A Rank-1 Sketch for Matrix Multiplicative Weights

Yair Carmon YAIRC@STANFORD.EDU
John C. Duchi JDUCHI@STANFORD.EDU
Aaron Sidford SIDFORD@STANFORD.EDU
Kevin Tian KJTIAN@STANFORD.EDU
Stanford University

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We show that a simple randomized sketch of the matrix multiplicative weight (MMW) update enjoys (in expectation) the same regret bounds as MMW, up to a small constant factor. Unlike MMW, where every step requires full matrix exponentiation, our steps require only a single product of the form $e^A b$, which the Lanczos method approximates efficiently. Our key technique is to view the sketch as a *randomized mirror projection*, and perform mirror descent analysis on the *expected projection*. Our sketch solves the online eigenvector problem, improving the best known complexity bounds by $\Omega(\log^5 n)$. We also apply this sketch to semidefinite programming in saddle-point form, yielding a simple primal-dual scheme with guarantees matching the best in the literature.

Keywords: Online learning, spectrahedron, matrix exponential, Lanczos method, mirror descent.

1. Introduction

Consider the problem of online learning over the spectrahedron Δ_n, the set of $n \times n$ symmetric positive semidefinite matrices with unit trace. At every time step t, a player chooses action $X_t \in \Delta_n$, an adversary supplies symmetric gain matrix G_t, and the player earns reward $\langle G_t, X_t \rangle := \mathrm{tr}(G_t X_t)$. We seek to minimize the regret with respect to the best single action (in hindsight),

$$\sup_{X \in \Delta_n} \sum_{t=1}^{T} \langle G_t, X \rangle - \sum_{t=1}^{T} \langle G_t, X_t \rangle = \lambda_{\max}\left(\sum_{t=1}^{T} G_t\right) - \sum_{t=1}^{T} \langle G_t, X_t \rangle. \tag{1}$$

Warmuth and Kuzmin (2008, 2012) solve this problem using the matrix exponentiated gradient algorithm (Tsuda et al., 2005), also known as matrix multiplicative weights (MMW). It is given by

$$X_t = \mathsf{P}^{\mathrm{mw}}\left(\eta \sum_{i=1}^{t-1} G_i\right), \text{ where } \mathsf{P}^{\mathrm{mw}}(Y) := \frac{e^Y}{\mathrm{tr} e^Y}, \tag{2}$$

and $\eta > 0$ is a step size parameter. If the operator norm $\|G_t\|_\infty \le 1$ for every t, using the MMW strategy (2) with $\eta = \sqrt{2\log(n)/T}$ guarantees that the regret (1) is bounded by $\sqrt{2\log(n)T}$; this guarantee is minimax optimal up to a constant (Arora et al., 2012).

Unlike standard (vector) multiplicative weights, MMW is computational expensive to implement in the high-dimensional setting $n \gg 1$. This is due ot the high cost of computing matrix exponentials; currently they require an eigen-decomposition which costs $\Theta(n^3)$ with practical general-purpose methods and $\Omega(n^\omega)$ in theory (Pan and Chen, 1999). This difficulty has led a number of

researchers to consider a rank-k sketch of P^{mw} of the form

$$\mathsf{P}_U(Y) := \frac{e^{Y/2}UU^Te^{Y/2}}{\langle e^Y, UU^T\rangle}, \text{ where } U \in \mathbb{R}^{n\times k} \tag{3}$$

and the elements of U are i.i.d. standard Gaussian. For $k \ll n$, P_U is much cheaper than P^{mw} to compute, since its computation requires only k products of the form e^Ab which can be evaluated efficiently via iterative methods (see Section 3). Since we play rank-deficient matrices, an adversary with knowledge of X_t may choose the gain G_t to be in its nullspace, incurring regret linear in T. To rule such an adversary out, we assume that G_t and X_t must be chosen simultaneously. We formalize this as

Assumption A *Conditionally on $X_1, G_1, \ldots, X_{t-1}, G_{t-1}$, the gain G_t is independent of X_t.*

This assumption is standard in the literature on adversarial bandit problems (Bubeck and Cesa-Bianchi, 2012) where it is similarly unavoidable. While it comes at significant loss of generality, Assumption A holds in two important applications, as described below.

The challenge of bias Assumption A allows us to write

$$\mathbb{E}\left[\left\langle G_t, \mathsf{P}_{U_t}\left(\eta\sum_{i=1}^{t-1}G_i\right)\right\rangle \,\middle|\, \{G_i\}_{i=1}^t\right] = \left\langle G_t, \mathbb{E}_U\mathsf{P}_U\left(\eta\sum_{i=1}^{t-1}G_i\right)\right\rangle.$$

However, even though U satisfies $\mathbb{E}_U UU^T = I$, we have $\mathbb{E}_U\mathsf{P}_U(Y) \neq \mathsf{P}^{\mathrm{mw}}(Y)$ for general Y. Therefore, the guarantees of MMW do not immediately apply to actions chosen according to the sketch (3), even in expectation. A common solution in the literature (Arora and Kale, 2007; Peng et al., 2016; Allen-Zhu et al., 2016) is to pick $k = \widetilde{O}(1/\epsilon^2)$ such that, by the Johnson-Lindenstrauss lemma, $\mathsf{P}_U(Y)$ approximates $\mathsf{P}^{\mathrm{mw}}(Y)$ to within multiplicative error ϵ. This makes the MMW guarantees applicable again, but requires considerable computation per step, that will match the cost of full matrix exponentiation for sufficiently small ϵ. Kalai and Vempala (2005) and Allen-Zhu and Li (2017) prove regret guarantees for sketches of fixed rank $k \leq 3$ with forms different from (3); we discuss their approaches in detail in Section 1.1.

Our approach In this work we use the sketch (3) with $k = 1$, playing the rank-1 matrix $X_t = \mathsf{P}_{u_t}(\eta\sum_{i=1}^{t-1}G_i)$ where $\mathsf{P}_u(Y) = vv^T/(v^Tv)$ for $v = e^{Y/2}u$ and $u_t \in \mathbb{R}^n$ standard Gaussian. Instead of viewing P_u as a biased estimator of P^{mw}, we define the deterministic function

$$\bar{\mathsf{P}}(Y) := \mathbb{E}_u\mathsf{P}_u(Y),$$

and view P_u as an unbiased estimator for $\bar{\mathsf{P}}$. Our primary contribution is in showing that

$$\bar{\mathsf{P}} \text{ is nearly as good a mirror projection as } \mathsf{P}^{\mathrm{mw}}.$$

More precisely, we show that replacing P^{mw} with $\bar{\mathsf{P}}$ leaves the regret bounds almost unchanged; if $\|G_t\|_\infty \leq 1$ for every t, the actions $\bar{X}_t = \bar{\mathsf{P}}(\eta\sum_{i=1}^{t-1}G_i)$ guarantee (with properly tuned η) regret of at most $\sqrt{6\log(4n)T}$, worse than MMW by only a factor of roughly $\sqrt{3}$. To prove this, we establish that $\bar{\mathsf{P}}$ possesses the geometric properties necessary for mirror descent analysis: it is Lipschitz continuous and its associated Bregman divergence is appropriately bounded. Since P_u is—by definition—an unbiased estimator of $\bar{\mathsf{P}}$, we immediately obtain (thanks to Assumption A) that $X_t = \mathsf{P}_{u_t}(\eta\sum_{i=1}^{t-1}G_i)$ satisfies the same regret bound in expectation. High-probability bounds follow immediately via martingale concentration.

Application to online PCA As our sketched actions are of the form $X_t = x_t x_t^T$, the regret they incur is $\lambda_{\max}\left(\sum_{t=1}^T G_t\right) - \sum_{t=1}^T x_t^T G_t x_t$. Therefore, the vectors x_t can be viewed as streaming approximations of the principal component[1] of the cumulative matrix $\sum_{i=1}^{t-1} G_i$. This online counterpart of the classical principal component analysis problem is the topic of a number of prior works (cf. Warmuth and Kuzmin, 2008; Garber et al., 2015; Allen-Zhu and Li, 2017). Our sketch offers regret bounds that are optimal up to constants, with computational cost per step as low as any known alternative, and overall computational cost better than any in the literature by a factor of at least $\log^5 n$ (see Section 1.1). Our regret bounds hold for gains G_t of any rank or sparsity, and our computational scheme (Section 3) naturally leverages low rank and/or sparsity in the gains.

Application to semidefinite programming (SDP) Any feasibility-form SDP is reducible to the matrix saddle-point game $\max_{X\in\Delta_n}\min_{y\in\sigma_m}\langle\sum_{i=1}^m y_i A_i, X\rangle$, where σ_m is the simplex in $\mathbb{R}^m$ and $A_1,\ldots,A_m \in \mathbb{R}^{n\times n}$ are symmetric matrices. A simple procedure for approximating a saddle-point (Nash equilibrium) for this game is to have each player perform online learning, where the max-player observes gains $G_t = \sum_{i=1}^m [y_t]_i A_i$ and the min-player observes costs $[c_t]_i = \langle A_i, X_t\rangle$. Using standard/matrix multiplicative weights for the min/max players, respectively, we may produce approximate solutions with additive error ϵ in $O(\log(nm)/\epsilon^2)$ iterations, with each iteration costing $O(n^3)$ time, due to the MMW computation. In Section 4 we show that by replacing MMW with our sketch we guarantee ϵ error in a similar number of iterations, but with each iteration costing $\widetilde{O}(N/\sqrt{\epsilon})$, where N is the problem description size, which is often significantly smaller than n^2. This guarantee matches the state-of-the-art in a number of settings.

Paper outline After surveying related work in Section 1.1, we present our main contribution in Section 2: regret bounds for our rank-1 randomized projections P_u and their proof via the geometry of $\bar{\mathsf{P}}$. In Section 3 we describe how to compute X_t in $\widetilde{O}(\sqrt{\eta t})$ matrix-vector products using the Lanczos method. In Section 4 we present in detail the application of our sketching scheme to semidefinite programming, as described above. We conclude the paper in Section 5 by discussing a number of possible extensions of our results along with the challenges they present.

1.1. Related work

MMW appears in a large body of work spanning optimization, theoretical computer science, and machine learning (e.g. Nemirovski, 2004; Warmuth and Kuzmin, 2008; Arora et al., 2012). Here, we focus on works that, like us, attempt to relieve the computational burden of computing the matrix exponential, while preserving the MMW regret guarantees. To our knowledge, the first proposal along these lines is due to Arora and Kale (2007), who apply MMW with a Johnson-Lindenstrauss sketch to semidefinite relaxations of combinatorial problems. Subsequent works on positive semidefinite programming adopted this technique (Peng et al., 2016; Allen-Zhu et al., 2016). To achieve ϵ-accurate solutions, these works require roughly ϵ^{-2} matrix exponential vector products per mirror projection.

Baes et al. (2013) apply the accelerated mirror-prox scheme of Nemirovski (2004) to matrix saddle-point problems and approximate P^{mw} using the rank-k sketch (3). Instead of appealing to the JL lemma, they absorb the bias and variance of this approximation directly into the algorithm's error estimates. This enables a more parsimonious choice of k; to attain additive error ϵ, they require $k = \widetilde{O}(\epsilon^{-1})$. See Appendix E.3 for additional discussion of the performance of this method.

1. For this reason we consider gain-maximization rather than loss-minimization, which is generally more conventional.

A different line of work, called Follow the Perturbed Leader (FTPL) (Kalai and Vempala, 2005), eschews matrix exponentiation, and instead produces rank-1 actions $X_t = x_t x_t^T$, where x_t is an approximate top eigenvector of a random perturbation of $\sum_{i=1}^{t-1} G_i$. While a single eigenvector computation has roughly the same cost as a single matrix-exponential vector product, the regret bounds for FTPL—and hence also the total work—scale polynomially in the problem dimension n: Garber et al. (2015) bound the regret by $\widetilde{O}(\sqrt{nT})$ and Dwork et al. (2014) improve the bound to $\widetilde{O}(\sqrt{n^{1/2}T})$ for gains of rank 1. In contrast, the regret of MMW and its sketches depends on n only logarithmically.

Allen-Zhu and Li (2017) give the first fixed-rank sketch with MMW-like regret, proposing a scheme called Follow the Compressed Leader (FTCL). Their approach is based on replacing the MMW mirror projection (2) with the projection corresponding to $\ell_{1-1/q}$ regularization, given by $\mathsf{P}^{\text{q-reg}}(Y) := (c(Y)I - Y)^{-q}$ where $c(Y)$ is the unique $c \in \mathbb{R}$ such that $cI - Y \succ 0$ and $\operatorname{tr}[(cI - Y)^{-q}] = 1$. They use a sketch of $\mathsf{P}^{\text{q-reg}}$ similar in spirit to (3) and prove that $k = 3$ suffices to obtain regret bounds within a polylogarithmic factor of MMW, with q chosen to be roughly $\log n$.

The basis of the FTCL proof strategy is a potential argument used to derive regret bounds for the exact $\mathsf{P}^{\text{q-reg}}$. Their analysis consists of carefully tracing this argument, and accounting for the errors caused by sketching in each step of the way. In comparison, we believe our analysis is more transparent; rather than control multiple series expansion error terms, we establish three simple geometric properties of our projection $\bar{\mathsf{P}}$. We also provide tighter bounds; to guarantee ϵ average regret, FTCL requires a factor of $\Omega(\log^5(n/\epsilon))$ more online learning steps than our method. The per-step computational cost of our method is similar to that of FTCL, with better polylogarithmic dependence on n. On a practical note, the computational scheme we describe in Section 3 is significantly simpler to implement than the one proposed for FTCL.

1.2. Notation

We use upper case letter for matrices and lower case letters for vectors and scalars. We let S_n denote the set of symmetric $n \times n$ matrices, and let $\Delta_n := \{X \in S_n \mid X \succeq 0, \operatorname{tr} X = 1\}$ denote the spectrahedron. We write $\mathbf{1}$ for the all-ones vector, and let $\sigma_n := \{x \in \mathbb{R}^n \mid x \geq 0, \mathbf{1}^T x = 1\}$ denote the simplex. We let $\langle Y, X \rangle = \operatorname{tr}(Y^T X)$ denote the Frobenius inner product between matrices. For $X \in S_n$, we let $\lambda_{\max}(X) = \lambda_1(X) \geq \lambda_2(X) \geq \ldots \geq \lambda_n(X) = \lambda_{\min}(X)$ denote the eigenvalues of X sorted in descending order. For $x \in \mathbb{R}^n$ and $p \geq 1$ we let $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ denote the ℓ_p norm, and for $X \in S_n$, we let $\|X\|_p := \|\lambda(X)\|_p$ be the standard Schatten p-norm. In particular, $\|X\|_\infty = \max\{\lambda_{\max}(X), -\lambda_{\min}(X)\}$ is the Euclidean operator norm and $\|X\|_1 = \sum_{i=1}^n |\lambda_i(X)|$ is the nuclear norm. We write $\mathsf{Uni}(\mathbb{S}^{n-1})$ for the uniform distribution over the unit sphere in $\mathbb{R}^n$.

2. A rank-1 sketch of matrix multiplicative weights

In this section, we state and prove our main result: regret bounds for a rank-1 sketch of the matrix multiplicative weights method. Let us recall our sketch. At time step t, having observed gain matrices $G_1, \ldots, G_{t-1} \in S_n$, we independently draw[2] $u_t \sim \mathsf{Uni}(\mathbb{S}^{n-1})$ and play the rank-1 matrix

$$X_t := \mathsf{P}_{u_t}\left(\eta \sum_{i=1}^{t-1} G_i\right), \text{ where } \mathsf{P}_u(Y) := \frac{e^{Y/2} u u^T e^{Y/2}}{u^T e^Y u} = \frac{v v^T}{v^T v} \text{ for } v = e^{Y/2} u. \tag{4}$$

2. Since P_u is invariant to scaling of u, it has the same distribution for u standard Gaussian or uniform on a sphere.

We call $\mathsf{P}_u : S_n \to \Delta_n$ the *randomized mirror projection*. The key computational consideration is that we can evaluate $\mathsf{P}_u(Y)$ efficiently, while on the analytic side, we show that the update (4) defines *on average* an efficient mirror descent procedure. The regret bounds for X_t then follow.

2.1. Expected regret bounds

The focus of our analysis is the *average mirror projection*

$$\bar{\mathsf{P}}(Y) := \mathbb{E}_u \mathsf{P}_u(Y) \quad \text{and action sequence} \quad \bar{X}_t := \bar{\mathsf{P}}\left(\eta \sum_{i=1}^{t-1} G_i\right), \tag{5}$$

where $\mathbb{E}_u$ denotes expectation w.r.t. to $u \sim \mathsf{Uni}(\mathbb{S}^{n-1})$. As we show in Section 2.3 to come, $\bar{\mathsf{P}}$ is the gradient of the function

$$\bar{\mathsf{p}}(Y) := \mathbb{E}_u \log\left(\left\langle e^Y, uu^T \right\rangle\right) = \mathbb{E}_u \log\left(u^T e^Y u\right),$$

which we also show[3] is a convex spectral function (Lewis, 1996). As a consequence, we can write the average action $\bar{X}_t$ in the familiar dual averaging (Nesterov, 2009) or Follow the Regularized Leader (e.g. Hazan, 2016, Ch. 5) form

$$\bar{X}_t = \operatorname*{argmax}_{X \in \Delta_n} \left\{ \eta \sum_{i=1}^{t-1} \langle G_i, X \rangle - \bar{\mathsf{r}}(X) \right\}$$

where $\bar{\mathsf{r}}(X) = \sup_{Y \in S_n} \{\langle Y, X \rangle - \bar{\mathsf{p}}(Y)\}$ is the convex conjugate of $\bar{\mathsf{p}}$. In this standard approach, the regularizer $\bar{\mathsf{r}}$ defines the scheme, and regret analysis proceeds by showing that $\bar{\mathsf{r}}$ is strongly convex and has bounded range. The former property is equivalent to the smoothness of $\bar{\mathsf{p}}$.

