SIAM J. Optim.
Vol. 27, No. 4, pp. 2634–2661

VARIATIONAL GRAM FUNCTIONS: CONVEX ANALYSIS AND OPTIMIZATION[*]

AMIN JALALI[†], MARYAM FAZEL[‡], AND LIN XIAO[§]

Abstract. We introduce a new class of convex penalty functions, called *variational Gram functions* (VGFs), that can promote pairwise relations, such as orthogonality, among a set of vectors in a vector space. These functions can serve as regularizers in convex optimization problems arising from hierarchical classification, multitask learning, and estimating vectors with disjoint supports, among other applications. We study convexity of VGFs, and give characterizations for their convex conjugates, subdifferentials, proximal operators, and related quantities. We discuss efficient optimization algorithms for regularized loss minimization problems where the loss admits a common, yet simple, variational representation and the regularizer is a VGF. These algorithms enjoy a simple kernel trick, an efficient line search, as well as computational advantages over first order methods based on the subdifferential or proximal maps. We also establish a general representer theorem for such learning problems. Last, numerical experiments on a hierarchical classification problem are presented to demonstrate the effectiveness of VGFs and the associated optimization algorithms.

Key words. convex optimization, regularization, inverse problems, matrix optimization, minimax problems

AMS subject classifications. 68Q32, 49K35, 68W40, 68U15, 90C25, 49N45, 90C46

DOI. 10.1137/16M1087424

1. Introduction.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be vectors in $\mathbb{R}^n$. It is well known that their pairwise inner products $\mathbf{x}_i^T \mathbf{x}_j$ for $i, j = 1, \ldots, m$, reveal essential information about their relative orientations, and can serve as a measure for various properties such as orthogonality. In this paper, we consider a class of functions that selectively aggregate the pairwise inner products in a variational form,

$$\Omega_{\mathcal{M}}(\mathbf{x}_1, \ldots, \mathbf{x}_m) = \max_{M \in \mathcal{M}} \textstyle\sum_{i,j=1}^{m} M_{ij} \mathbf{x}_i^T \mathbf{x}_j \,, \tag{1}$$

where $\mathcal{M}$ is a compact subset of the set of m by m symmetric matrices. Let $X = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_m]$ be an $n \times m$ matrix. Then the pairwise inner products $\mathbf{x}_i^T \mathbf{x}_j$ are the entries of the Gram matrix $X^T X$ and the function above can be written as

$$\Omega_{\mathcal{M}}(X) = \max_{M \in \mathcal{M}} \left\langle X^T X, M \right\rangle = \max_{M \in \mathcal{M}} \operatorname{tr}\left(X M X^T \right) \,, \tag{2}$$

where $\langle A, B \rangle = \operatorname{tr}(A^T B)$ denotes the matrix inner product. We call $\Omega_{\mathcal{M}}$ a *variational Gram function* (VGF) of the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$ induced by the set $\mathcal{M}$. If the set $\mathcal{M}$ is clear from the context, we may write $\Omega(X)$ to simplify notation.

[*] Received by the editors August 1, 2016; accepted for publication (in revised form) September 28, 2017; published electronically December 21, 2017. An earlier version of this work has appeared in A. Jalali, *Convex Optimization Algorithms and Statistical Bounds for Learning Structured Models*, Ph.D. thesis, University of Washington, Seattle, WA, 2016, pp. 57–95 [21].

http://www.siam.org/journals/siopt/27-4/M108742.html

Funding: The first author did this work while at the University of Washington. The research of the second author was supported in part by grants NSF CCF-1409836 and ONR N00014-16-1-2789.

[†] Optimization Theme, Wisconsin Institute for Discovery, Madison, WI 53715 (amin.jalali@wisc.edu).

[‡] Department of Electrical Engineering, University of Washington, Seattle, WA 98195 (mfazel@uw.edu).

[§] Machine Learning Group, Microsoft Research, Redmond, WA 98052 (lin.xiao@microsoft.com).

As an example, consider the case where $\mathcal{M}$ is given by a box constraint,

$$\mathcal{M} = \left\{ M : \, |M_{ij}| \le \overline{M}_{ij}, \; i,j = 1,\ldots,m \right\}, \tag{3}$$

where $\overline{M}$ is a symmetric nonnegative matrix. In this case, the maximization in the definition of $\Omega_{\mathcal{M}}$ picks either $M_{ij} = \overline{M}_{ij}$ or $M_{ij} = -\overline{M}_{ij}$ depending on the sign of $\mathbf{x}_i^T\mathbf{x}_j$ for all $i,j = 1,\ldots,m$ (if $\mathbf{x}_i^T\mathbf{x}_j = 0$, the choice is arbitrary). Therefore,

$$\Omega_{\mathcal{M}}(X) = \max_{M\in\mathcal{M}} \; \textstyle\sum_{i,j=1}^m M_{ij}\mathbf{x}_i^T\mathbf{x}_j = \sum_{i,j=1}^m \overline{M}_{ij}|\mathbf{x}_i^T\mathbf{x}_j|\,. \tag{4}$$

Equivalently, $\Omega_{\mathcal{M}}(X)$ is the weighted sum of the absolute values of pairwise inner products. This function was proposed in [47] as a regularization function to promote orthogonality between selected pairs of linear classifiers in the context of hierarchical classification.

Observe that the function $\operatorname{tr}(XMX^T)$ is a convex quadratic function of X if M is positive semidefinite. As a result, the variational form $\Omega_{\mathcal{M}}(X)$ is convex if $\mathcal{M}$ is a subset of the positive semidefinite cone $\mathbb{S}_+^m$, because then it is the pointwise maximum of a family of convex functions indexed by $M \in \mathcal{M}$ (see, e.g., [38, Theorem 5.5]). However, this is not a necessary condition. For example, the set $\mathcal{M}$ in (3) is not a subset of $\mathbb{S}_+^m$ unless $\overline{M} = 0$, but the VGF in (4) is convex provided that the *comparison matrix* of $\overline{M}$ (derived by negating the off-diagonal entries) is positive semidefinite [47]. In this paper, we study conditions under which different classes of VGFs are convex and provide unified characterizations for the subdifferential, convex conjugate, and the associated proximal operator for any convex VGF. Interestingly, a convex VGF defines a seminorm[1] as

$$\|X\|_{\mathcal{M}} := \sqrt{\Omega_{\mathcal{M}}(X)} = \max_{M\in\mathcal{M}} \; \left(\textstyle\sum_{i,j=1}^m M_{ij}\mathbf{x}_i^T\mathbf{x}_j\right)^{1/2}. \tag{5}$$

If $\mathcal{M} \subset \mathbb{S}_+^m$, then $\|X\|_{\mathcal{M}}$ is the pointwise maximum of the seminorms $\|XM^{1/2}\|_F$ over all $M \in \mathcal{M}$.

VGFs and the associated norms can serve as penalties or regularization functions in optimization problems to promote certain pairwise properties among a set of vector variables (such as orthogonality in the above example). In this paper, we consider optimization problems of the form

$$\underset{X\in\mathbb{R}^{n\times m}}{\text{minimize}} \quad \mathcal{L}(X) + \lambda\,\Omega_{\mathcal{M}}(X)\,, \tag{6}$$

where $\mathcal{L}(X)$ is a convex loss function of the variable $X = [\mathbf{x}_1 \;\cdots\; \mathbf{x}_m]$, $\Omega(X)$ is a convex VGF, and $\lambda > 0$ is a parameter to trade off the relative importance of these two functions. We will focus on problems where $\mathcal{L}(X)$ is smooth or has an explicit variational structure, and show how to exploit the structures of $\mathcal{L}(X)$ and $\Omega(X)$ together to derive efficient optimization algorithms. More specifically, we employ a unified variational representation for many common loss functions, as

$$\mathcal{L}(X) = \max_{\mathbf{g}\in\mathcal{G}} \; \langle X, \mathcal{D}(\mathbf{g})\rangle - \hat{\mathcal{L}}(\mathbf{g})\,, \tag{7}$$

where $\hat{\mathcal{L}} : \mathbb{R}^p \to \mathbb{R}$ is a convex function, $\mathcal{G}$ is a convex and compact subset of $\mathbb{R}^p$, and $\mathcal{D} : \mathbb{R}^p \to \mathbb{R}^{n\times m}$ is a linear operator. Exploiting the variational structure in both the loss function and the regularizer allows us to employ a variety of efficient primal-dual algorithms, such as mirror-prox [36], which now only require projections onto $\mathcal{M}$ and $\mathcal{G}$, instead of computing subgradients or proximal mappings for the loss function and

[1] A seminorm satisfies all the properties of a norm except that it can be zero for a nonzero input.

the regularizer. Our approach is specially helpful for regularization functions with proximal mappings that are expensive to compute [24].

Exploiting this structure for the loss function and the regularizer enables a simple preprocessing step for dimensionality reduction, presented in section 5.2, which can substantially reduce the per iteration cost of any optimization algorithm for (6). We also present a general representer theorem for problems of the form (6) in section 5.3 where the optimal solution is characterized in terms of the input data in a simple and interpretable way. This representer theorem can be seen as a generalization of the well-known results for quadratic functions [41].

Organization. In section 2, we give more examples of VGFs and explain the connections with functions of Euclidean distance matrices, diversification, and robust optimization. Section 3 studies the convexity of VGFs, as well as their conjugates, semidefinite representability, corresponding norms, and subdifferentials. Their proximal operators are derived in section 4. In section 5, we study a class of structured loss minimization problems with VGF penalties, and show how to exploit their structure, to get an efficient optimization algorithm using a variant of the mirror-prox algorithm with adaptive line search, to use a simple preprocessing step to reduce the computations in each iteration, and to provide a characterization of the optimal solution as a representer theorem. Finally, in section 6, we present a numerical experiment on hierarchical classification to illustrate the application of VGFs.

Notation. In this paper, $\mathbb{S}^m$ denotes the set of symmetric matrices in $\mathbb{R}^{m\times m}$, and $\mathbb{S}^m_+ \subset \mathbb{S}^m$ is the cone of positive semidefinite (PSD) matrices. We may omit the superscript m when the dimension is clear from the context. The symbol $\preceq$ represents the Loewner partial order and $\langle\cdot,\cdot\rangle$ denotes the inner product. We use capital letters for matrices and bold lower case letters for vectors. We use $X \in \mathbb{R}^{n\times m}$ and $\mathbf{x} = \operatorname{vec}(X) \in \mathbb{R}^{nm}$ interchangeably with $\mathbf{x}_i$ denoting the ith column of X; i.e., $X = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_m]$. By $\mathbf{1}$ and 0 we denote matrices or vectors of all ones and all zeros, respectively, whose sizes would be clear from the context. The entrywise absolute value of X is denoted by $|X|$. The ℓ_p norm of the input vector or matrix is denoted by $\|\cdot\|_p$, and $\|\cdot\|_F$ and $\|\cdot\|_{\mathrm{op}}$ denote the Frobenius norm and the operator norm, respectively. We overload the superscript $*$ for three purposes. For a linear mapping $\mathcal{D}$, the adjoint operator is denoted by $\mathcal{D}^*$. For a norm denoted by $\|\cdot\|$, with possible subscripts, the dual norm is defined as $\|\mathbf{y}\|^* = \sup\{\langle \mathbf{x},\mathbf{y}\rangle : \ \|\mathbf{x}\| \le 1\}$. For other functions, denoted by a letter, namely, f, the convex conjugate is defined as $f^*(y) = \sup_y \ \langle x, y\rangle - f(x)$. By $\arg\min$ ($\arg\max$), we denote an optimal point to a minimization (maximization) program, while $\operatorname{Arg}\min$ (or $\operatorname{Arg}\max$) is the set of all optimal points. The operator $\operatorname{diag}(\cdot)$ is used to put a vector on the diagonal of a zero matrix of corresponding size, to extract the diagonal entries of a matrix as a vector, or for zeroing out the off-diagonal entries of a matrix. We use $f \equiv g$ to denote $f(x) = g(x)$ for all $x \in \operatorname{dom}(f) = \operatorname{dom}(g)$.

2. Examples and connections.

In this section, we present examples of VGFs associated with different choices of the set $\mathcal{M}$. The list includes some well-known functions that can be expressed in the variational form of (1), as well as some new ones.

Vector norms. Any vector norm $\|\cdot\|$ on $\mathbb{R}^m$ is the square root of a VGF defined by $\mathcal{M} = \{\mathbf{u}\mathbf{u}^T : \ \|\mathbf{u}\|^* \le 1\}$. For a column vector $\mathbf{x} \in \mathbb{R}^m$, the VGF is given by

$$\Omega_{\mathcal{M}}\left(\mathbf{x}^T\right) = \max_{\mathbf{u}} \ \left\{\operatorname{tr}\left(\mathbf{x}^T\mathbf{u}\mathbf{u}^T\mathbf{x}\right) : \ \|\mathbf{u}\|^* \le 1\right\} = \max_{\mathbf{u}} \ \left\{\left(\mathbf{x}^T\mathbf{u}\right)^2 : \ \|\mathbf{u}\|^* \le 1\right\} = \|\mathbf{x}\|^2 .$$

As another example for when $n = 1$, consider the case where $\mathcal{M}$ is a compact convex set of diagonal matrices with positive diagonal entries. The corresponding

VGF (and norm) is defined as

$$\Omega_{\mathcal{M}}\left(\mathbf{x}^T\right) = \max_{\theta\in\operatorname{diag}(\mathcal{M})} \textstyle\sum_{i=1}^m \theta_i x_i^2 = \|\mathbf{x}\|_{\mathcal{M}}^2 \,, \tag{8}$$

which is a squared norm and the dual norm can be expressed as

$$(\|\mathbf{x}\|_{\mathcal{M}}^*)^2 = \inf_{\theta\in\operatorname{diag}(\mathcal{M})} \sum_{i=1}^m \frac{1}{\theta_i} x_i^2 \,.$$

This norm and its dual were first introduced in [34], in the context of regularization for structured sparsity, and later discussed in detail in [3]. The k-support norm [2], which is a norm used to encourage vectors to have k or fewer nonzero entries, is a special case of the dual norm given above, corresponding to

$$\mathcal{M} = \{\operatorname{diag}(\theta) :\ 0 \le \theta_i \le 1 \,,\ \mathbf{1}^T\theta \le k\} \,.$$

Our optimization approach for VGF regularized problems (section 5) requires projection onto $\mathcal{M}$. Projection onto the intersection of a box with a half-space is a special case of the continuous quadratic knapsack problem and can be performed in linear time; e.g., see [25].

Weighted norms of the Gram matrix. Given a symmetric nonnegative matrix $\overline{M}$, we can define a class of VGFs based on any norm $\|\cdot\|$ and its dual norm $\|\cdot\|^*$. Consider

$$\mathcal{M} = \left\{K \circ \overline{M} :\ \|K\|^* \le 1,\ K^T = K\right\} \,, \tag{9}$$

where $\circ$ denotes the matrix Hadamard product, $(K\circ\overline{M})_{ij} = K_{ij}\overline{M}_{ij}$ for all i, j. Then,

$$\Omega_{\mathcal{M}}(X) = \max_{\|K\|^*\le 1} \left\langle K\circ\overline{M}, X^TX\right\rangle = \max_{\|K\|^*\le 1} \left\langle K, \overline{M}\circ\left(X^TX\right)\right\rangle = \left\|\overline{M}\circ\left(X^TX\right)\right\| \,.$$

The following are several concrete examples.

(i) If we let $\|\cdot\|^*$ in (9) be the ℓ_∞ norm, then

$$\mathcal{M} = \{M :\ |M_{ij}/\overline{M}_{ij}| \le 1,\ i,j = 1,\ldots,m\} \,.$$

which is the same as in (3). Here we use the convention $0/0 = 0$, thus $M_{ij} = 0$ whenever $\overline{M}_{ij} = 0$. In this case, we obtain the VGF in (4):

$$\Omega_{\mathcal{M}}(X) = \left\|\overline{M}\circ\left(X^TX\right)\right\|_1 = \textstyle\sum_{i,j=1}^m \overline{M}_{ij}\left|\mathbf{x}_i^T\mathbf{x}_j\right| .$$

(ii) If we use the ℓ_2 norm in (9), then $\mathcal{M} = \{M :\ \sum_{i,j=1}^m (M_{ij}/\overline{M}_{ij})^2 \le 1\}$ and

$$\Omega_{\mathcal{M}}(X) = \left\|\overline{M}\circ\left(X^TX\right)\right\|_F = \left(\textstyle\sum_{i,j=1}^m \left(\overline{M}_{ij}\mathbf{x}_i^T\mathbf{x}_j\right)^2\right)^{1/2} . \tag{10}$$

This function has been considered in experiment design [8, 12].

