A randomized Mirror-Prox method for solving structured large-scale matrix saddle-point problems

Michel Baes*, Michael Bürgisser†, Arkadi Nemirovski‡

December 6, 2011

Abstract

In this paper, we derive a randomized version of the Mirror-Prox method for solving some structured matrix saddle-point problems, such as the maximal eigenvalue minimization problem. Deterministic first-order schemes, such as Nesterov's Smoothing Techniques or standard Mirror-Prox methods, require the exact computation of a matrix exponential at every iteration, limiting the size of the problems they can solve. Our method allows us to use stochastic approximations of matrix exponentials. We prove that our randomized scheme decreases significantly the complexity of its deterministic counterpart for large-scale matrix saddle-point problems. Numerical experiments illustrate and confirm our theoretical results.

Keywords: stochastic approximation, Mirror-Prox methods, matrix saddle-point problems, eigenvalue optimization, large-scale problems, matrix exponentiation

1 Introduction

Large-scale semidefinite programming attracts substantial research efforts nowadays. A vast set of applications can be modeled as such optimization problems, and many strategies have been studied theoretically and implemented in excellent softwares.

Arguably, general purpose semidefinite methods suffer from an intrinsic drawback. They forbid themselves, for the sake of generality, to exploit explicitly some structural features of the particular instance they are given to solve, hampering the resolution of very large-scale problems.

As a result, we are witnessing the development of special-purpose algorithms, designed for particular subclasses of semidefinite optimization problems, where the utilization of their specific structure is instrumental for aiming at large size problems. In this paper, we are addressing the problem of minimizing the maximal eigenvalue of an affine combination of given symmetric matrices, plus a linear function of the coefficients of this affine combination. Different strategies have

*Institute for Operations Research, ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland, michel.baes@ifor.math.ethz.ch.

†Institute for Operations Research, ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland, michael.buergisser@ifor.math.ethz.ch.

‡Georgia Institute of Technology, Atlanta, Georgia 30332, USA, nemirovs@isye.gatech.edu.

been devised to deal specifically with the maximal eigenvalue minimization problem. Among the first investigated techniques, Bundle methods were introduced in [HR00, Ous00], and subsequently refined in a number of further papers. Theoretical results on Bundle methods concern mainly asymptotic convergence properties; to the best of our knowledge, no complexity guarantees have been obtained so far in this direction.

Another strategy has been discovered in [Nes07] when Nesterov showed how his Smoothing Techniques can be specialized to the maximal eigenvalue minimization problem. As a result, he obtained worst-case complexity guarantees for his method: if $\epsilon > 0$ is the desired absolute accuracy on the objective value, $A_1, \ldots, A_m$ are real symmetric $n \times n$-matrices, $\Delta_m \subseteq \mathbb{R}^m$ is the $(m-1)$-dimensional simplex, and $\lambda_{\max}(A)$ denotes the maximal eigenvalue of any real symmetric matrix A, his algorithm solves the problem

$$\min \left\{ \lambda_{\max} \left(\sum_{j=1}^{m} A_j x_j \right) : \; x \in \Delta_m \right\} \tag{1}$$

in $\mathcal{O}((n^3 + mn^2) \max_j \lambda_{\max}(|A_j|) \sqrt{\ln(n) \ln(m)}/\epsilon)$ elementary operations, where $|A| = \sqrt{A^2}$ for any real symmetric matrix A. Almost simultaneously, the paper [Nem04a] develops a Mirror-Prox method, which can be particularized as well for our problem, and obtains equivalent complexity results. Two papers by Warmuth et al. [TRW05, WK06] present a scheme - called the Matrix Exponentiated Gradient Update method - that can basically be applied to problem (1). A form of this algorithm was independently discovered by Arora and Kale [AK07]. The methods of Arora et al. and Warmuth et al. essentially reduce to a subgradient method (provided that we adapt them to our problem), but with a worse complexity guarantee: in order to find a solution to problem (1) with absolute accuracy ϵ, these methods need $\mathcal{O}(\max_j \lambda^2_{\max}(|A_j|) \ln(n)/\epsilon^2)$ iterations. Each of these iterations requires the computation of a matrix exponential and further operations with a cost not exceeding $\mathcal{O}(mn^2)$.

In a nutshell, the methods introduced in [AK07, TRW05, WK06] present the same computational bottleneck as Smoothing Techniques and the Mirror-Prox method when applied to our problem: at every iteration, all these schemes require the determination of a symmetric matrix's exponential. Several efforts have been carried out to reduce the iteration computation cost. In [d'A08a], d'Aspremont analyzes the possibility of using approximate gradients, and thereby approximate matrix exponentials in Nesterov's Smoothing Techniques. In [JNT08], Mirror-Prox methods for variational problems where extended to situations where only some stochastic information is available from the instance to solve. These methods were particularized to the maximal eigenvalue minimization problem where all the input matrices share the same block-diagonal pattern. Albeit the problem is completely deterministic, an artificial randomization was introduced in the oracle of the method, which reduced the iteration cost while retaining some probabilistic guarantees on the output of the algorithm. Finally, Arora and Kale [AK07] obtain, by approximating the rows of $\exp(X/2)$, a substitute for the exact Gram matrix $\exp(X)$, where X is some real symmetric $n \times n$-matrix. The rows of $\exp(X/2)$ are approximated by projecting an appropriate truncation of the exponential Taylor series approximation on, roughly speaking, $\mathcal{O}(1/\epsilon^2)$ random directions.

In this paper, we apply the general results of [JNT08] to analyze another randomization strategy for computing matrix exponentials, which is also based on a vector sampling and on an appropriate truncation of the exponential Taylor series. Whereas we consider the same number of terms in the Taylor series approximation of the matrix exponential as Arora and Kale [AK07] do, we can

significantly reduce the number of required random vectors: roughly speaking, we project the truncated Taylor series on $\mathcal{O}(1/\epsilon)$ random directions.

The approximation strategy developed in this paper proves to be theoretically efficient for large-scale problems. In theory, it outperforms all its competitors on a reasonably large set of instances, described by the size of the input matrices, their number, their sparsity, and the desired accuracy. Our theoretical conclusions are demonstrated by numerical evidence: for solving problems (up to a relative accuracy of 0.2%) that involve a hundred matrices of size 800×800, the Mirror-Prox method equipped with our randomization procedure requires on average about, roughly speaking, half of the CPU time needed by the Mirror-Prox method with exact computations.

The paper is organized as follows. Section 2 contains the necessary notational conventions and a brief recall on existing results on Mirror-Prox methods for general convex problems with approximate oracle. We particularize these considerations in Section 3 to slightly structured matrix saddle-point problems and we analyze the stochastic exponential approximation strategy briefly described above for computing an approximate oracle. In Section 4, we derive the complexity of solving the maximal eigenvalue minimization problem. In Section 5, we test our method for solving large-scale eigenvalue optimization problems, comparing its efficiency with the provably best purely deterministic method in terms of worst-case complexity.

2 Mirror-Prox methods with approximate first-order information: a review

Let E be a Euclidean space with inner product $\langle \cdot, \cdot \rangle$. We endow E with a norm $\|\cdot\|$, which may differ from the one that is induced by this inner product. The *conjugate norm* $\|\cdot\|_*$ to $\|\cdot\|$ is defined as:

$$\|w\|_* := \max_{x \in E} \left[\langle w, x \rangle : \ \|x\| = 1 \right].$$

2.1 Variational inequalities and saddle-point problems

Variational inequalities. Let Q be a non-empty convex compact subset of E, and let $F : Q \to E$ be a Lipschitz continuous monotone mapping with Lipschitz constant $L > 0$:

$$\begin{array}{ll} \|F(z) - F(z')\|_* \leq L \|z - z'\| & \forall\ z, z' \in Q \\ \langle F(z) - F(z'), z - z' \rangle \geq 0 & \forall\ z, z' \in Q. \end{array}$$

The variational inequality associated with the set Q and the operator F reads as follows:

$$\text{find } z^* \in Q \text{ such that } \langle F(z), z^* - z \rangle \leq 0 \text{ for all } z \in Q. \tag{2}$$

In the sequel, in order to measure the inaccuracy of a point $\bar{z} \in Q$ as a candidate solution to (2), we use the *dual gap function*

$$\epsilon(\bar{z}) := \max_{z \in Q} \langle F(z), \bar{z} - z \rangle .$$

For $z \in Q$, we clearly have $\epsilon(z) \geq 0$, and $\epsilon(z) = 0$ if and only if z solves (2).

Saddle point problems. Assume that $E := E_1 \times E_2$ for Euclidean spaces E_1 and E_2, and that $Q := Q_1 \times Q_2$ is non-empty with two convex compact sets $Q_1 \subset E_1$ and $Q_2 \subset E_2$. Let $\phi : Q_1 \times Q_2 \to \mathbb{R}$ be a convex-concave function. We restrict ourselves to functions ϕ that are differentiable with Lipschitz continuous gradient. The function $\phi(\cdot,\cdot)$ is associated with the saddle-point problem

$$\min_{x\in Q_1} \max_{y\in Q_2} \phi(x,y). \tag{3}$$

Due to the standard Minimax Theorem in Convex Analysis (see Corollary 37.3.2 in [Roc70]), we have the following pair of primal-dual convex optimization problems:

$$\min_{x\in Q_1} \left[\overline{\phi}(x) := \max_{y\in Q_2} \phi(x,y)\right] = \max_{y\in Q_2} \left[\underline{\phi}(y) := \min_{x\in Q_1} \phi(x,y)\right].$$

It is well known that the solutions to the saddle point problem (3) are exactly the pairs (x_*, y_*) comprised of optimal solutions to the above two optimization problems, and that these pairs are exactly the solutions to the variational inequality given by $Q = Q_1 \times Q_2$ and the monotone operator

$$F(x,y) = \left(\frac{\partial \phi(x,y)}{\partial x}; -\frac{\partial \phi(x,y)}{\partial y}\right) : Q \to E_1 \times E_2.$$

We quantify the accuracy of a candidate solution $\bar{z} = (\bar{x}, \bar{y}) \in Q$ to the saddle point problem (3) by the value of the corresponding duality gap

$$\epsilon^{sad}(\bar{z}) := \overline{\phi}(\bar{x}) - \underline{\phi}(\bar{y}) = \max_{v\in Q_2} \phi(\bar{x}, v) - \min_{u\in Q_1} \phi(u, \bar{y}).$$

Due to the convex-concave structure of ϕ, any point $\bar{z} = (\bar{x}, \bar{y}) \in Q_1 \times Q_2$ constitutes an $\epsilon^{\rm sad}(\bar{z})$-approximate solution to the variational inequality that is associated with $Q = Q_1 \times Q_2$ and the above F:

$$\begin{aligned}
\epsilon^{sad}(\bar{x}, \bar{y}) &= \max_{(x,y)\in Q_1\times Q_2} [\phi(\bar{x}, y) - \phi(x, y) + \phi(x, y) - \phi(x, \bar{y})] \\
&\geq \max_{(x,y)\in Q_1\times Q_2} \left[\left\langle \frac{\partial \phi(x,y)}{\partial x}, \bar{x} - x \right\rangle + \left\langle -\frac{\partial \phi(x,y)}{\partial y}, \bar{y} - y \right\rangle\right] = \max_{z\in Q} \langle F(z), \bar{z} - z\rangle = \epsilon(\bar{x}, \bar{y}).
\end{aligned}$$

2.2 Mirror-Prox algorithm: preliminaries

In its basic form, the Mirror Prox (MP) algorithm is aimed at solving variational inequalities on a convex compact subset Q of a Euclidean space E equipped with a norm $\|\cdot\|$. The setup for the algorithm is given by a *distance-generating function* (d.-g.f.) ω: $Q \to \mathbb{R}$ which possesses the following properties:

- $\diamond$ ω is continuous and convex on Q. In particular, the domain $Q^o = \{x \in Q : \partial\omega(x) \neq \emptyset\}$ of the subdifferential of ω is nonempty.
- $\diamond$ ω is regular on Q^o, i.e., the subdifferential $\partial\omega(\cdot)$ admits a *continuous* selection $\omega'(\cdot)$ on Q^o.

◇ The function ω is strongly convex modulus 1 with respect to $\|\cdot\|$:

$$\langle \omega'(z) - \omega'(y), z - y \rangle \geq \|z - y\|^2 \qquad \forall\ y, z \in Q^o.$$

In the sequel, we refer to the latter property as the *compatibility* of the d.-g.f. $\omega(\cdot)$ and the norm $\|\cdot\|$.

Furthermore, we suppose that we choose ω such that we can easily solve problems of the form:

$$\min_{z \in Q} \left[\omega(z) - \langle e, z \rangle\right], \qquad e \in E. \tag{4}$$

Remark 2.1 *The optimal solution z_e to (4) clearly exists, is unique by continuity and strong convexity of ω, and belongs to Q^o (indeed, by evident reasons $e \in \partial\omega(z_e)$). From regularity of $\omega(\cdot)$, it immediately follows that*

$$\langle \omega'(z_e) - e, z - z_e \rangle \geq 0 \qquad \forall\ z \in Q. \tag{5}$$

A d.-g.f. $\omega(\cdot)$ gives rise to several entities:

◇ The *ω-center* $z^\omega := \arg\min_{z \in Q} \omega(z)$ of Q.

◇ The *Bregman distance* $V_z(w) = \omega(w) - \omega(z) - \langle \omega'(z), w - z \rangle$, where $z \in Q^o$ and $w \in Q$. By strong convexity, we have:

$$V_z(w) \geq \frac{1}{2}\|w - z\|^2 \qquad \forall (z \in Q^o, w \in Q). \tag{6}$$

◇ The *ω-diameter* Ω *of* Q, which is defined as:

$$\Omega := \sqrt{2 \max_{z \in Q} V_{z^\omega}(z)} \leq \sqrt{2\left(\max_{z \in Q} \omega(z) - \min_{z \in Q} \omega(z)\right)},$$

where the concluding inequality follows from the fact that $V_{z^\omega}(z) \leq \omega(z) - \omega(z^\omega)$ due to $\langle \omega'(z^\omega), z - z^\omega \rangle \geq 0$ for every $z \in Q$, see (5). Further, by (6), we have:

$$\|w - z^\omega\| \leq \Omega \text{ for any } w \in Q, \text{ whence } D := \max_{w,z \in Q} \|w - z\| \leq 2\Omega. \tag{7}$$

◇ For parameter $z \in Q^o$, we define the *Prox-mapping* as:

$$\mathrm{Prox}_z(\xi) = \arg\min_{w \in Q} \left[V_z(w) + \langle \xi, w \rangle\right] : E \to Q^o$$

(the arg min in question indeed belongs to Q^o, see Remark 2.1).

