Maximizing concave piecewise affine functions on the unitary group

Stephane Gaubert, Zheng Qu, Srinivas Sridharan

MAXIMIZING CONCAVE PIECEWISE AFFINE FUNCTIONS ON THE UNITARY GROUP

STÉPHANE GAUBERT, ZHENG QU, AND SRINIVAS SRIDHARAN

Abstract. We show that a convex relaxation, introduced by Sridharan, McEneaney, Gu and James to approximate the value function of an optimal control problem arising from quantum gate synthesis, is exact. This relaxation applies to the maximization of a class of concave piecewise affine functions over the unitary group.

1. Introduction

1.1. **Main result.** The main object of this paper is to show the equivalence between the following non-convex optimization problem

$$\max_{X\in U(n)} \min_{1\leqslant i\leqslant m} \langle P_i - P_0, X\rangle + c_i - c_0 \tag{1}$$

and its convex relaxation:

$$\max_{X\in B(n)} \min_{1\leqslant i\leqslant m} \langle P_i - P_0, X\rangle + c_i - c_0\,. \tag{2}$$

Here, $U(n)$ denotes the unitary group of degree n, i.e., the subset of unitary matrices of the space $M_n(\mathbb{C})$ of $n\times n$ complex matrices, $B(n)$ denotes the nth unitary ball, i.e., the set of matrices of $M_n(\mathbb{C})$ with singular values at most 1; $\langle\cdot,\cdot\rangle$ denotes the canonical inner product of $M_n(\mathbb{C})$, thought of as a real Hilbert space, so that $\langle X_1, X_2\rangle = \mathrm{Re}(\mathrm{trace}(X_1^* X_2))$, where X^* denotes the conjugate transpose of a matrix X, and finally

$$P_0,\dots,P_m \in U(n), \qquad c_0,\dots,c_1 \in \mathbb{R}\,,$$

are given. Since $B(n)$ is the convex hull of $U(n)$, Problem 2 is a convex optimization problem. Moreover, since the convex set $B(n)$ can be represented in terms of elementary matrix inequalities, Problem 2, and so, Problem 1, can be solved efficiently by convex programming techniques. We refer the reader to [BTN01] for more background on complexity results in convex programming, including general polynomial time complexity bounds which apply to the formulation (2).

The second author carried out parts of this work when she was with INRIA and CMAP, École Polytechnique, CNRS and subsequently with the School of Mathematics of University of Edinburgh. The third author carried out part of this work when he was with UMA, ENSTA, Palaiseau, France.

The first two authors were partially supported by the PGMO Program of FMJH and EDF, and by the program "Ingénierie Numérique & Sécurité" of the French National Agency of Research, project "MALTHY", number ANR-13-INSE-0003.

1.2. **Motivation from Optimal Control.** The present contribution arose from the *pruning step* of the *curse of dimensionality-free* numerical method, introduced by McEneaney [McE07], for solving first order Hamilton-Jacobi-Bellman Partial Differential Equations (HJB PDE). This method applies to a special class of optimal control problems in which the Hamiltonian is given or approximated by a pointwise supremum of a finite number of elementary Hamiltonians. Such Hamiltonians arise naturally when the control space comprises a discrete component, for example in switched systems. The value function V of the optimal control problem is approximated by the infimum (or supremum if the objective is maximized) of a finite number of elementary basis functions $\{\phi_0, \dots, \phi_m\}$, which are typically linear or quadratic forms:

$$V \simeq \min_{0 \leqslant i \leqslant m} \phi_i \,. \tag{3}$$

Theoretical estimates show that, to reach an approximation of a prescribed accuracy, the computational complexity grows polynomially with the space dimension [MK10, Qu14]. However, the number of basis functions is multiplied by the number of elementary Hamiltonians at each propagation step, the so called *curse of complexity*. Therefore in order to attenuate this curse of complexity in practice, a pruning operation is introduced. It achieves this improvement in efficiency by removing, at each step, a certain number of basis functions less useful than the others.

