Variational Gram Functions: Convex Analysis and Optimization

Maryam Fazel
University of Washington

Joint work with: Amin Jalali (UW), Lin Xiao (MSR)

SIAM Applied Linear Algebra Conference, Atlanta, Oct 2015

Variational Gram Functions

Let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be vectors in $\mathbb{R}^n$. Given a compact set $\mathcal{M} \subset \mathbb{S}^m$, define

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

which we call **variational Gram function (VGF)** of $\mathbf{x}_1, \ldots, \mathbf{x}_m$ induced by $\mathcal{M}$.

Variational Gram Functions

Let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be vectors in $\mathbb{R}^n$. Given a compact set $\mathcal{M} \subset \mathbb{S}^m$, define

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

which we call **variational Gram function (VGF)** of $\mathbf{x}_1, \ldots, \mathbf{x}_m$ induced by $\mathcal{M}$.

Let $X = [\mathbf{x}_1 \;\cdots\; \mathbf{x}_m] \in \mathbb{R}^{n \times m}$. Pairwise inner products $\mathbf{x}_i^T \mathbf{x}_j$ are entries of the Gram matrix $X^T X$,

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

Variational Gram Functions

Let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be vectors in $\mathbb{R}^n$. Given a compact set $\mathcal{M} \subset \mathbb{S}^m$, define

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

which we call **variational Gram function (VGF)** of $\mathbf{x}_1, \ldots, \mathbf{x}_m$ induced by $\mathcal{M}$.

Let $X = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_m] \in \mathbb{R}^{n \times m}$. Pairwise inner products $\mathbf{x}_i^T \mathbf{x}_j$ are entries of the Gram matrix $X^T X$,

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

a.k.a a support function of set $\mathcal{M}$, at $X^T X$

(support function of set $\mathcal{M}$ is $S_{\mathcal{M}}(Y) = \max_{M \in \mathcal{M}} \langle Y, M \rangle$)

Variational Gram Functions

Examples.

- norms on $\mathbb{R}^m$: for $\mathcal{M} = \{\mathbf{u}\mathbf{u}^T : \|\mathbf{u}\|^\star \leq 1\}$,

$$\Omega(\mathbf{x}) = \max_{M \in \mathcal{M}} \mathbf{x}^T M \mathbf{x} = \|\mathbf{x}\|^2$$

- for ellipsoid $\mathcal{M} = \left\{M : \sum_{i,j=1}^m (M_{ij}/\overline{M}_{ij})^2 \leq 1\right\}$,

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

- for box $\mathcal{M} = \{M : -\overline{M}_{ij} \leq M_{ij} \leq \overline{M}_{ij}\}$,

$$\Omega(X) = \max_{|M_{ij}| \leq \overline{M}_{ij}} \sum_{i,j=1}^m M_{ij}\, \mathbf{x}_i^T \mathbf{x}_j = \sum_{i,j=1}^m \overline{M}_{ij} |\mathbf{x}_i^T \mathbf{x}_j|$$

- for box $\mathcal{M}$ when $n = 1$, $\quad \Omega(\mathbf{x}) = |\mathbf{x}|^T \overline{M} |\mathbf{x}|$

Outline

- motivating applications; interpretations
- convex analysis of VGFs:
  representations, conjugate, subdifferential, prox operator
- optimization algorithms for regularized loss minimization

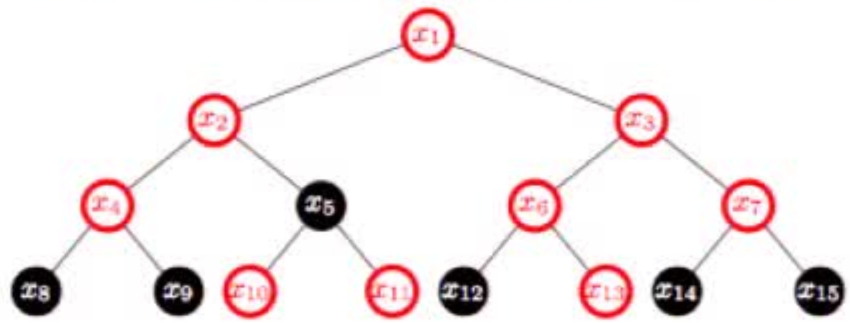
$$\min_X \mathcal{L}(X) + \lambda\Omega(X)$$