(iii) Using ℓ_1 norm for $\|\cdot\|^*$ in (9) gives $\mathcal{M} = \{M :\ \sum_{i,j=1}^m |M_{ij}/\overline{M}_{ij}| \le 1\}$ and

$$\Omega_{\mathcal{M}}(X) = \left\|\overline{M}\circ\left(X^TX\right)\right\|_\infty = \max_{i,j=1,\ldots,m} \overline{M}_{ij}\left|\mathbf{x}_i^T\mathbf{x}_j\right| \,. \tag{11}$$

This case can also be traced back to [8] in the statistics literature, where the maximum of $|\mathbf{x}_i^T\mathbf{x}_j|$ for $i \neq j$ is used as the measure to choose among supersaturated designs.

Many other interesting examples can be constructed this way. For example, one can model *sharing* versus *competition* using the group-ℓ_1 norm of the Gram matrix which was considered in vision tasks [22]. We will revisit the above examples to discuss their convexity conditions in section 3.

Spectral functions. From the definition, the value of a VGF is invariant under left multiplication of X by an orthogonal matrix, but this is not true for right multiplication. Hence, VGFs are *not* functions of singular values (e.g., see [29]) in general, and are functions of the row space of X as well. This also implies that in general $\Omega(X) \not\equiv \Omega(X^T)$. However, if the set $\mathcal{M}$ is closed under left and right multiplication by orthogonal matrices, then $\Omega_{\mathcal{M}}(X)$ becomes a function of squared singular values of X. For any matrix $M \in \mathbb{S}^m$, denote the sorted vector of its singular values, in descending order, by $\sigma(M)$ and let $\Theta = \{\sigma(M) : \ M \in \mathcal{M}\}$. Then we have

$$\Omega_{\mathcal{M}}(X) = \max_{M \in \mathcal{M}} \operatorname{tr}\left(XMX^T\right) = \max_{\theta \in \Theta} \textstyle\sum_{i=1}^{\min(n,m)} \theta_i \, \sigma_i(X)^2 \tag{12}$$

as a result of von Neumann's trace inequality [35]. Note the similarity of the above to the VGF in (8). As an example, consider

$$\mathcal{M} = \{M : \ \alpha_1 I \preceq M \preceq \alpha_2 I, \ \operatorname{tr}(M) \le \alpha_3\}, \tag{13}$$

where $0 < \alpha_1 < \alpha_2$ and $m\alpha_1 \le \alpha_3 \le m\alpha_2$ are given constants. Note that in this case $\mathcal{M} \subset \mathbb{S}_+^m$ which readily establishes the convexity of $\Omega_{\mathcal{M}}$. For

$$\mathcal{M}_r \coloneqq \{M : \ 0 \preceq M \preceq I, \ \operatorname{tr}(M) \le r\},$$

the corresponding norm $\|\cdot\|_{\mathcal{M}_r}$ is known as the Ky-Fan $(2, r)$-norm, and $\Omega_{\mathcal{M}_r}$ has been analyzed in the context of low-rank regression analysis [16]. For $\mathcal{M}$ in (13), the dual norm $\|\cdot\|_{\mathcal{M}}^*$ is referred to as the *spectral box norm* in [33], and $\Omega_{\mathcal{M}}^*$ has been considered in [20] for clustered multitask learning where it is presented as a convex relaxation for k-means. $\|\cdot\|_{\mathcal{M}_r}^*$ is considered in [14] for finding large low-rank submatrices in a given nonnegative matrix.

Finite set $\mathcal{M}$. For a finite set $\mathcal{M} = \{M_1, \ldots, M_p\} \subset \mathbb{S}_+^m$, the VGF is given by

$$\Omega_{\mathcal{M}}(X) = \max_{i=1,\ldots,p} \left\| X M_i^{1/2} \right\|_F^2,$$

i.e., the pointwise maximum of a finite number of squared weighted Frobenius norms.

In the following subsections, we consider classes of VGFs that can be used in promoting diversity, have connections to Euclidean distance matrices, or can be interpreted in a robust optimization framework.

2.1. Diversification.

VGFs can be used for *diversifying* certain pairs of columns of the input matrix, e.g., minimizing (4) pushes to zero the inner products $\mathbf{x}_i^T \mathbf{x}_j$ corresponding to the nonzero entries in $\overline{M}$ as much as possible. As another example, observe that two nonnegative vectors have disjoint supports if and only if they are orthogonal to each other. Hence, using a VGF as (4), $\Omega_{\mathcal{M}}(X) = \sum_{i,j=1}^m \overline{M}_{ij} |\mathbf{x}_i^T \mathbf{x}_j|$, that promotes orthogonality, we can define

$$\Psi(X) \coloneqq \Omega_{\mathcal{M}}(|X|) \tag{14}$$

to promote disjoint supports among certain columns of X, hence, diversifying the supports of columns of X. Convexity of (14) is discussed in section 3.6. Different approaches has been used in machine learning applications for promoting diversity; e.g., see [31, 27, 19] and references therein.

2.2. Functions of Euclidean distance matrix.

Consider a set $\mathcal{M} \subset \mathbb{S}^m$ with the property that $M\mathbf{1} = 0$ for all $M \in \mathcal{M}$. For every $M \in \mathcal{M}$, let $A = \operatorname{diag}(M) - M$ and observe that

$$\operatorname{tr}\left(XMX^T\right) = \textstyle\sum_{i,j=1}^m M_{ij} \mathbf{x}_i^T \mathbf{x}_j = \frac{1}{2} \sum_{i,j=1}^m A_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 .$$

This allows us to express the associated VGF as a function of the *Euclidean distance matrix* D, which is defined by $D_{ij} = \frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|_2^2$ for $i, j = 1, \dots, m$ (see, e.g., [9, section 8.3]). Let $\mathcal{A} = \{\mathrm{diag}(M) - M : M \in \mathcal{M}\}$. Then we have

$$\Omega_{\mathcal{M}}(X) = \max_{M \in \mathcal{M}} \mathrm{tr}\left(XMX^T\right) = \max_{A \in \mathcal{A}} \langle A, D \rangle .$$

A sufficient condition for the above function to be convex in X is that each $A \in \mathcal{A}$ is entrywise nonnegative, which implies that the corresponding $M = \mathrm{diag}(A\mathbf{1}) - A$ is diagonally dominant with nonnegative diagonal elements, hence, positive semidefinite. However, this is not a necessary condition and $\Omega_{\mathcal{M}}$ can be convex without all A's being entrywise nonnegative.

2.3. Connection with robust optimization. The VGF-regularized loss minimization problem has the following connection to robust optimization (see, e.g., [7]): the optimization program

$$\underset{X}{\mathrm{minimize}} \ \max_{M \in \mathcal{M}} \ \mathcal{L}(X) + \mathrm{tr}\left(XMX^T\right)$$

can be interpreted as seeking an X with minimal worst-case value over an uncertainty set $\mathcal{M}$. Alternatively, when $\mathcal{M} \subset \mathbb{S}_+^m$, this can be viewed as a problem with Tikhonov regularization $\|XM^{1/2}\|_F^2$, where the weight matrix $M^{1/2}$ is subject to errors characterized by the set $\mathcal{M}$.

3. Convex analysis of VGF.

3. Convex analysis of VGF. In this section, we study the convexity of VGFs, their conjugate functions, and subdifferentials, as well as the related norms.

First, we review some basic properties. Notice that $\Omega_{\mathcal{M}}$ is the *support function* of the set $\mathcal{M}$ at the Gram matrix $X^T X$, i.e.,

$$\Omega_{\mathcal{M}}(X) = \max_{M \in \mathcal{M}} \mathrm{tr}\left(XMX^T\right) = S_{\mathcal{M}}\left(X^T X\right) , \tag{15}$$

where the support function of a set $\mathcal{M}$ is defined as $S_{\mathcal{M}}(Y) = \sup_{M \in \mathcal{M}} \langle M, Y \rangle$ (see, e.g., [38, section 13]). By properties of the support function (see [38, section 15]),

$$\Omega_{\mathcal{M}} \equiv \Omega_{\mathrm{conv}(\mathcal{M})} ,$$

where $\mathrm{conv}(\mathcal{M})$ denotes the convex hull of $\mathcal{M}$. It is clear that the representation of a VGF (i.e., the associated set $\mathcal{M}$) is not unique. Henceforth, without loss of generality, we assume $\mathcal{M}$ is convex unless explicitly noted otherwise. Also, for simplicity we assume $\mathcal{M}$ is a compact set, while all we need is that the maximum in (1) is attained. For example, a noncompact $\mathcal{M}$ that is unbounded along any negative semidefinite direction is allowed. Last, we assume $0 \in \mathcal{M}$.

Moreover, VGFs are left unitarily invariant; for any $Y \in \mathbb{R}^{n \times m}$ and any orthogonal matrix $U \in \mathbb{R}^{n \times n}$, where $UU^T = U^T U = I$, we have $\Omega(Y) = \Omega(UY)$ and $\Omega^*(Y) = \Omega^*(UY)$; use (2) and (19). We use this property in simplifying computations involving VGFs (such as proximal mapping calculations in section 4) as well as in establishing a general kernel trick and representer theorem in section 5.2.

As we mentioned in the introduction, a sufficient condition for the convexity of a VGF is that $\mathcal{M} \subset \mathbb{S}_+^m$. In section 3.1, we discuss more concrete conditions for determining convexity when the set $\mathcal{M}$ is a polytope. In section 3.2, we describe a more tangible sufficient condition for general sets.

3.1. Convexity with polytope $\mathcal{M}$. Consider the case where $\mathcal{M}$ is a polytope with p vertices, i.e., $\mathcal{M} = \mathrm{conv}\{M_1, \dots, M_p\}$. The support function of this set is given as $S_{\mathcal{M}}(Y) = \max_{i=1,\dots,p} \langle Y, M_i \rangle$ and is piecewise linear [40, section 8.E]. For a polytope $\mathcal{M}$, we define $\mathcal{M}_{\text{eff}}$ as a subset of $\{M_1, \dots, M_p\}$ with the smallest possible size satisfying $S_{\mathcal{M}}(X^T X) = S_{\mathcal{M}_{\text{eff}}}(X^T X)$ for all $X \in \mathbb{R}^{n\times m}$.

As an example, for $\mathcal{M} = \{M : \ |M_{ij}| \le \overline{M}_{ij}, \ i,j = 1,\dots,m\}$ which gives the function defined in (4), we have

$$\mathcal{M}_{\text{eff}} \subseteq \left\{ M : \ M_{ii} = \overline{M}_{ii} \ , \ M_{ij} = \pm \overline{M}_{ij} \text{ for } i \neq j \ \right\} \ . \tag{16}$$

Whether the above inclusion holds with equality or not depends on n.

THEOREM 3.1. *For a polytope $\mathcal{M} \subset \mathbb{S}^m$, the associated VGF is convex if and only if $\mathcal{M}_{\text{eff}} \subset \mathbb{S}^m_+$.*

Proof. Obviously, $\mathcal{M}_{\text{eff}} \subset \mathbb{S}^m_+$ ensures convexity of $\max_{M\in\mathcal{M}_{\text{eff}}} \operatorname{tr}(XMX^T) = \Omega_{\mathcal{M}}(X)$. Next, we prove the necessity of this condition for any $\mathcal{M}_{\text{eff}}$. Take any $M_i \in \mathcal{M}_{\text{eff}}$. If for every $X \in \mathbb{R}^{n\times m}$ with $\Omega(X) = \operatorname{tr}(XM_iX^T)$ there exists another $M_j \in \mathcal{M}_{\text{eff}}$ with $\Omega(X) = \operatorname{tr}(XM_jX^T)$, then $\mathcal{M}_{\text{eff}}\backslash\{M_i\}$ is an effective subset of $\mathcal{M}$ which contradicts the minimality of $\mathcal{M}_{\text{eff}}$. Hence, there exists X_i such that $\Omega(X_i) = \operatorname{tr}(X_iM_iX_i^T) > \operatorname{tr}(X_iM_jX_i^T)$ for all $j \neq i$. Hence, Ω is twice continuously differentiable in a small neighborhood of X_i with Hessian $\nabla^2\Omega(\operatorname{vec}(X_i)) = M_i \otimes I_n$, where $\otimes$ denotes the matrix Kronecker product. Since Ω is assumed to be convex, the Hessian has to be PSD which gives $M_i \succeq 0$. ∎

Next we give a few examples to illustrate the use of Theorem 3.1.

Example 1. We begin with the example defined in (4). Authors in [47] provided the necessary (when $n \ge m-1$) and sufficient condition for convexity using results from M-matrix theory: First, define the comparison matrix $\widetilde{M}$ associated with the symmetric nonnegative matrix $\overline{M}$ as $\widetilde{M}_{ii} = \overline{M}_{ii}$ and $\widetilde{M}_{ij} = -\overline{M}_{ij}$ for $i \neq j$. Then $\Omega_{\mathcal{M}}$ is convex if $\widetilde{M}$ is PSD, and this condition is also necessary when $n \ge m-1$ [47]. Theorem 3.1 provides an alternative and more general proof. Denote the minimum eigenvalue of a symmetric matrix M by $\lambda_{\min}(M)$. From (16) we have

$$\begin{aligned}
\min_{M\in\mathcal{M}_{\text{eff}}} \lambda_{\min}(M) &= \min_{\substack{M\in\mathcal{M}_{\text{eff}}\\ \|\mathbf{z}\|_2=1}} \mathbf{z}^T M\mathbf{z} \ge \min_{\|\mathbf{z}\|_2=1} \sum_i \overline{M}_{ii} z_i^2 - \sum_{i\neq j} \overline{M}_{ij}|z_iz_j| \\
&= \min_{\|\mathbf{z}\|_2=1} |\mathbf{z}|^T \widetilde{M} |\mathbf{z}| \ge \lambda_{\min}(\widetilde{M})\,.
\end{aligned} \tag{17}$$

When $n \ge m-1$, one can construct $X \in \mathbb{R}^{n\times m}$ such that all off-diagonal entries of X^TX are negative (see the example in [47, Appendix A.2]). On the other hand, [11, Lemma 2.1(2)] states that the existence of such a matrix implies $n \ge m-1$. Hence, $\widetilde{M} \in \mathcal{M}_{\text{eff}}$ if and only if $n \ge m-1$. Therefore, both inequalities in (17) should hold with equality, which means that $\mathcal{M}_{\text{eff}} \subset \mathbb{S}^m_+$ if and only if $\widetilde{M} \succeq 0$. By Theorem 3.1, this is equivalent to the VGF in (4) being convex. If $n < m-1$, then $\mathcal{M}_{\text{eff}}$ may not contain $\widetilde{M}$, thus $\widetilde{M} \succeq 0$ is only a "sufficient" condition for convexity for general n. An illustration for this result is given in Figure 1.

Example 2. Similar to the set $\mathcal{M}$ above, consider a box that is not necessarily symmetric around the origin. More specifically, let

$$\mathcal{M} = \{M \in \mathbb{S}^m : \ M_{ii} = D_{ii}\,, \ |M - C| \le D\}\,,$$

where C (denoting the center) is a symmetric matrix with zero diagonal, and D is a symmetric nonnegative matrix. In this case, we have

$$\mathcal{M}_{\text{eff}} \subseteq \{M : \ M_{ii} = D_{ii}\,, \ M_{ij} = C_{ij} \pm D_{ij} \text{ for } i \neq j\}\,.$$

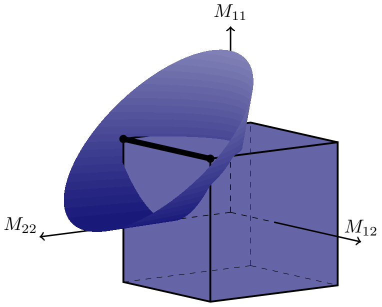


FIG. 1. *The PSD cone, and the set in* (3) *defined by* $\overline{M} = [1,\ 0.8;\ 0.8,\ 1]$, *where* 2×2 *symmetric matrices are embedded into* $\mathbb{R}^3$. *The thick edge of the cube is the set of all points with the same diagonal elements as* $\overline{M}$ *(see* (16)*), and the two endpoints constitute* $\mathcal{M}_{\rm eff}$. *Positive semidefiniteness of* $\widetilde{M}$ *is a necessary and sufficient condition for the convexity of* $\Omega_{\mathcal{M}} : \mathbb{R}^{n\times 2} \to \mathbb{R}$ *for all* $n \geq m-1 = 1$.