The pruning problem has received much attention [MDG08, McE09, SGJM10, GMQ11, SMGJ14, GQS14]. Actually, the experiments in [GMQ11] indicate that the main part of the execution time of the curse of dimensionality-free method is spent in pruning. Therefore, pruning is both a critical step and a bottleneck stage – necessitating the development of fast exact or approximate algorithms. The pruning problem appears as well in other approximate dynamic programming methods in which the value function is approximated by a maximum of affine functions, like the dual dynamic programming method of Pereira and Pinto [PP91] and its extensions, note in particular the discussion by Shapiro [Sha11].

Pruning was shown in [GMQ11] to be a continuous version of the facility location problem. Most pruning methods used in practice rely on calculating the *importance metric* δ_j of each basis function ϕ_j, with respect to a bounded state space $\mathcal{X}$. This metric is defined by

$$\delta_j := \max_{x \in \mathcal{X}} \min_{i \neq j} \phi_i(x) - \phi_j(x) \,. \tag{4}$$

The importance metric δ_j thus measures the maximal loss caused by removing the basis function ϕ_j. When the state space $\mathcal{X}$ is not bounded, δ_j is replaced by a variant, involving a normalization, see [MDG08]. In some sense the higher δ_j is, the more useful the function ϕ_j is for the approximation. In particular, if $\delta_j < 0$, then, the basis function j can be deleted without changing the approximation of V in (3). Once we get all the values $\{\delta_0, \dots, \delta_m\}$, the simplest algorithm consists of removing those basis functions with lowest importance metrics, as suggested in [MDG08]. One may also apply an iterative greedy algorithm, in which the importance metric is recomputed dynamically after each removal of basis function, or use the importance metric to construct a discretized problem to which combinatorial facility location algorithms can be applied [GMQ11].

As mentioned before, the curse of dimensionality-free method can be used to solve switched optimal control problems. In particular, it has been recently applied to the optimal quantum circuit synthesis, following a work by Gu, James, McEneaney and Sridharan, see [SGJM10, SMGJ14] and also [GQS14]. The related optimal control problem is to find a least path-length trajectory on the special unitary group $SU(n)$. For this particular application, the basis functions are chosen to be affine functions defined on the space $M_n(\mathbb{C})$:

$$\phi_i(X) = \langle P_i, X\rangle + c_i, \;\; \forall X \in M_n(\mathbb{C}), \;\; i = 0, \dots, m,$$

with $P_0, \dots, P_m \in U(n)$ and $c_0, \dots, c_m \in \mathbb{R}$. Then calculating the importance metric of ϕ_0 reduces to solving the following optimization problem:

$$\delta_0 := \max_{X \in SU(n)} \min_{1 \leqslant i \leqslant m} \langle P_i - P_0, X\rangle + c_i - c_0\,. \tag{5}$$

The importance metric of other functions $\{\delta_1, \dots, \delta_m\}$ can be obtained by solving similar programs (in which the index 0 is replaced by an index j). A key difficulty here is that the state space (the special unitary group $SU(n)$) is non-convex. Hence, in order to compute efficiently an importance metric, we need to perform a relaxation. In [SMGJ14], the special unitary group $SU(n)$ is first relaxed into the unitary group $U(n)$, then replaced by the unitary ball $B(n)$. Then, the value δ_0 is approximated by the optimal value $\bar{\delta}_0$ of the convex optimization problem (2). Since $\bar{\delta}_0 \geqslant \delta_0$, the condition $\bar{\delta}_0 < 0$ guarantees that the function ϕ_0 is not active in the infimum in (3) and can therefore be pruned. More generally, any relaxation leading to an importance metric $\delta_0' \geqslant \delta_0$ yields a sound pruning method. The closer the relaxed importance metric is from the true one, the more efficient is the pruning.

1.3. **Discussion of the results.** One question which arises from [SMGJ14] is whether the relaxation of (1) into (2) is exact. Our main result, Theorem 2.1, shows that the optimal solution of (2) contains always a unitary matrix. Thus, Program (1) is equivalent to its convex relaxation (2). This property is somehow surprising, as the maximum of a concave function over a compact convex set is generally *not* attained at an extreme point of a set. What makes this property valid for the special non-convex optimization problem (1) is that the gradients of the affine constituents of the objective function are differences $P_i - P_0$, $1 \leqslant i \leqslant m$, where each P_i is unitary, and P_0 is a *fixed* unitary matrix.

