Theorem 4.6 (Laguerre method rates)**.** *Let $\alpha > -1$ and μ_α be a Gamma distribution, i.e. $\mathrm{d}\mu_\alpha(\lambda) = \lambda^\alpha e^{-\lambda}/\Gamma(\alpha+1)\,\mathrm{d}\lambda$. The optimal rates are given by the Laguerre method of appropriate tuning and*

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] = \Theta(t^{-(\alpha+2)}). \tag{24}$$

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

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

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

Remark 4.7. 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., 2022)

5. Experiments

We generate random quadratic problems in two different ways. First, by sampling $\boldsymbol{H} = \boldsymbol{X}\boldsymbol{X}^\top$ where the entries of $\boldsymbol{X}$ are i.i.d. Gaussian. Since the ESD of $\boldsymbol{H}$ converges to the Marchenko-Pastur distribution as the number of samples and features, go to infinity. This corresponds to a Beta distribution with parameters $(\tau, \xi) = (1/2, -1/2)$.

The other process we use to generate random quadratic problems is by sampling $\Lambda \in \mathbb{R}^d$ from the corresponding Beta distribution and taking $\boldsymbol{H} = \boldsymbol{U}\,\mathbf{diag}(\Lambda)\boldsymbol{U}^\top$, where $\boldsymbol{U}$ is an independently sampled orthonormal matrix.

We let $\boldsymbol{x}^\star = \mathbf{0}$ and sample $\boldsymbol{x}_0$ from a standard Gaussian distribution. In all experiments, we use the problem's instance largest eigenvalue to calibrate each method (e.g., gradient descent's step size is $1/L$).

Our theoretical rates in Theorem 4.5 and Theorem 4.4 respectively for the Nesterov method and gradient descent are precise under the approximate range $-1 < \xi < 0$ as we

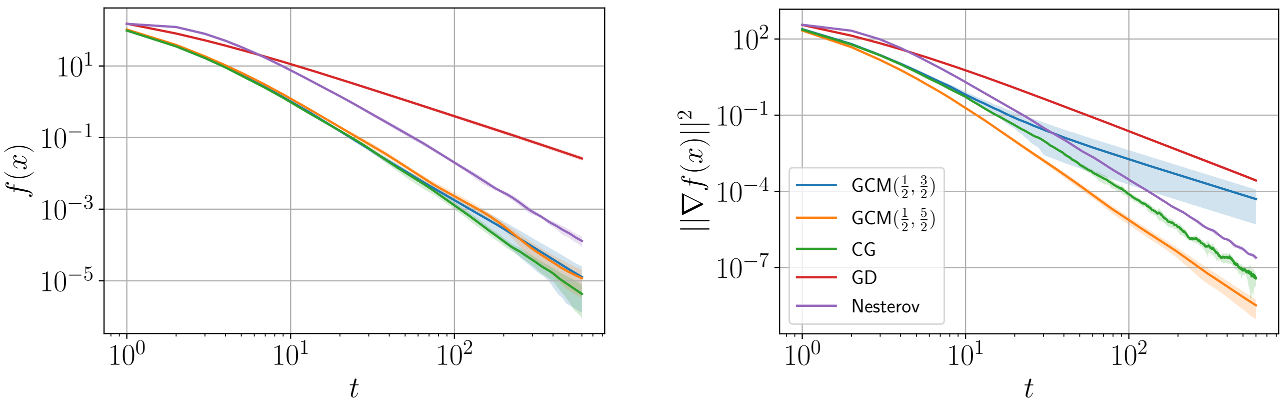

Figure 3. Rates for a synthetic problem, simulating the Marchenko-Pastur distribution. CG stands for conjugate gradient and GD for gradient descent. Shades are standard deviation for 8 different samplings of the distribution and of the random initialization. Note that both tunings of the GCM achieve performance in function value very close to the one of Conjugate Gradient, which is optimal for every draw of the problem.

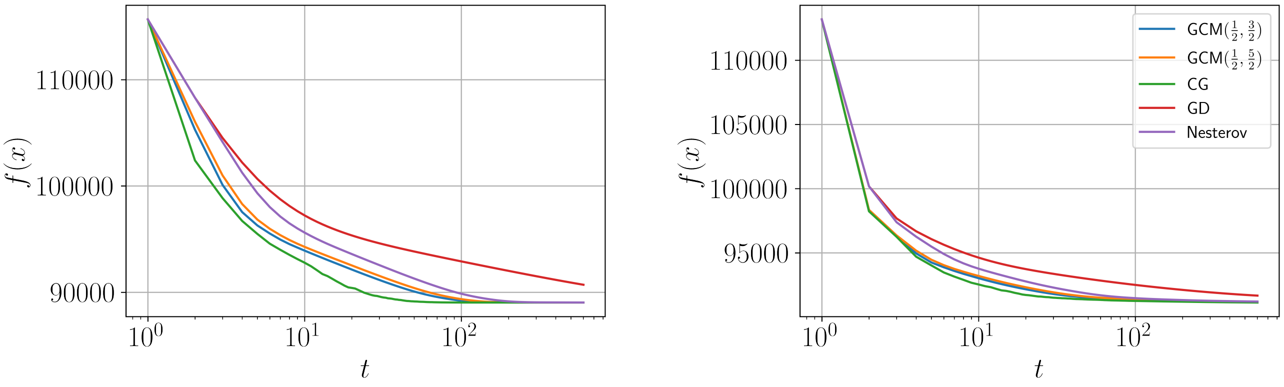

Figure 4. Function values for different methods where the data comes from CIFAR-10 Inception features *(left)* and MNIST features *(right)*. The properly tuned GCM achieves remarkable performance under these non-synthetic spectra.