In contrast, our starting point is the definition (5) of the projection $\bar{\mathsf{P}}$, and we find it more convenient to argue about $\bar{\mathsf{P}}$ and $\bar{\mathsf{p}}$ directly. Toward that end, for any $Y, Y' \in S_n$ we let

$$\bar{V}_Y(Y') := \bar{\mathsf{p}}(Y') - \bar{\mathsf{p}}(Y) - \left\langle Y' - Y, \bar{\mathsf{P}}(Y) \right\rangle \tag{6}$$

denote the Bregman divergence induced by $\bar{\mathsf{p}}$. We show that $\bar{V}_Y(\cdot)$ has the properties—analogous to those arising from duality in analyses of dual averaging (Nesterov, 2009)—necessary to establish our regret bounds.

Proposition 1 *The projection $\bar{\mathsf{P}}$ and divergence $\bar{V}$ satisfy*

1. *Smoothness: for every $Y, D \in S_n$, $\bar{V}_Y(Y+D) \le \frac{3}{2}\|D\|_\infty^2$.*

1′. *Refined smoothness for positive shifts: for every $Y, D \in S_n$ such that $D \succeq 0$ and $\|D\|_\infty \le \frac{1}{6}$, $\bar{V}_Y(Y+D) \le 3\|D\|_\infty \langle D, \bar{\mathsf{P}}(Y) \rangle$.*

2. *Diameter bound: for every $Y, Y' \in S_n$, $\bar{V}_Y(0) - \bar{V}_Y(Y') \le \log(4n)$.*

3. *Surjectivity: for every $X \in \operatorname{relint} \Delta_n$ there exists $Y \in S_n$ such that $\bar{\mathsf{P}}(Y) = X$.*

We return to Proposition 1 and prove it in Section 2.3. The proposition gives the following regret bounds for the averaged actions $\bar{X}_t$.

3. For *fixed* $u \in \mathbb{R}^n$, however, $\mathsf{P}_u \neq \nabla \log\left(u^T e^Y u\right)$ and we do not know if it is the gradient of any other function. Moreover, $Y \mapsto \log\left(u^T e^Y u\right)$ is not convex.

Theorem 2 *Let $G_1,\dots,G_T$ be* any *sequence of gain matrices in S_n and let $\bar{X}_t = \bar{\mathsf{P}}(\eta\sum_{i=1}^{t-1} G_i)$ as in Eq. (5). Then, for every $T \in \mathbb{N}$,*

$$\lambda_{\max}\left(\sum_{t=1}^{T} G_t\right) - \sum_{t=1}^{T}\left\langle G_t, \bar{X}_t\right\rangle \le \frac{\log(4n)}{\eta} + \frac{3\eta}{2}\cdot\sum_{t=1}^{T}\|G_t\|_\infty^2. \tag{7}$$

If additionally $0 \preceq G_t \preceq I$ for every t and $\eta \le \frac{1}{6}$,

$$\lambda_{\max}\left(\sum_{t=1}^{T} G_t\right) - \sum_{t=1}^{T}\left\langle G_t, \bar{X}_t\right\rangle \le \frac{\log(4n)}{\eta} + 3\eta\cdot\lambda_{\max}\left(\sum_{t=1}^{T} G_t\right). \tag{8}$$

We prove Theorem 2 in Appendix A. The proof is essentially the standard dual averaging telescoping argument (Nesterov, 2009), which we perform using only the properties in Proposition 1. Indeed, matrix multiplicative weights satisfies a version of Proposition 1 with slightly smaller constant factors, and its regret bounds follow similarly.

The projection $\bar{\mathsf{P}}$ is no easier to compute than the matrix multiplicative weights projection. However, P_u is easily computed and is unbiased for $\bar{\mathsf{P}}$. Consequently—under Assumption A—the sketch P_u inherits the regret guarantees in Theorem 2. To argue this formally, we define the σ-fields

$$\mathcal{F}_t := \sigma(G_1, X_1, \dots, G_t X_t, G_{t+1}),$$

so that $G_t \in \mathcal{F}_{t-1}$ and $\bar{X}_t \in \mathcal{F}_{t-1}$, while, under Assumption A, $\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = \bar{X}_t$ because $u_t \sim \mathsf{Uni}(\mathbb{S}^{n-1})$, independent of $\mathcal{F}_{t-1}$. Consequently, we have the following

Corollary 3 *Let $G_1, \dots, G_T$ be symmetric gain matrices satisfying Assumption A and let X_t be generated according to Eq. (4). Then*

$$\mathbb{E}\left[\lambda_{\max}\left(\sum_{t=1}^{T} G_t\right) - \sum_{t=1}^{T}\langle G_t, X_t\rangle\right] \le \frac{\log(4n)}{\eta} + \frac{3\eta}{2}\cdot\sum_{t=1}^{T}\mathbb{E}\left[\|G_t\|_\infty^2\right].$$

If additionally $0 \preceq G_t \preceq I$ for every t and $\eta \le \frac{1}{6}$,

$$\mathbb{E}\left[\lambda_{\max}\left(\sum_{t=1}^{T} G_t\right) - \sum_{t=1}^{T}\langle G_t, X_t\rangle\right] \le \frac{\log(4n)}{\eta} + 3\eta\cdot\mathbb{E}\left[\lambda_{\max}\left(\sum_{t=1}^{T} G_t\right)\right].$$

Proof Using $G_t \in \mathcal{F}_{t-1}$ and $\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = \bar{X}_t$, we have $\mathbb{E}\langle G_t, X_t\rangle = \mathbb{E}[\mathbb{E}[\langle G_t, X_t\rangle \mid \mathcal{F}_{t-1}]] = \mathbb{E}\langle G_t, \bar{X}_t\rangle$, and so the result is immediate from taking expectation in Theorem 2. ■

It is instructive to compare these guarantees to those for the full (non-approximate) matrix multiplicative weights algorithm. Let

$$\mathcal{R}[T] := \mathbb{E}\left[\lambda_{\max}\left(\frac{1}{T}\sum_{t=1}^{T} G_t\right) - \frac{1}{T}\sum_{t=1}^{T}\langle G_t, X_t\rangle\right]$$

denote the expected *average* regret at time T. If $\|G_t\|_\infty \le 1$ for every t, the bound (7) along with Corollary 3 imply, for $\eta = (2\log(4n)/(3T))^{1/2}$,

$$\mathcal{R}[T] \le \sqrt{\frac{6\log(4n)}{T}}, \text{ i.e. } \mathcal{R}[T] \le \epsilon \text{ for } T \ge \frac{6\log(4n)}{\epsilon^2}.$$

In contrast, the matrix multiplicative weights procedure (2) guarantees average regret below ϵ in $2\log(n)/\epsilon^2$ steps, so our guarantee is worse by a factor of roughly 3.

The bound (8) guarantees smaller *relative* average regret when we additionally assume $0 \preceq G_t \preceq I$ for every t and an a-priori upper bound of the form $\lambda^\star := \lambda_{\max}(\frac{1}{T}\sum_{t=1}^T G_t) \geq \lambda_0$. Here, a judicious choice of η guarantees $\mathcal{R}[T]/\lambda^\star \leq \varepsilon$ for $T \geq 12\log(4n)/(\lambda_0 \varepsilon^2)$. Again, this is slower than the corresponding guarantee for matrix multiplicative weights by a factor of roughly 3. Relative regret bounds of the form (8) are useful in several application of multiplicative weights and its matrix variant (Arora et al., 2012), e.g. width-independent solvers for linear and positive semidefinite programs (Peng et al., 2016).

2.2. High-probability regret bounds

Using standard martingale convergence arguments (cf. Cesa-Bianchi et al., 2004; Nemirovski et al., 2009), we can provide high-probability convergence guarantees for our algorithm. Indeed, we have already observed in Corollary 3 that $\mathbb{E}[\langle G_t, X_t\rangle \,|\, \mathcal{F}_{t-1}] = \langle G_t, \bar{X}_t\rangle$ and therefore $\langle G_t, X_t - \bar{X}_t\rangle$ is a martingale difference sequence adapted to the filtration $\mathcal{F}_t$. As $|\langle G_t, X_t\rangle| \leq \|G_t\|_\infty \|X_t\|_1 = \|G_t\|_\infty$, the martingale $\sum_{i=1}^t \langle G_i, X_i - \bar{X}_i\rangle$ has bounded differences whenever $\|G_t\|_\infty$ is bounded, so that the next theorem is an immediate consequence of the Azuma-Hoeffding inequality and its multiplicative variant (Allen-Zhu and Li, 2017, Lemma G.1)

Corollary 4 *Let $G_1, \ldots, G_T$ be symmetric gain matrices satisfying Assumption A and let X_t be generated according to Eq. (4). If $\|G_t\|_\infty \leq 1$ for every t, then for every $T \in \mathbb{N}$ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\lambda_{\max}\left(\sum_{i=1}^T G_t\right) - \sum_{t=1}^T \langle G_t, X_t\rangle \leq \frac{\log(4n)}{\eta} + \frac{3\eta}{2}T + \sqrt{2T\log\tfrac{1}{\delta}}. \tag{9}$$

If additionally $0 \preceq G_t \preceq I$ for every t and $\eta \leq \frac{1}{6}$, then with probability at least $1-\delta$,

$$\lambda_{\max}\left(\sum_{i=1}^T G_t\right) - \sum_{t=1}^T \langle G_t, X_t\rangle \leq \frac{\log(4n/\delta)}{\eta} + 4\eta\lambda_{\max}\left(\sum_{i=1}^T G_t\right). \tag{10}$$

We give the proof of Corollary 4 in Appendix B.

Our development uses Assumption A only through its consequence $\mathbb{E}[X_t \,|\, \mathcal{F}_{t-1}] = \bar{X}_t$. Therefore, our results apply to any adversary that produces gains with such martingale structure, a weaker requirement than Assumption A.

2.3. Analyzing the average mirror projection

In this section we outline the proof of Proposition 1, which constitutes the core technical contribution of our paper. Our general strategy is to relate the average mirror projection to the multiplicative weights projection, which satisfies a version of Proposition 1. Our principal mathematical tool is the theory of convex, twice-differentiable spectral functions (Lewis, 1996; Lewis and Sendov, 2001).

We begin with the vector log-sum-exp, or softmax, function

$$\mathrm{lse}(v) := \log\left(\sum_{j=1}^n e^{v_j}\right) \quad \text{and its gradient} \quad \nabla \mathrm{lse}(v) = \frac{e^v}{\mathbf{1}^T e^v},$$

where we write e^v for $\exp(\cdot)$ applied elementwise to v and $\mathbf{1}$ for the all-ones vector. Note that $\nabla\mathrm{lse}:\mathbb{R}^n\to\sigma_n$ is the mirror projection associated with (vector) multiplicative weights. Let $Y\in S_n$ have eigen-decomposition $Y=Q\mathrm{diag}(\lambda)Q^T$. The matrix softmax function is

$$\mathsf{p}^{\mathrm{mw}}(Y):=\log\mathrm{tr}e^Y=\mathrm{lse}(\lambda)\ \text{ and }\ \mathsf{P}^{\mathrm{mw}}(Y)=\nabla\mathsf{p}^{\mathrm{mw}}(Y)=\frac{e^Y}{\mathrm{tr}e^Y}=Q\mathrm{diag}(\nabla\mathrm{lse}(\lambda))Q^T$$

is the matrix multiplicative weights mirror projection.

We now connect the function $\bar{\mathsf{p}}(Y)=\mathbb{E}_u[\log\mathrm{tr}(e^Yuu^T)]$ and the projection $\bar{\mathsf{P}}(Y)=\mathbb{E}_u\frac{e^{Y/2}uu^Te^{Y/2}}{u^Te^Yu}$ to their counterparts p^{mw},P^{mw} and lse.

Lemma 5 *Let $Y\in S_n$ have eigen-decomposition $Y=Q\mathrm{diag}(\lambda)Q^T$. Let $w\in\sigma_n$ be drawn from a* $\mathrm{Dirichlet}(\frac{1}{2},\ldots,\frac{1}{2})$ *distribution. Then*

$$\bar{\mathsf{p}}(Y)=\mathbb{E}_w[\mathrm{lse}(\lambda+\log w)]=\mathbb{E}_w\mathsf{p}^{\mathrm{mw}}(Y+Q\mathrm{diag}(\log w)Q^T) \tag{11}$$

where log *is applied elementwise. The function $\bar{\mathsf{p}}$ is convex and its gradient is*

$$\bar{\mathsf{P}}(Y)=\nabla\bar{\mathsf{p}}(Y)=Q\mathrm{diag}(\mathbb{E}_w[\nabla\mathrm{lse}(\lambda+\log w)])Q^T=\mathbb{E}_w\mathsf{P}^{\mathrm{mw}}(Y+Q\mathrm{diag}(\log w)Q^T). \tag{12}$$

Proof Let u be uniformly distributed over the unit sphere in $\mathbb{R}^n$ and note that u and Q^Tu are identically distributed. Therefore, for $\Lambda=\mathrm{diag}(\lambda)$,

$$\bar{\mathsf{p}}(Y)=\mathbb{E}_u\log\big(u^Te^Yu\big)=\mathbb{E}_u\log\big((Q^Tu)^Te^\Lambda(Q^Tu)\big)=\mathbb{E}_u\log\big(u^Te^\Lambda u\big)=\bar{\mathsf{p}}(\Lambda).$$

Further, a vector w with coordinates[4] $w_i=u_i^2$ has a $\mathrm{Dirichlet}(\frac{1}{2},\ldots,\frac{1}{2})$ distribution. Hence,

$$\bar{\mathsf{p}}(\Lambda)=\mathbb{E}_u\log\left(\sum_{i=1}^n u_i^2e^{\lambda_i}\right)=\mathbb{E}_w\log\left(\sum_{i=1}^n e^{\lambda_i+\log w_i}\right)=\mathbb{E}_w\mathrm{lse}(\lambda+\log w),$$

establishing the identity (11).

Evidently, $\bar{\mathsf{p}}(Y)$ is a spectral function—a permutation-invariant function of the eigenvalues of Y. Moreover, since lse is convex, $\lambda\mapsto\mathbb{E}_w\mathrm{lse}(\lambda+\log w)$ is also convex, and Lewis (1996, Corollary 2.4) shows that $\bar{\mathsf{p}}$ is convex. Moreover, Lewis (1996, Corollary 3.2) gives

$$\nabla\bar{\mathsf{p}}(Y)=Q\mathrm{diag}(\nabla\mathbb{E}_w[\mathrm{lse}(\lambda+\log w)])Q^T=\mathbb{E}_w\mathsf{P}^{\mathrm{mw}}(Y+Q\log(w)Q^T).$$

It remains to show that $\bar{\mathsf{P}}(Y)=\nabla\bar{\mathsf{p}}(Y)$. Here we again use the rotational symmetry of u to write

$$\bar{\mathsf{P}}(Y)=\mathbb{E}_u\frac{e^{Y/2}uu^Te^{Y/2}}{u^Te^Yu}=Q\left(\mathbb{E}_u\frac{e^{\Lambda/2}(Q^Tu)(Q^Tu)^Te^{\Lambda/2}}{(Q^Tu)^Te^\Lambda(Q^Tu)}\right)Q^T=Q\bar{\mathsf{P}}(\Lambda)Q^T.$$

Moreover,

$$\bar{\mathsf{P}}(\Lambda)_{ij}=\mathbb{E}_u\frac{u_iu_je^{(\lambda_i+\lambda_j)/2}}{\sum_{k=1}^n u_k^2e^{\lambda_k}}\overset{(\star)}{=}\mathbb{E}_u\frac{u_i^2e^{\lambda_i}\mathbb{I}_{\{i=j\}}}{\sum_{k=1}^n u_k^2e^{\lambda_k}}=\mathbb{E}_w\nabla_i\mathrm{lse}(\lambda+\log w)\mathbb{I}_{\{i=j\}}$$

4. The letter w naturally denotes a vector of 'weights' in the simplex. Here, it is also double-u.

where the equality $(\star)$ above follows because u_i has a symmetric distribution, even conditional on $u_j, j \neq i$, so $\mathbb{E}\left[u_i u_j \,|\, u_1^2,\ldots,u_n^2,u_j\right] = 0$ for $i \neq j$. ■

Lemma 5 is all we need in order to prove parts 2 and 3 of Proposition 1.

Proof (Proposition 1, parts 2 and 3) We first observe the following simple lower bound on $\bar{\mathsf{p}}$, immediate from identity (11) in Lemma 5,

$$\bar{\mathsf{p}}(Y) = \mathbb{E}_w \log\Big(\sum_{i=1}^n e^{\lambda_i(Y)+\log w_i}\Big) \geq \lambda_{\max}(Y) + E_{w_1} \log w_1 \geq \lambda_{\max}(Y) - \log(4n), \tag{13}$$

where $\mathbb{E}_{w_1} \log w_1 \geq -\log(4n)$ comes from noting that $w_1 \sim \mathrm{Beta}(\frac{1}{2}, \frac{n-1}{2})$ (see Lemma 15 in Appendix C.4). For matrices $Y \in S_n$ and $X \in \Delta_n$,

$$\langle Y, X\rangle = \langle Y - \lambda_{\min}(Y) I, X\rangle + \lambda_{\min}(Y) \mathrm{tr} X \leq \|Y - \lambda_{\min}(Y) I\|_\infty \|X\|_1 + \lambda_{\min}(Y) \mathrm{tr} X = \lambda_{\max}(Y),$$

where the final equality is due to $\|Y - \lambda_{\min}(Y) I\|_\infty = \lambda_{\max}(Y) - \lambda_{\min}(Y)$ for every $Y \in S_n$ and $\|X\|_1 = \mathrm{tr} X = 1$ for every $X \in \Delta_n$. Combining this bound with (13), we have that

$$\langle Y, X\rangle - \bar{\mathsf{p}}(Y) \leq \log(4n) \tag{14}$$

for every $Y \in S_n$ and $X \in \Delta_n$. Part 2 follows since

$$\bar{V}_Y(0) - \bar{V}_Y(Y') = \bar{\mathsf{p}}(0) + \left\langle Y', \bar{\mathsf{P}}(Y)\right\rangle - \bar{\mathsf{p}}(Y') \leq \bar{\mathsf{p}}(0) + \log(4n) = \log(4n),$$

where we used the bound (14) with $X = \bar{\mathsf{P}}(Y)$ and the fact that $\bar{\mathsf{p}}(0) = \mathbb{E}_w \log(\mathbf{1}^T w) = 0$.