The proof of Theorem 2.1 exploits the strict convexity of the Frobenius norm. The proof also uses a regularization argument (replacing the piecewise affine objective function by a smooth log-exp type function), together with algebraic properties of the tangent cone at a given point of $B(n)$. The analogous result holds, with obvious changes, when $SU(n)$ is replaced by the nth real special orthogonal group $SO(n)$. The latter also arises in optimal estimation and filtering, in particular in optimal attitude estimation [SM13], and so, the present results apply as well to this case.

To get more insight on the assumptions which govern Theorem 2.1, we finally present a variation of this theorem, showing that a similar result of independent interest holds (Proposition 2.2) for the maximization over a sphere of a concave piecewise affine functions the affine constituents of which have unitary gradients.

We finally note that Theorem 2.1 has been announced, without proof, in our recent conference article [GQS14]. We showed there that this result can be combined with scalable bundle-type non-differentiable optimization methods to speed up the pruning procedure of [SMGJ14].

1.4. **Summary of notation.** It is convenient to list here the main notation used throughout the paper. As mentioned above, $M_n(\mathbb{C})$ denotes the space of $n \times n$ complex matrices. For $X \in M_n(\mathbb{C})$, X^* denotes its conjugate transpose. The space $M_n(\mathbb{C})$ is considered as a real Hilbert space endowed with the inner product $\langle \cdot, \cdot \rangle$ given by $\langle X_1, X_2 \rangle = \mathrm{Re}(\mathrm{trace}(X_1^* X_2)),\ \forall X_1, X_2 \in M_n(\mathbb{C})$. We denote by I_n the $n \times n$ identity matrix, so that the nth unitary group is given by $U(n) = \{X \in M_n(\mathbb{C}) \mid XX^* = I_n\}$. The space of $n \times n$ positive semidefinite (resp. positive definite) matrices is denoted by S_n^+ (resp. $\hat{\mathrm{S}}_n^+$). For two Hermitian matrices $A, B \in M_n(\mathbb{C})$, we write $A \succcurlyeq B$ (resp. $A \succ B$) if $A - B \in \mathrm{S}_n^+$ (resp. $A - B \in \hat{\mathrm{S}}_n^+$). Hence, the unitary ball $B(n)$ is given by $B(n) = \{X \in M_n(\mathbb{C}) \mid XX^* \preccurlyeq I_n\}$.

2. Maximizing a class of concave functions over the unitary group

In the following we show that (2) is an exact relaxation of (1). Our main result is Theorem 2.1. We begin with a few lemmas.

Lemma 2.1. *Let $P_0, \ldots, P_m$ be $n \times n$ unitary matrices and let $X \in M_n(\mathbb{C})$. If XP_0 is a point in the relative interior of the convex hull of $\{XP_i\}_{i=1,\ldots,m}$, then $XP_i = XP_0$ for all $i = 1, \ldots, m$.*

Proof. If XP_0 belongs to the relative interior of the convex hull of $\{XP_i\}_{i=1,\ldots,m}$, then there are $\alpha_1, \ldots, \alpha_m > 0$ such that $\sum_{i=1}^m \alpha_i = 1,\ \sum_{i=1}^m \alpha_i XP_i = XP_0$. The Frobenius norm of XP_i equals to that of X for all $i = 0, \ldots, m$. By the strict convexity of the Frobenius norm, we deduce that $XP_i = XP_0$ holds for all $i = 1, \ldots, m$. □

The *tangent cone* to $B(n)$ at $X \in B(n)$ is defined by

$$T_{B(n)}(X) = \mathrm{cl}\{\lambda(Z - X) : \lambda \geqslant 0, Z \in B(n)\},$$

where cl denotes the closure of a set, see [RW98, p. 204].