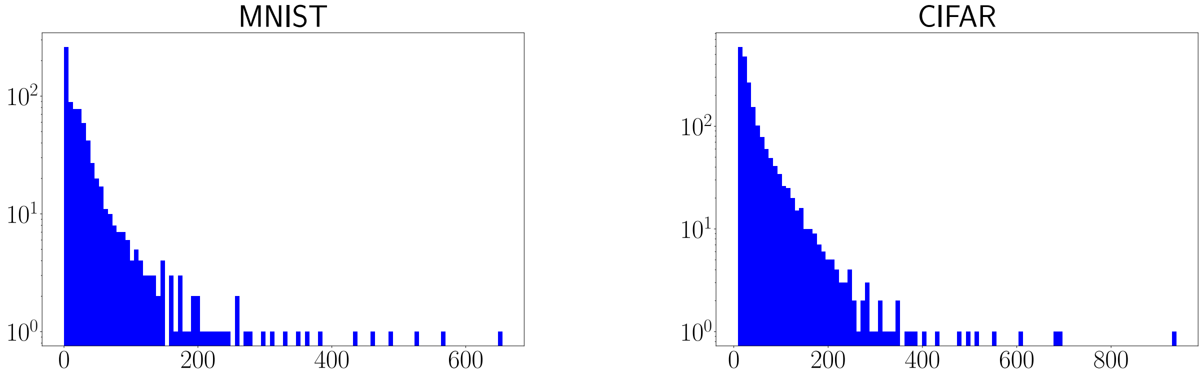

Figure 5. Empirical spectrum for the covariance matrix of the features. On the left we considered the MNIST raw features and on the right we considered the features of an inception net applied on the CIFAR10 dataset.

show in Figure 6. Distributions with higher ξ need many samples. Otherwise, they behave as strongly convex functions.

The Generalized Chebyshev Method shows large variance for the different objectives r_t in some settings. Consider $\beta^\star$ the optimal β value for a given ξ. If $\beta < \beta^\star$ or ξ is low the function value $f(\boldsymbol{x}_t)$, when seen as random variable w.r.t. the random initialization is slow to converge with increasing dimension. This is shown in appendix D.

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

6. Conclusion and further work

In this paper, we've established that the asymptotic 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.

Acknowledgements

Gauthier Gidel is supported by an IVADO grant and a Canada CIFAR AI Chair.

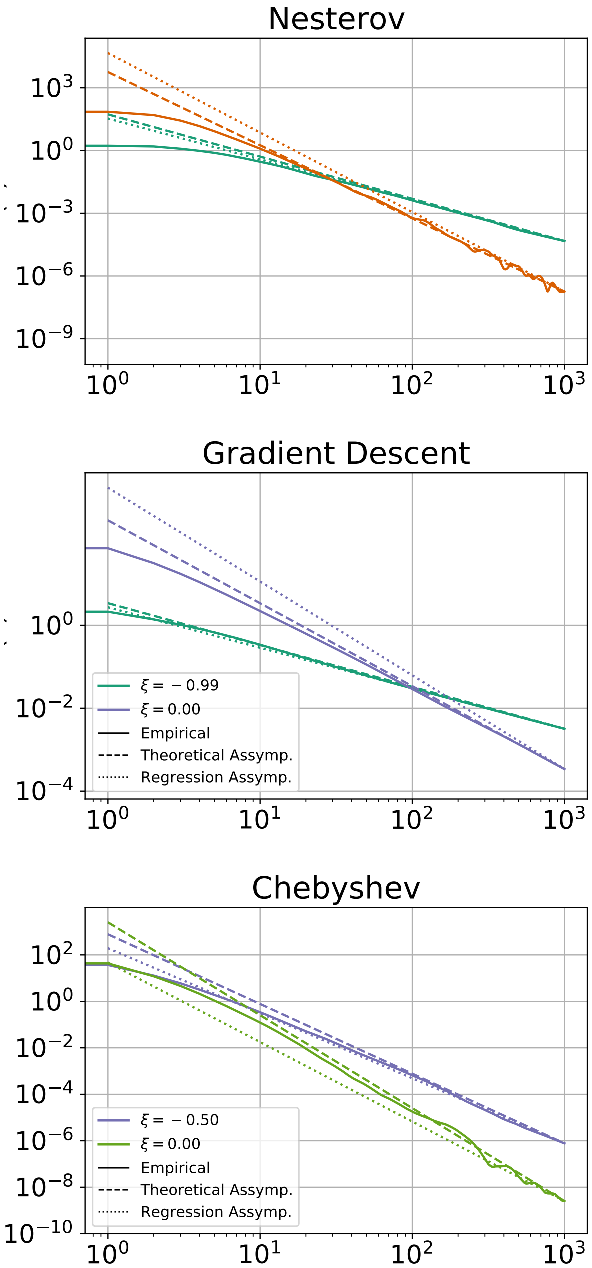

Figure 6. Comparison between experiments run on synthetic Beta distribution and theoretical asymptotic. Shades are standard deviation for 8 different samplings of the distribution and of the random initialization. Regression Assymptotic is the linear regression of the last 700 values, in logarithmic scale . Y-axis is the function value.