- application to a hierarchical classification problem

Motivating Applications

First, a toy example:

- linear measurements of $\mathbf{x} = [x_1 \ \cdots \ x_{15}]$ are given; i.e., $\mathbf{b} = \mathbf{A}\mathbf{x}$.
- $\mathbf{x}$ has at most one nonzero entry on any root-leaf path of this tree



- can minimize

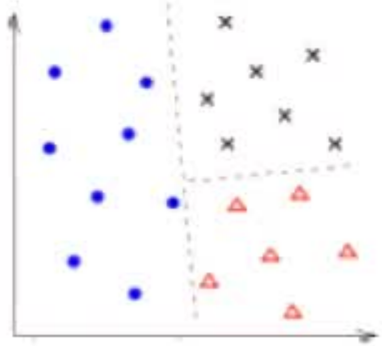
$$\Omega(\mathbf{x}) = \sum_{p} \sum_{(i,j)\in p} w_{ij}|x_i x_j|$$

over $\mathbf{A}\mathbf{x} = \mathbf{b}$.

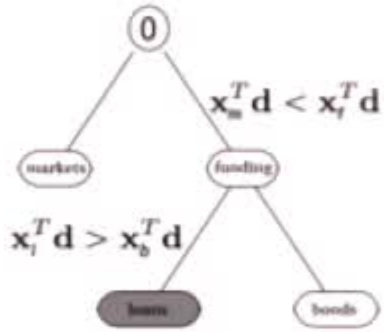
(e.g., exclusive lasso [Zhou,Jin,Hoi '10] nonoverlapping case)

Motivating Applications

A machine learning application: *hierarchical* classification vs flat classification



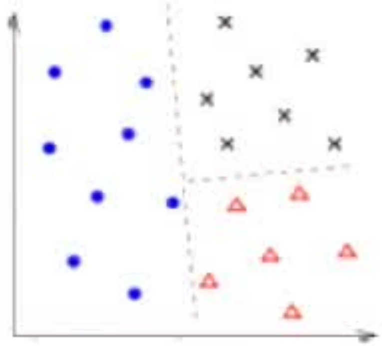
three classes



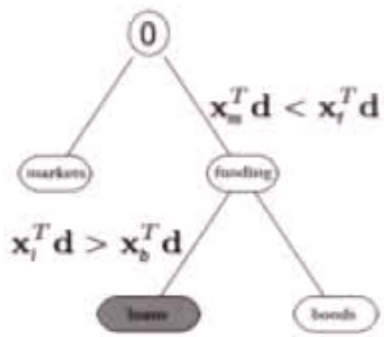
recursive labeling

Motivating Applications

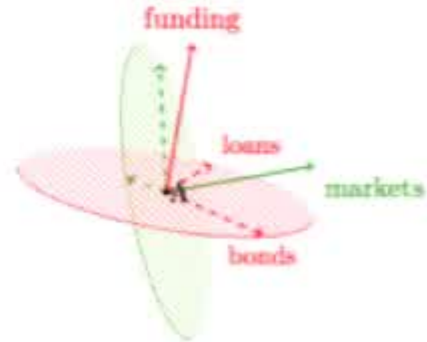
A machine learning application: *hierarchical* classification vs flat classification



three classes



recursive labeling



$\mathbf{x}_{\text{child}} \perp \mathbf{x}_{\text{parent}}$

- classifiers of different layers use different features (or different combinations of same features)
- subspace of classifiers desired to be *orthogonal* to parent classifiers (hierarchical via orthogonal transfer [Zhou,Xiao,Wu'11])
- $\mathbf{x}_l \perp \mathbf{x}_f$ and $\mathbf{x}_b \perp \mathbf{x}_f$ are desired

$$\Omega(\mathbf{x}_m, \mathbf{x}_f, \mathbf{x}_l, \mathbf{x}_b) = w_1 \left|\mathbf{x}_l^T \mathbf{x}_f\right| + w_2 \left|\mathbf{x}_b^T \mathbf{x}_f\right|$$

