# Variational Gram Functions: Convex Analysis and Optimization

Amin Jalali* Maryam Fazel† Lin Xiao‡

October 28, 2015

**Abstract**

We propose 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, estimating vectors with disjoint supports, and other applications. We study necessary and sufficient condtions under which a VGF is convex, and give a characterization of its subdifferential. In addition, we show how to compute its proximal operator, and discuss efficient optimization algorithms for some structured loss-minimization problems using VGFs. Numerical experiments are presented to demonstrate the effectiveness of VGFs and the associated optimization algorithms.

## 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 positions and orientations, and can serve as a measure for various properties such as orthogonality. In this paper, we consider a class of functions that aggregate the pairwise inner products in a variational form,

$$\Omega_{\mathcal{M}}(\mathbf{x}_1, \ldots, \mathbf{x}_m) = \max_{M \in \mathcal{M}} \sum_{i,j=1}^{m} M_{ij} \mathbf{x}_i^T \mathbf{x}_j \,, \tag{1.1}$$

where $\mathcal{M}$ is a compact subset 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

$$\Omega_{\mathcal{M}}(X) = \max_{M \in \mathcal{M}} \langle X^T X, M \rangle = \max_{M \in \mathcal{M}} \operatorname{tr}(XMX^T) \,, \tag{1.2}$$

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.

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

$$\mathcal{M} = \left\{ M : \ |M_{ij}| \leqslant \overline{M}_{ij}, \ i, j = 1, \ldots, m \right\}, \tag{1.3}$$

*Department of Electrical Engineering, University of Washington, Seattle, WA 98195. Email: amjalali@uw.edu
†Department of Electrical Engineering, University of Washington, Seattle, WA 98195. Email: mfazel@uw.edu
‡Machine Learning Groups, Microsoft Research, Redmond, WA 98053. Email: lin.xiao@microsoft.com

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,

$$\Omega_{\mathcal{M}}(X) = \max_{M\in\mathcal{M}} \sum_{i,j=1}^{m} M_{ij}\mathbf{x}_i^T\mathbf{x}_j = \sum_{i,j=1}^{m} \overline{M}_{ij}|\mathbf{x}_i^T\mathbf{x}_j|\,. \tag{1.4}$$

In other words, $\Omega_{\mathcal{M}}(X)$ is the weighted sum of the absolute values of the pairwise inner products. This function was proposed in [42] as a regularization function to promote orthogonality between linear classifiers in the context of hierarchical classification.

An important question from both the theoretical and algorithmic points of view is: what are the conditions on $\mathcal{M}$ such that a VGF is convex? Observe that the function $\mathrm{tr}(XMX^T)$ 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 parametrized by $M \in \mathcal{M}$ (see, e.g., [36, Theorem 5.5]). However, this is not a necessary condition. For example, the set $\mathcal{M}$ in (1.3) is not a subset of $\mathbb{S}_+^m$ unless $\overline{M} = 0$, but the VGF in (1.4) is convex provided that the *comparison matrix* of $\overline{M}$ (by negating the off-diagonal entries) is positive semidefinite [42]. In this paper, we give more careful analysis of the conditions for a VGF to be convex, and characterize its subdifferential and associated proximal operator.

Given a convex VGF, we can define a semi-norm[1] by taking its square root as

$$\|X\|_{\mathcal{M}} := \sqrt{\Omega_{\mathcal{M}}(X)} = \max_{M\in\mathcal{M}} \left( \sum_{i,j=1}^{m} M_{ij}\mathbf{x}_i^T\mathbf{x}_j \right)^{1/2}. \tag{1.5}$$

If $\mathcal{M} \subset \mathbb{S}_+^m$, then $\|X\|_{\mathcal{M}}$ is the pointwise maximum of the semi-norms $\|XM^{1/2}\|_F$ over all $M \in \mathcal{M}$. We call $\|X\|_{\mathcal{M}}$ a VGF-induced (semi-)norm.

VGFs and the associated norms can serve as penalty or regularization functions in optimization problem 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

$$\underset{X\in\mathbb{R}^{n\times m}}{\text{minimize}} \quad \mathcal{L}(X) + \lambda\,\Omega_{\mathcal{M}}(X)\,, \tag{1.6}$$

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 structure of $\mathcal{L}(X)$ and $\Omega(X)$ together to derive efficient optimization algorithms.

**Organization.** In Section 2, we give more examples of VGF and explain its connections with functions of Euclidean distance matrices and robust optimization. Section 3 studies the convexity of VGFs and their conjugates, semidefinite representability, VGF-induced norms and their subdifferentials. Their proximal operators are derived in Section 4. In Section 5, we study a class of structured loss minimization problems with VGF penalities, and show how to exploit their structure using the mirror-prox algorithm. Finally in Section 6, we present two numerical examples to illustrate the application of VGF: one on finding vectors with disjoint support, and the other on hierarchical classification.

[1] a semi-norm satisfies all the properties of a norm except definiteness; i.e. it can have zero value for a nonzero input.

**Notation.** We use $\mathbb{S}^m$ to denote 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. The symbol $\preceq$ represents the Loewner partial order unless subscripted by a specific cone, and $\langle\cdot,\cdot\rangle$ denotes the inner product of two vectors. We use capital letters for matrices and bold lower case letters for vectors. We use $X \in \mathbb{R}^{n\times m}$ and $\mathbf{x} = \mathrm{vec}(X) \in \mathbb{R}^{nm}$ interchangeably, with $\mathbf{x}_i$ denoting the $i$th column of $X$ and $\mathbf{x}_{(i)}^T$ being its $i$th row; i.e., $X = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_m]$ and $X^T = [\mathbf{x}_{(1)} \ \cdots \ \mathbf{x}_{(n)}]$. We use $\mathbf{1}$ and $\mathbf{0}$ to denote matrices or vectors of all ones and all zeros respectively, whose sizes would be clear from the context. The entry-wise absolute value of $X$ is denoted by $|X|$. We use $\|\cdot\|_p$ to denote the $\ell_p$ norm of the input vector or matrix, and $\|\cdot\|_F$ as the Frobenius norm (same as $\ell_2$ vector norm). The convex conjugate of a function $f$ is defined as $f^*(y) = \sup_y \langle x, y\rangle - f(x)$, and the dual norm of $\|\cdot\|$ is defined as $\|\mathbf{y}\|^* = \sup\{\langle \mathbf{x}, \mathbf{y}\rangle : \ \|\mathbf{x}\| \leqslant 1\}$. Finally $\arg\min$ ($\arg\max$) returns an optimal point to a minimization (maximization) program while $\mathrm{Arg}\min$ (or $\mathrm{Arg}\max$) is the set of all optimal points. The operator $\mathrm{diag}(\cdot)$ is interchangeably used to put a vector on the diagonal of a zero matrix of corresponding size, extract the diagonal entries of a matrix as a vector, or zeroing out the off-diagonal entries of a matrix. We use $f \equiv g$ to denote $f(x) = g(x)$ for all $x \in \mathrm{dom}(f) = \mathrm{dom}(g)$.

## 2 Examples and connections

In this section, we present examples of VGF corresponding to different choices of the set $\mathcal{M}$. The list includes some well known functions that can be expressed in the variational form of (1.1), as well as some new examples.

**Vector norms.** Any vector norm $\|\cdot\|$ on $\mathbb{R}^m$ is the square root of a VGF defined over $\mathbb{R}^{1\times m}$ with

$$\mathcal{M} = \{\mathbf{u}\mathbf{u}^T : \ \|\mathbf{u}\|^* \leqslant 1\}.$$

For a column vector $\mathbf{x} \in \mathbb{R}^m$, the VGF is given by

$$\Omega_{\mathcal{M}}(\mathbf{x}^T) = \max_{\mathbf{u}} \{\mathrm{tr}(\mathbf{x}^T\mathbf{u}\mathbf{u}^T\mathbf{x}) : \ \|\mathbf{u}\|^* \leqslant 1\} = \max_{\mathbf{u}} \{(\mathbf{x}^T\mathbf{u})^2 : \ \|\mathbf{u}\|^* \leqslant 1\} = \|\mathbf{x}\|^2.$$

Another example for $n = 1$, where $\mathcal{M}$ is a compact convex set of diagonal matrices with positive diagonals, has been first introduced in [30] and later discussed in [3]. The corresponding norm is defined as

$$\Omega_{\mathcal{M}}(\mathbf{x}^T) = \max_{\theta\in\mathrm{diag}(\mathcal{M})} \sum_{i=1}^m \theta_i x_i^2 = \|\mathbf{x}\|^2, \tag{2.1}$$

and $(\|\mathbf{x}\|^*)^2 = \min_{\theta\in\mathcal{H}} \sum_{i=1}^m \frac{1}{\theta_i} x_i^2$. The $k$-support norm [2], which is a norm used to encourage vectors to have $k$ or fewer nonzero entries, is an example of (2.1) corresponding to $\mathcal{M} = \{\mathrm{diag}(\theta) : 0 \leqslant \theta_i \leqslant 1, \ \mathbf{1}^T\theta = k\}$.

**Norms of the Gram matrix.** Given a symmetric nonnegative matrix $\overline{M}$, we can define a class of VGFs based on any vector norm $\|\cdot\|$ and its dual norm $\|\cdot\|_*$. Define the variational set as

$$\mathcal{M} = \{K \circ \overline{M} : \ \|K\|^* \leqslant 1, \ K^T = K\}, \tag{2.2}$$

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

$$\begin{aligned} \Omega_{\mathcal{M}}(X) = \max_{M \in \mathcal{M}} \langle M, X^T X \rangle &= \max_{\|K\|^* \leqslant 1} \langle K \circ \overline{M}, X^T X \rangle \\ = \max_{\|K\|^* \leqslant 1} \langle K, \overline{M} \circ (X^T X) \rangle &= \|\overline{M} \circ (X^T X)\|. \end{aligned} \tag{2.3}$$

The following are several concrete examples.

- If we let $\|\cdot\|^*$ in (2.2) be the $\ell_\infty$ norm, then $\mathcal{M} = \{M \ : \ |M_{ij}/\overline{M}_{ij}| \leqslant 1, \ i, j = 1, \ldots, m\}$, which is the same as in (1.3). 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 (1.4):

$$\Omega_{\mathcal{M}}(X) = \|\overline{M} \circ (X^T X)\|_1 = \sum_{i,j=1}^{m} \overline{M} |\mathbf{x}_i^T \mathbf{x}_j|$$

- If we use the $\ell_2$ norm in (2.2), then $\mathcal{M} = \left\{M \ : \ \sum_{i,j}^m (M_{ij}/\overline{M}_{ij})^2 \leqslant 1\right\}$. In this case, we have

$$\Omega(X) = \|\overline{M} \circ (X^T X)\|_F = \left( \sum_{i,j=1}^{m} \overline{M}_{ij} (\mathbf{x}_i^T \mathbf{x}_j) \right)^{1/2}. \tag{2.4}$$

  This function has been considered in multi-task learning [39], and also in the context of super-saturated designs [8, 13].

