Regularized Learning and Gram Matrices

Maryam Fazel
University of Washington

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

SAMSI Optimization Program, Opening Workshop
Aug 29, 2016
Regularized Learning of Structured Models

\[
\min_x \mathcal{L}(x; \text{data}) + \lambda \Omega(x)
\]

- \(x\): model parameters
- \(\mathcal{L}\): loss function, measures model fit
- \(\Omega\): \textit{regularizer}, often promotes desired \textit{structure} in \(x\) (Bayesian view: prior on \(x\))
Regularized Learning of Structured Models

\[
\min_x \mathcal{L}(x; \text{data}) + \lambda \Omega(x)
\]

- \textbf{x}: model parameters
- \textbf{\mathcal{L}}: loss function, measures model fit
- \textbf{\Omega}: regularizer, often promotes desired \textit{structure} in \textbf{x} (Bayesian view: prior on \textbf{x})

examples:

- \|x\|_1: sparsity (compressed sensing, LASSO,...)
- \|x\|_{1,2}: group sparsity (group-LASSO)
- \sum_i |x_{i+1} - x_i| = \|Dx\|_1: total variation (image processing, fused LASSO,...)
- \|X\|_*: low-rank (matrix recovery, completion,...)

Big data, machine learning, signal processing,...
"good" $\Omega$: 

- statistical guarantees
- computational efficiency (convex, scalable algorithms)
- interacts well with loss $\mathcal{L}$

More complicated structures need new $\Omega$
“good" $\Omega$:

- statistical guarantees
- computational efficiency (convex, scalable algorithms)
- interacts well with loss $\mathcal{L}$

More complicated structures need new $\Omega$

**This talk**: overview of regularization with *Gram matrices*
A Motivating Application

in machine learning: *hierarchical classification* vs flat classification

three classes
A Motivating Application

in machine learning: \textit{hierarchical classification} vs flat classification

three classes
A Motivating Application

in machine learning: *hierarchical classification* vs flat classification

three classes

recursive labeling
A Motivating Application

in machine learning: *hierarchical classification* vs flat classification

- classifiers of different levels use different features (or different combinations of features)
- different transfer learning methods e.g., [Cai, Hoffman’04; Dekel et al, 04]
A Motivating Application

in machine learning: *hierarchical classification* vs flat classification

- classifiers of different levels use different features (or different combinations of features)
- different transfer learning methods e.g., [Cai, Hoffman’04; Dekel et al, 04]
- orthogonal transfer: classifiers desired to be *orthogonal* to parent classifiers
- encourage $x_l \perp x_f$ and $x_b \perp x_f$ [Zhou,Xiao,Wu’11]
Application: multitask learning e.g., [Evgeniou, et al '05]

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

\[
X = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]
Promoting pairwise structure

For $x_1, \ldots, x_m \in \mathbb{R}^n$

- $x_i^T x_j$'s reveal information about relative orientations; can serve as a measure for properties such as orthogonality
- E.g., minimizing

$$
\Omega(x_1, \ldots, x_m) = \sum_{i,j=1}^{m} M_{ij} |x_i^T x_j|
$$

promotes pairwise orthogonality for certain pairs, specified by $M$.

Variational Gram Functions

For vectors $x_1, \ldots, x_m \in \mathbb{R}^n$ and a compact set $M \subset \mathbb{S}^m$, define the variational Gram function (VGF)

$$\Omega_M(x_1, \ldots, x_m) = \max_{M \in \mathcal{M}} \sum_{i,j=1}^{m} M_{ij} x_i^T x_j$$
Variational Gram Functions

For vectors $x_1, \ldots, x_m \in \mathbb{R}^n$ and a compact set $\mathcal{M} \subset \mathbb{S}^m$, define the variational Gram function (VGF)

$$\Omega_{\mathcal{M}}(x_1, \ldots, x_m) = \max_{M \in \mathcal{M}} \sum_{i,j=1}^{m} M_{ij} x_i^T x_j$$

Let $X = [x_1 \cdots x_m]$. Pairwise inner products $x_i^T x_j$ are entries of Gram matrix $X^TX$,

$$\Omega_{\mathcal{M}}(X) = \max_{M \in \mathcal{M}} \langle X^T X, M \rangle = \max_{M \in \mathcal{M}} \text{tr}(X M X^T)$$
Variational Gram Functions

