A. Proof of main results

For a polynomial P and a measure (resp. density) μ, we will denote $\mu[P] := \int_{\mathbb{R}} P(x)\mu(x)\mathrm{d}\mu(x)$ (resp. $\mu[P] := \int_{\mathbb{R}} P(x)\mu(x)\mathrm{d}x$). For a density μ, we stress the fact that $\mu[x]$ and $\mu(x)$ refer to different quantities.

A.1. Proof of Theorem 1

We recall that $\Pi_0(x)=1$, $\Pi_1(x)=1+b_1x$ and for $t \geqslant 2$,

$$\Pi_t(x) = (a_t + b_t x)\,\Pi_{t-1}(x) + (1-a_t)\Pi_{t-2}(x)\,. \tag{36}$$

First, we claim that for any $t \geqslant 0$, $\Delta_t = \Pi_t\big(C_S^{-1}\big)\,\Delta_0$, and we show it by induction. Since $\Pi_0(x) = 1$, we have that $\Delta_0 = \Pi_0\big(C_S^{-1}\big)\cdot\Delta_0$. Since $x_1 = x_0 + b_1 H_S^{-1}\nabla f(x_0)$, subtracting x^* and multiplying by $U^\top A$ the latter equation, we obtain that $\Delta_1 = \Pi_1\big(C_S^{-1}\big)\cdot\Delta_0$. Suppose that for some $t \geqslant 2$, the induction claim holds for $t-1$ and $t-2$. Subtracting x^* and multiplying by $U^\top A$ the update formula (11), we obtain that

$$\begin{aligned}
\Delta_t &= \Delta_{t-1} + (1-a_t)(\Delta_{t-2} - \Delta_{t-1}) + b_t C_S^{-1}\,\Delta_{t-1} \\
&= (a_t + b_t C_S^{-1})\Delta_{t-1} + (1-a_t)\Delta_{t-2} \\
&= \big((a_t + b_t C_S^{-1})\Pi_{t-1}\big(C_S^{-1}\big) + (1-a_t)\Pi_{t-2}\big(C_S^{-1}\big)\big)\,\Delta_0 \\
&= \Pi_t\big(C_S^{-1}\big)\,\Delta_0\,,
\end{aligned}$$

where we used the induction hypothesis for $t-1$ and $t-2$ in the third equality, and the recursion formula (36) in the last equality. Consequently, using Lemma 2.1, we obtain that

$$\lim_{n\to\infty} \frac{\mathbb{E}\|\Delta_t\|^2}{\mathbb{E}\|\Delta_0\|^2} = \int_a^b \Pi_t^2\big(\lambda^{-1}\big)\,\mathrm{d}\mu(\lambda) = \mathcal{L}^*_{\mu,t}\,.$$

A.2. Proof of Theorem 2

We have already argued that $\{x_t\}$ is asymptotically optimal. It remains to prove that $\mathcal{L}^*_{\mu_\rho,t} = \rho^t$.

Set $\lambda_\rho(x) = x^{-1}\mu_\rho(x)$. Let $\{\Pi_t\}$ be an orthogonal basis with respect to μ_ρ such that $\Pi_t(0) = 1$ and $\deg(\Pi_t) = t$. From Lemma 2.3, we have $\mathcal{L}^*_{\mu_\rho,t} = (1-\rho)\lambda_\rho[\Pi_t^2]$, so that it suffices to show that $\lambda_\rho[\Pi_t^2] = (1-\rho)^{-1}\rho^t$. On the other hand, in the proof of Lemma 2.4 in Appendix B.4, we establish that there exists a sequence of polynomials $\{T_k\}_{k\geqslant 1}$ such that for any $t \geqslant 1$ and $k, \ell \geqslant 1$,

$$\begin{aligned}
\Pi_t(x) &= 1 - \sum_{j=1}^t \lambda_\rho[T_t]\,T_t(x)\,, \\
\lambda_\rho[T_t] &= (-1)^{t-1}\sqrt{\rho}^{\,t-1}\,, \\
\lambda_\rho[T_k T_\ell] &= \delta_{k\ell}\,,
\end{aligned}$$

where $\delta_{k\ell} = 1$ if $k = \ell$, and 0 otherwise. Using the latter properties, it follows that

$$\begin{aligned}
\lambda_\rho[\Pi_t^2] &= \lambda_\rho[1] - 2\sum_{j=1}^t \lambda_\rho[T_j]^2 + \sum_{j=1}^t \lambda_\rho[T_j]^2 \underbrace{\lambda_\rho[T_j^2]}_{=1} \\
&= \lambda_\rho[1] - \sum_{j=1}^t \lambda_\rho[T_j]^2 \\
&= \frac{1}{1-\rho} - \sum_{j=0}^{t-1}\rho^j \\
&= \frac{\rho^t}{1-\rho}\,,
\end{aligned}$$

and, in the third equality, we used the standard inverse moment formula $\lambda_\rho[1] = \int x^{-1}\mu_\rho(x)\mathrm{d}x = (1-\rho)^{-1}$. Consequently, we obtain the claimed formula, that is, $\mathcal{L}^*_{\mu_\rho,t} = \rho^t$.

A.3. Proof of Theorem 3

According to Lemma 2.5, the support of F_h is included within the interval $(0,1)$. Therefore, we fix $x \in (0,1)$ and we consider the complex number $z = x + iy$, where $y > 0$. Our goal is to compute the quantity

$$\lim_{y \to 0^+} \frac{1}{\pi} |\mathrm{Im}(m_h(z))| \,.$$

If the above limit exists, then F_h is differentiable at x and its derivative is equal to this limit (Silverstein & Choi, 1995). Note that the absolute value is not necessary, since $\mathrm{Im}(m_h(z))$ is positive on $\mathbb{C}^+$. But it will avoid to specify explicitly the branch cut of the square-root considered later in this proof, and thus additional technicalities.