When used as a penalty function in applications, this can capture the prior information that when $\mathbf{x}_i^T\mathbf{x}_j$ is not zero, a particular range of acute or obtuse angles (depending on the sign of C_{ij}) between the vectors is preferred. Similarly to (17),

$$\min_{M\in\mathcal{M}_{\rm eff}} \lambda_{\min}(M) \geq \min_{\|\mathbf{z}\|_2=1} |\mathbf{z}|^T\widetilde{D}|\mathbf{z}| + \mathbf{z}^TC\mathbf{z} \geq \lambda_{\min}(\widetilde{D}) + \lambda_{\min}(C)\,,$$

where $\widetilde{D}$ is the comparison matrix associated with D . Note that C has zero diagonals and cannot be PSD. Hence, a sufficient condition for convexity of $\Omega_{\mathcal{M}}$ defined by an asymmetric box is that $\lambda_{\min}(\widetilde{D}) + \lambda_{\min}(C) \geq 0$.

Example 3. *Consider the VGF defined in* (11), *associated with*

$$\mathcal{M} = \left\{M \in \mathbb{S}^m :\ \textstyle\sum_{(i,j):\,\overline{M}_{ij}\neq 0} |M_{ij}/\overline{M}_{ij}| \leq 1\,,\ M_{ij} = 0 \textit{ if } \overline{M}_{ij} = 0\right\}, \tag{18}$$

where $\overline{M}$ *is a symmetric nonnegative matrix. Vertices of* $\mathcal{M}$ *are matrices with either only one nonzero value* $\overline{M}_{ii}$ *on the diagonal, or two nonzero off-diagonal entries at* (i,j) *and* (j,i) *equal to* $\frac{1}{2}\overline{M}_{ij}$ *or* $-\frac{1}{2}\overline{M}_{ij}$. *The second type of matrices cannot be PSD as their diagonal is zero, and according to Theorem 3.1, convexity of* $\Omega_{\mathcal{M}}$ *requires these vertices do not belong to* $\mathcal{M}_{\rm eff}$. *Therefore, the matrices in* $\mathcal{M}_{\rm eff}$ *should be diagonal. Hence, a convex VGF corresponding to the set* (18) *has the form* $\Omega(X) = \max_{i=1,\ldots,m} \overline{M}_{ii}\|\mathbf{x}_i\|_2^2$. *To ensure such a description for* $\mathcal{M}_{\rm eff}$ *we need* $\max\{\overline{M}_{ii}\|\mathbf{x}_i\|_2^2, \overline{M}_{jj}\|\mathbf{x}_j\|_2^2\} \geq \overline{M}_{ij}|\mathbf{x}_i^T\mathbf{x}_j|$ *for all* i, j, *and any* $X \in \mathbb{R}^{n\times m}$, *which is equivalent to* $\overline{M}_{ii}\overline{M}_{jj} \geq \overline{M}_{ij}^2$ *for all* i, j. *This is satisfied if* $\overline{M} \succeq 0$. *However, positive semidefiniteness is not necessary. For example, all of the three 2 by 2 principal minors of the following matrix are nonnegative as desired, but it is not PSD:* $\overline{M} = [1, 1, 2\,; 1, 2, 0\,; 2, 0, 5] \not\succeq 0$.

3.2. A spectral sufficient condition. As mentioned before, when $\mathcal{M}$ is not a polytope, it seems less clear how we can provide necessary and sufficient guarantees for convexity that are easy to check. However, simple sufficient conditions can be easily checked for certain sets $\mathcal{M}$, for example, spectral sets (Lemma 3.2). We first provide an example and consider a specialized approach to establish convexity, to illustrate the advantage of a simple guarantee as the one we present in Lemma 3.2.

(i) Consider the VGF defined in (10) and its associated set given in (9) when we plug in the Frobenius norm, i.e.,

$$\mathcal{M} = \left\{ K \circ \overline{M} : \ \|K\|_F \leq 1\,, \ K^T = K \right\}.$$

In this case, $\mathcal{M}$ is not a polytope, but we can proceed with a similar analysis as in the previous subsection. In particular, given any $X \in \mathbb{R}^{n\times m}$, the value of $\Omega_{\mathcal{M}}(X)$ is achieved by an optimal matrix $K_X = (\overline{M} \circ X^T X)/\|\overline{M} \circ X^T X\|_F$. We observe that

$$\overline{M} \succeq 0 \implies \overline{M} \circ \overline{M} \succeq 0 \iff K_X \circ \overline{M} \succeq 0 \ \ \forall X \implies \Omega_{\mathcal{M}} \text{ is convex}\,.$$

The first implication is by the Schur product theorem [18, Theorem 7.5.1] and does not hold in reverse. For example, besides obvious cases such as $\overline{M} = -I$, consider $\overline{M} \circ \overline{M} = [1, 1, 2; 1, 2, 3; 2, 3, 5.01] \succeq 0$, where $\overline{M} \not\succeq 0$. The second implication, from left to right, is again by the Schur product theorem. The right to left part is by observing that for any $n \geq 1$, X can always be chosen to select a principal minor of $\overline{M} \circ \overline{M}$. The third implication is straightforward: the pointwise maximum of convex quadratics is convex. All in all, a sufficient condition for $\Omega_{\mathcal{M}}$ being convex is that the Hadamard square of $\overline{M}$, namely, $\overline{M} \circ \overline{M}$, is PSD. It is worth mentioning that when $\overline{M} \circ \overline{M} \succeq 0$, hence, real, nonnegative, and PSD, it is referred to as a *doubly nonnegative matrix*.

Denote by M_+ the orthogonal projection of a symmetric matrix M onto the PSD cone, which is given by the matrix formed by only positive eigenvalues and their associated eigenvectors of M.

LEMMA 3.2 (a sufficient condition). *$\Omega_{\mathcal{M}}$ is convex if for any $M \in \mathcal{M}$ there exists $M' \in \mathcal{M}$ such that $M_+ \preceq M'$.*

Proof. For any X, $\operatorname{tr}(XMX^T) \leq \operatorname{tr}(XM_+X^T)$ clearly holds. Therefore,

$$\Omega_{\mathcal{M}}(X) = \max_{M\in\mathcal{M}} \operatorname{tr}\left(XMX^T\right) \leq \max_{M\in\mathcal{M}} \operatorname{tr}\left(XM_+X^T\right) .$$

On the other hand, the assumption of the lemma gives

$$\max_{M\in\mathcal{M}} \operatorname{tr}\left(XM_+X^T\right) \leq \max_{M'\in\mathcal{M}} \operatorname{tr}\left(XM'X^T\right) = \Omega_{\mathcal{M}}(X)$$

which implies that the inequalities have to hold with equality, which implies that $\Omega_{\mathcal{M}}(X)$ is convex. Note the assumption of the lemma can hold while $\mathcal{M}_+ \not\subseteq \mathcal{M}$. □

On the other hand, it is easy to see that the condition in Lemma 3.2 is not necessary. Consider $\mathcal{M} = \{M \in \mathbb{S}^2 : \ |M_{ij}| \leq 1\}$. Although the associated VGF is convex (because the comparison matrix is PSD), there is no matrix $M' \in \mathcal{M}$ satisfying $M' \succeq M_+$, where $M = [0, 1; 1, 1] \in \mathcal{M}$ and $M_+ \simeq [0.44, 0.72; 0.72, 1.17]$, as for any $M' \in \mathcal{M}$ we have $(M' - M_+)_{22} < 0$.

As discussed before, when $\mathcal{M}$ is a polytope, convexity of $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M}_{\text{eff}}}$ is equivalent to $\mathcal{M}_{\text{eff}} \subset \mathbb{S}^m_+$. For general sets $\mathcal{M}$, we showed that $\mathcal{M}_+ \subseteq \mathcal{M}$ is a sufficient condition for convexity. Similarly to the proof of Lemma 3.2, we can provide another sufficient condition for convexity of a VGF: that all of the maximal points of $\mathcal{M}$ with respect to the partial order defined by $\mathbb{S}^m_+$ (the Loewner order) are PSD. These are the points $M \in \mathcal{M}$ for which $(\mathcal{M} - M) \cap \mathbb{S}^m_+ = \{0_m\}$. In all of these pursuits, we are looking for a subset $\mathcal{M}'$ of the PSD cone such that $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M}'}$. When such a set exists, $\Omega_{\mathcal{M}}$ is convex and various optimization-related quantities can be computed for it. Hereafter, we assume there exists a set $\mathcal{M}' \subseteq \mathcal{M} \cap \mathbb{S}_+$ for which $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M}'}$, which in turn implies $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M}\cap\mathbb{S}_+}$. For example, based on Theorem 3.1, this property holds for all convex VGFs associated with a polytope $\mathcal{M}$.

3.3. Conjugate function. For any function Ω, the conjugate function is defined as $\Omega^*(Y) = \sup_X \langle X, Y\rangle - \Omega(X)$ and the transformation that maps Ω to Ω^* is called the Legendre–Fenchel transform (e.g., [38, section 12]). In this section, we derive a representation for the conjugate function for a VGF. First, we state a result on generalized Schur complements which will be used in the following sections.

LEMMA 3.3 (generalized Schur complement [1]). *For symmetric matrices M, C,*

$$\begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \succeq 0 \quad \Longleftrightarrow \quad M \succeq 0\ ,\ C - YM^{\dagger}Y^T \succeq 0\ ,\ Y(I - MM^{\dagger}) = 0\,,$$

where the last condition is equivalent to $\operatorname{range}(Y^T) \subseteq \operatorname{range}(M)$*.*

PROPOSITION 3.4 (conjugate VGF). *Consider a convex VGF associated with a compact convex set $\mathcal{M} \subset \mathbb{S}^m$ with $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M}\cap\mathbb{S}^m_+}$. The conjugate function is given by*

$$\Omega^*_{\mathcal{M}}(Y) = \tfrac{1}{4}\inf_{M} \left\{ \operatorname{tr}\left(YM^{\dagger}Y^T\right) \ :\ \operatorname{range}(Y^T) \subseteq \operatorname{range}(M)\,,\ M \in \mathcal{M}\cap\mathbb{S}^m_+ \right\}, \tag{19}$$

where $M^{\dagger}$ is the Moore–Penrose pseudoinverse of M.

Note that $\Omega^*(Y) = +\infty$ if the optimization problem in (19) is infeasible, i.e., if $\operatorname{range}(Y^T) \not\subseteq \operatorname{range}(M)$ for all $M \in \mathcal{M}\cap\mathbb{S}^m_+$. This condition is equivalent to $Y(I - MM^{\dagger}) \neq 0$ for all $M \in \mathcal{M}\cap\mathbb{S}^m_+$, where $MM^{\dagger}$ is the orthogonal projection onto the range of M. This can be seen using the generalized Schur complement.

Proof. From the assumption $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M}\cap\mathbb{S}_+}$, we get $\Omega^*_{\mathcal{M}} \equiv \Omega^*_{\mathcal{M}\cap\mathbb{S}_+}$. Define

$$f_{\mathcal{M}}(Y) = \tfrac{1}{4}\inf_{M,C} \left\{ \operatorname{tr}(C) \ :\ \begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \succeq 0\,,\ M \in \mathcal{M} \right\}. \tag{20}$$

The positive semidefiniteness constraint implies $M \succeq 0$, therefore, $f_{\mathcal{M}} \equiv f_{\mathcal{M}\cap\mathbb{S}_+}$. Its conjugate function is

$$\begin{aligned} f^*_{\mathcal{M}}(X) &= \sup_{Y}\ \sup_{M,C} \left\{ \langle X, Y\rangle - \tfrac{1}{4}\operatorname{tr}(C) \ :\ \begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \succeq 0\,,\ M \in \mathcal{M} \right\} \\ &= \sup_{M\in\mathcal{M}\cap\mathbb{S}_+}\ \sup_{Y,C} \left\{ \langle X, Y\rangle - \tfrac{1}{4}\operatorname{tr}(C) \ :\ \begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \succeq 0 \right\}. \end{aligned} \tag{21}$$

Consider the Lagrangian dual of the inner optimization problem over Y and C; e.g., see [43] for a review. Let $W \succeq 0$ be the dual variable with corresponding blocks, and write the Lagrangian as

$$L(Y, C, W) = \langle X, Y\rangle - \tfrac{1}{4}\operatorname{tr}(C) + \langle W_{11}, M\rangle + 2\langle W_{21}, Y\rangle + \langle W_{22}, C\rangle\,,$$

whose maximum value is finite only if $W_{21} = -\frac{1}{2}X$ and $W_{22} = \frac{1}{4}I$. Therefore, the dual problem is

$$\min_{W_{11}} \left\{ \langle W_{11}, M\rangle :\ \begin{bmatrix} W_{11} & -\frac{1}{2}X^T \\ -\frac{1}{2}X & \frac{1}{4}I \end{bmatrix} \succeq 0 \right\} = \min_{W_{11}} \left\{ \langle W_{11}, M\rangle :\ W_{11} \succeq X^TX \right\},$$

which is equal to $\langle M, X^TX\rangle$, and we used the generalized Schur complement in Lemma 3.3. Notice that the above dual problem is bounded below (nonnegative since $M \in \mathcal{M}\cap\mathbb{S}_+$) and strictly feasible; consider $W_{11} = 1 + \sigma^2_{\max}(X)$ which implies $[W_{11}, -\frac{1}{2}X^T; -\frac{1}{2}X, \frac{1}{4}I] \succ 0$. Therefore, the dual is attained and strong duality holds; e.g., see [46, Chapter 4]. By plugging the result into (21), we conclude $f^*_{\mathcal{M}} \equiv \Omega_{\mathcal{M}\cap\mathbb{S}_+}$.

The domain of optimization in the definition of f_M is a closed convex set, which we denote by $\mathcal{F}$. Then, $4f_M(Y) = \inf_{M,C} \operatorname{tr}(C) + \iota_{\mathcal{F}}(M, C)$ can be viewed as a parametric minimization. Denote this objective function by $g(M,C)$, which is convex. For any $\alpha \in \mathbb{R}$, the level set $\{(M,C) : \; g(M,C) \le \alpha\}$ is bounded because $g(M,C) \le \alpha$ implies $M \succeq 0$ and $C \succeq 0$, as well as $\operatorname{tr}(C) \le \alpha$, and $\mathcal{M}$ is compact. Therefore, by [40, Theorem 1.17 and Proposition 2.22], f_M is a proper, lower semicontinuous, convex function. This, by [40, Theorem 11.1], implies $f_{\mathcal{M}}^{**} = f_{\mathcal{M}}$. Therefore, $f_{\mathcal{M}}$ is equal to $\Omega^*_{\mathcal{M}\cap\mathbb{S}_+}$ which we showed to be equal to $\Omega^*_{\mathcal{M}}$. Using the generalized Schur complement, in Lemma 3.3, for the semidefinite constraint in (20) gives the desired representation in (19). ☐

While (19) can be cumbersome to implement, (20) is a convenient semidefinite representation of the same function. A set such as (3) illustrates this difference.

3.4. Related norms. Given a convex VGF $\Omega_{\mathcal{M}}$ with $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M}\cap\mathbb{S}_+}$, we have

$$\Omega_{\mathcal{M}}(X) = \sup_{M\in\mathcal{M}\cap\mathbb{S}_+} \operatorname{tr}\left(XMX^T\right) = \sup_{M\in\mathcal{M}\cap\mathbb{S}_+} \left\| XM^{1/2} \right\|_F^2 .$$

This representation shows that $\sqrt{\Omega_{\mathcal{M}}}$ is a seminorm: absolute homogeneity holds, and it is easy to prove the triangle inequality for the maximum of seminorms. The next lemma, which can be seen from [38, Corollary 15.3.2], generalizes this assertion.