2.3 Mirror-Prox algorithm with noisy first-order information

The prototype MP algorithm [Nem04a] is aimed at solving variational inequality (2) when exact values of F are available. In this paper, we use a modification of the original MP scheme, the *Stochastic Mirror-Prox* (SMP) algorithm proposed and investigated in [JNT08], which operates with noisy estimates of F. Specifically, the algorithm has access to a *Stochastic Oracle*: at the t-th call of the oracle, $z_t \in Q^o$ being the query point, the oracle returns an estimate $\hat{F}_{\xi_t}(z_t)$ of $F(z_t)$. Here, ξ_t is the t-th realization of the "oracle's noise", which is modeled as a random vector ξ, and $\hat{F}_\xi(z)$ is a Borel function of ξ and z. We assume that the realizations $(\xi_t)_{t\geq 1}$ of the random vector ξ are independent. From now on, we set $\xi_{[t]} = (\xi_1, \xi_2, ..., \xi_t)$. The algorithm is as follows.

Algorithm 1 [Mirror-Prox method with noisy first-order information]
1: Choose the number of iterations T. Set $z_0 = z^\omega \in Q^o$.
2: **for** $1 \leq t \leq T$ **do**
3: Given $z_{t-1} \in Q^o$, choose (a deterministic) $\gamma_t > 0$ such that:

$$\gamma_t \leq \frac{1}{\sqrt{2}L}. \tag{8}$$

4: Call Stochastic Oracle with query point z_{t-1} and receive $\eta_t := \hat{F}_{\xi_{2t-1}}(z_{t-1})$.
5: Set $w_t = \mathrm{Prox}_{z_{t-1}}(\gamma_t \eta_t) \in Q^o$.
6: Call Stochastic Oracle with query point w_t and receive $\zeta_t := \hat{F}_{\xi_{2t}}(w_t)$.
7: Set $z_t = \mathrm{Prox}_{z_{t-1}}(\gamma_t \zeta_t) \in Q^o$.
8: **end for**
9: Return $z^T := \left(\sum_{t=1}^T \gamma_t\right)^{-1} \sum_{t=1}^T \gamma_t w_t$.

Note that z_t is a deterministic function of $\xi_{[2t]}$, while w_t is a deterministic function of $\xi_{[2t-1]}$. In order to show expected convergence of Algorithm 1, we need to define the following quantities:

$$\begin{aligned}
\mu_{z_{t-1}} &:= \mathbb{E}_{\xi_{2t-1}}\left\{\hat{F}_{\xi_{2t-1}}(z_{t-1})|\ \xi_{[2t-2]}\right\} - F(z_{t-1}),\\
\mu_{w_t} &:= \mathbb{E}_{\xi_{2t}}\left\{\hat{F}_{\xi_{2t}}(w_t)|\xi_{[2t-1]}\right\} - F(w_t),\\
\sigma_{z_{t-1}} &:= \hat{F}_{\xi_{2t-1}}(z_{t-1}) - \mathbb{E}_{\xi_{2t-1}}\left\{\hat{F}_{\xi_{2t-1}}(z_{t-1})|\ \xi_{[2t-2]}\right\},\\
\sigma_{w_t} &:= \hat{F}_{\xi_{2t}}(w_t) - \mathbb{E}_{\xi_{2t}}\left\{\hat{F}_{\xi_{2t}}(w_t)|\xi_{[2t-1]}\right\},
\end{aligned} \tag{9}$$

where $1 \leq t \leq T$. Note that we define $\mathbb{E}_{\xi_1}\left\{\hat{F}_{\xi_1}(z_0)|\ \xi_{[0]}\right\} := \mathbb{E}_{\xi_1}\left\{\hat{F}_{\xi_1}(z_0)\right\}$. Note also that $\sigma_{z_{t-1}}$ and σ_{w_t} are martingale differences. The following result is proven in the Appendix.

Theorem 2.1 *Let*

$$\epsilon^T = \frac{\frac{\Omega^2}{2} + \Omega\|\sum_{t=1}^T \gamma_t \sigma_{w_t}\|_* + \sum_{t=1}^T \left\{\gamma_t D\|\mu_{w_t}\|_* + 2\gamma_t^2\left(\|\sigma_{w_t} - \sigma_{z_{t-1}}\|_*^2 + \|\mu_{w_t} - \mu_{z_{t-1}}\|_*^2\right)\right\}}{\sum_{t=1}^T \gamma_t}. \tag{10}$$

We have:

$$\mathbb{E}_{\xi_{[2T]}}\left\{\epsilon(z^T)\right\} \leq \mathbb{E}_{\xi_{[2T]}}\left\{\epsilon^T\right\}.$$

Moreover, for an operator F that is associated with saddle-point problem (3), the following inequality holds:

$$\mathbb{E}_{\xi_{[2T]}}\left\{\epsilon^{sad}(z^T)\right\} \le \mathbb{E}_{\xi_{[2T]}}\left\{\epsilon^T\right\}.$$

3 Mirror-Prox algorithm for matrix saddle-point problems

3.1 Matrix saddle-point problems

The problem of primary interest in this paper is the Eigenvalue Minimization problem

$$\mathrm{Opt} = \min_{x\in Q_1}\left[\lambda_{\max}(\mathcal{A}(x)+B) + \langle c, x\rangle\right], \tag{11}$$

where:

- Q_1 is a convex compact subset of the space $E_1 = \mathbb{R}^m$ equipped with the standard inner product $\langle x, y\rangle = x^T y$;
- $\mathcal{A}(x) = \sum_{i=1}^m x_i A_i$ is a linear mapping from $\mathbb{R}^m$ into the space $E_2 = \mathcal{S}^n$ of symmetric $n\times n$ matrices (so that $A_1, ..., A_m \in \mathcal{S}^n$), and $B \in \mathcal{S}^n$;
- $\lambda_{\max}(A)$ stands for the maximal eigenvalue of a symmetric matrix A;
- $c \in \mathbb{R}^m$.

We equip $E_2 = \mathcal{S}^n$ with the Frobenius inner product $\langle X, Y\rangle_F = \mathrm{Tr}(XY)$. Denoting by Δ_n^M the *spectahedron* $\{Y \in \mathcal{S}^n : Y \succeq 0, \mathrm{Tr}(Y) = 1\}$ and observing that $\lambda_{\max}(A) = \max\limits_{Y\in\Delta_n^M} \mathrm{Tr}(YA)$, we can reformulate (11) as the following bilinear saddle point problem:

$$\mathrm{Opt} := \min_{x\in Q_1}\max_{Y\in\Delta_n^M} \phi(x,Y), \qquad \phi(x,Y) = \langle B, Y\rangle_F + \langle \mathcal{A}(x), Y\rangle_F + \langle c, x\rangle. \tag{12}$$

The associated operator F is:

$$F(x,Y) := \left[F_x(Y) := \frac{\partial\phi(x,Y)}{\partial x} = \mathcal{A}^*(Y) + c; F_Y(x) := -\frac{\partial\phi(x,Y)}{\partial Y} = -\mathcal{A}(x) - B\right], \tag{13}$$

where the linear mapping $\mathcal{A}^*(Y) = [\mathrm{Tr}(A_1Y); \mathrm{Tr}(A_2Y); ...; \mathrm{Tr}(A_mY)] : \mathcal{S}^n \to \mathbb{R}^m$ is conjugate to the mapping $x \mapsto \mathcal{A}(x)$.

We are about to solve the Eigenvalue Minimization problem by applying to (12) the Mirror-Prox algorithm. While the problem is fully deterministic, we intend to use an appropriately constructed Stochastic Oracle, which is computationally cheaper than the exact deterministic oracle; our ultimate goal is to demonstrate that the resulting SMP algorithm significantly outperforms its deterministic counterpart in a meaningful range of problem's sizes.

We start with presenting the algorithm's setup.

3.2 Algorithm's setup

We assume that the space $E_1 = \mathbb{R}^m$ is equipped with a norm $\|\cdot\|_x$, the conjugate norm being $\|\cdot\|_{x,*}$, and that Q_1 is equipped with a d.-g.f. $\omega_x(x)$ that is compatible with $\|\cdot\|_x$. We denote by x^{ω_x} and Ω_x the ω_x-center of Q_1 and the ω_x-diameter of Q_1, respectively; see Section 2.2.

We equip the space $E_2 = \mathcal{S}^n$ with the trace-norm $\|W\|_Y := \sum_{i=1}^n |\lambda_i(W)|$, where the vector $\lambda(W) = [\lambda_1(W); \lambda_2(W); ...; \lambda_n(W)]$ consists of the eigenvalues $\lambda_1(W) \geq \ldots \geq \lambda_n(W)$ of $W \in \mathcal{S}^n$. As it is well-known, the conjugate norm is the usual spectral norm $\|W\|_{Y,*} = \max_{1\leq i\leq n} |\lambda_i(W)|$. Further, we equip the spectahedron $Q_2 := \Delta_n^M$ with the matrix entropy d.g.-f.:

$$\omega_Y(Y) = \mathrm{Ent}(Y) := \sum_{i=1}^n \lambda_i(Y) \ln(\lambda_i(Y)). \tag{14}$$

As shown in [Nes07], see also [BTN05], this d.-g.f. is compatible with $\|\cdot\|_Y$, and as it is immediately seen, the corresponding center and diameter of Q_2 are as follows:

$$Y^{\omega_Y} = \frac{1}{n} I_n; \quad \Omega_Y = \sqrt{2\ln(n)}. \tag{15}$$

Finally, we equip the embedding space $E = E_1 \times E_2$ of the domain $Q = Q_1 \times Q_2$ of (12) with the norm

$$\|(x, Y)\| = \sqrt{\frac{1}{\Omega_x^2}\|x\|_x^2 + \frac{1}{\Omega_Y^2}\|Y\|_Y^2}, \tag{16}$$

implying that the conjugate norm is

$$\|(x, Y)\|_* = \sqrt{\Omega_x^2\|x\|_{x,*}^2 + \Omega_Y^2\|Y\|_{Y,*}^2}. \tag{17}$$

The domain $Q = Q_1 \times Q_2$ of (12) is equipped with the d.-g.f.

$$\omega(x, Y) = \frac{1}{\Omega_x^2}\omega_x(x) + \frac{1}{\Omega_Y^2}\omega_Y(Y); \tag{18}$$

it is immediately seen that this indeed is a d.-g.f. for Q compatible with $\|\cdot\|$, and that the corresponding diameter of Q is $\Omega = \sqrt{2}$, while the ω-center of Q is $z^\omega = (x^{\omega_x}, Y^{\omega_Y})$.

Finally, let $\mathcal{L}$ be (an upper bound on) the norm of the linear mapping $x \mapsto \mathcal{A}(x) : E_1 \to E_2$ induced by the norms $\|\cdot\|_x$ and $\|\cdot\|_{Y,*}$ on the argument and the image spaces:

$$\forall x \in E_1 : \|\mathcal{A}(x)\|_{Y,*} \leq \mathcal{L}\|x\|_x. \tag{19}$$

It is immediately seen that the affine monotone operator F associated with (12) (see (13)) satisfies:

$$\forall (z, z' \in E = E_1 \times E_2) : \|F(z) - F(z')\|_* \leq L\|z - z'\|, \qquad \text{where } L := \Omega_x\Omega_Y\mathcal{L}. \tag{20}$$

3.3 Randomized Mirror-Prox method for (12)

3.3.1 Randomization: motivation and strategy

With the outlined setup, when applying the deterministic MP algorithm (i.e., Algorithm 1 with precise information: $\hat{F}_{\xi_t} \equiv F$) to the variational inequality associated with the saddle point problem (12), the computational effort at iteration t is dominated by the necessity

(A) to compute the value of F at two points, namely at the points $\bar{z} = z_{t-1} = (\bar{x}, \bar{Y}) \in Q^o$ and $\bar{w} = w_t \in Q^o$;

and

(B) to compute the value of the prox-mapping $\mathrm{Prox}_{\bar{z}}(\zeta) = \arg\min_{w \in Q} [\omega(w) + \langle \zeta - \omega'(\bar{z}), w - \bar{z} \rangle]$ at two different points $\zeta \in E$.