References

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

Brakhage, H. On ill-posed problems and the method of conjugate gradients. In *Inverse and ill-posed Problems*. Elsevier, 1987.

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

Flanders, D. A. and Shortley, G. Numerical determination of fundamental modes. *Journal of Applied Physics*, 1950.

Hestenes, M. R., Stiefel, E., et al. *Methods of conjugate gradients for solving linear systems*. NBS Washington, DC, 1952.

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

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

Paquette, C., van Merriënboer, B., Paquette, E., and Pedregosa, F. Halting time is predictable for large models: A universality property and average-case analysis. *Foundations of Computational Mathematics*, 2022.

Pedregosa, F. and Scieur, D. Acceleration through spectral density estimation. In *International Conference on Machine Learning*. PMLR, 2020.

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

Scieur, D. and Pedregosa, F. Universal average-case optimality of Polyak momentum. *International Conference on Machine Learning*, 2020.

Szegö, G. Orthogonal polynomials. In *American Mathematical Society Colloquium Publications*, 1975.

Van Assche, W. Weak convergence of orthogonal polynomials. *Indagationes Mathematicae*, 1995.

A. Proofs of Section 4

Theorem 2.5. *Let $\boldsymbol{x}_t$ be generated by a first-order method with associated residual polynomial P_t, μ be the ESD, and $\mathbb{E}[(\boldsymbol{x}_0 - \boldsymbol{x}^\star)(\boldsymbol{x}_0 - \boldsymbol{x}^\star)^\top] = R^2\boldsymbol{I}$ for some constant R. Then we have the following identities for different convergence metrics:*

$$\mathbb{E}[\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2] = R^2 \int P_t^2(\lambda)\,\mathrm{d}\mu(\lambda)\,, \tag{5}$$

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] = \frac{R^2}{2}\int P_t^2(\lambda)\lambda\,\mathrm{d}\mu(\lambda)\,, \tag{6}$$

$$\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|_2^2] = R^2 \int P_t^2(\lambda)\lambda^2\,\mathrm{d}\mu(\lambda)\,. \tag{7}$$

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

$$\mathbb{E}_H[\mathrm{tr}(g(\boldsymbol{H}))] = \int g(\lambda)\,\mathrm{d}\mu(\lambda). \tag{25}$$

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 ESD:

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 &= \mathbb{E}[\mathrm{tr}((\boldsymbol{x}_0 - \boldsymbol{x}^\star)^\top P_t(\boldsymbol{H})^2(\boldsymbol{x}_0 - \boldsymbol{x}^\star))], && (26)\\
&= \mathbb{E}_{\boldsymbol{H},\boldsymbol{x}_0 - \boldsymbol{x}^\star}[\mathrm{tr}(P_t(\boldsymbol{H})^2(\boldsymbol{x}_0 - \boldsymbol{x}^\star)(\boldsymbol{x}_0 - \boldsymbol{x}^\star)^\top] && (27)\\
&= \mathbb{E}_{\boldsymbol{H}}\left[\mathrm{tr}(P_t(\boldsymbol{H})^2\mathbb{E}_{\boldsymbol{x}_0 - \boldsymbol{x}^\star}[(\boldsymbol{x}_0 - \boldsymbol{x}^\star)(\boldsymbol{x}_0 - \boldsymbol{x}^\star)^\top]))\right], && (28)\\
&= R^2\mathbb{E}_{\boldsymbol{H}}[P_t(\mathrm{tr}(\boldsymbol{H}))^2] = R^2\int P_t(\lambda)^2\,\mathrm{d}\mu(\lambda)\,. && (29)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] &= \frac{R^2}{2}\mathbb{E}[\mathrm{tr}((\boldsymbol{x}_0 - \boldsymbol{x}^\star)^\top P_t(\boldsymbol{H})\boldsymbol{H}P_t(\boldsymbol{H})(\boldsymbol{x}_0 - \boldsymbol{x}^\star))] && (30)\\
&= \mathbb{E}_{\boldsymbol{H}}[\mathrm{tr}((\lambda P_t)(\boldsymbol{H})^2)] = \frac{R^2}{2}\int \lambda P_t(\lambda)^2\,\mathrm{d}\mu(\lambda), && (31)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}\|\nabla f(\boldsymbol{x}_t))\|^2 &= \mathbb{E}[\mathrm{tr}((\boldsymbol{x}_0 - \boldsymbol{x}^\star)^\top P_t(\boldsymbol{H})\boldsymbol{H}^2P_t(\boldsymbol{H})(\boldsymbol{x}_0 - \boldsymbol{x}^\star))] && (32)\\
&= \mathbb{E}_{\boldsymbol{H}}[\mathrm{tr}((\lambda^2 P_t)(\boldsymbol{H})^2)] = R^2\int \lambda^2 P_t(\lambda)^2\,\mathrm{d}\mu(\lambda)\,. && (33)
\end{aligned}$$

□

Proposition A.1. *Let ν be a distribution with continuous support and P_t^l be defined as*

$$P_t^l \overset{def}{=} \arg\min_{P_t(0)=1}\int P_t^2(\lambda)\lambda^l d\nu(\lambda)\,. \tag{8}$$

Then $(P_t^l)_{t=0,1,\ldots}$ is the family of residual orthogonal polynomials w.r.t. to $\lambda^{l+1}d\nu$.