- We can use the $\ell_1$ norm in (2.2) to define $\mathcal{M} = \left\{M \ : \ \sum_{i,j}^m |M_{ij}/\overline{M}_{ij}| \leqslant 1\right\}$, which results in

$$\Omega(X) = \|\overline{M} \circ (X^T X)\|_\infty = \max_{i,j=1,\ldots,m} \overline{M}_{ij} |\mathbf{x}_i^T \mathbf{x}_j|. \tag{2.5}$$

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

Many other interesting examples can be constructed this way. For example, using group-$\ell_1$ norm of the Gram matrix can model *sharing* vs *competition*, which was considered in vision tasks [21]. We will revisit the above examples for 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 [26]) 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 by $\sigma(M)$ and let $\Theta = \{\sigma(M) : \ M \in \mathcal{M}\}$. Then we have

$$\Omega_{\mathcal{M}}(X) = \max_{M \in \mathcal{M}} \operatorname{tr}(X M X^T) = \max_{\theta \in \Theta} \sum_{i=1}^{\min(n,m)} \theta_i \, \sigma_i(X)^2, \tag{2.6}$$

as a result of Von Neumann's trace inequality [31]. Notice the similarity of the above to the VGF in (2.1). As an example, consider

$$\mathcal{M} = \{M : \ \alpha_1 \mathbf{I} \preceq M \preceq \alpha_2 \mathbf{I}, \ \operatorname{tr}(M) = \alpha_3\}, \tag{2.7}$$

where $0 < \alpha_1 < \alpha_2$ and $\alpha_3 \in [m\alpha_1,\ m\alpha_2]$ are given constants. The so called *spectral box-norm* [29] is the dual to the norm in (1.5) defined via this $\mathcal{M}$. Note that in this case, $\mathcal{M} \subset \mathbb{S}^m_+$ so it is convex. The square of this norm has been considered in [20] for clustering.

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

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

which is the pointwise maximum of a finite number of squared weighted Frobenius norms.

### 2.1 Diversification

VGFs can help in *diversifying* the columns of the input matrix; e.g. minimizing (1.4) 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 non-negative vectors have disjoint supports if and only if they are orthogonal to each other. Hence, using a VGF as (1.4) that promotes orthogonality, we can define

$$\Psi(X) = \Omega(|X|) \tag{2.8}$$

to promote disjoint supports among the columns of $X$; hence diversifying the supports of columns of $X$. Convexity of (2.8) will be discussed in Section 3.6.

### 2.2 Functions of Euclidean distance matrix

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

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

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 $i, j = 1, \ldots, m$ (see, e.g., [9, Section 8.3]). Let $\mathcal{A} = \{\operatorname{diag}(M) - M :\ M \in \mathcal{M}\}$. Then we have

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

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 $M = \operatorname{diag}(A\mathbf{1}) - A$ is diagonally dominant with nonnegative diagonal elements, hence positive semidefinite. However, this is not a necessary condition, the function can be convex without all $A$'s being entrywise nonegative. In Section 3 we will discuss more general conditions for convexity of VGFs. See [16] and references therein for applications of this VGF.

### 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

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

can be interpreted as seeking an $X$ with minimal worst-case value over an uncertainty set $\mathcal{M}$. Alternatively, 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

In this section, we study the convexity of VGFs, their conjugate functions and subdifferentials.

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

$$\Omega_{\mathcal{M}}(X) = \max_{M\in\mathcal{M}} \operatorname{tr}(XMX^T) = S_{\mathcal{M}}(X^TX) \tag{3.1}$$

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., [36, Section 13]). Therefore, by properties of the support function, we have

$$\Omega_{\mathcal{M}} \equiv \Omega_{\operatorname{conv}(\mathcal{M})}, \tag{3.2}$$

where $\operatorname{conv}(\mathcal{M})$ denotes the convex hull of $\mathcal{M}$. It is clear that the representation (i.e., the associated set $\mathcal{M}$) of a VGF is not unique. Henceforth, without loss of generality we assume $\mathcal{M}$ is convex unless explicitly noted otherwise. Moreover, while we have assumed $\mathcal{M}\subset\mathbb{S}^m$ is compact, we only need the maximum in (1.1) to be attained, which for example allows for a closed set $\mathcal{M}$ which is unbounded along any negative semidefinite direction.

As we mentioned in the introduction, a sufficient condition for the convexity of a VGF is $\mathcal{M}\subset\mathbb{S}^m_+$. The following theorem gives a necessary and sufficient condition for a VGF to be convex. Basically, it only requires the VGF to admit a representation with a set of PSD matrices.

**Theorem 3.1** *Suppose that $\mathcal{M}$ is compact. Then $\Omega_{\mathcal{M}}$ is convex if and only if for every $X$ there exists an $M\in\mathcal{M}\cap\mathbb{S}_+$ that achieves the maximum value in the definition of $\Omega_{\mathcal{M}}(X)$. In other words, $\Omega_{\mathcal{M}}$ is convex if and only if $\Omega_{\mathcal{M}}\equiv\Omega_{\mathcal{M}\cap\mathbb{S}_+}$.*

The above theorem means that a convex VGF is essentially the point-wise maximum of a family of squared weighted Frobenius norms. Moreover, the condition is independent of the dimension $n$. That is, once $\Omega_{\mathcal{M}}$ is convex on $\mathbb{R}^{n\times m}$ for some $n$, it is convex on $\mathbb{R}^{q\times m}$ for any $q\geqslant 1$. We postpone the proof until after Lemma 3.4, where we derive the conjugate functions of VGFs, since our proof uses conjugate functions.

While Theorem 3.1 gives a necessary and sufficient condition for the convexity of VGFs, it does not provide an effective procedure to check whether or not such a condition holds. 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., $\mathcal{M}=\operatorname{conv}\{M_1,\dots,M_p\}$. 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 [38, Section 8.E]. We define $\mathcal{M}_{\text{eff}}$ as a subset of $\{M_1,\dots,M_p\}$ with *minimal cardinality* such that its support function coincides with $\Omega_{\mathcal{M}}(X)$ for every $X$. That is, $\mathcal{M}_{\text{eff}}\subseteq\{M_1,\dots,M_p\}$ and $S_{\mathcal{M}}(X^TX)=S_{\mathcal{M}_{\text{eff}}}(X^TX)$ for all $X\in\mathbb{R}^{n\times m}$, but deleting any of member from $\mathcal{M}_{\text{eff}}$ violates this property. As an example,

considering $\mathcal{M} = \{M : \ |M_{ij}| \leqslant \overline{M}_{ij}, \ i,j = 1,\ldots,m\}$, which gives the function defined in (1.4), we have

$$\mathcal{M}_{\text{eff}} \subseteq \{M : \ M_{ii} = \overline{M}_{ii} \ , \ M_{ij} = \pm\overline{M}_{ij} \text{ for } i \neq j \} . \tag{3.3}$$

**Theorem 3.2** *For a polytope $\mathcal{M} \subset \mathbb{R}^{m\times m}$ , 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 $\max_{M\in\mathcal{M}_{\text{eff}}} \text{tr}(XMX^T) = \Omega_{\mathcal{M}}(X)$. Next, we prove necessity for any $\mathcal{M}_{\text{eff}}$. Here is an observation. Take any $M_i \in \mathcal{M}_{\text{eff}}$. If for every $X \in \mathbb{R}^{n\times m}$ with $\Omega(X) = \text{tr}(XM_iX^T)$ there exists another $M_j \in \mathcal{M}_{\text{eff}}$ with $\Omega(X) = \text{tr}(XM_jX^T)$, 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 $\Omega(X_i) = \text{tr}(X_iM_iX_i^T) > \text{tr}(X_iM_jX_i^T)$ for all $j \neq i$. Hence, for this $X_i$, $\Omega$ is twice continuously differentiable in a small neighborhood of $X_i$ with Hessian $\nabla^2\Omega(\text{vec}(X_i)) = M_i \otimes \mathbf{I}_n$, where $\otimes$ denotes the matrix Kronecker product. Since $\Omega$ is assumed to be convex, the Hessian has to be PSD which gives $M_i \succeq \mathbf{0}$. ∎

The definition of $\mathcal{M}_{\text{eff}}$ requires $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M}_{\text{eff}}}$, and the condition in Theorem 3.2 is $\mathcal{M}_{\text{eff}} \subset \mathbb{S}^m_+$. Comparing with Theorem 3.1, here we have $\mathcal{M}_{\text{eff}} \subset \mathcal{M} \cap \mathbb{S}^m_+$, which can be a strict inclusion (even $\text{conv}(\mathcal{M}_{\text{eff}})$ can be a strict subset of $\mathcal{M} \cap \mathbb{S}^m_+$). Next we give a few examples to illustrate the use of Theorem 3.2.

- We continue with the example defined in (1.4). Authors in [42] provided the necessary (when $n \geqslant m-1$) and sufficient condition for convexity using results from M-matrix theory. First, define the comparison matrix $\widetilde{M}$ associated to the 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 positive semidefinite, and this condition is also necessary when $n \geqslant m-1$ [42]. Theorem 3.2 provides an alternative and more general proof. Let $\lambda_{\min}(M)$ be the minimum eigenvalue of a symmetric matrix $M$. Considering the characterization of $\mathcal{M}_{\text{eff}}$ in (3.3), we have

$$\begin{aligned}\min_{M\in\mathcal{M}_{\text{eff}}} \lambda_{\min}(M) &= \min_{\substack{M\in\mathcal{M}_{\text{eff}}\\ \|\mathbf{z}\|_2=1}} \mathbf{z}^T M\mathbf{z} \ \geqslant \ \min_{\|\mathbf{z}\|_2=1} \sum_i \overline{M}_{ii}z_i^2 - \sum_{i\neq j} \overline{M}_{ij}|z_iz_j| \\ &= \min_{\|\mathbf{z}\|_2=1} |\mathbf{z}|^T\widetilde{M}|\mathbf{z}| \ \geqslant \ \lambda_{\min}(\widetilde{M}).\end{aligned} \tag{3.4}$$

  When $n \geqslant m-1$, 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 Appendix A.2 of [42]). On the other hand, Lemma 2.1(2) of [12] states that the existence of such a matrix implies $n \geqslant m-1$. Hence, $\widetilde{M} \in \mathcal{M}_{\text{eff}}$ if and only if $n \geqslant m-1$. Therefore, both inequalities in (3.4) 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 3.2, this is equivalent to the VGF in (1.4) being convex. If $n < m-1$, then $\widetilde{M}$ may not belong to $\mathcal{M}_{\text{eff}}$, thus $\widetilde{M} \succeq 0$ is only a sufficient condition for convexity for general $n$.

- Consider a box similar to the set $\mathcal{M}$ above which is not necessarily symmetric around the origin. More specifically, let $\mathcal{M} = \{M \in \mathbb{S}^m : \ M_{ii} = C_{ii} \ , \ |M - C| \leqslant D\}$ where $C$ represents a center which is symmetric and has zero diagonals, and $D$ is a symmetric nonnegative matrix. In this case, we have $\mathcal{M}_{\text{eff}} \subseteq \{M : \ M_{ii} = D_{ii} \ , \ M_{ij} = C_{ij} \pm D_{ij} \text{ for } i \neq j\}$. When used as a penalty function in applications, this may 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 two vectors is preferred. Similar to (3.4), we have

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

where $\widetilde{D}$ is the comparison matrix associated to $D$. Notice that $C$ has zero diagonal and it cannot be PSD. Hence, a sufficient condition for convexity of $\Omega_{\mathcal{M}}$ defined via such an asymmetric box is that $\lambda_{\min}(\widetilde{D}) + \lambda_{\min}(C) \geqslant 0$.