With our d.-g.f. ω that is "separable" with respect to the x- and to the Y-component of $\bar{z}$, task (B) reduces ("at no cost") to solving the two optimization problems:

$$\begin{array}{ll} (a) & \arg\min_{u \in Q_1} \left[\omega_x(u) + u^T[\zeta_x - \omega_x'(\bar{x})]\right] \\ (b) & P_{\bar{Y}}(\zeta_Y) := \arg\min_{V \in Q_2} \left[\mathrm{Ent}(V) + \mathrm{Tr}\left(V\left[\zeta_Y - \mathrm{Ent}'(\bar{Y})\right]\right)\right] \end{array} \tag{21}$$

with $\zeta_x \in \mathbb{R}^m = E_1$ and $\zeta_Y \in \mathcal{S}_n = E_2$ readily given by ζ, specifically, $\zeta = (\Omega_x^{-2}\zeta_x, \Omega_Y^{-2}\zeta_Y)$. In the sequel, we assume that (a) is easy to solve. The solution of (b) can be written explicitly (see, e.g., [BTN05]). Specifically, since $\bar{z} \in Q^o$, we have $\bar{Y} \in Q_2^o = \{Y \in \mathcal{S}^n : Y \succ 0, \mathrm{Tr}(Y) = 1\}$, and the latter set clearly is the set of all matrices of the form

$$\mathcal{H}(V) = \frac{\exp\{V\}}{\mathrm{Tr}(\exp\{V\})}, \qquad V \in \mathcal{S}^n. \tag{22}$$

Assuming that we have at our disposal a representation $\bar{Y} = \mathcal{H}(\bar{V})$ with $\bar{V} \in \mathcal{S}_n$, the solution to (b) is just $\mathcal{H}(-\zeta_Y + \bar{V})$. In other words, when parameterizing points $Y \in Q_2^o$ according to $Y = \mathcal{H}(V)$, prox-mapping (21) becomes trivial - it reduces to a matrix addition. The Y-components of the points $w_t = (u_t, W_t)$ and $z_t = (x_t, Y_t)$ generated by the deterministic MP are of the form $W_t = P_{Y_{t-1}}(\Omega_Y^2[\eta_t]_Y)$ and $Y_t = P_{Y_{t-1}}(\Omega_Y^2[\zeta_t]_Y)$, where $\eta_t, \zeta_t \in \mathbb{R}^m \times \mathcal{S}^n$ are given (see Algorithm 1). When using parametric representations $W_t = \mathcal{H}(U_t)$ and $Y_t = \mathcal{H}(V_t)$, the matrices U_t and V_t are easy to update: $U_t = V_{t-1} - \Omega_Y^2[\eta_t]_Y$ and $V_t = V_{t-1} - \Omega_Y^2[\zeta_t]_Y$, respectively, with η_t, ζ_t as defined in Algorithm 1. Thus, when representing W_t and Y_t by their "matrix logarithms" U_t and V_t, it looks *as if* the computational effort per step of MP as applied to (12) were dominated by the necessity to resolve task (A), and in task (B) — to solve the problem (21.a) alone. This impression, however, is not fully true. Indeed, looking at (13), we observe that — while computing the Y-component $F_Y(x) = -\mathcal{A}(x) - B$ of F at a point $z = (x, Y)$ needs the knowledge of x only and is independent of how Y is represented — computing the Y-component $F_x(Y) = [\mathrm{Tr}(A_1 Y); ...; \mathrm{Tr}(A_m Y)] + c$ of $F(z)$ seemingly requires the explicit representation of Y. This latter observation makes it necessary to solve explicitly problems (21.b), or, which is the same, requires computation of the value of $\mathcal{H}$ at a given point V. The related computational effort is $\mathcal{O}(n^3)$ (the arithmetic cost of an eigenvalue decomposition of V), which, depending on the problem's structure and sizes, can by far dominate all other computational expenses at an iteration. The goal of this paper is to demonstrate that one can avoid the explicit solution of "troublemaking" problems (21.b), and use instead the easy-to-update "logarithmic" representations at the cost of a randomized computation of $F_x(\cdot)$. The idea of randomization is as follows: assume that we are given a "matrix logarithm" V of $Y \in Q_2^o$, so that $Y = \mathcal{H}(V)$. We need to compute a randomized estimate of the vector $F_x(Y) = \mathcal{A}^*(Y) + c$, that is, of:

$$\mathcal{A}^*(Y) = [\mathrm{Tr}(A_1 Y); \ldots; \mathrm{Tr}(A_m Y)] = \frac{1}{\mathrm{Tr}(\exp\{V\})}[\mathrm{Tr}(A_1 \exp\{V\}); \ldots; \mathrm{Tr}(A_m \exp\{V\})].$$

Imagine for a moment that we can multiply vectors by the matrix $\exp\{V/2\}$. Then, generating a sample ξ of N independent vectors $\xi^s \sim \mathcal{N}(0, I_n)$, $1 \le s \le N$, and setting $\chi^s = \exp\{V/2\}\xi^s$, $s = 1, 2, ..., N$, we have:

$$\mathbb{E}\left\{\frac{1}{N}\sum_{s=1}^{N}[[\chi^s]^T A_1 \chi^s; \ldots; [\chi^s]^T A_m \chi^s]\right\} = [\mathrm{Tr}(A_1 \exp\{V\}); ...; \mathrm{Tr}(A_m \exp\{V\})],$$

and:

$$\mathbb{E}\left\{\frac{1}{N}\sum_{s=1}^{N}[\chi^s]^T \chi^s\right\} = \mathrm{Tr}(\exp\{V\}),$$

so that we can use the random vector

$$g_\xi(V) := \frac{\sum_{s=1}^{N}[[\chi^s]^T A_1 \chi^s; ...; [\chi^s]^T A_m \chi^s]}{\sum_{s=1}^{N}[\chi^s]^T \chi^s} + c, \qquad \chi^s := \exp\{V/2\}\xi^s, \tag{23}$$

as a random (biased!) estimate of $F_x(\mathcal{H}(V))$. The last strategic question to be addressed is how indeed to compute, given V and a vector ξ, the vector $\chi = \exp\{V/2\}\xi$. We propose to build a high accuracy approximation $\bar{\chi}$ to χ by setting

$$\bar{\chi} = \sum_{j=0}^{J} \frac{1}{j!}(V/2)^j \xi \tag{24}$$

with J large enough to guarantee a desired accuracy, and to compute the terms $\upsilon_j = \frac{1}{j!}(V/2)^j\xi$ by successive matrix-vector multiplications: $\upsilon_0 = \xi$, $\upsilon_{j+1} = \frac{1}{2(j+1)} V\upsilon_j$. We then merely replace in (23) the vectors χ^s with their approximations $\bar{\chi}^s$, thus getting an estimate

$$\hat{g}_\xi(V) := \frac{\sum_{s=1}^{N}[[\bar{\chi}^s]^T A_1 \bar{\chi}^s; ...; [\bar{\chi}^s]^T A_m \bar{\chi}^s]}{\sum_{s=1}^{N}[\bar{\chi}^s]^T \bar{\chi}^s} + c \tag{25}$$

of $g_\xi(V)$. We also set

$$\hat{\mathcal{H}}_\xi(V) = \left[\sum_{s=1}^{N}[\bar{\chi}^s]^T \bar{\chi}^s\right]^{-1} \sum_{s=1}^{N}[\bar{\chi}^s[\bar{\chi}^s]^T]; \tag{26}$$

note that $\hat{\mathcal{H}}_\xi(V) \in Q_2 = \Delta_n^M$ can be considered as a random estimate of $\mathcal{H}(V)$ (see (22)), and that $\hat{g}_\xi(V) = F_x(\hat{\mathcal{H}}_\xi(V))$.

3.3.2 Randomized algorithm

Implementing the outlined randomization strategy with the setup presented in Section 3.2, Algorithm 1 becomes as follows:

Algorithm 2 [Randomized Mirror-Prox method applied to matrix saddle-point problem (12)]
1: Choose the number of iterations T, the sample size N, and a sequence of positive integers J_t, $1 \le t \le T$. Generate $2T$ independent samples $\xi_1, \xi_2, ..., \xi_{2T}$, each of them comprised of N

independent realizations $\xi_t^s \sim \mathcal{N}(0, I_n)$, $1 \le s \le N$.
2: Set $x_0 = x^{\omega_x}$ and let $V_0 \in \mathcal{S}_n$ be the all zero matrix.
3: **for** $1 \le t \le T$ **do**
4: Given (x_{t-1}, V_{t-1}), choose (deterministic) $\gamma_t > 0$ such that (cf. (20), (19)):

$$\gamma_t \le \frac{1}{\sqrt{2}L} = \frac{1}{\sqrt{2}\Omega_x\Omega_Y\mathcal{L}}. \tag{27}$$

5: Compute the approximation

$$\hat{F}_{\xi_{2t-1}}(x_{t-1}, V_{t-1}) = \left[\hat{g}_{\xi_{2t-1}}(V_{t-1}); -B - \mathcal{A}(x_{t-1})\right]$$

of $F(x_{t-1}, \mathcal{H}(V_{t-1})) = [\mathcal{A}^*(\mathcal{H}(V_{t-1})) + c; -B - \mathcal{A}(x_{t-1})]$, where $\hat{g}_{\xi_{2t-1}}(\cdot)$ is as explained in (25), with J_t in the role of J.
6: Set

$$\begin{aligned} \bar{x}_t &= \arg\min_{x \in Q_1} \left\{\left\langle \gamma_t \Omega_x^2 \hat{g}_{\xi_{2t-1}}(V_{t-1}) - \omega_x'(x_{t-1}), x\right\rangle + \omega_x(x)\right\} \\ \bar{V}_t &= V_{t-1} + \gamma_t \Omega_Y^2 \left(B + \mathcal{A}(x_{t-1})\right). \end{aligned} \tag{28}$$

7: Compute the approximation $\hat{\mathcal{H}}_t := \hat{\mathcal{H}}_{\xi_{2t}}(\bar{V}_t)$ of $\mathcal{H}(\bar{V}_t)$ and the approximation

$$\hat{F}_{\xi_{2t}}(\bar{x}_t, \bar{V}_t) = \left[\hat{g}_{\xi_{2t}}(\bar{V}_t); -B - \mathcal{A}(\bar{x}_t)\right]$$

of $F(\bar{x}_t, \mathcal{H}(\bar{V}_t))$, where $\hat{\mathcal{H}}_{\xi_{2t}}(\bar{V}_t)$ and $\hat{g}_{\xi_{2t}}(\cdot)$ are as explained in (26) and (25), respectively, and with J_t in the role of J.
8: Set

$$\begin{aligned} x_t &= \arg\min_{x \in Q_1} \left\{\left\langle \gamma_t \Omega_x^2 \hat{g}_{\xi_{2t}}(\bar{V}_t) - \omega_x'(x_{t-1}), x\right\rangle + \omega_x(x)\right\} \\ V_t &= V_{t-1} + \gamma_t \Omega_Y^2 \left(B + \mathcal{A}(\bar{x}_t)\right). \end{aligned} \tag{29}$$

9: **end for**
10: Return $x^T := \left(\sum_{t=1}^T \gamma_t\right)^{-1} \sum_{t=1}^T \gamma_t \bar{x}_t$.

3.3.3 Convergence and complexity analysis

Regularity assumption and preliminaries. In order for Algorithm 2 to be well behaved, we need certain additional assumption on $(E_1, \|\cdot\|_{x,*})$, specifically, the one of *regularity* with certain parameter $\kappa = \kappa_{E_1}$. Instead of stating this notion here in full generality (this is done in Section A of the Appendix), let us just hint that the property has to do with "good behavior" of sums of martingale differences taking values in E_1 and list the regularity parameters for the most important, in regard to applications, pairs $(E_1, \|\cdot\|_{x,*})$. Specifically, denoting by $|\cdot|_p$ the spectral ℓ_p-norm on the space $\mathcal{M}^m$ of $m \times m$ matrices, that is, $|A|_p = \|\sigma(A)\|_p$, where $\sigma(A)$ is the vector of singular values of A, the following holds true (from now on, all $\mathcal{O}(1)$'s are appropriate absolute constants):

> (!) *If $1 \le p \le 2$ and either $(E_1, \|\cdot\|_x) = (\mathbb{R}^m, \|\cdot\|_p)$, or $(E_1, \|\cdot\|_x) = (\mathcal{M}^m, |\cdot|_p)$, then the regularity parameter of $(E_1, \|\cdot\|_{x,*})$ is equal to 1 when $p = 2$, is bounded from above by $\frac{1}{p-1}$ when $p > 1$, and is bounded from above by $\mathcal{O}(1)\ln(m+1)$.*

From now on, if the opposite is not explicitly stated, it is assumed that $(E_1, \|\cdot\|_{x,*})$ is κ-regular. An instrumental role in the convergence analysis of Algorithm 2 is played by the following fact (proved in Appendix).

Proposition 3.1 *Let $(E_1, \|\cdot\|_{x,*})$ be κ-regular for some κ. With F given by (13) and $g_\xi(\cdot)$ given by (23), one has for every $V \in \mathcal{S}^n$:*

$$\begin{array}{ll} (a) & \|\mathbb{E}_\xi\{g_\xi(V)\} - F_x(\mathcal{H}(V))\|_{x,*} \le \mathcal{O}(1)\mathcal{L}\sqrt{\kappa}N^{-1} \\ (b) & \mathbb{E}_\xi\left\{\exp\left\{\frac{\sqrt{N}\|g_\xi(V)-\mathbb{E}_\xi\{g_\xi(V)\}\|_{x,*}}{\mathcal{O}(1)\mathcal{L}\sqrt{\kappa}}\right\}\right\} \le \exp\{1\}. \end{array} \tag{30}$$

Another component of our analysis is the following simple statement (recall that $\|\cdot\|_{Y,*}$ is the usual spectral norm on $\mathcal{S}^n$):

Proposition 3.2 *Let $W \in \mathcal{S}_n$ and $J \ge \exp(2)\|W\|_{Y,*}$. Then,*

$$\left\|\exp\{W\} - \sum_{j=0}^{J}\frac{W^j}{j!}\right\|_{Y,*} \le \exp\{-J\}. \tag{31}$$

This result can be proved by applying the same arguments as in the proof of Lemma 6 in [AK07].

Convergence analysis. Proposition 3.2 shows that in order to approximate the matrix exponent $\exp\{W\}$ by its Taylor polynomial within accuracy $\epsilon \ll 1$, it suffices to take for the degree of the polynomial the number $J = \mathcal{O}(1)\ln(1/\epsilon)\|W\|_{Y,*}$, so that J is "nearly independent" of ϵ. Now, when ϵ is really small – like 10^{-16} or even 10^{-100} – any ϵ-approximation of the matrix exponent is, "for all practical purposes," the same as the matrix exponent itself. Assuming that the choice of J indeed ensures "really small" inaccuracies in the approximation of the matrix exponent, we have all reasons to undertake a *simplified* convergence analysis of Algorithm 2, where we neglect the difference between the quantities $g_\xi(\cdot)$ as given by (23) and their estimates $\hat{g}_\xi(\cdot)$ (defined in (25)). Or, alternatively formulated: we analyze the idealized version of the algorithm with $g_\xi(\cdot)$ in place of $\hat{g}_\xi(\cdot)$.