Proof. This is a direct application of the following lemma w.r.t. to the distribution ν defined as $d\nu(\lambda) = \lambda^l\,\mathrm{d}\mu(\lambda)$.

Lemma A.2. *(Fischer, 1996) The residual polynomial of degree t that minimizes $\int P_t^2(\lambda)d\nu(\lambda)$ is given by the degree t residual orthogonal polynomial with respect to the weight function $\lambda d\nu(\lambda)$.*

Let $P_t(\lambda;(a_k))$ be the polynomial of degree k defined by the coefficients a_k and $F((a_k)) = \int \lambda^l P_t^2(\lambda;(a_k))\,\mathrm{d}\mu(\lambda)$. The restriction $P_t(0) = 1$ is expressed as $g(a_k) = a_0 - 1 = 0$. We then consider the minimization problem:

$$\begin{aligned} &\min_{a_k} F(a_k) \\ \text{s.t.}\quad & g(a_k) = 0 \end{aligned}$$

We observe that the problem is convex in the variables a_k and consider the lagrangian $\mathcal{L}(a_k;\alpha) = F(a_k) - \alpha g(a_k)$. By the first-order optimality conditions the derivatives w.r.t. the coefficients $a_k, k > 0$ of the polynomials should be 0:

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

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

□

B. GCM and Laguerre method derivation

We will first state two lemmas that allow us to construct the optimal polynomials.

Lemma B.1. *Let $(\tilde{P}_t)$ be a family polynomials with recurrence*

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

with $\tilde{P}_0$ a constant polynomial and $\tilde{P}_t(0) \neq 0$ for all t. Then,

$$P_t(\lambda) = (a_t + b_t\lambda)P_{t-1}(\lambda) + (1 - a_t)P_{t-2}(\lambda)\,, \tag{34}$$

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

$$a_t = \delta_t \alpha_t, \tag{35}$$

$$b_t = \delta_t \beta_t, \tag{36}$$

$$\delta_t = (\alpha_t + \gamma_t \delta_{t-1}) \qquad (\delta_0 = 0). \tag{37}$$

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 B.2. *Let $(\tilde{P}_t)$ be a family polynomials following*

$$\tilde{P}_t(\lambda) = (\alpha_t + \beta_t\lambda)\tilde{P}_{t-1}(\lambda) + \gamma_t \tilde{P}_{t-2}(\lambda), \tag{38}$$

and define polynomials P_t such that :

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

with $m(\lambda) = a\lambda + b$ a non-singular affine transform. Then P_t follows a recurrence like in eq. (38), with:

$$\alpha'_t = \alpha_t + b\beta_t, \tag{39}$$

$$\beta'_t = a\beta_t\,, \tag{40}$$

$$\gamma'_t = \gamma_t\,. \tag{41}$$

Proof. The result follows from considering Eq. (38) with argument $m^{-1}(\lambda)$. □

These results are sufficient to obtain 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 (Szegö, 1975):

$$\alpha_t = \frac{(2n+\alpha+\beta)(2n+\alpha+\beta-1)}{2n(n+\alpha+\beta)}, \tag{42}$$

$$\beta_t = \frac{(\alpha^2-\beta^2)(2n+\alpha+\beta-1)}{2n(n+\alpha+\beta)(2n+\alpha+\beta-2)}, \tag{43}$$

$$\gamma_t = \frac{-2(n+\alpha-1)(n+\beta-1)(2n+\alpha+\beta)}{2n(n+\alpha+\beta)(2n+\alpha+\beta-2)}\,. \tag{44}$$

We then shift the distribution according to $\eta(x)$, and then transform to the residual ones. We slightly simplify these computations and use remark 2.3 to get Algorithm 1.

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

$$L_t^\alpha(\lambda) = \left(\frac{2t+\alpha-1}{t} - \frac{1}{t}\lambda\right) L_{t-1}^\alpha(\lambda) + \frac{t+\alpha-1}{t} L_{t-2}^\alpha(\lambda)\,. \tag{45}$$

As we don't have to shift the domains, we have only to apply lemma B.1 to get the Laguerre method. Further, we can get a explicit expression for $\delta_t = \frac{t}{t+\alpha}$ by simplifying the expression.