To show Part 3, let $\bar{\mathsf{r}}(X) := \sup_{Y \in S_n} \{\langle Y, X\rangle - \bar{\mathsf{p}}(Y)\}$ be the convex conjugate of $\bar{\mathsf{p}}$. Eq. (14) implies that $\bar{\mathsf{r}}(X) < \infty$ for all $X \in \Delta_n$, and therefore $\mathrm{relint}\, \Delta_n \subseteq \mathrm{relint}\, \mathrm{dom} \bar{\mathsf{r}}$. Every convex function has nonempty subdifferential on the relative interior of its domain (Hiriart-Urruty and Lemaréchal, 1993, Theorem X.1.4.2), and thus for $X \in \mathrm{relint}\, \Delta_n$ there exists $Y \in \partial \bar{\mathsf{r}}(X)$. By definition of $\bar{\mathsf{r}}$, any such Y satisfies $X = \nabla \bar{\mathsf{p}}(Y) = \bar{\mathsf{P}}(Y)$, as required. ■

Proving parts 1 and 1′ requires second order information on $\bar{\mathsf{p}}$. For twice differentiable function f, we denote $\nabla^2 f(A)[B,B] = \frac{\partial^2}{\partial t^2} f(A + tB)|_{t=0}$. It is easy to verify that, for every $\lambda, \delta \in \mathbb{R}^n$,

$$\delta^T \nabla^2 \mathrm{lse}(\lambda) \delta = \nabla^2 \mathrm{lse}(\lambda)[\delta,\delta] \leq (\delta^2)^T \nabla \mathrm{lse}(\lambda),$$

where $[\delta^2]_i = \delta_i^2$; this concisely captures the pertinent second order structure of the multiplicative weights mirror projection. Nesterov (2007) shows that this property extends to the matrix case.

Lemma 6 *For any $Y, D \in S_n$, $\nabla^2 \mathsf{p}^{\mathrm{mw}}(Y)[D,D] \leq \left\langle D^2, \mathsf{P}^{\mathrm{mw}}(Y)\right\rangle$.*

In Appendix C.1 we explain how to find this result in Nesterov (2007), as it is not explicit there. In view of Lemma 5, it is natural to hope that $\nabla^2 \bar{\mathsf{p}}$ and $\nabla^2 \mathsf{p}^{\mathrm{mw}}$ are also related via simple expectation. Unfortunately, this fails; we can, however, derive a bound.

Lemma 7 *For any $Y, D \in S_n$, orthogonal eigenbasis Q for Y, and $w \sim \mathrm{Dirichlet}(\frac{1}{2},\ldots,\frac{1}{2})$,*

$$\nabla^2 \bar{\mathsf{p}}(Y)[D,D] \leq 3 \cdot \mathbb{E}_w \nabla^2 \mathsf{p}^{\mathrm{mw}}(Y + Q \mathrm{diag}(\log w) Q^T)[D,D] \tag{15}$$

$$\leq 3 \left\langle D^2, \bar{\mathsf{P}}(Y)\right\rangle. \tag{16}$$

Our proof of Lemma 7 is technical; we give it in Appendix C.2. The key ingredient in the proof is a formula for the Hessian of spectral functions (Lewis and Sendov, 2001). The remainder of Proposition 1 follow from the bound (16) via $\bar{V}_Y(Y + D) = \int_0^1 \int_0^t \nabla^2 \bar{\mathsf{p}}(Y + \tau D)[D,D] d\tau dt$; we give the details in Appendix C.3.

3. Efficient computation of matrix exponential-vector products

The main burden in computing the randomized mirror projections (4) lies in computing $e^A b$ for $A \in S_n$ and $b \in \mathbb{R}^n$. Matrix exponential-vector products have widespread use in solutions of differential equations (cf. Saad, 1992; Hochbruck and Ostermann, 2010), and also appear as core components in a number of theoretical algorithms (Arora and Kale, 2007; Orecchia et al., 2012; Jambulapati et al., 2018). Following a large body of literature (cf. Moler and Loan, 2003), we approximate $e^A b$ via the classic Lanczos method (Lanczos, 1950), an iterative process for computing $f(A)b$ for general real functions f applied to matrix A. The Lanczos approximation enjoys strong convergence guarantees upon which we base our analysis (Sachdeva and Vishnoi, 2014). It is also eminently practical: the only tunable parameter is the number of iterations, and each iteration accesses A via a single matrix-vector product.

Let $\widetilde{\exp}_k(A,b)$ be the result of k iterations of the Lanczos method for approximating $e^A b$. We provide a precise description of the method in Appendix D. Let

$$\tilde{X}_{t;k} = \widetilde{\mathsf{P}}_{u_t;k}\left(\eta \sum_{i=1}^{t-1} G_i\right), \text{ where } \widetilde{\mathsf{P}}_{u;k}(Y) = \frac{vv^T}{v^T v} \text{ for } v = \widetilde{\exp}_k(Y/2, u) \tag{17}$$

denote the *approximate* randomized mirror projection. Using the Lanczos method to compute full eigen-decompositions has well-documented numerical stability issues (Meurant, 2006). In contrast, the approximation (17) appears to be numerically stable. To provide a theoretical basis for this observation, we exhibit error bounds under finite floating point precision, leveraging the results of Musco et al. (2017), which in turn build on Druskin and Knizhnerman (1991, 1995). To account for computational cost, we denote by $\mathrm{mv}(Y)$ the cost of multiplying matrix Y by any vector.

Proposition 8 *Let $\epsilon, \delta \in (0,1)$ and $Y \in S_n$, and set $M \coloneqq \max\{\|A\|_\infty, \log(\frac{n}{\epsilon\delta}), 1\}$. Let u be uniformly distributed on the unit sphere in $\mathbb{R}^n$ and independent of Y. If the number of Lanczos iterations k satisfies $k \geq \Theta(1)\sqrt{M \log(\frac{nM}{\epsilon\delta})}$ then the approximation (17) satisfies*

$$\|\mathsf{P}_u(Y) - \widetilde{\mathsf{P}}_{u;k}(Y)\|_1 \leq \epsilon \text{ with probability } \geq 1-\delta \text{ over } u \sim \mathsf{Uni}(\mathbb{S}^{n-1})$$

when implemented using floating point operations with $B = \Theta(1)\log\frac{nM}{\epsilon\delta}$ bits of precision. The time to compute $\widetilde{\mathsf{P}}_{u;k}(Y)$ is $O(\mathrm{mv}(Y)k + k^2 B)$.

We prove Proposition 8 in Appendix D and describe here the main ingredients in the proof. First, we show by calculation that $\|\mathsf{P}_u(Y) - \widetilde{\mathsf{P}}_{u;k}(Y)\|_1 \leq \sqrt{8}\frac{\|e^{Y/2}u - \widetilde{\exp}_k(Y/2,u)\|_2}{\|e^{Y/2}u\|_2}$. Therefore, a multiplicative error guarantee for $\widetilde{\exp}_k$ would imply our result. Unfortunately, for such a guarantee to hold for *all* vectors u we must have $k = \Omega(\|Y\|_\infty)$ (Orecchia et al., 2012, Section 3.3). We circumvent that by using the randomness of u to argue that w.h.p. $\|e^{Y/2}u\|_2 \gtrsim \frac{1}{\sqrt{n}} e^{\lambda_{\max}(Y/2)}\|u\|_2$. This allows us to use existing additive error guarantees for $\widetilde{\exp}_k$ to obtain our result.

We connect the approximation to regret in Appendix D.6. In the setting of Corollary 4, we show that approximating X_t via $\tilde{X}_{t;k_t}$ with $k_t = O(\lceil\sqrt{\eta t}\rceil \log(nT/\delta))$ leaves the regret guarantees essentially unchanged. Therefore, we may achieve ϵ average regret with $O(\epsilon^{-2.5}\log^{2.5}(\frac{n}{\epsilon\delta}))$ matrix-vectors product, with probability at least $1-\delta$.

Finally, as we discuss in detail in Appendix D.5, computing matrix exponential-vector products (and hence P_u) reduces to solving $\widetilde{O}(1)$ linear systems. Since Allen-Zhu and Li (2017) propose to

compute their sketch using a similar reduction, the running time guarantees they establish for their sketch are also valid for ours.

4. Application to semidefinite programming

Here we describe how to use our rank-1 sketch to solve semidefinite programs (SDPs). The standard SDP formulation is, given $\tilde{C},\tilde{A}_1,\ldots,\tilde{A}_{\tilde{m}} \in S_{\tilde{n}}$ and $\tilde{b} \in \mathbb{R}^{\tilde{m}}$,

$$\underset{Z \succeq 0}{\text{minimize}} \langle \tilde{C}, Z \rangle \text{ subject to } \langle \tilde{A}_i, Z \rangle = \tilde{b}_i \quad \forall i \in [\tilde{m}].$$

A binary search over the optimum value reduces this problem to a sequence of feasibility problems. When the constraints imply $\operatorname{tr} Z \le r$ for some $r < \infty$, every intermediate feasibility problem is equivalent to deciding whether there exists X in the spectrahedron Δ_n s.t. $\langle A_i, X \rangle \le 0$ for all $i \in [m]$, with n, m and $A_i \in S_n$ constructed from $\tilde{n}, \tilde{m}, \tilde{A}_i, \tilde{b}, \tilde{C}$ and r. This decision problem is in turn equivalent (cf. Garber and Hazan, 2016) to determining the sign of

$$\mathfrak{s} = \min_{y \in \sigma_m} \max_{X \in \Delta_n} \langle \mathcal{A}^\star y, X \rangle, \text{ where } \mathcal{A}^\star y := \sum_{i \in [m]} [y]_i A_i. \tag{18}$$

and σ_m is the simplex in $\mathbb{R}^m$. We have that $\min_{y' \in \sigma_m} \langle \mathcal{A}^\star y', X \rangle \le \mathfrak{s} \le \max_{X' \in \Delta_n} \langle \mathcal{A}^\star y, X' \rangle$ for every $y \in \sigma_m$ and $X \in \Delta_n$. Therefore, to determine $\mathfrak{s}$ to additive error ϵ, it suffices to find y, X with $\operatorname{Gap}(X, y) \le \epsilon$, where

$$\operatorname{Gap}(X, y) := \max_{X' \in \Delta_n} \langle \mathcal{A}^\star y, X' \rangle - \min_{y' \in \sigma_m} \langle \mathcal{A}^\star y', X \rangle = \lambda_{\max}(\mathcal{A}^\star y) - \min_{i \in [m]} \langle A_i, X \rangle. \tag{19}$$

A basic approach to solving convex-concave games such as (18) is to apply online learning for X and y simultaneously, where at each round the gains/costs to the max/min player are determined by the actions of the opposite player in the previous round. Importantly, such dynamics satisfy Assumption A, and we use our rank-1 sketch as the online learning strategy of the (matrix) max player, and standard multiplicative weights for the (vector) min player. Algorithm 1 describes the resulting scheme. To factor problem scaling into the analysis, we define the *width parameter*

$$\omega := \max_{i \in [m]} \|A_i\|_\infty.$$

In Appendix E.1 we use the standard no-regret argument combined with Corollary 3 to prove the following converges guarantee for Algorithm 1 (a high-probability version follows via Corollary 4).

Theorem 9 *Let $\{X_t, y_t\}_{t=1}^T$ be the actions produced by Algorithm 1 and, define $X_T^{\text{avg}} = \frac{1}{T}\sum_{t=1}^T X_t$, $y_T^{\text{avg}} = \frac{1}{T}\sum_{t=1}^T y_t$. Then*

$$\mathbb{E}\left[\operatorname{Gap}\left(X_T^{\text{avg}}, y_T^{\text{avg}}\right)\right] \le \frac{\log(4mn)}{\eta T} + 2\eta\omega^2.$$

For $\eta = \frac{\log(4mn)}{\sqrt{2\omega^2 T}}$ and $T = \frac{8\log(4mn)\omega^2}{\epsilon^2}$, Theorem 9 guarantees $\mathbb{E}\left[\operatorname{Gap}(X_T^{\text{avg}}, y_T^{\text{avg}})\right] \le \epsilon$. Let $\operatorname{mv}(M)$ denote the time required to multiply M by any vector, let $\operatorname{mv}(\mathcal{A}) := \sum_{i \in [m]} \operatorname{mv}(A_i)$, and let $\operatorname{mv}(\mathcal{A}^\star) := \max_{\alpha \in \mathbb{R}^m} \{\operatorname{mv}(\mathcal{A}^\star \alpha)\} \le \min\{\operatorname{mv}(\mathcal{A}), n^2\}$. At each iteration, computing c_t and y_t takes $O(\operatorname{mv}(\mathcal{A}))$ time while computing X_t takes $\widetilde{O}((\omega/\epsilon)^{0.5}\operatorname{mv}(\mathcal{A}^\star))$ time by Section 3, and G_t need not be formed explicitly. The total computational cost is therefore at most

$$\widetilde{O}\left(\left[(\omega/\epsilon)^{0.5}\operatorname{mv}(\mathcal{A}^\star) + \operatorname{mv}(\mathcal{A})\right]T\right) = \widetilde{O}((\omega/\epsilon)^{2.5}\operatorname{mv}(\mathcal{A}^\star) + (\omega/\epsilon)^2\operatorname{mv}(\mathcal{A})).$$

In Appendices E.2 and E.3 we derive this bound in more detail and compare it with the literature.

Algorithm 1: Primal-dual SDP feasibility

Let $G_0 := 0$ and $c_0 := 0$

for $t = 1,...,T$ **do**

Sample vector u_t uniformly at random from the unit sphere

Play matrix $X_t := \mathsf{P}_{u_t}\left(\sum_{i=1}^{t-1} \eta G_i\right)$

Play vector $y_t := \nabla \mathrm{lse}\left(-\eta \sum_{i=1}^{t-1} c_i\right) = \frac{y_{t-1} \circ e^{-\eta c_{t-1}}}{\mathbf{1}^T (y_{t-1} \circ e^{-\eta c_{t-1}})}$.

Form gain matrix $G_t = \mathcal{A}^\star y_t = \sum_{i \in [m]} [y_t]_i A_i$

Form cost vector $[c_t]_i := \langle X_t, A_i \rangle$, $i \in [m]$

end

5. Discussion

We discuss four additional settings where our sketch might be beneficial. In the first two, the associated online learning problem involves adversaries that violate Assumption A, demonstrating a limitation of our analysis.

Online convex optimization In the online convex optimization problem, at every time step t the adversary provides a convex loss ℓ_t, the players pays a cost $\ell_t(X_t)$ and wishes to minimize the regret $\sum_{t=1}^T \ell_t(X_t) - \min_X \sum_{t=1}^T \ell_t(X)$. The standard reduction to the online learning problem is to construct an adversary with gains $G_t = -\nabla \ell_t(X_t)$. However, even if the losses ℓ_t follow Assumption A, the constructed gains G_t clearly violate it. Therefore, extensions of our results to online convex optimization will require additional work and probably depend on finer problem structure.

Positive semidefinite programming Peng et al. (2016) and Allen-Zhu et al. (2016) propose algorithms for solving positive (packing/covering) semidefinite programs with width independent running time, meaning that the computational cost of solving the problems to ϵ multiplicative error depends only logarithmically on the width parameter (ω in Section 4). Both algorithms rely on matrix exponentiation, which they approximate with a rank $\widetilde{O}(\epsilon^{-2})$ sketch using the Johnson-Lindenstrauss lemma. The algorithm of Peng et al. (2016) uses matrix multiplicative weights in essentially a black-box fashion, so one could hope to replace their high-rank sketch with our rank-1 technique. Unfortunately, the gain matrices that they construct violate Assumption A and so our results do not immediately apply. A rank-1 sketch for this setting remains an intriguing open problem.

Improved computational efficiency against an oblivious adversary An oblivious adversary produces gain matrices $G_1,...,G_T$ independent of the actions $X_1,...,X_T$; this is a stronger version of Assumption A. For such an adversary, if we draw $u \sim \mathsf{Uni}(\mathbb{S}^{n-1})$ and set $u_1 = u_2 = \cdots = u_T = u$, the average regret guarantee of Theorem 3 still applies, as Allen-Zhu and Li (2017) explain. In this setting, it may be possible to make the computation of X_t more efficient by reusing X_{t-1}. Such savings exist in the stochastic setting (when G_t are i.i.d.) via Oja's algorithm (Allen-Zhu and Li, 2017), and would be interesting to extend to the oblivious setting.

Online k eigenvectors Nie et al. (2013) show that a variant of matrix multiplicative weights is also capable of learning online the top k-dimensional eigenspace, with similar regret guarantees. As our rank-1 sketch solves the $k = 1$ leading eigenvector problem, it is interesting to study whether a rank-k sketch solves the k leading eigenvectors problem.

Acknowledgments

YC was supported by the Stanford Graduate Fellowship. JCD was supported by the NSF CAREER award 1553086, the Sloan Foundation and ONR-YIP N00014-19-1-2288. AS was supported by the NSF CAREER Award CCF-1844855. KT was supported by the NSF Graduate Fellowship DGE-1656518.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Follow the compressed leader: Faster online learning of eigenvectors and faster mmwu. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Zeyuan Allen-Zhu, Yin Tat Lee, and Lorenzo Orecchia. Using optimization to obtain a width-independent, parallel, simpler, and faster positive sdp solver. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2016.

Horst Alzer. On some inequalities for the gamma and psi functions. *Mathematics of Computation of the American Mathematical Society*, 66(217):373–389, 1997.

S. Arora, E. Hazan, and S. Kale. The multiplicative weights update method: a meta algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

Sanjeev Arora and Satyen Kale. A combinatorial, primal-dual approach to semidefinite programs. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on the Theory of Computing*. ACM, 2007.