Lemma 2.2. *Let $1 \leqslant k \leqslant n$ and Σ be a diagonal matrix with non-negative real diagonal entries $(\lambda_1, \cdots, \lambda_n)$ such that $\lambda_j = 1$ for all $j = 1, \ldots, k-1$ and $\lambda_j < 1$ for all $j = k, \cdots, n$. Then*

$$\left\{ \begin{pmatrix} X_1 & X_2 \\ X_3 & X_4 \end{pmatrix} \in M_n(\mathbb{C}) : -X_1 - X_1^* \in \hat{\mathrm{S}}_{k-1}^+ \right\} \subset T_{B(n)}(\Sigma).$$

Proof. If $k = 1$ then it is clear that Σ is an interior point of $B(n)$ and the tangent cone at Σ is the whole space $M_n(\mathbb{C})$. We next consider the case when $1 < k \leqslant n$. For ease of proof, we write Σ as a block matrix $\Sigma = \begin{pmatrix} I_{k-1} & 0 \\ 0 & \Sigma_4 \end{pmatrix}$,where Σ_4 is a diagonal matrix such that $I_{n-k+1} - \Sigma_4 \Sigma_4^* \succ 0$. Let $X = \begin{pmatrix} X_1 & X_2 \\ X_3 & X_4 \end{pmatrix} \in M_n(\mathbb{C})$, such that $-X_1 - X_1^* \in \hat{\mathrm{S}}_{k-1}^+$. For the simplicity of notation, let $H_1 \in M_{k-1}(\mathbb{C})$, $H_4 \in M_{n-k+1}(\mathbb{C})$ and $H_2 \in \mathbb{C}^{(k-1)\times(n-k+1)}$ such that $XX^* = \begin{pmatrix} H_1 & H_2 \\ H_2^* & H_4 \end{pmatrix}$.Take any scalar $\Delta > 0$. Then,

$$\begin{aligned}(\Sigma + \Delta X)(\Sigma + \Delta X)^* - I_n &= \Sigma\Sigma^* + \Delta(X\Sigma^* + \Sigma X^*) - I_n + \Delta^2 XX^* \\ &= \begin{pmatrix} \Delta(X_1 + X_1^*) & \Delta(X_2\Sigma_4^* + X_3^*) \\ \Delta(\Sigma_4 X_2^* + X_3) & \Sigma_4\Sigma_4^* - I_{n-k+1} + \Delta(\Sigma_4 X_4^* + X_4\Sigma_4^*) \end{pmatrix} \\ &\quad + \begin{pmatrix} \Delta^2 H_1 & \Delta^2 H_2 \\ \Delta^2 H_2^* & \Delta^2 H_4 \end{pmatrix}\end{aligned}$$

By Schur's complement lemma, we know that $(\Sigma + \Delta X)(\Sigma + \Delta X)^* \prec I_n$ if and only if $X_1 + X_1^* + \Delta H_1 \prec 0$ and

$$I_{n-k+1} - \Sigma_4\Sigma_4^* - \Delta(\Sigma_4 X_4^* + X_4\Sigma_4^*) - \Delta^2 H_4 + \Delta(\Sigma_4 X_2^* + X_3 + \Delta H_2^*)(X_1 + X_1^* + \Delta H_1)^{-1}(X_2\Sigma_4^* + X_3^* + \Delta H_2) \succ 0\,.$$

Since $X_1 + X_1^* \prec 0$ and $I_{n-k+1} - \Sigma_4\Sigma_4^* \succ 0$, there is $\Delta > 0$ such that the latter two inequalities hold thus $(\Sigma + \Delta X)(\Sigma + \Delta X)^* \prec I_n$. Hence, X is in the tangent cone of $B(n)$. □

In the sequel, let $P_0, \ldots, P_m$ be $n \times n$ unitary matrices, and $c_0, \ldots, c_m$ be real numbers. For all $\beta > 0$, define $\phi_\beta : M_n(\mathbb{C}) \to \mathbb{R}$ by:

$$\phi_\beta(X) = -\beta^{-1} \log\Big(\sum_{i=1}^m e^{-\beta(\langle P_i - P_0, X\rangle + c_i - c_0)}\Big), \qquad \forall X \in M_n(\mathbb{C})\,. \tag{6}$$

Lemma 2.3. *There is a constant $K > 0$ independent of $\beta > 0$ such that the map $\phi_\beta : M_n(\mathbb{C}) \to \mathbb{R}$ is K-Lipschitz continuous.*

Proof. Let $\beta > 0$. For all $X, Y \in M_n(\mathbb{C})$, the differential map of ϕ_β at point X, $D\phi_\beta(X)$ evaluated at Y, satisfies:

$$D\phi_\beta(X) \circ Y = \sum_{i=1}^m \alpha_i \langle P_i - P_0, Y\rangle\,,$$

where

$$\alpha_i = \Big(\sum_{j=1}^m e^{-\beta(\langle P_j - P_0, X\rangle + c_j - c_0)}\Big)^{-1} e^{-\beta(\langle P_i - P_0, X\rangle + c_i - c_0)},\ \ i = 1, \ldots, m.$$

