Matrix Support Functional and its Applications

James V Burke
Mathematics, University of Washington

Joint work with
Yuan Gao (UW) and Tim Hoheisel (McGill),

CORS, Banff 2016
June 1, 2016

Connections

What do the following topics have in common?

- Quadratic Optimization Problem with Equality Constraints
- The Matrix Fractional Function and its Generalization
- Ky Fan p-k Norms
- K-means Clustering
- Best Affine Unbiased Estimator
- Supervised Representation Learning
- Multi-task Learning
- Variational Gram Functions

Connections

What do the following topics have in common?

- Quadratic Optimization Problem with Equality Constraints
- The Matrix Fractional Function and its Generalization
- Ky Fan p-k Norms
- K-means Clustering
- Best Affine Unbiased Estimator
- Supervised Representation Learning
- Multi-task Learning
- Variational Gram Functions

Answer: They can all be represented using a matrix support function that is smooth on the interior of its domain.

A Matrix Support Functional (B-Hoheisel (2015))

Given $A \in \mathbb{R}^{p\times n}$ and $B \in \mathbb{R}^{p\times m}$ set

$$\mathcal{D}(A,B) := \left\{ \left(Y, -\frac{1}{2}YY^T \right) \in \mathbb{R}^{n\times m} \times \mathbb{S}^n \;\middle|\; Y \in \mathbb{R}^{n\times m} :\ AY = B \right\}$$

A Matrix Support Functional (B-Hoheisel (2015))

Given $A \in \mathbb{R}^{p \times n}$ and $B \in \mathbb{R}^{p \times m}$ set

$$\mathcal{D}(A,B) := \left\{ \left(Y, -\tfrac{1}{2}YY^T \right) \in \mathbb{R}^{n \times m} \times \mathbb{S}^n \;\middle|\; Y \in \mathbb{R}^{n \times m} :\; AY = B \right\}$$

We consider the support functional for $\mathcal{D}(A,B)$.

$$\sigma\left((X,V) \mid \mathcal{D}(A,B)\right) = \sup_{AY=B} \langle (X,V), (Y, -\tfrac{1}{2}YY^T) \rangle$$

A Matrix Support Functional (B-Hoheisel (2015))

Given $A \in \mathbb{R}^{p\times n}$ and $B \in \mathbb{R}^{p\times m}$ set

$$\mathcal{D}(A,B) := \left\{ \left(Y, -\tfrac{1}{2}YY^T \right) \in \mathbb{R}^{n\times m} \times \mathbb{S}^n \;\middle|\; Y \in \mathbb{R}^{n\times m} : \; AY = B \right\}$$

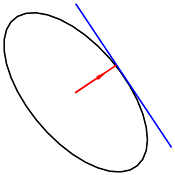
We consider the support functional for $\mathcal{D}(A,B)$.

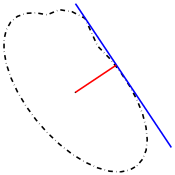
$$\sigma\left((X,V) \mid \mathcal{D}(A,B)\right) = \sup_{AY=B} \langle (X,V), (Y, -\tfrac{1}{2}YY^T) \rangle$$

$$= - \inf_{AY=B} \; \tfrac{1}{2}\mathrm{tr}\left(Y^T VY\right) - \langle X, \; Y \rangle$$

Support Functions

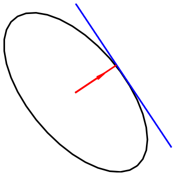
$$\sigma_S(x) := \sigma\left(x \mid S\right) := \sup_{y \in S} \langle x, y \rangle$$

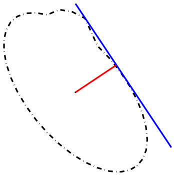




Support Functions

$$\sigma_S(x) := \sigma\left(x \mid S\right) := \sup_{y \in S} \langle x, y \rangle$$

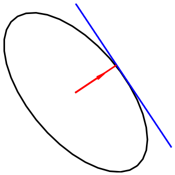


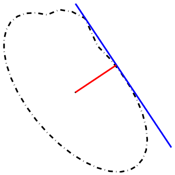


$S = \bigcap_x \{y \mid \langle x, y \rangle \leq \sigma\left(x \mid S\right)\}$

Support Functions

$$\sigma_S(x) := \sigma\left(x \mid S\right) := \sup_{y \in S} \langle x, y \rangle$$



