Universal Average-Case Optimality of Polyak Momentum

Damien Scieur [*1] **Fabian Pedregosa** [*2]

Abstract

We consider the average-case runtime analysis of algorithms for minimizing quadratic objectives. In this setting, and contrary to the more classical worst-case analysis, non-asymptotic convergence rates and optimal algorithms depend on the full spectrum of the Hessian through its expected spectral distribution. Under mild assumptions, we show that these optimal methods converge asymptotically towards Polyak momentum *independently* of the expected spectral density. This makes Polyak momentum universally (i.e., independent of the spectral distribution) asymptotically average-case optimal.

1. Introduction

Average-case analysis of optimization algorithm provides the *expected* complexity of an algorithm over class of problems. As such, it is more representative of the typical behavior of the algorithm than the more traditional worst-case analysis (Nemirovski, 1995; Nesterov, 2013).

However, the average-case complexity has received little attention in the optimization community due to its dependency on the probability distribution that is chosen over the problem class. For example, Pedregosa & Scieur (2019) develop optimal average-case algorithms, but require access to the expected spectral density ($\equiv$ law of a random Hessian eigenvalue). This is highly problem-dependent, resulting in a different algorithm for different problems. Furthermore, it may be hard to know in advance what is the density function of the class of problem we want to solve.

In this paper we focus instead on a single method, Polyak momentum. This method, also known as heavy-ball, was introduced by Polyak (1964) to solve quadratic optimizatio problems. Despite its poor theoretical properties on non-quadratics (Lessard et al., 2016), Polyak momentum is widely used in deep learning due to its good empirical performance (Sutskever et al., 2013; Goodfellow et al., 2016; Goh, 2017).

Contributions. In this paper, study the asymptotic behavior of optimal methods and the *average-case* optimality of Polyak momentum (Polyak, 1964; 1987) for solving quadratic optimization problems. Our main results are the following:

> **Main result 1 (informal)** Consider a class of quadratic problems where the probability density function of a random Hessian eigenvalue verifies $\mu(\lambda) > 0$ for $0 < \ell < \lambda < L$ and zero elsewhere. Then **there is not first-order method** with an asymptotic *expected* (where the expectation is taken over problem instances) convergence rate strictly better than Polyak momentum. See Theorem 3.1 for precise formulation.

> **Main result 2 (informal)** Consider same problem as above. Then, the expected rate of *any* optimal method in average-case converges asymptotically to the known worst-case rate of convergence of Polyak momentum. See Theorem 3.2 for precise formulation.

These results show a surprising universality property of Polyak momentum, which turns out to be optimal in the average-case analysis for a very wide class of problems. The only condition necessary on the problem class is the very mild *gapless* (i.e., non-zero in its support) property on the law of a random eigenvalue. This result brings us one step closer to understanding the remarkable empirical performance of this method.

Paper organization. In §2 we present the framework for the average-case analysis, while in §3 we present the theorem stating the asymptotic optimality of Polyak momentum and an outline of its proof. The rest of the paper is devoted to the proof of this result.

1.1. Related work

The framework of average-case analysis was briefly introduced in (Pedregosa & Scieur, 2019) with the goal of deriving non-asymptotic optimal (average-case) algorithms. In

[*]Equal contribution [1]SAIT AI Lab, Montreal [2]Google Research.

contrast, in this work we focus on the *asymptotic* behavior of optimal algorithms, and their links with Polyak momentum.

Another popular approach to avoid worst-case analysis is the smoothed analysis framework of (Spielman, 2005), for instance applied to conjugate gradient (Menon & Trogdon, 2016) or even manifold optimization (Pumir et al., 2018).

Notation. Throughout the paper we denote vectors in lowercase boldface ($\boldsymbol{x}$), matrices in uppercase boldface letters ($\boldsymbol{H}$). Probability density functions and eigenvalues are written in Greek letters (μ, λ), while polynomials are written in uppercase latin letter (P). We will often omit integration variable, with the understanding that $\int \varphi \, \mathrm{d}\mu$ is a shorthand for $\int \varphi(\lambda)\mu(\mathrm{d}\lambda)$.

2. Average-Case Analysis of Optimization Algorithms

In this section we review the framework for the average-case of quadratic optimization algorithm, which establishes a constructive way to design average-case optimal algorithms.

2.1. Setting

Let $\boldsymbol{H} \in \mathbb{R}^{d\times d}$ be a symmetric positive-definite matrix and $\boldsymbol{x}^\star$ a d-dimensional vector. We consider the quadratic minimization problem

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

where $\boldsymbol{x}_t$ is the t-th update of a first-order method starting from $\boldsymbol{x}_0$.

Remark 1. *Problem* (OPT) *subsumes the quadratic minimization problem* $\min_{\boldsymbol{x}} \boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x} + c$ *but the notation above will be more convenient for our purposes.*

A convenient way to collect statistics on the problem is through the *empirical spectral density*. Let λ_i be the eigenvalues of $\boldsymbol{H}$. We define the empirical spectral density $\hat{\mu}_{\boldsymbol{H}}$ as

$$\hat{\mu}_{\boldsymbol{H}}(\lambda) \stackrel{\text{def}}{=} \sum_{i=1}^{d} \delta_{\lambda_i}(\lambda) \,, \tag{1}$$

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

To solve (OPT), we will consider *first order methods*, which build $\boldsymbol{x}_t$ using a pre-defined linear combination of an initial guess and previous gradients:

$$\boldsymbol{x}_t \in \boldsymbol{x}_0 + \mathbf{span}\{\nabla f(\boldsymbol{x}_0), \, \ldots, \, \nabla f(\boldsymbol{x}_t)\}. \tag{2}$$

There is an intimate link between first order methods and polynomials that allows to simplify the analysis of gradient-based methods. The following proposition showcases this relation by relating the error at iteration t with the error at initialization and the residual polynomial.

