Low Rank Estimation of Smooth Kernels on Weighted Graphs

Vladimir Koltchinskii
Based on a joint work with Pedro Rangel

School of Mathematics
Georgia Institute of Technology

July 2013
Matrix Completion

- $V$ a set, $\text{card}(V) = m$
- $(X, X', Y)$ a random variable in $V \times V \times [-a, a]$, $a > 0$
- $X, X'$ independent points sampled from the uniform distribution $\Pi$ in $V$
- $Y \in [-a, a]$ a response variable
Matrix Completion

- $S_*(u, v) := \mathbb{E}(Y|X = u, X' = v)$
- Assume that $S_* \in S_V$, where

$$S_V := \left\{ S : S(u, v) = S(v, u), u, v \in V \right\}$$

- $(X_1, X_1', Y_1), \ldots, (X_n, X'_n, Y_n)$ i.i.d. copies of $(X, X', Y)$ (data)
- **Goal**: estimate $S_*$ based on the data
- **Assumption**: $S_*$ is low rank
Noiseless problem:

\[ Y_j := S_*(X_j, X'_j), \quad j = 1, \ldots, n \]

\[ \hat{S} := \text{argmin} \left\{ \|S\|_1 : Y_j = S(X_j, X'_j), j = 1, \ldots, n, S \in S_V \right\}, \]

where \( \| \cdot \|_1 \) is \textbf{the nuclear norm}:

\[ \|S\|_1 := \text{tr}(|S|) = \text{tr}\left(\sqrt{S^2}\right) \]

Noiseless Matrix Completion: Low Coherence Conditions

\[ S_* = \sum_{j=1}^{r} \mu_j (\psi_j \otimes \psi_j), \quad \text{sign}(S_*) := \sum_{j=1}^{r} \text{sign}(\mu_j)(\psi_j \otimes \psi_j) \]

\[ L := \text{l.s.}(\{\psi_j : 1 \leq j \leq r\}) \]

\{e_1, \ldots, e_m\} the canonical basis of \( \mathbb{R}^V \) with standard Euclidean inner product \( \langle \cdot, \cdot \rangle \)

**Low Coherence.** For some \( \nu > 0 \),

\[ \| P_L e_j \|^2 \leq \frac{\nu r}{m}, \quad j = 1, \ldots, m \]

and

\[ \left| \langle \text{sign}(S_*) e_i, e_j \rangle \right|^2 \leq \frac{\nu r}{m^2}, \quad i, j = 1, \ldots, m. \]
The following result is due to Candes and Tao (2010), Gross (2011):

**Theorem**

There exists a constant $C > 0$ such that

$$\mathbb{P}\{\hat{S} \neq S_*\} \leq m^{-2}$$

for all

$$n \geq C_{\nu} r m \log^2 m.$$

Let $S^a_V := \left\{ S : S \in S_V, \max_{u,v \in V} |S(u, v)| \leq a \right\}$

$$\hat{S}^\varepsilon := \arg\min_{S \in S^a_V} \left\{ n^{-1} \sum_{j=1}^n (Y_j - S(X_j, X'_j))^2 + \varepsilon \| S \|_1 \right\}$$

$\varepsilon > 0$ regularization parameter

The error of this estimator will be expressed in terms of the squared $L_2(\Pi^2)$-norm:

$$\| S \|_{L_2(\Pi^2)}^2 = m^{-2} \sum_{u,v \in V} |S(u, v)|^2.$$
A Low Rank Oracle Inequality

Koltchinskii (2012)

Given \( t > 0 \), \( \bar{t} := t + 3 \log \log \left( m \lor n \lor a \lor a^{-1} \lor 2 \right) \).