From Lemma 2.5, it holds that

$$2\gamma m_h(z) = \frac{2\gamma - 1}{1 - z} + \frac{\xi - \gamma}{z(1-z)} - \frac{R(z)}{z(1-z)} \,. \tag{37}$$

where $R(z) = \sqrt{(\gamma + \xi - 2 + z)^2 + 4(z-1)(1-\gamma)(1-\xi)}$, and the branch cut of the square-root is chosen such that $m_h > 0$ on $\mathbb{C}^+$, $m_h < 0$ on $\mathbb{C}^-$ (the complex numbers with negative imaginary parts), and $m_h > 0$ on $\mathbb{R}_-$ (the negative real numbers). Further, we have

$$\frac{1}{z(1-z)} = \frac{x(1-x) + y^2 + iy(2x-1)}{(x(1-x) + y^2)^2 + y^2(2x-1)^2} \,, \qquad \frac{1}{1-z} = \frac{1 - x + iy}{(1-x)^2 + y^2} \,,$$

from which we deduce that the imaginary parts of the first two terms in the expansion (37) of $2\gamma m_h(z)$ are given by

$$\begin{aligned}
\mathrm{Im}\left(\frac{2\gamma - 1}{1 - z}\right) &= \frac{(2\gamma - 1)y}{(1-x)^2 + y^2} \,, \\
\mathrm{Im}\left(\frac{\xi - \gamma}{z(1-z)}\right) &= \frac{(\xi - \gamma)(2x-1)y}{(x(1-x) + y^2)^2 + y^2(2x-1)^2} \,.
\end{aligned}$$

Since $x \in (0,1)$, the limits $y \to 0^+$ of the two above quantities exist and are equal to 0. Hence, provided it exists, we have

$$\lim_{y \to 0^+} 2\gamma |\mathrm{Im}(m_h(z))| = \lim_{y \to 0^+} \left| \mathrm{Im}\left(\frac{R(z)}{z(1-z)}\right) \right| \,. \tag{38}$$

We introduce the function $f(z) = (z - \alpha - \beta)^2 + 4(z-1)\alpha\beta$ where $\alpha = 1 - \xi$ and $\beta = 1 - \gamma$, so that $R(z) = \sqrt{f(z)}$. We have $f(z) = X + iY$ where

$$\begin{aligned}
X &= (x - \alpha - \beta)^2 - y^2 + 4(x-1)\alpha\beta \,, \\
Y &= 2(x - \alpha - \beta + 2\alpha\beta)y \,.
\end{aligned}$$

Thus, the absolute values of the real and imaginary parts of $R(z)$ are given by

$$\begin{aligned}
|\mathrm{Re}(R(z))| &= \frac{1}{\sqrt{2}} \sqrt{\sqrt{X^2 + Y^2} + X} \,, \\
|\mathrm{Im}(R(z))| &= \frac{1}{\sqrt{2}} \sqrt{\sqrt{X^2 + Y^2} - X} \,,
\end{aligned}$$

and they have respective limits

$$\begin{aligned}
\lim_{y \to 0^+} |\mathrm{Re}(R(z))| &= \sqrt{|\varphi(x)|} \cdot \mathbf{1}(\varphi(x) > 0) \,, \\
\lim_{y \to 0^+} |\mathrm{Im}(R(z))| &= \sqrt{|\varphi(x)|} \cdot \mathbf{1}(\varphi(x) < 0) \,,
\end{aligned}$$

where $\varphi(x) := (x - \alpha - \beta)^2 + 4(x-1)\alpha\beta$. Further, we have

$$\mathrm{Im}\left(\frac{R(z)}{z(1-z)}\right) = \frac{y(2x-1)\mathrm{Re}(R(z)) + (x(1-x) + y^2)\mathrm{Im}(R(z))}{g(x,y)} \,,$$

where $g(x,y) = (x(1-x)+y^2)^2 + y^2(2x-1)^2$. Note that $\lim_{y\to 0^+} g(x,y) = x^2(1-x)^2$, which is non-zero since $x \in (0,1)$. Then, we obtain

$$\lim_{y\to 0^+} \left|\mathrm{Im}\left(\frac{R(z)}{z(z-1)}\right)\right| = \frac{\sqrt{|\varphi(x)|}\mathbf{1}(\varphi(x)<0)}{x(1-x)}\,,$$

Using (38), it follows that for any $x \in (0,1)$, $\lim_{y\to 0^+} \frac{1}{\pi}|\mathrm{Im}(m_h(x))|$ exists. This implies that F_h admits a density over $(0,1)$, given by

$$f_h(x) = \frac{1}{2\gamma\pi}\frac{\sqrt{(\Lambda_h - x)_+(x-\lambda_h)_+}}{x(1-x)}\,,$$

where we used the fact that $\varphi(x) = (x-\Lambda_h)(x-\lambda_h)$, and we recall that the edge eigenvalues Λ_h and λ_h are given by

$$\lambda_h := \left(\sqrt{(1-\gamma)\xi} - \sqrt{(1-\xi)\gamma}\right)^2$$
$$\Lambda_h := \left(\sqrt{(1-\gamma)\xi} + \sqrt{(1-\xi)\gamma}\right)^2\,.$$

Using the fact that F_h is supported within the interval $(0,1)$, we have recovered the whole density of the limiting spectral distribution F_h of the matrix $U^\top S^\top S U$.

A.4. Proof of Theorem 5

We have already argued that $\{x_t\}$ is asymptotically optimal. It remains to show that $\mathcal{L}^*_{f_h,t} \asymp \frac{(1-\xi)^t}{(1-\gamma)^t}\rho^t$.

Using (25), we have

$$\begin{aligned}\mathcal{L}^*_{f_h,t} &= \frac{c\tau}{(1-\tau)\gamma}\min_{P\in\mathbb{R}^0_t[X]}\int_\alpha^\beta P^2(t)\frac{\mu_\tau(x)}{x-c}\,\mathrm{d}x \\ &= \frac{c\tau}{(1-\tau)\gamma}\min_{P\in\mathbb{R}^0_t[X]}\int_\alpha^\beta P^2(t)\frac{x}{x-c}\frac{\mu_\tau(x)}{x}\,\mathrm{d}x\,.\end{aligned}$$

For $x \in [\alpha,\beta]$, it holds that

$$\frac{\beta}{\beta-c} \leqslant \frac{x}{x-c} \leqslant \frac{\alpha}{\alpha-c}\,,$$

and consequently, we can lower and upper bound $\mathcal{L}^*_{f_h,t}$ as follows,

$$\frac{c\tau}{(1-\tau)\gamma}\frac{\beta}{\beta-c}\min_{P\in\mathbb{R}^0_t[X]}\int_\alpha^\beta P^2(t)\frac{\mu_\tau(x)}{x}\,\mathrm{d}x \leqslant \mathcal{L}^*_{f_h,t} \leqslant \frac{c\tau}{(1-\tau)\gamma}\frac{\alpha}{\alpha-c}\min_{P\in\mathbb{R}^0_t[X]}\int_\alpha^\beta P^2(t)\frac{\mu_\tau(x)}{x}\,\mathrm{d}x\,.$$