LEMMA 3.5. *Suppose a function $\Omega : \mathbb{R}^{n\times m} \to \mathbb{R}$ is homogeneous of order 2, i.e., $\Omega(\theta X) = \theta^2\Omega(X)$ for all $\theta \in \mathbb{R}$. Then its square root $\|X\| := \sqrt{\Omega(X)}$ is a seminorm if and only if Ω is convex. If Ω is strictly convex then $\sqrt{\Omega}$ is a norm.*

Proof. First, assume that Ω is convex. By plugging in X and $-X$ in the definition of convexity for Ω we get $\Omega(X) \ge 0$, so the square root is well-defined. We show the triangle inequality $\sqrt{\Omega(X+Y)} \le \sqrt{\Omega(X)} + \sqrt{\Omega(Y)}$ holds for any X, Y. If $\Omega(X+Y)$ is zero, the inequality is trivial. Otherwise, for any $\theta \in (0,1)$ let $A = \frac{1}{\theta}X$, $B = \frac{1}{1-\theta}Y$, and use the convexity and second-order homogeneity of Ω to get

$$\Omega(X+Y) = \Omega(\theta A + (1-\theta)B) \le \theta\Omega(A) + (1-\theta)\Omega(B) = \tfrac{1}{\theta}\Omega(X) + \tfrac{1}{1-\theta}\Omega(Y). \tag{22}$$

If $\Omega(X) \ge \Omega(Y) = 0$, set $\theta = (\Omega(X) + \Omega(X+Y))/(2\Omega(X+Y)) > 0$. Notice that $\theta \ge 1$ provides $\Omega(X) \ge \Omega(X+Y)$ as desired. On the other hand, if $\theta < 1$, we can use it in (22) to get the desired result as

$$\Omega(X+Y) \le \tfrac{1}{\theta}\,\Omega(X) = \frac{2\Omega(X+Y)\Omega(X)}{\Omega(X+Y)+\Omega(X)} \implies \Omega(X) \ge \Omega(X+Y) .$$

And if $\Omega(X), \Omega(Y) \neq 0$, set $\theta = \sqrt{\Omega(X)}/(\sqrt{\Omega(X)} + \sqrt{\Omega(Y)}) \in (0,1)$ to get

$$\Omega(X+Y) \le \tfrac{1}{\theta}\,\Omega(X) + \tfrac{1}{1-\theta}\,\Omega(Y) = \left(\sqrt{\Omega(X)} + \sqrt{\Omega(Y)}\right)^2 .$$

Since $\sqrt{\Omega}$ satisfies the triangle inequality and absolute homogeneity, it is a seminorm. Notice that $\Omega(X) = 0$ does not necessarily imply $X = 0$, unless Ω is strictly convex.

Now, suppose that $\sqrt{\Omega}$ is a seminorm, hence, convex. The function f defined by $f(x) = x^2$ for $x \ge 0$ and $f(x) = 0$ for $x \le 0$ is nondecreasing, so the composition of these two functions is convex and equal to Ω. It is worth mentioning that one can alternatively use [38, Corollary 15.3.2] to prove the first part of the lemma. ☐

Considering $\|\cdot\|_{\mathcal{M}} \equiv \sqrt{\Omega_{\mathcal{M}}}$, we have $\frac{1}{2}\|\cdot\|_{\mathcal{M}}^2 \equiv \frac{1}{2}\Omega_{\mathcal{M}}$. Taking the conjugate function of both sides yields $\frac{1}{2}(\|\cdot\|_{\mathcal{M}}^*)^2 \equiv 2\Omega_{\mathcal{M}}^*$ where we used the order-2 homogeneity of $\Omega_{\mathcal{M}}$. Therefore, $\|\cdot\|_{\mathcal{M}}^* \equiv 2\sqrt{\Omega_{\mathcal{M}}^*}$. Given the representation of $\Omega_{\mathcal{M}}^*$ in Proposition 3.4, one can derive a similar representation for $\sqrt{\Omega_{\mathcal{M}}^*}$ as follows.

PROPOSITION 3.6. *For a convex VGF $\Omega_{\mathcal{M}}$ associated with a nonempty compact convex set $\mathcal{M}$ with $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M}\cap\mathbb{S}_+}$,*

$$\|Y\|_{\mathcal{M}}^* = 2\sqrt{\Omega_{\mathcal{M}}^*(Y)} = \tfrac{1}{2}\inf_{M,C}\left\{\operatorname{tr}(C) + \gamma_{\mathcal{M}}(M) :\ \begin{bmatrix} M & Y^T \\ Y & C\end{bmatrix} \succeq 0\right\}, \tag{23}$$

where $\gamma_{\mathcal{M}}(M) = \inf\{\lambda \geq 0 :\ M \in \lambda\mathcal{M}\}$ is the gauge function associated with $\mathcal{M}$.

Proof. The square root function, over positive numbers, can be represented in a variational form as $\sqrt{y} = \inf\left\{\alpha + \frac{y}{4\alpha} :\ \alpha > 0\right\}$. Without loss of generality, suppose $\mathcal{M}$ is a compact convex set containing the origin. Provided that $\Omega_{\mathcal{M}}^*(Y) > 0$, from the variational representation of a conjugate VGF function we have

$$\begin{aligned}\sqrt{\Omega_{\mathcal{M}}^*(Y)} &= \tfrac{1}{4}\inf_{M,\alpha>0}\left\{\alpha + \tfrac{1}{\alpha}\operatorname{tr}\left(YM^\dagger Y^T\right)\ :\ \operatorname{range}\left(Y^T\right) \subseteq \operatorname{range}(M)\,,\ M \in \mathcal{M}\cap\mathbb{S}_+\right\}\\ &= \tfrac{1}{4}\inf_{M,\alpha>0}\left\{\alpha + \operatorname{tr}\left(YM^\dagger Y^T\right)\ :\ \operatorname{range}\left(Y^T\right) \subseteq \operatorname{range}(M)\,,\ M \in \alpha(\mathcal{M}\cap\mathbb{S}_+)\right\}\end{aligned}$$

where we observed $M^\dagger/\alpha = (\alpha M)^\dagger$ and changed the variable αM to M, to get the second representation. The last representation is the same as (23), as the constraint restricts M to the PSD cone, for which $\gamma_{\mathcal{M}}(M) = \gamma_{\mathcal{M}\cap\mathbb{S}_+}(M)$. On the other hand, when $\Omega_{\mathcal{M}}^*(Y) = 0$, the claimed representation returns 0 as well because $0 \in \mathcal{M}$. □

For example, $\mathcal{M} = \{M \succeq 0 : \operatorname{tr}(M) \leq 1\}$ gives $\gamma_{\mathcal{M}}(M) = \operatorname{tr}(M)$ which if plugged into (23) yields the well-known semidefinite representation for the nuclear norm; [43, section 3.1].

3.5. Subdifferentials. In this section, we characterize the subdifferential of VGFs and their conjugate functions, as well as that of their corresponding norms. Due to the variational definition of a VGF where the objective function is linear in M, and the fact that $\mathcal{M}$ is assumed to be compact, it is straightforward to obtain the subdifferential of $\Omega_{\mathcal{M}}$ (e.g., see [28, Theorem 2]).

PROPOSITION 3.7. *For a convex VGF with $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M}\cap\mathbb{S}_+}$, the subdifferential at X is given by*

$$\partial\,\Omega_{\mathcal{M}}(X) = \operatorname{conv}\left\{2XM :\ \operatorname{tr}\left(XMX^T\right) = \Omega(X)\,,\ M \in \mathcal{M}\cap\mathbb{S}_+\right\}.$$

When $\Omega_{\mathcal{M}}(X) \neq 0$, we have $\partial\|X\|_{\mathcal{M}} = \frac{1}{2\|X\|_{\mathcal{M}}}\partial\,\Omega_{\mathcal{M}}(X)$.

As an example, the subdifferential of $\Omega(X) = \sum_{i,j=1}^m \overline{M}_{ij}|\mathbf{x}_i^T\mathbf{x}_j|$, from (4), is given by

$$\begin{aligned}\partial\,\Omega(X) = \big\{2XM :\ M_{ij} = \overline{M}_{ij}\operatorname{sign}\left(\mathbf{x}_i^T\mathbf{x}_j\right)\ \text{ if }\ \langle\mathbf{x}_i,\mathbf{x}_j\rangle \neq 0\,,\\ M_{ii} = \overline{M}_{ii}\,,\ |M_{ij}| \leq \overline{M}_{ij}\ \text{ otherwise}\big\}\,.\end{aligned} \tag{24}$$

PROPOSITION 3.8. *For a convex VGF with $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M}\cap\mathbb{S}_+}$, the subdifferential of its conjugate function is given by*

$$\begin{aligned}\partial\,\Omega_{\mathcal{M}}^*(Y) = \big\{\tfrac{1}{2}\left(YM^\dagger + W\right) :\ \Omega\left(YM^\dagger + W\right) = 4\Omega^*(Y) = \operatorname{tr}\left(YM^\dagger Y^T\right)\,,\\ \operatorname{range}\left(W^T\right) \subseteq \ker(M) \subseteq \ker(Y)\,,\ M \in \mathcal{M}\cap\mathbb{S}_+\big\}\,.\end{aligned} \tag{25}$$

When $\Omega_{\mathcal{M}}^*(Y) \neq 0$, we have $\partial\|Y\|_{\mathcal{M}}^* = \frac{2}{\|Y\|_{\mathcal{M}}^*}\partial\,\Omega_{\mathcal{M}}^*(Y)$.

The proof of Proposition 3.8 is given in the appendix.

Since $\partial\Omega^*(Y)$ is nonempty, for any choice of M_0, there exists a W such that $\frac{1}{2}(YM_0^\dagger + W) \in \partial\Omega^*(Y)$. However, finding such a W is not trivial. The following lemma characterizes the subdifferential as the solution set of a convex optimization problem involving Ω and affine constraints.

LEMMA 3.9. *Given Y and any choice of $M_0 \in \mathcal{M}\cap\mathbb{S}_+$ satisfying $Y(I-M_0M_0^\dagger)=0$ and $\Omega^*(Y) = \frac{1}{4}\operatorname{tr}(YM_0^\dagger Y^T)$, we have*

$$\partial\Omega^*(Y) = \operatorname{Arg\,min}_Z \left\{\Omega(Z) :\ Z = \tfrac{1}{2}\left(YM_0^\dagger + W\right),\ WM_0 = 0\right\}.$$

Assuming the optimality of M_0 establishes the second equality and the second inclusion in (25). Moreover, $\Omega(Z) \geq \operatorname{tr}(ZM_0Z^T) = \frac{1}{4}\operatorname{tr}(YM_0^\dagger Y^T) = \Omega^*(Y)$ for all feasible Z in the above. Note that $WM_0 = 0$ is equivalent to $\operatorname{range}(W^T) \subseteq \ker(M_0)$.

The characterization of the whole subdifferential is helpful for understanding optimality conditions, but algorithms only need to compute a single subgradient, which is easier than computing the whole subdifferential.

3.6. Composition of VGF and absolute values.

The characterization of the subdifferential allows us to establish conditions for convexity of $\Psi(X) = \Omega(|X|)$, defined in (14) as a regularization function for diversity. Our result in Corollary 3.12 is based on the following lemma.

LEMMA 3.10. *Given a continuous function $f : \mathbb{R}^n \to \mathbb{R}$, consider $h(\mathbf{x}) := f(|\mathbf{x}|)$ and $g(\mathbf{x}) := \min_{\mathbf{y}\geq|\mathbf{x}|} f(\mathbf{y})$, where the absolute values and inequalities are all entrywise. Then, the following hold.*
(a) $h^{**} \leq g \leq h$.
(b) *If f is convex then g is convex and $g = h^{**}$.*
(c) *If f is convex then h is convex if and only if $g = h$.*
(d) *If f is convex and f has an entrywise nonnegative subgradient at any entrywise nonnegative $\mathbf{x}$, then h is convex and $g = h$.*

Proof. (a) In $h^*(\mathbf{y}) = \sup_{\mathbf{x}} \{\langle \mathbf{x},\mathbf{y}\rangle - f(|\mathbf{x}|)\}$, the optimal $\mathbf{x}$ should have the same sign pattern as $\mathbf{y}$; hence $h^*(\mathbf{y}) = \sup_{\mathbf{x}\geq 0} \{\langle \mathbf{x}, |\mathbf{y}|\rangle - f(\mathbf{x})\}$. Next, we have

$$\begin{aligned} h^{**}(\mathbf{z}) &= \sup_{\mathbf{y}} \left\{\langle \mathbf{y},\mathbf{z}\rangle - \sup_{\mathbf{x}\geq 0}\{\langle \mathbf{x},|\mathbf{y}|\rangle - f(\mathbf{x})\}\right\} = \sup_{\mathbf{y}\geq 0}\inf_{\mathbf{x}\geq 0}\left\{\langle \mathbf{y},|\mathbf{z}|\rangle - \langle \mathbf{x},\mathbf{y}\rangle + f(\mathbf{x})\right\} \\ &\leq \inf_{\mathbf{x}\geq 0}\sup_{\mathbf{y}\geq 0}\left\{\langle \mathbf{y},|\mathbf{z}|\rangle - \langle \mathbf{x},\mathbf{y}\rangle + f(\mathbf{x})\right\} = \inf_{\mathbf{x}\geq|\mathbf{z}|} f(\mathbf{x}) = g(\mathbf{z}), \end{aligned}$$

where we invoke the minimax inequality; e.g., [38, Lemma 36.1]. This shows the first inequality in (a). The second inequality follows directly from the definition of g and h.

(b) Consider $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ and $\theta \in [0,1]$. For any $\varepsilon > 0$, there exist some $\mathbf{y}_i \geq |\mathbf{x}_i|$ for $i = 1,2$, for which $f(\mathbf{y}_i) \leq g(\mathbf{x}_i) + \varepsilon$. Then,

$$\theta\mathbf{y}_1 + (1-\theta)\mathbf{y}_2 \geq \theta|\mathbf{x}_1| + (1-\theta)|\mathbf{x}_2| \geq |\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2|.$$

Hence, by the definition of g and convexity of f,

$$g(\theta\mathbf{x}_1+(1-\theta)\mathbf{x}_2) \leq f(\theta\mathbf{y}_1+(1-\theta)\mathbf{y}_2) \leq \theta f(\mathbf{y}_1)+(1-\theta)f(\mathbf{y}_2) \leq \theta g(\mathbf{x}_1)+(1-\theta)g(\mathbf{x}_2)+\varepsilon.$$

Therefore, $g(\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2) \leq \theta g(\mathbf{x}_1) + (1-\theta)g(\mathbf{x}_2) + \varepsilon$ for any $\varepsilon > 0$, which implies that g is convex. It is a classical result that the epigraph of the biconjugate h^{**} is the closed convex hull of the epigraph of h; in other words, h^{**} is the largest lower semicontinuous convex function that is no larger than h (e.g., [38, Theorem 12.2]). Since g is convex and $h^{**} \leq g \leq h$, we must have $h^{**} = g$.

(c) Assuming h is a closed convex function, we have $h = h^{**}$ [38, Theorem 12.2], thus part (a) implies $h = g$. On the other hand, given a convex function f, part (b) states that $g = h^{**}$ is also convex. Hence, $h = g$ implies convexity of h.

(d) For any $\mathbf{x}$, any $\mathbf{y} \geq |\mathbf{x}|$, and an entrywise nonnegative subgradient of f at $|\mathbf{x}| \geq 0$, we have $f(\mathbf{y}) \geq f(|\mathbf{x}|) + \langle \mathbf{y} - |\mathbf{x}|, \mathbf{g}\rangle \geq f(|\mathbf{x}|)$. Therefore, $h(\mathbf{x}) = f(|\mathbf{x}|) = \min_{\mathbf{y}\geq|\mathbf{x}|} f(\mathbf{y}) = g(\mathbf{x})$ holds. Part (c) establishes the convexity of h. □

We can provide a variation of Lemma 3.10(c) for norms.

Lemma 3.11. *Consider any norm $\|\cdot\|$. Then, $\||\cdot|\|$ is a norm itself if and only if we have $\||\mathbf{x}|\| = \min_{\mathbf{y}\geq|\mathbf{x}|} \|\mathbf{y}\|$.*

Proof. First, suppose $\|\cdot\|_a := \||\cdot|\|$ is a norm, hence, it is an absolute and monotonic norm; e.g., see [3, Proposition 1.7]. Therefore, for any $\mathbf{y} \geq |\mathbf{x}|$ we have $\|\mathbf{y}\|_a \geq \|\mathbf{x}\|_a$ which gives $\min_{\mathbf{y}\geq|\mathbf{x}|} \|\mathbf{y}\|_a \geq \|\mathbf{x}\|_a$. Since $|\mathbf{x}|$ is feasible in this optimization and $\||\mathbf{x}|\|_a = \|\mathbf{x}\|_a$, we get the desired result: $\||\mathbf{x}|\| = \|\mathbf{x}\|_a = \min_{\mathbf{y}\geq|\mathbf{x}|} \|\mathbf{y}\|$. On the other hand, consider $g(\mathbf{x}) := \min_{\mathbf{y}\geq|\mathbf{x}|} \|\mathbf{y}\|$. We show that it is a norm. Clearly, g is nonnegative and homogenous, and $g(\mathbf{x}) = 0$ implies that $\|\mathbf{y}\| = 0$ for some $\mathbf{y} \geq |\mathbf{x}| \geq 0$ which implies $\mathbf{x} = 0$. The triangle inequality can be verified as

$$\begin{aligned} g(\mathbf{x}+\mathbf{z}) &= \min_{\mathbf{y}\geq|\mathbf{x}+\mathbf{z}|} \|\mathbf{y}\| \leq \min_{\mathbf{y}\geq|\mathbf{x}|+|\mathbf{z}|} \|\mathbf{y}\| = \min_{\mathbf{y}_1\geq|\mathbf{x}|\,,\,\mathbf{y}_2\geq|\mathbf{z}|} \|\mathbf{y}_1+\mathbf{y}_2\| \\ &\leq \min_{\mathbf{y}_1\geq|\mathbf{x}|\,,\,\mathbf{y}_2\geq|\mathbf{z}|} \|\mathbf{y}_1\| + \|\mathbf{y}_2\| = g(\mathbf{x}) + g(\mathbf{z}). \end{aligned}$$ □

Corollary 3.12. *For a convex VGF $\Omega_{\mathcal{M}}$, consider $\Psi(X) := \Omega_{\mathcal{M}}(|X|)$. Then,*

(a) *$\Psi(X)$ is a convex function of X if and only if $\Omega_{\mathcal{M}}(|X|) = \min_{Y\geq|X|} \Omega_{\mathcal{M}}(Y)$;*

(b) *provided that $\Omega_{\mathcal{M}}$ has an entrywise nonnegative subgradient at any entrywise nonnegative X, then $\Psi(X) = \min_{Y\geq|X|} \Omega_{\mathcal{M}}(Y)$ and it is convex in X.*

For example, consider the VGF defined in (4), and assume $\overline{M} \geq 0$ is chosen in a way that $\Omega_{\mathcal{M}}$ is convex. The subdifferential $\partial\Omega_{\mathcal{M}}$ is given in (24). For any $X \geq 0$, the inner product of any two columns of X is nonnegative which implies $2X\overline{M} \in \partial\Omega_{\mathcal{M}}(X)$. Since $2X\overline{M} \geq 0$, the condition of Lemma 3.10(d) is satisfied, and $\Psi(X) = \Omega_{\mathcal{M}}(|X|)$ is convex with an alternative representation $\Psi(X) = \min_{Y\geq|X|} \Omega_{\mathcal{M}}(Y)$. This specific function Ψ has been used in [44] for learning matrices with disjoint supports.