Proposition 2.1. *(Hestenes et al., 1952) Let $\boldsymbol{x}_t$ be generated by a first-order method. Then there exists a polynomial P_t of degree t such that $P_t(0) = 1$ that verifies*

$$\boldsymbol{x}_t - \boldsymbol{x}^\star = P_t(\boldsymbol{H})(\boldsymbol{x}_0 - \boldsymbol{x}^\star) \,. \tag{3}$$

Following (Fischer, 1996), we will use thorough this paper the term *residual polynomial*.

Definition 1. ***(Residual polynomial)*** *A polynomial P is called residual if it satisfies $P(0) = 1$.*

Since there is a one-to-one correspondence between optimization algorithm on quadratics and polynomials, we can only use the latter for the design and analysis of novel methods. This simplifies considerably the theoretical study of optimization methods.

2.2. Average-case analysis

Consider now a probability space over instances of problem (OPT). This could be for example the space of different realizations of an experiment, or a collection of different datasets. In this case the error $\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|$ becomes a random variable and it is meaningful to consider the *expected error* of an algorithm $\mathbb{E}\,\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2$, where $\mathbb{E}$ is the expectation over the space of problems.

For simplicity, we assume that $\boldsymbol{x}_0$ and $\boldsymbol{x}^\star$ are sampled randomly and independently from $\boldsymbol{H}$ and satisfy

$$\mathbb{E}[(\boldsymbol{x}_0 - \boldsymbol{x}^\star)(\boldsymbol{x}_0 - \boldsymbol{x}^\star)^\top] = R^2 \boldsymbol{I} \tag{4}$$

This makes our analysis simpler, although it is possible to extend the results in this paper by considering a more general setting where the covariance is not the scaled identity.

Crucial to the analysis is the concept of *expected spectral density* $\mathbb{E}\,\hat{\mu}_{\boldsymbol{H}}$. First, we define the expectation of a random measure as follows.

Definition 2 (Tao (2012))**.** *Given a random measure $\hat{\omega}_\xi$, its expected measure $\mathbb{E}_\xi \hat{\omega}_\xi$ is the measure that satisfies*

$$\int_{\mathbb{R}} \varphi \, \mathrm{d}[\mathbb{E}_\xi \hat{\omega}_\xi] \;=\; \mathbb{E}_\xi \int_{\mathbb{R}} \varphi \, \mathrm{d}\hat{\omega}_\xi \,, \tag{5}$$

for any continuous φ with compact support on $\mathbb{R}$.

For notational convenience, we will denote by μ the *expected spectral density* function of $\boldsymbol{H}$, i.e.,

$$\mu(\lambda) \stackrel{\text{def}}{=} \mathbb{E}_{\boldsymbol{H}}[\hat{\mu}_{\boldsymbol{H}}(\lambda)]. \tag{6}$$

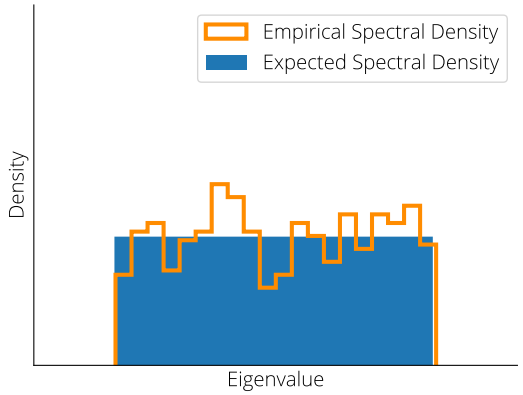

Figure 1. Empirical spectral density and the *expected* spectral density. Even if the eigenvalues of $\boldsymbol{H}$ does not fit exactly the spectral density, on average, its spectrum behave like the expected spectral density.

We now give an intuition about the expected spectral density function. Assume we have a probability distribution over the matrices $\boldsymbol{H}$ in problem (OPT), Then, the *expected spectral density* is the probability density function of a *random* eigenvalue of a *random* matrix $\boldsymbol{H}$. We illustrate this intuition in Figure 1.

We can now use the residual polynomial to derive a simple identity for the error at iteration t. Taking the expectation over $\boldsymbol{x}_0$, $\boldsymbol{x}^\star$ of the norm of both sides of (3), we obtain the following identity for the expected error:

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 &= \operatorname{tr}\left((\boldsymbol{x}_0 - \boldsymbol{x}^\star)^\top P_t^2(\boldsymbol{H})(\boldsymbol{x}_0 - \boldsymbol{x}^\star)\right)\\ &= \operatorname{tr}\left(P_t^2(\boldsymbol{H})(\boldsymbol{x}_0 - \boldsymbol{x}^\star)(\boldsymbol{x}_0 - \boldsymbol{x}^\star)^\top\right)\\ &= R^2\mathbb{E}\left[\int_{\mathbb{R}} P_t^2 \,\mathrm{d}\hat{\mu}_{\boldsymbol{H}}\,,\right]\\ &= R^2\int_{\mathbb{R}} \underbrace{P_t^2}_{\text{algorithm}} \overbrace{\mathrm{d}\mu}^{\text{problem}}\,. \end{aligned} \tag{7}$$

This identity will be instrumental for the average case analysis, as it relates in a simple way the two different aspects that determine the convergence of an optimization algorithm: the algorithm, which appears through the residual polynomial and the problem difficulty, which enters through the expected spectral density.