- other transfer learning methods e.g., [Cai, Hoffman'04; Dekel et al, 04]

Promoting pairwise structure

More generally, for $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$

- $\mathbf{x}_i^T \mathbf{x}_j$'s reveal essential information about relative positions and orientations; can serve as a measure for various properties such as orthogonality
- Minimizing

$$\Omega(\mathbf{x}_1, \ldots, \mathbf{x}_m) = \sum_{i,j=1}^{m} \overline{M}_{ij} \, |\mathbf{x}_i^T \mathbf{x}_j|$$

  promotes pairwise orthogonality for certain pairs specified by $\overline{M}$

[Zhou,Xiao,Wu, '11] introduced this penalty for hierarchical classification.

Promoting pairwise structure

when is it convex?

Theorem (Zhou, Xiao, Wu, '11)

Ω is convex if $\overline{M} \geqslant 0$ and $\widetilde{M}$, the comparison matrix of $\overline{M}$ is PSD, where

$$\widetilde{M} = \begin{cases} -\overline{M}_{ij} & i \neq j \\ \overline{M}_{ii} & i = j \end{cases} ;$$

condition is also necessary if $n \geqslant m - 1$.

proof: brute-force (verify def. of convexity)

question: when is a general VGF convex?

Promoting pairwise structure

More generally, for $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$

- $\mathbf{x}_i^T \mathbf{x}_j$'s reveal essential information about relative positions and orientations; can serve as a measure for various properties such as orthogonality
- Minimizing

$$\Omega(\mathbf{x}_1, \ldots, \mathbf{x}_m) = \sum_{i,j=1}^{m} \overline{M}_{ij} \, |\mathbf{x}_i^T \mathbf{x}_j|$$

promotes pairwise orthogonality for certain pairs specified by $\overline{M}$

[Zhou,Xiao,Wu, '11] introduced this penalty for hierarchical classification.

Convexity

Given compact set $\mathcal{M}$, $\Omega : \mathbb{R}^{n\times m} \to \mathbb{R}$

$$\Omega(X) = \max_{M \in \mathcal{M}} \operatorname{tr}(XMX^T)$$

Theorem

$\Omega(X)$ is convex, if and only if for every X there exists a positive semidefinite *$M \in \mathcal{M}$ satisfying $\Omega(X) = \operatorname{tr}(XMX^T)$.*

intuition: for every X, $\Omega(X)$ can be written as a convex quadratic, hence convex

corollary: when Ω is convex, $\sqrt{\Omega}$ is pointwise max of weighted Frobenius norms

$$\sqrt{\Omega(X)} = \max_{M \in \mathcal{M} \cap \mathcal{S}_+} \|XM^{1/2}\|_F$$

but when is the condition satisfied?

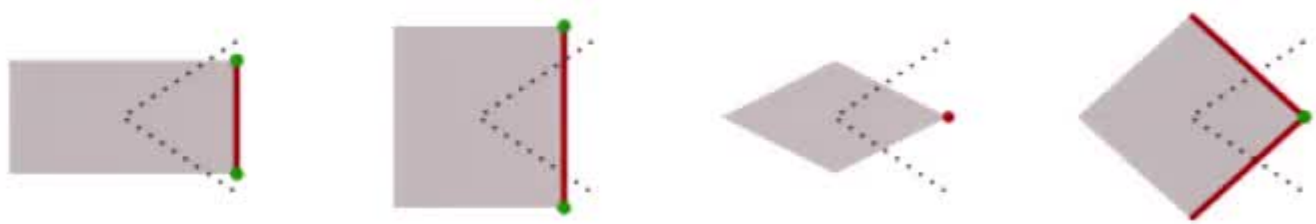
Convexity

polytope: $\mathcal{M} = \mathrm{conv}\{M_1, \ldots, M_p\}$. let $\mathcal{M}_{\text{eff}}$ be the smallest subset of vertices satisfying

$$\max_{M \in \mathcal{M}} \mathrm{tr}(XMX^T) = \max_{M \in \mathcal{M}_{\text{eff}}} \mathrm{tr}(XMX^T), \quad \forall X$$