4. Proximal operators.

The proximal operator of a closed convex function $h(\cdot)$ is defined as $\operatorname{prox}_h(\mathbf{x}) = \arg\min_{\mathbf{u}} \{h(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|_2^2\}$, which always exists and is unique (e.g., [38, section 31]). Computing the proximal operator is the essential step in the proximal point algorithm [32, 39] and the proximal gradient methods (e.g., [37]). In each iteration of such algorithms, we need to compute $\operatorname{prox}_{\tau h}(\cdot)$, where $\tau > 0$ is a step size parameter. To simplify the presentation, assume $\mathcal{M} \subset \mathbb{S}_+^m$ and consider the associated VGF. Then,

$$\operatorname{prox}_{\tau\Omega}(X) = \arg\min_{Y} \max_{M\in\mathcal{M}} \ \tfrac{1}{2}\|Y - X\|_F^2 + \tau \operatorname{tr}\left(YMY^T\right). \tag{26}$$

Since $\mathcal{M} \subset \mathbb{S}_+$ is a compact convex set, and the objective is convex-concave, one can change the order of min and max (e.g., [38, Corollary 37.3.2]) and first solve for Y in terms of any given X and M, which gives $Y = X(I + 2\tau M)^{-1}$. Then, by plugging this optimal Y in the above optimization program, and after some algebraic manipulations, the optimal value of (26) will be equal to the optimal value of

$$\max_{M\in\mathcal{M}}\ \tfrac{1}{2}\left\|X\right\|_F^2-\tfrac{1}{2}\operatorname{tr}\left(X\left(I+2\tau M\right)^{-1}X^T\right)$$

for which we can find an optimal $M_0\in\mathcal{M}$ via

$$M_0\in\operatorname*{Arg\,min}_{M\in\mathcal{M}}\ \operatorname{tr}\left(X(I+2\tau M)^{-1}X^T\right)\,.$$

Plugging M_0 into the expression we derived before for the optimal Y establishes that the pair $(Y_{\text{opt}},M_{\text{opt}})=(X(I+2\tau M_0)^{-1},M_0)$ is an optimal solution in (26). Therefore, $\operatorname{prox}_{\tau\Omega}(X)=Y_{\text{opt}}=X(I+2\tau M_0)^{-1}$. To compute the proximal operator for the conjugate function Ω^*, one can use Moreau's formula (see, e.g., [38, Theorem 31.5]):

$$\operatorname{prox}_{\tau\Omega}(X)+\tau^{-1}\operatorname{prox}_{\tau^{-1}\Omega^*}(X)=X\,. \tag{27}$$

Next we discuss proximal operators of norms induced by VGFs (section 3.4). Since computing the proximal operator of a norm is related to the orthogonal projection onto the dual norm ball, i.e., $\operatorname{prox}_{\tau\|\cdot\|}(X)=X-\Pi_{\|\cdot\|^*\le\tau}(X)$, we can express the proximal operator of the norm $\|\cdot\|\equiv\sqrt{\Omega_{\mathcal{M}}(\cdot)}$ as

$$\operatorname{prox}_{\tau\|\cdot\|}(X)=X-\operatorname*{arg\,min}_{Y}\min_{M,C}\left\{\|Y-X\|_F^2:\ \operatorname{tr}(C)\le\tau^2,\ M\in\mathcal{M},\ \begin{bmatrix}M & Y^T\\ Y & C\end{bmatrix}\succeq 0\right\},$$

using (20) and (23). Moreover, plugging (23) into the definition of the proximal operator gives

$$\operatorname{prox}_{\tau\|\cdot\|^*}(X)=\operatorname*{arg\,min}_{Y}\min_{M,C}\left\{\|Y-X\|_F^2+\tau(\operatorname{tr}(C)+\gamma_{\mathcal{M}}(M)):\ \begin{bmatrix}M & Y^T\\ Y & C\end{bmatrix}\succeq 0\right\},$$

where $\gamma_{\mathcal{M}}(M)=\inf\{\lambda\ge 0:\ M\in\lambda\mathcal{M}\}$ is the gauge function associated with the nonempty convex set $\mathcal{M}$. The computational cost for computing proximal operators can be high in general (involving solving semidefinite programs); however, they may be simplified for special cases of $\mathcal{M}$. For example, a fast algorithm for computing the proximal operator of the VGF associated with the set $\mathcal{M}$ defined in (13) is presented in [33]. For general problems, due to the convex-concave saddle point structure in (26), we may use the mirror-prox algorithm [36] to obtain an inexact solution.

Left unitary invariance and QR factorization. As mentioned before, VGFs and their conjugates are left unitarily invariant. We can use this fact to simplify the computation of corresponding proximal operators when $n\ge m$. Consider the QR decomposition of a matrix $Y=QR$ where Q is an orthogonal matrix, $Q^TQ=QQ^T=I$, and $R=[R_Y^T\ 0_{m\times(n-m)}]^T$ is an upper triangular matrix with $R_Y\in\mathbb{R}^{m\times m}$. From the definition, we have $\Omega(Y)=\Omega(R_Y)$ and $\Omega^*(Y)=\Omega^*(R_Y)$. For the proximal operators, we can use the QR decomposition $X=Q[R_X^T\ 0]^T$ to get

$$\begin{aligned}\operatorname{prox}_{\tau\Omega^*}(X)&=\operatorname*{arg\,min}_{Y}\min_{M,C}\left\{\|Y-X\|_2^2+\tfrac{1}{2}\tau\operatorname{tr}(C):\ \begin{bmatrix}M & Y^T\\ Y & C\end{bmatrix}\succeq 0\,,\ M\in\mathcal{M}\right\}\\ &=Q\cdot\operatorname*{arg\,min}_{R\in\mathcal{R}}\min_{M,C}\left\{\|R-R_X\|_2^2+\tfrac{1}{2}\tau\operatorname{tr}(C):\ \begin{bmatrix}M & R^T\\ R & C\end{bmatrix}\succeq 0\,,\ M\in\mathcal{M}\right\},\end{aligned}$$

where $\mathcal{R}$ is the set of upper triangular matrices and the new PSD matrix is of size $2m$ instead of $n+m$ that we had before. The above equality uses two facts. First,

$$\begin{bmatrix}I_m & 0\\ 0 & Q^T\end{bmatrix}\begin{bmatrix}M & Y^T\\ Y & C\end{bmatrix}\begin{bmatrix}I_m & 0\\ 0 & Q\end{bmatrix}=\begin{bmatrix}M & R^T\\ R & Q^TCQ\end{bmatrix}\succeq 0\,, \tag{28}$$

where the right and left matrices in the multiplication are positive definite. Second, $\operatorname{tr}(C) = \operatorname{tr}(C')$, where $C' = Q^T C Q$ and assuming C' to be zero outside the first $m \times m$ block can only reduce the objective function. Therefore, we can ignore the last $n - m$ rows and columns of the above PSD matrix.

More generally, because of left unitary invariance, the optimal Y's in all of the optimization problems in this section have the same column space as the input matrix X; otherwise, a rotation as in (28) produces a feasible Y with a smaller value for the objective function.

5. Algorithms for optimization with VGF. In this section, we discuss optimization algorithms for solving convex minimization problems, in the form of (6), with VGF penalties. The proximal operators of VGFs we studied in the previous section are the key parts of proximal gradient methods (see, e.g., [5, 6, 37]). More specifically, when the loss function $\mathcal{L}(X)$ is smooth, we can iteratively update the variables $X^{(t)}$ as follows:

$$X^{(t+1)} = \operatorname{prox}_{\gamma_t \Omega}(X^{(t)} - \gamma_t \nabla \mathcal{L}(X^{(t)}))\,, \qquad t = 0\,, 1, 2, \dots,$$

where γ_t is a step size at iteration t. When $\mathcal{L}(X)$ is not smooth, then we can use subgradients of $\mathcal{L}(X^{(t)})$ in the above algorithm, or use the classical subgradient method on the overall objective $\mathcal{L}(X) + \lambda\Omega(X)$. In either case, we need to use diminishing step size and the convergence can be very slow. Even when the convergence is relatively fast (in terms of number of iterations), the computational cost of the proximal operator in each iteration can be very high.

In this section, we focus on loss functions that have a special form shown in (29). This form comes up in many common loss functions, some of which are listed later in this section, and allows for faster algorithms. We assume that the loss function $\mathcal{L}$ in (6) has the following representation:

$$\mathcal{L}(X) = \max_{\mathbf{g} \in \mathcal{G}} \; \langle X, \mathcal{D}(\mathbf{g}) \rangle - \hat{\mathcal{L}}(\mathbf{g})\,, \tag{29}$$

where $\hat{\mathcal{L}} : \mathbb{R}^p \to \mathbb{R}$ is a convex function, $\mathcal{G}$ is a convex and compact subset of $\mathbb{R}^p$, and $\mathcal{D} : \mathbb{R}^p \to \mathbb{R}^{n \times m}$ is a linear operator. This is also known as a Fenchel-type representation (see, e.g., [24]). Moreover, consider the infimal postcomposition [4, Definition 12.33] of $\hat{\mathcal{L}} : \mathcal{G} \to \mathbb{R}$ by $\mathcal{D}(\cdot)$, defined as

$$(\mathcal{D} \triangleright \hat{\mathcal{L}})(Y) = \inf\{\hat{\mathcal{L}}(G) : \; \mathcal{D}(G) = Y\,,\; G \in \mathcal{G}\}\,.$$

Then, the conjugate to this function is equal to $\mathcal{L}$. In other words, $\mathcal{L}(X) = \hat{\mathcal{L}}^*(\mathcal{D}^*(X))$, where $\hat{\mathcal{L}}^*$ is the conjugate function and $\mathcal{D}^*$ is the adjoint operator. The composition of a nonlinear convex loss function and a linear operator is very common for optimization of linear predictors in machine learning (e.g., [17]), which we will demonstrate with several examples later in this section.

With the variational representation of $\mathcal{L}$ in (29), and assuming $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M} \cap \mathbb{S}_+}$, we can write the VGF-penalized loss minimization problem (6) as a convex-concave saddle point optimization problem:

$$J_{\text{opt}} = \min_X \max_{M \in \mathcal{M} \cap \mathbb{S}_+,\, \mathbf{g} \in \mathcal{G}} \; \langle X, \mathcal{D}(\mathbf{g}) \rangle - \hat{\mathcal{L}}(\mathbf{g}) + \lambda \operatorname{tr}\left(X M X^T\right)\,. \tag{30}$$

If $\hat{\mathcal{L}}$ is smooth (while $\mathcal{L}$ may be nonsmooth) and the sets $\mathcal{G}$ and $\mathcal{M}$ are simple (e.g., admitting simple projections), we can solve problem (30) using a variety of primal-dual optimization techniques such as the *mirror-prox* algorithm [36, 24]. In section

5.1, we present a variant of the mirror-prox algorithm equipped with an adaptive line search scheme. Then in section 5.2, we present a preprocessing technique to transform problems of the form (30) into smaller dimensions, which can be solved more efficiently under favorable conditions.

Before diving into the algorithmic details, we examine some common loss functions and derive the corresponding representation (29) for them. This discussion will provide intuition for the linear operator $\mathcal{D}$ and the set $\mathcal{G}$ in relation to data and prediction.

Norm loss. Given a norm $\|\cdot\|$ and its dual $\|\cdot\|^*$, consider the squared norm loss

$$\mathcal{L}(\mathbf{x}) = \tfrac{1}{2}\|A\mathbf{x}-\mathbf{b}\|^2 = \max_{\mathbf{g}} \langle \mathbf{g}, A\mathbf{x}-\mathbf{b}\rangle - \tfrac{1}{2}(\|\mathbf{g}\|^*)^2 ,$$

where $A \in \mathbb{R}^{p\times n}$. In terms of the representation in (29), here we have $\mathcal{D}(\mathbf{g}) = A^T\mathbf{g}$, $\hat{\mathcal{L}}(\mathbf{g}) = \frac{1}{2}(\|\mathbf{g}\|^*)^2 + \mathbf{b}^T\mathbf{g}$, and $\mathcal{G} = \mathbb{R}^p$. Similarly, a norm loss can be represented as

$$\mathcal{L}(\mathbf{x}) = \|A\mathbf{x}-\mathbf{b}\| = \max_{\mathbf{g}} \left\{ \langle \mathbf{x}, A^T\mathbf{g}\rangle - \mathbf{b}^T\mathbf{g} : \|\mathbf{g}\|^* \le 1 \right\} ,$$

where we have $\mathcal{D}(\mathbf{g}) = A^T\mathbf{g}$, $\hat{\mathcal{L}}(\mathbf{g}) = \mathbf{b}^T\mathbf{g}$, and $\mathcal{G} = \{\mathbf{g} : \|\mathbf{g}\|^* \le 1\}$.

ε-insensitive (deadzone) loss. Another variant of the absolute loss function is called the ε-insensitive loss (e.g., see [42] for more details and applications) and can be represented, similarly to (29), as

$$\mathcal{L}_\varepsilon(x) = \max\{0, |x|-\varepsilon\} = \max_{\alpha,\beta} \{\alpha(x-\varepsilon) + \beta(-x-\varepsilon) : \alpha,\beta \ge 0,\ \alpha+\beta \le 1\} .$$

Hinge loss for binary classification. In binary classification problems, we are given a set of training examples $(\mathbf{a}_1, b_1), \ldots, (\mathbf{a}_N, b_N)$, where each $\mathbf{a}_s \in \mathbb{R}^n$ is a feature vector and $b_s \in \{+1,-1\}$ is a binary label. We would like to find $\mathbf{x} \in \mathbb{R}^n$ such that the linear function $\mathbf{a}_s^T\mathbf{x}$ can predict the sign of label b_s for each $s = 1,\ldots,N$. The hinge loss $\max\{0, 1-b_s(\mathbf{a}_s^T\mathbf{x})\}$ returns 0 if $b_s(\mathbf{a}_s^T\mathbf{x}) \ge 1$ and a positive loss growing with the absolute value of $b_s(\mathbf{a}_s^T\mathbf{x})$ when it is negative. The average hinge loss over the whole data set can be expressed as

$$\mathcal{L}(\mathbf{x}) = \frac{1}{N}\sum_{s=1}^{N} \max\left\{0, 1-b_s\left(\mathbf{a}_s^T\mathbf{x}\right)\right\} = \max_{\mathbf{g}\in\mathcal{G}} \langle \mathbf{g}, \mathbf{1}-\mathbf{D}\mathbf{x}\rangle ,$$

where $\mathbf{D} = [b_1\mathbf{a}_1, \ldots, b_N\mathbf{a}_N]^T$. Here, in terms of (29), we have $\mathcal{D}(\mathbf{g}) = -\mathbf{D}^T\mathbf{g}$, $\hat{\mathcal{L}}(\mathbf{g}) = -\mathbf{1}^T\mathbf{g}$, and $\mathcal{G} = \{\mathbf{g} \in \mathbb{R}^N : 0 \le g_s \le 1/N\}$.