From Lemma 2.3, we know that $\mathcal{L}^*_{\mu_\tau,t} = (1-\tau)\min_{P\in\mathbb{R}^0_t[X]}\int_\alpha^\beta P^2(t)\frac{\mu_\tau(x)}{x}\,\mathrm{d}x$. Thus,

$$\frac{c\tau}{(1-\tau)^2\gamma}\frac{\beta}{\beta-c}\mathcal{L}^*_{\mu_\tau,t} \leqslant \mathcal{L}^*_{f_h,t} \leqslant \frac{c\tau}{(1-\tau)^2\gamma}\frac{\alpha}{\alpha-c}\mathcal{L}^*_{\mu_\tau,t}\,.$$

From Theorem 2, we know that $\mathcal{L}^*_{\mu_\tau,t} = \tau^t$. Thus, we obtain that

$$\mathcal{L}^*_{f_h,t} \asymp \tau^t\,.$$

A simple calculation gives that $\tau = \frac{1-\xi}{1-\gamma}\rho$, which yields the claimed result. As for the Gaussian case, an exact calculation of $\mathcal{L}^*_{f_h,t}$ is actually possible. But, after investigation, the resulting expression is lengthy and fairly difficult to simplify, whereas we are primarily interested in the scaling in terms of the iteration number t.

B. Proofs of intermediate results

B.1. Proof of Lemma 2.1

Suppose that $\{x_t\}$ is generated by a first-order method (6). Fix $t \geqslant 1$, then there exists $\alpha_{0,t}, \dots, \alpha_{t-1,t}$ such that

$$x_t = x_{t-1} + \sum_{j=0}^{t-1} \alpha_{j,t} H_S^{-1} A^\top (Ax_j - b)\,. \tag{39}$$

Multiplying both sides of (39) by $U^\top A$, subtracting $U^\top Ax^*$ and using the normal equation $A^\top Ax^* = A^\top b$, we find that

$$\Delta_t = \Delta_{t-1} + \sum_{j=0}^{t-1} \alpha_{j,t} C_S^{-1} \Delta_j\,. \tag{40}$$

First, we aim to show that there exists a polynomial $p_t \in \mathbb{R}_t^0[X]$ such that $\Delta_t = p_t\big(C_S^{-1}\big)\,\Delta_0$. We proceed by induction over $t \geqslant 0$. For $t=0$, the claim is true. Suppose that for some $t \geqslant 1$, it holds that $\Delta_j = p_j\big(C_S^{-1}\big)\,\Delta_0$ with $p_j \in \mathbb{R}_j^0[X]$ for $j=0,\dots,t-1$. Then, we have from (40) that

$$\Delta_t = p_{t-1}\big(C_S^{-1}\big)\,\Delta_0 + \sum_{j=0}^{t-1} \alpha_{j,t} C_S^{-1} p_j\big(C_S^{-1}\big)\,\Delta_0 \tag{41}$$

$$= \left(p_{t-1}\big(C_S^{-1}\big) + \sum_{j=0}^{t-1} \alpha_{j,t} C_S^{-1} p_j\big(C_S^{-1}\big) \right) \Delta_0\,. \tag{42}$$

We set $p_t(x) = p_{t-1}(x) + \sum_{j=0}^{t-1} \alpha_{j,t} x p_j(x)$. It holds that $p_t(0) = p_{t-1}(0) + 0 = 1$, and $\deg(p_t) \leqslant t$ since $\deg(p_{t-1}) \leqslant t-1$ and $\deg(xp_j(x)) \leqslant j+1 \leqslant t$ for $j = 0, \dots, t-1$. Then, from (42), we have $\Delta_t = p_t\big(C_S^{-1}\big)\,\Delta_0$, which concludes the induction.

Second, we aim to show that $\lim_{n\to\infty} \frac{\mathbb{E}[\|\Delta_t\|^2]}{\mathbb{E}[\|\Delta_0\|^2]} = \int_{\mathbb{R}} p^2\left(\lambda^{-1}\right) \mathrm{d}\mu(\lambda)$, where μ is the l.s.d. of C_S for S an $m \times n$ Gaussian or SRHT embedding. The Gaussian case is straightforward to prove, by using the rotational invariance of the Gaussian distribution. The SRHT case is more involved, and we leverage tools from free probability theory.

B.1.1. The Gaussian case

Let S be an $m \times n$ random matrix with i.i.d. entries $\mathcal{N}(0, 1/m)$. Then, by rotational invariance, SU is an $m \times d$ matrix with i.i.d. entries $\mathcal{N}(0, 1/m)$. Write the eigenvalue decomposition $C_S = V \Sigma V^\top$ where V is a $d \times d$ orthogonal matrix, and Σ a diagonal matrix with positive entries $\lambda_1, \dots, \lambda_d$. A standard result states that V and Σ are independent matrices, and V is Haar-distributed.

Fix $t \geqslant 0$, and let $p_t \in \mathbb{R}_t^0[X]$ such that $\Delta_t = p_t\big(C_S^{-1}\big)\,\Delta_0$. Taking the squared norm and the expectation, we obtain that

$$\begin{aligned}\mathbb{E}\left[\|\Delta_t\|^2\right] &= \mathbb{E}\left[\Delta_0^\top p_t^2\big(C_S^{-1}\big)\,\Delta_0\right] \\ &= \mathbb{E}\left[\Delta_0^\top V p_t^2\big(\Sigma^{-1}\big)\,V^\top \Delta_0\right]\,.\end{aligned}$$

Using the independence of Σ, V and Δ_0 and writing $V = [v_1, \dots, v_d]$, we further obtain that

$$\begin{aligned}\mathbb{E}\left[\|\Delta_t\|^2\right] &= \mathbb{E}\big[\Delta_0^\top V \mathbb{E}\left[p_t^2\big(\Sigma^{-1}\big)\right] V^\top \Delta_0\big] \\ &= \sum_{i=1}^{d} \mathbb{E}\big[(v_i^\top \Delta_0)^2\big]\, \mathbb{E}[p_t^2(\lambda_i^{-1})]\,.\end{aligned}$$