Remark 2. *The expected error $\mathbb{E}\,\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2$ is over the inputs ($\boldsymbol{H}$, $\boldsymbol{x}_0$, $\boldsymbol{x}^\star$) and not over any randomness of the algorithm, as is common in the stochastic literature. In this paper we will only consider deterministic algorithms.*

Remark 3. *It is possible to derive similar results for the objective suboptimality $f(\boldsymbol{x}) - f^\star$ or the gradient norm $\|\nabla f(\boldsymbol{x})\|^2$, but for simplicity in this paper we focus only to be optimal to $\|\boldsymbol{x} - \boldsymbol{x}^\star\|$.*

Remark 4. *Instead of the independence assumption on $\boldsymbol{x}_0$ and $\boldsymbol{x}^\star$, we can take the Cauchy-Schwartz bound combined with with $\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\| \leq R$. The results of this paper still holds, but with inequalities on the rates of convergence (which corresponds to a mix between an average case over the input $\boldsymbol{H}$ and a worse case on the initialization $\boldsymbol{x}_0 - \boldsymbol{x}^\star$).*

2.3. Optimal average-case algorithms

In light of the previous identity for the expected error, the method that is optimal with respect to this expected error is given by the residual polynomial that solves

$$P_t^\star \in \operatorname*{arg\,min}_{P:\deg(P)\leq t,\, P(0)=1} \int_{\mathbb{R}} P^2 \mathrm{d}\mu. \tag{8}$$

It is known that the solution to this problem can be expressed trough the use of *residual orthogonal polynomials* w.r.t. the weight function $\lambda\mu(\lambda)$, see (Fischer, 1996, Corollary 2.4.7) for instance. We describe this framework in this section.

We first define the notions or orthogonal and residual polynomials.

Definition 3. *A sequence of polynomials $\{P_i\}$ is orthogonal with respect to a weight function ω if*

$$\int_{\mathbb{R}} P_i P_j \,\mathrm{d}\omega \begin{cases} = 0 & \text{if } i \neq j \\ > 0 & \text{if } i = j \end{cases}. \tag{9}$$

*Moreover, if $P_t(0) = 1$ for all t, we call these **residual orthogonal polynomials**.*

It is known result (see for instance Gautschi (1996, Sec. 2.1)) that orthogonal polynomials follow a three-term recursion of the form

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

Due to the number of degrees of freedom, the polynomial P_t can be defined up to a constant. However, in the case where we consider *residual polynomial* (i.e., $P_t(0) = 1$), this constraint can be incorporated in the coefficients, as shown in the following theorem.

Theorem 2.1. *Any sequence of residual orthogonal polynomials $P_1, P_2, \ldots$ can be written under the form*

$$P_t(\lambda) = (a_t + b_t\lambda)P_{t-1}(\lambda) + (1 - a_t)P_{t-2}(\lambda) \tag{10}$$

where $P_0(\lambda) = 1$, $a_{-1} = a_0 = 1$ and $b_{-1} = 0$.

We now write explicitly the optimal first order method associated to the optimal polynomial from Theorem 1. As for orthogonal polynomials, the optimal method also uses a three-terms recurrence as shown in the following theorem.

Lemma 1. *(Fischer, 1996, Corollary 2.4.7) Let the sequence of residual polynomials $\{P_i^\star\}$ be orthogonal w.r.t. $\lambda\mu(\lambda)$. Then, these polynomials solve for all t*

$$P_t^\star \in \operatorname*{arg\,min}_{P:P(0)=1,\,\deg(P)\leq t} \int P^2 \,\mathrm{d}\mu. \tag{11}$$

From this result, it is straightforward to design the optimal algorithm associated to this polynomial, as shown in the next theorem.

Theorem 2.2. *(Pedregosa & Scieur, 2019) A method to solve (OPT) with optimal average-case complexity is given by the following algorithm:*

$$\begin{aligned} \boldsymbol{x}_1 &= \boldsymbol{x}_0 + b_1 \nabla f(\boldsymbol{x}_0), \\ \boldsymbol{x}_t &= \boldsymbol{x}_{t-1} + (1-a_t)(\boldsymbol{x}_{t-2} - \boldsymbol{x}_{t-1}) + b_t \nabla f(\boldsymbol{x}_{t-1}) \end{aligned} \tag{12}$$

where a_t and b_t are the coefficients that build the sequence of residual polynomials P_t, orthogonal w.r.t $\lambda\mu(\lambda)$. Its rate of convergence is given by

$$\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 = R^2 \int_{\mathbb{R}} P_t \,\mathrm{d}\mu\,. \tag{13}$$

The square in the integral (11) can be simplified in (13) thank to the orthogonality property. This simplifies sometimes the computation of the analytic expression of the integral.

2.4. Optimal worst-case and Polyak momentum

The Polyak momentum algorithm (Polyak, 1964) takes as argument the smallest and largest eigenvalue of the matrix $\boldsymbol{H}$ (denoted ℓ and L respectively) in (OPT) and iterates as follow,

$$\begin{aligned} &\boldsymbol{x}_t = \boldsymbol{x}_{t-1} + m(\boldsymbol{x}_{t-2} - \boldsymbol{x}_{t-1}) + h\nabla f(\boldsymbol{x}_{t-1}) \\ &\text{where } m = -\left(\tfrac{\sqrt{L}-\sqrt{\ell}}{\sqrt{L}+\sqrt{\ell}}\right)^2 \text{ and } h = \tfrac{-4}{(\sqrt{\ell}+\sqrt{L})^2} \end{aligned} \tag{PM}$$

Originally, this methods was designed to be *optimal* with *fixed coefficients* in the worst case, over the whole class of first order methods. Informally, it considers the family of algorithms

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + m(\boldsymbol{x}_{t-2} - \boldsymbol{x}_{t-1}) + h\nabla f(\boldsymbol{x}_{t-1}) \tag{14}$$

and solve over $m,\ h$ the following problem,

$$\min_{m,\,h}\ \max_{t>0}\ \max_{\ell\mathbf{I}\preceq\boldsymbol{H}\preceq L\mathbf{I}}\ \max_{\|\boldsymbol{x}_0-\boldsymbol{x}^\star\|\leq R} \sqrt[t]{\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_2^2}$$

(see e.g. (Scieur et al., 2017)). Since the Polyak momentum method is optimal under the worst-case analysis, it is agnostic of what happens *inside* the edges of the spectrum, thus it does not exploit the structure (7).

3. Asymptotic Optimality of Polyak Momentum

In this section we present the main contributions in this paper. First, we show that, asymptotically and under mild assumptions, the recurrence of optimal methods converges to (PM) (Theorem 3.1). Moreover, we also show that asymptotically, the optimal rate of convergence converges to the worst-case rate (Theorem 3.2).