K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 68:357–367, 1967.

Michel Baes, Michael Bürgisser, and Arkadi Nemirovski. A randomized mirror-prox method for solving structured large-scale matrix saddle-point problems. *SIAM Journal on Optimization*, 23 (2):934–962, 2013.

Sébastien Bubeck and Nicoló Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, September 2004.

Michael B. Cohen, Yin Tat Lee, Gary L. Miller, Jakub W. Pachocki, and Aaron Sidford. Geometric median in nearly linear time. *arXiv:1606.05225 [cs.DS]*, 2016.

Alexandre d'Aspremont. Subsampling algorithms for semidefinite programming. *Stochastic Systems*, 1(2):209–436, 2011.

Vladimir Druskin and Leonid Knizhnerman. Error bounds in the simple Lanczos procedure for computing functions of symmetric matrices and eigenvalues. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 31(7):970–983, 1991.

Vladimir Druskin and Leonid Knizhnerman. Krylov subspace approximation of eigenpairs and matrix functions in exact and computer arithmetic. *Numerical Linear Algebra with Applications*, 2(3):205–217, 1995.

Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze Gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the Forty-Sixth Annual ACM Symposium on the Theory of Computing*. ACM, 2014.

Dan Garber and Elad Hazan. Sublinear time algorithms for approximate semidefinite programming. *Math. Program.*, 158(1-2):329–361, 2016.

Dan Garber, Elad Hazan, and Tengyu Ma. Online learning of eigenvectors. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Ming Gu and Stanley C. Eisenstat. A divide-and-conquer algorithm for the symmetric tridiagonal eigenproblem. *SIAM Journal on Matrix Analysis and Applications*, 16(1):172–191, 1995.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3–4):157–325, 2016.

J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I & II*. Springer, New York, 1993.

Marlis Hochbruck and Alexander Ostermann. Exponential integrators. *Acta Numerica*, 19:209–286, 2010.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

Arun Jambulapati, Kirankumar Shiragur, and Aaron Sidford. Efficient structured matrix recovery and nearly-linear time algorithms for solving inverse symmetric M-matrices. *arXiv:1812.06295 [cs.DS]*, 2018.

A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Cornelius Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards*, 45(4), 1950.

Adrian Lewis. Convex analysis on the Hermitian matrices. *SIAM Journal on Optimization*, 6: 164–177, 1996.

Adrian S. Lewis and Hristo S. Sendov. Twice differentiable spectral functions. *SIAM Journal on Matrix Analysis and Applications*, 23(2):368–386, 2001.

Gérard Meurant. *The Lanczos and Conjugate Gradient Algorithms: From Theory to Finite Precision Computations*. Society for Industrial and Applied Mathematics, 2006.

Cleve B. Moler and Charles Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003.

Cameron Musco, Christopher Musco, and Aaron Sidford. Stability of the Lanczos method for matrix function approximation. *arXiv:1708.07788 [cs.DS]*, 2017.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Arkadi Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):261–283, 2009.

Yurii Nesterov. Smoothing technique and its applications in semidefinite optimization. *Mathematical Programming, Series A*, 110:245–259, 2007.

Jiazhong Nie, Wojciech Kotłowski, and Manfred K Warmuth. Online PCA with optimal regrets. In *Proceedings of the Twenty Sixth Annual Conference on Computational Learning Theory*, 2013.

Lorenzo Orecchia, Sushant Sachdeva, and Nisheeth K. Vishnoi. Approximating the exponential, the lanczos method and an $\tilde{o}(m)$-time spectral algorithm for balanced separator. In *Proceedings of the Forty-Fourth Annual ACM Symposium on the Theory of Computing*, 2012.

Victor Y Pan and Zhao Q Chen. The complexity of the matrix eigenproblem. In *Proceedings of the Thirty-First Annual ACM Symposium on the Theory of Computing*. ACM, 1999.

Richard Peng, Kanat Tangwongsan, and Peng Zhang. Faster and simpler width-independent parallel algorithms for positive semidefinite programming. *arXiv:1201.5135v3 [cs.DS]*, 2016.

Yousef Saad. Analysis of some Krylov subspace approximations to the matrix exponential operator. *SIAM Journal on Numerical Analysis*, 29(1):209–228, 1992.

Sushant Sachdeva and Nisheeth K. Vishnoi. Faster algorithms via approximation theory. *Foundations and Trends in Theoretical Computer Science*, 9(2):125–210, 2014.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

Koji Tsuda, Gunnar Rätsch, and Manfred K Warmuth. Matrix exponentiated gradient updates for on-line learning and bregman projection. *Journal of Machine Learning Research*, 6:995–1018, 2005.

Lieven Vandenberghe, Martin S Andersen, et al. Chordal graphs and semidefinite optimization. *Foundations and Trends® in Optimization*, 1(4):241–433, 2015.

Manfred K. Warmuth and Dima Kuzmin. Randomized online PCA algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, 9:2287–2320, 2008.

Manfred K. Warmuth and Dima Kuzmin. Online variance minimization. *Machine Learning*, 87(1): 1–32, 2012.

Appendix

Appendix A. Dual averaging regret bounds

Theorem 2 *Let $G_1,\ldots,G_T$ be* any *sequence of gain matrices in S_n and let $\bar{X}_t = \bar{\mathsf{P}}(\eta\sum_{i=1}^{t-1} G_i)$ as in Eq. (5). Then, for every $T \in \mathbb{N}$,*

$$\lambda_{\max}\left(\sum_{t=1}^{T} G_t\right) - \sum_{t=1}^{T}\left\langle G_t, \bar{X}_t\right\rangle \le \frac{\log(4n)}{\eta} + \frac{3\eta}{2}\cdot\sum_{t=1}^{T}\|G_t\|_\infty^2. \tag{7}$$

If additionally $0 \preceq G_t \preceq I$ for every t and $\eta \le \frac{1}{6}$,

$$\lambda_{\max}\left(\sum_{t=1}^{T} G_t\right) - \sum_{t=1}^{T}\left\langle G_t, \bar{X}_t\right\rangle \le \frac{\log(4n)}{\eta} + 3\eta\cdot\lambda_{\max}\left(\sum_{t=1}^{T} G_t\right). \tag{8}$$

Proof We start with the well-known Bregman 3-point identity, valid for any $\Phi_0,\Phi_1,\Phi_2 \in S_n$,

$$\left\langle \Phi_2 - \Phi_1, \bar{\mathsf{P}}(\Phi_0) - \bar{\mathsf{P}}(\Phi_1)\right\rangle = \bar{V}_{\Phi_0}(\Phi_1) - \bar{V}_{\Phi_0}(\Phi_2) + \bar{V}_{\Phi_1}(\Phi_2); \tag{20}$$

the identity follows from the definition (6) of $\bar{V}$ by direct substitution. Fix some $S \in \operatorname{relint}\Delta_n$ and $S \in S_n$ such that $S = \bar{\mathsf{P}}(\Psi)$ (which exists by Proposition 1.3). Let $Y_t = \eta\sum_{i=1}^{t-1} G_i$ so that $\bar{X}_t = \bar{\mathsf{P}}(Y_t)$. For a given t, we use the 3-point identity with $\Phi_0 = \Psi, \Phi_1 = Y_t$ and $\Phi_2 = Y_{t+1}$, yielding

$$\eta\left\langle G_t, S - \bar{X}_t\right\rangle = \left\langle Y_{t+1} - Y_t, \bar{\mathsf{P}}(\Psi) - \bar{\mathsf{P}}(Y_t)\right\rangle = \bar{V}_\Psi(Y_t) - \bar{V}_\Psi(Y_{t+1}) + \bar{V}_{Y_t}(Y_{t+1}).$$

Summing these equalities over $t = 1,\ldots,T$ and dividing by η gives

$$\left\langle \sum_{t=1}^{T} G_t, S\right\rangle - \sum_{t=1}^{T}\left\langle G_t, \bar{X}_t\right\rangle = \frac{\bar{V}_\Psi(Y_1) - \bar{V}_\Psi(Y_{T+1})}{\eta} + \frac{1}{\eta}\sum_{t=1}^{T}\bar{V}_{Y_t}(Y_{t+1}) \tag{21}$$

$$\le \frac{\log 4n}{\eta} + \frac{3\eta}{2}\sum_{t=1}^{T}\|G_t\|_\infty^2. \tag{22}$$

Above, we used $\bar{V}_{Y_t}(Y_{t+1}) = \bar{V}_{Y_t}(Y_t + \eta G_t) \le \frac{3}{2}\eta^2\|G_t\|_\infty^2$ (Proposition 1.1) along with $Y_1 = 0$ and $\bar{V}_\Psi(0) - \bar{V}_\Psi(Y_{T+1}) \le \log 4n$ (Proposition 1.2).

Since the bound (22) is valid for any $S \in \operatorname{relint}\Delta_n$, we may supremize it over S. The result (7) follows from noting that $\sup_{S\in\operatorname{relint}\Delta_n}\left\langle \sum_{t=1}^{T} G_t, S\right\rangle = \lambda_{\max}\left(\sum_{i=1}^{T} G_t\right)$.

To see the second bound (8), we return to the identity (21) and note that the assumptions $0 \preceq G_t \preceq I$ and $\eta \le \frac{1}{6}$ imply $\|\eta G_t\|_\infty \le \frac{1}{6}$. Therefore we may use Proposition 1.1′ to obtain

$$\bar{V}_{Y_t}(Y_{t+1}) = \bar{V}_{Y_t}(Y_t + \eta G_t) \le 3\|\eta G_t\|_\infty\left\langle \eta G_t, \bar{\mathsf{P}}(Y_t)\right\rangle = 3\eta^2\left\langle G_t, \bar{X}_t\right\rangle.$$

Substituting back into (21), rearranging and taking the supremum over S as before, we obtain

$$\lambda_{\max}\left(\sum_{i=1}^{T} G_t\right) \le (1+3\eta)\sum_{t=1}^{T}\left\langle G_t, \bar{X}_t\right\rangle + \frac{\log(4n)}{\eta}. \tag{23}$$

Dividing through by $(1+3\eta)$ and noting that $1 - x \le \frac{1}{1+x} \le 1$ for every $x \ge 0$, we obtain the result (8), concluding the proof. ∎

Appendix B. High probability regret bounds

Corollary 10 *Let $G_1,\dots,G_T$ be symmetric gain matrices satisfying Assumption A and let X_t be generated according to Eq. (4). If $\|G_t\|_\infty \le 1$ for every t, then for every $T\in\mathbb{N}$ and $\delta\in(0,1)$, with probability at least $1-\delta$,*

$$\lambda_{\max}\left(\sum_{i=1}^{T} G_t\right) - \sum_{t=1}^{T}\langle G_t, X_t\rangle \le \frac{\log(4n)}{\eta} + \frac{3\eta}{2}T + \sqrt{2T\log\tfrac{1}{\delta}}. \tag{9}$$

If additionally $0\preceq G_t\preceq I$ for every t and $\eta\le\frac{1}{6}$, then with probability at least $1-\delta$,

$$\lambda_{\max}\left(\sum_{i=1}^{T} G_t\right) - \sum_{t=1}^{T}\langle G_t, X_t\rangle \le \frac{\log(4n/\delta)}{\eta} + 4\eta\lambda_{\max}\left(\sum_{i=1}^{T} G_t\right). \tag{10}$$

Proof We start with the first claim (9). Recall that a random process D_t adapted to a filtration $\mathcal{F}_t$ is σ^2-sub-Gaussian if $\mathbb{E}[\exp(\lambda D_t)\,|\,\mathcal{F}_{t-1}] \le \exp(\lambda^2\sigma^2/2)$ for all $\lambda\in\mathbb{R}$. Then using the boundedness assumption that $\langle G_t, X_t\rangle \le \|G_t\|_\infty \le 1$, Hoeffding's lemma on bounded random variables (Hoeffding, 1963) implies that the martingale difference sequence $\langle G_t, X_t - \bar{X}_t\rangle$ is 1-sub-Gaussian. Consequently, the Azuma-Hoeffding inequality (Azuma, 1967) immediately implies that

$$\sum_{t=1}^{T}\langle G_t, X_t\rangle \ge \sum_{t=1}^{T}\langle G_t, \bar{X}_t\rangle - \sqrt{2T\log\tfrac{1}{\delta}} \text{ w.p. } \ge 1-\delta.$$

The bound (7) in Theorem 2 thus gives the result (9).

For the multiplicative bound (10), we require a slightly different relative martingale convergence guarantee.

Lemma 11 (Allen-Zhu and Li (2017), Lemma G.1) *Let $\{D_t\}$ be adapted to the filtration $\{\mathcal{F}_t\}$ and satisfy $0\le D_t\le 1$. Then, for any $\delta,\mu\in(0,1)$, and any $T\in\mathbb{N}$,*

$$\mathbb{P}\left(\sum_{t=1}^{T} D_t \ge (1-\mu)\sum_{t=1}^{T}\mathbb{E}[D_t\,|\,\mathcal{F}_{t-1}] - \frac{\log\frac{1}{\delta}}{\mu}\right) \ge 1-\delta.$$

Similarly, the assumption $0\preceq G_t\preceq I$, along with $X_t\in\Delta_n$, imply $0\le\langle G_t, X_t\rangle\le 1$. Therefore, the conditions of Lemma 11 hold for $D_t = \langle G_t, X_t\rangle$, and we use it with $\mu=\eta\le 1$, obtaining

$$\sum_{t=1}^{T}\langle G_t, X_t\rangle \ge (1-\eta)\sum_{t=1}^{T}\langle G_t, \bar{X}_t\rangle - \frac{\log\frac{1}{\delta}}{\eta} \text{ w.p. } \ge 1-\delta.$$

The bound (8) in Theorem 2 thus yields that with probability at least $1-\delta$ over the randomness in X_t and G_t,

$$\sum_{t=1}^{T}\langle G_t, X_t\rangle \ge (1-\eta)(1-3\eta)\lambda_{\max}\left(\sum_{t=1}^{T} G_t\right) - \frac{\log(4n/\delta)}{\eta}.$$

Noting that $(1-\eta)(1-3\eta)\ge 1-4\eta$ completes the proof. ∎

Appendix C. Proofs from Section 2.3

C.1. Proof of Lemma 6

Lemma 6 *For any $Y,D \in S_n$, $\nabla^2 \mathsf{p}^{\mathrm{mw}}(Y)[D,D] \le \langle D^2, \mathsf{P}^{\mathrm{mw}}(Y)\rangle$.*

Proof While the result is evident from the development in (Nesterov, 2007), it is not stated there formally. We therefore derive it here using our notation and one key lemma from (Nesterov, 2007). First, note that

$$\langle D, \nabla \mathsf{p}^{\mathrm{mw}}(Y)\rangle = \langle D, \mathsf{P}^{\mathrm{mw}}(Y)\rangle = \frac{\langle D, e^Y\rangle}{\mathrm{tr} e^Y},$$

where throughout ∇ denotes differentiation with respect to Y and D is viewed as fixed. Applying ∇ again gives,

$$\nabla^2 \mathsf{p}^{\mathrm{mw}}(Y)[D,D] = \left\langle D, \nabla\left(\frac{\langle D, e^Y\rangle}{\mathrm{tr} e^Y}\right)\right\rangle = \frac{\langle D, \nabla\langle D, e^Y\rangle\rangle}{\mathrm{tr} e^Y} - \left(\frac{\langle D, e^Y\rangle}{\mathrm{tr} e^Y}\right)^2 \le \frac{\langle D, \nabla\langle D, e^Y\rangle\rangle}{\mathrm{tr} e^Y}.$$

Note that $\nabla\langle D, e^Y\rangle \neq De^Y$ when D and Y do not commute. However, using the Taylor series for the exponential and the formula $\nabla\langle D, Y^k\rangle = \sum_{i=0}^{k-1} Y^i D Y^{k-1-i}$ gives,

$$\nabla\langle D, e^Y\rangle = \sum_{k=0}^{\infty} \frac{1}{k!} \nabla\left\langle D, Y^k\right\rangle = \sum_{k=1}^{\infty}\sum_{i=0}^{k-1} \frac{1}{k!} Y^i D Y^{k-1-i}.$$

Consequently, we may write

$$\langle D, \nabla\langle D, e^Y\rangle\rangle = \sum_{k=1}^{\infty}\sum_{i=0}^{k-1} \frac{1}{k!}\left\langle D, Y^i D Y^{k-1-i}\right\rangle = \sum_{k=1}^{\infty}\sum_{i=0}^{k-1} \frac{1}{2(k!)}\left\langle D, Y^i D Y^{k-1-i} + Y^{k-1-i} D Y^i\right\rangle.$$

Lemma 1 in (Nesterov, 2007) shows that, when $Y \succeq 0$,

$$\left\langle D, Y^i D Y^{k-1-i} + Y^{k-1-i} D Y^i\right\rangle \le 2\left\langle D^2, Y^{k-1}\right\rangle.$$

Substituting back, this gives

$$\langle D, \nabla\langle D, e^Y\rangle\rangle \le \sum_{k=1}^{\infty} \frac{1}{(k-1)!}\left\langle D^2, Y^{k-1}\right\rangle = \langle D^2, e^Y\rangle,$$

and consequently

$$\nabla^2 \mathsf{p}^{\mathrm{mw}}(Y)[D,D] \le \frac{\langle D^2, e^Y\rangle}{\mathrm{tr} e^Y} = \langle D^2, \mathsf{P}^{\mathrm{mw}}(Y)\rangle = \langle D^2, \nabla \mathsf{p}^{\mathrm{mw}}(Y)\rangle$$

as required. Finally, note that the assumption $Y \succeq 0$ is without loss of generality, as $\mathsf{P}^{\mathrm{mw}}(Y) = \mathsf{P}^{\mathrm{mw}}(Y + cI)$ for every $c \in \mathbb{R}$, and therefore $\nabla^2 \mathsf{p}^{\mathrm{mw}}$ is also invariant to scalar shifts. ■