Theorem

If $\mathcal{M}$ is a polytope, Ω is convex ***if and only if*** $\mathcal{M}_{\text{eff}} \subset \mathbb{S}_+^m$.



gray: set $\mathcal{M}$; red: maximal points w.r.t. PSD cone; green: $\mathcal{M}_{\text{eff}}$

convexity test: check whether green vertices are PSD...

Convexity

Examples.

- For $\mathcal{M} = \{M : \ |M_{ij}| \leqslant \overline{M}_{ij}\}\,, ; \quad \Omega(X) = \sum_{i,j=1}^{m} \overline{M}_{ij} |\mathbf{x}_i^T \mathbf{x}_j|$

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

if $n \geqslant m-1$, $\mathcal{M}_{\text{eff}} \subset \mathbb{S}_+^m$ is equivalent to: comparison matrix of $\overline{M}$ is PSD.

Convexity

Examples.

- For $\mathcal{M} = \{M : |M_{ij}| \leqslant \overline{M}_{ij}\}, ; \quad \Omega(X) = \sum_{i,j=1}^m \overline{M}_{ij} |\mathbf{x}_i^T \mathbf{x}_j|$

  $$\mathcal{M}_{\text{eff}} \subset \{M : M_{ii} = \overline{M}_{ii}, \; M_{ij} = \pm \overline{M}_{ij} \text{ for } i \neq j\}$$

  if $n \geqslant m-1$, $\mathcal{M}_{\text{eff}} \subset \mathbb{S}_+^m$ is equivalent to: comparison matrix of $\overline{M}$ is PSD.

- For $\mathcal{M} = \{M : \sum_{i,j=1}^m (M_{ij}/\overline{M}_{ij})^2 \leqslant 1\}$,

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

  $\overline{M}_{ij} \geqslant 0$ ensures convexity (proof by examining $\mathcal{M}_{\text{eff}}$).

Convexity

Examples.

- Squared norm $\|\mathbf{x}\|^2$ for $\mathbf{x} \in \mathbb{R}^m$ are convex VGFs corresponding to $\mathcal{M} = \{\mathbf{u}\mathbf{u}^T : \|\mathbf{u}\|^\star \leqslant 1\}$

Convexity

Examples.

- Squared norm $\|\mathbf{x}\|^2$ for $\mathbf{x} \in \mathbb{R}^m$ are convex VGFs corresponding to $\mathcal{M} = \{\mathbf{u}\mathbf{u}^T : \|\mathbf{u}\|^\star \leqslant 1\}$
- As a function of *Euclidean distance matrix* $D_{ij} = \frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|_2^2$

$$\Omega_{\mathcal{M}}(X) = \max_{M \in \mathcal{M}} \operatorname{tr}(XMX^T) = \max_{A \in \mathcal{A}} \sum_{i,j} A_{ij}\|\mathbf{x}_i - \mathbf{x}_j\|_2^2$$

where $\mathcal{M} = \{\operatorname{diag}(A\mathbf{1}) - A : \ A \in \mathcal{A}\}$.

simple sufficient condition: $A \geqslant 0$ for all $A \in \mathcal{A} \implies M \succeq 0$ for all $M \in \mathcal{M}$ $\implies \Omega_{\mathcal{M}}$ is convex in X .