**Theorem**

There exist constants \( C, D > 0 \) such that, for all \( t > 0 \) and for all \( \varepsilon \geq D a \left( \sqrt{\frac{t + \log(2m)}{mn}} \lor \sqrt{\frac{t + \log(2m)}{n}} \right) \),

with probability at least \( 1 - e^{-t} \),

\[
\| \hat{S}^\varepsilon - S_* \|_{L^2(\mathbb{P}^2)}^2 \leq \inf_{S \in S^a_v} \left[ \| S - S_* \|_{L^2(\mathbb{P}^2)}^2 + C m^2 \varepsilon^2 \text{rank}(S) + \frac{a^2 \bar{t}}{n} \right].
\]
Remarks

- Take

\[ \varepsilon = Da\sqrt{\frac{t + \log(2m)}{nm}} \]

and suppose that

\[ \frac{m(t + \log(2m))}{n} \leq 1, \quad \bar{t} \leq m(t + \log(2m)) \]

- Then, with probability at least \( 1 - e^{-t} \)

\[ \| \hat{S}^\varepsilon - S_* \|_{L_2(\mathbb{P}^2)}^2 \leq C \frac{a^2 m \text{rank}(S_*)(t + \log(2m))}{n} \]
Koltchinskii, Lounici and Tsybakov (2011)

Let $1 \leq r \leq m$ and let $\mathcal{P}_{r,a}$ be the set of all distributions $P$ of $(X, X', Y)$ such that $X, X'$ are independent, $X, X' \sim \Pi$, $|Y| \leq a$ a.s. and $\mathbb{E}(Y|X, X') = S_P(X, X')$, where $S_P \in S_V$, $\text{rank}(S_P) \leq r$.

**Theorem**

There exist constants $c_1, c_2 > 0$ such that

$$\inf_{\hat{S}} \sup_{P \in \mathcal{P}_{r,a}} \mathbb{P}_P \left\{ \| \hat{S} - S_P \|_{L_2(\Pi^2)}^2 \geq c_1 \frac{a^2 m r}{n} \right\} \geq c_2,$$

where the infimum is taken over all estimators $\hat{S}$ based on i.i.d. data $(X_j, X'_j, Y_j), j = 1, \ldots, n$ sampled from $P$. 
Weighted Graphs and Laplacians

- \((V, A)\) **weighted graph** with a weight matrix \(A := (a(u, v))_{u, v \in V}\), where \(a(u, v) \geq 0\), \(a(u, v) = a(v, u)\), \(u, v \in V\).

- \(\text{deg}(u) = \sum_{v \in V} a(u, v)\) the degree of vertex \(u \in V\).

- \(\Delta := D - A\) (the **Laplacian**), \(D\) the diagonal matrix with \(\text{deg}(u), u \in V\) on the diagonal.

- \(\Delta\) can be viewed as an operator from \(\mathbb{R}^V\) into itself.

\[
\langle \Delta f, f \rangle = \frac{1}{2} \sum_{u, v \in V} a(u, v)(f(u) - f(v))^2, f : V \mapsto \mathbb{R}
\]
Discrete Sobolev Norms and Smooth Kernels

- $S \in S_V$ a kernel with spectral representation

$$S = \sum_{j=1}^{m} \mu_j (\psi_j \otimes \psi_j),$$

- $\{\mu_j\}$ are its eigenvalues
- $\{\psi_j\}$ are its orthonormal (in $\mathbb{R}^V$) eigenfunctions
- The "smoothness" of $S$ can be characterized by discrete Sobolev norms $\|\Delta^{p/2} S\|_{L^2(\Pi^2)}$, $p > 0$:

$$\|\Delta^{p/2} S\|_{L^2(\Pi^2)}^2 = \sum_{j=1}^{m} \mu_j^2 \|\Delta^{p/2} \psi_j\|_{L^2(\Pi)}^2.$$
\( W = \Delta^p, \ p > 0 \) is a fixed constant.

**Spectral Representation**

\[
W = \sum_{j=1}^{m} \lambda_j (\phi_j \otimes \phi_j)
\]

0 \( \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m \) the eigenvalues of \( W \)

\( \phi_1, \phi_2, \ldots, \phi_m \) the corresponding orthonormal (in \( \mathbb{R}^V \))
eigenfunctions of \( W \) (and of \( \Delta \))

\( \lambda_{k+1} \leq c\lambda_k \) (if \( \lambda_k > 0 \))
Goal: estimate the kernel $S_*$ based on the training data $(X_1, X'_1, Y_1), \ldots, (X_n, X'_n, Y_n)$, assuming that $S_*$ is low rank and "smooth" on the graph.