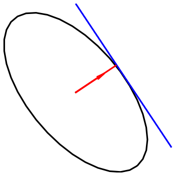


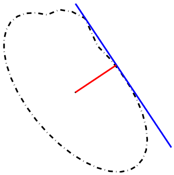
$S = \bigcap_x \left\{y \mid \langle x, y \rangle \leq \sigma\left(x \mid S\right)\right\}$

$\sigma_S = \sigma_{\mathrm{cl}\, S} = \sigma_{\mathrm{conv}\, S} = \sigma_{\overline{\mathrm{conv}}\, S}$

Support Functions

$$\sigma_S(x) := \sigma\left(x \mid S\right) := \sup_{y \in S} \langle x, y \rangle$$





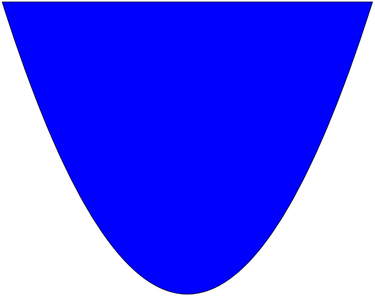
$S = \bigcap_x \left\{ y \mid \langle x, y \rangle \leq \sigma\left(x \mid S\right) \right\}$

$\sigma_S = \sigma_{\mathrm{cl}\, S} = \sigma_{\mathrm{conv}\, S} = \sigma_{\overline{\mathrm{conv}}\, S}$

When S is a closed convex set, then

$$\partial \sigma_S(x) = \arg\max_{y \in S} \langle x,\ y \rangle\ .$$

Epigraph



$$\operatorname{epi} f := \{(x, \mu) \mid f(x) \leq \mu\}$$

$$f^*(y) := \sigma\left((y, -1) \mid \operatorname{epi} f\right)$$

A Representation for $\sigma\left((X,V)\mid\mathcal{D}(A,B)\right)$

Let $A\in\mathbb{R}^{p\times n}$ and $B\in\mathbb{R}^{p\times m}$ such that $\operatorname{rge}B\subset\operatorname{rge}A$. Then

$$\sigma\left((X,V)\mid\mathcal{D}(A,B)\right)=\begin{cases}\frac{1}{2}\operatorname{tr}\left(\binom{X}{B}^T M(V)^\dagger\binom{X}{B}\right) & \text{if } \operatorname{rge}\binom{X}{B}\subset\operatorname{rge}M(V),\ V\succeq_{\ker A}0,\\ +\infty & \text{else.}\end{cases}$$

where

$$M(V):=\begin{pmatrix} V & A^T\\ A & 0\end{pmatrix}.$$

A Representation for $\sigma\left((X,V)\mid\mathcal{D}(A,B)\right)$

Let $A\in\mathbb{R}^{p\times n}$ and $B\in\mathbb{R}^{p\times m}$ such that $\operatorname{rge} B\subset\operatorname{rge} A$. Then

$$\sigma\left((X,V)\mid\mathcal{D}(A,B)\right)=\begin{cases}\frac{1}{2}\operatorname{tr}\left(\binom{X}{B}^T M(V)^\dagger\binom{X}{B}\right) & \text{if } \operatorname{rge}\binom{X}{B}\subset\operatorname{rge}M(V),\ V\succeq_{\ker A}0,\\ +\infty & \text{else.}\end{cases}$$

where

$$M(V):=\begin{pmatrix}V & A^T\\ A & 0\end{pmatrix}.$$

In particular,

$$\begin{aligned}\operatorname{dom}\sigma\left(\cdot\mid D(A,B)\right)&=\operatorname{dom}\partial\sigma\left(\cdot\mid D(A,B)\right)\\ &=\left\{(X,V)\in\mathbb{R}^{n\times m}\times\mathbb{S}^n\ \middle|\ \operatorname{rge}\begin{pmatrix}X\\ B\end{pmatrix}\subset\operatorname{rge}M(V),\ V\succeq_{\ker A}0\right\},\end{aligned}$$

with $\operatorname{int}\left(\operatorname{dom}\sigma\left(\cdot\mid D(A,B)\right)\right)=\left\{(X,V)\in\mathbb{R}^{n\times m}\times\mathbb{S}^n\mid V\succ_{\ker A}0\right\}.$