Conjugate Function

Conjugate function of $\Omega(X) = \max_{M \in \mathcal{M}} \operatorname{tr}(XMX^T)$ is

$$\Omega^\star(Y) = \frac{1}{2} \inf_{M \in \mathcal{M}} \left\{ \operatorname{tr}(YM^\dagger Y^T) : \operatorname{range}(Y^T) \subseteq \operatorname{range}(M) \right\}$$
$$= \frac{1}{2} \inf_{M,C} \left\{ \operatorname{tr}(C) \ : \begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \succeq 0 \ , M \in \mathcal{M} \right\}$$

and is "semidefinite representable"

Conjugate Function

Conjugate function of $\Omega(X) = \max_{M \in \mathcal{M}} \operatorname{tr}(XMX^T)$ is

$$\Omega^\star(Y) = \frac{1}{2} \inf_{M \in \mathcal{M}} \left\{ \operatorname{tr}(YM^\dagger Y^T) : \operatorname{range}(Y^T) \subseteq \operatorname{range}(M) \right\}$$

$$= \frac{1}{2} \inf_{M,C} \left\{ \operatorname{tr}(C) : \begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \succeq 0 \ , M \in \mathcal{M} \right\}$$

and is "semidefinite representable"

the dual norm (if M's invertible):

$$\sqrt{2\Omega^\star(X)} = \inf_{M \in \mathcal{M}} \|XM^{-1/2}\|_F$$

special case.

- with $\mathcal{M} = \{M : \alpha \mathbf{I} \preceq M \preceq \beta \mathbf{I}, \operatorname{tr}(M) = \gamma\}$, gives *cluster norm* defined by [Jacob, Bach,Vert '08]; can be interpreted as a convex relaxation of k-means.

Subdifferential

$$\Omega(X) = \max_{M \in \mathcal{M}} \operatorname{tr}(XMX^T) = \max_{M \in \mathcal{M}} \sum_{i,j=1}^{m} M_{ij} \mathbf{x}_i^T \mathbf{x}_j$$

subdifferential: $\partial \Omega(X) = \{2XM : M \in \mathcal{M}, \operatorname{tr}(XMX^T) = \Omega(X)\}$

Example:
For $\Omega(X) = \sum_{i,j=1}^{m} \overline{M}_{ij} |\mathbf{x}_i^T \mathbf{x}_j|$,

$$\partial \Omega(X) = \operatorname{conv}\{2XM : M_{ij} = \overline{M}_{ij} \operatorname{sign}(\mathbf{x}_i^T \mathbf{x}_j) \text{ if } \langle \mathbf{x}_i, \mathbf{x}_j \rangle \neq 0, \quad |M_{ij}| \leqslant \overline{M}_{ij} \text{ otherwise}\}$$

([Zhou et al '11] give just one subgradient)

Outline.

- convex analysis of VGFs
- **optimization problems and algorithms**
- connections & applications; numerical experiment

Regularized Loss Minimization

solve regularized loss minimization problem

$$J_{\text{opt}} = \min_{X} \; \mathcal{L}(X\,; \text{data}) + \lambda\,\Omega(X)$$

common losses include: norm loss, Huber loss, hinge, logistic, etc.

- when loss $\mathcal{L}(X)$ is smooth: e.g., can iteratively update variables $X^{(t)}$:

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

γ_t is step size

Regularized Loss Minimization

solve regularized loss minimization problem

$$J_{\text{opt}} = \min_{X} \; \mathcal{L}(X\,;\text{data}) + \lambda\,\Omega(X)$$

common losses include: norm loss, Huber loss, hinge, logistic, etc.

- when loss $\mathcal{L}(X)$ is smooth: e.g., can iteratively update variables $X^{(t)}$:

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

  γ_t is step size
- when $\mathcal{L}(X)$ is not smooth: subgradient-based methods; e.g. Regularized Dual Averaging [Xiao '11]
- convergence can be very slow

Regularized Loss Minimization

solve regularized loss minimization problem

$$J_{\text{opt}} = \min_{X} \ \mathcal{L}(X\,;\text{data}) + \lambda\,\Omega(X)$$

common losses include: norm loss, Huber loss, hinge, logistic, etc.