Multiclass hinge loss. For multiclass classification problems, each sample $\mathbf{a}_s$ has a label $b_s \in \{1,\ldots,m\}$ for $s = 1,\ldots,N$. Our goal is to learn a set of classifiers $\mathbf{x}_1,\ldots,\mathbf{x}_m$, that can predict the labels b_s correctly. For any given example $\mathbf{a}_s$ with label b_s, we say the prediction made by $\mathbf{x}_1,\ldots,\mathbf{x}_m$ is correct if

$$\mathbf{x}_i^T\mathbf{a}_s \ge \mathbf{x}_j^T\mathbf{a}_s \quad \text{for all } (i,j) \in \mathcal{I}(b_s) , \tag{31}$$

where $\mathcal{I}(k)$, for $k = 1,\ldots,m$, characterizes the required comparisons to be made for any example with label k. Here are two examples.

1. *Flat multiclass classification:* $\mathcal{I}(k) = \{(k,j) : j \ne k\}$. In this case, the constraints in (31) are equivalent to the label $b_s = \arg\max_{i\in\{1,\ldots,m\}} \mathbf{x}_i^T\mathbf{a}_s$; see [45].
2. *Hierarchical classification.* In this case, the labels $\{1,\ldots,m\}$ are organized in a tree structure, and each $\mathcal{I}(k)$ is a special subset of the edges of the tree depending on the class label k; see section 6 and [13, 47] for further details.

Given the labeled data set $(\mathbf{a}_1, b_1), \ldots, (\mathbf{a}_N, b_N)$, we can optimize $X = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_m]$ to minimize the averaged multiclass hinge loss

$$\mathcal{L}(X) = \frac{1}{N}\sum_{s=1}^{N} \max\left\{0, 1 - \max_{(i,j)\in\mathcal{I}(b_s)} \left\{\mathbf{x}_i^T\mathbf{a}_s - \mathbf{x}_j^T\mathbf{a}_s\right\}\right\}, \tag{32}$$

which penalizes the amount of violation for the inequality constraints in (31).

In order to represent the loss function in (32) in the form of (29), we need some more notations. Let $p_k = |\mathcal{I}(k)|$, and define $E_k \in \mathbb{R}^{m\times p_k}$ as the incidence matrix for the pairs in $\mathcal{I}(k)$, i.e., each column of E_k, corresponding to a pair $(i,j) \in \mathcal{I}(k)$, has only two nonzero entries: -1 at the ith entry and $+1$ at the jth entry. Then the p_k constraints in (31) can be summarized as $E_k^T X^T \mathbf{a}_s \leq 0$. It can be shown that the multiclass hinge loss $\mathcal{L}(X)$ in (32) can be represented in the form (29) via

$$\mathcal{D}(\mathbf{g}) = -A\,\mathcal{E}(\mathbf{g}) \qquad \text{and} \qquad \hat{\mathcal{L}}(\mathbf{g}) = -\mathbf{1}^T\mathbf{g}\,,$$

where $A = [\mathbf{a}_1 \ \cdots \ \mathbf{a}_N]$ and $\mathcal{E}(\mathbf{g}) = [E_{b_1}\mathbf{g}_1 \ \cdots \ E_{b_N}\mathbf{g}_N]^T \in \mathbb{R}^{N\times m}$. Moreover, the domain of maximization in (29) is defined as

$$\mathcal{G} = \mathcal{G}_{b_1} \times \cdots \times \mathcal{G}_{b_N}, \quad \text{where} \quad \mathcal{G}_k = \left\{\mathbf{g} \in \mathbb{R}^{p_k} \ : \ \mathbf{g} \geq 0\,,\ \mathbf{1}^T\mathbf{g} \leq 1/N\right\}\,. \tag{33}$$

Combining the above variational form for multiclass hinge loss and a VGF as a penalty on X, we can reformulate the nonsmooth convex optimization problem $\min_X \{\mathcal{L}(X) + \lambda\Omega_{\mathcal{M}}(X)\}$ as the convex-concave saddle point problem

$$\min_X \ \max_{M\in\mathcal{M}\cap\mathbb{S}_+,\, \mathbf{g}\in\mathcal{G}} \ \mathbf{1}^T\mathbf{g} - \langle X, A\,\mathcal{E}(\mathbf{g})\rangle + \lambda\,\mathrm{tr}\left(XMX^T\right). \tag{34}$$

5.1. Mirror-prox algorithm with adaptive line search. The mirror-prox (MP) algorithm was proposed by Nemirovski [36] for approximating the saddle points of smooth convex-concave functions and solutions of variational inequalities with Lipschitz continuous monotone operators. It is an extension of the extragradient method [26], and more variants are studied in [23]. In this section, we first present a variant of the MP algorithm equipped with an adaptive line search scheme. Then explain how to apply it to solve the VGF-penalized loss minimization problem (30).

We describe the MP algorithm in the more general setup of solving variational inequality problems. Let $\mathcal{Z}$ be a convex compact set in Euclidean space $\mathcal{E}$ equipped with inner product $\langle\cdot,\cdot\rangle$, and $\|\cdot\|$ and $\|\cdot\|^*$ be a pair of dual norms on $\mathcal{E}$, i.e., $\|\xi\|^* = \max_{\|z\|\leq 1}\langle\xi, z\rangle$. Let $F : \mathcal{Z} \to \mathcal{E}$ be a Lipschitz continuous monotone mapping:

$$\forall\, z, z' \in \mathcal{Z} : \ \ \|F(z) - F(z')\|^* \leq L\|z - z'\| \ \text{ and, } \ \langle F(z) - F(z'), z - z'\rangle \geq 0\,. \tag{35}$$

The goal of the MP algorithm is to approximate a (strong) solution to the variational inequality associated with $(\mathcal{Z}, F)$: $\langle F(z^*), z - z^*\rangle \geq 0 \ \forall\, z \in \mathcal{Z}$. Let $\phi(x,y)$ be a smooth function that is convex in x and concave in y, and $\mathcal{X}$ and $\mathcal{Y}$ be closed convex sets. Then the convex-concave saddle point problem

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} \ \phi(x,y)$$

can be posed as a variational inequality problem with $z = (x,y)^T$, $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$, and

$$F(z) = \begin{bmatrix} \nabla_x\phi(x,y) \\ -\nabla_y\phi(x,y) \end{bmatrix}. \tag{36}$$

```
Algorithm: Mirror-Prox($z_1, \gamma_1, \varepsilon$)
  repeat
    $t := t + 1$
    repeat
      $\gamma_t := \gamma_t / c_{\text{dec}}$
      $w_t := P_{z_t}(\gamma_t F(z_t))$
      $z_{t+1} := P_{z_t}(\gamma_t F(w_t))$
    until $\delta_t \le 0$
    $\gamma_{t+1} := c_{\text{inc}} \gamma_t$
  until $V_{z_t}(z_{t+1}) \le \varepsilon$
  return $\bar{z}_t := (\sum_{\tau=1}^{t} \gamma_\tau)^{-1} \sum_{\tau=1}^{t} \gamma_\tau w_\tau$
```

FIG. 2. *Mirror-prox algorithm with adaptive line search. Here $c_{\text{dec}} > 1$ and $c_{\text{inc}} > 1$ are parameters controlling the decrease and increase of the step size γ_t in the line search trials. The stopping criterion for the line search is $\delta_t \le 0$, where $\delta_t = \gamma_t \langle F(w_t), w_t - z_{t+1} \rangle - V_{z_t}(z_{t+1})$. With an abuse of notation, we refer to this algorithm as MP.*

The setup of the MP algorithm requires a distance-generating function $h(z)$ which is compatible with the norm $\| \cdot \|$. In other words, $h(z)$ is subdifferentiable on the relative interior of $\mathcal{Z}$, denoted $\mathcal{Z}^o$, and is strongly convex with modulus 1 with respect to $\| \cdot \|$, i.e., for all $z, z' \in \mathcal{Z}$, we have $\langle \nabla h(z) - \nabla h(z'), z - z' \rangle \ge \|z - z'\|^2$. For any $z \in \mathcal{Z}^o$ and $z' \in \mathcal{Z}$, we can define the Bregman divergence at z as

$$V_z(z') = h(z') - h(z) - \langle \nabla h(z), z' - z \rangle \,,$$

and the associated proximity mapping as

$$P_z(\xi) = \operatorname*{arg\,min}_{z' \in \mathcal{Z}} \left\{ \langle \xi, z' \rangle + V_z(z') \right\} = \operatorname*{arg\,min}_{z' \in \mathcal{Z}} \left\{ \langle \xi - \nabla h(z), z' \rangle + h(z') \right\} \,.$$

With these definitions, we are now ready to present the MP algorithm in Figure 2. Compared with the original MP algorithm [36, 23], our variant employs an adaptive line search procedure to determine the step sizes γ_t for $t = 1, 2, \ldots$. We can exit the algorithm whenever $V_{z_t}(z_{t+1}) \le \epsilon$ for some $\epsilon > 0$. Under the assumptions in (35), the MP algorithm in Figure 2 enjoys the same $O(1/t)$ convergence rate as the one proposed in [36], but performs much faster in practice. The proof requires only simple modifications of the proof in [36, 23].

When $\hat{\mathcal{L}}$ is smooth and $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M} \cap \mathbb{S}_+}$, we can apply the MP algorithm to solve the saddle point problem in (30). Then, the gradient mapping in (36) becomes

$$F(X, M, \mathbf{g}) = \begin{bmatrix} \operatorname{vec}(2\lambda X M + \mathcal{D}(\mathbf{g})) \\ -\lambda \operatorname{vec}(X^T X) \\ \operatorname{vec}(\nabla \hat{\mathcal{L}}(\mathbf{g}) - \mathcal{D}^*(X)) \end{bmatrix} , \tag{37}$$

where $\mathcal{D}^*(\cdot)$ is the adjoint operator to $\mathcal{D}(\cdot)$. Assuming $\mathbf{g}$ lives in $\mathbb{R}^p$, computing F requires $O(nm^2 + nmp)$ operations for matrix multiplications. In section 5.2, we present a method that can potentially reduce the problem size by replacing n with $\min\{mp, n\}$. In the case of a support vector machine with the hinge loss as in our real-data numerical example, one can replace n by $\min\{N, mp, n\}$, where N is the number of samples.

The assumption $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M} \cap \mathbb{S}_+}$ provides us with a convex-concave saddle point optimization problem in (30). However, MP iterations for (30) require a projection onto $\mathcal{M} \cap \mathbb{S}_+$ (or, more generally, computation of the proximity mapping $P_z(\xi)$ corresponding to the mirror map we choose and a set $\mathcal{Z}$ defined via $\mathcal{M} \cap \mathbb{S}_+$), and such projections might be much more complicated than projection onto $\mathcal{M}$. In fact, while $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M}\cap\mathbb{S}_+}$ implies that the achieving matrix in $\sup_{M\in\mathcal{M}}\langle M, X^TX\rangle$ is always in $\mathcal{M} \cap \mathbb{S}_+$, we need a separate guarantee to be able to project onto $\mathcal{M}$ and $\mathcal{M} \cap \mathbb{S}_+$ interchangeably. We remark on a guarantee for this in the following, where Lemma 5.1 and Corollary 5.2 provide sufficient conditions for when projection of a PSD matrix onto $\mathcal{M}$ is equivalent to projection onto $\mathcal{M} \cap \mathbb{S}_+$.

LEMMA 5.1. *For any $G \succeq 0$, consider $P = \Pi_{\mathcal{M}}(G)$ and its Moreau decomposition with respect to the positive semidefinite cone as $P = P_+ - P_-$, where $P_+, P_- \succeq 0$ and $\langle P_+, P_-\rangle = 0$. Then, $P_+ \in \mathcal{M}$ implies $P_- = 0$.*

Proof. Apply the firm nonexpansive property of the projection operator onto a convex set [40] to $P = \Pi_{\mathcal{M}}(G)$ and $P_+ = \Pi_{\mathcal{M}}(P_+)$ (implied by $P_+ \in \mathcal{M}$). We get $\|P - P_+\|_F^2 \leq \langle P - P_+, G - P_+\rangle$ which implies $\langle P_-, G\rangle + \|P_-\|_F^2 \leq 0$. Moreover, for two PSD matrices G and P_- we have $\langle G, P_-\rangle \geq 0$. All in all, $P_- = 0$. □

COROLLARY 5.2. *Provided that for any $M \in \mathcal{M}$ we have $M_+ \in \mathcal{M}$, then $\Omega_{\mathcal{M}}$ is convex. Moreover, $\Pi_{\mathcal{M}}(G) \succeq 0$ for all $G \succeq 0$.*

Corollary 5.2 establishes an important property about the iterates of the MP algorithm with $h(\cdot) = \frac{1}{2}\|\cdot\|_2^2$ as the mirror map, corresponding to $P_z(\xi) = \Pi_{\mathcal{Z}}(z-\xi)$. If in the MP Algorithm in Figure 2 we initialize the part of z_1 corresponding to M's to be a PSD matrix, all of such parts in the iterations z_t and w_t remain PSD as (1) we add a PSD matrix (λX^TX from (37)) to the previous iteration, and (2) the projection onto $\mathcal{M}$ (which is not necessarily a subset of the PSD cone) ends up being a PSD matrix (by Corollary 5.2), hence it is equivalent to projection onto $\mathcal{M} \cap \mathbb{S}_+$. Notice that such a condition is required for applying the MP algorithm: the objective has to be convex-concave and the positive semidefiniteness of all iterations guarantees this property.

The above provides a glimpse into a more general approach in optimization with composite functions. While every proper closed convex function has a variational representation in terms of its conjugate function, namely, $\Omega_{\mathcal{M}}(X) = \sup_Y \ \langle X, Y\rangle - \Omega_{\mathcal{M}}^*(Y)$, such expressions do not necessarily offer any computational advantage. With a more clever exploitation of the structure, $\Omega_{\mathcal{M}}(X)$ can be seen as a composition of the support function $S_{\mathcal{M}}(\cdot)$ with a structure mapping $g(X) = X^TX$, as in (15). Then,

$$\begin{aligned}\min_X \quad \mathcal{L}(X) + \Omega_{\mathcal{M}}(X) &\equiv \min_X \sup_Y \quad \mathcal{L}(X) + \langle g(X), Y\rangle - S_{\mathcal{M}}^*(Y)\\ &\equiv \min_X \sup_{Y\in\mathcal{M}} \quad \mathcal{L}(X) + \langle X^TX, Y\rangle\,,\end{aligned}$$

where we use the fact that $S_{\mathcal{M}}^*(Y)$ is the indicator function for the set $\mathcal{M}$. This can be seen as an interpretation of how our proposed algorithm replaces proximal mapping computations for $\Omega_{\mathcal{M}}$ with projections onto $\mathcal{M}$ (proximal mapping for the indicator function for $\Omega_{\mathcal{M}}$). Of course, to be able to use convex optimization algorithms, we will need to establish results similar to Lemma 5.1 and Corollary 5.2.

