Efficient Globally Convergent Stochastic Optimization for Canonical Correlation Analysis

Weiran Wang[‡⋆] Jialei Wang[†⋆] Dan Garber[‡] Nathan Srebro[‡]

[‡]Toyota Technological Institute at Chicago [†]University of Chicago

Canonical correlation analysis (CCA)

Given data: $\mathbf{X} \in \mathbb{R}^{d_x \times N}, \mathbf{Y} \in \mathbb{R}^{d_y \times N}$, CCA solves

$$\max_{\mathbf{u},\mathbf{v}} \quad \mathbf{u}^\top \mathbf{\Sigma}_{xy} \mathbf{v}, \qquad \text{s.t.} \quad \mathbf{u}^\top \mathbf{\Sigma}_{xx} \mathbf{u} = \mathbf{v}^\top \mathbf{\Sigma}_{yy} \mathbf{v} = 1,$$

where $\mathbf{\Sigma}_{xy} = \frac{1}{N}\mathbf{X}\mathbf{Y}^\top$, $\mathbf{\Sigma}_{xx} = \frac{1}{N}\mathbf{X}\mathbf{X}^\top + \gamma_x \mathbf{I}$, $\mathbf{\Sigma}_{yy} = \frac{1}{N}\mathbf{Y}\mathbf{Y}^\top + \gamma_y \mathbf{I}$, $(\gamma_x, \gamma_y) \geq 0$.

Goal: efficiently solve CCA problem when N, d are large.

Our approach: reduce CCA to a sequence of least square problems, and solve each problem to sufficient accuracy using fast solver.

Closed-form Solution

Form matrix $\mathbf{T}$:

$$\mathbf{T} = \mathbf{\Sigma}_{xx}^{-1/2} \mathbf{\Sigma}_{xy} \mathbf{\Sigma}_{yy}^{-1/2} \in \mathbb{R}^{d_x \times d_y}$$

Compute SVD:

$$\mathbf{T} = [\boldsymbol{\phi}, \mathbf{a}_2, \ldots, \mathbf{a}_r] \begin{bmatrix} \rho_1 & & & \\ & \rho_2 & & \\ & & \cdots & \\ & & & \rho_r \end{bmatrix} [\boldsymbol{\psi}, \mathbf{b}_2, \ldots, \mathbf{b}_r]^\top$$

Obtain $(\mathbf{u}^*, \mathbf{v}^*) = (\mathbf{\Sigma}_{xx}^{-\frac{1}{2}} \boldsymbol{\phi}, \mathbf{\Sigma}_{yy}^{-\frac{1}{2}} \boldsymbol{\psi})$

Forming $\mathbf{T}$ costs $\mathcal{O}(Nd^2 + d^3)$

Alternating least squares (ALS) [Golub and Zha 1995]

The ALS Algorithm: For $t = 1, \ldots,$

$$\tilde{\mathbf{u}}_t \leftarrow \mathbf{\Sigma}_{xx}^{-1} \mathbf{\Sigma}_{xy} \mathbf{v}_{t-1}, \qquad \tilde{\mathbf{v}}_t \leftarrow \mathbf{\Sigma}_{yy}^{-1} \mathbf{\Sigma}_{xy}^\top \mathbf{u}_{t-1}$$
$$\mathbf{u}_t \leftarrow \tilde{\mathbf{u}}_t / \sqrt{\tilde{\mathbf{u}}_t^\top \mathbf{\Sigma}_{xx} \tilde{\mathbf{u}}_t}, \qquad \mathbf{v}_t \leftarrow \tilde{\mathbf{v}}_t / \sqrt{\tilde{\mathbf{v}}_t^\top \mathbf{\Sigma}_{yy} \tilde{\mathbf{v}}_t}$$

Alternative view in $\boldsymbol{\phi}_t = \mathbf{\Sigma}_{xx}^{\frac{1}{2}} \mathbf{u}_t$ and $\boldsymbol{\psi}_t = \mathbf{\Sigma}_{yy}^{\frac{1}{2}} \mathbf{v}_t$:

$$\tilde{\boldsymbol{\phi}}_t \leftarrow \mathbf{\Sigma}_{xx}^{-\frac{1}{2}} \mathbf{\Sigma}_{xy} \mathbf{\Sigma}_{yy}^{-\frac{1}{2}} \boldsymbol{\psi}_{t-1}, \qquad \tilde{\boldsymbol{\psi}}_t \leftarrow \mathbf{\Sigma}_{yy}^{-\frac{1}{2}} \mathbf{\Sigma}_{xy}^\top \mathbf{\Sigma}_{xx}^{-\frac{1}{2}} \boldsymbol{\phi}_{t-1}$$
$$\boldsymbol{\phi}_t \leftarrow \tilde{\boldsymbol{\phi}}_t / \left\| \tilde{\boldsymbol{\phi}}_t \right\|, \qquad \boldsymbol{\psi}_t \leftarrow \tilde{\boldsymbol{\psi}}_t / \left\| \tilde{\boldsymbol{\psi}}_t \right\|$$