- when loss $\mathcal{L}(X)$ is smooth: e.g., can iteratively update variables $X^{(t)}$:

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

  γ_t is step size
- when $\mathcal{L}(X)$ is not smooth: subgradient-based methods; e.g. Regularized Dual Averaging [Xiao '11]
- convergence can be very slow

we focus on loss functions with special conjugate structure, that can be exploited together with the structure of the VGF penalty

VGF with Structured Loss Functions

First, exploit the smooth variational representation of a VGF,

$$J_{\text{opt}} = \min_{X} \max_{M \in \mathcal{M}} \; \mathcal{L}(X\,; \text{data}) + \lambda \, \text{tr}(XMX^T)$$

note: robust optimization interpretation

VGF with Structured Loss Functions

First, exploit the smooth variational representation of a VGF,

$$J_{\text{opt}} = \min_{X} \max_{M \in \mathcal{M}} \; \mathcal{L}(X\,; \text{data}) + \lambda \operatorname{tr}(XMX^T)$$

note: robust optimization interpretation

Second, consider loss functions with "nice" representation (called Fenchel-type):

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

where $\hat{\mathcal{L}}(\cdot)$ is convex, $\mathcal{G}$ is compact, and $\mathcal{D}(\cdot)$ is a linear operator.

- luckily, covers many important cases:
  norm loss, Huber loss, binary and muti-class hinge loss...
- Then,

$$J_{\text{opt}} = \min_{X} \max_{\substack{M \in \mathcal{M} \\ G \in \mathcal{G}}} \; \langle X, \mathcal{D}(G) \rangle - \hat{\mathcal{L}}(G) + \lambda \operatorname{tr}(XMX^T)$$

  convex-concave saddle-point problem!

Mirror-Prox Algorithm

$$J_{\text{opt}} = \min_{X} \max_{\substack{M \in \mathcal{M} \\ G \in \mathcal{G}}} \langle X, \mathcal{D}(G) \rangle - \hat{\mathcal{L}}(G) + \lambda \operatorname{tr}(XMX^T)$$

Setup. find the saddle points of smooth convex-concave functions

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

Mirror-prox [Nemirovski '04].

- $O(1/t)$ convergence
- $O(1/t^2)$ convergence if $\mathcal{M} \subset \mathbb{S}_{++}$
- can be used if we can project onto $\mathcal{X}$, $\mathcal{Y}$
- can remove the tuning requirement by an adaptive line search

```
repeat for t = 1, 2, ...
```

$w_t := \operatorname{prox}_{z_t} (\gamma_t F(z_t))$

$z_{t+1} := \operatorname{prox}_{z_t} (\gamma_t F(w_t))$

```
output
```

$$\bar{z}_t := \Big(\sum_{\tau=1}^{t} \gamma_\tau\Big)^{-1} \sum_{\tau=1}^{t} \gamma_\tau w_\tau$$

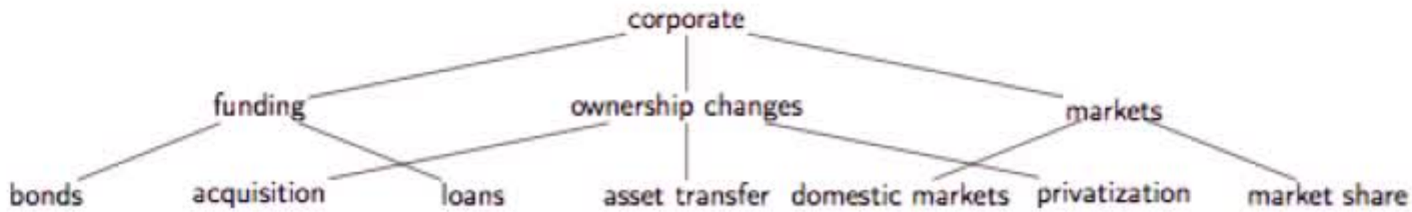
Preprocessing: Reduced Form

$$J_{\text{opt}} = \min_{X \in \mathbb{R}^{n \times m}} \max_{\substack{M \in \mathcal{M} \\ G \in \mathcal{G}}} \langle X, \mathcal{D}(G) \rangle - \hat{\mathcal{L}}(G) + \lambda \operatorname{tr}(XMX^T)$$

- $\mathcal{D}$ determined by the sampled data and the estimation method (regression, classification, etc).
- VGF's variational form can allow reducing the problem; i.e. solve the problem in smaller dimension.