Theorem 3.1. *Consider the iterative algorithm*

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

whose rate of convergence is average-case optimal for problem OPT, then asymptotically ***the recurrence converges to Polyak momentum*** *(PM).*

More formally, let the method be optimal for problem OPT, and let μ be the expected spectral density of $\boldsymbol{H}$. If $\mu(\lambda) > 0$ for $0 < \ell < \lambda < L$ and zero elsewhere (i.e., bounded convex support), then as $t \to \infty$ we have

$$(1-a_t) \to \underbrace{-\left(\frac{\sqrt{L}-\sqrt{\ell}}{\sqrt{L}+\sqrt{\ell}}\right)^2}_{=m \text{ in (PM)}},\quad b_t \to \underbrace{\frac{-4}{(\sqrt{\ell}+\sqrt{L})^2}}_{=h \text{ in (PM)}}. \tag{15}$$

The technical assumption on μ can be weaken by using the term "regular n^{th} root asymptotic behavior" from (Stahl, 1991, Thm 2.1). However, in practice it may be hard to prove this condition, while asking μ to be positive on its interval is more intuitive and implies to have a regular asymptotic behavior.

Theorem 3.2. *Under the same assumptions of Theorem 3.1, the asymptotic expected rate of convergence of the optimal method converges to the one worst-case rate of convergence, i.e.,*

$$\lim_{t\to\infty} \sqrt[t]{\mathbb{E}\left[\frac{\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2}{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2}\right]} = \left(\frac{\sqrt{L}-\sqrt{\ell}}{\sqrt{L}+\sqrt{\ell}}\right)^2. \tag{16}$$

4. Numerical Evidences

Before proving the result of Theorem 3.1, we provide a numerical illustration of the convergence of optimal algorithms towards Polyak momentum. We compare the (non-asymptotic) difference between of Polyak momentum parameters and the ones of the optimal polynomial in Figure 2 for several expected spectral density functions. In our experiment, we compare

- Difference of momentum: $|(1-a_t) - m|$
- Difference of stepsize: $|b_t - h|$

- Difference of the average rate:

$$\left(\frac{\sqrt{L}-\sqrt{\ell}}{\sqrt{L}+\sqrt{\ell}}\right)^2 - \left(\frac{\mathbb{E}[\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^2]}{\|\boldsymbol{x}_0-\boldsymbol{x}^\star\|^2}\right)^{1/t}$$

where m and h are defined in (PM), a_t and b_t are the coefficients of the optimal first order method, P_t the t-th optimal polynomial. We use the following family of weight function as a basis to build the the expected spectral density,

$$\nu(\xi) = \frac{(1-\xi)^\alpha(1+\xi)^\beta}{C(\alpha,\beta)}, \quad \alpha,\beta > -1,$$

for ξ in the interval $[-1,1]$, where C is a constant such that the integral of ν in the interval $[-1,\,1]$ is equal to one. These weight functions are known to generate the family of Jacobi polynomials $J_t(\alpha,\beta)$. In particular, we use a shifted version of ν, such that it is defined in the interval $[\ell,\,L]$,

$$\mu(\lambda) = \nu(m(\lambda)), \quad m(\lambda) = \frac{L+\ell}{\ell-L} + \frac{2\lambda}{L-\ell}.$$

In our experiment, we fix $\ell = 0.1$ and $L = 10$ in Figure 2 and $\ell = 0.01$ and $L = 100$ in Figure 3. The optimal residual polynomial w.r.t. the expected spectral density function μ is build using a modification of the recurrence of Jacobi polynomials, explained in (Gautschi, 1996, Thm. 7). We tried all combination between $\alpha = \{-0.5, 0, 1\}$ and $\beta = \{-0.5, 0, 1\}$. In all cases, and as expected by the theory, we see that the gap between the Polyak momentum method and the optimal one closes as t increases.

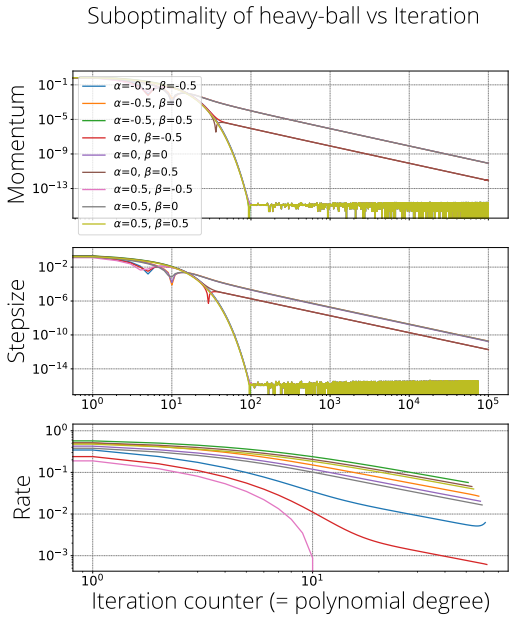

Figure 2. Difference between the momentum (top line), the stepsize (second line) and the rate (third line) of Polyak method compared to the optimal algorithm, for several density functions. The computation of the error on the rate was stop prematurely, when the integral was below machine precision (the curve report the integral power $1/\text{iteration}$). The distinction between curves is not important: what matter is the convergence of all quantities to zero.

