Variational Functions of Gram Matrices: Convex Analysis and Applications

Maryam Fazel
University of Washington

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

IMA Workshop on Resource Tradeoffs: Computation, Communication, Information
May 2016

Variational Gram Functions

For vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ and a compact set $\mathcal{M} \subset \mathbb{S}^m$, **define** the variational Gram function (VGF)

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

Variational Gram Functions

For vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ and a compact set $\mathcal{M} \subset \mathbb{S}^m$, **define** the variational Gram function (VGF)

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

let $X = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_m]$. pairwise inner products $\mathbf{x}_i^T \mathbf{x}_j$ are entries of 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

For vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ and a compact set $\mathcal{M} \subset \mathbb{S}^m$, **define** the variational Gram function (VGF)

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

let $X = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_m]$. pairwise inner products $\mathbf{x}_i^T \mathbf{x}_j$ are entries of 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 support function of set $\mathcal{M}$, at $X^T X$

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

Variational Gram Functions: examples

- box $\mathcal{M} = \{M : \ -\overline{M}_{ij} \leqslant M_{ij} \leqslant \overline{M}_{ij}\}$

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

(...more on this later)

Variational Gram Functions: examples

- box $\mathcal{M} = \{M : \ -\overline{M}_{ij} \leqslant M_{ij} \leqslant \overline{M}_{ij}\}$

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

(...more on this later)

- box with $n = 1$: $\quad \Omega(\mathbf{x}) = |\mathbf{x}|^T \overline{M} |\mathbf{x}|$

Outline

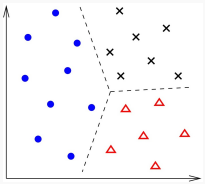
- motivating applications
- **convex analysis** of VGFs: convexity, conjugate, subdifferential
- **optimization algorithms** for regularized loss minimization

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

- application to a **hierarchical classification** problem

Applications

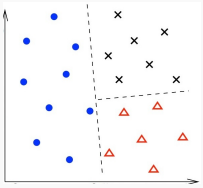
a machine learning application: *hierarchical classification* vs flat classification



three classes

Applications

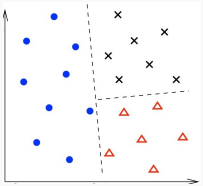
a machine learning application: *hierarchical classification* vs flat classification



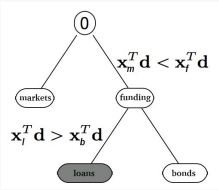
three classes

Applications

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



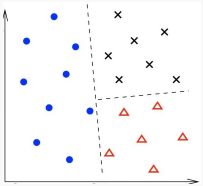
three classes



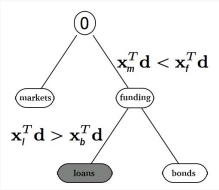
recursive labeling

Applications

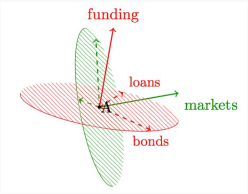
a machine learning application: *hierarchical classification* vs flat classification



three classes



recursive labeling



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

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

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

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

Application: multitask learning

learn multiple tasks simultaneously using shared information among tasks, e.g. structural assumptions on a matrix of classifiers

$$X = \begin{array}{c} \text{tasks} \\ \begin{array}{ccccccc} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \mathbf{x}_4 & \mathbf{x}_5 & \mathbf{x}_6 & \mathbf{x}_7 \end{array} \\ \left[\begin{array}{ccccccc} \times & \times & 0 & 0 & 0 & 0 & 0 \\ \times & \times & 0 & 0 & 0 & 0 & \times \\ \times & \times & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \times & \times & \times & \times & 0 \\ 0 & 0 & \times & \times & \times & \times & \times \\ 0 & 0 & 0 & 0 & 0 & 0 & \times \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \end{array} \text{features}$$

Promoting pairwise structure

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

- $\mathbf{x}_i^T \mathbf{x}_j$'s reveal information about relative orientations; can serve as a measure for properties such as orthogonality
- e.g., 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

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

when is it convex?

Theorem (Zhou,Xiao,Wu, '11)

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

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

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

Promoting pairwise structure

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

when is it convex?

Theorem (Zhou,Xiao,Wu, '11)

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

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

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

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

question: when is a general VGF convex?

Convexity

given compact (not necessarily convex) set $\mathcal{M}$,

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

Theorem (Jalali, F., Xiao)

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

Convexity

given compact (not necessarily convex) set $\mathcal{M}$,

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

Theorem (Jalali, F., Xiao)

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

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

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

Convexity

given compact (not necessarily convex) set $\mathcal{M}$,

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

Theorem (Jalali, F., Xiao)

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

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?