For vectors $x_1, \ldots, x_m \in \mathbb{R}^n$ and a compact set $\mathcal{M} \subset \mathbb{S}^m$, define the variational Gram function (VGF)

$$
\Omega_{\mathcal{M}}(x_1, \ldots, x_m) = \max_{M \in \mathcal{M}} \sum_{i,j=1}^{m} M_{ij} x_i^T x_j
$$

let $X = [x_1 \cdots x_m]$. pairwise inner products $x_i^T x_j$ are entries of Gram matrix $X^T X$,

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

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

(support function of set $\mathcal{M}$: $S_{\mathcal{M}}(Y) = \max_{M \in \mathcal{M}} \langle Y, M \rangle$)
\[ \Omega_M(X) = \sup_{M \in \mathcal{M}} \text{tr}(XMX^T) \]

<table>
<thead>
<tr>
<th>(\Omega_M)</th>
<th>(\mathcal{M})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\sum_{i,j=1}^{m} M_{ij}</td>
<td>x_i^T x_j</td>
</tr>
<tr>
<td>dual “cluster norm”</td>
<td>({ M : \alpha I \preceq M \preceq \beta I , \text{tr}(M) = \gamma })</td>
</tr>
<tr>
<td>(|UXV^T| = |X| = g(\sigma(X)))</td>
<td>(M \in \mathcal{M} \iff UMV^T \in \mathcal{M})</td>
</tr>
<tr>
<td>(</td>
<td>x</td>
</tr>
<tr>
<td>(|x|^2)</td>
<td>({ uu^T : |u|^* \leq 1})</td>
</tr>
<tr>
<td>(\max_{\theta \in \text{diag}(M)} \sum_{i=1}^{m} \theta_i x_i^2)</td>
<td>a subset of diagonal matrices</td>
</tr>
<tr>
<td>dual (k)-support norm</td>
<td>({ \text{diag}(\theta) : 0 \leq \theta_i \leq 1 , \ 1^T \theta = k })</td>
</tr>
<tr>
<td>dual (d)-valued norm</td>
<td>({ \bigoplus_{i=1}^{d} \frac{1}{</td>
</tr>
</tbody>
</table>
Promoting pairwise structure

\[ \Omega(x_1, \ldots, x_m) = \sum_{i,j} M_{ij} |x_i^T x_j| \]

when is it convex?

**Theorem (Zhou, Xiao, Wu, ‘11)**

\( \Omega \) is convex if \( \overline{M} \succeq 0 \), 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 \( n \geq m - 1 \).
Promoting pairwise structure

\[ \Omega(x_1, \ldots, x_m) = \sum_{i,j} M_{ij} |x_i^T x_j| \]

when is it convex?

**Theorem (Zhou, Xiao, Wu, '11)**

\( \Omega \) is convex if \( \widetilde{M} \geq 0 \), and \( \widetilde{M} \), the comparison matrix of \( M \), is PSD.

\[ \widetilde{M} = \begin{cases} -M_{ij} & i \neq j \\ M_{ii} & i = j \end{cases} \]

Condition is also necessary if \( n \geq m - 1 \).

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

**question:** when is a general VGF convex?
Convexity

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

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

**Theorem (Jalali, F., Xiao)**

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

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

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

**Theorem (Jalali, F., Xiao)**

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

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

$$\sqrt{\Omega}(X) = \max_{M \in \mathcal{M} \cap S_+} \|XM^{1/2}\|_F$$
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 $\Omega$ is convex, $\sqrt{\Omega}$ is pointwise max of weighted Frobenius norms

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

but when is the condition satisfied?
Convexity

**Polytope:** \( \mathcal{M} = \text{conv}\{M_1, \ldots, M_p\} \). Let \( \mathcal{M}_{\text{eff}} \) be the smallest subset satisfying

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

**Theorem**

If \( \mathcal{M} \) is a polytope, \( \Omega \) is convex if and only if \( \mathcal{M}_{\text{eff}} \subset S^m_{+} \).

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 $\mathcal{M} = \{ M : |M_{ij}| \leq \overline{M}_{ij} \}$, $\Omega(X) = \sum_{i,j} \overline{M}_{ij} |x_i^T x_j|$

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

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

examples

- for $\mathcal{M} = \{M : |M_{ij}| \leq \overline{M}_{ij}\}$, $\Omega(X) = \sum_{i,j} \overline{M}_{ij} |x_i^T x_j|$

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

if $n \geq m - 1$, $\mathcal{M}_{\text{eff}} \subset S_+^m$ is equivalent to: comparison matrix of $\overline{M}$ is PSD.
Convex analysis of VGF

- conjugate function is “semidefinite representable"
- dual norm
- proximal mapping is solution to a (related) SDP
- explicit form for subdifferential (complete characterization, not just one subgradient)
- VGF calculus
Regularized Learning