Let $r := \text{rank}(S_*)$

The "smoothness" of $S_*$ will be characterized by

$$\rho := \| W^{1/2} S_* \|_{L^2(\Pi^2)}.$$
Heuristic Considerations

- $S_{*,l} := \sum_{i,j=1}^{l} \langle S_{*} \phi_i, \phi_j \rangle (\phi_i \otimes \phi_j)$
- $\| S_{*} - S_{*,l} \|_{L_2(\Pi)}^2 \leq \frac{2\rho^2}{\lambda_{l+1}}, \ l = 1, \ldots, m$
- If, for each $l$, one can estimate $S_{*,l}$ with the squared $L_2(\Pi^2)$-error

$$\sim \frac{a^2 (r \wedge l) l}{n},$$

then one can expect that the squared $L_2(\Pi^2)$-error of estimation of $S_{*}$ would be

$$\min_{1 \leq l \leq m} \left[ \frac{a^2 (r \wedge l) l}{n} \lor \frac{\rho^2}{\lambda_{l+1}} \right]$$

- It is easy to see that

$$\min_{1 \leq l \leq m} \left[ \frac{a^2 (r \wedge l) l}{n} \lor \frac{\rho^2}{\lambda_{l+1}} \right] \approx \max_{1 \leq l \leq m} \left[ \frac{a^2 (r \wedge l) l}{n} \land \frac{\rho^2}{\lambda_l} \right]$$
$S_{r,\rho} := \{ S \in S_V : \text{rank}(S) \leq r, \| W^{1/2}S \|_{L_2(\Pi^2)} \leq \rho \}$

$\mathcal{P}_{r,\rho,a} := \left\{ P : (X, X', Y) \sim P, X, X' \text{ independent } \sim \Pi, |Y| \leq a, \mathbb{E}(Y|X, X') = S_*(X, X') = S_P(X, X'), S_P \in S_{r,\rho} \right\}$

$\bar{\phi}_j := \sqrt{m}\phi_j, \ j = 1, \ldots, m \ (\{\bar{\phi}_j\} \text{ are orthonormal in } L_2(\Pi))$.

$Q_p := \max_{1 \leq j \leq m} \| \bar{\phi}_j \|_{L_p(\Pi)}^2$. 
Lower Bounds for Smooth and Low Rank Kernels

\[ \delta_n(r, \rho, a) := \max_{1 \leq l \leq m} \left[ \frac{a^2 (r \wedge l)}{n} \wedge \frac{\rho^2}{\lambda_l} \wedge \frac{1}{(p - 1) Q_p^2 m^{4/p}} \frac{a^2 (r \wedge l)}{l} \right]. \]

**Theorem**

There exist constants \( c_1, c_2 > 0 \) such that

\[ \inf_{\hat{S}} \sup_{P \in \mathcal{P}_{r, \rho, a}} \mathbb{P}_P \left\{ \| \hat{S} - S_P \|_{L_2(\Pi^2)}^2 \geq c_1 \delta_n(r, \rho, a) \right\} \geq c_2, \]

where the infimum is taken over all estimators \( \hat{S} \) based on i.i.d. data \((X_j, X'_j, Y_j), j = 1, \ldots, n\) sampled from \( P \).
Lower Bounds for Smooth and Low Rank Kernels: Example

- $\lambda_k \asymp k^{2\beta}$ for some $\beta > 1/2$.
- $n \geq C' Q_{\infty}^{(\beta+1)/\beta} (\log m)^{(\beta+1)/2\beta} \left( \frac{\rho}{a} \right)^{1/\beta}$

$$
\delta_n(r, \rho, a) \asymp \left( \left( \frac{a^2 \rho^{1/\beta} r}{n} \right)^{2\beta/(2\beta+1)} \right) \wedge \left( \frac{a^2 \rho^{2/\beta}}{n} \right)^{\beta/(\beta+1)} \wedge \left( \frac{a^2 rm}{n} \right) \vee \frac{a^2}{n}.
$$
Least Squares Estimators with Nonconvex Penalties

\[ S_r(l; a) := \left\{ S \in S_V : \text{rank}(S) \leq r, \|S\|_{L_2(\Pi)} \leq a, S = \sum_{i,j=1}^{l} s_{ij}(\phi_i \otimes \phi_j) \right\} \]