- Consider the VGF defined in (2.5), whose associated variational set is

$$\mathcal{M} = \{M \in \mathbb{S}^m : \sum_{(i,j):\, \overline{M}_{ij}\neq 0} |M_{ij}/\overline{M}_{ij}| \leqslant 1\,,\ M_{ij} = 0 \text{ if } \overline{M}_{ij} = 0\},$$

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 symmetric 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. Therefore, according to Theorem (3.2), convexity of $\Omega_{\mathcal{M}}$ requires these vertices do not belong to $\mathcal{M}_{\text{eff}}$. To this end, we need $\max\{\overline{M}_{ii}\|\mathbf{x}_i\|_2^2, \overline{M}_{jj}\|\mathbf{x}_j\|_2^2\} \geqslant \overline{M}_{ij}\mathbf{x}_i^T\mathbf{x}_j$ for all $i$ and $j$, and any $X \in \mathbb{R}^{n\times m}$, which is equivalent to $\overline{M}_{ii}\overline{M}_{jj} \geqslant \overline{M}_{ij}^2$ for all $i, j$. This can be satisfied if $\overline{M} \succeq \mathbf{0}$. However, regardless of the positive semidefiniteness of $\overline{M}$ and by the above argument, a convex $\Omega$ corresponds to $\mathcal{M}_{\text{eff}}$ that only contains diagonal matrices. Hence, any such function has to be simply expressible as $\Omega(X) = \max_{i=1,\ldots,m} \overline{M}_{ii}\|\mathbf{x}_i\|_2^2$.

### 3.2 A sufficient condition for more general sets

For the VGF defined in (2.4), the associated set $\mathcal{M}$ is given in (2.2) with the Frobenius norm, i.e.,

$$\mathcal{M} = \{K \circ \overline{M} :\ \|K\|_F \leqslant 1,\ K^T = K\},$$

In this case, $\mathcal{M}$ is not a polytope, but we can proceed with 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 $K = (\overline{M} \circ X^TX)/\|\overline{M} \circ X^TX\|_F$. Theorem 3.1 implies that convexity of $\Omega_{\mathcal{M}}$ requires $K \circ \overline{M} \succeq 0$, which is equivalent to $\overline{M} \circ \overline{M} \circ X^TX \succeq 0$. Since this needs to hold for every $X$, we must have $\overline{M} \circ \overline{M} \succeq 0$. Schur Product Theorem [19, Theorem 7.5.1] states that $\overline{M} \succeq 0$ is sufficient for this requirement to hold, hence it is also a sufficient condition for convexity of $\Omega_{\mathcal{M}}$.

As a concrete example of the above analysis, choosing $\overline{M} = \mathbf{1}_{m\times m}$ leads to $\Omega_{\mathcal{M}}(X) = \|X^TX\|_F = \|\sigma^2(X)\|_2$, which is convex in $X \in \mathbb{R}^{n\times m}$. Alternatively, we notice that the corresponding set is $\mathcal{M} = \{K : \|K\|_F \leqslant 1,\ K^T = K\}$, which is closed under left and right multiplications by orthogonal matrices. So $\Omega_{\mathcal{M}}$ is a spectral function as given by (2.6). In this case, the set $\Theta$ in (2.6) is the unit $\ell_2$ ball in $\mathbb{R}^m$, and we obtain $\Omega_{\mathcal{M}}(X) = \|\sigma^2(X)\|_2$.

In fact, a similar result can be stated for any *absolute norm*: if $\|\mathbf{x}\| = \||\mathbf{x}|\|$ for all $\mathbf{x} \in \mathbb{R}^m$, then $\Omega(X) = \|\sigma^2(X)\|$ is convex in $X \in \mathbb{R}^{n\times m}$. This is because the dual of an absolute norm is also an absolute norm [19] and

$$\Omega(X) = \|\sigma^2(X)\| = \max_{\|\theta\|^*\leqslant 1} \langle\theta, \sigma^2(X)\rangle = \max_{\|\theta\|^*\leqslant 1, \mathbf{0}\leqslant\theta} \langle\theta, \sigma^2(X)\rangle = \max_{\|\theta\|^*\leqslant 1, \mathbf{0}\leqslant\theta} \operatorname{tr}(X\operatorname{diag}(\theta)X^T)$$

is convex by Theorem 3.1. For general norms (not absolute), $\Omega(X) = \|\sigma^2(X)\|$ is convex in $X \in \mathbb{R}^{n\times m}$ if the projection of the dual norm ball on to the nonnegative orthant is a subset of the dual norm ball itself. This can be seen as a result of the following more general lemma.

**Lemma 3.3 (a sufficient condition)** *Let $\mathcal{M}_+$ be the orthogonal projection of all matrices in $\mathcal{M}$ onto the PSD cone, and $\mathcal{M} - \mathbb{S}_+$ be the Minkowski difference $\mathcal{M} - \mathbb{S}_+ = \{M - S : M \in \mathcal{M}, S \in \mathbb{S}_+\}$. Then $\Omega_{\mathcal{M}}$ is convex provided that $\mathcal{M}_+ \subseteq \mathcal{M} - \mathbb{S}_+$.*

*Proof.* 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 $M$. It is easy to see that for any $X$ we have $\operatorname{tr}(XMX^T) \leqslant \operatorname{tr}(XM_+X^T)$. Therefore,

$$\Omega_{\mathcal{M}}(X) = \max_{M\in\mathcal{M}} \operatorname{tr}(XMX^T) \leqslant \max_{M\in\mathcal{M}_+} \operatorname{tr}(XMX^T).$$

If $\mathcal{M}_+ \subseteq \mathcal{M} - \mathbb{S}_+$, then we get equality in the above equation and $\Omega_{\mathcal{M}}(X)$ is convex by Theorem 3.1. Notice that $\mathcal{M}_+ \subseteq \mathcal{M} - \mathbb{S}_+$ can hold while $\mathcal{M}_+ \nsubseteq \mathcal{M}$. ∎

In the previous example, for a general norm $\|\cdot\|$, we have $\Omega(X) = \|\sigma^2(X)\| = \max\{\langle\theta, \sigma^2(X)\rangle : \|\theta\|^* \leqslant 1\}$, and Lemma 3.3 provides a sufficient condition for convexity.

Similar to the proof of Lemma 3.3, one can check that another sufficient condition for convexity of a VGF is that all of the maximal points of $\mathcal{M}$ with respect to $\mathbb{S}_+$ are PSD. On the other hand, it is easy to see that the condition in Lemma 3.3 is not necessary. Consider $\mathcal{M} = \{M \in \mathbb{S}^2 : |M_{ij}| \leqslant 1\}$. Although the associated VGF is convex (because the comparison matrix is PSD), we have $\mathcal{M}_+ \nsubseteq \mathcal{M}$. As an example,

$$M = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \in \mathcal{M}, \quad \text{but} \quad M_+ \simeq \begin{bmatrix} 0.44 & .72 \\ .72 & 1.17 \end{bmatrix} \notin \mathcal{M}.$$

### 3.3 Conjugate function and proof of Theorem 3.1

For any function $\Omega$, the conjugate function is defined as $\Omega^*(Y) = \sup_X \langle X, Y\rangle - \Omega(X)$ and the transformation that maps $\Omega$ to $\Omega^*$ is called the Legendre-Fenchel transform (e.g., [36, Section 12]).

**Lemma 3.4 (conjugate VGF)** *Consider a convex VGF associated to a compact convex set $\mathcal{M}$. The conjugate function is given by*

$$\Omega^*_{\mathcal{M}}(Y) = \tfrac{1}{4} \inf_M \left\{\operatorname{tr}(YM^\dagger Y^T) \ : \ \operatorname{range}(Y^T) \subseteq \operatorname{range}(M),\ M \in \mathcal{M} \cap \mathbb{S}^m_+\right\} \tag{3.5}$$

*where $M^\dagger$ is the Moore-Penrose pseudoinverse of $M$.*

Note that $\Omega^*(Y)$ is $+\infty$ if the optimization problem in (3.5) is infeasible; i.e. if $Y(\mathbf{I} - MM^\dagger)$ is nonzero for all $M \in \mathcal{M}$, where $MM^\dagger$ is the orthogonal projection onto the range of $M$. This can be seen from results on generalized Schur complement; e.g. see Appendix A.5.5 in [9] or [11].

*Proof.* Applying the definition of conjugate function to a VGF gives

$$\Omega^*_{\mathcal{M}}(Y) = \sup_X \inf_{M\in\mathcal{M}} \langle X, Y\rangle - \operatorname{tr}(XMX^T).$$

Since $\mathcal{M}$ is compact and convex, we can change the order of sup and inf. Next, consider splitting the minimization over $\mathcal{M}$ into minimization over $\mathcal{M} \cap \mathbb{S}^m_+$ and $\mathrm{cl}(\mathcal{M} \backslash \mathbb{S}_+)$. For any fixed non-PSD $M$ in the latter set, the maximization with respect to $X$ is unbounded from above. Therefore,

$$\Omega^*_{\mathcal{M}}(Y) = \inf_{M \in \mathcal{M} \cap \mathbb{S}^m_+} \sup_X \langle X, Y \rangle - \mathrm{tr}(XMX^T)\,.$$

This gives $\Omega^*_{\mathcal{M}} \equiv \Omega^*_{\mathcal{M} \cap \mathbb{S}_+}$. Next we define

$$f_{\mathcal{M}}(Y) = \tfrac{1}{4} \inf_{M,C} \left\{ \mathrm{tr}(C) \;:\; \begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \succeq \mathbf{0}\,,\; M \in \mathcal{M} \right\}. \tag{3.6}$$

We note that the constraint on the right-hand side of the above definition automatically implies $M \succeq \mathbf{0}$. Therefore we have $f_{\mathcal{M}} \equiv f_{\mathcal{M} \cap \mathbb{S}_+}$. Its conjugate function is

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

Replacing the optimization problem over $Y$ and $C$ with the dual problem gives $f^*_{\mathcal{M}} \equiv \Omega_{\mathcal{M} \cap \mathbb{S}_+}$: consider a dual variable $W \succeq \mathbf{0}$ with corresponding blocks and write down the Lagrangian as

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

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

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

which is equal to $\langle M, X^T X \rangle$. Next, convexity and lower semi-continuity of $f_{\mathcal{M}}$ implies $f^{**}_{\mathcal{M}} = f_{\mathcal{M}}$ (e.g. [38, Theorem 11.1]). Therefore, $f_{\mathcal{M}}$ is equal to $\Omega^*_{\mathcal{M} \cap \mathbb{S}_+}$ which we showed to be the same as $\Omega^*_{\mathcal{M}}$. Taking the generalized Schur complement of the semidefinite constraint in (3.6) gives the desired representation in (3.5). ∎

Following Lemma 3.4, we are ready to prove Theorem 3.1.