C.2. Proof of Lemma 7

Lemma 7 *For any $Y, D \in S_n$, orthogonal eigenbasis Q for Y, and $w \sim \mathrm{Dirichlet}(\frac{1}{2},\ldots,\frac{1}{2})$,*

$$\nabla^2 \bar{\mathsf{p}}(Y)[D,D] \le 3 \cdot \mathbb{E}_w \nabla^2 \mathsf{p}^{\mathrm{mw}}(Y + Q\mathrm{diag}(\log w)Q^T)[D,D] \tag{15}$$

$$\le 3\left\langle D^2, \bar{\mathsf{P}}(Y)\right\rangle. \tag{16}$$

C.2.1. Proof overview

Before going into the lengthy argument, let us briefly survey its key components. Our starting point is the formula Lewis and Sendov (2001) provide for the Hessian of spectral functions. Combined with combined with the spectral characterization (11), the formula gives that

$$\nabla^2 \bar{\mathsf{p}}(Y)[D,D] = \mathrm{diag}(\tilde{D})^T \left[\mathbb{E}_w \nabla^2 \mathrm{lse}(\lambda + \log w)\right] \mathrm{diag}(\tilde{D}) + \left\langle \mathbb{E}_w A^w(\lambda), \tilde{D} \circ \tilde{D} \right\rangle.$$

where $\tilde{D} = Q^T D Q$, $\mathrm{diag}(\tilde{D}) \in \mathbb{R}^n$ is the vector containing the diagonal entries of $\tilde{D}$, $A \circ B$ denotes elementwise multiplication of A and B, and $A^w_{ij}(\lambda) := \frac{\nabla_i \mathrm{lse}(\lambda + \log(w)) - \nabla_j \mathrm{lse}(\lambda + \log(w))}{\lambda_i - \lambda_j} \mathbb{I}_{\{i \neq j\}}$. With the shorthand $Y_{\{w\}} := Y + Q\mathrm{diag}(\log w)Q^T$, we use the formula of Lewis and Sendov (2001) again to express $\nabla^2 \mathsf{p}^{\mathrm{mw}}(Y_{\{w\}})$ as

$$\nabla^2 \mathsf{p}^{\mathrm{mw}}(Y_{\{w\}})[D,D] = \mathrm{diag}(\tilde{D})^T \left[\nabla^2 \mathrm{lse}(\lambda + \log w)\right] \mathrm{diag}(\tilde{D}) + \left\langle A^{\mathbf{1}}(\lambda + \log w), \tilde{D} \circ \tilde{D} \right\rangle,$$

where $A^{\mathbf{1}} = A^{\tilde{w}}$ evaluated at $\tilde{w} = \mathbf{1}$. The bulk of the proof is dedicated to establishing the entry-wise bounds

$$\mathbb{E}_w A^w_{ij}(\lambda) \le \mathbb{E}_w \left[\left(1 + \frac{\tanh\left(\frac{\lambda_i - \lambda_j}{2}\right) \left| \log \frac{w_i}{w_j} \right|}{\lambda_i - \lambda_j} \right) A^{\mathbf{1}}_{ij}(\lambda + \log w) \right] \le 3 \cdot \mathbb{E}_w A^{\mathbf{1}}_{ij}(\lambda + \log w).$$

The first inequality follows from pointwise analysis of a symmetrized version of A^w_{ij}. The second inequality follows from piecewise monotonicity of A^w_{ij} as a function of $\log \frac{w_i}{w_j} \sim \mathrm{logitBeta}(\frac{1}{2}, \frac{1}{2})$, combined with tight exponential tail bounds for the latter. Substituting the bound on $\mathbb{E}_w A^w_{ij}(\lambda)$ into the expression for $\nabla^2 \bar{\mathsf{p}}(Y)$ and comparing with $\mathbb{E}_w \nabla^2 \mathsf{p}^{\mathrm{mw}}(Y_{\{w\}})$ yields the desired result (15). Applying Lemma 6 and recalling the identity (12) yields

$$\mathbb{E}_w \nabla^2 \mathsf{p}^{\mathrm{mw}}(Y + Q\mathrm{diag}(\log w)Q^T)[D,D] \le \left\langle D^2, \mathbb{E}_w \mathsf{P}^{\mathrm{mw}}(Y + Q\mathrm{diag}(\log w)Q^T) \right\rangle = \left\langle D^2, \bar{\mathsf{P}}(Y) \right\rangle,$$

establishing the final bound (16).

C.2.2. Full proof

Let $\tilde{D} = Q^T D Q$, where as before $Y = Q \Lambda Q^T$ is an eigen-decomposition and $\Lambda = \mathrm{diag}(\lambda)$. Recall that $\mathrm{lse} : \mathbb{R}^n \to \mathbb{R}$ denotes the vector softmax function, $\mathrm{lse}(y) := \log(\sum_{i=1}^n e^{y_i}) = \mathsf{p}^{\mathrm{mw}}(\mathrm{diag}\, y)$. Similarly, define $\overline{\mathrm{lse}}(y) := \mathbb{E}_w \mathrm{lse}(y + \log w)$ for $w \sim \mathrm{Dirichlet}(\frac{1}{2},\ldots,\frac{1}{2})$. By Lemma 5, $\bar{\mathsf{p}}(Y) = \overline{\mathrm{lse}}(\lambda)$ is a spectral function. Lewis and Sendov (2001, Theorem 3.3) prove that

$$\nabla^2 \bar{\mathsf{p}}(Y)[D,D] = \nabla^2 \overline{\mathrm{lse}}(\lambda)[\mathrm{diag}\tilde{D}, \mathrm{diag}\tilde{D}] + \left\langle \bar{A}(\lambda), \tilde{D} \circ \tilde{D} \right\rangle, \tag{24}$$

where $\circ$ denotes elementwise multiplication, $\mathrm{diag}(\tilde{D})$ is a vector comprised of the diagonal of $\tilde{D}$, and the matrix $\bar{A}$ is given by

$$\bar{A}_{ij}(\lambda) = \frac{\nabla_i \overline{\mathrm{lse}}(\lambda) - \nabla_j \overline{\mathrm{lse}}(\lambda)}{\lambda_i - \lambda_j} = \mathbb{E}_w \underbrace{\frac{\nabla_i \mathrm{lse}(\lambda + \log w) - \nabla_j \mathrm{lse}(\lambda + \log w)}{\lambda_i - \lambda_j}}_{:= A^w_{ij}(\lambda)}$$

for $i \neq j$ and 0 otherwise, whenever λ has distinct elements. This distinctiveness assumption is without loss of generality, as $\bar{\mathsf{p}}$ is $\mathcal{C}^2$ (Lewis and Sendov, 2001, Theorem 4.2) so we may otherwise consider an arbitrarily small perturbation of λ and appeal to continuity of $\nabla^2 \bar{\mathsf{p}}$.

We now use the spectral function Hessian formula to write down $\nabla^2 \mathsf{p}^{\mathrm{mw}}(Y_{\{w\}})[D,D]$ where $Y_{\{w\}} := Y + Q\mathrm{diag}(\log w)Q^T$ (noting that Y and $Y_{\{w\}}$ have the same eigenvectors),

$$\nabla^2 \mathsf{p}^{\mathrm{mw}}(Y_{\{w\}})[D,D] = \nabla^2 \mathrm{lse}(\lambda + \log w)[\mathrm{diag}\tilde{D}, \mathrm{diag}\tilde{D}] + \left\langle A^{\mathrm{mw}}(\lambda + \log w), \tilde{D} \circ \tilde{D} \right\rangle, \tag{25}$$

where

$$A^{\mathrm{mw}}_{ij}(\lambda) := \frac{\nabla_i \mathrm{lse}(\lambda) - \nabla_j \mathrm{lse}(\lambda)}{\lambda_i - \lambda_j} = A^{\mathbf{1}}_{ij}(\lambda)$$

for $i \neq j$ and 0 otherwise. Taking the expectation over w in (25) and recalling the definition $\overline{\mathrm{lse}}(\lambda) = \mathbb{E}_w \mathrm{lse}(\lambda + \log w)$ gives

$$\mathbb{E}_w \nabla^2 \mathsf{p}^{\mathrm{mw}}(Y_{\{w\}})[D,D] = \nabla^2 \overline{\mathrm{lse}}(\lambda)[\mathrm{diag}\tilde{D}, \mathrm{diag}\tilde{D}] + \left\langle \mathbb{E}_w A^{\mathrm{mw}}(\lambda + \log w), \tilde{D} \circ \tilde{D} \right\rangle. \tag{26}$$

Comparing Eq. (26) to (24) and the desired bound (15), we see that it remains to upper bound $\bar{A}(\lambda) = \mathbb{E}_w A^w(\lambda)$ in terms of $\mathbb{E}_w A^{\mathrm{mw}}(\lambda + \log w)$. Fix indices $i, j \in [n]$ such that $i \neq j$, and let

$$\delta := \frac{\lambda_i - \lambda_j}{2} \quad \text{and} \quad \rho := \frac{1}{2} \log \frac{w_i}{w_j}.$$

Since $\bar{A}$ and A^{mw} are both symmetric matrices, we may assume that $\lambda_i > \lambda_j$ and so $\delta > 0$ (recall we assumed $\lambda_i \neq \lambda_j$ without loss of generality). Let $w^{i \leftrightarrow j}$ denote a vector identical to w except coordinates i and j are swapped. With this notation, Lemma 12, which we prove in Section C.2.3, yields the bound

$$A^w_{ij}(\lambda) + A^{w^{i \leftrightarrow j}}_{ij}(\lambda) \leq \left(1 + \frac{|\rho| \tanh(\delta)}{\delta}\right) \left[A^{\mathrm{mw}}_{ij}(\lambda + \log w) + A^{\mathrm{mw}}_{ij}(\lambda + \log w^{i \leftrightarrow j})\right].$$

Taking the expectation over w and using the fact that $\mathrm{Dirichlet}(\frac{1}{2}, \ldots, \frac{1}{2})$ is invariant to permutations, we have

$$\bar{A}_{ij}(\lambda) \leq \mathbb{E}_w \left[\left(1 + \frac{|\rho| \tanh(\delta)}{\delta}\right) A^{\mathrm{mw}}_{ij}(\lambda + \log w)\right]. \tag{27}$$

We now focus on the term $\mathbb{E}_w \frac{|\rho| \tanh(\delta)}{\delta} A^{\mathrm{mw}}_{ij}(\lambda + \log w)$. We have

$$\begin{aligned}
\mathbb{E}_w \frac{|\rho| \tanh(\delta)}{\delta} A^{\mathrm{mw}}_{ij}(\lambda + \log w) &= \mathbb{E}_w \frac{|\rho| \tanh(\delta)}{\delta} A^{\mathrm{mw}}_{ij}(\lambda + \log w) \left[\mathbb{I}_{\{|\rho| \leq \delta\}} + \mathbb{I}_{\{|\rho| > \delta\}}\right] \\
&\leq (\tanh \delta) \mathbb{E}_w A^{\mathrm{mw}}_{ij}(\lambda + \log w) \mathbb{I}_{\{|\rho| \leq \delta\}} + \frac{\tanh \delta}{\delta} \mathbb{E}_w |\rho| A^{\mathrm{mw}}_{ij}(\lambda + \log w) \mathbb{I}_{\{|\rho| > \delta\}},
\end{aligned} \tag{28}$$

where the final transition uses $|\rho|\mathbb{I}_{\{|\rho|\le\delta\}} \le \delta\mathbb{I}_{\{|\rho|\le\delta\}}$ and $A^{\mathrm{mw}}_{ij}(\zeta)\ge 0$ for every $\zeta\in\mathbb{R}^n$. The latter is a consequence of the convexity of lse and is also evident from Eq. (33) in Section C.2.3.

Since $w\sim\mathrm{Dirichlet}(\frac{1}{2},\ldots,\frac{1}{2})$, $\rho=\frac{1}{2}\log\frac{w_i}{w_j}$ is independent of $w_{\setminus ij}:=\{w_k\}_{k\neq i,j}$. Moreover, w_i,w_j are completely determined by ρ and $w_{\setminus ij}$ (see explicit expression in Section C.2.4). Therefore, conditional on $w_{\setminus ij}$, $A^{\mathrm{mw}}_{ij}(\lambda+\log w)$ is a function of ρ. In Lemma 13 we prove that for every λ and $w_{\setminus ij}$, this function is decreasing in ρ for $\rho>\delta$. Hence, conditionally on $w_{\setminus ij}$ and the event $\rho>\delta$, the random variables $|\rho|$ and $A^{\mathrm{mw}}_{ij}(\lambda+\log w)$ are *negatively correlated*: the expectation of their product at most the product of their expectations. Let $\mathbb{E}_\rho$ denote expectation conditional on $w_{\setminus ij}$. Lemma 14, with $f(\rho)=|\rho|$, $g(\rho)=A^{\mathrm{mw}}_{ij}(\lambda+\log w)$, and $\mathcal{S}=\{\rho\,|\,\rho>\delta\}$ gives that

$$\mathbb{E}_\rho|\rho|A^{\mathrm{mw}}_{ij}(\lambda+\log w)\mathbb{I}_{\{\rho>\delta\}} \le (\mathbb{E}_\rho[|\rho|\,|\,\rho>\delta])\big(\mathbb{E}_\rho A^{\mathrm{mw}}_{ij}(\lambda+\log w)\mathbb{I}_{\{\rho>\delta\}}\big). \tag{29}$$

Similarly, Lemma 13 also gives that (conditional on $w_{\setminus ij}$) $A^{\mathrm{mw}}_{ij}(\lambda+\log w)$ is increasing in ρ for $\rho<-\delta$, and therefore, by Lemma 14,

$$\mathbb{E}_\rho|\rho|A^{\mathrm{mw}}_{ij}(\lambda+\log w)\mathbb{I}_{\{\rho<-\delta\}} \le (\mathbb{E}_\rho[|\rho|\,|\,\rho<-\delta])\big(\mathbb{E}_\rho A^{\mathrm{mw}}_{ij}(\lambda+\log w)\mathbb{I}_{\{\rho<-\delta\}}\big). \tag{30}$$

Let $z\sim\mathrm{Beta}(\frac{1}{2},\frac{1}{2})$. The random variable $\rho=\frac{1}{2}\log\frac{w_i}{w_j}$ is symmetric and distributed as $\frac{1}{2}\log(\frac{1-z}{z})$. Therefore

$$\mathbb{E}_\rho[|\rho|\,|\,\rho<-\delta]=\mathbb{E}_\rho[|\rho|\,|\,\rho>\delta]=\frac{1}{2}\mathbb{E}\big[\log\tfrac{1-z}{z}\,|\,\log\tfrac{1-z}{z}>2\delta\big]\overset{(\star)}{\le}\delta+\sqrt{1+e^{-2\delta}},$$

where we prove the inequality $(\star)$ in Lemma 17. Substituting this bound into inequalities (29) and (30) and summing them, we obtain

$$\mathbb{E}_\rho|\rho|A^{\mathrm{mw}}_{ij}(\lambda+\log w)\mathbb{I}_{\{|\rho|>\delta\}} \le \left(\delta+\sqrt{1+e^{-2\delta}}\right)\mathbb{E}_\rho A^{\mathrm{mw}}_{ij}(\lambda+\log w)\mathbb{I}_{\{|\rho|>\delta\}}.$$

Taking expectation over $w_{\setminus ij}$ and substituting back into (28) therefore gives,

$$\mathbb{E}_w\frac{|\rho|\tanh(\delta)}{\delta}A^{\mathrm{mw}}_{ij}(\lambda+\log w) \le \left(\tanh(\delta)+\sqrt{1+e^{-2\delta}}\cdot\frac{\tanh(\delta)}{\delta}\right)\mathbb{E}_w A^{\mathrm{mw}}_{ij}(\lambda+\log w),$$

where we used again $A^{\mathrm{mw}}_{ij}(\cdot)\ge 0$ in order to increase the multiplier of $\mathbb{E}_w A^{\mathrm{mw}}_{ij}(\lambda+\log w)\mathbb{I}_{\{|\rho|\le\delta\}}$. Computation shows that $\tanh(\delta)+\sqrt{1+e^{-2\delta}}\cdot\frac{\tanh(\delta)}{\delta}\le 1.58\le 2$ for every $\delta\ge 0$. Therefore, by the bound (27) we have

$$\bar{A}_{ij}(\lambda)\le 3\cdot\mathbb{E}_w A_{ij}(\lambda+\log w). \tag{31}$$

Returning to (24), we write

$$\begin{aligned}\nabla^2\bar{\mathsf{p}}(Y)[D,D] &\le \nabla^2\overline{\mathrm{lse}}(\lambda)[\mathrm{diag}\tilde{D},\mathrm{diag}\tilde{D}]+3\left\langle\mathbb{E}_w A^{\mathrm{mw}}(\lambda+\log w),\tilde{D}\circ\tilde{D}\right\rangle\\ &\le 3\left[\nabla^2\overline{\mathrm{lse}}(\lambda)[\mathrm{diag}\tilde{D},\mathrm{diag}\tilde{D}]+\left\langle\mathbb{E}_w A^{\mathrm{mw}}(\lambda+\log w),\tilde{D}\circ\tilde{D}\right\rangle\right].\end{aligned}$$

In the first inequality above, we substituted the bound (31), using the fact that all the entries of $\tilde{D}\circ\tilde{D}$ are nonnegative. In the second inequality, we used that fact that $\nabla^2\overline{\mathrm{lse}}(\lambda)[\mathrm{diag}\tilde{D},\mathrm{diag}\tilde{D}]\ge 0$ since $\overline{\mathrm{lse}}$ is convex. Recalling the expression (26) gives (15). The final bound (16) follows from applying Lemma 6 to the right side of (15) and using the identity (12).