- **S^a truncation of** \( S \in S_V : \)
  \[
  S^a(u, v) = S(u, v)I(|S(u, v)| \leq a) + aI(S(u, v) > a) \\
  -aI(S(u, v) < -a)
  \]

- \( \bar{S}_r(l; a) := \{ S^a : S \in S_r(l; a) \} \)
\[ \hat{S}_{r,l,a} := \arg \min_{S \in S_{r,l,a}} \frac{1}{n} \sum_{j=1}^{n} (Y_j - S(X_j, X'_j))^2 \]

\[ (\hat{r}, \hat{l}) := \arg \min_{r,l} \left[ \frac{1}{n} \sum_{j=1}^{n} (Y_j - \hat{S}_{r,l,a}(X_j, X'_j))^2 + B \frac{a^2 (r \wedge l) l}{n} \log \left( \frac{Bnm}{(r \wedge l) l} \right) \right], \]

\[ B > 0 \text{ is a constant} \]

\[ \hat{S} := \hat{S}_{\hat{r}, \hat{l}, a} \]
Recall that

\[ S_{r, \rho} := \{ S \in S_V : \text{rank}(S) \leq r, \| W^{1/2} S \|_{L_2(\Pi^2)} \leq \rho \} \]

\[ P_{r, \rho, a} := \left\{ P : (X, X', Y) \sim P, X, X' \text{ independent } \sim \Pi, |Y| \leq a, \right. \]
\[ \left. \mathbb{E}(Y|X, X') = S_*(X, X') = S_P(X, X'), S_P \in S_{r, \rho} \right\} \]

**Theorem**

For all \( P \in P_{r, \rho, a} \) and for all \( t > 0 \) with probability at least \( 1 - e^{-t} \)

\[
\| \hat{S} - S_P \|_{L_2(\Pi^2)}^2 \leq C \left( \min_{1 \leq l \leq m} \left[ \frac{a^2 (r \wedge l) l}{n} \log \left( \frac{Bnm}{(r \wedge l)l} \right) \lor \frac{\rho^2}{\lambda_{l+1}} \right] \lor \frac{a^2 (t + \log m)}{n} \right).
\]
Upper Bound: Example

- \( \lambda_k \asymp k^{2\beta} \) for some \( \beta > 1/2 \).

\[
\| \hat{S} - S_P \|^2_{L_2(\Pi^2)} \leq C \left[ \left( \left( \frac{a^2 \rho^{1/\beta}}{n} \log \frac{Bnm}{r} \right)^{2\beta/(2\beta+1)} \wedge \left( \frac{a^{2\rho^{2/\beta}} \log(Bnm)}{n} \right)^{\beta/(\beta+1)} \wedge \frac{a^2 r m \log(Bnm)}{n} \right) \vee \frac{a^2 (t + \log m)}{n} \right].
\]

- Compare with the lower bound:

\[
\left( \left( \frac{a^2 \rho^{1/\beta}}{n} \right)^{2\beta/(2\beta+1)} \wedge \left( \frac{a^{2\rho^{2/\beta}}}{n} \right)^{\beta/(\beta+1)} \wedge \frac{a^2}{n} \right) \vee \frac{a^2}{n}.
\]
Least Squares Estimators with Double Penalization: Nuclear Norm and Discrete Sobolev Norm

\[ \hat{S}_{\epsilon, \bar{\epsilon}} := \arg\min_{S \in S^a_V} \left\{ \sum_{j=1}^{n} (Y_j - S(X_j, X'_j))^2 + \epsilon \| S \|_1 + \bar{\epsilon} \| W^{1/2} S \|_{L_2(\Sigma^2)}^2 \right\} \]

- \( \| S \|_1 := \text{tr}(\| S \|) \), \( S := \sqrt{S^2} \) nuclear norm
- \( \epsilon, \bar{\epsilon} \) regularization parameters
- If \( \bar{\epsilon} = 0 \), then \( \hat{S}_{\epsilon, 0} \) is a matrix LASSO estimator
Choice of $\bar{\epsilon}$ : aggregation