*Proof.* [of Theorem 3.1] We know that $\Omega_{\mathcal{M} \cap \mathbb{S}_+}$ is convex because it is the pointwise maximum of convex quadratic functions parametrized by $M \in \mathcal{M} \cap \mathbb{S}_+$. On the other hand, the proof of Lemma 3.4 shows that $\Omega^*_{\mathcal{M}} \equiv \Omega^*_{\mathcal{M} \cap \mathbb{S}_+}$, which in turn gives $\Omega^{**}_{\mathcal{M}} \equiv \Omega^{**}_{\mathcal{M} \cap \mathbb{S}_+}$. Since both $\Omega_{\mathcal{M}}$ and $\Omega_{\mathcal{M} \cap \mathbb{S}_+}$ are proper, lower semi-continuous convex functions, they are equal to their biconjugates [38, Theorem 11.1]. Therefore, $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M} \cap \mathbb{S}_+}$. ∎

### 3.4 VGF-induced norms

Given a convex VGF $\Omega_{\mathcal{M}}$, Theorem 3.1 states that $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M} \cap \mathbb{S}_+}$, which in turn implies

$$\Omega_{\mathcal{M}}(X) = \sup_{M \in \mathcal{M} \cap \mathbb{S}_+} \mathrm{tr}(XMX^T) = \sup_{M \in \mathcal{M} \cap \mathbb{S}_+} \|XM^{1/2}\|_F^2 \geqslant 0\,.$$

The above representation of $\Omega_{\mathcal{M}}$ shows that $\sqrt{\Omega_{\mathcal{M}}}$ is a semi-norm: absolute homogeneity holds trivially and it is easy to prove the triangle inequality for the maximum of semi-norms. The next lemma generalizes this assertion, and we provide a proof in Appendix A.

**Lemma 3.5** *Suppose a function $\Omega : \mathbb{R}^{n\times m} \to \mathbb{R}$ is homogeneous of order 2, i.e., $\Omega(\theta X) = \theta^2\Omega(X)$. Then its square root $\|X\| = \sqrt{\Omega(X)}$ is a semi-norm if and only if $\Omega$ is convex. Moreover, provided that $\Omega$ is strictly convex, $\sqrt{\Omega}$ is a norm.*

**Dual Norm.** Given the representation of $\Omega^*_{\mathcal{M}}$ in Lemma 3.4, one can derive a similar representation for $\sqrt{\Omega^*_{\mathcal{M}}}$ as follows.

**Theorem 3.6** *Consider a convex VGF $\Omega_{\mathcal{M}}$ where $\mathcal{M}$ is a compact convex set. We have*

$$\sqrt{\Omega^*_{\mathcal{M}}(Y)} = \tfrac{1}{4} \inf_{M,\,C,\,\alpha} \left\{ \operatorname{tr}(C) + \alpha \;:\; \begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \succeq \mathbf{0}\,,\; \tfrac{1}{\alpha} M \in \mathcal{M} \cap \mathbb{S}_+ \right\}. \tag{3.7}$$

*Proof.* Let $(M_1^\star, C_1^\star)$ be an optimal solution to the minimization problem in (3.6), and let $(M_2^\star, C_2^\star, \alpha_2^\star)$ be an optimal solution to the minimization problem in (3.7). In addition, let $\mathrm{OPT}_1$ and $\mathrm{OPT}_2$ be the optimal values to these two minimization problems respectively. First, observe that $M_1 = \frac{1}{\alpha_2^*} M_2^*$ and $C_1 = \alpha_2^* C_2^*$ are feasible for the optimization program in (3.6). Hence,

$$\mathrm{OPT}_1 \leqslant \frac{1}{4}\alpha_2^* \operatorname{tr}(C_2^*) \leqslant \frac{1}{16}(\alpha_2^* + \operatorname{tr}(C_2^*))^2 = \mathrm{OPT}_2^2\,.$$

On the other hand, $\alpha_2 = 2\sqrt{\mathrm{OPT}_1}$, $M_2 = \alpha_2 M_1^*$ and $C_2 = \frac{1}{\alpha_2} C_1^*$ are feasible for the optimization program in (3.7). Hence,

$$\mathrm{OPT}_2 \leqslant \frac{1}{4}\left(\frac{4}{\alpha_2}\,\mathrm{OPT}_1 + \alpha_2\right) = \sqrt{\mathrm{OPT}_1}.$$

These two results give $\mathrm{OPT}_2 = \sqrt{\mathrm{OPT}_1}$. From the proof of Lemma 3.4, we have $\mathrm{OPT}_1 = \Omega^*_{\mathcal{M}}(Y)$, which implies that $\mathrm{OPT}_2 = \sqrt{\Omega^*_{\mathcal{M}}(Y)}$. This is the desired result. ∎

Considering $\|\cdot\| \equiv \sqrt{\Omega_{\mathcal{M}}}$, we have $\frac{1}{2}\Omega_{\mathcal{M}} \equiv \frac{1}{2}\|\cdot\|^2$. Taking the conjugate function of both sides yields $2\Omega^*_{\mathcal{M}} \equiv \frac{1}{2}(\|\cdot\|^*)^2$ where we used the order-2 homogeneity of $\Omega_{\mathcal{M}}$. Therefore,

$$\|\cdot\|^* \equiv 2\sqrt{\Omega^*_{\mathcal{M}}}\,. \tag{3.8}$$

### 3.5 Subdifferentials

In this section, we characterize the subdifferential of VGFs and their conjugate functions, as well as that of their induced 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 [18, Theorem 4.4.2]).

**Proposition 3.7** *The subdifferential of a convex VGF $\Omega_{\mathcal{M}}$ at a given matrix $X$ is given by*

$$\partial\,\Omega_{\mathcal{M}}(X) = \operatorname{conv}\left\{2XM :\; M \in \mathcal{M} \cap \mathbb{S}_+\,,\, \operatorname{tr}(XMX^T) = \Omega(X)\right\}. \tag{3.9}$$

*For its induced norm $\|x\|_{\mathcal{M}} \equiv \sqrt{\Omega_{\mathcal{M}}}$, we have $\partial\|X\|_{\mathcal{M}} = \frac{1}{\|X\|_{\mathcal{M}}}\partial\,\Omega_{\mathcal{M}}(X)$ if $\Omega_{\mathcal{M}}(X) \neq 0$.*

As an example, the subdifferential of $\Omega(X) = \sum_{i,j=1}^m \overline{M}_{ij}|\mathbf{x}_i^T\mathbf{x}_j|$, defined in (1.4), is given by

$$\partial\,\Omega(X) = \{XM :\; M_{ij} = \overline{M}_{ij}\mathrm{sign}(\mathbf{x}_i^T\mathbf{x}_j) \;\text{ if }\; \langle \mathbf{x}_i, \mathbf{x}_j\rangle \neq 0\,,\; |M_{ij}| \leqslant \overline{M}_{ij} \;\text{ otherwise}\}. \tag{3.10}$$

**Proposition 3.8** *For a convex VGF $\Omega_{\mathcal{M}}$, the subdifferential of its conjugate function is given by*

$$\partial\,\Omega^*_{\mathcal{M}}(Y) = \big\{\tfrac{1}{2}(YM^\dagger + W) : \Omega(YM^\dagger + W) = 4\Omega^*(Y) = \operatorname{tr}(YM^\dagger Y^T)\,, \\ \operatorname{range}(W^T) \subseteq \ker(M) \subseteq \ker(Y)\,,\ M \in \mathcal{M} \cap \mathbb{S}_+\big\}\,. \tag{3.11}$$

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

*Proof.* We use the results on subdifferentiation in parametric minimization [38, Section 10.C]. First, let's fix some notation. Throughout the proof, we denote $\frac{1}{2}\Omega$ by $\Omega$, and $2\Omega^*$ by $\Omega^*$. Moreover, denote by $I_{\mathcal{M}}(M)$ the indicator function of the set $\mathcal{M}$ which is 1 when $M \in \mathcal{M}$ and $+\infty$ otherwise. We will use $\mathcal{M}$ instead of $\mathcal{M} \cap \mathbb{S}_+$ for simplicity of the notation. Considering

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

we have $\Omega^*(Y) = \inf_M\ f(Y, M) + I_{\mathcal{M}}(M)$. For such a function, we can use results in [10, Theorem 4.8] to show that

$$\partial f(Y, M) = \operatorname{conv}\big\{(Z, -\tfrac{1}{2}Z^TZ) :\ Z = YM^\dagger + W\,,\ \operatorname{range}(W^T) \subseteq \ker(M)\big\}\,.$$

Since $g(Y, M) := f(Y, M) + I_{\mathcal{M}}(M)$ is convex, we can use the second part of Theorem 10.13 in [38]: for any choice of $M_0$ which is optimal in the definition of $\Omega^*(Y)$,

$$\partial\Omega^*(Y) = \{Z :\ (Z, \mathbf{0}) \in \partial g(Y, M_0)\}\,.$$

Therefore, for any $Z \in \partial\Omega^*(Y)$ we have

$$\tfrac{1}{2}Z^TZ \in \partial I_{\mathcal{M}}(M_0) = \{G :\ \langle G, M' - M_0\rangle \leqslant 0\,,\ \forall\, M' \in \mathcal{M}\}$$

(Here $\partial I_{\mathcal{M}}(M_0)$ is the normal cone of $\mathcal{M}$ at $M_0$.) This implies

$$\tfrac{1}{2}\operatorname{tr}(ZM'Z^T) \leqslant \tfrac{1}{2}\operatorname{tr}(ZM_0Z^T)$$

for all $M' \in \mathcal{M}$. Taking the supremum of the left hand side over all $M' \in \mathcal{M}$, we get

$$\Omega(Z) = \tfrac{1}{2}\operatorname{tr}(ZM_0Z^T) = \tfrac{1}{2}\operatorname{tr}(YM_0^\dagger Y^T) = \Omega^*(Y)\,.$$

where the second equality is implied by the condition $\operatorname{range}(W^T) \subseteq \ker(M_0)$ (which is equivalent to $M_0W^T = \mathbf{0}$). Alternatively, for any matrix $Z$ from the right hand side of (3.11), and any $W \in \mathbb{R}^{n\times m}$ we have

$$\Omega^*(W) \geqslant \langle W, Z\rangle - \Omega(Z) = \langle W, Z\rangle - \Omega^*(Y) = \langle W - Y, Z\rangle + \Omega^*(Y)$$

where we used Fenchel's inequality, as well as the characterization of $Z$. Therefore, $Z \in \partial\Omega^*(Y)$. This finishes the proof. Notice that, for an achieving $M$, $\ker(M) \subseteq \ker(Y)$ (or equivalently, $\operatorname{range}(W^T) \subseteq \operatorname{range}(M)$) has to hold for the conjugate function to be defined. ∎

Since $\partial\Omega^*(Y)$ is non-empty, for any choice of $M_0$, we can always find a $W$ such that $\frac{1}{2}(YM_0^\dagger + W) \in \partial\Omega^*(Y)$. However, finding such $W$ is not trivial. The following remark provides us with a computational, but not necessarily simple, approach to compute the subdifferential. 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; we will need to compute a subgradient of $\Omega^*$ as a part of the reduction technique presented in Section 5.2, which is also helpful in computing the proximal operator for $\Omega$ in Section 4.