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

C. Proofs of section 4

In the following we will consider shifted versions of the spectral distributions. This shift is the affine function $m(\lambda)$: $[0, L] \to [-1, 1]$ 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 $\mu'(x) = \tilde{\mu}'(m(x))$, then,

$$\int P_t^2(x)\mu'(x)\,\mathrm{d}x \propto \int \tilde{P}_t^2(x)\tilde{\mu}'(x)\,\mathrm{d}x, \tag{46}$$

so all the asymptotics are the same. The Jacobi polynomials $P_t^{\alpha,\beta}$ are those orthogonal w.r.t $\mathrm{d}\mu(x) = (1-x)^\alpha(1+x)^\beta\,\mathrm{d}x$. 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 4.1 (GCM average-case rates)**.** *The Generalized Chebyshev Method with parameters (α, β) applied to a problem with expected spectral distribution $\nu_{\tau,\xi}$ has average-case rates $\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] \sim L \cdot C_{1,\nu}^{\alpha,\beta} t^{e_1}$ and $\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|_2^2] \sim L^2 \cdot C_{2,\nu}^{\alpha,\beta} t^{e_2}$, with the exponents*

$$e_1 = \begin{cases} -1-2\beta & \textit{if} \quad \alpha < \tau + \frac{1}{2} \quad \wedge \quad \beta < \xi + \frac{3}{2}, \\ -2(\xi+2)\log t & \textit{if} \quad \alpha = \tau + \frac{1}{2} \quad \wedge \quad \beta = \xi + \frac{3}{2}, \\ 2(\max\{\alpha - \beta - \tau, -\xi - 1\} - 1) & \textit{otherwise}, \end{cases}$$

$$e_2 = \begin{cases} -1-2\beta & \textit{if} \quad \alpha < \tau + \frac{1}{2} \quad \wedge \quad \beta < \xi + \frac{5}{2}, \\ -2(\xi+3)\log t & \textit{if} \quad \alpha = \tau + \frac{1}{2} \quad \wedge \quad \beta = \xi + \frac{5}{2}, \\ 2(\max\{\alpha - \beta - \tau, -\xi - 2\} - 1) & \textit{otherwise}\,, \end{cases}$$

where $\wedge$ denotes the logical "and" operator and $C_{\cdot,\nu}^{\alpha,\beta}$ are distribution dependent constants.

Proof. We will prove that for any α and β, $\xi, \tau > -1$, $l > 0$ and distributions $\nu_{\tau,\xi-l}$, 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):

$$\int_{-1}^{1} \tilde{P}_t^{\alpha,\beta}(x)(1-x)^\alpha(1+x)^\beta \,\mathrm{d}x = \frac{2^{\alpha+\beta+1}}{2t+\alpha+\beta+1} \frac{\Gamma(t+\alpha+1)\Gamma(t+\beta+1)}{\Gamma(t+1)\Gamma(t+\alpha+\beta+1)} = \Theta(t^{-1})\,. \tag{47}$$

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 C.1 (Szegö (1975))**.** *We have,*

$$\int_0^1 (1-x)^\tau P_t^{\alpha,\beta}(x)^2 dx \sim \Theta(h_\tau^\alpha), \tag{48}$$

$$h_\tau^\alpha \overset{def}{=} \begin{cases} t^{2(\alpha-\tau-1)} & \textit{if } \alpha > \tau + 1/2, \\ t^{-1} \log t & \textit{if } \alpha = \tau + 1/2, \\ t^{-1} & \textit{if } \alpha < \tau + 1/2\,. \end{cases} \tag{49}$$

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

$$\int_{-1}^{1} \tilde{P}_t(x)^2(1-x)^\tau(1+x)^\xi dx = \Theta\left(\int_0^1 (1-x)^\tau |\tilde{P}_t^{\alpha,\beta}(x)|^2 dx\right) + \Theta\left(\int_0^1 (1-x)^\xi |\tilde{P}_t^{\beta,\alpha}(x)|^2 dx\right). \tag{50}$$

We can then show our result for $\mathrm{d}\nu_{\tau,\xi-l}(x) = x^{\xi-l}(L-x)^\alpha \,\mathrm{d}x$ by carefully considering each of the cases on Lemma C.1 and the maximum of each term in eq. 50, 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:

$$\int_0^1 \tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}\nu_{\tau,\xi}(x) = \Theta\left(\int_0^1 (1-x)^\tau \tilde{P}_t^{\alpha,\beta}(x)^2 dx\right). \tag{51}$$

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 C.2 (Van Assche (1995))**.** *Let μ be a measure and (p_t) a family of associated orthonormal polynomials such that 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 we have:

$$\int f(x) p_t^2(x) \,\mathrm{d}\mu(x) \to \frac{1}{\pi} \int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}} dx\,. \tag{52}$$

Let ϵ such that

$$x \geq 1 - \epsilon \Rightarrow |\,\mathrm{d}\nu_{\tau,\xi} - A(1-x)^\tau| \leq B(1-x)^\tau. \tag{53}$$

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

$$\underbrace{\int_0^1 (1-x)^{\tau} \tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}x}_{\Theta(h_\tau^\alpha)} = \underbrace{\int_0^{1-\epsilon} (1-x)^{\tau} \tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}x}_{\Theta(t^{-1})} + \int_{1-\epsilon}^1 (1-x)^{\tau} \tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}x \Rightarrow \tag{54}$$

$$\int_{1-\epsilon}^1 (1-x)^{\tau} \tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}x = \Theta(h_\tau^\alpha), \tag{55}$$

$$\begin{aligned} \int_0^1 \tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}\nu_{\tau,\xi}(x) &= \underbrace{\int_0^{1-\epsilon} \tilde{P}_t^{\alpha,\beta}(x)^2 f(x)(1-x)^{\alpha}(1+x)^{\beta} \,\mathrm{d}x)}_{\Theta(t^{-1})} \\ &\quad + \Theta\left(\underbrace{\int_{1-\epsilon}^1 (1-x)^{\tau} \tilde{P}_t^{\alpha,\beta}(x)^2 \,\mathrm{d}x}_{\Theta(h_\tau^\alpha)}\right). \end{aligned} \tag{56}$$