Since each v_i is uniformly distributed on the unit sphere, we have that $\mathbb{E}\left[(v_i^\top \Delta_0)^2\right] = \frac{1}{d}\mathbb{E}\|\Delta_0\|^2$, so that

$$\begin{aligned}\mathbb{E}\left[\|\Delta_t\|^2\right] &= \frac{1}{d}\mathbb{E}\|\Delta_0\|^2 \mathbb{E}\left[\sum_{i=1}^{d} p_t^2(\lambda_i^{-1})\right] \\ &= \mathbb{E}\|\Delta_0\|^2\, \frac{1}{d}\,\mathrm{trace}\,\mathbb{E}\big[p_t^2\big(C_S^{-1}\big)\big]\,.\end{aligned}$$

Dividing both sides of the above equation by $\mathbb{E}\|\Delta_0\|^2$ and taking the limit $d \to \infty$, we obtain the claimed result,

$$\frac{\mathbb{E}\left[\|\Delta_t\|^2\right]}{\mathbb{E}\left[\|\Delta_0\|^2\right]} = \int_{\mathbb{R}} p_t^2\left(\lambda^{-1}\right) \mathrm{d}\mu(\lambda)\,.$$

B.1.2. THE SRHT CASE

The SRHT does not satisfy rotational invariance as the Gaussian distribution (or Haar matrices), and we need to use a different approach for this proof, based on *asymptotically liberating sequences of unitary matrices* (Anderson & Farrell, 2014).

Let S be an $m \times n$ SRHT embedding. We denote by μ the l.s.d. of the matrix C_S. Following the same first steps as for the Gaussian case, we have that

$$\mathbb{E}\left[\|\Delta_t\|^2\right] = \mathbb{E}\,\mathrm{trace}\left[p_t^2\left(C_S^{-1}\right)\Delta_0\Delta_0^\top\right] = \mathbb{E}\,\mathrm{trace}\left[p_t^2\left(C_S^{-1}\right)\Sigma_0\right]\,, \tag{43}$$

where $\Sigma_0 := \mathbb{E}\Delta_0\Delta_0^\top$. Writing $p_t^2(x) = \sum_{k=0}^t a_k x^{2k}$, it follows that

$$\mathbb{E}\left[\|\Delta_t\|^2\right] = \sum_{k=0}^t a_k\,\mathbb{E}\,\mathrm{trace}\left[C_S^{-2k}\Sigma_0\right]\,. \tag{44}$$

Introducing the matrix $\widetilde{\Sigma}_0 = \frac{\Sigma_0}{\mathrm{trace}\,\Sigma_0/d}$, we further obtain

$$\frac{\mathbb{E}\left[\|\Delta_t\|^2\right]}{\mathbb{E}\left[\|\Delta_0\|^2\right]} = \sum_{k=0}^t a_k\,\frac{1}{d}\mathbb{E}\,\mathrm{trace}\left[C_S^{-2k}\widetilde{\Sigma}_0\right]\,. \tag{45}$$

We use the following result, whose proof leverages some notions from free probability theory. We defer the proof to Appendix B.2.

Lemma 1. *It holds that for any $k \geqslant 0$,*

$$\lim_{n\to\infty}\frac{1}{d}\mathbb{E}\,\mathrm{trace}\left[C_S^{-2k}\widetilde{\Sigma}_0\right] = \lim_{n\to\infty}\frac{1}{d}\mathbb{E}\,\mathrm{trace}\left[C_S^{-2k}\right] = \int_{\mathbb{R}}\lambda^{-2k}\mathrm{d}\mu(\lambda)\,. \tag{46}$$

Combining (45) and the result of Lemma 1, it follows that

$$\lim_{n\to\infty}\frac{\mathbb{E}\left[\|\Delta_t\|^2\right]}{\mathbb{E}\left[\|\Delta_0\|^2\right]} = \sum_{k=0}^t a_k\int_{\mathbb{R}}\lambda^{-2k}\mathrm{d}\mu(\lambda) = \int_{\mathbb{R}} p_t^2\left(\lambda^{-1}\right)\mathrm{d}\mu(\lambda)\,, \tag{47}$$

which is the claimed result.

B.2. Proof of Lemma 1

We introduce a few needed concepts from free probability that will be used in this proof. We refer the reader to (Voiculescu et al., 1992; Hiai & Petz, 2006; Nica & Speicher, 2006; Anderson et al., 2010) for an extensive introduction to this field. Consider the algebra $\mathcal{A}_n$ of $n \times n$ random matrices. For $X_n \in \mathcal{A}_n$, we define the linear functional $\tau_n(X_n) := \frac{1}{n}\mathbb{E}\left[\mathrm{trace}\,X_n\right]$. Then, we say that a family $\{X_{n,1}, \ldots, X_{n,I}\}$ of random matrices in $\mathcal{A}_n$ is *asymptotically free* if for every $i \in \{1, \ldots, I\}$, $X_{n,i}$ has a limiting spectral distribution, and if $\tau_n\left(\prod_{j=1}^m P_j\left(X_{n,i_j} - \tau\left(P_j(X_{n,i_j})\right)\right)\right) \to 0$ almost surely for any positive integer m, any polynomials $P_1, \ldots, P_m$ and any indices $i_1, \ldots, i_m \in \{1, \ldots, I\}$ with $i_1 \neq i_2, \ldots, i_{m-1} \neq i_m$.

Let S be an $n \times n$ SRHT embedding (we consider the SRHT before discarding its zero rows). By definition, we can write $S = BHW$, where B is an $n \times n$ matrix with i.i.d. Bernoulli entries on the diagonal, with success probability m/n, $H = H_n$ is the n-th Walsh-Hadamard matrix. The matrix W is an $n \times n$ bi-signed permutation, i.e., $W = DP$, where D is a diagonal matrix with i.i.d. random signs, and P is an $n \times n$ uniformly random permutation matrix.

We aim to show that for any $k \geqslant 0$,

$$\lim_{d\to\infty}\tau_d\left(C_S^{-k}\widetilde{\Sigma}_0\right) = \lim_{d\to\infty}\tau_d\left(C_S^{-k}\right)\,. \tag{48}$$

We reduce the problem of proving (48) to the following, which is more simple to treat. The proof of this reduction is deferred to Appendix D.1.

Lemma 2. *Suppose that for any $k \geqslant 0$, we have*

$$\lim_{d\to\infty} \tau_d\left(C_S^k \widetilde{\Sigma}_0\right) = \lim_{d\to\infty} \tau_d\left(C_S^k\right) . \tag{49}$$

Then, the claim (48) *is true for any $k \geqslant 0$.*

Thus, we aim to show (49) for all $k \geqslant 0$.