**Remark 3.9** *Given $Y$ and an optimal $M_0$, which by optimality satisfies $\ker(M_0) \subseteq \ker(Y)$, we have*

$$\begin{aligned}\partial\Omega^*(Y) &= \operatorname*{Arg\,min}_Z \left\{\Omega(Z):\ Z = \tfrac{1}{2}(YM_0^\dagger + W),\ \operatorname{range}(W^T) \subseteq \ker(M_0) \subseteq \ker(Y)\right\} \\ &= \operatorname*{Arg\,min}_Z \left\{\Omega(Z):\ Z = \tfrac{1}{2}(YM_0^\dagger + W),\ WM_0M_0^\dagger = \mathbf{0}\right\}.\end{aligned}$$

*This is because for all feasible $Z$ we have $\Omega(Z) \geqslant \operatorname{tr}(ZM_0Z^T) = 4\Omega^*(Y)$.*

### 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 (2.8). Our result is based on the following Lemma.

**Lemma 3.10** *Given a function $f : \mathbb{R}^n \to \mathbb{R}$, consider $g(\mathbf{x}) = \min_{\mathbf{y}\geqslant|\mathbf{x}|} f(\mathbf{y})$, and $h(\mathbf{x}) = f(|\mathbf{x}|)$.*

*(a) $h^{**} \leqslant g \leqslant h$.*

*(b) If $f$ is convex then $g$ is convex and $g = h^{**}$.*

*Proof.* (a) First, $h^*(\mathbf{y}) = \sup_{\mathbf{x}} \{\langle \mathbf{x}, \mathbf{y}\rangle - f(|\mathbf{x}|)\} = \sup_{\mathbf{x}\geqslant\mathbf{0}} \{\langle \mathbf{x}, |\mathbf{y}|\rangle - f(\mathbf{x})\}$. Next, we have

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

This shows the first inequality in part (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 $\theta \in [0, 1]$. Suppose $g(\mathbf{x}_i) = f(\mathbf{y}_i)$ for some $\mathbf{y}_i \geqslant |\mathbf{x}_i|$, for $i = 1, 2$. In other words, $\mathbf{y}_i$ is the minimizer in the definition of $g(\mathbf{x}_i)$, for $i = 1, 2$. Then,

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

where the absolute values and inequalities are all entry-wise. By definition of $g$ and convexity of $f$

$$g(\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2) \leqslant f(\theta\mathbf{y}_1 + (1-\theta)\mathbf{y}_2) \leqslant \theta f(\mathbf{y}_1) + (1-\theta)f(\mathbf{y}_2) = \theta g(\mathbf{x}_1) + (1-\theta)g(\mathbf{x}_2),$$

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 semi-continuous convex function that is no larger than $h$ (e.g., [36, Theorem 12.2]). Since $g$ is convex and $h^{**} \leqslant g \leqslant h$, we must have $h^{**} = g$. ∎

**Corollary 3.11** *Let $\Omega_{\mathcal{M}}$ be a convex VGF. Then, $\Omega_{\mathcal{M}}(|X|)$ is a convex function of $X$ if and only if $\Omega_{\mathcal{M}}(|X|) = \min_{Y\geqslant|X|} \Omega_{\mathcal{M}}(Y)$.*

*Proof.* Let $\Omega_{\mathcal{M}}$ be the function $f$ in Lemma 3.10. Then we have $g(X) = \min_{Y\geqslant|X|} \Omega_{\mathcal{M}}(Y)$ and $h(X) = \Omega_{\mathcal{M}}(|X|)$. Since here $h$ is a closed convex function, we have $h = h^{**}$ [36, Theorem 12.2], thus part (a) of Lemma 3.10 implies $h = g$. On the other hand, given a convex function $f$, part (b) of Lemma 3.10 states that $g = h^{**}$ is also convex. Hence, $h = g$ implies convexity of $h$. ∎

**Lemma 3.12** *Let $\Omega_{\mathcal{M}}$ be a convex VGF. If $\partial\Omega_{\mathcal{M}}(X) \cap \mathbb{R}^{n\times m}_{+} \neq \varnothing$ holds for any $X \succcurlyeq \mathbf{0}$, then $\Psi(X) = \Omega_{\mathcal{M}}(|X|)$ is convex.*

*Proof.* Using the definition of subgradients for $\Omega$ at $|X|$ we have

$$\Omega(|X| + \Delta) \geqslant \Omega(|X|) + \sup\{\langle G, |X| + \Delta\rangle : \ G \in \partial\Omega \text{ at } |X|\}$$

where the right-most term is the directional derivative of $\Omega$ at $|X|$ in the direction $\Delta$. Provided that the assumption of the lemma holds, we get $\Omega(Y) \succcurlyeq \Omega(|X|)$ for all $Y \succcurlyeq |X|$. Therefore, $\Psi(X) = \Omega_{\mathcal{M}}(|X|) = \min_{Y\succcurlyeq|X|} \Omega_{\mathcal{M}}(Y)$. Corollary 3.11 establishes the convexity of $\Psi$. ∎

For example, consider the VGF $\Omega_{\mathcal{M}}$ defined in (1.4), and assume that it is convex. Its subdifferential $\partial\Omega_{\mathcal{M}}$ given in (3.10). For each $X \succcurlyeq 0$, the matrix product $X\overline{M} \succcurlyeq \mathbf{0}$ since $\overline{M}$ is also a nonnegative matrix, hence it belongs to $\partial\Omega_{\mathcal{M}}(X)$. Therefore the condition in the above lemma is satisfied, and as a consequence, the function $\Psi(X) = \Omega_{\mathcal{M}}(|X|)$ is convex and has an alternative representation $\Psi(X) = \min_{Y\succcurlyeq|X|} \Omega_{\mathcal{M}}(Y)$. This specific function $\Psi$ has been used in [40] for learning matrices with disjoint supports.

# 4 Proximal operators

The proximal operator of a closed convex function $h(\cdot)$ is defined as

$$\operatorname{prox}_h(\mathbf{x}) = \arg\min_{\mathbf{u}} \left\{h(\mathbf{u}) + \tfrac{1}{2}\|\mathbf{u} - \mathbf{x}\|_2^2\right\},$$

which always exists and is unique (e.g., [36, Section 31]). Computing the proximal operator is the essential step in the proximal point algorithm ([28, 37]) and the proximal gradient methods (e.g., [35]). In each iteration of such algorithms, we need to compute $\operatorname{prox}_{\tau h}()$ where $\tau > 0$ is a step size parameter. For a convex VGF $\Omega$ (for which, without loss of generality, we assume $\mathcal{M} \subset \mathbb{S}^m_+$), we have

$$\operatorname{prox}_{\tau\Omega}(X) = \arg\min_{Y} \max_{M\in\mathcal{M}} \left\{\tfrac{1}{2}\|Y - X\|_F^2 + \tau \operatorname{tr}(YMY^T)\right\}. \tag{4.1}$$

If $\mathcal{M}$ is a compact convex set, one can change the min and max and first solve for $Y$ in terms of any given $X$ and $M$, which gives $Y = X(I + 2\tau M)^{-1}$. Then we can find the optimal $M_0 \in \mathcal{M}$ given $X$ as

$$M_0 = \arg\min_{M\in\mathcal{M}} \operatorname{tr}\left(X(\mathbf{I} + 2\tau M)^{-1}X^T\right).$$

which gives $\operatorname{prox}_{\tau\Omega}(X) = X(\mathbf{I} + 2\tau M_0)^{-1}$. To compute the proximal operator for the conjugate function $\Omega^*$, one can use the Moreau's formula (see, e.g., [36, Theorem 31.5]):

$$\operatorname{prox}_{\tau\Omega}(X) + \tau^{-1}\operatorname{prox}_{\tau^{-1}\Omega^*}(X) = X. \tag{4.2}$$

Next we discuss proximal operators of VGF-induced norms. Since computing the proximal operator of a norm is equivalent to projection onto the dual norm ball, we can express the proximal operator of the norm $\|\cdot\| \equiv \sqrt{\Omega(\cdot)}$ as

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

where we used the representation of conjugate VGF in (3.6), and the characterization of the dual norm is (3.8). On the other hand, using the definition of proximal operator for the dual norm computed via (3.7) we have

$$\operatorname{prox}_{\tau\|\cdot\|^*}(X) = \arg\min_{Y} \min_{M,C,\alpha} \left\{ \|Y - X\|_F^2 + \tau(\operatorname{tr}(C) + \alpha) : \begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \succeq \mathbf{0}\,,\ \tfrac{1}{\alpha} M \in \mathcal{M} \right\}.$$

If $M$'s are invertible we can use the Schur complement and solve for $\alpha$ to get

$$\operatorname{prox}_{\tau\|\cdot\|^*}(X) = \arg\min_{Y} \min_{M} \left\{ \|Y - X\|_F^2 + 2\tau \|Y M^{-1/2}\|_F : \ M \in \mathcal{M} \right\}.$$

The computational cost involved in the above expressions for proximal operators can be high in general (involving solving semidefinite programs). However, we might be able to simplify them 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 (2.7) is presented in [29]. For general problems, due to the convex-concave saddle point structure in (4.1), we may use the mirror-prox algorithm [33] to obtain an inexact solution.

**Left unitarily invariance and QR factorization.** Suppose $n \geqslant m$. Consider the QR decomposition of a matrix $Y = QR$ where $Q$ is an orthogonal matrix with $Q^TQ = QQ^T = \mathbf{I}$ and $R = [R_Y^T\ \mathbf{0}]^T$ is an upper triangular matrix with $R_Y \in \mathbb{R}^{m\times m}$. Then for any VGF $\Omega$, we have

$$\Omega(Y) = \Omega(R_Y), \qquad \Omega^*(Y) = \Omega^*(R_Y)\,. \tag{4.3}$$

Notice that the semidefinite matrix in computing $\Omega^*(R_Y)$ via (3.6) is now $2m$-dimensional (independent of $n$). In fact, the matrix in (3.6) is PSD if and only if

$$\begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & Q^T \end{bmatrix} \begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & Q \end{bmatrix} = \begin{bmatrix} M & R^T \\ R & Q^TCQ \end{bmatrix} \succeq \mathbf{0}\,. \tag{4.4}$$

Moreover, $\operatorname{tr}(C) = \operatorname{tr}(C')$ where $C' = Q^TCQ$ 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 to get $\Omega^*(Y) = \Omega^*(R_Y)$.

Similarly for the proximal operators, we can simply plug in $R_X$ from the QR decomposition $X = Q[R_X^T\ \mathbf{0}]^T$ and get

$$\operatorname{prox}_{\tau\Omega^*}(X) = Q \cdot \arg\min_{R} \min_{M} \left\{ \|R - R_X\|_2^2 + \tfrac{1}{2}\tau \operatorname{tr}(C) : \begin{bmatrix} M & R^T \\ R & C \end{bmatrix} \succeq \mathbf{0}\,,\ M \in \mathcal{M} \right\}, \tag{4.5}$$

where $R$ is restricted to be an upper triangular matrix and the semidefinite matrix is of size $2m$ instead of $n + m$ we had before.