□

Theorem 4.2 (Optimal Rates)**.** *Consider the distribution $\nu_{\tau,\xi}$. The optimal asymptotic average-case rates for $\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)]$ and $\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|_2^2]$ are attained by the GCM with parameters $(\tau, \xi+2)$ and $(\tau, \xi+3)$, respectively, and read*

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

$$\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|_2^2] = \Theta(t^{-2(\xi+3)})\,. \tag{17}$$

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

$$\min_{P_t(0)=1} \int P_t^2(\lambda)\lambda^l d\nu(\lambda) = \Theta\left(\int_0^l \tilde{P}_t^{\alpha,\beta}(\lambda)^2 (L-\lambda)^{\tau} \lambda^{\xi+l} \,\mathrm{d}\lambda\right) = \Theta(t^{-2(\xi+l+1)}). \tag{57}$$

Showing the second equality is easy by considering theorem 4.1, and that is further the minimum asymptotic rate for the Beta distribution $\mu_{\tau,\xi}$.
By setting p_t^ν and $P_t^\nu = \frac{p_t^\nu}{p_t^\nu(0)}$ the optimal orthonormal and residual and polynomials w.r.t. ν we show that P_t^ν must have the same rate on $\mu_{\tau,\xi}$ as it does on ν, thus the optimal rate of ν cannot be lower than the optimal rate of $\mu_{\tau,\xi}$. Indeed, setting ϵ_1, ϵ_2 as in eq. 53:

$$\int_{1-\epsilon_2}^1 P_t^\nu(x)^2 d\nu(x) = \Theta\left(\int_{1-\epsilon_2}^1 P_t^\nu(x)^2 \,\mathrm{d}\mu_{\tau,\xi}(x)\right), \tag{58}$$

$$\int_{-1}^{-1+\epsilon_1} P_t^\nu(x)^2 d\nu(x) = \Theta\left(\int_{-1}^{-1+\epsilon_1} P_t^\nu(x)^2 \,\mathrm{d}\mu_{\tau,\xi}(x)\right), \tag{59}$$

$$\int_{-1+\epsilon_1}^{1-\epsilon_2} P_t^\nu(x)^2 d\nu(x) = \Theta\left(\int_{-1+\epsilon_1}^{1-\epsilon_2} P_t^\nu(x)^2 \,\mathrm{d}\mu_{\tau,\xi}(x)\right) = \Theta\left(\frac{1}{p_t^\nu(-1)^2}\right), \tag{60}$$

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

Proposition C.3 (GCM worst-case rates)**.** *Let f be a convex, L-smooth quadratic function. Then, for the Generalized Chebyshev Method with parameters (α, β), we have worst-case rates $f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) \leq C_1 L t^{e_3}$ and $\|\nabla f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)\| \leq$*

$C_2 L^2 t^{e_4}$, with the exponents

$$e_3 = \begin{cases} 2(\alpha-\beta) & \textit{if} \quad \alpha > \beta - 1, \\ -1-2\beta & \textit{if} \quad \alpha \leq \beta - 1 \wedge \beta \leq 1/2, \\ -2 & \textit{otherwise}, \end{cases} \tag{18}$$

$$e_4 = \begin{cases} 2(\alpha-\beta) & \textit{if} \quad \alpha > \beta - 2, \\ -1-2\beta & \textit{if} \quad \alpha \leq \beta - 2 \wedge \beta \leq 3/2, \\ -4 & \textit{otherwise}. \end{cases} \tag{19}$$

The logical and is noted $\wedge$.

Proof. 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:

$$v(\alpha,\beta,l) = \begin{cases} 2(\alpha-\beta) & \text{if} \quad \alpha > \beta - l \\ -1-2\beta & \text{if} \quad \alpha \leq \beta - l \quad \beta \leq l - \frac{1}{2}, \\ -2l, & \text{if} \quad \alpha \leq \beta - l \quad \beta \geq l - \frac{1}{2}. \end{cases} \tag{61}$$

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

$$\tilde{P}_t^{\alpha,\beta}(\cos\theta) = \begin{cases} O(t^{-1/2}) & \text{if} \quad \alpha < -\frac{1}{2}, \\ O(t^{\alpha}) & \text{if} \quad \alpha \geq -\frac{1}{2} \wedge 0 \leq \theta \leq ct^{-1}, \\ \theta^{-\alpha-1/2} O(t^{-1/2}) & \text{if} \quad \alpha \geq -\frac{1}{2} \wedge \theta > ct^{-1}. \end{cases} \tag{62}$$

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

$$\sup_{x\in[0,L]} x^l P_t^{\alpha,\beta}(x)^2 = \Theta\left(\max\left\{\sup_{x\in[0,1]} x\tilde{P}_t^{\alpha,\beta}(x)^2, \sup_{x\in[0,1]} (1-x)^l \tilde{P}_t^{\beta,\alpha}(x)^2\right\}\right). \tag{63}$$