The convergence analysis of the idealized algorithm is as follows. Let $1 \le t \le T$, and let γ_t satisfy (27). Note that in the notation of Algorithms 1 and 2 and of definitions (9) we have:

$$\begin{array}{l} z_{t-1} = (x_{t-1}, \mathcal{H}(V_{t-1})), \quad w_t = (\bar{x}_t, \mathcal{H}(\bar{V}_t)), \\ \mu_{z_{t-1}} = \left[\mathbb{E}_{\xi_{2t-1}}\left\{g_{\xi_{2t-1}}(V_{t-1})\middle|\xi_{[2t-2]}\right\} - F_x(z_{t-1}); 0\right] =: [\mu^x_{z_{t-1}}; 0], \\ \mu_{w_t} = [\mathbb{E}_{\xi_{2t}}\left\{g_{\xi_{2t}}(\bar{V}_t)\middle|\xi_{[2t-1]}\right\} - F_x(w_t); 0] =: [\mu^x_{w_t}; 0], \\ \sigma_{z_{t-1}} = \left[g_{\xi_{2t-1}}(V_{t-1}) - \mathbb{E}_{\xi_{2t-1}}\left\{g_{\xi_{2t-1}}(V_{t-1})\middle|\xi_{[2t-2]}\right\}; 0\right] =: [\sigma^x_{z_{t-1}}; 0], \\ \sigma_{w_t} = \left[g_{\xi_{2t}}(\bar{V}_t) - \mathbb{E}\left\{g_{\xi_{2t}}(\bar{V}_t)\middle|\xi_{[2t-1]}\right\}; 0\right] =: [\sigma^x_{w_t}; 0]. \end{array}$$

By (30.a) combined with (17), we obtain:

$$\begin{array}{l} \mathbb{E}_{\xi_{2t-1}}\left\{\|\mu_{z_{t-1}}\|_*\middle|\xi_{[2t-2]}\right\} = \Omega_x\mathbb{E}_{\xi_{2t-1}}\left\{\|\mu^x_{z_{t-1}}\|_{x,*}\middle|\xi_{[2t-2]}\right\} \le \mathcal{O}(1)\Omega_x\mathcal{L}\sqrt{\kappa}N^{-1}, \\ \mathbb{E}_{\xi_{2t}}\left\{\|\mu_{w_t}\|_*\middle|\xi_{[2t-1]}\right\} = \Omega_x\mathbb{E}_{\xi_{2t}}\left\{\|\mu^x_{w_t}\|_{x,*}\middle|\xi_{[2t-1]}\right\} \le \mathcal{O}(1)\Omega_x\mathcal{L}\sqrt{\kappa}N^{-1}, \end{array} \tag{32}$$

whence also:

$$\begin{array}{l}\mathbb{E}_{\xi_{[2T]}}\left\{\sum_{t=1}^T \gamma_t\|\mu_{w_t}\|_*\right\} = \Omega_x\mathbb{E}_{\xi_{[2T]}}\left\{\sum_{t=1}^T \gamma_t\|\mu^x_{w_t}\|_{x,*}\right\} \le \mathcal{O}(1)\Omega_x\mathcal{L}\sqrt{\kappa}N^{-1}\sum_{t=1}^T\gamma_t,\\ \mathbb{E}_{\xi_{[2T]}}\left\{\sum_{t=1}^T \gamma_t^2\|\mu_{w_t}-\mu_{z_{t-1}}\|_*^2\right\} = \Omega_x^2\mathbb{E}_{\xi_{[2T]}}\left\{\sum_{t=1}^T \gamma_t^2\|\mu_{w_t}-\mu_{z_{t-1}}\|_{x,*}^2\right\}\\ \qquad\qquad\qquad\qquad \le \mathcal{O}(1)\Omega_x^2\mathcal{L}^2\kappa N^{-2}\sum_{t=1}^T\gamma_t^2.\end{array} \tag{33}$$

Further, inequality (30.b) implies:

$$\mathbb{E}_{\xi_{2t-1}}\left\{\|\sigma^x_{z_{t-1}}\|_{x,*}^2\big|\xi_{[2t-2]}\right\} \le \mathcal{O}(1)\mathcal{L}^2\kappa/N, \qquad \mathbb{E}_{\xi_{2t}}\left\{\|\sigma^x_{w_t}\|_{x,*}^2\big|\xi_{[2t-1]}\right\} \le \mathcal{O}(1)\mathcal{L}^2\kappa/N. \tag{34}$$

Moreover, by the definition of $\sigma^x_{z_{t-1}}$ and $\sigma^x_{w_t}$, we have:

$$\mathbb{E}_{\xi_{2t-1}}\{\sigma^x_{z_{t-1}}\big|\xi_{[2t-2]}\} = 0, \qquad \mathbb{E}_{\xi_{2t}}\{\sigma^x_{w_t}\big|\xi_{[2t-1]}\} = 0. \tag{35}$$

Since $(E_1, \|\cdot\|_{x,*})$ is κ-regular, relations (34) and (35) imply by Proposition 3 of [Nem04b] that:

$$\mathbb{E}_{\xi_{[2T]}}\left\{\left\|\sum_{t=1}^T\gamma_t\sigma^x_{w_t}\right\|_{x,*}^2\right\} \le \mathcal{O}(1)\kappa^2\mathcal{L}^2N^{-1}\sum_{t=1}^T\gamma_t^2,$$

which results in:

$$\begin{array}{rl}\mathbb{E}_{\xi_{[2T]}}\left\{\left\|\sum_{t=1}^T\gamma_t\sigma_{w_t}\right\|_*\right\} &= \Omega_x\mathbb{E}_{\xi_{[2T]}}\left\{\left\|\sum_{t=1}^T\gamma_t\sigma^x_{w_t}\right\|_{x,*}\right\} \le \Omega_x\sqrt{\mathbb{E}_{\xi_{[2T]}}\left\{\left\|\sum_{t=1}^T\gamma_t\sigma^x_{w_t}\right\|_{x,*}^2\right\}}\\ &\le \mathcal{O}(1)\Omega_x\mathcal{L}\kappa N^{-1/2}\sqrt{\sum_{t=1}^T\gamma_t^2}.\end{array} \tag{36}$$

Besides this, (34) implies that:

$$\begin{array}{rl}\mathbb{E}_{\xi_{[2T]}}\left\{\sum_{t=1}^T\gamma_t^2\|\sigma_{w_t}-\sigma_{z_{t-1}}\|_*^2\right\} &= \Omega_x^2\mathbb{E}_{\xi_{[2T]}}\left\{\sum_{t=1}^T\gamma_t^2\|\sigma^x_{w_t}-\sigma^x_{z_{t-1}}\|_{x,*}^2\right\}\\ &\le \mathcal{O}(1)\Omega_x^2\mathcal{L}^2\kappa N^{-1}\sum_{t=1}^T\gamma_t^2.\end{array} \tag{37}$$

Combining (33), (36), (37), and taking into account that with our setup $\Omega = \sqrt{2}$ and $D \le 2\Omega$, we conclude from Theorem 2.1 that:

$$\mathbb{E}\left\{\overline{\phi}(x^T) - \min_{x\in Q_1}\overline{\phi}(x)\right\} \le \mathcal{O}(1)\frac{1+\Omega_x\mathcal{L}\sqrt{\kappa}\left[N^{-1}\sum_{t=1}^T\gamma_t+\sqrt{\kappa}N^{-1/2}\sqrt{\sum_{t=1}^T\gamma_t^2}+\Omega_x\mathcal{L}\sqrt{\kappa}N^{-1}\sum_{t=1}^T\gamma_t^2\right]}{\sum_{t=1}^T\gamma_t},$$

where $\phi(x,Y)$ is the cost function of the saddle point problem (12), $\overline{\phi}(x) = \max\limits_{Y\in\Delta_n^M}\phi(x,Y)$ is the objective in the problem of interest (11), and $\mathrm{Opt} = \min\limits_{x\in Q_1}\overline{\phi}(x)$ is the saddle point value in (12), or, equivalently, the optimal value in (11).

Optimizing the resulting efficiency estimate in the stepsizes γ_t satisfying (27), it is immediately seen that with

$$N = \mathrm{floor}\left(\frac{T\kappa}{2\Omega_Y^2}\right) = \mathrm{floor}\left(\frac{T\kappa}{4\ln(n)}\right), \qquad \gamma_t = \frac{1}{\sqrt{2}\Omega_x\Omega_Y\mathcal{L}} = \frac{1}{2\sqrt{\ln(n)}\Omega_x\mathcal{L}},\ 1\le t\le T, \tag{38}$$

the above inequality implies:

$$\mathbb{E}\left\{\overline{\phi}(x^T) - \min_{x\in Q_1}\overline{\phi}(x)\right\} \le \mathcal{O}(1)\frac{\Omega_x\mathcal{L}}{T}\left[\frac{\ln(n)}{\sqrt{\kappa}} + \sqrt{\kappa\ln(n)}\right]. \tag{39}$$

4 An application: minimizing the maximal eigenvalue of a convex combination of symmetric matrices

Consider the special case of problem (11) where Q_1 is the standard simplex in $\mathbb{R}^m$:

$$Q_1 = \Delta_m := \left\{x \in \mathbb{R}^m : x \ge 0, \sum_{j=1}^m x_j = 1\right\} \qquad [m \ge 2]$$

and $B = 0$, $c = 0$, so that the problem is

$$\mathrm{Opt} = \min_{x\in\Delta_m} \lambda_{\max}(\mathcal{A}(x)), \qquad \mathcal{A}(x) = \sum_{j=1}^m x_j A_j. \tag{40}$$

In other words, we want to minimize the largest eigenvalue of a convex combination $\sum_j x_j A_j$ of given symmetric $m \times m$ matrices. Note that the problem of minimizing the maximal eigenvalue of $B + \mathcal{A}(x)$ over Q_1 reduces to (40) by replacing the matrices A_j with $A_j + B$. The operator (13) associated with the problem is

$$F(x, Y) = [F_x(Y); F_y(x)] := [[\mathrm{Tr}(A_1 Y); ...; \mathrm{Tr}(A_m Y)]; -\mathcal{A}(x)]. \tag{41}$$

We equip $E_1 = \mathbb{R}^m$ with the norm $\|\cdot\|_x = \|\cdot\|_1$; the conjugate norm is $\|\cdot\|_{x,*} = \|\cdot\|_\infty$. It is well known that $(\mathbb{R}^m, \|\cdot\|_\infty)$ is κ-regular with $\kappa = \mathcal{O}(1)\ln(m)$ (see, e.g., Example 2.1 in [Nem04b]). Note that this choice of $\|\cdot\|_1$ results in

$$\mathcal{L} = \max_{1\le j\le m} \|A_j\|_{Y,*} \tag{42}$$

(see (20)), where $\|A\|_{Y,*}$ is the spectral norm of a matrix A.

We equip Q_1 with the d.-g.f. function:

$$\omega_x(x) = \mathrm{Ent}(x) := \sum_{j=1}^m x_j \ln(x_j),$$

which, as it is well known (and immediately seen), is compatible with $\|\cdot\|_1$. The associated entities are

$$Q_1^o = \left\{x \in \mathbb{R}^m : x > 0, \sum_{j=1}^m x_j = 1\right\}, \qquad x^\omega = \left[\frac{1}{m}; ...; \frac{1}{m}\right], \qquad \Omega_x = \sqrt{2\ln(m)}, \tag{43}$$

and the prox-mapping is given by an explicit formula:

$$\left(\arg\min_{w\in Q_1}\{\omega_x(w) + \langle\xi - \omega_x'(x), w\rangle\}\right)_j = \frac{\exp\{\ln(x_j) - \xi_j\}}{\sum_{\ell=1}^m \exp\{\ln(x_\ell) - \xi_\ell\}}, \qquad 1 \le j \le m,$$

see Section 3.3.1.

Let us solve (40) by T-step Algorithm 2 associated with the outlined setup and the parameters N and γ chosen according to (38), that is, as:

$$N = \mathcal{O}(1)\frac{\ln(m)}{\ln(n)}T, \qquad \gamma_t \equiv \gamma = \frac{1}{2\mathcal{L}\sqrt{2\ln(n)\ln(m)}}. \tag{44}$$

The efficiency estimate (39) now reads:

$$\mathbb{E}\left\{\lambda_{\max}(\mathcal{A}(x^T)) - \mathrm{Opt}\right\} \le \mathcal{O}(1)\left(\ln(n) + \ln(m)\sqrt{\ln(n)}\right)\mathcal{L}/T. \tag{45}$$

Choosing truncation levels J_t. Let us specify the "truncation levels" J_t for $1 \le t \le T$. In view of (28), (29), (42) and taking into account that $V_0 = 0$, $\Omega_Y = \sqrt{2\ln(n)}$, and $\Omega_x = \sqrt{2\ln(m)}$, we conclude that:

$$\|V_t\|_{Y,*} \le \mathcal{O}(1)\sqrt{\ln(n)/\ln(m)}t, \qquad \|\bar{V}_t\|_{Y,*} \le \mathcal{O}(1)\sqrt{\ln(n)/\ln(m)}t.$$

From Proposition 3.2, we deduce that the matrix exponentials we need to use can be approximated with accuracy $\delta \ll 1$ by a truncated Taylor series with $J_t = \mathcal{O}(1)\sqrt{\ln(n)/\ln(m)}\ln(1/\delta)t$ terms. Specifying δ as, say, machine accuracy, we see that "for all practical purposes" it suffices to take

$$J_t = \mathcal{O}(1)\sqrt{\ln(n)/\ln(m)}t, \qquad t \ge 1, \tag{46}$$

with a moderate absolute constant $\mathcal{O}(1)$.

Overall complexity. Assume that we want to solve (40) within accuracy ϵ in terms of the objective. This task is typically unreachable with a randomized algorithm. Instead, we need to content ourselves with a procedure returning an ϵ-solution with a prescribed probability of at least $1-\beta$, where $0 < \beta \ll 1$. To build such a procedure, we can specify $T = T(\epsilon)$ in such a way that the right hand side in (45) is at most $\epsilon/4$. We run the above $T(\epsilon)$-step algorithm k times, each time computing an accurate, within the margin $\epsilon/2$, estimate of the value $\lambda_{\max}(\mathcal{A}(x^{T,i}))$ of the objective at the corresponding output $x^{T,i}$, $1 \le i \le k$, and then select among the k outputs $x^{T,1}, ..., x^{T,k}$ the one with the smallest estimate of the objective value. Since with our choice of $T(\epsilon)$ we have $\mathrm{Prob}\{\lambda_{\max}(x^{T,i}) - \mathrm{Opt} > \epsilon/2\} \le \frac{1}{2}$ and $x^{T,1}$,...,$x^{T,k}$ are independent, this procedure yields an ϵ-solution to the problem of interest with a probability of at least $1-\beta$ for a "small" $k = \mathcal{O}(1)\ln(1/\beta)$.

Now, let us evaluate the computational complexity of a single $T(\epsilon)$-step run of Algorithm 2. Assume that every matrix A_i has at most S nonzero entries. We assume that $mS \ge n^2$, meaning that the matrices $\mathcal{A}(x)$ can be fully dense. In order to avoid intricate expressions, we omit in the sequel all factors that are logarithmic in m, n and $1/\beta$ (in particular, all absolute constant factors) and write down the statement "P is, within logarithmic factors, bounded from above by Q" as $P \lesssim Q$. We also write $P \sim Q$ when both $P \lesssim Q$ and $Q \lesssim P$. Finally, we set $\nu = \epsilon/\mathcal{L}$; this quantity can be naturally interpreted as the relative accuracy of an ϵ-solution. To establish the complexity of our procedure, note the following.

(A) By (45), the required number of steps $T = T(\epsilon)$ admits the bound $T \lesssim 1/\nu$, whence, by (44), $N \lesssim 1/\nu$.