5.2. A kernel trick (reduced formulation).

As we discussed earlier, when the loss function has the structure (29), we can write the VGF-penalized minimization problem as a convex-concave saddle point problem

$$J_{\text{opt}} = \min_{X\in\mathbb{R}^{n\times m}} \max_{\mathbf{g}\in\mathcal{G}} \ \langle X, \mathcal{D}(\mathbf{g})\rangle - \hat{\mathcal{L}}(\mathbf{g}) + \lambda\,\Omega(X)\,. \tag{38}$$

Since $\mathcal{G}$ is compact, Ω is convex in X, and $\hat{\mathcal{L}}$ is convex in $\mathbf{g}$, we can use a minimax theorem (e.g., [38, Corollary 37.3.2]) to interchange the max and min. Then, for any orthogonal matrix Q we have

$$\begin{aligned} J_{\text{opt}} &= \max_{\mathbf{g}\in\mathcal{G}} \min_{X} \ \langle X, \mathcal{D}(\mathbf{g})\rangle - \hat{\mathcal{L}}(\mathbf{g}) + \lambda\,\Omega(X) \\ &= \max_{\mathbf{g}\in\mathcal{G}} \min_{X} \ \left\langle Q^T X, Q^T\mathcal{D}(\mathbf{g})\right\rangle - \hat{\mathcal{L}}(\mathbf{g}) + \lambda\,\Omega\left(Q^T X\right) \\ &= \max_{\mathbf{g}\in\mathcal{G}} \min_{X} \ \left\langle X, Q^T\mathcal{D}(\mathbf{g})\right\rangle - \hat{\mathcal{L}}(\mathbf{g}) + \lambda\,\Omega(X)\,, \end{aligned} \tag{39}$$

where the second equality is due to the left unitary invariance of Ω, and we renamed the variable X to get the third equality. Observe that Q is an arbitrary orthogonal matrix in (39) and can be chosen in a clever way to simplify $\mathcal{D}$ as described in the following. Since $\mathcal{D}(\mathbf{g})$ is linear in $\mathbf{g}$, consider a representation as

$$\mathcal{D}(\mathbf{g}) = [D_1\mathbf{g} \ \cdots \ D_m\mathbf{g}] = [D_1 \ \cdots \ D_m](I_m\otimes\mathbf{g}) = \mathbf{D}(I_m\otimes\mathbf{g}) \tag{40}$$

for some $D_i \in \mathbb{R}^{n\times p}$ and $\mathbf{D}\in\mathbb{R}^{n\times mp}$. Then, express $\mathbf{D}$ as the product of an orthogonal matrix and a residue matrix, such as in QR decomposition $\mathbf{D} = QR$, where provided that $n > mp$, only the first mp rows of R can be nonzero (which will be denoted by R_1). Define $\mathcal{D}'(\mathbf{g}) = R_1(I_m\otimes\mathbf{g}) \in \mathbb{R}^{q\times m}$ for $q = \min\{mp, n\}$. Plugging the above choice of Q into (39) gives

$$J_{\text{opt}} = \max_{\mathbf{g}\in\mathcal{G}} \min_{X_1,X_2} \ \left\langle \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \begin{bmatrix} \mathcal{D}'(\mathbf{g}) \\ 0 \end{bmatrix} \right\rangle - \hat{\mathcal{L}}(\mathbf{g}) + \lambda\,\Omega\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\right).$$

Observe that setting X_2 to zero does not increase the value of Ω which allows for restricting the above to the subspace $X_2 = 0$ and getting

$$J_{\text{opt}} = \min_{X\in\mathbb{R}^{q\times m}} \max_{\mathbf{g}\in\mathcal{G}} \langle X, \mathcal{D}'(\mathbf{g})\rangle - \hat{\mathcal{L}}(\mathbf{g}) + \lambda\Omega(X) \tag{41}$$

whose X variable has $q = \min\{mp, n\}$ rows compared to n rows in (38).

Notice that while the evaluation of J_{opt} via (41) can potentially be more efficient, we are interested in recovering an *optimal point* X in (38) which is different from the optimal points in (41). Tracing back the steps we took from (38) to (41), we get

$$X_{\text{opt}}^{(38)} = Q\begin{bmatrix} X_{\text{opt}}^{(41)} \\ 0 \end{bmatrix}.$$

The special case of regularization with squared Euclidean norm has been understood and used before; e.g., see [41]. However, the above derivations show that we can get similar results when the regularization can be represented as a maximum of squared weighted Euclidean norms.

It is worth mentioning that the reduced formulation in (41) can be similarly derived via a dual approach; one has to take the dual of the loss-regularized optimization problem (e.g., see [40, Example 11.41]), use the left unitary invariance of the conjugate VGF to reduce $\mathcal{D}$ to $\mathcal{D}'$, and dualize the problem again, to get (41).

5.3. A representer theorem. A general loss-regularized optimization problem as in (6) where the loss admits a Fenchel-type representation and the regularizer is a strongly convex VGF (including all squared vector norms) enjoys a representer theorem (see, e.g., [41]). More specifically, the optimal solution is linearly related to the linear operator $\mathcal{D}$ in the representation of the loss. As mentioned before, for many common loss functions, $\mathcal{D}$ encodes the samples, which reduces the following proposition to the usual representer theorem.

THEOREM 5.3. *For a loss-regularized minimization problem as in* (6) *where* $\mathcal{M} \subset \mathbb{S}^m_{++}$ *and* $\mathcal{L}$ *admits a Fenchel-type representation as*

$$\mathcal{L}(X) = \max_{\mathbf{g}\in\mathcal{G}} \ \langle X, \mathcal{D}(\mathbf{g})\rangle - \hat{\mathcal{L}}(\mathbf{g}) = \max_{\mathbf{g}\in\mathcal{G}} \ \langle X, \mathbf{D}(I_m \otimes \mathbf{g})\rangle - \hat{\mathcal{L}}(\mathbf{g})\,,$$

the optimal solution $X_{\rm opt}$ *admits a representation of the form*

$$X_{\rm opt} = \mathbf{DC}$$

with a coefficient matrix $\mathbf{C}$ *given by* $\mathbf{C} = -\frac{1}{2\lambda}M_{\rm opt}^{-1}\otimes\mathbf{g}_{\rm opt}$ *(optimal solutions of* (30)*).*

Proof. Denote the optimal solution of (30) by $(X_{\rm opt}, \mathbf{g}_{\rm opt}, M_{\rm opt})$, which shares $(X_{\rm opt}, \mathbf{g}_{\rm opt})$ with (38). Consider the optimality condition as $-\frac{1}{\lambda}\mathcal{D}(\mathbf{g}_{\rm opt}) \in \partial\Omega(X_{\rm opt})$ which implies $X_{\rm opt} \in \partial\Omega^*(-\frac{1}{\lambda}\mathcal{D}(\mathbf{g}_{\rm opt}))$; e.g., see [40, Proposition 11.3]. Now, suppose $\mathcal{M} \subset \mathbb{S}^m_{++}$ which implies $\Omega_{\mathcal{M}}$ is strongly convex. Considering the characterization of a subdifferential for Ω^* from Proposition 3.8 as well as the representation of $\mathcal{D}(\mathbf{g})$ in (40) we get

$$X_{\rm opt} = -\tfrac{1}{2\lambda}\mathcal{D}(\mathbf{g}_{\rm opt})M_{\rm opt}^{-1} = -\tfrac{1}{2\lambda}\mathbf{D}(I_m\otimes\mathbf{g}_{\rm opt})M_{\rm opt}^{-1} = -\tfrac{1}{2\lambda}\mathbf{D}\left(M_{\rm opt}^{-1}\otimes\mathbf{g}_{\rm opt}\right)\,. \qquad \square$$

This representer theorem allows us to apply our methods in more general reproducing kernel Hilbert spaces by choosing a problem specific reproducing kernel; e.g., see [41, 47].

6. Numerical example.

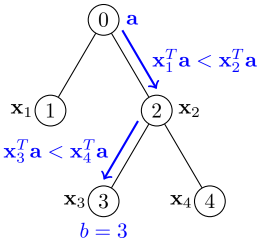
In this section, we discuss the application of VGFs in hierarchical classification to demonstrate the effectiveness of the presented approach in a real data experiment. More specifically, we compare the modified MP algorithm with adaptive line search presented in section 5.1 with the variant of the regularized dual averaging (RDA) method used in [47] in the text categorization application discussed in [47].

Let $(\mathbf{a}_1, b_1), \dots, (\mathbf{a}_N, b_N)$ be a set of labeled data, where each $\mathbf{a}_i \in \mathbb{R}^n$ is a feature vector and the associated $b_i \in \{1, \dots, m\}$ is a class label. The goal of multiclass classification is to learn a classification function $f : \mathbb{R}^n \to \{1, \dots, m\}$ so that, given any sample $\mathbf{a} \in \mathbb{R}^n$ (not necessarily in the training set), the prediction $f(\mathbf{a})$ attains a small classification error compared with the true label.

In hierarchical classification, the class labels $\{1, \dots, m\}$ are organized in a category tree, where the root of the tree is given the fictitious label 0 (see Figure 3(a)). For each node $i \in \{0, 1, \dots, m\}$, let $\mathcal{C}(i)$ be the set of children of i, $\mathcal{S}(i)$ be the set of siblings of i, and $\mathcal{A}(i)$ be the set of ancestors of i excluding 0 but including itself. A hierarchical linear classifier $f(\mathbf{a})$ is defined in Figure 3(b), which is parameterized by the vectors $\mathbf{x}_1, \dots, \mathbf{x}_m$ through a recursive procedure. In other words, an instance is labeled sequentially by choosing the category for which the associated vector outputs the largest score among its siblings, until a leaf node is reached. An example of this recursive procedure is shown in Figure 3(a). For the hierarchical classifier defined above, given an example $\mathbf{a}_s$ with label b_s, a correct prediction made by $f(\mathbf{a})$ implies that (31) holds with

$$\mathcal{I}(k) = \big\{(i,j)\,:\, j\in\mathcal{S}(i),\ i\in\mathcal{A}(k)\big\}\,.$$

Given a set of examples $(\mathbf{a}_1, b_1), \dots, (\mathbf{a}_N, b_N)$, we can train a hierarchical classifier parameterized by $X = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_m]$ by solving the problem $\min_X\{\mathcal{L}(X)+\lambda\Omega(X)\}$ with


0
a
$\mathbf{x}_1^T\mathbf{a} < \mathbf{x}_2^T\mathbf{a}$
$\mathbf{x}_1$ 1
2 $\mathbf{x}_2$
$\mathbf{x}_3^T\mathbf{a} < \mathbf{x}_4^T\mathbf{a}$
$\mathbf{x}_3$ 3
$\mathbf{x}_4$ 4
$b = 3$

(a)

$$f(\mathbf{a}) = \left\{ \begin{array}{l} \textbf{initialize } i := 0 \\ \textbf{while } \mathcal{C}(i) \text{ is not empty} \\ \qquad i := \underset{j\in\mathcal{C}(i)}{\arg\max}\ \mathbf{x}_j^T\mathbf{a} \\ \textbf{return } i \end{array} \right\}$$

(b)

FIG. 3. (a) *An example of hierarchical classification with four class labels* $\{1, 2, 3, 4\}$. *The instance* $\mathbf{a}$ *is classified recursively until it reaches the leaf node* $b = 3$, *which is its predicted label.* (b) *Definition of the hierarchical classification function.*

the loss function $\mathcal{L}(X)$ defined in (32) and an appropriate VGF penalty function $\Omega(X)$. As discussed in section 5, the training optimization problem can be reformulated as a convex-concave saddle point problem of the form (34) and solved by the MP algorithm described in section 5.1. In addition, we can use the reduction procedure discussed in section 5.2 to reduce computational costs.

As discussed in [47], one can assume a model where classification at different levels of the hierarchy rely on different features or different combinations of features. Therefore, authors in [47] proposed regularization with $|\mathbf{x}_i^T\mathbf{x}_j|$ whenever $j \in \mathcal{A}(i)$. A convex formulation of such a regularization function can be given in the form (4) with

$$\mathcal{M} = \left\{ M: \ M_{ii} = \overline{M}_{ii}\ ,\ |M_{ij}| = |\overline{M}_{ij}| \right\},$$

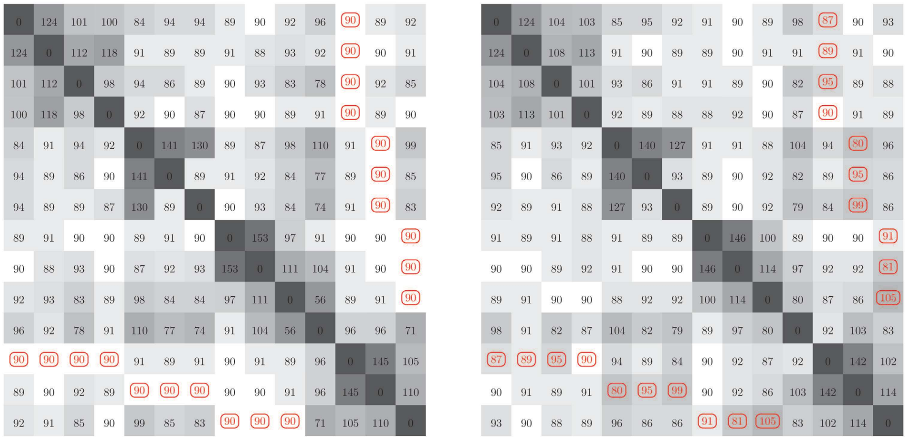
where the nonzero pattern of $\overline{M}$ corresponds to the pairs of ancestor-descendant nodes. According to (17), we have $\mathcal{M} \subset \mathbb{S}_+^m$ provided that $\lambda_{\min}(\widetilde{M}) \geq 0$.

As a real-world example, we consider the classification dataset Reuters Corpus Volume I, RCV1-v2 [30], which is an archive of over 800,000 manually categorized newswire stories and is available in libSVM. A subset of the hierarchy of labels in RCV1-v2, with $m = 23$ labels (18 leaves), is called ECAT and is used in our experiments. The samples and the classifiers are of dimension $n = 47236$. Last, there are 2196 training, and 69160 test samples available.

We solve the same loss-regularized problem as in [47], but using MP (discussed in section 5.1) instead of RDA. The regularization function is a VGF and is given in (4). A reformulation of the whole problem as a smooth convex-concave problem is given in (34). To obtain comparable results, we use the same matrix $\overline{M}$ and regularization parameter $\lambda = 1$ as in [47]. Note that in this experiment, $n = 47236$ while $m = 23$ and $p > 2196$, so the kernel trick is not particularly useful since n is not larger than mp.

Since we are solving the same problem as [47], the prediction error on test data will be the same as the error reported in this reference, which is better than the other methods. Moreover, one can look at the estimated classifiers and how well they validate the orthogonality assumption. Figure 4 compares the pairwise inner products of classifiers estimated by our approach for hierarchical classification and those estimated by the "transfer" method (see [47] for details on this method).

In the setup of the MP algorithm, we use $\frac{1}{2}\|\cdot\|_2^2$ as the mirror map which requires the least knowledge about the optimization problem (see [23] for the requirements when combining a number of mirror maps corresponding to different constraint sets



0	124	101	100	84	94	94	89	90	92	96	90	89	92
124	0	112	118	91	89	89	91	88	93	92	90	90	91
101	112	0	98	94	86	89	90	93	83	78	90	92	85
100	118	98	0	92	90	87	90	90	89	91	90	89	90
84	91	94	92	0	141	130	89	87	98	110	91	90	99
94	89	86	90	141	0	89	91	92	84	77	89	90	85
94	89	89	87	130	89	0	90	93	84	74	91	90	83
89	91	90	90	89	91	90	0	153	97	91	90	90	90
90	88	93	90	87	92	93	153	0	111	104	91	90	90
92	93	83	89	98	84	84	97	111	0	56	89	91	90
96	92	78	91	110	77	74	91	104	56	0	96	96	71
90	90	90	90	91	89	91	90	91	89	96	0	145	105
89	90	92	89	90	90	90	90	90	91	96	145	0	110
92	91	85	90	99	85	83	90	90	90	71	105	110	0

0	124	104	103	85	95	92	91	90	89	98	87	90	93
124	0	108	113	91	90	89	89	90	91	91	89	91	90
104	108	0	101	93	86	91	91	89	90	82	95	89	88
103	113	101	0	92	89	88	88	92	90	87	90	91	89
85	91	93	92	0	140	127	91	91	88	104	94	80	96
95	90	86	89	140	0	93	89	90	92	82	89	95	86
92	89	91	88	127	93	0	89	90	92	79	84	99	86
91	89	91	88	91	89	89	0	146	100	89	90	90	91
90	90	89	92	91	90	90	146	0	114	97	92	92	81
89	91	90	90	88	92	92	100	114	0	80	87	86	105
98	91	82	87	104	82	79	89	97	80	0	92	103	83
87	89	95	90	94	89	84	90	92	87	92	0	142	102
90	91	89	91	80	95	99	90	92	86	103	142	0	114
93	90	88	89	96	86	86	91	81	105	83	102	114	0

FIG. 4. *Pairwise angles (in degrees) between the estimated classifiers for dataset MCAT (part of RCV*1 − *v*2 [30]*) via (left) regularization by the VGF in* (4) *and (right) the "transfer" method (see* [47] *and references therein). The circled entries in red correspond to ancestor-descendant relations in the hierarchy of MCAT labels.*

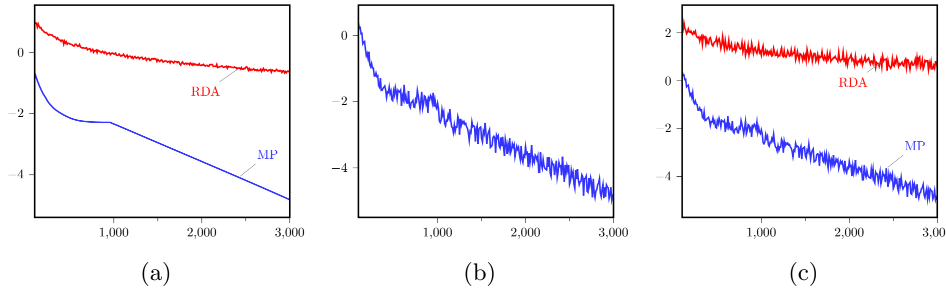


FIG. 5. *Convergence behavior for mirror-prox and RDA in our numerical experiment. (a) Average error over the m classifiers between each iteration and the final estimate,* $\|X_t - X_{final}\|_F$ *. (b) MP's gap* $V_{z_t}(z_{t+1})$*. (c) The value of loss function relative to the final value. For visualization purposes, all of the plots show data points at every 10 iterations. All vertical axes have a logarithmic scale.*

in the saddle point optimization problem). With this mirror map, the steps of MP only require orthogonal projection onto $\mathcal{G}$ and $\mathcal{M}$. The projection onto $\mathcal{G}$ in (33) boils down to separate projections onto N scaled simplexes (where the summation of entries is bounded by 1 and not necessarily equal to 1). Each projection amounts to zeroing out the negative entries followed by a projection onto the ℓ_1 unit norm ball (e.g., using the simple process described in [15]).