The $(1-x)^l$ term, corresponds to $(2\sin(\frac{\theta}{2}))^{2l}$ in the variable θ, which is $O(\theta^{2l})$. The rest follows from carefully considering the expressions given by eq. 62. □

Theorem 4.4 (Nesterov average-case rates)**.** *Consider the distribution $\nu_{\tau,\xi}$. Then for the Nesterov method, we have average-case rates*

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] \sim C'_{1,\nu} \begin{cases} t^{-2(\xi+2)} & \textit{if}\, \xi < -1/2, \\ t^{-3}\log t & \textit{if}\, \xi = -1/2, \\ t^{-(\xi+7/2)} & \textit{if}\, \xi > -1/2, \end{cases} \tag{20}$$

$$\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|_2^2] \sim C'_{2,\nu} t^{-(\xi+9/2)}. \tag{21}$$

Proof. We will prove:

$$\int_0^1 P_t^{\text{Nes}}(\lambda)^2 \lambda^l \, \mathrm{d}\nu_{\tau,\xi-l} \sim C'_{\nu} \begin{cases} t^{-2(\xi+1)} & \text{if } 0 < \xi < 1/2, \\ t^{-3}\log t & \text{if } \xi = 1/2, \\ t^{-(\xi+5/2)} & \text{if } \xi > 1/2. \end{cases} \tag{64}$$

(Paquette et al., 2022) has shown that the nesterov polynomials P_t are asymptotically, in t:

$$P_t(\lambda) \sim \frac{2J_1(t\sqrt{\alpha\lambda})}{t\sqrt{\alpha\lambda}} e^{-\alpha\lambda t/2}, \tag{65}$$

in the sense that:

$$\int_0^1 u^l \left[P_t^{\tilde{2}}(u) - \frac{4J_1^2(t\sqrt{u})}{t^2 u} e^{-ut}\right] \mathrm{d}\mu_{MP}(u) = O(t^{-(l+25/12))}). \tag{66}$$

The arguments can be easily used to show that such an integral is $O(t^{-(\alpha+l+31/12)})$ when evaluated w.r.t. 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 we will bound the others.
We consider first, the "middle region",i .e. for some $\epsilon > 0$:

$$\int_{\frac{\epsilon}{t}}^{\frac{\epsilon}{\sqrt{t}}} u^{\xi} \frac{4J_1^2(t\sqrt{u})}{t^2 u} e^{-ut} \,\mathrm{d}u. \tag{67}$$

We note the asymptotic for J_1^2:

$$J_1^2(\sqrt{tv}) \sim \frac{1}{\pi\sqrt{tv}}(1 + \cos(2\sqrt{tv} + 2\gamma)). \tag{68}$$

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

$$\begin{aligned} \int_{\frac{\epsilon}{t}}^{\frac{\epsilon}{\sqrt{t}}} u^{\xi} \frac{4J_1^2(t\sqrt{u})}{t^2 u} e^{-ut} \,\mathrm{d}u &= \Theta\left(t^{-2-\xi} \int_{\epsilon}^{\infty} v^{\xi-1} J_1^2(\sqrt{tv}) e^{-v} \,\mathrm{d}v\right), &(69)\\ &= \Theta\left(t^{-2-\xi} \int_{\epsilon}^{\infty} v^{\xi-1} \frac{1}{\pi\sqrt{tv}} e^{-v} \,\mathrm{d}v\right), &(70)\\ &= \Theta\left(t^{-\frac{5}{2}-\xi} \underbrace{\int_{\epsilon}^{\infty} v^{\xi-\frac{3}{2}} \frac{1}{\pi\sqrt{tv}} e^{-v} \,\mathrm{d}v}_{\Gamma(\xi-\frac{1}{2},\epsilon)}\right). &(71) \end{aligned}$$

Where, from the Riemann-Lebesgue, the cosine term goes to 0 lemma and Γ is the incomplete Gamma function.
The mass from the rightmost region ,i.e, the term corresponding to the interval $[\epsilon t^{-1/2}, 1]$ is exponentially small. Indeed, because of the exponential e^{-ut} it is $O(e^{-\epsilon\sqrt{t}})$. Lastly, we have for the $[0, \frac{\epsilon}{t}]$ region, doing the change of variables $v = t^2 u$:

$$\int_0^{\frac{\epsilon}{t}} u^{\xi} \frac{4J_1^2(t\sqrt{u})}{t^2 u} e^{-ut} \,\mathrm{d}u = \Theta\left(t^{-2(\xi+1)} \int_0^{t\epsilon} v^{\xi} \frac{J_1^2(\sqrt{v})}{v} e^{-\frac{v}{t}} \,\mathrm{d}v\right). \tag{72}$$