Thus $\alpha_1, \ldots, \alpha_m > 0$ and $\sum_{i=1}^m \alpha_i = 1$. Hence, for all $X, Y \in M_n(\mathbb{C})$, by the Cauchy-Schwarz inequality,

$$|D\phi_\beta(X) \circ Y| \leqslant \sum_{i=1}^m \alpha_i \|P_i - P_0\|\|Y\| \leqslant \max_{1 \leqslant i \leqslant m} \|P_i - P_0\|\|Y\|\,,$$

where $\|\cdot\|$ denotes the Euclidean norm associated to the Frobenius scalar product $\langle \cdot, \cdot \rangle$. It follows that the function ϕ_β is K-Lipschitz with respect to the norm $\|\cdot\|$ for all $\beta > 0$, with $K = \max_{1 \leqslant i \leqslant m} \|P_i - P_0\|$. □

Proposition 2.1. *For every $\beta > 0$, the set of optimal solutions of the optimization problem*

$$\max_{X \in B(n)} \phi_\beta(X) \tag{7}$$

contains a unitary matrix.

Proof. Let $U_0 \in B(n)$ be an optimal solution of (7). Suppose that U_0 is not unitary. Consider the SVD decomposition of U_0 given by $U_0 = V_1 \Sigma V_2$, where Σ is a diagonal matrix with non-negative real diagonal entries $(\lambda_1, \ldots, \lambda_n)$, listed in non-increasing order. Let $k \in \{1, \ldots, n\}$ such that $\lambda_i = 1$ for all $i = 1, \ldots, k-1$ and $\lambda_j < 1$ for all $j = k, \ldots, n$. Then Σ is an optimal solution of the following optimization problem:

$$\max_{X \in B(n)} \phi_\beta(V_1 X V_2)\,. \tag{8}$$

The first-order optimality condition [RW98, p.207] implies that

$$D\phi_\beta(V_1\Sigma V_2)\circ(V_1YV_2)\leqslant 0,\quad \forall Y\in T_{B(n)}(\Sigma)\,. \tag{9}$$

We have:

$$D\phi_\beta(V_1\Sigma V_2)\circ(V_1YV_2)=\sum_{i=1}^m\alpha_i\langle P_i-P_0,V_1YV_2\rangle\,,$$

where $\alpha_1,\ldots,\alpha_m>0$ and $\sum_{i=1}^m\alpha_i=1$. Therefore,

$$D\phi_\beta(V_1\Sigma V_2)\circ(V_1YV_2)=\langle\sum_{i=1}^m\alpha_i(V_1^*P_iV_2^*-V_1^*P_0V_2^*),Y\rangle$$

By the first-order optimality condition (9) and Lemma 2.2, we deduce that

$$\langle\sum_{i=1}^m\alpha_i(V_1^*P_iV_2^*-V_1^*P_0V_2^*),X\rangle\leqslant 0\,,$$

for all $X=\left(\begin{smallmatrix}X_1 & X_2\\ X_3 & X_4\end{smallmatrix}\right)\in M_n(\mathbb{C})$ such that $-X_1-X_1^*\in\hat{\mathrm{S}}_{k-1}^+$. Hence,

$$\sum_{i=1}^m\alpha_i(V_1^*P_iV_2^*-V_1^*P_0V_2^*)=\begin{pmatrix}Z & 0\\ 0 & 0\end{pmatrix}$$

for some $Z\in M_{k-1}(\mathbb{C})$. Therefore,

$$(I_n-\Sigma)\sum_{i=1}^m\alpha_i(V_1^*P_iV_2^*-V_1^*P_0V_2^*)=0\,.$$

By Lemma 2.1, we know that

$$(I_n-\Sigma)V_1^*P_iV_2^*=(I_n-\Sigma)V_1^*P_0V_2^*,\ \ i=1,\ldots,m.$$

This implies that

$$\langle P_i-P_0,V_1V_2\rangle=\langle P_i-P_0,V_1\Sigma V_2\rangle,\ \ i=1,\ldots,m.$$