C.2.3. A pointwise bound for Lemma 7

In this section we prove an elementary inequality that plays a central role in the proof of Lemma 7. Let $i,j\in[n]$ be such that $i\neq j$. For $\lambda\in\mathbb{R}^n$, we define

$$N_{ij}(\lambda) := \nabla_i \mathrm{lse}(\lambda) - \nabla_j \mathrm{lse}(\lambda) = \frac{e^{\lambda_i}-e^{\lambda_j}}{\sum_{k=1}^n e^{\lambda_k}} = \frac{\sinh\left(\frac{\lambda_i-\lambda_j}{2}\right)}{\cosh\left(\frac{\lambda_i-\lambda_j}{2}\right)+\frac{1}{2}\sum_{k\neq i,j} e^{\lambda_k-\frac{\lambda_i+\lambda_j}{2}}} \tag{32}$$

and

$$A_{ij}^{\mathrm{mw}}(\lambda) = \frac{N_{ij}(\lambda)}{\lambda_i-\lambda_j} \text{ and } A_{ij}^{w}(\lambda) = \frac{N_{ij}(\lambda+\log w)}{\lambda_i-\lambda_j}. \tag{33}$$

Additionally, for any vector $w\in\mathbb{R}^n$, let $w^{i\leftrightarrow j}$ denote a vector identical to w except coordinates i and j are swapped. With this notation in hand, we state and prove our bound.

Lemma 12 *Let $\lambda\in\mathbb{R}^n$, $w\in\mathbb{R}^n_+$ and $i,j\in[n]$, $i\neq j$. Set $\delta=\frac{\lambda_i-\lambda_j}{2}$ and $\rho=\frac{1}{2}\log\frac{w_i}{w_j}$. Then,*

$$A_{ij}^{w}(\lambda)+A_{ij}^{w^{i\leftrightarrow j}}(\lambda) \le \left(1+\frac{|\rho|\tanh(\delta)}{\delta}\right)\left[A_{ij}^{\mathrm{mw}}(\lambda+\log w)+A_{ij}^{\mathrm{mw}}(\lambda+\log w^{i\leftrightarrow j})\right].$$

Proof Define

$$r = \frac{1}{2}\sum_{k\notin\{i,j\}} e^{\lambda_k+\log w_k-\frac{\lambda_i+\log w_i+\lambda_j+\log w_j}{2}} \ge 0.$$

Observe that if we swap w_i and w_j, δ and r remain unchanged and the sign of ρ reverses. For $x\in\mathbb{R}$, let $f(x) := \frac{\sinh(x)}{\cosh(x)+r}$. Using (32), we may write

$$q_1 := 2A_{ij}^{w}(\lambda)+2A_{ij}^{w^{i\leftrightarrow j}}(\lambda) = \frac{f(\delta+\rho)}{\delta}+\frac{f(\delta-\rho)}{\delta}$$

and

$$q_2 := 2A_{ij}^{\mathrm{mw}}(\lambda+\log w)+2A_{ij}^{\mathrm{mw}}(\lambda+\log w^{i\leftrightarrow j}) = \frac{f(\delta+\rho)}{\delta+\rho}+\frac{f(\delta-\rho)}{\delta-\rho}.$$

With these definitions, our goal is to prove that $\frac{q_1-q_2}{q_2}\le\frac{|\rho|\tanh(\delta)}{\delta}$. Since $f(x)$ is an odd function of x, the terms q_1 and q_2 are invariant to sign flips in either δ or ρ. Therefore, we may assume both

$$\delta\ge 0 \text{ and } \rho\ge 0$$

without loss of generality.

Substituting back the expressions for q_1,q_2 and using that $|\rho|=\rho$ by assumption yields

$$\frac{q_1-q_2}{q_2} = \frac{\rho}{\delta}\cdot\frac{\frac{f(\delta+\rho)}{\delta+\rho}-\frac{f(\delta-\rho)}{\delta-\rho}}{\frac{f(\delta+\rho)}{\delta+\rho}+\frac{f(\delta-\rho)}{\delta-\rho}} = \frac{\rho}{\delta}\cdot\frac{g(\delta+\rho)-g(\delta-\rho)}{g(\delta+\rho)+g(\delta-\rho)}, \tag{34}$$

where

$$g(x) := \frac{f(x)}{x} = \frac{\tanh(x)}{x}\cdot\frac{\cosh(x)}{\cosh(x)+r}.$$

Note that $\frac{\tanh(x)}{x}$ is decreasing in $|x|$. Since $|\delta-\rho|\le|\delta+\rho|$ by the assumption $\rho,\delta\ge 0$, we have

$$g(\delta-\rho)\ge\frac{\tanh(\delta+\rho)}{\delta+\rho}\cdot\frac{\cosh(\delta-\rho)}{\cosh(\delta-\rho)+r}.$$

and therefore

$$g(\delta+\rho)-g(\delta-\rho)\le\frac{\tanh(\delta+\rho)}{\delta+\rho}\left(\frac{\cosh(\delta+\rho)}{\cosh(\delta+\rho)+r}-\frac{\cosh(\delta-\rho)}{\cosh(\delta-\rho)+r}\right)$$

and similarly,

$$g(\delta+\rho)+g(\delta-\rho)\ge\frac{\tanh(\delta+\rho)}{\delta+\rho}\left(\frac{\cosh(\delta+\rho)}{\cosh(\delta+\rho)+r}+\frac{\cosh(\delta-\rho)}{\cosh(\delta-\rho)+r}\right).$$

As $g(x)>0$ for every x, we may divide these bounds and obtain via elementary manipulation,

$$\begin{aligned}\frac{g(\delta+\rho)-g(\delta-\rho)}{g(\delta+\rho)+g(\delta-\rho)}&\le\frac{\frac{\cosh(\delta+\rho)}{\cosh(\delta+\rho)+r}-\frac{\cosh(\delta-\rho)}{\cosh(\delta-\rho)+r}}{\frac{\cosh(\delta+\rho)}{\cosh(\delta+\rho)+r}+\frac{\cosh(\delta-\rho)}{\cosh(\delta-\rho)+r}}\\&=\frac{r[\cosh(\delta+\rho)-\cosh(\delta-\rho)]}{2\cosh(\delta+\rho)\cosh(\delta-\rho)+r[\cosh(\delta+\rho)+\cosh(\delta-\rho)]}\\&\le\frac{\cosh(\delta+\rho)-\cosh(\delta-\rho)}{\cosh(\delta+\rho)+\cosh(\delta-\rho)}=\tanh(\rho)\tanh(\delta)\le\tanh(\delta).\end{aligned}$$

Substituting back into (34) establishes the desired bound. Examining the proof, we see that the bound is tight for large values of r and $|\rho|$. ■

C.2.4. Piecewise monotonicity of A^{mw}

Lemma 13 *Let $\lambda\in\mathbb{R}^n$, $w\in\sigma_n$ (the simplex in $\mathbb{R}^n$), and $i,j\in[n]$ such that $\delta:=\frac{1}{2}(\lambda_i-\lambda_j)>0$, and set $\rho:=\frac{1}{2}\log\frac{w_i}{w_j}$. When λ and $\{w_k\}_{k\neq i,j}$ are held fixed, $A^{\mathrm{mw}}_{ij}(\lambda+\log w)$ is increasing in ρ for $\rho<-\delta$, and decreasing in ρ for $\rho>\delta$.*

Proof First, we write $A^{\mathrm{mw}}_{ij}(\lambda+\log w)$ explicitly as a function of ρ, with λ and $\{w_k\}_{k\neq i,j}$ as fixed parameters. By (33) we have

$$A^{\mathrm{mw}}_{ij}(\lambda+\log w)=\frac{\sinh(\rho+\delta)}{2(\rho+\delta)\left[\cosh(\rho+\delta)+\frac{1}{2}\sum_{k\notin\{i,j\}}\frac{w_k}{\sqrt{w_iw_j}}e^{\lambda_k-\frac{\lambda_i+\lambda_j}{2}}\right]}.$$

Let $m=w_i+w_j=1-\sum_{k\neq i,j}w_k$. Since $\frac{w_i}{w_j}=e^{2\rho}$ and $w\in\sigma_n$, we have that $w_i=\frac{m}{1+e^{-2\rho}}$ and $w_j=\frac{m}{1+e^{2\rho}}$. Therefore,

$$\frac{1}{\sqrt{w_iw_j}}=\frac{1}{m}\sqrt{(1+e^{-2\rho})(1+e^{2\rho})}=\frac{2}{m}\cosh(\rho).$$

Thus,

$$A^{\mathrm{mw}}_{ij}(\lambda+\log w)=\frac{\sinh(\rho+\delta)}{2(\rho+\delta)[\cosh(\rho+\delta)+r_0\cosh(\rho)]},$$

where $r_0 = \sum_{k\notin\{i,j\}} \frac{w_k}{m} e^{\lambda_k - \frac{\lambda_i+\lambda_j}{2}}$ is a function of only λ and $\{w_k\}_{k\neq i,j}$, and therefore $A_{ij}^{\mathrm{mw}}(\lambda + \log w)$ can be viewed as a function of ρ as claimed.

Writing $x = \rho + \delta$, showing the desired monotonicity properties is equivalent to showing that

$$b(x) := \frac{\sinh(x)}{x(\cosh(x) + r_0\cosh(x-\delta))}$$

is decreasing for $x > 2\delta$ and increasing for $x < 0$. The derivative of $b(x)$ is

$$b'(x) = \frac{\cosh(x) - \frac{1}{x}\sinh(x)}{x(\cosh(x) + r_0\cosh(x-\delta))} - \frac{\sinh(x)[\sinh(x) + r_0\sinh(x-\delta)]}{x[\cosh(x) + r_0\cosh(x-\delta)]^2},$$

and has, for all $x \in \mathbb{R}$, the same sign as

$$s := \frac{x[\cosh(x) + r_0\cosh(x-\delta)]}{\sinh(x)} b'(x) = \coth(x) - \frac{1}{x} - \frac{\sinh(x) + r_0\sinh(x-\delta)}{\cosh(x) + r_0\cosh(x-\delta)}. \tag{35}$$

For $x > 2\delta$, we have by Dan's favorite inequality ($\frac{a_1+a_2}{b_1+b_2} \geq \min\{\frac{a_1}{b_1}, \frac{a_2}{b_2}\}$ for all $a_1, a_2, b_1, b_2 \geq 0$),

$$\frac{\sinh(x) + r_0\sinh(x-\delta)}{\cosh(x) + r_0\cosh(x-\delta)} \geq \min\{\tanh(x), \tanh(x-\delta)\} = \tanh(x-\delta) > \tanh(x/2),$$

where in the last transition we used the fact that $x > 2\delta$ implies $x - \delta > x/2$. Therefore, for $x > 2\delta$ we have the following bound for s,

$$s \leq \coth(x) - \frac{1}{x} - \tanh(x/2) = \frac{1}{\sinh(x)} - \frac{1}{x} < 0,$$

so we have that $b(x)$ is decreasing for $x > 2\delta$ as required, since s has the same sign as $b'(x)$.

Similarly, for $x < 0$, we have by Dan's favorite inequality,

$$\frac{-\sinh(x) - r_0\sinh(x-\delta)}{\cosh(x) + r_0\cosh(x-\delta)} \geq \min\{-\tanh(x), -\tanh(x-\delta)\} = -\tanh(x).$$

Therefore, for $x < 0$ we have

$$s \geq \coth(x) - \frac{1}{x} - \tanh(x) = \frac{1}{-x} - \frac{2}{\sinh(-2x)} > 0,$$

which shows that $b(x)$ is increasing for $x < 0$, concluding the proof. ■

The following Lemma proves the intuitive fact that decreasing and increasing functions of the same random variable are negatively correlated.

Lemma 14 *Let ρ be a real-valued random variable, let f,g be functions from $\mathbb{R}$ to $\mathbb{R}$ and let $\mathcal{S} \subset \mathbb{R}$ be an interval. If $f(x)$ is non-decreasing in x for $x \in \mathcal{S}$ and $g(x)$ is non-increasing in x for $x \in \mathcal{S}$, then*

$$\mathbb{E} f(\rho) g(\rho) \mathbb{I}_{\{\rho\in\mathcal{S}\}} \leq (\mathbb{E}[f(\rho) \,|\, \rho \in \mathcal{S}]) \cdot \left(\mathbb{E} g(\rho) \mathbb{I}_{\{\rho\in\mathcal{S}\}}\right).$$

Proof For every $x,x' \in \mathcal{S}$ we have $(f(x)-f(x'))\cdot(g(x)-g(x')) \le 0$. Hence, for every $x,x' \in \mathbb{R}$, the bound $(f(x)-f(x'))\cdot(g(x)-g(x'))\cdot \mathbb{I}_{\{x\in\mathcal{S}\}}\mathbb{I}_{\{x'\in\mathcal{S}\}} \le 0$ holds as well. Let ρ' be an independent copy of ρ, then

$$\mathbb{E}\big[(f(\rho)-f(\rho'))\cdot(g(\rho)-g(\rho'))\cdot\mathbb{I}_{\{\rho\in\mathcal{S}\}}\mathbb{I}_{\{\rho'\in\mathcal{S}\}}\big] \le 0.$$

Rearranging and using the fact that ρ,ρ' are i.i.d., we have

$$\big(\mathbb{E}f(\rho)g(\rho)\mathbb{I}_{\{\rho\in\mathcal{S}\}}\big)\cdot\big(\mathbb{E}\mathbb{I}_{\{\rho'\in\mathcal{S}\}}\big) \le \big(\mathbb{E}\big[f(\rho)\mathbb{I}_{\{\rho\in\mathcal{S}\}}\big]\big)\cdot\big(\mathbb{E}\big[g(\rho')\mathbb{I}_{\{\rho'\in\mathcal{S}\}}\big]\big).$$

Dividing by $\mathbb{E}\mathbb{I}_{\{\rho'\in\mathcal{S}\}} = \mathbb{P}(\rho\in\mathcal{S})$ yields the desired bound. ■

C.3. Proof of Proposition 1, parts 1 and 1′

Proof We begin with Proposition 1.1: *for every $Y,D\in S_n$, $\bar{V}_Y(Y+D) \le \frac{3}{2}\|D\|_\infty^2$.* To show this, fix $Y,D\in S_n$ and let $p(t) \coloneqq \bar{\mathsf{p}}(Y+tD)$. The Bregman divergence (6) admits the integral form

$$\begin{aligned}\bar{V}_Y(Y+D) &= p(1)-p(0)-p'(0) = \int_0^1 (p'(t)-p'(0))dt = \int_0^1\int_0^t p''(\tau)d\tau dt \\ &= \int_0^1\int_0^t \nabla^2\bar{\mathsf{p}}(Y+\tau D)[D,D]d\tau dt. \end{aligned} \tag{36}$$

Note that since $\bar{\mathsf{P}}(Y)\in\Delta_n$ for every $Y\in S_n$, $\langle D^2,\bar{\mathsf{P}}(Y)\rangle \le \|D^2\|_\infty\|\bar{\mathsf{P}}(Y)\|_1 = \|D\|_\infty^2$. Therefore, the bound (16) gives

$$\nabla^2\bar{\mathsf{p}}(Y+\tau D)[D,D] \le 3\|D\|_\infty^2.$$

Substituting back into (36) and using $\int_0^1\int_0^t d\tau dt = \frac{1}{2}$ gives Proposition 1.1.

Next, we show Proposition 1.1′: *for every $Y,D\in S_n$ such that $D\succeq 0$ and $\|D\|_\infty \le \frac{1}{6}$, we have $\bar{V}_Y(Y+D) \le 3\|D\|_\infty\langle D,\bar{\mathsf{P}}(Y)\rangle$.* When $D\succeq 0$, we have

$$\langle D^2,\bar{\mathsf{P}}(Y)\rangle = \left\langle D,D^{1/2}\bar{\mathsf{P}}(Y)D^{1/2}\right\rangle \le \|D\|_\infty\|D^{1/2}\bar{\mathsf{P}}(Y)D^{1/2}\|_1 = \|D\|_\infty\langle D,\nabla\bar{\mathsf{p}}(Y)\rangle.$$

Plugging the bound above into the bound (16) and substituting back into (36) gives

$$\bar{V}_Y(Y+D) \le 3\|D\|_\infty\int_0^1\int_0^t\langle D,\nabla\bar{\mathsf{p}}(Y+\tau D)\rangle d\tau dt. \tag{37}$$

Moreover,

$$\int_0^t\langle D,\nabla\bar{\mathsf{p}}(Y+\tau D)\rangle d\tau = \int_0^t p'(\tau)d\tau = p(t)-p(0) = \bar{V}_Y(Y+tD)+\langle tD,\bar{\mathsf{P}}(Y)\rangle, \tag{38}$$

where the final equality uses the definition (6) of the Bregman divergence. Note also that $v(t) \coloneqq \bar{V}_Y(Y+tD)$ is increasing for $t\ge 0$ due to convexity of $\bar{\mathsf{p}}$; $tv'(t) = \langle tD,\nabla\bar{\mathsf{p}}(Y+tD)-\nabla\bar{\mathsf{p}}(Y)\rangle \ge 0$. Therefore, the equality (38) implies $\int_0^t\langle D,\nabla\bar{\mathsf{p}}(Y+\tau D)\rangle d\tau \le \bar{V}_Y(Y+D)+t\cdot\langle D,\bar{\mathsf{P}}(Y)\rangle$ for every $0\le t\le 1$. Substituting this back into (37) and rearranging gives

$$\Big(1-3\|D\|_\infty\Big)\bar{V}_Y(Y+D) \le \frac{3}{2}\|D\|_\infty\langle D,\bar{\mathsf{P}}(Y)\rangle.$$

establishing part 1′ of the proposition, as $1-3\|D\|_\infty \ge \frac{1}{2}$ by assumption. ■

C.4. Facts about the Beta distribution

Here we collect properties of Beta-distributed random variables, which we use in our development.

Lemma 15 *Let $n \in \mathbb{N}$ and let $z \sim \mathrm{Beta}(\frac{1}{2}, \frac{n-1}{2})$. Then*

$$\mathbb{E}\log\tfrac{1}{z} = \psi\left(\tfrac{n}{2}\right) - \psi\left(\tfrac{1}{2}\right) \le \log(n) + \log(2) + \gamma \le \log(4n),$$

where $\psi(x) = \frac{d}{dx}\log\Gamma(x)$ is the digamma function, and γ is the Euler-Mascheroni constant.