A Representation for $\sigma\left((X,V) \mid \mathcal{D}(A,B)\right)$

Let $A \in \mathbb{R}^{p\times n}$ and $B \in \mathbb{R}^{p\times m}$ such that $\mathrm{rge}\, B \subset \mathrm{rge}\, A$. Then

$$\sigma\left((X,V) \mid \mathcal{D}(A,B)\right) = \begin{cases} \frac{1}{2}\mathrm{tr}\left(\binom{X}{B}^T M(V)^\dagger \binom{X}{B}\right) & \text{if } \mathrm{rge}\binom{X}{B} \subset \mathrm{rge}\, M(V),\ V \succeq_{\ker A} 0, \\ +\infty & \text{else.} \end{cases}$$

where

$$M(V) := \begin{pmatrix} V & A^T \\ A & 0 \end{pmatrix}.$$

In particular,

$$\begin{aligned} \mathrm{dom}\,\sigma\left(\cdot \mid D(A,B)\right) &= \mathrm{dom}\,\partial\sigma\left(\cdot \mid D(A,B)\right) \\ &= \left\{ (X,V) \in \mathbb{R}^{n\times m} \times \mathbb{S}^n \;\middle|\; \mathrm{rge}\begin{pmatrix} X \\ B \end{pmatrix} \subset \mathrm{rge}\, M(V),\ V \succeq_{\ker A} 0 \right\}, \end{aligned}$$

with $\mathrm{int}\left(\mathrm{dom}\,\sigma\left(\cdot \mid D(A,B)\right)\right) = \left\{(X,V) \in \mathbb{R}^{n\times m} \times \mathbb{S}^n \mid V \succ_{\ker A} 0\right\}$.

The inverse $M(V)^{-1}$ exists when $V \succ_{\ker A} 0$ and A is surjective.

Relationship to Equality Constrained QP

Consider a equality constrained QP:

$$\nu(x, V) := \inf_{u \in \mathbb{R}^n} \left\{ \tfrac{1}{2} u^T V u - x^T u \mid Au = b \right\}.$$

Relationship to Equality Constrained QP

Consider a equality constrained QP:

$$\nu(x, V) := \inf_{u \in \mathbb{R}^n} \left\{ \frac{1}{2} u^T V u - x^T u \mid Au = b \right\}.$$

The Lagrangian is $L(u, \lambda) = \frac{1}{2} u^T V u - x^T u + \lambda^T (Au - b)$.

Optimality conditions are

$$\begin{aligned} Vu + A^T \lambda - x &= 0 \\ Au &= b \end{aligned}$$

Relationship to Equality Constrained QP

Consider a equality constrained QP:

$$\nu(x, V) := \inf_{u \in \mathbb{R}^n} \left\{ \tfrac{1}{2} u^T V u - x^T u \mid Au = b \right\}.$$

The Lagrangian is $L(u, \lambda) = \frac{1}{2} u^T V u - x^T u + \lambda^T (Au - b)$.

Optimality conditions are

$$\begin{aligned} Vu + A^T \lambda - x &= 0 \\ Au &= b \end{aligned}$$

This is equivalent to

$$\begin{pmatrix} V & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} u \\ \lambda \end{pmatrix} = \begin{pmatrix} x \\ b \end{pmatrix}.$$

Relationship to Equality Constrained QP

Consider a equality constrained QP:

$$\nu(x, V) := \inf_{u \in \mathbb{R}^n} \left\{ \tfrac{1}{2} u^T V u - x^T u \mid Au = b \right\}.$$

The Lagrangian is $L(u, \lambda) = \frac{1}{2} u^T V u - x^T u + \lambda^T (Au - b)$.