It holds that

$$\begin{aligned} C_S = U^\top S^\top S U &= (U^\top W^\top H B)(B H W U) \\ &= U^\top W^\top H B^2 H W U \\ &= U^\top W^\top H B H W U \,, \end{aligned}$$

where we used $B^2 = B$ in the fourth equality. Further, we have the following equality in distribution, whose proof is deferred to Appendix B.3.

Lemma 3. *It holds that*

$$U^\top W^\top H B H W U \overset{\mathrm{d}}{=} U^\top W^\top H W B W^\top H W U \,. \tag{50}$$

Consequently,

$$C_S \overset{\mathrm{d}}{=} U^\top W^\top H W B W^\top H W U \,. \tag{51}$$

Let $k \geqslant 0$. We have $W^\top W = I_n$, $U^\top U = I_d$, $B^2 = B$, $H^2 = H$ and $\tau_d(\widetilde{\Sigma}_0) = 1$. Using (50), we find

$$\tau_d\Big(C_S^k \widetilde{\Sigma}_0\Big) = \tau_d\Big((U^\top W^\top H W B W^\top H W U)^k \widetilde{\Sigma}_0\Big) = \frac{n}{d} \cdot \tau_n\big(X_1 (Y X_2)^k\big) \,, \tag{52}$$

where we introduced the matrices $X_1 := W U \widetilde{\Sigma}_0 U^\top W^\top$, $X_2 := W U U^\top W^\top$ and $Y := H W B W^\top H$. These matrices satisfy the following collection of properties, whose proof is deferred to Appendix D.2.

Lemma 4. *It holds that $X_1 X_2 = X_2 X_1 = X_1$, $X_2^2 = X_2$, $Y^2 = Y$,*

$$\lim_{n\infty} \tau_n(X_1) = \lim_{n\infty} \tau_n(X_2) \,, \tag{53}$$

and the sets of matrices $\{X_1, X_2\}$ and $\{Y\}$ are asymptotically free.

Further, for any $k \geqslant 1$, we have

$$\lim_{n\to\infty} \tau_n(X_1 (Y X_2)^k) = \lim_{n\to\infty} \tau_n(X_2 (Y X_2)^k) \,. \tag{54}$$

Now, observe that

$$\begin{aligned} \tau_n(X_2 (Y X_2)^k) &= \tau_n(W U U^\top W^\top (H W B W^\top H W U U^\top W^\top)^k) \\ &= \frac{d}{n} \tau_d((U^\top W^\top H W B W^\top H W U)^k) \\ &= \frac{d}{n} \tau_d(C_S^k) \,, \end{aligned}$$

where we used the commutativity of the trace in the second equality, and the equality in distribution (50) for the third equality. Consequently,

$$\lim_{n\to\infty} \tau_n(X_1 (Y X_2)^k) = \gamma \lim_{d\to\infty} \tau_d(C_S^k) \,. \tag{55}$$

Combining the above equality (55) with equality (52), we obtain the claimed result (49).

B.3. Proof of Lemma 3

Note that both B and D are diagonal matrices whose diagonal entries are i.i.d. random variables, and P is a permutation matrix. Define $\tilde{B} = PBP^\top$ and $\tilde{D} = P^\top DP$, then $\tilde{B} \overset{d}{=} B$, $\tilde{D} \overset{d}{=} D$,

$$DP = P\tilde{D}, \quad P^\top D = \tilde{D}P^\top . \tag{56}$$

It follows that

$$\begin{aligned} U^\top W^\top HWBW^\top HWU &= U^\top P^\top DHDPBP^\top DHDPU \\ &= U^\top P^\top DHP\tilde{D}B\tilde{D}P^\top HDPU \\ &= U^\top P^\top DH_nPB\tilde{D}^2P^\top H_nDPU \\ &= U^\top P^\top DH_nPBP^\top H_nDPU \\ &= U^\top P^\top DH_n\tilde{B}H_nDPU \\ &\overset{d}{=} U^\top P^\top DH_nBH_nDPU, \end{aligned}$$

where the second equation follows from (56), the third equation holds because $\tilde{D}$ and B are diagonal so they commute, while the fourth equation holds because $\tilde{D}^2 = I_n$.

B.4. Proof of Lemma 2.4 – Alternative construction of the polynomials $\{\Pi_k\}$

We recall that for a polynomial P and a measure (resp. density) μ, we will denote $\mu[P] := \int_{\mathbb{R}} P(x)\mu(x)\mathrm{d}\mu(x)$ (resp. $\mu[P] := \int_{\mathbb{R}} P(x)\mu(x)\mathrm{d}x$). Thus, for a density μ, the reader should be aware that $\mu[x]$ and $\mu(x)$ refer to different quantities.

We present an alternative construction of the orthogonal family $\{\Pi_k\}$ with respect to μ_ρ, explicitly based on the Chebyshev polynomials of the second kind. This explicit construction allows us to leverage several properties of the polynomials $\{\Pi_k\}$ which are useful to perform calculations and prove Lemma 2.4, as well as Theorem 2.

We introduce the shifted Chebyshev polynomials of the second kind, which are defined by the recurrence

$$Q_0(x) = 1\,, \quad Q_1(x) = \frac{x-(1+\rho)}{\sqrt{\rho}}\,, \quad Q_{k+1}(x) = \frac{x-(1+\rho)}{\sqrt{\rho}}\,Q_k(x) - Q_{k-1}(x)\,. \tag{57}$$

A standard result states that the polynomials Q_k are orthonormal with respect to the measure $\nu(x)\mathrm{d}x := x\mu_\rho(x)\mathrm{d}x$. We set $\widehat{\Pi}_0(x) = 1$, and for $k \geqslant 1$,

$$\widehat{\Pi}_k(x) := 1 - \sum_{j=1}^{k} (-1)^{j-1}\sqrt{\rho}^{j-1}\, xQ_{k-1}(x)\,. \tag{58}$$

For instance, we have $\widehat{\Pi}_1(x) = 1 - x$ and $\widehat{\Pi}_2(x) = 1 - (2+\rho)x + x^2$.

We aim to show that $\{\widehat{\Pi}_k\}$ is an orthogonal family with respect to μ_ρ and then, that $\widehat{\Pi}_k = \Pi_k$.