(B) As it is immediately seen, when $mS \geq n^2$, the computational effort at step $t \leq T$ of the algorithm is, within factor $\mathcal{O}(1)$, dominated by the necessity

1. to compute $\mathcal{A}(x)$ at a given point x, ($\lesssim mS$ arithmetic operations (a.o.));
2. to generate N samples $\xi_t^s \sim \mathcal{N}(0, I_n)$ with $1 \leq s \leq N$, (totally $\lesssim nN$ a.o.);
3. to compute for every $s \leq N$ the vectors $\bar{\chi}_t^s = \sum_{j=0}^{J_t} (V/2)^j \xi_t^s / j!$, where $V \in \mathcal{S}^n$ is a given matrix (totally $\lesssim Ntn^2$ a.o., see (46));
4. to build the matrix $H = \left[\sum_{s=1}^N [\bar{\chi}_t^s]^T \bar{\chi}_t^s\right]^{-1} \sum_{s=1}^N \bar{\chi}_t^s [\bar{\chi}_t^s]^T$ ($\lesssim Nn^2$ a.o.);
5. to compute the vector $[\mathrm{Tr}(HA_1); \mathrm{Tr}(HA_2); ...; \mathrm{Tr}(HA_n)]$ ($\lesssim mS$ a.o.).

We see that the complexity of step t is $\lesssim Ntn^2 + mS$ a.o., implying that the overall complexity of a single run of the algorithm is $\lesssim NT^2n^2 + TmS \lesssim n^2/\nu^3 + mS/\nu$ a.o. We then should compute the value of the objective at the resulting approximate solution, that is, the maximal eigenvalue of a symmetric matrix with the spectral norm not exceeding $\mathcal{L}$. For our purposes, it suffices to approximate this value $(1 - \beta/k)$-reliably within accuracy $\mathcal{O}(\epsilon)$, which can be done by the Power method at the cost of $\lesssim n^2/\nu$ a.o. Finally, we should repeat this procedure $\mathcal{O}(1)\ln(1/\beta)$ times. Omitting constants and factors logarithmic in m, n, and $1/\beta$, our randomized algorithm yields an $(1 - \beta)$-reliable ϵ-solution to (40) at the cost of

$$\mathcal{C}_{\mathrm{SMP}} = \frac{n^2}{\nu^3} + \frac{mS}{\nu} \text{ a.o.} \qquad [\nu = \epsilon/\mathcal{L}].$$

Discussion. Let us compare the complexity of our algorithm with those of its existing competitors. To the best of our knowledge, the best existing complexity bounds for large-scale problems (40) are as follows (we again skip logarithmic factors):

- ◇ The complexity for *Interior Point methods* without any assumptions on A_j aside of their sparsity is
$$\mathcal{C}_{\mathrm{IPM}} = \sqrt{\max[n, m]}[n^3 + m^3 + m^2n^2]\ln(1/\nu) \text{ a.o.}$$
- ◇ *Advanced deterministic first-order algorithms*, like Nesterov's Smoothing [Nes07] or deterministic Mirror-Prox, require
$$\mathcal{C}_{\mathrm{FOM}} = \frac{n^3 + mS}{\nu} \text{ a.o.}$$
- ◇ We can also consider minimizing the original objective function $x \mapsto \lambda_{\max}(\sum_j x_j A_j)$ over the standard simplex using a *"slightly randomized" Mirror Descent method* [d'A08b]. This method requires
$$\mathcal{C}_{\mathrm{MD}} = \frac{n^2}{\nu^{5/2}} + \frac{mS}{\nu^2} \text{ a.o.}$$
The iteration count in this method is $\sim 1/\nu^2$. The computational effort per iteration reduces to assembling $\mathcal{A}(x)$ at a given point ($\sim mS$ a.o.) and computing an ϵ-subgradient of the objective and an ϵ-approximation of the value of the objective at x by applying the Power method to the matrix $\mathcal{A}(x)$ in order to approximate its maximal singular value and leading eigenvector. With a straightforward implementation of the Power method this task requires $\sim n^2/\nu$ a.o., and with an advanced implementation $\sim n^2/\sqrt{\nu}$ a.o. only.

	CPU time [sec]			number of iterations			CPU time per iteration
N	mean	std	95% conf	mean	std	95% conf	[sec/iteration]
1	66	9	[60, 72]	2948	327	[2746, 3151]	0.0224
3	90	14	[81, 98]	2970	343	[2757, 3183]	0.0302
5	86	11	[79, 93]	2900	271	[2732, 3068]	0.0298
10	87	12	[80, 94]	2860	207	[2732, 2988]	0.0305
50	92	7	[87, 96]	2840	232	[2696, 2984]	0.0323
100	98	9	[93, 104]	2850	207	[2722, 2978]	0.0344
500	141	15	[131, 150]	2860	222	[2722, 2998]	0.0491
1000	178	18	[167, 189]	2860	222	[2722, 2998]	0.0622
5000	533	47	[504, 562]	2877	204	[2750, 3003]	0.1851

Table 1: CPU time (mean, standard deviation, and 95% confidence interval), number of iterations (mean, standard deviation, and 95% confidence interval), and average CPU time per iteration required by the stochastic Mirror-Prox method for solving random instances of problem (47) with parameter values $n = 100$, $m = 100$, $\epsilon = 0.002$, and for different samples sizes N. The matrices A_j have a joint sparsity pattern and, in expectation, 10% of the entries are non-zero.

It turns out that *there exists a meaningful range of values of m, n, S, and ν where our stochastic algorithm significantly outperforms the outlined competitors.* For example, consider the case when n is large, and assume that we have for some $0 < \kappa < 1/4$:

$$mS \sim n^\beta \text{ with } 2 + 2\kappa \le \beta \le 3 - 2\kappa,\ n^{\max[2-\beta,-1/2]+\kappa} \le \nu \le n^{1-\beta/2}$$

(note that the outlined range of values of ν is nonempty; e.g., this range is $n^{-1/2+\kappa} \le \nu \le n^{-1/4}$ with $\beta = 2.5$). It is immediately seen that in the case in question we have $\mathcal{C}_{\rm SMP} \sim n^2/\nu^3$ and:

$$\frac{\mathcal{C}_{\rm IPM}}{\mathcal{C}_{\rm SMP}} \gtrsim n^{3\kappa},\ \frac{\mathcal{C}_{\rm FOM}}{\mathcal{C}_{\rm SMP}} \gtrsim n^{2\kappa},\ \frac{\mathcal{C}_{\rm MD}}{\mathcal{C}_{\rm SMP}} \ge n^{\kappa},$$

that is, our algorithm progressively outperforms its competitors as the sizes grow.

5 Numerical experiments

We consider randomly generated instances of the problem

$$\text{Opt} = \min_{x \in \Delta_m} \lambda_{\max}(\mathcal{A}(x)), \qquad \mathcal{A}(x) = \sum_{j=1}^m x_j A_j, \qquad A_j = j^{3/2} C_j, \tag{47}$$

where C_j is a sparse symmetric random $n \times n$-matrix and $j = 1, \ldots, m$, i.e., we are confronted with instances of problem (40) that we studied in the last section. We solve these problem instances up to a (relative) accuracy of $\delta := \epsilon\mathcal{L}$, where $0 < \epsilon < 1$ is the target accuracy and $\mathcal{L}$ is defined as in (42). In all the numerical experiments that we perform, the target accuracy ϵ is set to $2 \cdot 10^{-3}$.

We implement Algorithm 2 with constant step-sizes $\gamma_t = \gamma$, $t \ge 1$, and γ has the form described in (44). Given a matrix $W \in \mathcal{S}_n$, we choose the truncation level J_W of the matrix exponential

Taylor series approximation according to the following formula (compare with Proposition 3.2):

$$J = \left\lfloor \max\left\{ \log(1/\rho), \exp(1)\, \|W\|_{(\infty)} \right\} \right\rfloor, \qquad \rho := 10^{-3}.$$

Note that this setting slightly deviates from the truncation level derived in Proposition 3.2. The ∞-norm of W is computed approximately using the Power method. In accordance to (44) and (45), we need to choose the sample size N as:

$$N = \frac{\mathcal{O}(1) \ln^2(m)}{\epsilon \sqrt{\ln(n)}}. \tag{48}$$

In Table 1, we give the CPU time (mean, standard deviation, and corresponding 95% confidence interval), the number of iterations needed to find a solution with relative accuracy $\epsilon\mathcal{L}$ (mean, standard deviation, and corresponding 95% confidence interval), and the average CPU time per iteration for different samples sizes N. The matrices A_j are all of size 100×100, they follow the same sparsity pattern, and, on average, 10% of the entries are different from zero. In total, we have a hundred matrices A_j. All the numerical results that we present in this paper are averaged over ten runs and are obtained on a computer with 24 processors, each of them with 2.67 GHz, and with 96 GB of RAM. We observe that the smaller the sample size the lower the CPU time that is required to approximately solve the problem instances. Surprisingly, we can choose a very small sample size without sacrificing too many iterations. Let us illustrate this observation with an example. According to (48) and with an absolute constant of 1, we are supposed to choose N as about 5000. With this parameter choice, we need an average CPU time of 533 seconds. Using only one sample for each matrix exponential approximation, we observe that we can reduce the average CPU time by 87.6% (with a slight increase of 2.5% in the average number of iterations). For the subsequent tests, we will thus choose a sample size that deviates from its theoretical value and use only one sample for every matrix exponential approximation.

Given a pair $(\bar{x}, \bar{Y})$ of a primal and a dual feasible solution, we can compute the corresponding duality gap

$$\lambda_{\max}(\mathcal{A}(\bar{x})) - \min_{x \in \Delta_m} \left\langle \mathcal{A}(x), \bar{Y} \right\rangle_F, \tag{49}$$

which we use as stopping criterion for our algorithm and which we check at every 100th iteration of the method. The first term is approximated by an adapted version of the Power method and the second term is simply $\min\{\langle A_j, \bar{Y} \rangle_F : 1 \leq j \leq m\}$. As the Power method typically returns a lower bound on the eigenvalue of largest absolute value, we recompute the duality gap using the Matlab built-in functions *max()* and *eig()* when the first approximation obtained by our version of the Power method yields to a value that is smaller than $\epsilon\mathcal{L}$. We denote by $\hat{\mathcal{H}}(V)$ the approximation of $\exp(V)/\mathrm{Tr}(\exp(V))$ by the truncated Taylor development. The pair $(\bar{x}, \bar{Y})$ considered at iteration t is the average

$$\frac{1}{\sum_{\tau=1}^{t} \gamma_\tau} \sum_{\tau=1}^{t} \gamma_\tau (\bar{x}_\tau, \hat{\mathcal{H}}(\bar{V}_\tau)) = \frac{1}{t} \sum_{\tau=1}^{t} (\bar{x}_\tau, \hat{\mathcal{H}}(\bar{V}_\tau)),$$

where $\bar{x}_\tau$ and $\bar{V}_\tau$ are defined in Algorithm 2, equations (28). In principle, the criterion (49) gives theoretically a desirable solution only if we use exact scaled exponentials instead of $\hat{\mathcal{H}}(V_\tau)$. Nevertheless, $\hat{\mathcal{H}}(V)$ is in the matrix simplex by construction, and the number of terms we use in the

CPU time (mean [sec], standard deviation [sec])

	Mirror-Descent		det Mirror-Prox		stoch Mirror-Prox		ratios CPU time		
(n, S)	mean	std	mean	std	mean	std	$\frac{\text{det MP}}{\text{MD}}$	$\frac{\text{stoch MP}}{\text{MD}}$	$\frac{\text{stoch MP}}{\text{det MP}}$
$(100, 955)$	307	47	128	16	71	6	0.42	0.23	0.55
$(200, 3813)$	766	79	307	15	237	17	0.40	0.31	0.77
$(400, 15255)$	2522	120	1101	42	744	39	0.44	0.30	0.68
$(800, 60971)$	10262	746	4983	126	2814	74	0.49	0.27	0.56

number of iterations (mean, standard deviation) and average CPU time per iteration [sec / iteration]

	Mirror-Descent			det Mirror-Prox			stoch Mirror-Prox			ratios CPU time / iteration		
(n, S)	mean	std	$\frac{\text{CPU time}}{\text{iteration}}$	mean	std	$\frac{\text{CPU time}}{\text{iteration}}$	mean	std	$\frac{\text{CPU time}}{\text{iteration}}$	$\frac{\text{det MP}}{\text{MD}}$	$\frac{\text{stoch MP}}{\text{MD}}$	$\frac{\text{stoch MP}}{\text{det MP}}$
$(100, 955)$	21504	3287	0.0143	3120	349	0.0411	3000	313	0.0235	2.88	1.65	0.57
$(200, 3813)$	21388	1858	0.0358	2700	141	0.1137	2740	171	0.0866	3.17	2.42	0.76
$(400, 15255)$	22293	924	0.1131	2680	103	0.4109	2620	103	0.2841	3.63	2.51	0.69
$(800, 60971)$	22327	445	0.4596	2740	52	1.8187	2600	47	1.0822	3.96	2.35	0.60

Truncation level J of Taylor series approximation (only stochastic Mirror-Prox)

(n, S)	mean	std
$(100, 955)$	9	< 1
$(200, 3813)$	9	< 0.5
$(400, 15255)$	10	< 0.5
$(800, 60971)$	10	< 0.5

Table 2: CPU time (mean and standard deviation), number of iterations (mean and standard deviation), and average CPU time per iteration needed by the Mirror-Descent (MD), the deterministic Mirror-Prox (det MP), and the stochastic Mirror-Prox (stoch MP) method for solving random instances of problem (47) with parameter values $m = 100$ and $\epsilon = 0.002$, with S non-zero entries, and for different values of the matrix size n. The performance ratios express the CPU time (CPU time per iteration) required by "method A" as percentage of the corresponding quantity used by "method B". The stochastic Mirror-Prox method is implemented with $N = 1$, and the used truncation levels J are shown in the table at the bottom.

Taylor exponential is large enough to justify a very accurate approximation, so that $\hat{Y}$ can be considered as an adequate approximate solution to our problem.