Optimality conditions are

$$\begin{aligned} Vu + A^T \lambda - x &= 0 \\ Au &= b \end{aligned}$$

This is equivalent to

$$M(V) = \begin{pmatrix} V & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} u \\ \lambda \end{pmatrix} = \begin{pmatrix} x \\ b \end{pmatrix}.$$

Relationship to Equality Constrained QP

Consider a equality constrained QP:

$$\nu(x, V) := \inf_{u \in \mathbb{R}^n} \left\{ \tfrac{1}{2} u^T V u - x^T u \mid Au = b \right\}.$$

The Lagrangian is $L(u, \lambda) = \frac{1}{2} u^T V u - x^T u + \lambda^T (Au - b)$.

Optimality conditions are

$$\begin{aligned} Vu + A^T \lambda - x &= 0 \\ Au &= b \end{aligned}$$

This is equivalent to

$$M(V) = \begin{pmatrix} V & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} u \\ \lambda \end{pmatrix} = \begin{pmatrix} x \\ b \end{pmatrix}.$$

Hence

$$\nu(x, V) = -\sigma\left((x, V) \mid \mathcal{D}(A, b)\right).$$

Maximum Likelihood Estimation

$$L(\mu, \Sigma; Y) := (2\pi)^{-mN/2} |\Sigma|^{-N/2} \prod_{i=1}^{N} \exp((y_i - \mu)^T \Sigma^{-1} (y_i - \mu))$$

Up to a constant, the negative log-likelihood is

$$\begin{aligned} -\ln L(\mu, \Sigma; Y) &= \tfrac{1}{2} \ln\det\Sigma \;+\; \tfrac{1}{2}\mathrm{tr}\left((Y - M)^T \Sigma^{-1} (Y - M)\right) \\ &= \sigma\left((Y - M), \Sigma) \mid \mathcal{D}(0, 0)\right) - \tfrac{1}{2}(-\ln\det\Sigma). \end{aligned}$$

Maximum Likelihood Estimation

$$L(\mu, \Sigma; Y) := (2\pi)^{-mN/2}|\Sigma|^{-N/2} \prod_{i=1}^{N} \exp((y_i - \mu)^T \Sigma^{-1}(y_i - \mu))$$

Up to a constant, the negative log-likelihood is

$$\begin{aligned} -\ln L(\mu, \Sigma; Y) &= \tfrac{1}{2}\ln\det\Sigma \;+\; \tfrac{1}{2}\mathrm{tr}\left((Y-M)^T\Sigma^{-1}(Y-M)\right) \\ &= \sigma\left((Y-M),\Sigma\right) \mid \mathcal{D}(0,0)\,) - \tfrac{1}{2}(-\ln\det\Sigma). \end{aligned}$$

The Matrix Fractional Function: Take $A = 0$ and $B = 0$, and set

$$\gamma(X, V) := \sigma\left((x, V) \mid \mathcal{D}(0,0)\right)$$

$$= \begin{cases} \frac{1}{2}X^T V^\dagger X & \text{if } \mathrm{rge}\, X \subset \mathrm{rge}\, V,\, V \succeq 0, \\ +\infty & \text{else.} \end{cases}$$

$\overline{\text{conv}}\, \mathcal{D}(A, B)$

Recall $$\partial \sigma_C(x) = \{y \in \overline{\text{conv}}\, C \mid \langle x, y \rangle = \sigma_C(x)\}\,.$$

$\overline{\operatorname{conv}}\,\mathcal{D}(A,B)$

Recall $\qquad \partial\sigma_C(x) = \{y \in \overline{\operatorname{conv}}\, C \mid \langle x, y\rangle = \sigma_C(x)\}\,.$

For $\partial\sigma\,((X,V)\ \mid \mathcal{D}(A,B))$ we need $\overline{\operatorname{conv}}\,(\mathcal{D}(A,B))$.

$\overline{\mathrm{conv}}\, \mathcal{D}(A, B)$

Recall $\quad \partial\sigma_C(x) = \{y \in \overline{\mathrm{conv}}\, C \mid \langle x, y\rangle = \sigma_C(x)\}$.

For $\partial\sigma\left((X, V) \mid \mathcal{D}(A, B)\right)$ we need $\overline{\mathrm{conv}}\,(\mathcal{D}(A, B))$.