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

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

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

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

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

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

\(\gamma_t\) is step size
Regularized Learning

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

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

- with smooth loss (and cheap prox/projection): iteratively update \( X^{(t)} \):
  \[
  X^{(t+1)} = \text{prox}_{\gamma_t \Omega} \left( X^{(t)} - \gamma_t \nabla \mathcal{L}(X^{(t)}) \right), \quad t = 0, 1, 2, \ldots ,
  \]

\( \gamma_t \) 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,

\[ J_{\text{opt}} = \min_{X} \max_{M \in M} \mathcal{L}(X; \text{data}) + \lambda \text{tr}(XX^T) \]
VGF with Structured Loss Functions

exploit smooth variational representation of a VGF,

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

for losses with “nice” representation (called Fenchel-type):

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

where $\hat{\mathcal{L}}(\cdot)$ is convex, $G$ is compact, $\mathcal{D}(\cdot)$ is a linear operator.
VGF with Structured Loss Functions

exploit smooth variational representation of a VGF,

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

for losses with “nice” representation (called Fenchel-type):

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

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

- luckily, covers many important cases:
  - norm loss, Huber loss, binary and multi-class hinge loss...
VGF with Structured Loss Functions

exploit smooth variational representation of a VGF,

\[ J_{\text{opt}} = \min_X \max_{M \in \mathcal{M}} \mathcal{L}(X; \text{data}) + \lambda \text{tr}(XX^T) \]

for 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 multi-class hinge loss.

- then,

\[ J_{\text{opt}} = \min_X \max_{M \in \mathcal{M}} \max_{G \in \mathcal{G}} \langle X, \mathcal{D}(G) \rangle - \hat{\mathcal{L}}(G) + \lambda \text{tr}(XX^T) \]

smooth convex-concave saddle-point problem!
VGF with Structured Loss Functions

exploit smooth variational representation of a VGF,

for 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,

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

smooth convex-concave saddle-point problem!

Fenchel-type representation (Loss) + left unitary invariance (VGF) \(\rightarrow\) Kernel trick \(\rightarrow\) reduced-size problem
Mirror-Prox Algorithm

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

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

![Diagram of categories hierarchy]

$$\begin{align*}
\text{minimize} & \quad \frac{1}{N} \sum_{s=1}^{N} \xi_s + \lambda \Omega(X) \\
\text{subject to} & \quad x_i^T y_s - x_j^T y_s \geq 1 - \xi_s, \quad \forall j \in S(i), \forall i \in A^+(z_s), \forall s \in \{1, \ldots, N\} \\
& \quad \xi_s \geq 0, \quad \forall s \in \{1, \ldots, N\}
\end{align*}$$

where $y_s \in \mathbb{R}^n$ are the samples, and $z_s \in \{1, \ldots, m\}$ are the labels, $s = 1, \ldots, N$. 
sanity check: angles between pairs of classifiers

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

red: pairs desired to be orthogonal
Experiment: Text Categorization

<table>
<thead>
<tr>
<th>Method</th>
<th>Objective Function</th>
<th>Convergence Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgradient Method</td>
<td>non-smooth, convex</td>
<td>$\mathcal{O}(1/\sqrt{t})$</td>
</tr>
<tr>
<td>Regularized Dual Averaging</td>
<td>non-smooth, strongly cvx ($\sigma$)</td>
<td>$\mathcal{O}(\ln(t)/\sigma t)$</td>
</tr>
<tr>
<td>Mirror-prox</td>
<td>smooth var. form, convex</td>
<td>$\mathcal{O}(1/t)$</td>
</tr>
<tr>
<td>Mirror-prox</td>
<td>smooth var. form, strongly convex</td>
<td>$\mathcal{O}(1/t^2)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Prediction Error on Test Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>FlatMult</td>
<td>21.39(±0.29)</td>
</tr>
<tr>
<td>HierMult</td>
<td>21.41(±0.29)</td>
</tr>
<tr>
<td>Transfer</td>
<td>21.91(±0.31)</td>
</tr>
<tr>
<td>TreeLoss</td>
<td>26.32(±0.39)</td>
</tr>
<tr>
<td>Orthogonal Transfer</td>
<td>17.46(±0.74)</td>
</tr>
</tbody>
</table>

(prediction error on test data (from [Zhou, Xiao, Wu ’11])
Summary

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

Future work:
- design \( \mathcal{M} \) for different applications
- more general "composed" penalties \( p(s(x)) \)