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

$$\begin{aligned} \frac{J_1^2(\sqrt{v})}{v} &\sim \frac{1}{4}, & v \to 0, &\qquad(73)\\ \frac{J_1^2(\sqrt{v})}{v} &= O(v^{-3/2}), & v \to \infty, &\qquad(74) \end{aligned}$$

so we divide this integral aswell:

$$\begin{aligned} t^{-2(\xi+1)} \int_1^{t\epsilon} v^{\xi} \frac{J_1^2(\sqrt{v})}{v} e^{-\frac{v}{t}} \,\mathrm{d}v, &= \Theta\left(t^{-2(\xi+1)} \int_{\epsilon}^{t\epsilon} v^{\xi-\frac{3}{2}} \,\mathrm{d}v\right) = \Theta\left(I_{\xi}(t) t^{-\xi-\frac{5}{2}}\right) &(75)\\ t^{-2(\xi+1)} \int_0^1 v^{\xi} \frac{J_1^2(\sqrt{v})}{v} e^{-\frac{v}{t}} \,\mathrm{d}v &= \Theta\left(t^{-2(\xi+1)} \int_0 \epsilon^1 v^{\xi} \,\mathrm{d}v\right) = \Theta\left(t^{-2(\xi+1)}\right), &(76) \end{aligned}$$

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 4.5 (Gradient descent average-case rates)**.** *Consider the distribution $\nu_{\tau,\xi}$. Then for gradient descent*

$$\begin{aligned} \mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] &= \Theta(t^{-(\xi+2)}), &(22)\\ \mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|_2^2] &= \Theta(t^{-(\xi+3)}). &(23) \end{aligned}$$

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

$$\int_0^1 (1-\lambda)^{2t} \lambda^l \,\mathrm{d}\nu_{\tau,\xi-l} = \Theta(t^{-(\xi+l+1)}). \tag{77}$$

We know, for the Beta weights, that:

$$\int_0^1 (1-\lambda)^{2t+\tau}\lambda^{\xi+l}\,\mathrm{d}\lambda = \frac{\Gamma(l+\xi+1)\Gamma(2t+\tau+1)}{\Gamma(2t+l+\xi+\tau+2)} = \Theta(t^{-(\xi+l+1)}). \tag{78}$$

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

$$\int_\epsilon^1 (1-\lambda)^{2t+\tau}\lambda^{\xi+l}\,\mathrm{d}\lambda = \mathcal{O}((1-\epsilon)^{2t}), \tag{79}$$

then:

$$\int_\epsilon^1 (1-\lambda)^{2t}\lambda^{l}\,\mathrm{d}\nu_{\tau,\xi-l} = \mathcal{O}((1-\epsilon)^{2t}), \tag{80}$$

$$\int_0^\epsilon (1-\lambda)^{2t}\lambda^{l}\,\mathrm{d}\nu_{\tau,\xi-l} = \Theta\left(\int_0^\epsilon (1-\lambda)^{2t+\tau}\lambda^{\xi+l}\,\mathrm{d}\lambda\right) = \Theta(t^{-(\xi+l+1)}). \tag{81}$$

□

Theorem 4.6 (Laguerre method rates)**.** *Let* $\alpha > -1$ *and* μ_α *be a Gamma distribution, i.e.* $\mathrm{d}\mu_\alpha(\lambda) = \lambda^\alpha e^{-\lambda}/\Gamma(\alpha+1)\,\mathrm{d}\lambda$. *The optimal rates are given by the Laguerre method of appropriate tuning and*

$$\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] = \Theta(t^{-(\alpha+2)})\,. \tag{24}$$

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

$$\int L_t^\alpha(x)^2\,\mathrm{d}\mu_\alpha(x) = L_t^\alpha(0) = \binom{n+\alpha}{n} \tag{82}$$

Further [(Szegö, 1975) (5.1.13)]]:

$$\sum_{k=0}^{T} L_t^\alpha(x) = L_t^{\alpha+1}(x). \tag{83}$$

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

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] &= \int P_t^{\alpha+2}(\lambda)^2\,\mathrm{d}\mu_{\alpha+1}(\lambda) = \binom{t+\alpha+2}{t}^{-2}\int L_t^{\alpha+2}\,\mathrm{d}\mu_{\alpha+1}(\lambda),\\
&= \binom{t+\alpha+2}{t}^{-2}\sum_{k=0}^{t}\left[\int L_k^{\alpha+1}(\lambda)\,\mathrm{d}\mu_{\alpha+1}(\lambda)\right],\\
&= \binom{t+\alpha+2}{t}^{-2}\sum_{k=0}^{\top}\binom{k+\alpha+1}{k} = \binom{t+\alpha+2}{t}^{-2}\binom{t+\alpha+2}{t},\\
&= \binom{t+\alpha+2}{t}^{-1} = \Theta(t^{-(\alpha+2)}).
\end{aligned} \tag{84}$$

□

D. Additional Experiments

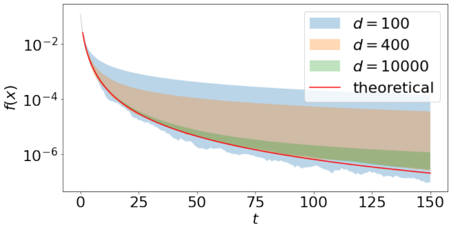

Figure 7. Function value for $\beta < \beta^\star$ for different dimensions of the problem

We observed that in the regimes $\beta < \beta^\star$ and $\xi \approx -1$ the objectives exhibit much higher variance than outside these regimes, where they concentrate tightly around the expected value. Despite this the objectives still converge with increasing dimension.