Set

$$\mathbb{S}^n_+(\ker A) := \left\{ W \in \mathbb{S}^n \;\middle|\; u^T W u \geq 0 \;\forall\; u \in \ker A \right\} = \{W \succeq_{\ker A} 0\}.$$

Then $\mathbb{S}^n_+(\ker A)$ is a closed convex cone whose polar is given by

$$\mathbb{S}^n_+(\ker A)^\circ = \{W \in \mathbb{S}^n \mid W = PWP \preceq 0\},$$

where P is the orthogonal projection onto $\ker A$.

$\overline{\mathrm{conv}}\, \mathcal{D}(A,B)$

Recall $\qquad \partial\sigma_C(x) = \{ y \in \overline{\mathrm{conv}}\, C \mid \langle x, y\rangle = \sigma_C(x) \}\,.$

For $\partial\sigma\left((X,V) \mid \mathcal{D}(A,B)\right)$ we need $\overline{\mathrm{conv}}\,(\mathcal{D}(A,B))$.

Set

$$\mathbb{S}^n_+(\ker A) := \left\{ W \in \mathbb{S}^n \;\middle|\; u^T W u \geq 0 \;\forall\; u \in \ker A \right\} = \{ W \succeq_{\ker A} 0 \}.$$

Then $\mathbb{S}^n_+(\ker A)$ is a closed convex cone whose polar is given by

$$\mathbb{S}^n_+(\ker A)^\circ = \{ W \in \mathbb{S}^n \mid W = PWP \preceq 0 \}\,,$$

where P is the orthogonal projection onto $\ker A$.

For $\mathcal{D}(A,B) := \{ (Y, -\frac{1}{2}YY^T) \in \mathbb{R}^{n\times m} \times \mathbb{S}^n \mid Y \in \mathbb{R}^{n\times m} :\; AY = B \}$,

$$\overline{\mathrm{conv}}\, \mathcal{D}(A,B) = \Omega(A,B)$$

$$:= \left\{ (Y,W) \in \mathbb{R}^{n\times m} \times \mathbb{S}^n_- \;\middle|\; AY = B \text{ and } \tfrac{1}{2}YY^T + W \in \mathbb{S}^n_+(\ker A)^\circ \right\}.$$

Applications

- Quadratic Optimization Problem with Equality Constraints
- The Matrix Fractional Function and its Generalization
- Ky Fan p-k Norms
- K-means Clustering
- Best Affine Unbiased Estimator
- Supervised Representation Learning
- Multi-task Learning
- Variational Gram Functions
- ...

Motivating Examples

Recall the matrix fractional function

$$\gamma(X, V) := \sigma\left((X, V) \mid \mathcal{D}(0,0)\right)$$

$$= \begin{cases} \frac{1}{2}\mathrm{tr}\,(X^T V^\dagger X) & \text{if} \quad \mathrm{rge}\, X \in \mathrm{rge}\, V,\ V \succeq 0, \\ +\infty & \text{else.} \end{cases}$$

Motivating Examples

Recall the matrix fractional function

$$\gamma(X, V) := \sigma\left((X, V) \mid \mathcal{D}(0,0)\right)$$

$$= \begin{cases} \frac{1}{2}\mathrm{tr}\left(X^T V^\dagger X\right) & \text{if} \quad \mathrm{rge}\, X \in \mathrm{rge}\, V,\, V \succeq 0, \\ +\infty & \text{else.} \end{cases}$$

We have the following two representations of the nuclear norm:

$$\|X\|_* = \min_V \gamma(X, V) + \frac{1}{2}\mathrm{tr}\, V$$

$$\frac{1}{2}\|X\|_*^2 = \min_V \gamma(X, V) + \delta\left(V \mid tr(V) \leq 1\right),$$

Infimal Projections

For a closed proper convex function h, define the infimal projection:

$$\varphi(X) := \inf_{V} \sigma\left((X, V) \mid \mathcal{D}(A, B)\right) + h(V).$$

Infimal Projections

For a closed proper convex function h, define the infimal projection:

$$\varphi(X) := \inf_V \sigma\left((X, V) \mid \mathcal{D}(A, B)\right) + h(V).$$