Experiment: Text Categorization

Experiment. Text Categorization for Reuters corpus volume 1: archive of manually categorized news stories. A part of the categories hierarchy:



$$\begin{aligned} \underset{X,\xi}{\text{minimize}} \quad & \frac{1}{N}\sum_{s=1}^{N} \xi_s + \lambda\Omega(X) \\ \text{subject to} \quad & \mathbf{x}_i^T \mathbf{y}_s - \mathbf{x}_j^T \mathbf{y}_s \geqslant 1 - \xi_s\,, \quad \forall j \in \mathcal{S}(i)\,, \forall i \in \mathcal{A}^+(z_s)\,, \forall s \in \{1,\ldots,N\} \\ & \xi_s \geqslant 0\,, \quad \forall s \in \{1,\ldots,N\} \end{aligned}$$

where $\mathbf{y}_s \in \mathbb{R}^n$ are the samples, and $z_s \in \{1,\ldots,m\}$ are the labels, $s = 1,\ldots,N$.

Experiment: Text Categorization

	objective function	convergence rate
Subgradient Method	non-smooth, convex	$\mathcal{O}(1/\sqrt{t})$
Regularized Dual Averaging	non-smooth, **strongly cvx** (σ)	$\mathcal{O}(\ln(t)/\sigma t)$
Mirror-prox	smooth var. form, convex	$\mathcal{O}(1/t)$
Mirror-prox	smooth var. form, strongly convex	$\mathcal{O}(1/t^2)$

FlatMult	HierMult	Transfer	TreeLoss	Orthogonal Transfer
$21.39(\pm 0.29)$	$21.41(\pm 0.29)$	$21.91(\pm 0.31)$	$26.32(\pm 0.39)$	$17.46(\pm 0.74)$

Prediction Error on Test Data

Summary, future work

- VGFs: functions of Gram matrix, defined via weight set $\mathcal{M}$
- unify special cases; lead to new functions
- convex analysis: conjugate, subdifferential, prox
- efficient algorithms

future work:

- design $\mathcal{M}$ for different applications
- other applications:
  multitask learning (with clustered or diverse sets of tasks); disjoint visial features (vision);...

Reference: A. Jalali, L. Xiao, M. Fazel, "Variational Gram Functions: Convex Analysis and Optimization", from website: `faculty/washington.edu/mfazel`

Conjugate Function

Conjugate function of $\Omega(X) = \max_{M \in \mathcal{M}} \operatorname{tr}(XMX^T)$ is

$$\begin{aligned} \Omega^\star(Y) &= \tfrac{1}{2} \inf_{M \in \mathcal{M}} \{\operatorname{tr}(YM^\dagger Y^T) : \operatorname{range}(Y^T) \subseteq \operatorname{range}(M)\} \\ &= \tfrac{1}{2} \inf_{M, C} \left\{ \operatorname{tr}(C) \ : \ \begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \succeq 0 \ , M \in \mathcal{M} \right\} \end{aligned}$$

and is "semidefinite representable"

the dual norm (if M's invertible):

$$\sqrt{2\Omega^\star(X)} = \inf_{M \in \mathcal{M}} \|XM^{-1/2}\|_F$$

special case.

- with $\mathcal{M} = \{M : \alpha \mathbf{I} \preceq M \preceq \beta \mathbf{I}, \ \operatorname{tr}(M) = \gamma\}$, gives *cluster norm* defined by [Jacob, Bach,Vert '08]; can be interpreted as a convex relaxation of k-means.