In Table 2, we compare the performance of our randomized version of the Mirror-Prox method with the efficiency of its deterministic counterpart and of the Mirror-Descent scheme for random problem instances (47). As before, we have a hundred matrices A_j, but this time their size is varying. They are sparse with a joint sparsity pattern and with S non-zero values; the values for S can be found in Table 5. In this table, we show the CPU time (mean and standard deviation), the number of iterations required to find a solution with accuracy $\epsilon\mathcal{L}$ (mean and standard deviation), and the average CPU time per iteration. Moreover, we express the average (total) CPU time and the average CPU time per iteration of the Mirror-Descent method (deterministic Mirror-Prox) in percentage of the stochastic and the deterministic Mirror-Prox method (stochastic Mirror-Prox). We observe that the stochastic Mirror-Prox method has an average CPU time that corresponds to 23 to 31% of the running of the Mirror-Descent scheme and to 55 to 77% of the CPU time required by the deterministic Mirror-Prox method for problem instances involving matrices of size 100×100 up to size 800×800. For the stochastic Mirror-Prox method, we also give the truncation levels J, which are nine or ten on average for these problem instances.

Acknowledgments: We would like to thank Hans-Jakob Lüthi for having initiated this collarobation and for many helpful discussions. This research was partially funded NSF Grant DMS-0914785 and by Swiss National Science Foundation (SNF), project no 200021-129743, "First-order methods for Semidefinite Optimization".

References

[AK07] S. Arora and S. Kale, *A combinatorial, primal-dual approach to semidefinite programs*, Proceedings of the 39th Annual ACM Symposium on Theory of Computing, San Diego, California, USA, June 2007 (D. Johnson and U. Feige, eds.), ACM, 2007, pp. 227–236.

[BTN05] A. Ben-Tal and A. Nemirovski, *Non-Euclidean restricted memory level method for large-scale convex optimization*, Mathematical Programming **102** (2005), no. 3, 407–456.

[d'A08a] A. d'Aspremont, *Smooth optimization with approximate gradient*, SIAM Journal on Optimization **19** (2008), no. 3, 1171–1183.

[d'A08b] A. d'Aspremont, *Subsampling Algorithms for Semidefinite Programming*, Tech. report, March 2008, Available at http://www.princeton.edu/∼aspremon/ColSubSamp.pdf.

[HR00] C. Helmberg and F. Rendl, *A Spectral Bundle Method for Semidefinite Programming*, SIAM Journal on Optimization **10** (2000), no. 3, 673–696.

[JNT08] A. Juditsky, A. Nemirovski, and C. Tauvel, *Solving Variational Inequalities with Stochastic Mirror-Prox Algorithm*, Tech. report, 2008, Available at http://www2.isye.gatech.edu/∼nemirovs/SMP.pdf.

[Nem04a] A. 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** (2004), no. 1, 229–251.

[Nem04b] ———, *Regular Banach Spaces and Large Deviations of Random Sums*, Tech. report, 2004, Available at http://www2.isye.gatech.edu/∼nemirovs/LargeDev2004.pdf.

[Nes07] Y. Nesterov, *Smoothing technique and its applications in semidefinite optimization*, Mathematical Programming **110** (2007), no. 2, 245–259.

[Ous00] F. Oustry, *A second-order bundle method to minimize the maximum eigenvalue function*, Mathematical Programming **89** (2000), 1–33.

[Roc70] R. T. Rockafellar, *Convex Analysis*, Princeton Mathematics Series, vol. 28, Princeton University Press, 1970.

[TRW05] K. Tsuda, G. Rätsch, and M. K. Warmuth, *Matrix exponentiated gradient updates for online learning and bregman projections*, Journal of Machine Learning Research **6** (2005), 995–1018.

[WK06] M. K. Warmuth and D. Kuzmin, *Online variance minimization*, In Proceedings of the 19th Annual Conference on Learning Theory, Springer, 2006, pp. 514–528.

A Large deviations of random sums

Let E be a Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and let $\kappa \geq 1$. We endow E with a norm denoted by $\|\cdot\|$, which might differ from the one associated with this inner product.

We say that the space $(E, \|\cdot\|)$ is κ-smooth, if the function $p(\cdot) := \|\cdot\|^2$ is continuously differentiable on E and

$$p(x+y) \leq p(x) + \langle p'(x), y \rangle + \kappa p(y) \qquad \forall\ x, y \in E.$$

Both the space $(E, \|\cdot\|)$ and the norm $\|\cdot\|$ are called κ-regular, if there exist $\kappa_+ \in [1, \kappa]$ and a norm $\|\cdot\|_+$ on E with the following two characteristics:

- ◇ The space $(E, \|\cdot\|_+)$ is κ_+-smooth.
- ◇ The norm $\|\cdot\|_+$ is κ/κ_+-compatible with $\|\cdot\|$, that is:

$$\|x\|^2 \leq \|x\|_+^2 \leq \frac{\kappa}{\kappa_+}\|x\|^2 \qquad \forall\ x \in E.$$

The regularity constant κ_E of the space $(E, \|\cdot\|)$ is defined as:

$$\kappa_E := \inf\{\kappa \geq 1 :\ (E, \|\cdot\|) \text{ is } \kappa\text{-regular}\}.$$

From now on, assume that $(E, \|\cdot\|)$ has regularity constant κ_E.

We define an n-dimensional martingale difference sequence $\xi_1, \ldots, \xi_T$, that is, a sequence of n-dimensional random vectors such that ξ_{t-1} is $\sigma(\xi_t)$-measurable and $\mathbb{E}_{\xi_t}\{\xi_t | \xi_{t-1}\} = 0$ for every t. In our context, the following result on regular Euclidean spaces is of particular interest.

Theorem A.1 *[Nem04b] Choose $\chi \in (0,2]$ and reals $\sigma_1, \ldots, \sigma_T > 0$ such that:*

$$\mathbb{E}_{\xi_t}\left\{\exp\left\{\frac{\|\xi_t\|^\chi}{\sigma_t^\chi}\right\} \mid \xi_{t-1}\right\} \leq \exp\{1\} \qquad \forall\ t = 1, \ldots, T.$$

(a) For all $c \geq 0$, we have:

$$\mathbb{P}\left[\left\|\sum_{t=1}^T \xi_t\right\| > c\sqrt{\kappa_E \sum_{t=1}^T \sigma_t^2}\right] \leq C_\chi \exp\left(-c^\chi / C_\chi\right),$$

where $C_\chi \geq 2$ is a properly chosen constant that solely depends on χ and that is continuous in χ.

(b) With a properly chosen constant $c_\chi > 0$ that solely depends on χ and that is continuous in χ, we have:

$$\mathbb{E}_{\xi_{[T]}}\left\{\exp\left\{\left(\frac{\|\sum_{t=1}^T \xi_t\|}{c_\chi\sqrt{\kappa_E \sum_{t=1}^T \sigma_t^2}}\right)^\chi\right\}\right\} \leq \exp\{1\}.$$

■

For a proof of the first part of the above theorem, we refer to [Nem04b]. Statement b) follows from a) by integration.

B Proofs

B.1 Proof of Theorem 2.1

The proof of Theorem 2.1 requires a result from [JNT08], which we reproduce below.

Lemma B.1 *[JNT08] Let $z \in Q^o$, $\gamma > 0$, and $\eta, \zeta \in E$. Consider the points*

$$\begin{aligned} x &:= \arg\min_{y \in Q}\{\langle \gamma\eta - \omega'(z), y\rangle + \omega(y)\} \\ z_+ &:= \arg\min_{y \in Q}\{\langle \gamma\zeta - \omega'(z), y\rangle + \omega(y)\}. \end{aligned}$$

Then,

$$\langle \gamma\zeta, x - y\rangle \leq V_z(y) - V_{z_+}(y) + \frac{\gamma^2}{2}\|\eta - \zeta\|_*^2 - \frac{1}{2}\|x - z\|^2 \qquad \forall\ y \in Q.$$

■

Let us show Theorem 2.1:

Proof
We can represent the random elements $F(z_{t-1})$ and $F(w_t)$ as follows:

$$F(z_{t-1}) = \hat{F}_{\xi_{2t-1}}(z_{t-1}) - \sigma_{z_{t-1}} - \mu_{z_{t-1}} \qquad \text{and} \qquad F(w_t) = \hat{F}_{\xi_{2t}}(w_t) - \sigma_{w_t} - \mu_{w_t},$$

respectively. Let $u \in Q$. As F is a monotone operator, we have:

$$\sum_{t=1}^{T} \gamma_t \left\langle F(u), z^T - u \right\rangle = \sum_{t=1}^{T} \gamma_t \left\langle F(u), w_t - u \right\rangle \leq \sum_{t=1}^{T} \gamma_t \left\langle F(w_t), w_t - u \right\rangle$$
$$= \sum_{t=1}^{T} \gamma_t \left\langle \hat{F}_{\xi_{2t}}(w_t) - \sigma_{w_t} - \mu_{w_t}, w_t - u \right\rangle. \quad (50)$$

By Theorem B.1, we obtain:

$$\sum_{t=1}^{T} \gamma_t \left\langle \hat{F}_{\xi_{2t}}(w_t), w_t - u \right\rangle \leq V_{z^\omega}(u) + \sum_{t=1}^{T} \left(\frac{\gamma_t^2}{2} \left\| \hat{F}_{\xi_{2t}}(w_t) - \hat{F}_{\xi_{2t-1}}(z_{t-1}) \right\|_*^2 - \frac{1}{2} \left\| w_t - z_{t-1} \right\|^2 \right).$$

Furthermore,

$$\begin{aligned} \left\| \hat{F}_{\xi_{2t}}(w_t) - \hat{F}_{\xi_{2t-1}}(z_{t-1}) \right\|_*^2 &= \left\| F(w_t) + \sigma_{w_t} + \mu_{w_t} - F(z_{t-1}) - \sigma_{z_{t-1}} - \mu_{z_{t-1}} \right\|_*^2 \\ &\leq 2 \left(\left\| F(w_t) - F(z_{t-1}) \right\|_*^2 + \left\| \sigma_{w_t} - \sigma_{z_{t-1}} + \mu_{w_t} - \mu_{z_{t-1}} \right\|_*^2 \right) \\ &\leq 2 \left(L^2 \left\| w_t - z_{t-1} \right\|^2 + \left\| \sigma_{w_t} - \sigma_{z_{t-1}} + \mu_{w_t} - \mu_{z_{t-1}} \right\|_*^2 \right), \end{aligned}$$

where the concluding inequality is due to the Lipschitz continuity of F. Observe that $\gamma_t^2 L^2 \leq \frac{1}{2}$ because of the step-size choice (8). Thus,

$$\begin{aligned} \sum_{t=1}^{T} \gamma_t \left\langle \hat{F}_{\xi_{2t}}(w_t), w_t - u \right\rangle &\leq V_{z^\omega}(u) + \sum_{t=1}^{T} \gamma_t^2 \left\| \sigma_{w_t} - \sigma_{z_{t-1}} + \mu_{w_t} - \mu_{z_{t-1}} \right\|_*^2 \\ &\leq \frac{\Omega^2}{2} + 2 \sum_{t=1}^{T} \gamma_t^2 \left(\left\| \sigma_{w_t} - \sigma_{z_{t-1}} \right\|_*^2 + \left\| \mu_{w_t} + \mu_{z_{t-1}} \right\|_*^2 \right). \end{aligned} \quad (51)$$

Additionally, the following inequality holds:

$$\begin{aligned} \sum_{t=1}^{T} \gamma_t \left\langle \sigma_{w_t} + \mu_{w_t}, u - w_t \right\rangle &= \sum_{t=1}^{T} \gamma_t \left(\left\langle \sigma_{w_t}, u - z^\omega \right\rangle + \left\langle \sigma_{w_t}, z^\omega - w_t \right\rangle + \left\langle \mu_{w_t}, u - w_t \right\rangle \right) \\ &\leq \Omega \left\| \sum_{t=1}^{T} \gamma_t \sigma_{w_t} \right\|_* + \sum_{t=1}^{T} \gamma_t \left(\left\langle \sigma_{w_t}, z^\omega - w_t \right\rangle + D \left\| \mu_{w_t} \right\|_* \right). \end{aligned}$$

We combine the above inequality with (50) and (51):

$$\begin{aligned} \sum_{t=1}^{T} \gamma_t \left\langle F(u), z^T - u \right\rangle \leq\ & \frac{\Omega^2}{2} + 2 \sum_{t=1}^{T} \gamma_t^2 \left(\left\| \sigma_{w_t} - \sigma_{z_{t-1}} \right\|_*^2 + \left\| \mu_{w_t} + \mu_{z_{t-1}} \right\|_*^2 \right) \\ & + \Omega \left\| \sum_{t=1}^{T} \gamma_t \sigma_{w_t} \right\|_* + \sum_{t=1}^{T} \gamma_t \left(\left\langle \sigma_{w_t}, z^\omega - w_t \right\rangle + D \left\| \mu_{w_t} \right\|_* \right). \end{aligned}$$

Maximizing the left-hand side of the above inequality with respect to $u \in Q$ and taking expectations on both sides, we obtain:

$$\sum_{t=1}^{T} \gamma_t \mathbb{E}_{\xi_{[2T]}} \left\{ \epsilon(z^T) \right\} \leq \frac{\Omega^2}{2} + 2 \sum_{t=1}^{T} \gamma_t^2 \mathbb{E}_{\xi_{[2T]}} \left\{ \left\| \sigma_{w_t} - \sigma_{z_{t-1}} \right\|_*^2 + \left\| \mu_{w_t} + \mu_{z_{t-1}} \right\|_*^2 \right\}$$

$$+\Omega\mathbb{E}_{\xi_{[2T]}}\left\|\sum_{t=1}^{T}\gamma_t\sigma_{w_t}\right\|_* + \sum_{t=1}^{T}\gamma_t\mathbb{E}_{\xi_{[2T]}}\left\{\langle\sigma_{w_t}, z^\omega - w_t\rangle + D\left\|\mu_{w_t}\right\|_*\right\}.$$

It remains to observe that $\mathbb{E}_{\xi_{[2T]}}\{\langle\sigma_{w_t}, z^\omega - w_t\rangle\} = 0$, as σ_{w_t} is a martingale difference.