Theorem

If $\operatorname{dom} h \cap \mathbb{S}^n_{++}(\ker A) \neq \emptyset$, *then*

$$\varphi^*(Y) = \inf\left\{h^*(-W) \mid (Y, W) \in \overline{\operatorname{conv}}\, \mathcal{D}(A, B)\right\}.$$

Infimal Projections with Indicators

When h is an indicator of a closed convex set $\mathcal{V}$,

$$\varphi_{\mathcal{V}}(X) := \inf_{V\in\mathcal{V}} \sigma\left((X,V) \mid \mathcal{D}(A,B)\right),$$

then

$$\varphi_{\mathcal{V}}^{*}(Y) = \tfrac{1}{2}\sigma\left(YY^T \mid \{V\in\mathcal{V} \mid V \succeq_{\ker A} 0\}\right) + \delta\left(Y \mid AY = B\right)$$

$$= \tfrac{1}{2}\sigma\left(YY^T \mid \mathcal{V}\cap\mathbb{S}^n_+(\ker A)\right) + \delta\left(Y \mid AY = B\right)$$

Infimal Projections with Indicators

When h is an indicator of a closed convex set $\mathcal{V}$,

$$\varphi_{\mathcal{V}}(X) := \inf_{V \in \mathcal{V}} \sigma\left((X, V) \mid \mathcal{D}(A, B)\right),$$

then

$$\varphi^*_{\mathcal{V}}(Y) = \tfrac{1}{2}\sigma\left(YY^T \mid \{V \in \mathcal{V} \mid V \succeq_{\ker A} 0\}\right) + \delta\left(Y \mid AY = B\right)$$

$$= \tfrac{1}{2}\sigma\left(YY^T \mid \mathcal{V} \cap \mathbb{S}^n_+(\ker A)\right) + \delta\left(Y \mid AY = B\right)$$

Note that when $B = 0$, both $\varphi_{\mathcal{V}}$ and $\varphi^*_{\mathcal{V}}$ are positively homogeneous of degree 2.
When $A = 0$ and $B = 0$, $\varphi^*_{\mathcal{V}}$ is called a variational Gram functionin Jalali-Xiao-Fazel (2016?).

Ky Fan (p,k) norm

For $p \geq 1$, $1 \leq k \leq \min\{m, n\}$, the *Ky Fan (p,k)-norm* of a matrix $X \in \mathbb{R}^{n \times m}$ is given by

$$\|X\|_{p,k} = \left(\sum_{i=1}^{k} \sigma_i^p \right)^{1/p},$$

where σ_i are the singular values of X sorted in nonincreasing order.

Ky Fan (p,k) norm

For $p \geq 1$, $1 \leq k \leq \min\{m, n\}$, the *Ky Fan (p,k)-norm* of a matrix $X \in \mathbb{R}^{n \times m}$ is given by

$$\|X\|_{p,k} = \left(\sum_{i=1}^{k} \sigma_i^p \right)^{1/p},$$

where σ_i are the singular values of X sorted in nonincreasing order.

- The Ky Fan $(p, \min\{m, n\})$-norm is the Schatten-p norm.
- The Ky Fan $(1, k)$-norm is the standard Ky Fan k-norm.

Ky Fan (p,k) norm

For $p \geq 1$, $1 \leq k \leq \min\{m, n\}$, the *Ky Fan (p,k)-norm* of a matrix $X \in \mathbb{R}^{n \times m}$ is given by

$$\|X\|_{p,k} = \left(\sum_{i=1}^{k} \sigma_i^p \right)^{1/p},$$

where σ_i are the singular values of X sorted in nonincreasing order.

- The Ky Fan $(p, \min\{m, n\})$-norm is the Schatten-p norm.
- The Ky Fan $(1, k)$-norm is the standard Ky Fan k-norm.