Therefore, $\phi_\beta(V_1V_2)=\phi_\beta(V_1\Sigma V_2)=\max_{X\in B(n)}\phi_\beta(X)$. □

Theorem 2.1. *Let $P_0,\ldots,P_m$ be $n\times n$ unitary matrices, and let $c_0,\ldots,c_m$ be real numbers. The set of optimal solutions of the following optimization problem:*

$$\max_{X\in B(n)}\min_{1\leqslant i\leqslant m}\langle P_i-P_0,X\rangle+c_i-c_0 \tag{10}$$

contains a unitary matrix.

Proof. Denote $\phi(X)=\min_{1\leqslant i\leqslant m}\langle P_i-P_0,X\rangle+c_i-c_0$, for $X\in M_n(\mathbb{C})$. By Lemma 2.3 and the Arzelà-Ascoli theorem, the function ϕ_β defined in (6), which converges pointwise to ϕ as β goes to $+\infty$, converges uniformly to the same function. For each β, by Proposition 2.1 the intersection $O_\beta:=U(n)\cap\arg\max_{X\in B(n)}\phi_\beta$ is not empty. Since the convergence of ϕ_β to ϕ is uniform, each cluster point of a sequence $\{U_\beta\}_{\beta\geqslant 0}$ with $U_\beta\in O_\beta$ for all $\beta>0$ is an optimal solution of the problem (10), see [RW98, p.266]. The cluster point is unitary because $U(n)$ is closed. Thus the optimization problem (10) must have a unitary optimal solution. □

By Theorem 2.1, solving (2) is equivalent to solving (1).

Theorem 2.1, which concern matrices with complex entries, has a real analogue, in which $SU(n)$ is replaced by the real special orthogonal group $SO(n)$. The latter group appeared for instance in the application of max-plus methods to a deterministic filtering problem [SM13]. Then, $U(n)$ is replaced by the nth real orthogonal group $O(n)$, and $B(n)$ is replaced by $B_{\mathbb{R}}(n)$, the set of $n \times n$ matrices X with real entries such that $XX^\top \preccurlyeq I_n$. We next state the real analogue of Theorem 2.1.

Theorem 2.2. *Let $P_0, \dots, P_m$ be $n \times n$ orthogonal matrices, and let $c_0, \dots, c_m$ be real numbers. Then, the set of optimal solutions of the following optimization problem:*

$$\max_{X \in B_{\mathbb{R}}(n)} \min_{1 \leqslant i \leqslant m} \langle P_i - P_0, X \rangle + c_i - c_0 \tag{11}$$

contains an orthogonal matrix.

Proof. The proof is identical to the one of Theorem 2.1, up to trivial changes. □

Example 2.1. The elementary situation in which $n = 1$, so that $O(n) = \{1, -1\}$, $B_{\mathbb{R}}(n) = [-1, 1]$, $U(n)$ is the unit circle, and $B(n)$ is the unit disk, will allow us to see that our restrictive assumptions in Theorem 2.1 and 2.2 are useful. The latter result states that as soon as $P_0, \dots, P_m \in \{-1, 1\}$, the maximum of the function ϕ on the interval $[-1, 1]$ is always attained at the boundary of this interval. Indeed, if $P_0 = -1$, the gradients of the affine constituents of ϕ belong to $\{0, 2\}$, and the maximum of ϕ is attained at point 1, whereas if $P_0 = 1$, the same gradients belong to $\{-2, 0\}$, and the maximum of ϕ is now attained at point -1. A consideration of the $n = 1$ case shows that Theorem 2.2 does not carry over to piecewise linear concave functions of the form

$$\phi(X) = \min_{1 \leqslant i \leqslant m} \langle P_i - Q_i, X \rangle + d_i \,, \tag{12}$$

where P_i, Q_i are orthogonal, and $d_i \in \mathbb{R}$. Indeed, since $P_i, Q_i \in \{\pm 1\}$, $P_i - Q_i$ can achieve any value in $\{-2, 0, 2\}$, and so,