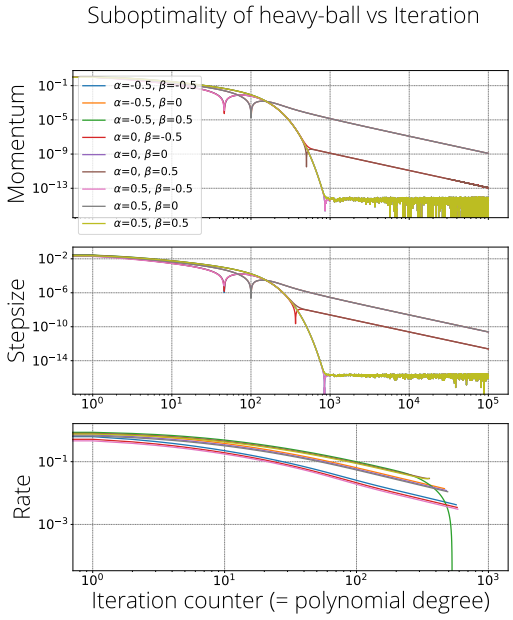

Figure 3. Same experiment as in Figure 2 with $\ell = 0.01$ and $L = 100$. The conditioning of the problem affect slightly the convergence of the suboptimality of Polyak momentum.

5. Proof sketch

Before presenting the proof in full detail, we give a broad overview on the techniques used in it.

As optimal method can be viewed as orthogonal polynomials, the crucial ingredient of the proof are asymptotic results on orthogonal polynomials. Unfortunately, most existing results are derived for for *orthonormal* polynomials ($\equiv$ the integral in Definition 3 is 1 for $i = j$) defined in the interval $[-1,1]$, while we will be interested in *residual* orthogonal polynomials defined in an arbitrary interval $[\ell, L]$. This means The first part of the proof (§7.1.1) relies on adapting existing results through a change of variable $m(\lambda)$ that translates the interval $[\ell,\,L]$ to $[-1,1]$.

Then, using the change of variable, we deduce a recurrence for orthogonal polynomials in the interval $[\ell,\,L]$ (§7.1.2). It remains to ensure that $P_t(0) = 1$, as required by Theorem 3.1. Finally, in §7.1.3, we use the properties of the asymptotic recurrence for orthonormal polynomials, combined with the asymptotic value of orthonormal polynomial when evaluated *outside* the interval of orthogonality, to obtain Theorem 3.1.

The rate of convergence in Theorem 3.2 is ensured by evaluating the integral (13). We show in §7.2 this integral can be evaluated using the recurrence

$$\int_{\mathbb{R}} P_t = a_t \int_{\mathbb{R}} P_{t-1}\,\mathrm{d}\mu + (1-a_t)\int_{\mathbb{R}} P_{t-2}\,\mathrm{d}\mu$$

Using the again some properties on the asymptotic behavior of a_t, the recurrence converges to a stationary one, whose solution gives the worst-case rate of Theorem 3.2.

6. Asymptotic results for orthonormal polynomials

In this section, we review some results on asymptotic of orthonormal polynomials. The proof of Theorem 3.1 and Proposition 3.2 heavily depends on two important results (Theorems 6.1 and 6.2), describing the asymptotic behavior of orthogonal polynomials inside and outside the interval of orthogonality.

Notations. We define new notations that will be used in the rest of the paper. We will jump between polynomials and variables defined in the interval $[-1,\,1]$ or in the interval $[\ell,\,L]$. We thus make the distinction between $Q(\xi)$, where $\xi \in [-1,\,1]$ and $P(\lambda)$, where $\lambda \in [\ell,\,L]$

First, we define orthonormal polynomials.

Definition 4. *We say that the sequence of polynomials $Q_t(\xi)$ is orthonormal w.r.t. the weight function ν in the interval $[-1,\,1]$ if*

$$\int_{-1}^{1} \mathrm{d}\nu = 1 \quad \textit{and} \quad \int_{-1}^{1} Q_i Q_j\,\mathrm{d}\nu = \Delta_{ij}, \tag{17}$$

where Δ_{ij} is the Kronecker delta function, defined as 1 if $i=j$ and 0 otherwise.

We have also the following assumptions on the distribution.

Assumption 1. *We assume that ν satisfies*

$$\nu(\xi)\begin{cases} >0 & \textit{if } \xi \in [-1,1] \\ =0 & \textit{if } \xi \notin [-1,1] \end{cases} \quad \textit{and} \quad \int_{-1}^{1} \mathrm{d}\nu = 1.$$

More specifically, we need that ν is strictly positive on a continuous, compact interval. This regularity assumption is important in this section, otherwise the next asymptotic results may not hold. Nevertheless, this remains a sufficient condition, and it is possible to find distributions that do not satisfy this condition while enjoying the same properties (Ullman, 1984).

The sequence of optimal polynomials P_t, solving (11) (associated to the optimal method for solving problem (OPT)), is actually orthogonal to the measure $\lambda\mu(\lambda)$, where μ is the expected spectral density of $\boldsymbol{H}$. To prove Theorem 3.1, we link the polynomials P_t and Q_t, where the sequence $\{Q_i\}$ is *orthonormal* w.r.t. the distribution ν. A sequence of orthonormal polynomials follows also a tree-terms recurrence, but with the following special structure,

$$\lambda Q_{t-1}(\xi) = A_t Q_t(\xi) + B_{t-1}Q_{t-1}(\xi) + A_{t-1}Q_{t-2}(\xi). \tag{18}$$

For instance, see (Lubinsky, 2000). The study of the asymptotic behavior of A_t and B_t will lead to the result of Theorem 3.1.

There are surprisingly strong results on asymptotic properties of orthonormal polynomials, and we will list here the most important ones for our purpose. The first one describes the limits of the coefficients of A_t and B_t.

Theorem 6.1. *Let A_t and B_t be the coefficients that defines the sequence of orthonormal polynomials (17). Then,*

$$A_t \to \tfrac{1}{2} \quad \textit{and} \quad B_t \to 0 \quad \textit{when } t \to \infty. \tag{19}$$

The second theorem also plays an important role in our result, as it describes the value of the polynomial *outside* the interval $[-1,\,1]$. The proof of this result can be found for example in (Rakhmanov, 1983; Máté et al., 1985)

Theorem 6.2. *(Rakhmanov, 1983). Let $\{Q_t\}$ be the sequence of orthonormal polynomials (17). Then,*

$$\lim_{t\to\infty} \frac{Q_t(\xi)}{Q_{t-1}(\xi)} = \xi + \sqrt{\xi^2-1} \quad \textit{for } \xi \notin [-1,1].$$

Corollary 1. *Under the same assumptions of Theorem 6.2,*

$$\lim_{t\to\infty} \sqrt[t]{Q_t(\xi)} = \xi + \sqrt{\xi^2-1} \quad \textit{for } \xi \notin [-1,1].$$

7. Asymptotic results on optimal methods

In this section, we present the proofs for Theorem 3.1 and Proposition 3.2.