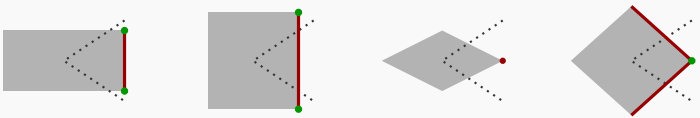
Convexity

polytope: $\mathcal{M} = \mathrm{conv}\{M_1, \ldots, M_p\}$. let $\mathcal{M}_{\text{eff}}$ be the smallest subset 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 $\mathcal{M} = \{M : \ |M_{ij}| \leqslant \overline{M}_{ij}\}\,, \quad \Omega(X) = \sum_{i,j} \overline{M}_{ij} |\mathbf{x}_i^T \mathbf{x}_j|$

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

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

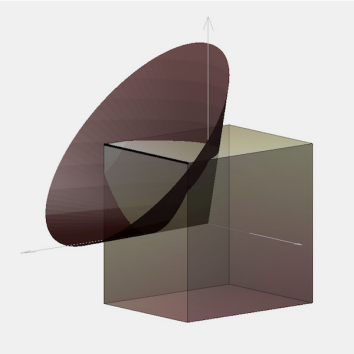
Convexity

examples

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

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

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



Convexity

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

Convexity

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

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

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

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

Conjugate Function

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

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

Conjugate Function

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

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

and is "semidefinite representable"

Conjugate Function

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

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

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 $\mathcal{M} = \{M : \ \alpha \mathbf{I} \preceq M \preceq \beta \mathbf{I}, \ \mathrm{tr}(M) = \gamma\}$, gives *cluster norm* defined by [Jacob, Bach,Vert '08]

Proximal Operator

proximal operator: $\operatorname{prox}_{\tau h}(\mathbf{x}) = \arg\min_{\mathbf{u}} \; h(\mathbf{u}) + \frac{1}{2\tau}\|\mathbf{u}-\mathbf{x}\|_2^2$

$\operatorname{prox}_{\tau\Omega^\star}(X)$ can be computed by an SDP

$$\begin{aligned} \arg\min_{Y} \min_{M,C} \quad & \|Y - X\|_2^2 + \tau \operatorname{tr}(C) \\ \text{subject to} \quad & \begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \succeq \mathbf{0}\,,\; M \in \mathcal{M} \end{aligned}$$

use QR decomposition $X = Q[R_X^T, \mathbf{0}]^T$ to get SDP of size $2m$

Proximal Operator

proximal operator: $\operatorname{prox}_{\tau h}(\mathbf{x}) = \arg\min_{\mathbf{u}} \; h(\mathbf{u}) + \frac{1}{2\tau}\|\mathbf{u}-\mathbf{x}\|_2^2$

$\operatorname{prox}_{\tau\Omega^\star}(X)$ can be computed by an SDP

$$\begin{aligned} \arg\min_{R} \min_{M,C} \quad & \|R - R_X\|_2^2 + \tau \operatorname{tr}(C) \\ \text{subject to} \quad & \begin{bmatrix} M & R^T \\ R & C \end{bmatrix} \succeq \mathbf{0}, \; M \in \mathcal{M} \end{aligned}$$

use QR decomposition $X = Q[R_X^T, \mathbf{0}]^T$ to get SDP of size $2m$

$\operatorname{prox}_{\tau\Omega(\cdot)}(X) = X - \operatorname{prox}_{\tau\Omega^\star(\cdot)}(X)$

Subdifferential

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

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

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

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

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

Outline

- motivating applications
- **convex analysis** of VGFs: convexity, conjugate, subdifferential
- **optimization algorithms** for regularized loss minimization

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

- application to a **hierarchical classification** problem

Regularized Loss Minimization

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

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

Regularized Loss Minimization

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

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

- with smooth loss (and prox cheap): iteratively update $X^{(t)}$:

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

  γ_t is step size

Regularized Loss Minimization

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

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

- with smooth loss (and prox cheap): iteratively update $X^{(t)}$:

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

  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

VGF with Structured Loss Functions

exploit smooth variational representation of a VGF,

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

VGF with Structured Loss Functions

exploit smooth variational representation of a VGF,

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

and consider losses 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, $\mathcal{D}(\cdot)$ is a linear operator

VGF with Structured Loss Functions

exploit smooth variational representation of a VGF,

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

and consider losses 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, $\mathcal{D}(\cdot)$ is a linear operator

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

VGF with Structured Loss Functions

exploit smooth variational representation of a VGF,

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

and consider losses 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, $\mathcal{D}(\cdot)$ is a linear operator

- luckily, covers many important cases:
  norm loss, Huber loss, binary and muti-class hinge loss...
- then,
  $$J_{\text{opt}} = \min_{X} \max_{\substack{M \in \mathcal{M} \\ G \in \mathcal{G}}} \; \langle X, \mathcal{D}(G) \rangle - \hat{\mathcal{L}}(G) + \lambda \operatorname{tr}(XMX^T)$$
  smooth 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

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

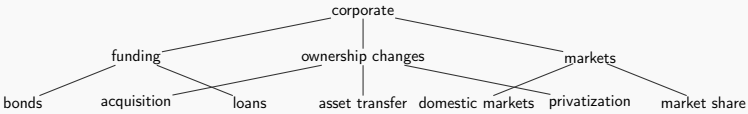
Mirror-prox [Nemirovski '04]:

- $O(1/t)$ convergence
- $O(1/t^2)$ convergence if strongly convex
- useful if projection (or prox) is cheap

often can preprocess to reduce to smaller dimension

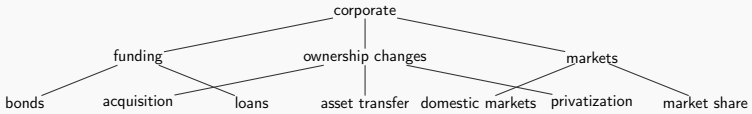
Experiment: Text Categorization

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



Experiment: Text Categorization

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

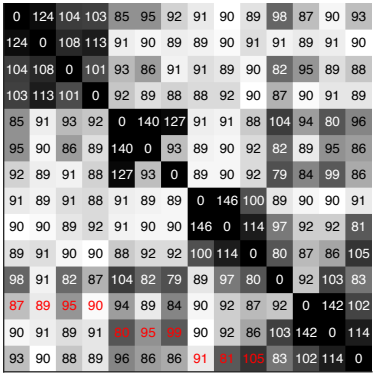


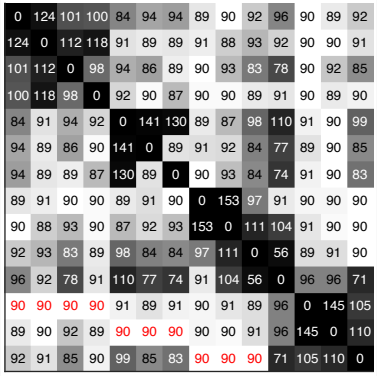
$$\begin{aligned} \underset{X,\,\xi}{\text{minimize}} \quad & \tfrac{1}{N}\sum_{s=1}^{N}\xi_s + \lambda\Omega(X) \\ \text{subject to} \quad & \mathbf{x}_i^T\mathbf{y}_s - \mathbf{x}_j^T\mathbf{y}_s \geqslant 1-\xi_s\,, \quad \forall j \in \mathcal{S}(i)\,, \forall i \in \mathcal{A}^+(z_s)\,, \forall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

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

sanity check: angles between pairs of classifiers

left: flat classification; **right:** hierarchical classification





red: pairs desired to be orthogonal

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 (from [Zhou,Xiao, Wu'11])

Summary, future work

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

future work:

- design $\mathcal{M}$ for different applications
- other applications:
  multitask learning (with clustered/diverse tasks); disjoint visual features;...

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

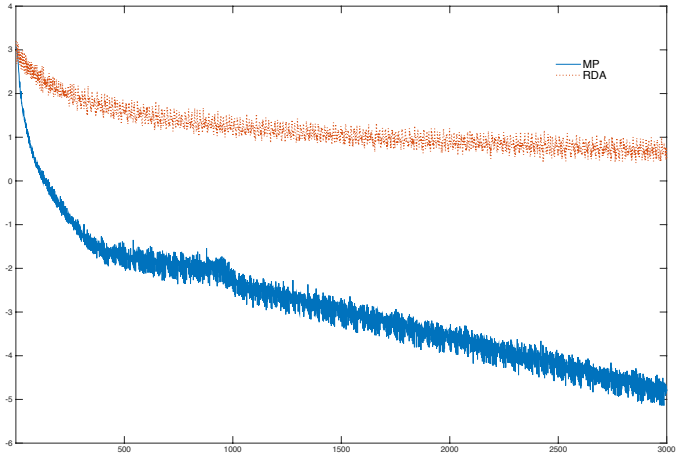


Figure: Value of loss function vs iteration t for MP and RDA algorithms (log scale)

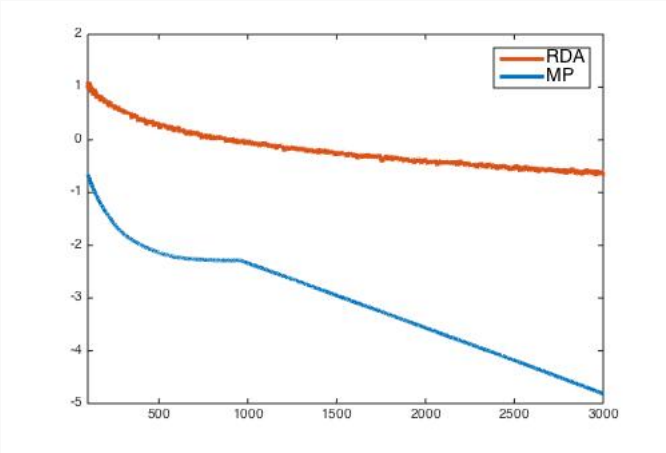


Figure: $\|X_t - X_{\text{final}}\|_F$ vs iteration t for MP and RDA algorithms (log scale)

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`