Proof $\mathbb{E}\log\frac{1}{z} = \psi\left(\frac{n}{2}\right) - \psi\left(\frac{1}{2}\right)$ by the well-known formula for expectation of the logarithm of a Beta random variable. We have $\psi(x) \le \log(x)$ (Alzer, 1997) and $\psi(\frac{1}{2}) = -\log(4) - \gamma$. Moreover, $\gamma \le \log 2$, giving the final bound. ■

Lemma 16 *Let $z \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$ and $\ell \ge 0$. Then*

$$\frac{2}{\pi}\frac{e^{-\ell/2}}{\sqrt{1+e^{-\ell}}} \le \mathbb{P}\left(\log\frac{1-z}{z} \ge \ell\right) \le \frac{2}{\pi}e^{-\ell/2}.$$

Proof The distribution $\mathrm{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$ has density $\frac{1}{\pi}x^{-1/2}(1-x)^{-1/2}$. Therefore

$$\mathbb{P}\left(\log\frac{1-z}{z} \ge \ell\right) = \mathbb{P}\left(z \le \frac{1}{1+e^{\ell}}\right) = \frac{1}{\pi}\int_0^{\left(1+e^{\ell}\right)^{-1}} x^{-1/2}(1-x)^{-1/2}dx.$$

To obtain a lower bound, we use $(1-x)^{-1/2} \ge 1$ for every $x \in [0,1]$, and therefore,

$$\mathbb{P}\left(\log\frac{1-z}{z} \ge \ell\right) \ge \frac{1}{\pi}\int_0^{\left(1+e^{\ell}\right)^{-1}} x^{-1/2}dx = \frac{2}{\pi\sqrt{1+e^{\ell}}} = \frac{2}{\pi}\frac{e^{-\ell/2}}{\sqrt{1+e^{-\ell}}}.$$

For the upper bound, we use $(1-x)^{-1/2} \le \left(1 - \frac{1}{1+e^{\ell}}\right)^{-1/2}$ for every $0 \le x \le (1+e^{\ell})^{-1}$, giving

$$\mathbb{P}\left(\log\frac{1-z}{z} \ge \ell\right) \le \frac{1}{\pi}\sqrt{\frac{1+e^{\ell}}{e^{\ell}}}\int_0^{\left(1+e^{\ell}\right)^{-1}} x^{-1/2}dx = \frac{2}{\pi}e^{-\ell/2}.$$

■

Lemma 17 *Let $z \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$ and $\ell \ge 0$. Then*

$$\mathbb{E}\left[\log\frac{1-z}{z} \,\middle|\, \log\frac{1-z}{z} \ge \ell\right] \le \ell + 2\sqrt{1+e^{-\ell}}.$$

Proof Conditional on $\log\frac{1-z}{z} \ge \ell$, $\log\frac{1-z}{z}$ is a nonnegative random variable, and we may therefore write

$$\begin{aligned}\mathbb{E}\left[\log\frac{1-z}{z} \,\middle|\, \log\frac{1-z}{z} \ge \ell\right] &= \int_{x=0}^{\infty}\mathbb{P}\left(\log\frac{1-z}{z} \ge x \,\middle|\, \log\frac{1-z}{z} \ge \ell\right)dx \\ &= \ell + \int_{x=\ell}^{\infty}\frac{\mathbb{P}\left(\log\frac{1-z}{z} \ge x\right)}{\mathbb{P}\left(\log\frac{1-z}{z} \ge \ell\right)}dx.\end{aligned}$$

By Lemma 16,

$$\frac{\mathbb{P}\left(\log\frac{1-z}{z}\geq x\right)}{\mathbb{P}\left(\log\frac{1-z}{z}\geq \ell\right)}\leq\sqrt{1+e^{-\ell}}\cdot e^{-(x-\ell)/2}.$$

Integrating, we obtain the desired bound. ■

Lemma 18 *Let $3\leq n\in\mathbb{N}$ and let $z\sim\mathrm{Beta}(\frac{1}{2},\frac{n-1}{2})$. For every $\delta\in(0,1)$,*

$$\mathbb{P}\left(z\geq\frac{\delta^2}{n}\right)>1-\delta.$$

Proof The random variable z has density

$$\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n-1}{2})}x^{-1/2}(1-x)^{(n-3)/2}\leq\sqrt{\frac{n}{2\pi x}},$$

where we used $\Gamma(\frac{1}{2})=\sqrt{\pi}$ and Gautschi's inequality $\Gamma(m+1)/\Gamma(m+s)\leq(m+1)^{1-s}$ with $m=\frac{n}{2}-1$ and $s=\frac{1}{2}$. Integrating the upper bound on the density, we find $\mathbb{P}(z\leq\delta^2/n)\leq\sqrt{\frac{2}{\pi}}\delta<\delta$. ■

Appendix D. Efficient computation of matrix exponential-vector products

In this section we give a more detailed discussion of matrix exponential-vector product approximation using the Lanczos method, and prove the results stated in Section 3. In Section D.1 we formally state the Lanczos method. In Section D.2 we survey known approximation guarantees and derive simple corollaries. In Section D.3 we show that we can apply the matrix exponential to a random vector with a multiplicative error guarantee, and in Section D.4 we prove it implies Proposition 8. In Section D.5 we discuss some possible improvement to our guarantees via modifications and alternatives to the Lanczos method. Finally, in Section D.6 we prove Corollary 25.

Throughout this section we use $\mathrm{mv}(A)$ to denote the time required to multiply the matrix A with any vector.

D.1. Description of the Lanczos method

Ignoring numerical precision issues, each iteration in the for loop requires $O(\mathrm{mv}(A))$ time, and that for a k-by-k tridiagonal matrix, eigen-decomposition requires $O(k^2)$ time (Gu and Eisenstat, 1995), and so the total complexity is $O(\mathrm{mv}(A)k+k^2)$. In practical settings $k\ll n\leq\mathrm{mv}(A)$ and the cost of the eigen-decomposition is negligible. Nevertheless, there are ways to avoid performing it, which we discuss briefly in Section D.5.

D.2. Known approximation results, and some corollaries

We begin with a result on uniform polynomial approximation of the exponential due to Sachdeva and Vishnoi (2014).

Theorem 19 (Sachdeva and Vishnoi (2014), Theorem 4.1 Restated) *For every $b>0$ and every $\epsilon\in(0,1]$ there exists polynomial $p:\mathbb{R}\to\mathbb{R}$ of degree $O(\sqrt{\max\{b,\log(1/\epsilon)\}\log(1/\epsilon)})$ such that*

$$\sup_{x\in[0,b]}|\exp(-x)-p(x)|\leq\epsilon.$$

Algorithm 2: Lanczos method for computing matrix exponential vector product $\widetilde{\exp}_k(A,b)$

input : $A \in S_n$, number of iterations k, vector $b \in \mathbb{R}^n$

$q_0 \leftarrow 0 \in \mathbb{R}^n$, $q_1 \leftarrow b/\|b\|_2$, $\beta_1 \leftarrow 1$

for $i = 1$ *to* k **do**

  $q_{i+1} \leftarrow Aq_i - \beta q_{i-1}$ and $\alpha_i \leftarrow q_{i+1}^T q_i$

  $q_{i+1} \leftarrow q_{i+1} - \alpha q_i$ and $\beta_{i+1} = \|q_{i+1}\|_2$

  if $\beta_{i+1} = 0$ **then break else** $q_{i+1} \leftarrow q_{i+1}/\beta_{i+1}$

end

Let

$$Q = [q_1 \cdots q_k] \quad \text{and} \quad T = \begin{bmatrix} \alpha_1 & \beta_2 & & 0 \\ \beta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_k \\ 0 & & \beta_k & \alpha_k \end{bmatrix}$$

Compute tridiagonal eigen-decomposition $T = V \Lambda V^T$

return: $\widetilde{\exp}_k(A,b) = \|b\|_2 \cdot QV \exp(\Lambda) V^T e_1$

As an immediate corollary of this we obtain the following bounds for approximating $\exp(x)$ over arbitrary values

Corollary 20 *For every $a < b \in \mathbb{R}$ and every $\epsilon \in (0,1]$ there exists polynomial $p : \mathbb{R} \to \mathbb{R}$ of degree $O(\sqrt{\max\{b-a, \log(1/\epsilon)\}\log(1/\epsilon)})$ polynomial such that*

$$\sup_{x \in [a,b]} |\exp(x) - p(x)| \leq \epsilon \exp(b).$$

Proof For all $x \in [a,b]$ we have $b - x \in [0, b-a]$ and therefore by Theorem 19 there is a degree $O(\sqrt{\max\{b-a, \log(1/\epsilon)\}\log(1/\epsilon)})$ polynomial $q : \mathbb{R} \to \mathbb{R}$ such that

$$\sup_{x \in [a,b]} |\exp(-(b-x)) - q(b-x)| \leq \epsilon.$$

Since $\exp(-(b-x)) = \exp(-b)\exp(x)$, the polynomial $p(x) = \exp(b) q(b-x)$ is as desired. ■

The classical theory on the Lanczos method tells us that its error is bounded by twice that of any uniform polynomial approximation. However, this theory does not account for finite precision. A recent result (Musco et al., 2017) ties polynomial approximation to the error of the Lanczos method using finite bitwidth floating point operations.

Theorem 21 (Musco et al. (2017), Theorem 1) *Let $A \in S_n$, $u \in \mathbb{R}^n$, and $f : \mathbb{R} \to \mathbb{R}$. Suppose $k \in \mathbb{N}$, $\eta \in (0, \|A\|_\infty]$ and a polynomial p for degree $< k$ satisfy,*

$$\sup_{x \in [\lambda_{\min}(A) - \eta, \lambda_{\max}(A) + \eta]} |f(x) - p(x)| \leq \epsilon_k \quad \textit{and} \quad \sup_{x \in [\lambda_{\min}(A) - \eta, \lambda_{\max}(A) + \eta]} |f(x)| \leq C.$$

For any $\mu \in (0,1)$, let $y_{k,\mu}$ be the output of k iterations of the Lanczos method for approximating $f(A)v$, using floating point operations with $B \geq c\log(\frac{nk\|A\|_\infty}{\mu\eta})$ bits precision (for numerical constant $c<\infty$). Then $y_{k,\mu}$ satisfies

$$\|f(A)u-y_{k,\mu}\|_2 \leq (7k\cdot\epsilon_k+\mu\cdot C)\|u\|_2.$$

If arithmetic operations with B bits of precision can be performed in $O(1)$ time then the method can be implemented in time $O(\mathrm{mv}(A)k+kB\max\{k,B\})$.

Specializing to the matrix exponential and using the uniform approximation guarantee of Corollary 20, we immediately obtain the following.

Corollary 22 *Let $A \in S_n$, $u \in \mathbb{R}^n$, and $\epsilon > 0$, and set $M = \max\{\|A\|_\infty, \log(1/\epsilon), 1\}$. There exists numerical constants $c, c' < \infty$ such that, for $k \geq c\sqrt{M\log(M/\epsilon)}$ and $B \geq c'\log(\frac{nM}{\epsilon})$, computing $y = \widetilde{\exp}_k(A,u)$ with B bits of floating point precision guarantees*

$$\|\exp(A)u-y\|_2 \leq \epsilon\exp(\lambda_{\max}(A))\|u\|_2.$$

The computation takes time

$$O\left(\mathrm{mv}(A)\sqrt{M\log(M/\epsilon)}+M\log^2(nM/\epsilon)\right)$$

provided $\Theta(\log(\frac{nM}{\epsilon}))$ bit arithmetic operations can be performed in time $O(1)$.

Proof Let $\eta = 1$. Using $\lambda_{\max}(A)-\lambda_{\min}(A) \leq 2\|A\|$, Corollary 20 yields that for all $\alpha \in (0,1]$ there exists a degree $O\left(\sqrt{\max\{1+\|A\|_\infty, \log(\frac{1}{\alpha})\}\log(\frac{1}{\alpha})}\right)$ polynomial $p : \mathbb{R} \to \mathbb{R}$ such that

$$\sup_{x \in [\lambda_{\min}(A)-\eta, \lambda_{\max}(A)+\eta]} |\exp(x)-p(x)| \leq \alpha\exp(\eta)\exp(\lambda_{\max}(A)) \ .$$

Further, since $|\exp(x)| \leq \exp(\eta)\exp(\lambda_{\max}(A))$ for all $x \in [\lambda_{\min}(A)-\eta, \lambda_{\max}(A)+\eta]$, Theorem 21 with $f(x) = e^x$ and $\eta = 1$ implies that for all $\mu \in (0,1)$, after applying Lanczos for $k = O(\sqrt{\max\{\|A\|_\infty, \log(1/\alpha)\}\log(1/\alpha)})$ iterations on a floating point machine with $\Theta(B)$ bits of precision for $B = \log(\frac{nk\|A\|}{\mu})$ returns y with

$$\|f(A)u-y\|_2 \leq \left(\mu+\alpha\cdot O(\sqrt{\max\{\|A\|_\infty, \log(1/\alpha)\}\log(1/\alpha)})\right)\exp(\eta)\exp(\lambda_{\max}(A)))$$

in time $O((\mathrm{mv}(A)+n)k + kB\max\{k,B\})$. Choosing, $\alpha = O(\epsilon/(M\log(M/\epsilon)))$ and $\mu = O(\epsilon)$ yields the result. ■

D.3. Multiplicative approximation for random vectors

We now combine the known results cited in the previous section with the randomness of the vector fed to the matrix exponential, to obtain a multiplicative guarantee that holds with high-probability over the choice of u, but not for all $u \in \mathbb{S}^{n-1}$.

Proposition 23 *Let $\epsilon \in (0,1)$, $\delta \in (0,1)$, and $A \in S_n$. If u is sampled uniformly at random from the unit sphere and for $k = \Omega(\sqrt{M\log(nM/(\epsilon\delta))} \in \mathbb{N}$ for $M = \max\{\|A\|_\infty, \log(n/(\epsilon\delta)), 1\}$ we let $y = \widetilde{\exp}_k(A,u)$ (See Algorithm 2) then*

$$\|\exp(A)u - y\|_2 \le \epsilon \|\exp(A)u\|_2 \text{ with probability } \ge 1-\delta.$$

This can be implemented in time $O\left(\mathrm{mv}(A)\sqrt{M\log(nM/(\epsilon\delta)} + M\log^2(nM/(\epsilon\delta))\right)$ on a floating point machine with $O(\log(nM/(\epsilon\delta)))$ bits of precision where arithmetic operations take $O(1)$ time.

Proof Consider an application of Corollary 22 to compute y such that

$$\|\exp(A)u - y\|_2 \le \epsilon' \exp(\lambda_{\max}(A)) \|u\|_2 .$$

Now let v be a unit eigenvector of A with eigenvalue $\lambda_{\max}(A)$. Since v is an eigenvector or the PSD matrix $\exp(A)$ with eigenvalue $\exp(\lambda_{\max}(A))$ we have that $\|\exp(A)u\| \ge \exp(\lambda_{\max}) |v^T u|$. However, since u is a random unit vector we have that $|v^T u|^2 / \|u\|_2^2 \sim \mathrm{Beta}(\frac{1}{2}, \frac{n-1}{2})$. Lemma 18 therefore gives that $|v^T u|^2 / \|u\|_2^2 \ge \frac{\delta^2}{n}$ with probability at least $1-\delta$. Consequently, $\exp(\lambda_{\max}(A)) \|u\|_2 \le \frac{\sqrt{n}}{\delta} \|\exp(A)u\|_2$ with the same probability. Choosing $\epsilon' = \epsilon\delta/\sqrt{n}$ and invoking Corollary 22 yields the result. ■

D.4. Proof of Proposition 8

The following lemma relates the multiplicative approximation error for matrix exponential vector products with the additive approximation error for $\mathsf{P}_u(Y)$ under trace norm. Combining it with Proposition 23 immediately yields Proposition 8.

Lemma 24 *Let $Y \in S_n$, $u, y \in \mathbb{R}^n$ and $\epsilon \in [0,1)$. If $y \in \mathbb{R}^n$ satisfies*

$$\|\exp(Y/2)u - y\|_2 \le \frac{\epsilon}{\sqrt{8}} \|\exp(Y/2)u\|_2$$

then

$$\left\| \mathsf{P}_u(Y) - \frac{yy^T}{\|y\|_2^2} \right\|_1 \le \epsilon .$$

Proof Let $z := \exp(Y/2)u$ so that by assumption $\|z-y\|_2 \le \epsilon \|z\|_2$. Further, let $\bar{z} := z/\|z\|_2$ and $\bar{y} := y/\|y\|_2$. Direct calculation (see e.g. Lemma 27 of Cohen et al. (2016)) yields that the eigenvalues of $\bar{z}\bar{z}^T - \bar{y}\bar{y}^T$ are $\pm\sqrt{1-(\bar{z}^T\bar{y})^2} = \pm\frac{1}{2}\|\bar{z}+\bar{y}\|_2 \|\bar{z}-\bar{y}\|_2$ and therefore the definition of $\mathsf{P}_u(Y)$ yields

$$\left\| \mathsf{P}_u(Y) - \frac{yy^T}{\|y\|_2^2} \right\|_1 = \left\| \bar{z}\bar{z}^T - \bar{y}\bar{y}^T \right\|_1 = \|\bar{z}+\bar{y}\|_2 \cdot \|\bar{z}-\bar{y}\|_2 \le \sqrt{2}\|\bar{z}-\bar{y}\|_2, \tag{39}$$

where in the last inequality we used that $\bar{z}$ and $\bar{y}$ are unit vectors. Further, by the triangle inequality and the definitions of $\bar{y}$ and $\bar{z}$ we have

$$\begin{aligned} \|\bar{z}-\bar{y}\|_2 &\le \left\| \frac{z}{\|z\|_2} - \frac{y}{\|z\|_2} \right\|_2 + \left\| \frac{y}{\|z\|_2} - \frac{y}{\|y\|_2} \right\|_2 \\ &= \frac{\|z-y\|_2}{\|z\|_2} + \frac{|\|y\|_2 - \|z\|_2|}{\|z\|_2} \le 2\frac{\|z-y\|_2}{\|z\|_2} \end{aligned} \tag{40}$$

Combining (39) and (40) with the fact that $\|z-y\|_2 \le (\epsilon/\sqrt{8})\|z\|_2$ then yields

$$\left\| \mathsf{P}_u(Y) - \frac{yy^T}{\|y\|_2^2} \right\|_1 \le \sqrt{2}\cdot 2\cdot(\epsilon/\sqrt{8}) = \epsilon.$$

■

Therefore, Proposition 8 follows immediately by invoking 23 with slightly smaller ϵ.