7.1. Proof of Theorem 2.1

We now present the main contribution of this paper. In short, we merge the results from the two previous sections. In particular, we show that if we have an optimal method that solve (OPT), where the expected spectral density function of $\boldsymbol{H}$ is μ, then under mild assumption on μ, the methods converges to Polyak momentum.

7.1.1. From orthonormal to residual polynomials on different domains

First, we have make the link between the residual polynomials P_t, orthogonal to $\lambda\mu(\lambda)$, and orthogonal polynomials Q_t. We first need to define the two linear mappings

m and m^{-1}, which map the interval $[\ell,\, L]$ to $[-1,\, 1]$ and inversely,

$$m(\lambda) = \frac{L+\ell}{\ell - L} + \frac{2\lambda}{L-\ell}, \quad m^{-1}(\xi) = \frac{L+\ell}{2} + \xi\frac{L+\ell}{2}. \tag{20}$$

Now, we define the normalized density functions

$$\nu(\xi) = \frac{m^{-1}(\xi)\mu(m^{-1}(\xi))}{\int_\ell^L m^{-1}(\xi)\mu(m^{-1}(\xi))\,\mathrm{d}\xi}, \tag{21}$$

which corresponds to the normalization of the density $\lambda\mu(\lambda)$ after the change of variable $\lambda = m^{-1}(\xi)$. Finally, we define the sequence of polynomials P_t to be orthogonal w.r.t. $\lambda\mu(\lambda)$, i.e.,

$$P_t(0) = 0 \quad \text{and} \quad \int_\ell^L P_i P_j \lambda\mu(\mathrm{d}\lambda) \begin{cases} > 0 \text{ if } i = j \\ = 0 \text{ otherwise} \end{cases},$$

and the sequence of polynomials Q to be orthonormal w.r.t. ν,

$$\int_\ell^L Q_i Q_j \nu(\mathrm{d}\lambda) = \Delta_{i,\,j},$$

where $\Delta_{i,j} = 1$ if $i = j$ and 0 otherwise. The main relation between P and Q is their orthogonality after a change of variable. Indeed, take the definition of $Q_t(\xi)$ and make let $\xi = m(\lambda)$. Then,

$$\int_\ell^L Q_i(m(\lambda))Q_j(m(\lambda))\, \underbrace{\nu((m(\mathrm{d}\lambda)))}_{=\text{Constant}\cdot\mu(\mathrm{d}\lambda)} \begin{cases} > 0 \text{ if } i = j \\ = 0 \text{ otherwise} \end{cases}$$

This means that $P_t(\lambda)$ and $Q_t(m(\lambda))$ are the same polynomial, up to a constant. Since $P_t(0) = 1$, we thus can write the relation

$$P_t(\lambda) = \frac{Q_t(m(\lambda))}{Q_t(m(0))}. \tag{22}$$

The idea here is to characterize the asymptotic behavior of P_t using the asymptotic properties of Q_t.

7.1.2. Three-terms recurrence for the residual polynomial

Because Q_t is orthonormal, it follows the recurrence (18),

$$\begin{aligned} Q_t(m(\lambda)) = &\frac{m(\lambda)}{A_t}Q_{t-1}(m(\lambda)) \\ &- \frac{B_{t-1}}{A_t}Q_{t-1}(m(\lambda)) - \frac{A_{t-1}}{A_t}Q_{t-2}(m(\lambda)). \end{aligned}$$

This only gives a recurrence for $Q_t(m(\lambda))$, but not for P_t. To do so, we use the relation (22) between these two polynomials and obtain

$$\begin{aligned} P_t(\lambda)Q_t(m(0)) = &\frac{m(\lambda)}{A_t}Q_{t-1}(m(0))P_{t-1}(\lambda) \\ &- \frac{B_{t-1}}{A_t}Q_{t-1}(m(0))P_{t-1}(\lambda) \\ &- \frac{A_{t-1}}{A_t}Q_{t-2}(m(0))P_{t-2}(\lambda). \end{aligned}$$

7.1.3. Asymptotic recurrence

We now use Theorem 6.1, describing the asymptotic value of A_t and B_t. Thus, for $t \to \infty$,

$$\begin{aligned} P_t(\lambda) = &2m(\lambda)\frac{Q_{t-1}(m(0))}{Q_t(m(0))}P_{t-1}(m(\lambda)) \\ &- \frac{Q_{t-2}(m(0))}{Q_t(m(0))}P_{t-2}(m(\lambda)). \end{aligned}$$

Now, we need to estimate the asymptotic value of the ratio $Q_{t-1}(m(0))/Q_t(m(0))$. Since $m(0)$ is *outside* the interval $[-1,\, 1]$ (because $0 < \ell < L$), we can apply Theorem 6.2. We thus obtain, for $t \to \infty$,

$$\frac{Q_{t-1}(m(0))}{Q_t(m(0))} = \frac{1}{m(0) + \sqrt{m(0)^2 - 1}} = -\frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}}. \tag{23}$$

Similarly,

$$\begin{aligned} \frac{Q_{t-2}(m(0))}{Q_t(m(0))} &= \frac{Q_{t-2}(m(0))}{Q_{t-1}(m(0))}\frac{Q_{t-1}(m(0))}{Q_t(m(0))}, \\ &= \left(\frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}}\right)^2. \end{aligned}$$

It remains to evaluate $2m(\lambda)\frac{Q_{t-1}(m(0))}{Q_t(m(0))}$, which gives

$$\begin{aligned} &2m(\lambda)\frac{Q_{t-1}(m(0))}{Q_t(m(0))} = \\ &1 + \left(\frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}}\right)^2 - \lambda\frac{4}{(\sqrt{\ell} + \sqrt{L})^2} \end{aligned}$$

All together, the recursion becomes

$$\begin{aligned} P_t(\lambda) = P_{t-1}(\lambda) &- \frac{4}{(\sqrt{\ell} + \sqrt{L})^2}\lambda P_{t-1}(\lambda) \\ &+ \left(\frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}}\right)^2 (P_{t-1} - P_{t-2}). \end{aligned}$$

which corresponds *exactly* to the recursion of Polyak momentum methods from Polyak (1964).