Assume that F is associated with the saddle-point problem (3), that is:

$$F(x,y) = \left(\frac{\partial\phi(x,y)}{\partial x}; -\frac{\partial\phi(x,y)}{\partial y}\right).$$

Recall that $t \in \{1,\ldots,T\}$. Let $w_t = (x_t, y_t) \in Q := Q_1 \times Q_2$, $u = (u_x, u_y) \in Q_1 \times Q_2$, and $\lambda_t := \gamma_t / \sum_{i=1}^T \gamma_i$. As the function ϕ is convex in the first and concave in the second argument, we have:

$$\begin{aligned}\sum_{t=1}^{T}\lambda_t\langle F(w_t), w_t - u\rangle &\geq \sum_{t=1}^{T}\lambda_t\left(\phi(x_t,y_t) - \phi(u_x,y_t) + \phi(x_t,u_y) - \phi(x_t,y_t)\right)\\ &= \sum_{t=1}^{T}\lambda_t\left(\phi(x_t,u_y) - \phi(u_x,y_t)\right)\\ &\geq \phi\left(\sum_{t=1}^{T}\lambda_t x_t, u_y\right) - \phi\left(u_x, \sum_{t=1}^{T}\lambda_t y_t\right).\end{aligned}$$

It remains to apply the same arguments as above in order to complete the proof. ∎

B.2 Proof of Proposition 3.1

For $V \in \mathcal{S}_n$, we define

$$g(V) := h(V) + c, \qquad \text{where } h(V) := \mathcal{A}^*\left(\frac{\exp\{V\}}{\mathrm{Tr}(\exp\{V\})}\right). \tag{52}$$

Recall that $\xi^1,\ldots,\xi^N$ are independent $\mathcal{N}(0,I_n)$-distributed random vectors and $\xi = (\xi^1,\ldots,\xi^N)$. Let $g_\xi(V)$ be defined as in (23), that is:

$$g_\xi(V) := h_\xi(V) + c, \qquad \text{where } h_\xi(V) := \mathcal{A}^*\left(\frac{G_\xi(V)}{\theta_\xi(V)}\right) \tag{53}$$

with

$$G_\xi(V) := \frac{\sum_{i=1}^N\left(\exp\{V/2\}\,\xi^i\right)\left(\exp\{V/2\}\,\xi^i\right)^T}{N} \qquad \text{and} \qquad \theta_\xi(V) := \frac{\sum_{i=1}^N[\xi^i]^T\exp\{V\}\,\xi^i}{N}. \tag{54}$$

We start with the observation:

Assumption B.1 *As the standard multivariate normal distribution $\mathcal{N}(0,I_n)$ is orthogonal invariant, and as both $h(V)$ and $h_\xi(V)$ are invariant under positive scaling, we can assume, without loss of generality, that $\exp\{V\}$ is diagonal (with positive diagonal entries) and of trace 1. This assumption shall hold for the rest of this section.*

For $i = 1, \ldots, N$, we set:

$$D_{\xi^i} := \left(\exp\{V/2\}\xi^i\right)\left(\exp\{V/2\}\xi^i\right)^T - \exp\{V\}, \qquad d_{\xi^i} := \mathcal{A}^* D_{\xi^i}, \qquad d_\xi := \frac{1}{N}\sum_{i=1}^N d_{\xi^i},$$

and:

$$f_{\xi^i} := [\xi^i]^T \exp\{V\}\xi^i - 1, \qquad f_\xi := \frac{1}{N}\sum_{i=1}^N f_{\xi^i}.$$

Lemma B.2 *For an appropriately chosen constant $c > 0$, we have for any $i = 1, \ldots, N$.:*

a) $\mathbb{E}_{\xi^i} D_{\xi^i} = \mathbb{E}_{\xi^i} d_{\xi^i} = \mathbb{E}_\xi d_\xi = \mathbb{E}_{\xi^i} f_{\xi^i} = \mathbb{E}_\xi f_\xi = 0.$

b) $\mathbb{E}_{\xi^i}\left\{\exp\left\{\frac{\|D_{\xi^i}\|_Y}{c}\right\}\right\} \le \exp\{1\}.$

c) $\mathbb{E}_{\xi^i}\left\{\exp\left\{\frac{\|d_{\xi^i}\|_{x,*}}{c\mathcal{L}}\right\}\right\} \le \exp\{1\}.$

d) $\mathbb{E}_\xi\left\{\exp\left\{\frac{\sqrt{N}\|d_\xi\|_{x,*}}{c\mathcal{L}\sqrt{\kappa}}\right\}\right\} \le \exp\{1\}.$

e) $\mathbb{E}_{\xi^i}\left\{\exp\left\{\frac{|f_{\xi^i}|}{c}\right\}\right\} \le \exp\{1\}.$

f) $\mathbb{E}_\xi\left\{\exp\left\{\frac{\sqrt{N}|f_\xi|}{c}\right\}\right\} \le \exp\{1\}.$

Proof

Assume that $diag(\exp\{V\}) = (v_1, \ldots, v_n)^T$, where $v_i \ge 0$ for any $i = 1, \ldots, n$ and $\sum_{i=1}^n v_i = 1$.

a) Let $i \in \{1, \ldots, n\}$. We have:

$$\mathbb{E}_{\xi^i} D_{\xi^i} = \exp\{V/2\}\mathbb{E}_{\xi^i}\left\{\xi^i[\xi^i]^T\right\}\exp\{V/2\} - \exp\{V\} = 0,$$

where the concluding equality holds as $\xi^i \sim \mathcal{N}(0, I_n)$. Moreover,

$$\mathbb{E}_{\xi^i}\left\{[\xi^i]^T \exp\{V\}\xi^i\right\} = \sum_{k=1}^n v_k = 1,$$

which proves $\mathbb{E}_{\xi^i} f_{\xi^i} = 0$. The remaining equalities follow immediately.

b) Assume that the n-dimensional random vector ζ is $\mathcal{N}(0, I_n)$-distributed. Then,

$$\left\|\left(\exp\{V/2\}\zeta\right)\left(\exp\{V/2\}\zeta\right)^T\right\|_Y = \sum_{i=1}^n v_i\zeta_i^2. \tag{55}$$

For any $0 < c_1 < \frac{1-\exp\{-2\}}{2}$ $\left(< \frac{1}{2}\right)$, it holds that:

$$\mathbb{E}_\zeta \exp\left\{c_1\left\|\left(\exp\{V/2\}\zeta\right)\left(\exp\{V/2\}\zeta\right)^T\right\|_Y\right\} = \prod_{i=1}^n \mathbb{E}_\zeta \exp\left\{c_1 v_i \zeta_i^2\right\}$$

$$
\begin{aligned}
&= \prod_{i=1}^{n} (1 - 2c_1 v_i)^{-1/2} \\
&= \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \ln(1 - 2c_1 v_i) \right\} \\
&\leq \exp\left\{ -\frac{1}{2} \ln(1 - 2c_1) \right\} \\
&\leq \exp\{1\}, \qquad (56)
\end{aligned}
$$

where the first inequality holds as the maximum of $-\sum_{i=1}^{n} \ln(1 - 2c_1 v_i)$ over the probability simplex is attained at an extreme point (note that the function $v \mapsto -\sum_{i=1}^{n} \ln(1 - 2c_1 v_i)$ is separable and each of its components is convex). We obtain:

$$
\exp\{1\} \geq c_1 \mathbb{E}_\zeta \left\| (\exp\{V/2\}\zeta)(\exp\{V/2\}\zeta)^T \right\|_Y .
$$

Using Jensen's inequality, it follows that:

$$
\mathbb{E}_\zeta \exp\left\{ \frac{c_1}{2\exp\{1\}} \left\| \chi\chi^T - \mathbb{E}_\zeta \chi\chi^T \right\|_Y \right\} \leq \exp\{1\}, \qquad \chi := \exp\{V/2\}\zeta.
$$

We conclude by observing that $(\exp\{V/2\}\zeta)(\exp\{V/2\}\zeta)^T - \mathbb{E}_\zeta (\exp\{V/2\}\zeta)(\exp\{V/2\}\zeta)^T$ has the same distribution than any of the matrices D_{ξ^i}, $i = 1, \ldots, N$.

c) Let $i = 1, \ldots, N$ and $0 < c_2 \leq \frac{c_1}{2\exp\{1\}}$. Due to a), we obtain:

$$
\mathbb{E}_{\xi^i} \exp\left\{ \frac{c_2 \left\| d_{\xi^i} \right\|_{x,*}}{\mathcal{L}} \right\} = \mathbb{E}_{\xi^i} \exp\left\{ \frac{c_2 \left\| \mathcal{A}^* D_{\xi^i} \right\|_{x,*}}{\mathcal{L}} \right\} \leq \mathbb{E}_{\xi^i} \exp\left\{ \frac{c_2 \left\| D_{\xi^i} \right\|_Y \mathcal{L}}{\mathcal{L}} \right\} \leq \exp\{1\}.
$$

d) According to Theorem A.1 (note that we apply Theorem A.1 with $\chi = 1$ and $\sigma_i = \mathcal{L}/c_2$ for any $i = 1, \ldots, N$), there exists $c_3 > 0$ such that:

$$
\mathbb{E}_\xi \exp\left\{ \frac{c_2 \sqrt{N} \left\| d_\xi \right\|_{x,*}}{c_3 \mathcal{L} \sqrt{\kappa}} \right\} \leq \exp\{1\}.
$$

e) Here, the n-dimensional random vector ζ is $\mathcal{N}(0, I_n)$-distributed. Due to (55) and (56), we obtain:

$$
\mathbb{E}_\zeta \exp\left\{ \frac{c_1}{2} \left| \zeta^T \exp\{V\}\zeta - 1 \right| \right\} \leq \exp\left\{ \frac{1}{2} \right\} \left(\mathbb{E}_\zeta \exp\left\{ c_1 \sum_{i=1}^{n} v_i \zeta_i \right\} \right)^{1/2} \leq \exp\{1\},
$$

where $0 < c_1 < \frac{1 - \exp\{-2\}}{2}$. It remains to note that $\zeta^T \exp\{V\}\zeta - 1$ has the same distribution as any of the random variables f_{ξ^i}, $i = 1, \ldots, N$.

f) We observe that the space $(\mathbb{R}, \|\cdot\|_1)$ has a regularity constant of 1. Due to Theorem A.1 (we apply Theorem A.1 with $\chi = 1$ and $\sigma_i = 2/c_1$ for any $i = 1, \ldots, N$), there exists a constant $c_4 > 0$ such that:

$$
\mathbb{E}_\xi \exp\left\{ \frac{c_1 \sqrt{N} \left| f_\xi \right|}{2c_4} \right\} = \mathbb{E}_\xi \exp\left\{ \frac{c_1 \left| \sum_{i=1}^{N} f_{\xi^i} \right|}{2c_4 \sqrt{N}} \right\} \leq \exp\{1\}.
$$

It remains to choose $c \geq \max\{2\exp\{1\}/c_1, c_3/c_2, 2c_4/c_1\}$. ■

We are ready to prove Proposition 3.1:

Proof
Let $V \in \mathcal{S}_n$ and recall definitions (52)-(54). Consider the random element:

$$\beta_\xi := d_\xi - f_\xi h(V).$$

Lemma B.2 implies $\mathbb{E}_\xi \beta_\xi = 0$. Moreover, by the same lemma, there exists a constant $c_1 > 0$ such that:

$$\begin{aligned}
\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,\|\beta_\xi\|_{x,*}}{2c_1\mathcal{L}\sqrt{\kappa}}\right\} &\leq \mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,\|d_\xi\|_{x,*}}{2c_1\mathcal{L}\sqrt{\kappa}}\right\} \exp\left\{\frac{\sqrt{N}\,\|f_\xi h(V)\|_{x,*}}{2c_1\mathcal{L}\sqrt{\kappa}}\right\} \\
&\leq \left[\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,\|d_\xi\|_{x,*}}{c_1\mathcal{L}\sqrt{\kappa}}\right\}\right]^{1/2} \left[\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,\|f_\xi h(V)\|_{x,*}}{c_1\mathcal{L}\sqrt{\kappa}}\right\}\right]^{1/2} \\
&\leq \exp\left\{\frac{1}{2}\right\} \left[\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,|f_\xi|}{c_1\sqrt{\kappa}}\right\}\right]^{1/2} \\
&\leq \exp\{1\}, \qquad (57)
\end{aligned}$$

where we use Hölder's inequality and the facts $\|h(V)\|_{x,*} \leq \mathcal{L}$ and $\kappa \geq 1$. Let

$$\gamma_\xi := h_\xi(V) - \beta_\xi - h(V).$$

As $\theta_\xi(V) = f_\xi + 1$ and $\mathcal{A}^*\left(G_\xi(V)\right) = d_\xi + h(V)$, it holds that:

$$h_\xi(V) = \frac{\mathcal{A}^*\left(G_\xi(V)\right)}{\theta_\xi(V)} = \frac{d_\xi + h(V)}{f_\xi + 1}.$$

We obtain:

$$\gamma_\xi = \frac{d_\xi + h(V)}{f_\xi + 1} - h(V) - d_\xi + f_\xi h(V) = \frac{f_\xi^2 h(V) - d_\xi f_\xi}{1 + f_\xi}.$$

Consider the sets:

$$\Pi := \{\xi :\ |f_\xi| \leq 1/2\} \qquad \text{and} \qquad \hat{\Pi} := \mathbb{R}^{n\times N} \setminus \Pi.$$

When $\xi \in \Pi$, we have:

$$\begin{aligned}
\|\gamma_\xi\|_{x,*} &= \frac{\left\|f_\xi^2 h(V) - d_\xi f_\xi\right\|_{x,*}}{|1 + f_\xi|} \leq 2\left\|f_\xi^2 h(V) - d_\xi f_\xi\right\|_{x,*} \\
&\leq 2\left(|f_\xi|^2\,\mathcal{L} + |f_\xi|\,\|d_\xi\|_{x,*}\right) \leq |f_\xi|\,\mathcal{L} + \|d_\xi\|_{x,*}. \qquad (58)
\end{aligned}$$

More generally, it holds that:

$$\begin{aligned}
\|\gamma_\xi\|_{x,*} &\leq \|h_\xi(V)\|_{x,*} + \|\beta_\xi\|_{x,*} + \|h(V)\|_{x,*} \\
&\leq \|h_\xi(V)\|_{x,*} + \|d_\xi\|_{x,*} + \|f_\xi h(V)\|_{x,*} + \|h(V)\|_{x,*} \\
&= \|h_\xi(V)\|_{x,*} + \|d_\xi\|_{x,*} + |f_\xi|\,\|h(V)\|_{x,*} + \|h(V)\|_{x,*}
\end{aligned}$$

$$\leq (2 + |f_\xi|)\,\mathcal{L} + \|d_\xi\|_{x,*}\,, \tag{59}$$

which is due to $\|h(V)\|_{x,*} \leq \mathcal{L}$ and to the following inequality:

$$\|h_\xi(V)\|_{x,*} \leq \mathcal{L}\frac{\left\|\sum_{i=1}^N \left(\exp\{V/2\}\xi^i\right)\left(\exp\{V/2\}\xi^i\right)^T\right\|_Y}{\sum_{i=1}^N [\xi^i]^T \exp\{V\}\xi^i} = \mathcal{L}.$$