D.5. Improvements to the Lanczos method

In this paper we focused on the Lanczos method for approximating matrix exponential vector products because of its excellent practicality and clean analysis. However, there are several modifications to the method with appealing features, which we now describe briefly. A common theme among these modifications is the use of rational approximations to the exponential, which converge far faster than polynomial approximations (Orecchia et al., 2012; Sachdeva and Vishnoi, 2014). Consequently, it suffices to perform $\widetilde{O}(1)$ Lanczos iterations on a carefully shifted and inverted version of the matrix. Each of these iterations then involves solving a linear system, and the efficacy of the shift-invert scheme will depend on how quickly they are solved.

One basic approach to solving these systems is via standard iterative methods, e.g. conjugate gradient. We expect such approach to offer little to no advantage over applying the Lanczos approximation directly, as both methods produce vectors in the same Krylov subspace. However, the approach renders the number of Lanczos iterations k logarithmic in $\|A\|_\infty$, and therefore the cost k^2 will never dominate the cost of the matrix-vector products (Orecchia et al., 2012; Musco et al., 2017, Corollary 17).

There is, however, a simpler way of avoiding the eigen-decomposition—simply use the rational approximation on the tridiagonal matrix formed by running the ordinary Lanczos method, as Saad (1992) proposes. With an appropriate rational function, computing a highly accurate approximation to $\exp(T)e_1$ requires $\widetilde{O}(1)$ tridiagonal system solves, each costing $O(k)$ time. We leave the derivation of explicit error bounds for this technique (similar to Corollary 20) to future work. In practice, the cost $O(k^2)$ of tridiagonal eigen-decomposition will often be very small compared to the cost $O(\mathrm{mv}(A)k)$ of the matrix-vector products.

More significant improvements are possible if the linear system solving routine is able to exploit information beyond matrix-vector products. For example, consider the case where the matrix to be exponentiated is a sum of very sparse matrices—this will happen for our sketch whenever the G_t matrices are much sparser than their cumulative sum. Then, it is possible to use stochastic variance reduced optimization methods to solve the linear system, as Allen-Zhu and Li (2017) describe. Another scenario of interest is when the input matrix has a Laplacian/SDD structure and in this case the performance of specialized linear system solvers implies approximation guarantees where the polynomial dependence on $\|A\|_\infty$ is removed altogether (Orecchia et al., 2012). A final useful structure is a chordal sparsity pattern (Vandenberghe et al., 2015), which enables efficient linear system solving through fast Cholesky decomposition.

D.6. Proof of Corollary 25

Corollary 25 *Let $G_1,\ldots,G_T$ be symmetric gain matrices satisfying $\|G_t\|_\infty \le 1$ for every t. There exists a numerical constant $k_0 < \infty$, such that for every $T \in \mathbb{N}$ and $\delta \in (0,1)$, $\tilde{X}_{t;k_t}$ defined in (17)*

with $k_t = \left\lceil k_0(\sqrt{1+\eta t})\log(\frac{nT}{\delta})\right\rceil$, *and* X_t *defined in* (4) *satisfy*

$$\sum_{t=1}^{T}\left\langle G_t, \tilde{X}_{t;k_t}\right\rangle \geq -1+\sum_{t=1}^{T}\langle G_t, X_t\rangle \;\; \textit{w.p.} \geq 1-\delta/2. \tag{41}$$

Let $\epsilon \in (0,1]$, $T = \frac{16\log(4en/\delta)}{\epsilon^2}$ *and* $\eta = \sqrt{\frac{2\log(4en)}{3T}}$. *If Assumption A holds with respect to the actions* $\tilde{X}_{t;k_t}$, *then with probability at least* $1-\delta$, $\frac{1}{T}\lambda_{\max}\left(\sum_{i=1}^{T} G_t\right) - \frac{1}{T}\sum_{t=1}^{T}\left\langle G_t, \tilde{X}_{t;k_t}\right\rangle \leq \epsilon$. *Computing the actions* $\tilde{X}_{1;k_1},\ldots,\tilde{X}_{T;k_T}$ *requires* $O(\epsilon^{-2.5}\log^{2.5}(\frac{n}{\epsilon\delta}))$ *matrix-vector products.*

Proof To obtain the bound (41) we use Proposition 8 with $\epsilon \leftarrow \frac{1}{T}$ and $\delta \leftarrow \delta/(2T)$ (since we will use a union bound). At iteration t, $\|G_i\|_\infty \leq 1$ for all $i < t$, the quantity M appearing in Proposition 8 can be bounded as

$$M \leq \left(1+\left\|\frac{\eta}{2}\sum_{i=1}^{t-1} G_i\right\|_\infty\right)\log\frac{nT^2}{\delta} \leq O(1)(1+\eta t)\log\frac{nT}{\delta}.$$

Therefore, our choice of k_t suffices to guarantee, for $Y_t = \eta\sum_{i=1}^{t-1} G_i$,

$$\|\mathsf{P}_{u_t}(Y_t) - \widetilde{\mathsf{P}}_{u_t;k_t}(Y_t)\|_1 \leq \frac{1}{T} \text{ with probability } \geq 1-\frac{\delta}{2T},$$

and so by the union bound the inequality above holds for all $t = 1,\ldots,T$ with probability at least $1-(\delta/2)$. Note that when using Proposition 8 we use the fact that u_t is independent of Y_t. Thus, we have

$$\sum_{t=1}^{T}\left\langle G_t, X_t - \tilde{X}_{t;k_t}\right\rangle \leq \sum_{t=1}^{T}\|G_t\|_\infty\|X_t - \tilde{X}_{t;k_t}\|_1 \leq \sum_{t=1}^{T}\|\mathsf{P}_{u_t}(Y_t) - \widetilde{\mathsf{P}}_{u_t;k_t}(Y_t)\|_1 = 1,$$

giving (41), where we have used $\|G_t\|_\infty \leq 1$ for every t.

Note that if Assumption A holds with respect to the actions $\tilde{X}_{t;k_t}$ then we have $G_t \perp u_t \mid \mathcal{F}_{t-1}$ and therefore $\mathbb{E}[\langle G_t, X_t\rangle \mid \mathcal{F}_{t-1}] = \left\langle G_t, \bar{X}_t\right\rangle$ so that Corollary 4 holds. Thus, to obtain the second part of the corollary, we use the bound (9) with $\delta \leftarrow \delta/2$ and η and T as specified; using a union bound again we have that (41) and (9) hold together with probability at least $1-\delta$. Note that $\eta \leq \epsilon \leq 1$ and therefore $1/T \leq 1/(\eta T)$. This gives,

$$\begin{aligned}\frac{1}{T}\lambda_{\max}\left(\sum_{i=1}^{T} G_t\right) - \frac{1}{T}\sum_{t=1}^{T}\left\langle G_t, \tilde{X}_{t;k_t}\right\rangle &\leq \frac{1}{T} + \frac{3\eta}{2} + \frac{\log(4n)}{\eta T} + \sqrt{\frac{2\log\frac{2}{\delta}}{T}} \\ &\leq \frac{3\eta}{2} + \frac{\log(4en)}{\eta T} + \sqrt{\frac{2\log\frac{2}{\delta}}{T}} = \sqrt{\frac{6\log(4en)}{T}} + \sqrt{\frac{2\log\frac{2}{\delta}}{T}} \leq \epsilon,\end{aligned}$$

as required. Finally note that $1+\eta T = O(\epsilon^{-1}\log(\frac{n}{\delta}))$ and consequently

$$k_T = O\left(\epsilon^{-1/2}\log^{1/2}(\tfrac{n}{\delta})\log^{1/2}(\tfrac{nT}{\delta})\right) = O\left(\epsilon^{-1/2}\log^{1.5}(\tfrac{n}{\epsilon\delta})\right).$$

Since $k_1 \leq k_2 \leq \cdots k_T$, the total number of matrix-vector products is bounded by $T \cdot k_T = O(\epsilon^{-2.5}\log^{2.5}(\frac{n}{\epsilon\delta}))$, which concludes the proof. ■

Appendix E. Application to semidefinite programming

E.1. Proof of Theorem 9

Theorem 9 *Let $\{X_t,y_t\}_{t=1}^T$ be the actions produced by Algorithm 1 and, define $X_T^{\text{avg}}=\frac{1}{T}\sum_{t=1}^T X_t$, $y_T^{\text{avg}}=\frac{1}{T}\sum_{t=1}^T y_t$. Then*

$$\mathbb{E}\left[\text{Gap}\left(X_T^{\text{avg}},y_T^{\text{avg}}\right)\right]\le\frac{\log(4mn)}{\eta T}+2\eta\omega^2.$$

Proof Recalling the definition (19) of the duality gap, and that $G_t=\mathcal{A}^\star y_t$ and $[c_t]_i=\langle A_i,X_t\rangle$, we have

$$\text{Gap}(X_T^{\text{avg}},y_T^{\text{avg}})=\frac{1}{T}\lambda_{\max}\left(\sum_{t=1}^T G_t\right)-\frac{1}{T}\min_{i\in[m]}\left\{\sum_{t=1}^T[c_t]_i\right\}.$$

Note that $y_t=\nabla\text{lse}(-\eta\sum_{i=1}^{t-1}c_i)$ is a function of $X_1,\ldots,X_{t-1}$. Therefore, $G_t=\mathcal{A}^\star y_t$ satisfies Assumption A and we may use Corollary 3 to write

$$\mathbb{E}\left[\lambda_{\max}\left(\sum_{t=1}^T G_t\right)-\sum_{t=1}^T\langle G_t,X_t\rangle\right]\le\frac{\log(4n)}{\eta}+\frac{3\eta}{2}\cdot\sum_{t=1}^T\mathbb{E}\left[\|G_t\|_\infty^2\right]\le\frac{\log(4n)}{\eta}+\frac{3\eta\omega^2T}{2},$$

where in the second inequality we used $\omega=\max_{i\in[m]}\|A_i\|_\infty$ and $y\in\sigma_m$ to bound $\|G_t\|_\infty=\|\mathcal{A}^\star y_t\|_\infty\le\omega\cdot\mathbf{1}^Ty_t=\omega$. Similarly, we use the standard multiplicative weights regret bound (cf. Shalev-Shwartz, 2012, Theorem 2.21) to write

$$\sum_{t=1}^T c_t^Ty_t-\min_{i\in[m]}\left\{\sum_{t=1}^T[c_t]_i\right\}\le\frac{\log(m)}{\eta}+\frac{\eta}{2}\cdot\sum_{t=1}^T\|c_t\|_\infty^2\le\frac{\log(m)}{\eta}+\frac{\eta\omega^2T}{2},$$

where the second inequality again follows from $|[c_t]_i|=|\langle A_i,X_t\rangle|\le\|A_i\|_\infty\le\omega$ since $X_t\in\Delta_n$.

Finally,

$$c_t^Ty_t=\sum_{i=1}^m[y_t]_i\langle A_i,X_t\rangle=\langle\mathcal{A}^\star y_t,X_t\rangle=\langle G_t,X_t\rangle.$$

Hence, summing the two regret bounds and dividing by T gives the result. ■

E.2. Derivation of total computational cost bound

Recall that for $\eta=\log(4mn)/\sqrt{2\omega^2T}$ and $T=8\log(4mn)\omega^2/\epsilon^2$, Theorem 9 guarantees $\mathbb{E}\left[\text{Gap}(X_T^{\text{avg}},y_T^{\text{avg}})\right]\le\epsilon$. Let $\text{mv}(M)$ denote the time required to multiply the matrix M by any vector, and let $\text{mv}(\mathcal{A}):=\sum_{i\in[m]}\text{mv}(A_i)$. Except for the computation of X_t, every step in the for loop in Algorithm 1 takes $O(\text{mv}(\mathcal{A}))$ work to execute (we may assume $\text{mv}(\mathcal{A})\ge\max\{n,m\}$ without loss of generality). Let $Y_t=\eta\sum_{i=1}^{t-1}G_i=\mathcal{A}^\star(\sum_{i=1}^{t-1}\eta y_i)$, and note that, with the values of η and T above, $\|Y_t\|_\infty\le\eta T\omega=\widetilde{O}(\omega/\epsilon)$ for every $t\le T$. Per Section 3, the computation of X_t costs $\widetilde{O}(\|Y_t\|_\infty^{0.5}\text{mv}(Y_t))=\widetilde{O}((\omega/\epsilon)^{0.5}\text{mv}(Y_t))$. Writing $\text{mv}(\mathcal{A}^\star):=\max_{\alpha\in\mathbb{R}^m}\{\text{mv}(\mathcal{A}^\star\alpha)\}\le\min\{\text{mv}(\mathcal{A}),n^2\}$, the total computational cost of our algorithm is

$$\widetilde{O}(\left[(\omega/\epsilon)^{0.5}\text{mv}(\mathcal{A}^\star)+\text{mv}(\mathcal{A})\right]T)=\widetilde{O}((\omega/\epsilon)^{2.5}\text{mv}(\mathcal{A}^\star)+(\omega/\epsilon)^2\text{mv}(\mathcal{A})).$$

In many settings of interest—namely when the A_is have mostly non-overlapping sparsity patterns and yet the Y_ts are sparse—we have $\text{mv}(\mathcal{A}^\star)\approx\text{mv}(\mathcal{A})$, so that the computational cost is dominated by the first term.

E.3. Comparison with other algorithms

Let $\mathrm{nnz}(M)$ denote the number of nonzero entries of matrix M, and let $\mathrm{nnz}(\mathcal{A}) := \sum_{i\in[m]} \mathrm{nnz}(A_i) \geq \mathrm{mv}(\mathcal{A})$. If in Algorithm 1 we replace the randomized projection P_u with the matrix multiplicative weights projection P^{mw}, the regret bound of Theorem 9 still holds, but the overall computational cost becomes $\widetilde{O}((\omega/\epsilon)^2(n^3+\mathrm{nnz}(\mathcal{A})))$ due to full matrix exponentiation. Nemirovski (2004) accelerates this scheme using extra-gradient steps, guaranteeing duality gap below ϵ in $\widetilde{O}(\omega/\epsilon)$ iterations, with each iteration involving two full matrix exponential computations. The overall computational cost of such scheme is consequently $\widetilde{O}((\omega/\epsilon)(n^3+\mathrm{nnz}(\mathcal{A})))$. Nesterov (2007) attains the same rate by using accelerated gradient descent on a smoothed version of the dual problem. Our scheme improves on this rate for sufficiently sparse problems, with $n^3/\mathrm{nnz}(\mathcal{A}) \gg (\omega/\epsilon)^{-1.5}$.

d'Aspremont (2011) applies a subgradient method to the dual problem, approximating the subgradients using the Lanczos method to compute a leading eigenvector of $\mathcal{A}^\star y$. The method solves the dual problem to accuracy ϵ with total work $\widetilde{O}((\omega/\epsilon)^{2.5}\mathrm{mv}(\mathcal{A}^\star)+(\omega/\epsilon)^2\mathrm{mv}(\mathcal{A}))$, essentially the same as us. However, it is not clear how to efficiently recover a primal solution from this method. Moreover, the surrogate duality gap d'Aspremont (2011) proposes will not always be 0 at the global optimum, whereas with our approach the true duality gap is readily computable.

Baes et al. (2013) replace the full matrix exponentiation in the accelerated scheme of Nemirovski (2004) with a rank-k sketch of the form (3), where $k = \widetilde{O}(\omega/\epsilon)$. Similarly to Nemirovski (2004), they require $\widetilde{O}(\omega/\epsilon)$ iterations to attain duality gap below ϵ. Baes et al. (2013) approximate matrix exponential vector products by truncating a Taylor series, costing $\widetilde{O}(k(\omega/\epsilon)\,\mathrm{mv}(\mathcal{A}^\star)) = \widetilde{O}((\omega/\epsilon)^2\,\mathrm{mv}(\mathcal{A}^\star))$ work per iteration. With the Lanczos method, the cost improves to $\widetilde{O}((\omega/\epsilon)^{1.5}\mathrm{mv}(\mathcal{A}^\star))$ work per iteration. Every step of their method also computes $\langle A_i, X\rangle$ for all $i\in[m]$ and a rank-k matrix $X = \sum_{j=1}^k v_j v_j^T$; this costs either $k\cdot\mathrm{mv}(\mathcal{A})$ work (computing $\langle A_i, v_j\rangle$ for every i,j) or $\mathrm{nnz}(\mathcal{A})+n^2k$ (when forming X explicitly). The former option yields total complexity identical to our method. The latter option is preferable only when $\mathrm{nnz}(\mathcal{A}) \gg n^2 \geq \mathrm{mv}(\mathcal{A}^\star)$, and can result in an improvement over the running time of our method if $\mathrm{mv}(\mathcal{A}^\star) \ll \mathrm{nnz}(\mathcal{A})(\omega/\epsilon)^{-1.5} + n^2(\omega/\epsilon)^{-0.5}$. Baes et al. (2013) report that $k=1$ often gave the best result in their experiment, which is not predicted by their theory. A hypothetical explanation for this finding is that, with $k=1$, they are essentially running Algorithm 1.

Finally, d'Aspremont (2011) and Garber and Hazan (2016) propose sub-sampling based algorithms for approximate SDP feasibility with runtimes potentially sublinear in $\mathrm{mv}(\mathcal{A}^\star)$. However, because of their significantly worse dependence on ω/ϵ, as well as dependence on Frobenius norms, we match or improve upon their runtime guarantees in a variety of settings; see (Garber and Hazan, 2016) for a detailed comparison.