More generally, notice that VGFs are left unitarily invariant. Therefore, 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 (4.4) produces a feasible $Y$ with a smaller value for the objective function.

## 5 Structured optimization with VGF

In this section, we discuss optimization algorithms for solving convex minimization problems with VGF penalties in the form of (1.6). The proximal operators of VGFs we studied in the previous section are the key parts of proximal gradient methods (see, e.g., [5, 6, 35]). More specifically, when the loss function $\mathcal{L}(X)$ is smooth, we can iteratively update the variables $X^{(t)}$ as follows:

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

where $\gamma_t$ 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 conjugate structure that can be exploited together with the structure of the VGF penalty functions. We assume that the loss function $\mathcal{L}$ in (1.6) has the following representation:

$$\mathcal{L}(X) = \max_{\mathbf{g} \in \mathcal{G}} \ \langle X, \mathcal{D}(\mathbf{g}) \rangle - \hat{\mathcal{L}}(\mathbf{g}), \tag{5.1}$$

where $\hat{\mathcal{L}} : \mathbb{R}^p \to \mathbb{R}$ is a convex function, $\mathcal{G}$ is 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 (e.g. see [24]). Moreover, consider the infimal post-composition [4, Def. 12.33] of $\hat{\mathcal{L}} : \mathcal{G} \to \mathbb{R}$ by $\mathcal{D}(\cdot)$, defined as

$$(\mathcal{D} \rhd \hat{\mathcal{L}})(Y) = \inf \{ \hat{\mathcal{L}}(G) : \ \mathcal{D}(G) = Y \ , \ G \in \mathcal{G} \} . \tag{5.2}$$

Then, the conjugate to this function is equal to $\mathcal{L}$. 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.

With the variational representation of $\mathcal{L}$ in (5.1), we can write the VGF-penalized loss minimization problem (1.6) as a convex-concave saddle-point optimization problem:

$$J_{\mathrm{opt}} = \min_{X} \max_{M \in \mathcal{M}, \ \mathbf{g} \in \mathcal{G}} \ \left\{ \langle X, \mathcal{D}(\mathbf{g}) \rangle - \hat{\mathcal{L}}(\mathbf{g}) + \lambda \, \mathrm{tr}(XMX^T) \right\} . \tag{5.3}$$

If $\hat{\mathcal{L}}$ is smooth (while $\mathcal{L}$ may still be nonsmooth) and the sets $\mathcal{G}$ and $\mathcal{M}$ are simple (e.g., admitting simple projections), we can solve problem (5.3) using the *mirror-prox* algorithm [33, 24]. In section 5.1, we present a variant of the miror-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 (5.3) 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 (5.1) for them. This discussion will provide intuition about the linear operator $\mathcal{D}$ and set $\mathcal{G}$ in relation with data and prediction.

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

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

In terms of the representation in (5.1), here we have $\mathcal{D}(\mathbf{g}) = A^T\mathbf{g}$ and $\hat{\mathcal{L}}(\mathbf{g}) = \frac{1}{2}(\|\mathbf{g}\|^*)^2 + \mathbf{b}^T\mathbf{g}$. Similarly, a norm loss can be represented as

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

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

**$\varepsilon$-insensitive loss.** The Huber loss, defined as

$$H(u) = \begin{cases} \frac{u^2}{2\delta} + \frac{\delta}{2} & \text{if } |u| \leqslant \delta \\ |u| & \text{otherwise} \end{cases} \tag{5.4}$$

can be represented in a variational form as $H(u) = \frac{1}{2}\min_{\theta \geqslant \delta} \ \frac{u^2}{\theta} + \theta$. A variant of Huber loss function (e.g., see [32, Section 14.5.1] for more details and applications) is called the $\varepsilon$-insensititive loss and is given as

$$\mathcal{L}_\varepsilon(x) = (|x| - \varepsilon)_+ = \max_{\alpha,\beta}\{\alpha(x - \varepsilon) + \beta(-x - \varepsilon) : \ \alpha, \beta \geqslant 0, \ \alpha + \beta \leqslant 1\}\,. \tag{5.5}$$

**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 $\mathbf{a}_s \in \mathbb{R}^p$ is a feature vector and $b_s \in \{+1, -1\}$ is a binary label. We would like to find $x \in \mathbb{R}^p$ such that the linear function ${\mathbf{a_s}}^T x$ 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 $b_s(\mathbf{a}_s^T\mathbf{x}) \geqslant 1$ 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

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

where $\mathbf{D} = [b_1\mathbf{a}_1, \ldots, b_N\mathbf{a}_N]^T$. Here, in terms of (5.1), we have, $\mathcal{G} = \{\mathbf{g} \in \mathbb{R}^N : \ 0 \leqslant g_s \leqslant 1/N\}$, $\mathcal{D}(\mathbf{g}) = -\mathbf{D}^T\mathbf{g}$, and $\hat{\mathcal{L}}(\mathbf{g}) = -\mathbf{1}^T\mathbf{g}$.

**Multi-class 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 $\mathbf{x}_i, \ldots, \mathbf{x}_m$, 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

$$\mathbf{x}_i^T\mathbf{a}_s \geqslant \mathbf{x}_j^T\mathbf{a}_s \quad \text{for all } (i, j) \in \mathcal{I}(b_s), \tag{5.6}$$

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.

- *Flat multiclass classification:* $\mathcal{I}(k) = \{(k, j) : j \neq k\}$. In this case, the constraints in (5.6) are equivalent to the label $b_s = \arg\max_{i\in\{1,\ldots,m\}} \mathbf{x}_i^T\mathbf{a}_s$; see [41].
- *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 in the tree depending on the class label $k$; see Section 6 and [14, 42] for further details.

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

$$\mathcal{L}(X) = \frac{1}{N} \sum_{s=1}^{N} \max \left\{ 0, 1 - \max_{(i,j) \in \mathcal{I}(b_s)} \{ \mathbf{x}_i^T \mathbf{a}_s - \mathbf{x}_j^T \mathbf{a}_s \} \right\}, \tag{5.7}$$

which penalizes the amount of violation for the inequality constraints in (5.6).

In order to represent the loss function in (5.7) in the form of (5.1), 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 $i$th entry and $+1$ at the $j$th entry. Then the $p_k$ constraints in (5.6) can be summarized as $E_k^T X^T \mathbf{a}_s \leqslant \mathbf{0}$. It can be shown that the multi-class hinge loss $\mathcal{L}(X)$ in (5.7) can be represented in the form (5.1) via

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

where $A = [\mathbf{a}_1 \ \cdots \ \mathbf{a}_N]$ and $\mathcal{E}(\mathbf{g}) = [E_{b_1} \mathbf{g}_1 \ \cdots \ E_{b_N} \mathbf{g}_N] \in \mathbb{R}^{m \times N}$. Moreover, the domain of maximization in (5.1) is defined as

$$\mathcal{G} = \mathcal{G}_{b_1} \times \ldots \times \mathcal{G}_{b_N} \quad \text{where} \quad \mathcal{G}_k = \{ \mathbf{g} \in \mathbb{R}^{p_k} \ : \ \mathbf{g} \geqslant 0 \ , \ \mathbf{1}^T \mathbf{g} \leqslant 1/N \}. \tag{5.8}$$

Combining the above variational form for multi-class hinge loss and a VGF as penality 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

$$\min_{X} \max_{M \in \mathcal{M}, \mathbf{g} \in \mathcal{G}} \left\{ \mathbf{1}^T \mathbf{g} - \langle X, A\, \mathcal{E}(\mathbf{g})^T \rangle + \lambda \operatorname{tr}(X M X^T) \right\}. \tag{5.9}$$

### 5.1 Mirror-prox algorithm with adaptive line search

The mirror-prox (MP) algorithm was proposed by Nemirovski [34] for approximating the saddle points of smooth convex-concave functions and solutions of variational inequalities with Lipschitz continuous monotone operators. It is extension of the extra-gradient method [25], and more variants are studied in [23]. In the sequel, 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 (5.3).

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

$$\forall\, z, z' \in Z : \qquad \|F(z) - F(z')\|_* \leqslant L \|z - z'\|, \tag{5.10}$$

$$\forall\, z, z' \in Z : \qquad \langle F(z) - F(z'), z - z' \rangle \geqslant 0. \tag{5.11}$$

The goal of the MP algorithm is to approximate a (strong) solution to the variational inequality associated with $(Z, F)$:

$$\langle F(z^\star), z - z^\star \rangle \geqslant 0, \quad \forall\, z \in Z.$$

Let $\phi(x, y)$ be a smooth function that is convex in $x$ and concave in $y$, and $X$ and $Y$ are closed convex sets. Then the convex-concave saddle point problem

$$\min_{x \in X} \max_{y \in Y} \ \phi(x, y), \tag{5.12}$$

**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** $\delta_t \leqslant 0$
  $\gamma_{t+1} := c_{\text{inc}} \gamma_t$
**until** $V_{z_t}(z_{t+1}) \leqslant \varepsilon$
**return** $\bar{z}_t := (\sum_{\tau=1}^{t} \gamma_\tau)^{-1} \sum_{\tau=1}^{t} \gamma_\tau w_\tau$

Figure 1: Mirror-Prox algorithm with adaptive line search. Here $c_{\text{dec}} > 1$ and $c_{\text{inc}} > 1$ are two fixed parameters controlling how much to decrease and increase the step size $\gamma_t$ during the line search trials (the inner **repeat** loop). The stopping criterion for the line search is $\delta_t \leqslant 0$ where $\delta_\tau = \gamma_\tau \langle F(w_\tau), w_\tau - z_{\tau+1} \rangle - V_{z_\tau}(z_{\tau+1})$.

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

$$F(z) = \begin{bmatrix} \nabla_x \phi(x, y) \\ -\nabla_y \phi(x, y) \end{bmatrix}. \tag{5.13}$$

The setup of the Mirror-Prox 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 $Z$, denoted $Z^o$, and is strongly convex with modulus 1 with respect to $\|\cdot\|$, i.e.,

$$\forall\, z, z' \in Z: \qquad \langle \nabla h(z) - \nabla h(z'), z - z' \rangle \;\geqslant\; \|z - z'\|^2. \tag{5.14}$$

For any $z \in Z^o$ and $z' \in Z$, 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

$$P_z(\xi) \;=\; \arg\min_{z' \in Z} \left\{ \langle \xi, z' \rangle + V_z(z') \right\} \;=\; \arg\min_{z' \in Z} \left\{ \langle \xi - \nabla h(z), z' \rangle + h(z') \right\}.$$

With the above definitions, we are now ready to present the MP algorithm in Figure 1. Compared with the original MP algorithm [33, 23], our variant in Figure 1 employs an adaptive line search procedure to determine the step sizes $\gamma_t$, for $t = 1, 2, \ldots$. We can use exit the algorithm whenever $V_{z_t}(z_{t+1}) \leqslant \epsilon$ for some $\epsilon > 0$. Under the assumptions (5.10) and (5.11), the MP algorithm in Figure 1 enjoys the same $O(1/t)$ convergence rate as the one proposed in [33], but performs much faster in practice. The proof requires only simple modifications of the proof in [33, 23], and we leave it as an exercise for the reader.