First, we show that the polynomials $\widehat{\Pi}_k$ form an orthonormal family with respect to μ_ρ such that $\deg(\widehat{\Pi}_k) = k$ and $\widehat{\Pi}_k(0) = 1$. For $k \geqslant 1$, we define the polynomial $T_k(x) = xQ_{k-1}(x)$ and the measure $\lambda_\rho(x) = x^{-1}\mu_\rho(x)$. We have that $\lambda_\rho[T_kT_\ell] = \nu_\rho[Q_{k-1}Q_{\ell-1}] = \delta_{k\ell}$, so that the T_k are orthonormal with respect to λ_ρ. Since $\deg(Q_{k-1}) = k-1$, we have $\deg(T_k) = k$. We also have $T_k(0) = 0 \cdot Q_{k-1}(0) = 0$.

Second, we show that $\mu_\rho[Q_k] = (-1)^k\sqrt{\rho}^k$, which will immediately imply that

$$\lambda_\rho[T_k] = \lambda_\rho[xQ_{k-1}(x)] = \mu_\rho[Q_{k-1}] = (-1)^{k-1}\sqrt{\rho}^{k-1}\,. \tag{59}$$

We denote $u_k := \mu_\rho[Q_k]$. The measure μ_ρ is a probability measure, so that $u_0 = 1$. Further, we have

$$u_1 = \mu_\rho[Q_1] = \int_a^b \frac{x-(1+\rho)}{\sqrt{\rho}}\mu_\rho(x)\,\mathrm{d}x = \frac{-1-\rho+\int_a^b x\mu_\rho(x)\,\mathrm{d}x}{\sqrt{\rho}}\,.$$

The first moment $\mu_\rho[x]$ is equal to 1, so that $u_1 = -\sqrt{\rho}$. From the recurrence relationship (57), we obtain $u_{k+1} = -\frac{1+\rho}{\sqrt{\rho}} u_k - u_{k-1}$. The characteristic equation $x^2 + \frac{1+\rho}{\sqrt{\rho}} x + 1 = 0$ has roots $-1/\sqrt{\rho}$ and $-\sqrt{\rho}$. Therefore, $u_k = \alpha \frac{(-1)^k}{\sqrt{\rho}^k} + \beta (-1)^k \sqrt{\rho}^k$ for some $\alpha, \beta \in \mathbb{R}$. Using the initial values u_0 and u_1, we find $\alpha = 0$ and $\beta = 1$. This yields the claimed formula for u_k.

Then, using the definition (58), we have

$$\widehat{\Pi}_k = 1 - \sum_{j=1}^{k} (-1)^{j-1} \sqrt{\rho}^{j-1}\, x Q_{k-1}(x) \tag{60}$$

$$= 1 - \sum_{j=1}^{k} \lambda_\rho[T_k]\, T_k(x)\,. \tag{61}$$

Hence, recognizing the Gram-Schmidt orthogonalization of the constant polynomial 1 with respect to $\{T_1, \ldots, T_k\}$, we deduce that the family $\{\widehat{\Pi}_k, T_1, \ldots, T_k\}$ is orthogonal with respect to λ_ρ, and is a basis of $\mathbb{R}_k[X]$. Consider now the variational problem

$$\min_{p \in \mathbb{R}_k^0[X]} \int p^2(x) \lambda_\rho(x)\, \mathrm{d}x\,. \tag{62}$$

Let $p \in \mathbb{R}_k^0[X]$ and decompose p as $p = \alpha_0 \widehat{\Pi}_k + \sum_{j=1}^k \alpha_j T_k$. Using $p(0) = 1$, $\widehat{\Pi}_k(0) = 1$ and $T_j(0) = 0$, we get that α_0 must be equal to 1. Then,

$$\int p^2(x) \lambda_\rho(x)\, \mathrm{d}x = \int \widehat{\Pi}_k^2(x) \lambda_\rho(x)\, \mathrm{d}x + 2 \sum_{j=1}^{k} \alpha_j \int \widehat{\Pi}_k(x) T_j(x) \lambda_\rho(x)\, \mathrm{d}x$$
$$+ \int (\sum_{j=1}^{k} \alpha_j T_j(x))^2 \lambda_\rho(x)\, \mathrm{d}x\,.$$

The cross-term is equal to 0 by orthogonality of the family $\{\widehat{\Pi}_k, T_1, \ldots, T_k\}$. The third-term is non-negative, and equal to 0 if and only if $p = \widehat{\Pi}_k$. Therefore, the minimizer of the variational problem (62) is exactly $\widehat{\Pi}_k$. On the other hand, applying Lemma 2.2 with $\nu = x\lambda_\rho = \mu_\rho$, we know that the solution of each of the problems (62) (for varying k) is unique, and the solutions form an orthogonal family with respect to $x\lambda_\rho(x)\mathrm{d}x = \mu_\rho(x)\mathrm{d}x$. Thus, we obtain that the family $\{\widehat{\Pi}_k\}$ is orthogonal with respect to μ_ρ.

Finally, we show that the sequence $\{\widehat{\Pi}_k\}$ satisfies the recurrence relationship (15). Observe that

$$x\widehat{\Pi}_k(x) = x - \sum_{j=1}^{k} \lambda_\rho[T_j] x T_j(x) = x - \lambda_\rho[T_1] x T_1(x) - \sum_{j=2}^{k} \lambda_\rho[T_j] x T_j(x)$$
$$= x - x^2 - \sum_{j=2}^{k} \lambda_\rho[T_j] x T_j(x)\,.$$

Multiplying (57) by x and using the definition $T_k(x) = xQ_{k-1}(x)$, we find that for $k \geqslant 2$,

$$x\, T_j(x) = \sqrt{\rho}\, (T_{j-1}(x) + T_{j+1}(x)) + (1+\rho) T_j(x)\,.$$

Using the above decomposition of $xT_j(x)$, it obtain $\sum_{j=2}^n \lambda_\rho(T_j) x T_j(x) = s_1 + s_2 + s_3$, where

$$s_1 := \sqrt{\rho} \sum_{j=2}^{k} \lambda_\rho(T_j) T_{j+1}(x) = \sum_{j=2}^{k} (-1)^{j-1} \sqrt{\rho}^j T_{j+1}(x)$$
$$= \sum_{j=3}^{k+1} (-1)^j \sqrt{\rho}^{j-1} T_j(x)$$
$$= \widehat{\Pi}_{k+1}(x) - 1 + T_1(x) - \sqrt{\rho}\, T_2(x)$$
$$= \widehat{\Pi}_{k+1}(x) - 1 + x - x^2 + (1+\rho)\, x\,,$$