Corollary

$$\frac{1}{2}\|X\|^2_{\frac{2p}{p+1},\min\{m,n\}} = \inf_{\|V\|_{p,\min\{m,n\}} \leq 1} \gamma(X, V).$$

Ky Fan (p,k) norm

Corollary

$$\frac{1}{2}\|X\|^2_{\frac{2p}{p+1},\min\{m,n\}} = \inf_{\|V\|_{p,\min\{m,n\}}\leq 1} \gamma(X,V).$$

$$\gamma(X,V) = \sigma\left((X,V) \mid \mathcal{D}(0,0)\right), \qquad \varphi_{\mathcal{V}}(X) := \inf_{V\in\mathcal{V}} \sigma\left((X,V) \mid \mathcal{D}(A,B)\right)$$

$$\text{and} \qquad \gamma_{\mathcal{V}}(X) := \inf_{V\in\mathcal{V}} \sigma\left((X,V) \mid \mathcal{D}(0,0)\right)$$

Proof.

$$\begin{aligned}\left(\inf_{\|V\|_{p,\min\{m,n\}}\leq 1} \gamma(X,V)\right)^* &= \sigma\left(\tfrac{1}{2}XX^T \mid \left\{V \succeq 0 \mid \|V\|_{p,\min\{m,n\}} \leq 1\right\}\right) \\ &= \tfrac{1}{2}\|XX^T\|_{\frac{p}{p-1},\min\{m,n\}} = \tfrac{1}{2}\|X\|^2_{\frac{2p}{p+1},\min\{m,n\}}.\end{aligned}$$

Ky Fan (p,k) norm

As a special case when $p = 1$,

Corollary

$$\tfrac{1}{2}\|X\|_*^2 = \min_{\operatorname{tr} V \le 1} \gamma(X, V).$$

Lemma

Let $\mathcal{V}$ to be the set of rank-k orthogonal projection matrices

$$\mathcal{V} = \left\{ UU^T \mid U \in \mathbb{R}^{n \times k}, U^T U = I_k \right\}, \textit{then} \;\; \tfrac{1}{2}\|X\|_{2,k}^2 = \sigma_{\mathcal{V}}\left(\tfrac{1}{2}XX^T\right).$$

Proof.

A consequence of the following fact [Fillmore-Williams 1971]:

$$\operatorname{conv}\left\{ UU^T \mid U \in \mathbb{R}^{n \times k}, U^T U = I_k \right\} = \left\{ V \in \mathbb{S}^n \mid I \succeq V \succeq 0, \operatorname{tr} V = k \right\}.$$

$\square$

K-means Clustering [Zha-He-Ding-Gu-Simon 2001]

Consider $X \in \mathbb{R}^{n\times m}$, the k-means objective is

$$K(X) := \min_{C,E} \tfrac{1}{2}\|X - EC\|_2^2,$$

where $C \in \mathbb{R}^{k\times m}$ represents the k centers, and E is a $n \times k$ matrix where each row is one of $e_1^T, \ldots, e_k^T$ which correspond to the k cluster assignments.

The optimal C is given by $C = (E^T E)^{-1} E^T X$.

Define $P_E = E(E^T E)^{-1} E^T$, then P_E is an orthogonal projection.

$$\begin{aligned} K(X) &= \min_E \tfrac{1}{2}\|(I - P_E)X\|_2^2 = \tfrac{1}{2}\min_E \operatorname{tr}\left((I - P_E)XX^T\right) \\ &= \tfrac{1}{2}\|X\|_2^2 - \sigma_{\mathcal{P}_k}\left(\tfrac{1}{2}XX^T\right) \geq \tfrac{1}{2}\left(\|X\|_2^2 - \|X\|_{2,k}^2\right). \end{aligned}$$

Best Affine Unbiased Estimator

For a linear regression model $y = A^T\beta + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2 V)$, and a given matrix B, an affine unbiased estimator of $B^T\beta$ is an estimator of the form $\hat{\theta} = X^T y + c$ satisfying $\mathrm{E}\hat{\theta} = B^T\beta$.

$$\boxed{\text{Best: } \mathrm{Var}(\hat{\theta}^*) \preceq \mathrm{Var}(\hat{\theta}),\ \forall\hat{\theta}}$$

Best Affine Unbiased Estimator

For a linear regression model $y = A^T\beta + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2 V)$, and a given matrix B, an affine unbiased estimator of $B^T\beta$ is an estimator of the form $\hat{\theta} = X^T y + c$ satisfying $\mathrm{E}\hat{\theta} = B^T\beta$.