$$\phi(X) = \min(2X, -2X)$$

is of the form (12). The maximum of the latter function over $B_{\mathbb{R}} = [-1, 1]$ is equal to 0, whereas its maximum over $O(1) = \{1, -1\}$ is equal to -1. A similar counter example holds in the complex case. Consider $P_1 = 1, Q_1 = -1$, $P_2 = -1, Q_2 = 1$, $P_3 = i, Q_3 = -i$, $P_4 = -i, Q_4 = i$, we get

$$\phi(X) = \min(-2|\operatorname{Re} X|, -2|\operatorname{Im} X|) \,.$$

The maximum of this function over $B(1)$ is equal to 0, whereas it is readily seen that its maximum over $U(1)$ is equal to $-\sqrt{2}$ (attained by the unitary vector $X = (1 + i)/\sqrt{2}$).

We next state an elementary variation of Theorem 2.1, in which a concave piecewise affine function is maximized over a sphere.

Let $\|\cdot\|$ denote a norm on $\mathbb{R}^n$, and let $\|\cdot\|^*$ denote the dual norm, so that

$$\|p\|^* = \max_{\|y\| \leqslant 1} p \cdot y = \max_{\|y\| = 1} p \cdot y, \qquad \forall p \in \mathbb{R}^n \,, \tag{13}$$

and

$$\|x\| = \max_{\|q\|^* \leqslant 1} q \cdot x = \max_{\|q\|^* = 1} q \cdot x, \qquad \forall x \in \mathbb{R}^n \,,$$

where $\cdot$ denotes the standard scalar product of $\mathbb{R}^n$, and the notation "max" means that each supremum is attained. Let S, B denote the primal unit sphere and the primal ball, so that $S := \{x \in \mathbb{R}^n \mid \|x\| = 1\}$ and $B := \{x \in \mathbb{R}^n \mid \|x\| \leqslant 1\}$. Similarly, let $S^* := \{p \in \mathbb{R}^n \mid \|p\|^* = 1\}$ and $B^* := \{p \in \mathbb{R}^n \mid \|p\|^* \leqslant 1\}$ denote the dual unit sphere and ball. We refer the reader to [AB99, Ch. 6] for more background on convex duality and norms.

Proposition 2.2. *Let $p_0, \ldots, p_m \in S^*$, $c_0, \ldots, c_m \in \mathbb{R}$, and let:*

$$\phi(x) = \inf_{1 \leqslant i \leqslant m} \langle p_i - p_0, x \rangle + c_i - c_0 \,.$$

Then,

$$\max_{x \in B} \phi(x) = \max_{x \in S} \phi(x) \,.$$

Proof. Let ϕ_β denote the log-exp regularization of the function ϕ, defined as in (6). Then, arguing as in the proof of Proposition 2.1, we see that any maximum point x of ϕ_β on B satisfies

$$\sum_{1 \leqslant i \leqslant m} \alpha_i \langle p_i - p_0, z \rangle \leqslant 0, \qquad \forall z \in T_B(x) \,,$$

where $\alpha_1, \ldots, \alpha_m$ are positive numbers with sum 1. If $\|x\| < 1$, we have $T_B(x) = \mathbb{R}^n$, and so,

$$\sum_{1 \leqslant i \leqslant m} \alpha_i (p_i - p_0) = 0$$

i.e.,

$$p_0 = \sum_{1 \leqslant i \leqslant m} \alpha_i p_i \tag{14}$$

Since $\|p_0\|^* = 1$, by (13), we can find a vector $y \in \mathbb{R}^n$ such that $\|y\| = 1$ and $p_0 \cdot y = 1$. Moreover, $p_i \cdot y \leqslant \|p_i\|^* \|y\| = 1$ holds for all i. We deduce from (14) that $1 = p_0 \cdot y = \sum_{1 \leqslant i \leqslant m} \alpha_i p_i \cdot y$, and so, $1 = p_0 \cdot y = p_i \cdot y$, for $1 \leqslant i \leqslant m$. It follows that for all $t \geqslant 0$, $\phi_\beta(x + ty) = \phi_\beta(x)$. Since $\|x + ty\| = \|x\| < 1$ for $t = 0$, and $\|x + ty\| \to \infty$ when $t \to \infty$, using the intermediate value theorem, we deduce that there is a parameter $\bar{t}$ such that $x + \bar{t}y \in S$ minimizes ϕ_β over B. Then, we conclude as in the proof of Theorem 2.1, letting β tend to ∞. □