The variant of RDA proposed in [47] has a convergence rate of $O(\ln(t)/\sigma t)$ for the objective value, where σ is the strong convexity parameter of the objective. On the other hand, MP enjoys a convergence rate of $O(1/t)$ as given in [36]. Although there is a clear advantage to the MP method compared to RDA in terms of the theoretical guarantee, one should be aware of the difference between the notions of gap for the two methods. Figure 5(a) compares $\|X_t - X_{\text{final}}\|_F$ for MP and RDA using each one's own final estimate X_{final} . In terms of the runtime, we empirically observe that each iteration of MP takes about 3 times longer compared to RDA. However as is evident from Figure 5(a), MP is still much faster in generating a fixed-accuracy solution. Figure 5(b) illustrates the decay in the value of the gap for the MP method, $V_{z_t}(z_{t+1})$, which confirms the theoretical convergence rate of $O(1/t)$.

7. Discussion. In this paper, we introduce variational Gram functions, which include many existing regularization functions as well as important new ones. Convexity properties of this class, conjugate functions, subdifferentials, semidefinite representability, proximal operators, and other convex analysis properties are studied. By exploiting the structure in the loss function and the regularizer, namely, $\mathcal{L}(X) = \hat{\mathcal{L}}^*(\mathcal{D}^*(X))$ and $\Omega_{\mathcal{M}}(X) = S_{\mathcal{M}}(X^TX)$, we provide various tools and insight into such regularized loss minimization problems: By adapting the MP method [36], we provide a general and efficient optimization algorithm for VGF-regularized loss minimization problems. We establish a general kernel trick and a representer theorem for such problems. Finally, the effectiveness of VGF regularization as well as the efficiency of our optimization approach is illustrated by a numerical example on hierarchical classification for text categorization.

There are numerous directions for future research on this class of functions. One issue to address is how to systematically pick an appropriate set $\mathcal{M}$ when defining a new VGF for some new application. Statistical properties of VGFs, for example, the corresponding sample complexity, are of interest from a learning theory perspective. The presented kernel trick (which uses the left unitary invariance property of VGFs) can be potentially extended to other invariant regularizers. And last but not least, it is interesting to see if there is a variational Gram representation for any squared left unitarily invariant norm.

Appendix A. Proof of Proposition 3.8. First, let us simplify some notation. Throughout the proof, we denote $\frac{1}{2}\Omega$ by Ω, and $2\Omega^*$ by Ω^*. Denote by $\iota_{\mathcal{M}}(M)$ the indicator function of the set $\mathcal{M}$ which is 1 when $M \in \mathcal{M}$ and $+\infty$ otherwise. Since $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M}\cap\mathbb{S}_+}$, we assume $\mathcal{M} \subset \mathbb{S}_+$, with no loss of generality. Observe that $\Omega^*(Y) = \inf_M\ f(Y,M) + \iota_{\mathcal{M}}(M)$, where

$$f(Y,M) := \begin{cases} \frac{1}{2}\operatorname{tr}(YM^{\dagger}Y^T) & \text{if } \operatorname{range}(Y^T) \subseteq \operatorname{range}(M)\ ,\ M \succeq 0\,, \\ +\infty & \text{otherwise}\,. \end{cases}$$

Function $f(Y,M)$ coincides with $\sigma_{\mathcal{D}(A,B)}$, for $A = 0$ and $B = 0$ in [10, (2)]. Then, by [10, Corollary 4 and (8)], we get

$$\partial f(Y,M) = \left\{(Z,H):\ \tfrac{1}{2}Z^TZ + H \preceq 0\ ,\ Y = ZM\ ,\ \left\langle M, \tfrac{1}{2}Z^TZ + H\right\rangle = 0\right\}\ . \tag{42}$$

Since $g(Y,M) := f(Y,M) + \iota_{\mathcal{M}}(M)$ is convex, we can use results from parametric minimization, [40, Theorem 10.13], to get, for Y with $\Omega^*(Y) \neq +\infty$ and for any choice of $M_0 \in \mathcal{M}$ satisfying $\Omega^*(Y) = \frac{1}{2}\operatorname{tr}(YM_0^{\dagger}Y^T)$ and $Y(I - M_0M_0^{\dagger}) = 0$,

$$\begin{align}
\partial\Omega^*(Y) &= \{Z:\ (Z\,,0) \in \partial g(Y\,,M_0)\} \tag{43}\\
&= \{Z:\ (Z\,,-H) \in \partial f(Y\,,M_0)\,,\ H \in \partial\iota_{\mathcal{M}}(M_0)\,,\ \text{for some } H\} \tag{44}\\
&= \{Z:\ \tfrac{1}{2}Z^TZ \preceq H\,,\ Y = ZM_0\,, \nonumber\\
&\qquad \textstyle\sup_{M\in\mathcal{M}}\langle M\,,H\rangle = \langle M_0\,,H\rangle = \tfrac{1}{2}\operatorname{tr}\left(ZM_0Z^T\right)\ ,\ \text{for some } H\} \tag{45}\\
&= \{Z:\ \tfrac{1}{2}Z^TZ \preceq H\,,\ Y = ZM_0\,, \nonumber\\
&\qquad \textstyle\sup_{M\in\mathcal{M}}\langle M\,,H\rangle = \langle M_0\,,H\rangle = \tfrac{1}{2}\operatorname{tr}\left(ZM_0Z^T\right) = \Omega(Z)\,,\ \text{for some } H\} \tag{46}\\
&= \left\{Z:\ Y = ZM_0\,,\ \Omega(Z) = \tfrac{1}{2}\operatorname{tr}\left(ZM_0Z^T\right)\right\}\ . \tag{47}
\end{align}$$

Let us elaborate on these derivations. For (45), we used (42) as well as $\partial\iota_{\mathcal{M}}(M_0) \coloneqq \{G : \langle G, M - M_0\rangle \leq 0\,,\ \forall\, M \in \mathcal{M}\} = \{G : \ \sup_{M\in\mathcal{M}}\langle M, G\rangle = \langle M_0, G\rangle\}$, as $M_0 \in \mathcal{M}$. For (46), consider any $Z \in \partial\Omega^*(Y)$ and any H corresponding to Z in (45), and observe

$$\Omega(Z) \leq \sup\nolimits_{M\in\mathcal{M}}\langle M, H\rangle = \langle M_0, H\rangle = \tfrac{1}{2}\operatorname{tr}\left(ZM_0Z^T\right) \leq \Omega(Z)\,, \tag{48}$$

where the first inequality is due to $\frac{1}{2}Z^TZ \preceq H$. Hence inequalities in (48) hold with equality and (46) is established. Ignoring H in (46) establishes the forward inclusion for (47). On the other hand, for any Z in the right-hand side of (47) and for any Y'

$$\Omega^*(Y') \geq \langle Y', Z\rangle - \Omega(Z) = \langle Y', Z\rangle - \Omega^*(Y) = \langle Y' - Y, Z\rangle + \Omega^*(Y)\,,$$

where we used Fenchel's inequality, as well as the characterization of Z. Therefore, $Z \in \partial\Omega^*(Y)$. This establishes (47). Last, recall that M_0 is an achieving matrix in $\Omega^*(Y)$, which implies $Y(I - M_0M_0^\dagger) = 0$. This in turn implies that (e.g., see [1])

$$Y = ZM_0 \iff \exists W\ ;\ \ Z = YM_0^\dagger + W\ ,\ WM_0 = 0\,. \tag{49}$$

Moreover, (48) (with equalities), property $M_0 = M_0M_0^\dagger M_0$, and $Y = ZM_0$, imply

$$\Omega(Z) = \tfrac{1}{2}\operatorname{tr}\left(ZM_0Z^T\right) = \tfrac{1}{2}\operatorname{tr}\left(ZM_0M_0^\dagger M_0Z^T\right) = \tfrac{1}{2}\operatorname{tr}\left(YM_0^\dagger Y^T\right) = \Omega^*(Y)\,. \tag{50}$$

Combining (47), (49), and (50), yields

$$\begin{aligned}\partial\Omega^*(Y) = \Big\{Z = YM_0^\dagger + W :\ \Omega(Z) = \tfrac{1}{2}\operatorname{tr}\left(ZM_0Z^T\right) = \Omega^*(Y)\ ,\\ \operatorname{range}\left(Y^T\right) \subseteq \operatorname{range}(M_0) \subseteq \ker(W)\ ,\ M_0 \in \mathcal{M}\Big\}\end{aligned} \tag{51}$$

which is the claimed characterization (after we adjust for the $\frac{1}{2}$-rescaling we did in the beginning). Note that for an achieving M_0, range$(Y^T) \subseteq$ range(M_0) has to hold for the conjugate function to have a finite value.

It is worth mentioning that the introduction of H and then omitting it hints to the possibility of simpler proofs. We postpone this to future examinations.

Acknowledgment. We are grateful to the reviewers as their comments led us to improve both the content and the presentation of this final version.

REFERENCES

[1] A. Albert, *Conditions for positive and nonnegative definiteness in terms of pseudoinverses*, SIAM J. Appl. Math., 17 (1969), pp. 434–440.

[2] A. Argyriou, R. Foygel, and N. Srebro, *Sparse prediction with the k-support norm*, in Advances in Neural Information Processing Systems, Curran Associates, Red Hook, NY, 2012, pp. 1457–1465.

[3] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, *Optimization with sparsity-inducing penalties*, Found. Trends Mach. Learn., 4 (2012), pp. 1–106.

[4] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, New York, 2011.

[5] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[6] A. Beck and M. Teboulle, *Gradient-based algorithms with applications to signal-recovery problems*, in Convex Optimization in Signal Processing and Communications, Cambridge University Press, Cambridge, 2009, pp. 42–88.

[7] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust Optimization*, Princeton Ser. Appl. Math., Princeton University Press, Princeton, NJ, 2009.

[8] K. H. V. Booth and D. R. Cox, *Some systematic supersaturated designs*, Technometrics, 4 (1962), pp. 489–495.

[9] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

[10] J. V. Burke, Y. Gao, and T. Hoheisel, *Convex Geometry of the Generalized Matrix-Fractional Function*, preprint, arxiv.org/abs/1703.01363, 2017.

[11] X. Cai and X. Wang, *A note on the positive semidefinite minimum rank of a sign pattern matrix*, Electron. J. Linear Algebra, 26 (2013), pp. 345–356.

[12] C.-S. Cheng, *$E(s^2)$-optimal supersaturated designs*, Statist. Sinica, 7 (1997), pp. 929–939.

[13] O. Dekel, J. Keshet, and Y. Singer, *Large margin hierarchical classification*, in Proceedings of the 21st International Conference on Machine Learning, ACM, NY, 2004, pp. 27–34.

[14] X. V. Doan and S. Vavasis, *Finding the largest low-rank clusters with Ky Fan 2-k-norm and ℓ_1-norm*, SIAM J. Optim., 26 (2016), pp. 274–312.

[15] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, *Efficient projections onto the ℓ_1-ball for learning in high dimensions*, in Proceedings of the 25th International Conference on Machine Learning, ACM, New York, 2008, pp. 272–279.

[16] C. Giraud, *Low rank multivariate regression*, Electron. J. Stat., 5 (2011), pp. 775–799.

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Springer Ser. Statist., Springer, New York, 2009.

[18] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed., Cambridge University Press, Cambridge, 2013.

[19] R. Iyer and J. Bilmes, *Submodular point processes with applications to machine learning*, in Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, Proc. Mach. Learn. Res. 38, Proceedings of Machine Learning Research, 2015, pp. 388-397.

[20] L. Jacob, F. Bach, and J.-P. Vert, *Clustered multi-task learning: A convex formulation*, in Advances in Neural Information Processing Systems, Curran Associates, Red Hook, NY, 2008, pp. 745–752.

[21] A. Jalali, *Convex Optimization Algorithms and Statistical Bounds for Learning Structured Models*, Ph.D. thesis, University of Washington, Seattle, WA, 2016, pp. 57–95.

[22] D. Jayaraman, F. Sha, and K. Grauman, *Decorrelating semantic visual attributes by resisting the urge to share*, in 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, IEEE, Piscataway, NJ, 2014, pp. 1629–1636.

[23] A. Juditsky and A. Nemirovski, *First-order methods for nonsmooth convex large-scale optimization,* II*: Utilizing problems's structure*, in Optimization for Machine Learning, S. Sra, S. Nowozin, and S. J. Wright, eds., The MIT Press, Cambridge, MA, 2011, pp. 149–184.

[24] A. Juditsky and A. Nemirovski, *Solving variational inequalities with monotone operators on domains given by linear minimization oracles*, Math. Program., 156 (2016), pp. 221–256.

[25] K. C. Kiwiel, *Breakpoint searching algorithms for the continuous quadratic knapsack problem*, Math. Program., 112 (2008), pp. 473–491.

[26] G. M. Korpelevič, *An extragradient method for finding saddle points and for other problems*, Ekonom. Mat. Metody, 12 (1976), pp. 747–756.

[27] A. Kulesza and B. Taskar, *Determinantal point processes for machine learning*, Found. Trends Mach. Learn., 5 (2012), pp. 123–286.

[28] V. L. Levin, *The application of E. Helly's theorem in convex programming, problems of best approximation, and related questions*, Mat. Sb. (N.S.), 79 (1969), pp. 250–263.

[29] A. S. Lewis, *The convex analysis of unitarily invariant matrix functions*, J. Convex Anal., 2 (1995), pp. 173–183.

[30] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, *RCV*1*: A new benchmark collection for text categorization research*, J. Mach. Learn. Res., 5 (2004), pp. 361–397.

[31] J. Malkin and J. Bilmes, *Ratio semi-definite classifiers*, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Las Vegas, NV, IEEE, Piscataway, NJ, 2008, pp. 4113–4116.

[32] B. Martinet, *Régularisation d'inéquations variationnelles par approximations successives*, Rev. Française Inform. Rech. Opér., 4 (1970), pp. 154–158.

[33] A. M. McDonald, M. Pontil, and D. Stamos, *New perspectives on k-support and cluster norms*, J. Mach. Learn. Res., 17 (2016), pp. 5376–5413.

[34] C. A. Micchelli, J. M. Morales, and M. Pontil, *Regularizers for structured sparsity*, Adv. Comput. Math., 38 (2013), pp. 455–489.

[35] L. Mirsky, *A trace inequality of John von Neumann*, Monatsh. Math., 79 (1975), pp. 303–306.

[36] 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 J. Optim., 15 (2004), pp. 229–251.

[37] Yu. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program., 140 (2013), pp. 125–161.

[38] R. T. Rockafellar, *Convex Analysis*, Princeton Math. Ser. 28, Princeton University Press, Princeton, NJ, 1970.

[39] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.

[40] R. T. Rockafellar and Roger J.-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.

[41] B. Schölkopf, R. Herbrich, and A. J. Smola, *A generalized representer theorem*, in Proceedings of the Fourteenth Annual Conference on Computational Learning Theory, Amsterdam, The Netherlands, Springer, Berlin, 2001, pp. 416–426.

[42] A. J. Smola and B. Schölkopf, *A tutorial on support vector regression*, Statist. Comput., 14 (2004), pp. 199–222.

[43] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.

[44] K. Vervier, P. Mahé, A. D'Aspremont, J.-B. Veyrieras, and J.-P. Vert, *On learning matrices with orthogonal columns or disjoint supports*, in Machine Learning and Knowledge Discovery in Databases, Springer, Cham, Switzerland, 2014, pp. 274–289.

[45] J. Weston and C. Watkins, *Support vector machines for multi-class pattern recognition*, in Proceedings of the Sixth European Symposium on Artificial Neural Networks (ESANN), D-FAC70, Brussels, 1998, pp. 219–224.

[46] H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., *Handbook of Semidefinite Programming*, Internat. Ser. Oper. Res. Management Sci. 27, Kluwer Academic, Boston, MA, 2000.

[47] D. Zhou, L. Xiao, and M. Wu, *Hierarchical classification via orthogonal transfer*, in Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, International Machine Learning Society, Madison, WI, 2011, pp. 801–808.