The Generalized Trace Norm

Jason D. M. Rennie
jrennie@gmail.com

April 19, 2006

1 Introduction

Let $X \in \mathbb{R}^{m \times n}$. Define the orthogonal group, denoted $O(n)$, as the set of orthogonal $n \times n$ matrices. Then, we define the generalized trace norm with respect to $V \in O(n)$ as the sum of lengths of the columns of XV,

$$\mathrm{GTN}_V(X) = \sum_j \sqrt{\sum_i \left(\sum_k X_{ik} V_{kj} \right)^2}. \tag{1}$$

Note that rows of XV are the rows of X represented in the V basis. The sum of lengths of columns of XV depend on the choice of basis, V. The trace norm is the minimum over bases of column vector lengths,

$$\|X\|_\Sigma = \min_{V \in V_{n,p}} \mathrm{GTN}_V(X). \tag{2}$$

The GTN can be viewed as a sum of projections. XV is the projection of rows of X onto the basis vectors (columns of V). A column of XV is the vector of projection lengths for a basis vector. Its length is the extent to which rows of X point in the direction of the corresponding basis vector.

Note that GTN_V also corresponds to the sum of diagonal entries of Σ in the decomposition $X = U\Sigma V^T$, where $U^T \in O(m)$ and Σ is an $m \times n$ matrix with zeros everywhere but the diagonal. The Singular Value Decomposition (SVD) corresponds to the choice of V used for calculation of the trace norm, (2).

2 Nature of the (Generalized) Trace Norm

Consider a pair of unit-length vectors in $\mathbb{R}^2$. Up to a rotation, we can define these with a single parameter, x. Stacking the two vectors as rows of a matrix, we get

$$X(x) = \begin{bmatrix} 1 & 0 \\ x & \sqrt{1-x^2} \end{bmatrix}. \tag{3}$$

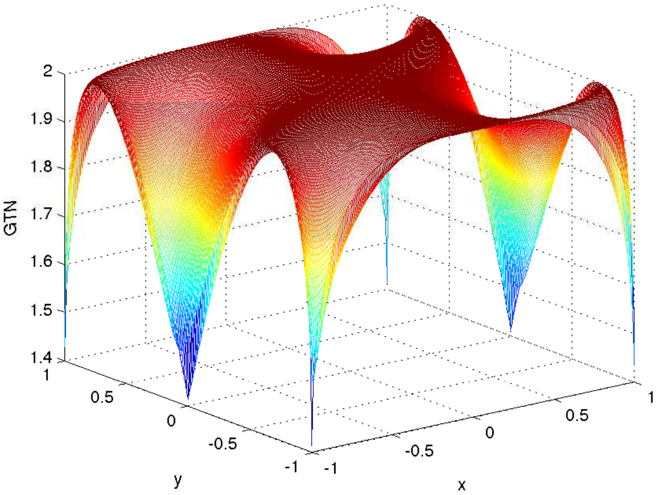

Figure 1: Shown is the generalized trace norm (GTN) of $X(x)$ with respect to $V(y)$. When the vectors are nearly orthogonal ($x \approx 0$), the GTN is near its maximum value of 2. When the vectors are nearly co-linear ($x \approx \pm 1$), the GTN value depends heavily on the choice of basis. When the basis is aligned with the vectors ($y \in \{-1, 0, 1\}$), the GTN is near its minimum value of $\sqrt{2}$; when the alignment is poor, the GTN nears its maximum value of 2.

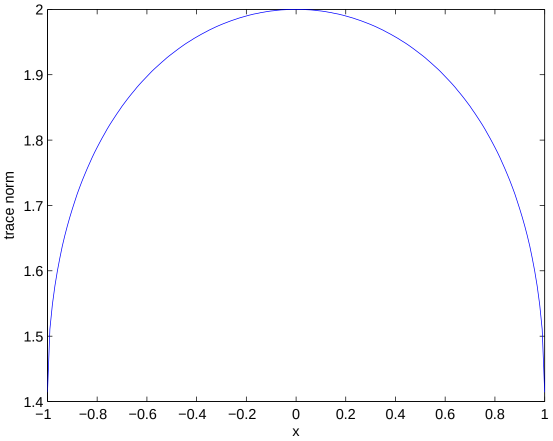

Figure 2: Shown is the trace norm of the matrix X (where x is the only parameter).