In order to solve the saddle-point problem we consider in (5.3), assuming that $\hat{\mathcal{L}}$ is smooth, we can apply the MP algorithm directly. In particular, the gradient mapping in (5.13) becomes

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

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

### 5.2 Reduced Formulation

As we discussed early, when the loss function has the structure (5.1), we can formulate the VGF-penalized minimization problem as a convex-concave saddle point problem

$$J_{\text{opt}} = \min_{X} \max_{\mathbf{g}\in\mathcal{G}} \left\{ \langle X, \mathcal{D}(\mathbf{g})\rangle - \hat{\mathcal{L}}(\mathbf{g}) + \lambda\, \Omega(X) \right\}. \tag{5.15}$$

Since $\mathcal{G}$ is compact, $\Omega$ is convex in $X$, and $\hat{\mathcal{L}}$ is convex in $\mathbf{g}$, we can use the minimax theorem to interchange the max and min, and then use the definition of the conjugate function $\Omega^*$ to obtain

$$J_{\text{opt}} = -\min_{\mathbf{g}\in\mathcal{G}} \left\{ \hat{\mathcal{L}}(\mathbf{g}) + \lambda\Omega^*(-\tfrac{1}{\lambda}\mathcal{D}(\mathbf{g})) \right\}. \tag{5.16}$$

While the dimensions of $\mathcal{D}(\mathbf{g})$ are the same as those of $X$, the structure of a VGFs allows for making the optimization program in (5.16) independent of $n$ (the ambient dimension of columns of $X$), and getting a *reduced formulation*.

First, observe that both the VGF and its conjugate depend only on the Gram matrix of their input matrix. Therefore, if possible, we can replace $\mathcal{D}(\mathbf{g}) \in \mathbb{R}^{n\times m}$ with a smaller matrix $\mathcal{D}'(\mathbf{g})$ as long as $\mathcal{D}(\mathbf{g})^T\mathcal{D}(\mathbf{g}) = \mathcal{D}'(\mathbf{g})^T\mathcal{D}'(\mathbf{g})$. Since $\mathcal{D}(\mathbf{g})$ is linear in $\mathbf{g}$, consider a representation as

$$\mathcal{D}(\mathbf{g}) = [D_1\mathbf{g} \ \cdots \ D_m\mathbf{g}] = [D_1 \ \cdots \ D_m](\mathbf{I}_m \otimes \mathbf{g}) = \mathbf{D}(\mathbf{I}_m \otimes \mathbf{g}),$$

where $D_i \in \mathbb{R}^{n\times p}$ and $\mathbf{D} \in \mathbb{R}^{n\times mp}$. Considering $\mathbf{B} \in \mathbb{R}^{q\times mp}$ as the reduced Cholesky factor of $\mathbf{D}^T\mathbf{D} \in \mathbb{R}^{mp\times mp}$, i.e. $\mathbf{B}^T\mathbf{B} = \mathbf{D}^T\mathbf{D}$ and letting $q = \operatorname{rank}(\mathbf{D})$, we can define

$$\mathcal{D}'(\mathbf{g}) := \mathbf{B}(\mathbf{I}_m \otimes \mathbf{g}) \in \mathbb{R}^{q\times m}. \tag{5.17}$$

Notice that $q = \operatorname{rank}(\mathbf{D}) \leqslant mp$ can potentially be much smaller than $n$ depending on the application. In such cases, we can state the *reduced reformulation* of (5.16) as

$$J_{\text{opt}} = -\min_{\mathbf{g}\in\mathcal{G}} \left\{ \hat{\mathcal{L}}(\mathbf{g}) + \tfrac{1}{\lambda}\Omega^*\left(\mathcal{D}'(\mathbf{g})\right) \right\}, \tag{5.18}$$

where $\hat{\mathcal{L}}(\cdot)$ is convex, $\mathcal{G} \subset \mathbb{R}^p$ is convex and compact, and $\mathcal{D}' : \mathbb{R}^p \to \mathbb{R}^{q\times m}$ is a linear operator with $q \leqslant mp$. Observe that when $\Omega$ is convex on $\mathbb{R}^{n\times m}$, it is also convex on $\mathbb{R}^{q\times m}$ for any $q \leqslant n$ (see discussions after Theorem 3.1). Now, we can either directly solve (5.18) or a smaller version of (5.15) as

$$J_{\text{opt}} = \min_{X'\in\mathbb{R}^{q\times m}} \max_{\mathbf{g}\in\mathcal{G}} \left\{ \langle X', \mathcal{D}'(\mathbf{g})\rangle - \hat{\mathcal{L}}(\mathbf{g}) + \lambda\, \Omega(X') \right\}. \tag{5.19}$$

Both have reduced dimensions.

**Recovering $X$.** The linear operator $\mathcal{D}'$ encodes the information that is essential for the loss minimization problem with a VGF penalty. This reduction allows us to carry out the computations in possibly much smaller dimensions than the original formulation. However, we still need to recover $X$ will be still recoverable after solving (5.18). To this end, recall that the reduction step used definition of the conjugate function:

$$\Omega^*(-\tfrac{1}{\lambda}\mathcal{D}(\mathbf{g})) = \sup_X \left\{\langle X, \tfrac{-1}{\lambda}\mathcal{D}(\mathbf{g})\rangle - \Omega(X)\right\}.$$

This implies the conjugacy relations (e.g. see [38, Proposition 11.3])

$$X_{\text{opt}} \in \partial\Omega^*(-\tfrac{1}{\lambda}\mathcal{D}(\mathbf{g}_{\text{opt}}))\,, \qquad -\tfrac{1}{\lambda}\mathcal{D}(\mathbf{g}_{\text{opt}}) \in \partial\Omega(X_{\text{opt}}),$$

and

$$\Omega(X_{\text{opt}}) + \Omega^*(-\tfrac{1}{\lambda}\mathcal{D}(\mathbf{g}_{\text{opt}})) = \langle X, -\tfrac{1}{\lambda}\mathcal{D}(\mathbf{g}_{\text{opt}})\rangle\,.$$

The subdifferential of $\Omega^*$ has been characterized in Section 3.5. According to Remark 3.9, we can recover an optimal solution as

$$X_{\text{opt}} = \arg\min_Z \min_Y \left\{\Omega(Z) :\ Z = \tfrac{1}{2}(-\tfrac{1}{\lambda}\mathcal{D}(\mathbf{g}_{\text{opt}})M_{\text{opt}}^\dagger + Y)\,,\ Y M_{\text{opt}} M_{\text{opt}}^\dagger = \mathbf{0}\right\}. \tag{5.20}$$

Solving this problem might in general require the same effort as directly solving (5.15). However, in some cases $Y = 0$ gives a valid subgradient in (3.11) and we do not need to solve the above optimization problem. Moreover, if $M_{\text{opt}}$ is positive definite, we can output $X_{\text{opt}} = -\frac{1}{2\lambda}\mathcal{D}(\mathbf{g}_{\text{opt}})M_{\text{opt}}^{-1}$. Notice that a VGF with $\mathcal{M} \subset \mathbb{S}_{++}$ always lie in this category.

**Summary.** If $n \gg m$, we can use the following steps to solved a problem with reduced size and then recover a solution to the original problem.

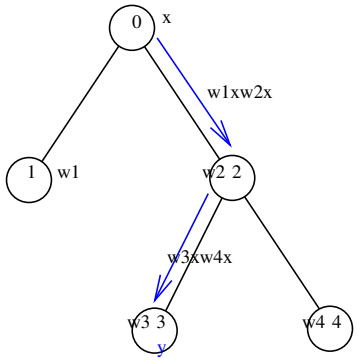
1. Compute $\mathcal{D}'$ from (5.17) by performing a Cholesky factorization once.
2. Solve (5.19) using the MP algorithm in Section 5.1, and keep the optimal $M_{\text{opt}}$ and $\mathbf{g}_{\text{opt}}$.
3. Compute a subgradient of $\Omega^*(-\frac{1}{\lambda}\mathcal{D}(\mathbf{g}_{\text{opt}}))$ as in (5.20), and output it as an optimal solution for (5.15). If $M_{\text{opt}}$ is positive definite, we can output $X_{\text{opt}} = -\frac{1}{2\lambda}\mathcal{D}(\mathbf{g}_{\text{opt}})M_{\text{opt}}^{-1}$.

# 6 Numerical Examples

In this section, we discuss the application of VGFs in hierarchical classification to demonstrate the effectiveness of the presented algorithms in a real data experiment.

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

In hierarchical classification, the class labels $\{1, \ldots, m\}$ are organized in a category tree, where the root of the tree is given the fictious label 0 (see Figure 2a). For each node $i \in \{0, 1, \ldots, m\}$, 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




(a)

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

(b)

Figure 2: (2a): 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. (2b): Definition of the hierarchical classification function.

of $i$ excluding 0 but including itself. A hierarchical linear classifier $f(\mathbf{a})$ is defined in Figure 2b, which is parameterized by the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$ through a recursive procedure. In other words, an instance is labeled sequentially by choosing the category of 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 2a.

For the hierarchical classifier defined above, given any example $\mathbf{a}_s$ with label $b_s$, a correct prediction made by $f(\mathbf{a})$ implies that (5.6) holds with

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

Given a set of examples $(\mathbf{a}_1, b_1), \ldots, (\mathbf{a}_N, b_N)$, we can train a hierarchical classifier parametrized by $X = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$ by solving the problem $\min_X \big\{\mathcal{L}(X) + \lambda\Omega(X)\big\}$, with the loss function $\mathcal{L}(X)$ defined in (5.7) 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 (5.3) and solved by the mirror-prox algorithm described in Section 5.1. In addition, we can use the reduction procedure descussed in Section 5.2 to reduce computational cost.

As a real-world example, we consider the classification dataset Reuters Corpus Volume I, RCV1-v2 [27], 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 $m = 14$ labels, is called CCAT and is used in our experiments. The samples and the classifiers are of dimension $n = 47236$ .

| | Nodes | Leaves | Total | Train | Test |
|---|---|---|---|---|---|
| CCAT | 14 | 10 | 59835 | 1797 | 58038 |

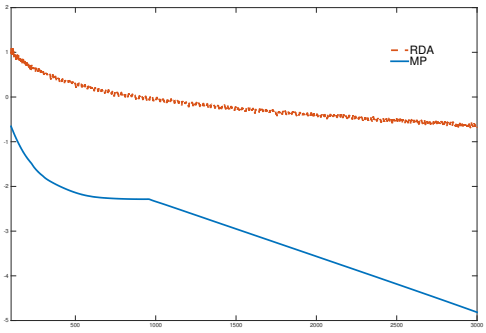
Table 1: Some statistics from RCV1-v2 [27].

We solve a VGF regularized hinge loss minimization problem as in (5.9), with $\lambda = 1$ , using the mirror-prox method. We use the $\ell_2$ norm as the mirror map to require the least knowledge about the optimization problem (see [22] for the requirements when combining different mirror maps).