$$\boxed{\text{Best: } \mathrm{Var}(\hat{\theta}^*) \preceq \mathrm{Var}(\hat{\theta}),\ \forall\hat{\theta}}$$

If a solution to

$$v(A, B, V) := \min_{X:AX=B} \tfrac{1}{2}\mathrm{tr}\, X^T V^\dagger X.$$

exists and unique, then $\hat{\theta}^* = (X^*)^T y$.

$$v(A, B, V) = -\sigma_{\mathcal{D}(A,B)}(0, V).$$

The optimal solution X^* satisfies

$$M(V)\begin{pmatrix} X^* \\ W \end{pmatrix} = \begin{pmatrix} 0 \\ B \end{pmatrix}.$$

Supervised Representation Learning

Consider a binary classification problem where we are given the training data: $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^m \times \{-1, 1\}$, and test data: $x_{n+1}, \ldots, x_{n+t} \in \mathbb{R}^m$.

Representation learning aims to learn a feature mapping $\Phi : \mathbb{R}^m \to \mathcal{H}$ that maps the data points to a feature space where points between the two classes are well separated.

Kernel Methods: Instead of specifying the function Φ explicitly, kernel methods consider mapping the data points to a reproducing kernel Hilbert space $\mathcal{H}$ so that the kernel matrix $K \in \mathbb{S}_+^{n+t}$, where $K_{ij} = \langle \Phi(x_i), \Phi(x_j) \rangle$, implicitly determines the mapping Φ.

Supervised Representation Learning

Let $\mathcal{K} \subset \mathbb{S}_+^{n+t}$ be a set of candidate kernels. The best $K \in \mathcal{K}$ can be selected by maximizing its alignment with the kernel specified by the training labels:

$$\varphi^*_{\mathcal{V}}(y) = \max_{K \in \mathcal{V}} \tfrac{1}{2} \left\langle K_{1:n,1:n},\, yy^T \right\rangle,$$

where

$$\mathcal{V} = \mathcal{K} \cap \mathbb{B}_2 \cap \mathbb{S}_+^n,\ \ A = \begin{bmatrix} 0_{n \times n} & 0 \\ 0 & I_{t \times t} \end{bmatrix}, \quad \text{and } B = 0_{(n+t) \times 1}.$$

Multi-task Learning

In multi-task learning, T sets of labelled training data $(x_{t1}, y_{t1}), \ldots, (x_{tn}, y_{tn}) \in \mathbb{R}^m \times \mathbb{R}$ are given, representing T learning tasks.
Assumption: A linear feature map $h_i(x) = \langle u_i, x \rangle$, $i = 1, \ldots, m$, where $U = (u_1, \ldots, u_m)$ is an $m \times m$ orthogonal matrix, and the predictor for each task is $f_t(x) := \langle a_t, h(x) \rangle$.
The multi-task learning problem is then

$$\min_{A,U} \sum_{t=1}^{T} \sum_{i=1}^{m} L_t \left(y_{ti}, \langle a_t, U^T x_{ti} \rangle \right) + \mu \|A\|_{2,1}^2,$$

where $A = (a_1, \ldots, a_T)$, $\|A\|_{2,1}^2$ is the square of the sum of the 2-norm of the rows of A, and L_t is a loss function for each task.
Denote $W = UA$, then the nonconvex problem is equivalent to the following convex problem [Argyriou-Evgeniou-Pontil 2006]:

$$\min_{W,D} \sum_{t=1}^{T} \sum_{i=1}^{m} L \left(y_{ti}, \langle w_t, x_{ti} \rangle \right) + 2\mu \gamma(W, D) \quad \text{s.t.} \quad \mathrm{tr}\, D \leq 1.$$

It is equivalent to

$$\min_{W} \sum_{t=1}^{T} \sum_{i=1}^{m} L \left(y_{ti}, \langle w_t, x_{ti} \rangle \right) + \mu \|W\|_*^2.$$

Thank you !

References I

J. V. Burke and T. Hoheisel: *Matrix Support Functionals for Inverse Problems, Regularization, and Learning*. SIAM Journal on Optimization, 25(2):1135-1159, 2015.