7.2. Proof of Theorem 3.2

We now show that, asymptotically, the rate of convergence of the optimal method converges (asymptotically) to the worst-case rate of Polyak momentum. First, we use the result of Theorem 2.2, equation 13, which link the expected rate of convergence with the integral of the polynomial P_t,

$$\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 = R^2 \int_{\mathbb{R}} P_t \,\mathrm{d}\mu \,.$$

To simplify the evaluation of this integral, we use the recurrence of P_t,

$$\int_{\mathbb{R}} P_t \,\mathrm{d}\mu = \int_{\mathbb{R}} (a_t + \lambda b_t) P_{t-1} + (1 - a_t) P_{t-2} \,\mathrm{d}\mu$$

However, because the sequence $\{P_i\}$ is orthogonal w.r.t. $\lambda\mu(\lambda)$, we have that

$$\int_{\mathbb{R}} P_t \lambda\mu(\lambda) \,\mathrm{d}\lambda = \int_{\mathbb{R}} (P_t \cdot 1)\lambda\mu(\lambda) \,\mathrm{d}\lambda = 0 \quad \text{if } t > 0.$$

Let $r_t = \int_{\mathbb{R}} P_t \,\mathrm{d}\mu$. This means we have

$$r_t = a_t r_{t-1} + (1 - a_t) r_{t-2}, \quad r_0 = \int_{\mathbb{R}} \mathrm{d}\mu, \quad r_1 = \int_{\mathbb{R}} P_1 \,\mathrm{d}\mu.$$

Now, we analyse the asymptotic behavior of the recurrence, already studied in the previous section. From Theorem 3.1, we have that the sequence a_t converges asymptotically to

$$\lim_{t\to\infty} a_t = 1 + \left(\frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}}\right)^2 = a_\infty.$$

We can thus study the solution of the asymptotic recurrence

$$r_t = a_\infty r_{t-1} + (1 - a_\infty) r_{t-2}.$$

The general solution of this recurrence equation reads

$$r_t = c_1 (1 - a_\infty)^t + c_2$$

where c_1 and c_2 are constants. However, we know that $r_\infty = 0$ because the method is convergent, thus $c_2 = 0$. The asymptotic solution reads

$$r_t = c_1 \left(\frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}}\right)^{2t},$$

More precisely, we recover *exactly* to the one of Polyak momentum after taking the t^{th} root,

$$\begin{aligned}\lim_{t\to\infty} \sqrt[t]{c_1 \left(\frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}}\right)^{2t}} &= \lim_{t\to\infty} \sqrt[t]{c_1} \left(\frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}}\right)^2, \\ &= \left(\frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}}\right)^2.\end{aligned}$$

8. Conclusion and future work

In this work, we linked the Polyak momentum method to the asymptotic behavior of optimal methods in average-case for minimizing quadratics, under mild assumptions on the expected spectral distribution. Surprisingly, those methods are asymptotically equivalent, with average-case complexity equal to the known worst-case bound. Our numerical experiments suggest that the conditioning of the problem (measured by L/ℓ) plays a role in how quickly optimal methods behave like Polyak momentum, as shown in Figure 2 and Figure 3.

A first research direction may be the analysis of the rate of convergence of optimal method to Polyak momentum algorithm. It seems the convergence of the stepsize and momentum parameters are bounded polynomially in function of the number of iterations.

A second research direction is the study of optimal polynomials on the *complex* plane. In this case, we are no longer solving the optimization problem (OPT). Instead, we aim to solve the linear system $Ax = b$, where the matrix A is non-symmetric, with potentially complex eigenvalues. This has implication in the study of optimal algorithm in game theory (Azizian et al., 2020) or in the acceleration of primal-dual algorithms (Bollapragada et al., 2018).

References

Azizian, W., Scieur, D., Mitliagkas, I., Lacoste-Julien, S., and Gidel, G. Accelerating smooth games by manipulating spectral shapes. *arXiv preprint arXiv:2001.00602*, 2020.

Bollapragada, R., Scieur, D., and d'Aspremont, A. Nonlinear acceleration of momentum and primal-dual algorithms. *arXiv preprint arXiv:1810.04539*, 2018.

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

Gautschi, W. Orthogonal polynomials: applications and computation. *Acta numerica*, 1996.

Goh, G. Why momentum really works. *Distill*, 2017.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

Hestenes, M. R., Stiefel, E., et al. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 1952.

Lessard, L., Recht, B., and Packard, A. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

Lubinsky, D. Asymptotics of orthogonal polynomials: some old, some new, some identities. *Acta Applicandae Mathematica*, 2000.

Máté, A., Nevai, P., and Totik, V. Asymptotics for the ratio of leading coefficients of orthonormal polynomials on the unit circle. *Constructive Approximation*, 1985.

Menon, G. and Trogdon, T. Smoothed analysis for the conjugate gradient algorithm. *arXiv preprint arXiv:1605.06438*, 2016.

Nemirovski, A. Information-based complexity of convex programming. *Lecture Notes*, 1995.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013.

Pedregosa, F. and Scieur, D. Acceleration through spectral modeling. *Beyond First Order Methods in Machine Learning Workshop*, 2019.

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

Polyak, B. T. Introduction to optimization. *Optimization Software, Inc. Publications Division, New York*, 1987.