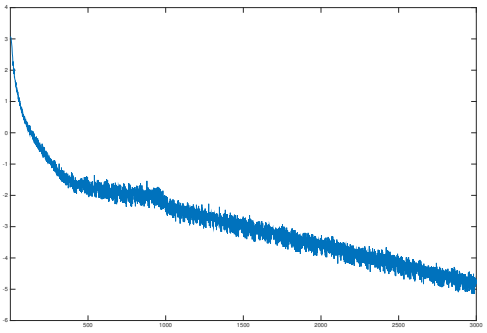
With such choice of a mirror map, the projection onto $\mathcal{G}$ in (5.8) boils down to separate projections onto $N$ full-dimensional simplexes. Each projection can be done by zeroing out the negative entries followed by a simple projection onto $\ell_1$ unit norm ball which can be implemented using the simple process described in [15].

To compare the prediction error of estimated classifiers from regularization with this VGF on test data, see [42]).

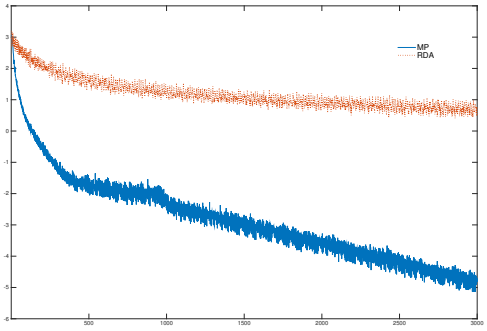
Although there is a clear advantage to the MP method compared to RDA in terms of the theoretical guarantee, one should be aware of the difference between the notion of gaps in the two methods. Figure 3a compares $\|X_t - X_{\text{final}}\|_F$ for MP and RDA using each one's own final estimate $X_{\text{final}}$. In terms of the runtime, we have observed that each iteration of MP takes 3 times more time compared to RDA. However, as evident from Figure 3a, MP is still much faster in generating a fixed-accuracy solution. Figure 3b illustrates the decay in the value of the gap for mirror-prox method, $V_{z_t}(z_{t+1})$, which confirms the theoretical findings.




(a) Average $\ell_2$ error of classifiers, between each iteration and the final estimate, $\|X_t - X_{\text{final}}\|_F$, for the MP and RDA algorithms; on a logarithmic scale.



(b) $V_{z_t}(z_{t+1})$ on a logarithmic scale.




(c) The value of loss function, relative to the final value, on a logarithmic scale for MP.

Figure 3: Convergence behavior for mirror-prox and RDA in our numerical experiment.

# A Proof of Lemma 3.5

First, assume that $\Omega$ is convex. By plugging in $X$ and $-X$ in the definition of convexity for $\Omega$ we get $\Omega(X) \geqslant 0$. Therefore, it is well-defined to consider the square root of $\Omega$. Now, for any given $X$ and $Y$, we are going to establish the triangle inequality $\sqrt{\Omega(X+Y)} \leqslant \sqrt{\Omega(X)} + \sqrt{\Omega(Y)}$. If $\Omega(X+Y)$ be zero, then the inequality is trivially satisfied; hence, assume it is not equal to zero. Given $X$ and $Y$, for any value $\theta \in (0,1)$ define $A = \frac{1}{\theta}X$, and $B = \frac{1}{1-\theta}Y$ and use the definition of convexity for $\Omega$ to get

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

where we used the second order homogeneity.

- if $\Omega(X) \geqslant \Omega(Y) = 0$, set $\theta = (\Omega(X) + \Omega(X+Y))/(2\Omega(X+Y)) > 0$. Assuming that $\theta < 1$, from (A.1) we get

  $$\Omega(X+Y) \leqslant \tfrac{1}{\theta}\,\Omega(X) + \tfrac{1}{1-\theta}\,\Omega(Y) = \frac{2\Omega(X+Y)\Omega(X)}{\Omega(X+Y)+\Omega(X)}$$

  which is equivalent to $\theta \geqslant 1$ and is a contradiction. Therefore, $\theta \geqslant 1$ which establishes the desired inequality.

- if $\Omega(X), \Omega(Y) \neq 0$, plug in $\theta = \sqrt{\Omega(X)}/(\sqrt{\Omega(X)} + \sqrt{\Omega(Y)}) \in (0,1)$ in (A.1) to get

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

Since $\sqrt{\Omega}$ satisfies the triangle inequality, as well as absolute homogeneity, it is a semi-norm. Notice that $\Omega(X) = 0$ does not necessarily imply $X = 0$. However, for a strictly convex $\Omega$, plugging in $X \neq 0$ and $-X$ in the definition of convexity gives $\Omega(X) > 0$. Hence, $\sqrt{\Omega}$ will be a norm.

Now, suppose that $\sqrt{\Omega}$ is a semi-norm; hence convex and non-negative. Moreover, $f$ defined by $f(x) = x^2$ for $x \geqslant 0$ and $f(x) = 0$ for $x \leqslant 0$ is a non-decreasing function. Composition of these two functions will be convex and is equal to $\Omega$.

# B Calculus of VGF Representations

Considering a VGF as a function of the associated set, one can derive a calculus over this argument as given in Table 2.

| $\mathcal{M}$ | $\Omega_{\mathcal{M}}$ |
|---|---|
| $\alpha\mathcal{A}\,,\ \alpha > 0$ | $\alpha\Omega_{\mathcal{A}}$ |
| $\mathcal{A} \cup \mathcal{B}$ | $\Omega_{\mathcal{A}} \vee \Omega_{\mathcal{B}}$ |
| $\bigcup_{\alpha\in\Delta}(\alpha_1\mathcal{A} : \alpha_2\mathcal{B})$ | $\Omega_{\mathcal{A}} \triangledown \Omega_{\mathcal{B}}$ |
| $\mathcal{A} + \mathcal{B}$ | $\Omega_{\mathcal{A}} + \Omega_{\mathcal{B}}$ |
| $\mathcal{A} : \mathcal{B} = \{A : B\}$ | $\Omega_{\mathcal{A}} \square \Omega_{\mathcal{B}}$ |

Table 2: Calculus of Representations for VGFs.

In the above table, we have used the following notations:

- The *parallel sum* [1] for two Hermitian positive semidefinite matrices is defined as $A : B = B : A = A(A + B)^{\dagger}B$. If both $A$ and $B$ are non-singular, $A : B = (A^{-1} + B^{-1})^{-1}$.
- $\Delta$ is the flat simplex.
- The *infimal convolution* is defined as $(\Omega_{\mathcal{A}} \square \Omega_{\mathcal{B}})(X) = \inf\{\Omega_{\mathcal{A}}(X_1) + \Omega_{\mathcal{B}}(X_2) : \ X = X_1 + X_2\}$.
- The *sublevel convolution* is defined as $(\Omega_{\mathcal{A}} \triangledown \Omega_{\mathcal{B}})(X) = \inf\{\Omega_{\mathcal{A}}(X_1) \vee \Omega_{\mathcal{B}}(X_2) : \ X = X_1 + X_2\}$.

# References


[1] W. N. Anderson and R. J. Duffin. Series and parallel addition of matrices. *Journal of Mathematical Analysis and Applications*, 26(3):576–594, 1969.

[2] A. Argyriou, R. Foygel, and N. Srebro. Sparse prediction with the $k$-support norm. In *Advances in Neural Information Processing Systems*, pages 1457–1465, 2012.

[3] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

[4] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer Science & Business Media, 2011.

[5] A. Beck and M. Teboulle. A fast iterative shrinkage-threshold algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[6] A. Beck and M. Teboulle. Gradient-based algorithms with applications to signal recovery. In D. P. Palomar and Y. C. Eldar, editors, *Convex Optimization in Signal Processing and Communications*, chapter 2, pages 42–88. Cambridge University Press, 2010.

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

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

[9] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[10] J. V. Burke and T. Hoheisel. Matrix support functionals for inverse problems, regularization, and learning. *SIAM Journal on Optimization*, 2015.

[11] F. Burns, D. Carlson, E. Haynsworth, and T. Markham. Generalized inverse formulas using the schur complement. *SIAM Journal on Applied Mathematics*, 26(2):254–259, 1974.

[12] X. Cia and X. Wang. A note on the positive semidefinite minimum rank of a sign pattern matrix. *Electronic Journal of Linear Algebra*, 26:345–356, 2013.

[13] C.-S. Cheng. $E(s^2)$-optimal supersaturated designs. *Statistica Sinica*, 7(4):929–939, 1997.

[14] O. Dekel, J. Keshet, and Y. Singer. Large margin hierarchical classification. In *Proceedings of the 21st International Conference on Machine Learning*, pages 27–34, 2004.

[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*, pages 272–279. ACM, 2008.

[16] T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. In *Journal of Machine Learning Research*, pages 615–637, 2005.

[17] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition, 2009.

[18] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms I: fundamentals*, volume 305. Springer Science & Business Media, 2013.

[19] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 2012.

[20] L. Jacob, F. Bach, and J.-P. Vert. Clustered multi-task learning: A convex formulation. In *NIPS*, volume 21, pages 745–752, 2008.

[21] D. Jayaraman, F. Sha, and K. Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1629–1636. IEEE, 2014.

[22] A. Juditsky and A. Nemirovski. First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure. *Optimization for Machine Learning*, pages 149–183, 2011.

[23] A. Juditsky and A. Nemirovski. First-order methods for nonsmooth convex large-scale optimization, II: Utilizing problems's structure. In S. Sra, S. Nowozin, and S. J. Wright, editors, *Optimization for Machine Learning*, chapter 6, pages 149–184. The MIT Press, Cambridge, MA., 2011.

[24] A. Juditsky and A. Nemirovski. Solving variational inequalities with monotone operators on domains given by linear minimization oracles. *arXiv preprint arXiv:1312.1073*, 2013.

[25] G. M. Korpelevič. An extragradient method for finding saddle points and for other problems. *Èkonom. i Mat. Metody*, 12(4):747–756, 1976.

[26] A. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1/2):173–183, 1995.

[27] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004.

[28] B. Martinet. Regularisation, d'inéquations variationelles par approximations succesives. *Revue Francaise d'Informatique et de Recherche Operationelle*, 4:154–159, 1970.

[29] A. M. McDonald, M. Pontil, and D. Stamos. New perspectives on k-support and cluster norms. *arXiv preprint arXiv:1403.1481*, 2014.

[30] C. A. Micchelli, J. M. Morales, and M. Pontil. Regularizers for structured sparsity. *Advances in Computational Mathematics*, 38(3):455–489, 2013.

[31] L. Mirsky. A trace inequality of john von neumann. *Monatshefte für Mathematik*, 79(4):303–306, 1975.

[32] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2012.

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

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

[35] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming, Ser. B*, 140:125–161, 2013.

[36] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[37] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14:877–898, 1976.

[38] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[39] B. Romera-Paredes, A. Argyriou, N. Berthouze, and M. Pontil. Exploiting unrelated tasks in multi-task learning. In *International Conference on Artificial Intelligence and Statistics*, pages 951–959, 2012.

[40] 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*, pages 274–289. Springer, 2014.

[41] J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. In *Proceedings of the 6th European Symposium on Artificial Neural Networks (ESANN)*, pages 219–224, 1999.

[42] D. Zhou, L. Xiao, and M. Wu. Hierarchical classification via orthogonal transfer. *Proceedings of the 28th International Conference on Machine Learning (ICML)*, June 2011.