- Divide the sample $(X_1, X'_1, Y_1), \ldots, (X_n, X'_n, Y_n)$ into two parts, $(X_j, X'_j, Y_j), j = 1, \ldots, n'$ and $(X_{n'+j}, X'_{n'+j}, Y_{n'+j}), j = 1, \ldots, n - n'$, where $n' := [n/2] + 1$.
- $\hat{S}_l := \hat{S}_{\epsilon, \bar{\epsilon}}, \bar{\epsilon}_l := \lambda_l^{-1}, l = 1, \ldots, m$ is based only on the first $n'$ observations.

$$\hat{l} := \arg\min_{l=1, \ldots, m} \frac{1}{n - n'} \sum_{j=1}^{n-n'} (Y_{n'+j} - \hat{S}_l(X_{n'+j}, X'_{n'+j}))^2.$$

- $\hat{S} := \hat{S}_l$. 
Spectral Function: $F(\lambda) := \sum_{j=1}^{m} I(\lambda_j \leq \lambda)$

$S_* = \sum_{k=1}^{r} \mu_k (\psi_k \otimes \psi_k)$

$L := l.s. (\psi_1, \ldots, \psi_r), \dim(L) = \text{rank}(S_*) = r$

$L$ the orthogonal projector onto $L$

$E(\lambda) := \sum_{\lambda_j \leq \lambda} (\phi_j \otimes \phi_j)$

Coherence Function

$\varphi(S_*; \lambda) := \langle P_L, E(\lambda) \rangle = \sum_{\lambda_j \leq \lambda} \|P_L \phi_j\|^2, \lambda \geq 0$

$\varphi(S_*; \lambda) \leq F(\lambda)$

A Low Coherence Assumption: for some $\nu(S_*) \geq 1$,

$\varphi(S_*; \lambda) \leq \frac{\nu(S_*) r F(\lambda)}{m}, \lambda \geq 0$. 

Vladimir Koltchinskii (Georgia Tech)
Bounds under Low Coherence Assumption

Given $t > 0$, let $t_{n,m} := t + 3 \log \left( 2 \log_2 n + \frac{1}{2} \log_2 \frac{\lambda m}{\lambda_1} + 2 \right)$.

Suppose $\frac{m(t+\log(2m))}{n} \leq 1$ and $\varepsilon = Da \sqrt{\frac{t+\log(2m)}{nm}}$.

Theorem

With probability at least $1 - e^{-t}$,

$$\| \hat{S} - S_* \|^2_{L_2(\Pi^2)} \leq$$

$$C \min_{1 \leq l \leq m} \left( \frac{\nu(S_*) \text{rank}(S_*) F(\lambda_l)(t + \log(2m))}{n} + \frac{\| W^{1/2} S_* \|^2_{L_2(\Pi^2)}}{\lambda_l} \right)$$

$$+ C \frac{a^2(\log(m + 1) + t_{n,m})}{n}.$$
\( \lambda_k \asymp k^{2\beta} \) for some \( \beta > 1/2 \)

Then

\[
\| \hat{S} - S^* \|_{L_2(\mathbb{P}^2)}^2 \leq C \left( \left( \left( \frac{\nu a^2 \rho^{1/\beta} r \log(2m)}{n} \right)^{2\beta/(2\beta+1)} \right) \wedge \left( \frac{a^2 \rho^{2/\beta} \log(2m)}{n} \right)^{\beta/(\beta+1)} \right) \\
\wedge \frac{a^2 rm}{n} \right) \vee \frac{a^2 (\log(m+1) + t_{n,m})}{n}
\]

Compare with the lower bound:

\[
\left( \left( \frac{a^2 \rho^{1/\beta} r}{n} \right)^{2\beta/(2\beta+1)} \right) \wedge \left( \frac{a^2 \rho^{2/\beta}}{n} \right)^{\beta/(\beta+1)} \wedge \frac{a^2 rm}{n} \right) \vee \frac{a^2}{n}.
\]