Similarly, we can define a basis in $\mathbb{R}^2$ with a single parameter, y,

$$V(y) = \begin{bmatrix} y & \sqrt{1-y^2} \\ \sqrt{1-y^2} & -y \end{bmatrix}. \tag{4}$$

The generalized trace norm with respect to V is the sum of the column lengths of XV,

$$\begin{aligned}\mathrm{GTN}_{V(y)}(X(x)) = &\sqrt{1-x^2+2x^2y^2+2xy\sqrt{1-x^2}\sqrt{1-y^2}} \\ &+\sqrt{1+x^2-2x^2y^2-2xy\sqrt{1-x^2}\sqrt{1-y^2}}\end{aligned} \tag{5}$$

Figure 1 plots $\mathrm{GTN}_V(X)$ as a function of x and y.

The trace norm is the minimum GTN over bases V. In our 2×2 example, this corresponds to minimization over the parameter y; one can visualize a plot of the trace norm as the lower boundary of the plot in Figure 1 rotated so that the line of sight is parallel to the y-axis. Figure 2 provides this visualization—the plot of trace norm as a function of x. Note the similarity to the shape of the binary entropy (as a function of $p \in [0, 1]$), and the sine function (on $\theta \in [0, \pi]$). The trace norm, in fact, is more rounded, and rises more steeply at the edges than either entropy or sine. All three functions can be used to judge the degree to which multiple factors exist in the data.

For further understanding of the trace norm, we consider the general 2×2 case. Note that the trace norm (1) is rotation invariant (i.e. $\|X\|_\Sigma = \|XR\|_\Sigma$ for $R \in O(n)$), and (2) and the trace norm is associative with scalar multiplication (i.e. $\|cX\|_\Sigma = c\|X\|_\Sigma$). Thus, orientation and scale of one row of X is arbitrary.

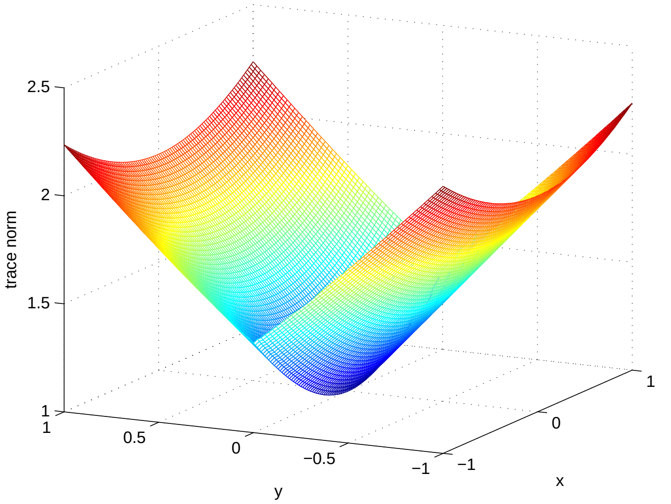

Figure 3: Shown is the trace norm of the matrix $X(x, y)$.

So, we only parameterize the second row:

$$X(x,y) = \begin{bmatrix} 1 & 0 \\ x & y \end{bmatrix}. \tag{6}$$

The parameterization of the basis is unchanged. The generalized trace norm is

$$\begin{aligned}\mathrm{GTN}_{V(w)}(X(x,y)) = {}& \sqrt{w^2 + y^2 - y^2w^2 + x^2w^2 + 2xyw\sqrt{1-w^2}} \\ & + \sqrt{1 - w^2 + x^2 - x^2w^2 + y^2w^2 - 2xyw\sqrt{1-w^2}}.\end{aligned} \tag{7}$$

As noted earlier, the trace norm is the minimum generalized trace norm over bases. The trace norm is plotted in Figure 3 as a function of x and y. For any fixed $x = c$, the trace norm is approximately $\|X(c,y)\|_\Sigma \approx |y| + \sqrt{1+c^2}$. For fixed $y = c$, the trace norm is approximately $\|X(x,c)\|_\Sigma \approx |c| + \sqrt{1+x^2}$. When $y = 0$, the matrix has rank 1. When $y \neq 0$, the matrix has rank 2. The fact that trace norm behaves like an absolute value function as y changes (but x is held fixed) means that when the trace norm is used in a combined objective, a low rank solution is encouraged. Thus the trace norm is an effective regularizer for learning; it provides a continuous, convex penalty which specifically discourages a high rank solution.