the second term is $s_2 := \sqrt{\rho}\sum_{j=2}^{k} \lambda_\rho[T_j]T_{j-1}(x) = \sum_{j=2}^{k}(-1)^{j-1}\sqrt{\rho}^j T_{j-1}(x) = \rho\Big(\widehat{\Pi}_{k-1}(x)-1\Big)$ and the third term is $s_3 := (1+\rho)\sum_{j=2}^{k}\lambda_\rho[T_j]T_j(x) = -(1+\rho)\Big(\widehat{\Pi}_k(x)-1+x\Big)$. Consequently,

$$\begin{aligned} x\Pi_k(x) &= x - x^2 - s_1 - s_2 - s_3 \\ &= x - x^2 - \widehat{\Pi}_{k+1}(x) + 1 - x + x^2 - (1+\rho)x - \rho\,(\widehat{\Pi}_{k-1}(x)-1) + (1+\rho)(\widehat{\Pi}_k(x)-1+x) \\ &= -\widehat{\Pi}_{k+1}(x) - \rho\,\widehat{\Pi}_{k-1}(x) + (1+\rho)\widehat{\Pi}_k(x)\,, \end{aligned}$$

which is the claimed recurrence. We deduce that $\widehat{\Pi}_k = \Pi_k$, and that the family $\{\Pi_k\}$ is orthogonal with respect to μ_ρ.

C. Description of numerical experiments

Numerical simulations are run in *Python* with the numerical linear algebra module *NumPy* and the scientific computation module *SciPy*, on a machine with 256Gb of memory.

To generate an $m \times n$ Haar matrix S_h, we sample an $m \times n$ matrix G with i.i.d. Gaussian entries $\mathcal{N}(0,1)$, and we set S_h to be its $m \times n$ matrix of right singular vectors. To generate an $m \times n$ SRHT matrix, we follow the description given in Section 1. The plots correspond to one trial for each embedding.

C.1. Figure 1

We set $n = 8192$, $d = 1640$ and $m \in \{1720, 3280, 4915\}$. We generate the plots of μ_ρ and $f_{h,r}$ by discretizing their respective supports with step size $1e{-}5$.

C.2. Figures 2 and 3

We generate an $n \times d$ Gaussian matrix G with i.i.d. entries, and we compute its left singular matrix U and right singular matrix V. Then, we set $A = U\Sigma V^\top$, where Σ is a $d \times d$ diagonal matrix with entries $\Sigma_j = 0.98^j$ for $j = 1, \dots, d$. We generate a vector b using a planted model $b = Ax_{\text{pl}} + \frac{1}{\sqrt{n}}\mathcal{N}(0, I_n)$, and $x_{\text{pl}} \sim \frac{1}{\sqrt{d}}\mathcal{N}(0, I_d)$. Note that, although the performance of the algorithms do not depend on the data A and b, we choose a standard statistical model to generate the data, and a data matrix with a very large condition number.

Algorithms 1 and 2 are implemented following their pseudo-code description. We use small perturbations of the algorithmic parameters by setting $a_t^\delta = (1+\delta)a_t$ and $b_t^\delta = (1-\delta)b_t$ with $\delta = 0.01$ – where a_t and b_t correspond to the parameters as described in Theorem 1. Similarly, for the Heavy-ball method with fixed SRHT embeddings and parameters derived based on our new asymptotitc edge eigenvalues ("SRHT (edge eig.)"), we use instead the slightly perturbed edge eigenvalues $\lambda_h^\delta = (1-\delta)\lambda_h$ and $\Lambda_h^\delta = (1+\delta)\Lambda_h$, with $\delta = 0.01$. These small perturbations of the parameters are necessary in practice due to the finite-sample approximations. For the Heavy-ball method with fixed SRHT embeddings based on the bounds of Tropp (2011) ("SRHT (baseline)"), we use the parameters prescribed in (Lacotte & Pilanci, 2019). For the Heavy-ball method with refreshed SRHT embeddings ("SRHT (refreshed)"), we use the parameters prescribed in (Lacotte et al., 2020). For each algorithm, results are averaged over 20 independent trials (using the same data A and b).

D. Proofs of auxiliary results

D.1. Proof of Lemma 2

Suppose that (49) holds. Let $k \geqslant 0$. We have

$$C_S^{-k} = (I_d - (I_d - C_S))^{-k} = \left(\sum_{j=0}^{\infty}(I_d - C_S)^j\right)^k\,, \tag{63}$$

where the series expansion $(I_d - (I_d - C_S))^{-1} = \sum_{j=0}^{\infty}(I_d - C_S)^j$ holds almost surely, due to the fact that C_S has spectrum within $(0,1)$ almost surely. There exist coefficients $\{a_\ell\}$ such that $\left(\sum_{j=0}^{\infty} x^j\right)^k = \sum_{\ell=0} a_\ell x^\ell$, and such that the

sum is absolutely convergent, i.e., $\sum_{\ell=0} |a_\ell||x|^\ell < +\infty$, for any $x \in (0,1)$. Consequently,

$$C_S^{-k} = \sum_{\ell=0}^{\infty} a_\ell C_S^\ell \tag{64}$$

Then, by absolute convergence of $\sum_\ell a_\ell x^\ell$ and using the fact that $\|C_S\|_2 < 1$, we can exchange the operator τ_d and the infinite sum, so that

$$\tau_d\big(C_S^{-k}\big) = \tau_d\left(\sum_{\ell=0}^{\infty} a_\ell C_S^\ell\right) = \sum_{\ell=0}^{\infty} a_\ell \tau_d\big(C_S^\ell\big) \ . \tag{65}$$

and writing the latter as a series in C_S, we obtain the claimed result. Due to the fact that $\sup_\ell \lim_{d\infty} \tau_d\big(C_S^\ell\big) < 1$, and using again the absolute convergence of $\sum_\ell a_\ell x^\ell$ for $|x| < 1$, it follows that

$$\lim_{d\to\infty} \tau_d\big(C_S^{-k}\big) = \lim_{d\to\infty} \sum_{\ell=0}^{\infty} a_\ell \tau_d\big(C_S^\ell\big) \tag{66}$$

$$= \sum_{\ell=0}^{\infty} a_\ell \lim_{d\to\infty} \tau_d\big(C_S^\ell\big) \tag{67}$$

$$= \sum_{\ell=0}^{\infty} a_\ell \lim_{d\to\infty} \tau_d\Big(C_S^\ell \widetilde{\Sigma}_0\Big) \ . \tag{68}$$

Using the same arguments, we find that

$$\sum_{\ell=0}^{\infty} a_\ell \lim_{d\to\infty} \tau_d\Big(C_S^\ell \widetilde{\Sigma}_0\Big) = \lim_{d\to\infty} \tau_d\Big(C_S^{-k} \widetilde{\Sigma}_0\Big) \ , \tag{69}$$

and we conclude that

$$\tau_d\Big(C_S^{-k} \widetilde{\Sigma}_0\Big) = \tau_d\big(C_S^{-k}\big) \tag{70}$$

D.2. Proof of Lemma 4

We have

$$X_1 X_2 = WU\widetilde{\Sigma}_0 U^\top W^\top WUU^\top W^\top = WU\widetilde{\Sigma}_0 U^\top W^\top = X_1$$

where we used in the second equality $U^\top W^\top WU = I_d$. Similarly, we obtain $X_2 X_1 = X_1$.