Furthermore, denoting by $\mathbb{P}$ the probability measure of the random matrix ξ:

$$\begin{aligned}
\mathbb{P}\left[\hat{\Pi}\right] &= \mathbb{P}\left[|f_\xi| > 1/2\right] = \mathbb{P}\left[\exp\left\{\frac{\sqrt{N}\,|f_\xi|}{c_1}\right\} > \exp\left\{\frac{\sqrt{N}}{2c_1}\right\}\right] \\
&< \exp\left\{-\frac{\sqrt{N}}{2c_1}\right\}\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,|f_\xi|}{c_1}\right\} \\
&\leq \exp\left\{-\frac{\sqrt{N}}{2c_1}\right\}\exp\{1\}\,,
\end{aligned} \tag{60}$$

where the inequalities follow from Markov's inequality and from statement e) of Lemma B.2, respectively. Choose $c_2 \geq 4c_1$ and observe:

$$\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,\|\gamma_\xi\|_{x,*}}{3c_2\mathcal{L}\sqrt{\kappa}}\right\} = \int_\Pi \exp\left\{\frac{\sqrt{N}\,\|\gamma_\xi\|_{x,*}}{3c_2\mathcal{L}\sqrt{\kappa}}\right\} d\mathbb{P}(\xi) + \int_{\hat{\Pi}} \exp\left\{\frac{\sqrt{N}\,\|\gamma_\xi\|_{x,*}}{3c_2\mathcal{L}\sqrt{\kappa}}\right\} d\mathbb{P}(\xi).$$

By (58), Cauchy-Schwarz inequality, and Lemma B.2, we obtain:

$$\begin{aligned}
\int_\Pi \exp\left\{\frac{\sqrt{N}\,\|\gamma_\xi\|_{x,*}}{3c_2\mathcal{L}\sqrt{\kappa}}\right\} d\mathbb{P}(\xi) &\leq \int_\Pi \exp\left\{\frac{\sqrt{N}\left(|f_\xi|\,\mathcal{L} + \|d_\xi\|_{x,*}\right)}{3c_2\mathcal{L}\sqrt{\kappa}}\right\} d\mathbb{P}(\xi) \\
&\leq 1 \cdot \left[\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,|f_\xi|}{c_2\sqrt{\kappa}}\right\}\right]^{1/3}\left[\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,\|d_\xi\|_{x,*}}{c_2\mathcal{L}\sqrt{\kappa}}\right\}\right]^{1/3} \\
&\leq \exp\left\{\frac{1}{3}\right\}\exp\left\{\frac{1}{3}\right\} = \exp\left\{\frac{2}{3}\right\}.
\end{aligned}$$

Additionally, by (59), we have:

$$\int_{\hat{\Pi}} \exp\left\{\frac{\sqrt{N}\,\|\gamma_\xi\|_{x,*}}{3c_2\mathcal{L}\sqrt{\kappa}}\right\} d\mathbb{P}(\xi) \leq \int_{\hat{\Pi}} \exp\left\{\frac{\sqrt{N}\left(2\mathcal{L} + |f_\xi|\,\mathcal{L} + \|d_\xi\|_{x,*}\right)}{3c_2\mathcal{L}\sqrt{\kappa}}\right\} d\mathbb{P}(\xi).$$

Let

$$A := \int_{\hat{\Pi}} \exp\left\{\frac{\sqrt{N}\left(|f_\xi|\,\mathcal{L} + \|d_\xi\|_{x,*}\right)}{3c_2\mathcal{L}\sqrt{\kappa}}\right\} d\mathbb{P}(\xi).$$

Cauchy-Schwarz inequality, bound (60), and Lemma B.2 imply:

$$A \leq \left[\int_{\hat{\Pi}} 1 d\mathbb{P}(\xi)\right]^{1/3}\left[\int_{\hat{\Pi}} \exp\left\{\frac{\sqrt{N}\,|f_\xi|}{c_2\sqrt{\kappa}}\right\} d\mathbb{P}(\xi)\right]^{1/3}\left[\int_{\hat{\Pi}} \exp\left\{\frac{\sqrt{N}\,\|d_\xi\|_{x,*}}{c_2\mathcal{L}\sqrt{\kappa}}\right\} d\mathbb{P}(\xi)\right]^{1/3}$$

$$
\begin{aligned}
&\leq \left[\mathbb{P}\left[\hat{\Pi}\right]\right]^{1/3}\left[\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,|f_\xi|}{c_2\sqrt{\kappa}}\right\}\right]^{1/3}\left[\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,\|d_\xi\|_{x,*}}{c_2\mathcal{L}\sqrt{\kappa}}\right\}\right]^{1/3}\\
&\leq \left[\mathbb{P}\left[\hat{\Pi}\right]\right]^{1/3}\left[\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,|f_\xi|}{c_2}\right\}\right]^{1/3}\left[\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,\|d_\xi\|_{x,*}}{c_2\mathcal{L}\sqrt{\kappa}}\right\}\right]^{1/3}\\
&\leq \exp\left\{-\frac{\sqrt{N}}{6c_1}\right\}\exp\left\{\frac{2}{3}\right\}.
\end{aligned}
$$

As $c_2 \geq 4c_1$, we obtain:

$$
\begin{aligned}
\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,\|\gamma_\xi\|_{x,*}}{3c_2\mathcal{L}\sqrt{\kappa}}\right\} &\leq \exp\left\{\frac{2}{3}\right\} + \exp\left\{\frac{2\sqrt{N}}{3c_2\sqrt{\kappa}}\right\}\exp\left\{-\frac{\sqrt{N}}{6c_1}\right\}\exp\left\{\frac{2}{3}\right\}\\
&\leq \exp\left\{\frac{2}{3}\right\} + \exp\left\{\frac{2\sqrt{N}}{3c_2} - \frac{\sqrt{N}}{6c_1}\right\}\exp\left\{\frac{2}{3}\right\} \leq 2\exp\left\{\frac{2}{3}\right\}. \qquad (61)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,\|h_\xi(V)-h(V)\|_{x,*}}{6c_2\mathcal{L}\sqrt{\kappa}}\right\} &= \mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,\|\beta_\xi+\gamma_\xi\|_{x,*}}{6c_2\mathcal{L}\sqrt{\kappa}}\right\}\\
&\leq \left[\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,\|\beta_\xi\|_{x,*}}{3c_2\mathcal{L}\sqrt{\kappa}}\right\}\right]^{1/2}\left[\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,\|\gamma_\xi\|_{x,*}}{3c_2\mathcal{L}\sqrt{\kappa}}\right\}\right]^{1/2}\\
&\leq \left[\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,\|\beta_\xi\|_{x,*}}{c_1\mathcal{L}\sqrt{\kappa}}\right\}\right]^{1/2}\left[\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,\|\gamma_\xi\|_{x,*}}{3c_2\mathcal{L}\sqrt{\kappa}}\right\}\right]^{1/2}\\
&\leq \sqrt{2}\exp\left\{\frac{5}{6}\right\}, \qquad (62)
\end{aligned}
$$

where the inequalities are due to Hölder's inequality, the fact that $c_2 \geq c_1$, bound (57), and inequality (61), respectively.

It remains to find an appropriate bound on the norm of $h(V) - \mathbb{E}_\xi h_\xi(V)$. Recall that $\mathbb{E}_\xi \beta_\xi = 0$. Therefore,

$$
\begin{aligned}
\|\mathbb{E}_\xi h_\xi(V) - h(V)\|_{x,*} &= \|\mathbb{E}_\xi \gamma_\xi\|_{x,*} \leq \mathbb{E}_\xi \|\gamma_\xi\|_{x,*}\\
&= \int_\Pi \|\gamma_\xi\|_{x,*}\, d\mathbb{P}(\xi) + \int_{\hat{\Pi}} \|\gamma_\xi\|_{x,*}\, d\mathbb{P}(\xi)\\
&\leq 2\int_\Pi |f_\xi|^2\,\mathcal{L} + |f_\xi|\,\|d_\xi\|_{x,*}\, d\mathbb{P}(\xi) + \int_{\hat{\Pi}} \|\gamma_\xi\|_{x,*}\, d\mathbb{P}(\xi),
\end{aligned}
$$

where the concluding inequality follows from (58). As $2\exp(x) \geq x^2$ for any $x \geq 0$, we obtain by Lemma B.2:

$$
2\mathcal{L}\int_\Pi |f_\xi|^2\, d\mathbb{P}(\xi) \leq 2\mathcal{L}\mathbb{E}_\xi |f_\xi|^2 \leq \frac{4c_1^2\mathcal{L}}{N}\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\,|f_\xi|}{c_1}\right\} \leq \frac{4\exp\{1\}c_1^2\mathcal{L}}{N}.
$$

Furthermore, the same arguments imply:

$$\begin{aligned}
2\int_{\Pi} |f_\xi| \, \|d_\xi\|_{x,*} \, d\mathbb{P}(\xi) &\leq \int_{\Pi} \sqrt{\kappa}\mathcal{L} \, |f_\xi|^2 \, d\mathbb{P}(\xi) + \int_{\Pi} \frac{\|d_\xi\|_{x,*}^2}{\sqrt{\kappa}\mathcal{L}} d\mathbb{P}(\xi) \\
&\leq \mathbb{E}_\xi \left\{ \sqrt{\kappa}\mathcal{L} \, |f_\xi|^2 \right\} + \mathbb{E}_\xi \left\{ \frac{\|d_\xi\|_{x,*}^2}{\mathcal{L}\sqrt{\kappa}} \right\} \\
&\leq \frac{2c_1^2\mathcal{L}\sqrt{\kappa}}{N} \left(\mathbb{E}_\xi \left\{ \exp\left\{ \frac{\sqrt{N} \, |f_\xi|}{c_1} \right\} \right\} + \mathbb{E}_\xi \left\{ \exp\left\{ \frac{\sqrt{N} \, \|d_\xi\|_{x,*}}{c_1\mathcal{L}\sqrt{\kappa}} \right\} \right\} \right) \\
&\leq \frac{4\exp\{1\}c_1^2\mathcal{L}\sqrt{\kappa}}{N}.
\end{aligned}$$

Finally,

$$\begin{aligned}
\int_{\hat{\Pi}} \|\gamma_\xi\|_{x,*} \, d\mathbb{P}(\xi) &\leq \left[\mathbb{P}\left[\hat{\Pi}\right] \right]^{1/2} \left[\mathbb{E}_\xi \|\gamma_\xi\|_{x,*}^2 \right]^{1/2} \\
&\leq \exp\left\{ \frac{1}{2} - \frac{\sqrt{N}}{4c_1} \right\} \left[\mathbb{E}_\xi \|\gamma_\xi\|_{x,*}^2 \right]^{1/2} \\
&\leq \exp\left\{ \frac{1}{2} - \frac{\sqrt{N}}{4c_1} \right\} \left[\frac{18c_2^2\mathcal{L}^2\kappa}{N} \mathbb{E}_\xi \exp\left\{ \frac{\sqrt{N} \, \|\gamma_\xi\|_{x,*}}{3c_2\mathcal{L}\sqrt{\kappa}} \right\} \right]^{1/2} \\
&\leq \frac{6\exp\left\{\frac{5}{6}\right\} c_2\mathcal{L}\sqrt{\kappa}}{\sqrt{N}} \exp\left\{ -\frac{\sqrt{N}}{4c_1} \right\} \\
&\leq \frac{6\exp\left\{\frac{5}{6}\right\} c_2\mathcal{L}\sqrt{\kappa}}{\sqrt{N}} \exp\left\{ -\frac{\sqrt{N}}{c_2} \right\} \\
&\leq \frac{6\exp\left\{\frac{5}{6}\right\} c_2^2\mathcal{L}\sqrt{\kappa}}{N},
\end{aligned}$$

where the inequalities hold due to Hölder's inequality, the fact that $\|\gamma_\xi\|_{x,*}^2$ is nonnegative for any $\xi \in \mathbb{R}^{n\times N}$, bound (60), the fact that $2\exp(x) \geq x^2$ for any $x \geq 0$, inequality (61), and the assumption $c_2 \geq 4c_1$, respectively. We obtain:

$$\begin{aligned}
\|\mathbb{E}_\xi h_\xi(V) - h(V)\|_{x,*} &\leq \frac{4\exp\{1\}c_1^2\mathcal{L}}{N} + \frac{4\exp\{1\}c_1^2\mathcal{L}\sqrt{\kappa}}{N} + \frac{6\exp\{1\}c_2^2\mathcal{L}\sqrt{\kappa}}{N} \\
&\leq \frac{8\exp\{1\}c_1^2\mathcal{L}\sqrt{\kappa}}{N} + \frac{6\exp\{1\}c_2^2\mathcal{L}\sqrt{\kappa}}{N} \\
&\leq \frac{\exp\{1\}c_2^2\mathcal{L}\sqrt{\kappa}}{2N} + \frac{6\exp\{1\}c_2^2\mathcal{L}\sqrt{\kappa}}{N} \\
&= \frac{13\exp\{1\}c_2^2\mathcal{L}\sqrt{\kappa}}{2N},
\end{aligned}$$

which proves statement a). The above inequality ensures together with bound (62):

$$\mathbb{E}_\xi \exp\left\{ \frac{\sqrt{N} \, \|h_\xi(V) - \mathbb{E}_\xi h_\xi(V)\|_{x,*}}{6c_2\mathcal{L}\sqrt{\kappa}} \right\} \leq \sqrt{2}\exp\left\{\frac{5}{6}\right\} \exp\left\{ \frac{13\exp\{1\}c_2}{12\sqrt{N}} \right\}$$

$$\leq \sqrt{2}\exp\left\{\frac{5}{6}+\frac{13\exp\{1\}c_2}{12}\right\}.$$

Let:

$$c_3 := \frac{5}{6}+\frac{\ln(2)}{2}+\frac{13\exp\{1\}c_2}{12}.$$

By Jensen's inequality, we conclude that:

$$\mathbb{E}_\xi \exp\left\{\frac{\sqrt{N}\left\|h_\xi(V)-\mathbb{E}_\xi h_\xi(V)\right\|_{x,*}}{6c_2c_3\mathcal{L}\sqrt{\kappa}}\right\} \leq \exp\{1\}.$$

■