Pumir, T., Jelassi, S., and Boumal, N. Smoothed analysis of the low-rank approach for smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pp. 2281–2290, 2018.

Rakhmanov, E. A. On the asymptotics of the ratio of orthogonal polynomials. II. *Mathematics of the USSR-Sbornik*, 1983.

Scieur, D., Roulet, V., Bach, F., and d'Aspremont, A. Integration methods and optimization algorithms. In *Advances in Neural Information Processing Systems*, 2017.

Spielman, D. A. The smoothed analysis of algorithms. In *International Symposium on Fundamentals of Computation Theory*. Springer, 2005.

Stahl, H. On regular nth root asymptotic behavior of orthonormal polynomials. *Journal of Approximation Theory*, 1991.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, 2013.

Tao, T. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

Ullman, J. L. A survey of exterior asymptotics for orthogonal polynomials associated with a finite interval and a study of the case of the general weight measures. In *Approximation theory and spline functions*. Springer, 1984.

A. Proof of Theorem 2.2

For completeness, we write the proof of Theorem 2.2.

Theorem 2.2. *(Pedregosa & Scieur, 2019) A method to solve* (OPT) *with optimal average-case complexity is given by the following algorithm:*

$$\begin{aligned} \boldsymbol{x}_1 &= \boldsymbol{x}_0 + b_1 \nabla f(\boldsymbol{x}_0), \\ \boldsymbol{x}_t &= \boldsymbol{x}_{t-1} + (1-a_t)(\boldsymbol{x}_{t-2} - \boldsymbol{x}_{t-1}) + b_t \nabla f(\boldsymbol{x}_{t-1}) \end{aligned} \tag{12}$$

where a_t and b_t are the coefficients that build the sequence of residual polynomials P_t, orthogonal w.r.t $\lambda\mu(\lambda)$. Its rate of convergence is given by

$$\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 = R^2 \int_{\mathbb{R}} P_t \,\mathrm{d}\mu\,. \tag{13}$$

Proof. We start with the definition of the algorithm, then we show that it actually builds the optimal polynomial (11). Indeed, by removing $\boldsymbol{x}^\star$ on both sides,

$$\boldsymbol{x}_t - \boldsymbol{x}^\star = a_t(\boldsymbol{x}_{t-1} - \boldsymbol{x}^\star) - (1-a_t)(\boldsymbol{x}_{t-2} - \boldsymbol{x}^*) - b_t \nabla f(\boldsymbol{x}_{t-1})$$

By definition of our quadratic function, its gradient reads

$$\nabla f(\boldsymbol{x}_{t-1}) = \boldsymbol{H}(\boldsymbol{x}_{t-1} - \boldsymbol{x}^\star).$$

All together, the recursion becomes

$$\boldsymbol{x}_t - \boldsymbol{x}^\star = a_t(\boldsymbol{x}_{t-1} - \boldsymbol{x}^\star) - (1-a_t)(\boldsymbol{x}_{t-2} - \boldsymbol{x}^\star) - b_t \boldsymbol{H}(\boldsymbol{x}_{t-1} - \boldsymbol{x}^\star)$$

However, we have that $\boldsymbol{x}_0 - \boldsymbol{x}^\star$ and $\boldsymbol{x}_1 - \boldsymbol{x}^\star$ are polynomials, since

$$\begin{aligned} \boldsymbol{x}_0 - \boldsymbol{x}^\star &= \boldsymbol{H}^0(\boldsymbol{x}_0 - \boldsymbol{x}^\star) = p_0^\star(\boldsymbol{H})(x_0 - x^\star) \\ \boldsymbol{x}_1 - \boldsymbol{x}^\star &= x_0 - x^\star - b_1 \boldsymbol{H}^0(\boldsymbol{x}_0 - \boldsymbol{x}^\star) = p_1^\star(\boldsymbol{H})(\boldsymbol{x}_0 - \boldsymbol{x}^\star). \end{aligned}$$

Using this argument recursively, we have that

$$\boldsymbol{x}_t - \boldsymbol{x}^\star = \left(a_t P_{t-1}^\star(\boldsymbol{H}) - (1-a_t)P_{t-2}^\star(\boldsymbol{H}) - b_t P_{t-1}^\star(\boldsymbol{H})\right)(\boldsymbol{x}_0 - \boldsymbol{x}^\star) = P_t^\star(\boldsymbol{H})(x_0 - x^*).$$

Since, by construction, $p^\star$ is the solution of (11), the algorithm is optimal. We now show that the square in the integral can be simplified. Without any loss of generality we consider $\beta = 0$. Indeed,

$$\int_{\mathbb{R}^+} P_t^2(\lambda)\mu(\lambda)\,\mathrm{d}\lambda = \int_{\mathbb{R}^+} P_t(\lambda)\left(\lambda \frac{P_t(\lambda) - P_t(0)}{\lambda} + P_t(0)\right)\mu(\lambda)\,\mathrm{d}\lambda.$$

Let $Q(\lambda) = \frac{P_t(\lambda) - P_t(0)}{\lambda}$. This is a polynomial of degree $t-1$ since we took P_t, then we removed the independent term, then divided by λ. Since the sequence $\{P_i\}$ forms a basis of polynomial, we have that P_t is orthogonal to all polynomials of degree less than t w.r.t. $\lambda\mu(\lambda)$. In this case, P_t is orthogonal to Q, thus

$$\int_{\mathbb{R}^+} P_t^2(\lambda)\mu(\lambda)\,\mathrm{d}\lambda = \int_{\mathbb{R}^+} P_t(\lambda)Q(\lambda)\lambda\mu(\lambda)\,\mathrm{d}\lambda + \int_{\mathbb{R}^+} P_t(\lambda)P_t(0)\mu(\lambda)\,\mathrm{d}\lambda = \int_{\mathbb{R}^+} P_t(\lambda)\mu(\lambda)\,\mathrm{d}\lambda,$$

where the last equality follows from the fact that P is a normalized polynomial, thus $P(0) = 1$. □