We have

$$Y^2 = (HWBW^\top H)(HWBW^\top H) = HWBW^\top H = Y$$

where we used in the second equality $BW^\top HHWB = B$.

We have

$$X_2^2 = WUU^\top W^\top WUU^\top W^\top = WUU^\top W^\top = X_2 \, ,$$

where we used in the second equality $U^\top W^\top WU = I_d$.

Further, it holds that

$$\lim_{n\infty} \tau_n(X_1) = \gamma \lim_{d\infty} \tau_d(\widetilde{\Sigma}_0) = \gamma = \lim_{n\infty} \tau_n(X_2), \, .$$

We show asymptotic freeness. Note that the matrices $UU^\top$, B and $\widetilde{\Sigma}_0$ have l.s.d. compactly supported. For the latter, this directly follows from our initial assumption that the condition number of the matrix $U^\top A\mathbb{E}[x_0 x_0^\top]A^\top U + U^\top bb^\top U$

remains bounded. Then, applying Corollary 3.2 from (Anderson & Farrell, 2014) with the set of asymptotically liberating matrices $\{W, HW\}$, we immediately obtain asymptotic freeness of $\{X_1, X_2\}$ and $\{Y\}$.

It remains to show that for any $k \geqslant 0$,

$$\lim_{n\to\infty} \tau_n(X_1(YX_2)^k) = \lim_{n\to\infty} \tau_n(X_2(YX_2)^k)\,. \tag{71}$$

For the rest of this proof, we use the more compact notations $a := X_1$, $b := Y$, $c := X_2$ and $\varphi = \lim_{n\to\infty} \tau_n$. We show (71) by induction over $k \geqslant 0$. For $k = 0$, the claim is true because $\varphi(a) = \varphi(c)$ as shown above. Fix $k \geqslant 1$ and suppose that the claim is true for $j = 0, \dots, k-1$. By asymptotic freeness, we have

$$\varphi\Big((a - \varphi(a))\big((b - \varphi(b))(c - \varphi(c))\big)^k\Big) = 0\,. \tag{72}$$

We expand the left-hand side of the above equation as

$$\begin{aligned}
&\varphi\Big((a - \varphi(a))\big((b - \varphi(b))(c - \varphi(c))\big)^k\Big) \\
= \;& \varphi\big(a(bc)^k\big) + \sum_{\substack{\delta_1,\dots,\delta_{2k}\in\{0,1\}\\(\delta_1,\dots,\delta_{2k})\neq(1,\dots,1)}} \varphi\Big(ab^{\delta_1}c^{\delta_2}b^{\delta_3}\dots c^{\delta_{2k}}(-\varphi(b))^{1-\delta_1}\dots(-\varphi(c))^{1-\delta_{2k}}\Big) \\
= \;& \varphi\big(a(bc)^k\big) + \sum_{\substack{\delta_1,\dots,\delta_{2k}\in\{0,1\}\\(\delta_1,\dots,\delta_{2k})\neq(1,\dots,1)}} (-\varphi(b))^{1-\delta_1}\dots(-\varphi(c))^{1-\delta_{2k}}\varphi\Big(ab^{\delta_1}c^{\delta_2}\dots c^{\delta_{2k}}\Big)\,.
\end{aligned}$$

For binary exponents $(\delta_1, \dots, \delta_{2k}) \neq (1, \dots, 1)$, the product of non-commutative matrices $b^{\delta_1}c^{\delta_2}b^{\delta_3}\dots c^{\delta_{2k}}$ must have a sub-product of the form bb or cc. Using the fact that $b^2 = b$ and $c^2 = c$, it follows that there exists some integer ℓ such that $0 \leqslant \ell < k$, and

$$b^{\delta_1}c^{\delta_2}b^{\delta_3}\dots c^{\delta_{2k}} = (bc)^\ell\,.$$

Using the induction hypothesis, we have

$$\varphi\Big(ab^{\delta_1}c^{\delta_2}\dots c^{\delta_{2k}}\Big) = \varphi(a(bc)^\ell) = \varphi(c(bc)^\ell) = \varphi\Big(cb^{\delta_1}c^{\delta_2}\dots c^{\delta_{2k}}\Big)\,.$$

Consequently, we get

$$\begin{aligned}
&\varphi\Big((a - \varphi(a))\big((b - \varphi(b))(c - \varphi(c))\big)^k\Big) \\
= \;& \varphi\big(a(bc)^k\big) + \sum_{\substack{\delta_1,\dots,\delta_{2k}\in\{0,1\}\\(\delta_1,\dots,\delta_{2k})\neq(1,\dots,1)}} \varphi\Big(cb^{\delta_1}c^{\delta_2}b^{\delta_3}\dots c^{\delta_{2k}}(-\varphi(b))^{1-\delta_1}\dots(-\varphi(c))^{1-\delta_{2k}}\Big)
\end{aligned}$$

On the other hand, using asymptotic freeness again, we have

$$\begin{aligned}
0 = \;& \varphi\Big((c - \varphi(c))\big((b - \varphi(b))(c - \varphi(c))\big)^k\Big) \\
= \;& \varphi\big(c(bc)^k\big) + \sum_{\substack{\delta_1,\dots,\delta_{2k}\in\{0,1\}\\(\delta_1,\dots,\delta_{2k})\neq(1,\dots,1)}} \varphi\Big(cb^{\delta_1}c^{\delta_2}b^{\delta_3}\dots c^{\delta_{2k}}(-\varphi(b))^{1-\delta_1}\dots(-\varphi(c))^{1-\delta_{2k}}\Big)
\end{aligned}$$

Combining the two above sets of equalities, we obtain

$$\varphi\big(a(bc)^k\big) = \varphi\big(c(bc)^k\big)\,,$$

which concludes the induction, and the proof.