Remark 2.2. An inspection of the $n = 1$ case shows that the optimization problem (1) is not equivalent to the optimization problem (5), so that the present results only yield an exact convex relaxation for the former problem. An exceptional case in which there is an exact convex relaxation of a problem of type (5) is when the state space of the control problem is the group $SO(2)$. This yields a program of type (5) where the matrices $\{P_0, \ldots, P_m\}$ are all in $SO(2)$. Since $SO(2)$ can be identified to the unit circle of $\mathbb{R}^2$, Proposition 2.2 yields an exact convex relaxation. We are not aware of exact convex relaxations of (5) in general situations.

References

[AB99] C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis. A Hitchiker's Guide*. Springer, 1999.

[BTN01] A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization*. MPS/SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Programming Society (MPS), Philadelphia, PA, 2001. Analysis, algorithms, and engineering applications.

[GMQ11] S. Gaubert, W. M. McEneaney, and Z. Qu. Curse of dimensionality reduction in max-plus based approximation methods: Theoretical estimates and improved pruning algorithms. In *Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC 11)*, pages 1054–1061. IEEE, 2011.

[GQS14] S. Gaubert, Z. Qu, and S. Sridharan. Bundle-based pruning in the max-plus curse of dimensionality free method. In *21st International Symposium on Mathematical Theory of Networks and Systems*, Groningen, The Netherlands, July 2014.

[McE07] W. M. McEneaney. A curse-of-dimensionality-free numerical method for solution of certain HJB PDEs. *SIAM J. Control Optim.*, 46(4):1239–1276, 2007.

[McE09] W. M. McEneaney. Complexity reduction, cornices and pruning. In *Tropical and idempotent mathematics*, volume 495 of *Contemp. Math.*, pages 293–303. Amer. Math. Soc., Providence, RI, 2009.

[MDG08] W. M. McEneaney, A. Deshpande, and S. Gaubert. Curse-of-complexity attenuation in the curse-of-dimensionality-free method for HJB PDEs. In *Proc. of the 2008 American Control Conference*, pages 4684–4690, Seattle, Washington, USA, June 2008.

[MK10] W. M. McEneaney and L. J. Kluberg. Convergence rate for a curse-of-dimensionality-free method for a class of HJB PDEs. *SIAM J. Control Optim.*, 48(5):3052–3079, 2009/10.

[PP91] M.V.F. Pereira and L.M.V.G. Pinto. Multi-stage stochastic optimization applied to energy planning. *Mathematical Programming*, 52:359375, 1991.

[Qu14] Z. Qu. Contraction of Riccati flows applied to the convergence analysis of a max-plus curse-of-dimensionality-free method. *SIAM J. Control Optim.*, 52(5):2677–2706, 2014.

[RW98] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*. Springer-Verlag, Berlin, 1998.

[SGJM10] S. Sridharan, M. Gu, M. R. James, and W. M. McEneaney. Reduced-complexity numerical method for optimal gate synthesis. *Phys. Rev. A*, 82:042319, Oct 2010.

[Sha11] Alexander Shapiro. Analysis of stochastic dual dynamic programming method. *European J. Oper. Res.*, 209(1):63–72, 2011.

[SM13] S. Sridharan and W.M. McEneaney. Deterministic filtering for optimal attitude estimation on SO(3) using max-plus methods. In *Proceedings of the European Control Conference ECC 2013*, pages 2220–2225, Zurich, July 2013.

[SMGJ14] S. Sridharan, W. McEneaney, M. Gu, and M. R. James. A reduced complexity min-plus solution method to the optimal control of closed quantum systems. *Appl. Math. Optim.*, 2014. published on line.

(Stéphane Gaubert) INRIA and CMAP, École Polytechnique, CNRS. Postal address: CMAP, École Polytechnique, 91128 Palaiseau Cedex, France.

E-mail address: Stephane.Gaubert@inria.fr

(Zheng Qu) Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong

E-mail address: zhengqu@hku.hk

(Srinivas Sridharan) Informatics Department at University of Sussex, Brighton, BN1 9RH, United Kingdom.

E-mail address: ss771@sussex.ac.uk