ALS w.r.t $(\mathbf{u}, \mathbf{v})$ $\Longleftrightarrow$ Power iterations w.r.t $(\boldsymbol{\phi}, \boldsymbol{\psi})$

$$\begin{bmatrix} \tilde{\boldsymbol{\phi}}_t \\ \tilde{\boldsymbol{\psi}}_t \end{bmatrix} \leftarrow \begin{bmatrix} \mathbf{0} & \mathbf{T} \\ \mathbf{T}^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\phi}_{t-1} \\ \boldsymbol{\psi}_{t-1} \end{bmatrix}, \qquad \begin{bmatrix} \boldsymbol{\phi}_t \\ \boldsymbol{\psi}_t \end{bmatrix} \leftarrow \begin{bmatrix} \tilde{\boldsymbol{\phi}}_t / ||\tilde{\boldsymbol{\phi}}_t|| \\ \tilde{\boldsymbol{\psi}}_t / ||\tilde{\boldsymbol{\psi}}_t|| \end{bmatrix}$$

Every two steps of ALS perform one step power iteration on $\mathbf{T}\mathbf{T}^\top$ for the $\boldsymbol{\phi}$-sequence, and $\mathbf{T}^\top\mathbf{T}$ for the $\boldsymbol{\psi}$-sequence.

Theorem: After $T = \mathcal{O}\left(\frac{\rho_1^2}{\rho_1^2 - \rho_2^2} \log\left(\frac{1}{\eta}\right)\right)$ iterations, we have for all $t \geq T$:

$$\min\left((\mathbf{u}_t^\top \mathbf{\Sigma}_{xx} \mathbf{u}^*)^2,\ (\mathbf{v}_t^\top \mathbf{\Sigma}_{yy} \mathbf{v}^*)^2\right) \geq 1 - \eta, \qquad \mathbf{u}_t^\top \mathbf{\Sigma}_{xy} \mathbf{v}_t \geq \rho_1 (1 - 2\eta).$$

Inexact ALS: inexact power iteration on $\begin{bmatrix} \mathbf{0} & \mathbf{T} \\ \mathbf{T}^\top & \mathbf{0} \end{bmatrix}$

Non-zero eigenvalues: $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_r \geq -\rho_r \geq \cdots \geq -\rho_1$

with eigenvectors

$$\begin{bmatrix} \boldsymbol{\phi} \\ \boldsymbol{\psi} \end{bmatrix}, \begin{bmatrix} \mathbf{a}_2 \\ \mathbf{b}_2 \end{bmatrix}, \ldots, \begin{bmatrix} \mathbf{a}_r \\ \mathbf{b}_r \end{bmatrix}, \begin{bmatrix} \mathbf{a}_r \\ -\mathbf{b}_r \end{bmatrix}, \ldots, \begin{bmatrix} \mathbf{a}_2 \\ -\mathbf{b}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\phi} \\ -\boldsymbol{\psi} \end{bmatrix}$$

Relative eigengap: $\frac{\rho_1^2}{\rho_1^2 - \rho_2^2}$

Algorithm: For $t = 1, \ldots, T$

– Approximately solve $\min_{\mathbf{u}} \ f_t(\mathbf{u}) := \frac{1}{N} \sum_{i=1}^N \frac{1}{2} (\mathbf{u}^\top \mathbf{x}_i - \mathbf{v}_{t-1}^\top \mathbf{y}_i)^2 + \frac{\gamma_x}{2} \|\mathbf{u}\|^2$ s.t. $f_t(\tilde{\mathbf{u}}_t) \leq \min_{\mathbf{u}} \ f_t(\mathbf{u}) + \epsilon$

– Approximately solve $\min_{\mathbf{v}} \ g_t(\mathbf{v}) := \frac{1}{N} \sum_{i=1}^N \frac{1}{2} (\mathbf{v}^\top \mathbf{y}_i - \mathbf{u}_{t-1}^\top \mathbf{x}_i)^2 + \frac{\gamma_y}{2} \|\mathbf{v}\|^2$ s.t. $g_t(\tilde{\mathbf{v}}_t) \leq \min_{\mathbf{v}} \ g_t(\mathbf{v}) + \epsilon$

– Normalization: $\mathbf{u}_t \leftarrow \tilde{\mathbf{u}}_t / \sqrt{\tilde{\mathbf{u}}_t^\top \mathbf{\Sigma}_{xx} \tilde{\mathbf{u}}_t}, \qquad \mathbf{v}_t \leftarrow \tilde{\mathbf{v}}_t / \sqrt{\tilde{\mathbf{v}}_t^\top \mathbf{\Sigma}_{yy} \tilde{\mathbf{v}}_t}$

Subproblem conditioning: κ' for NAG, κ for SVRG

Theorem: After $T = \mathcal{O}\left(\frac{\rho_1^2}{\rho_1^2 - \rho_2^2} \log\left(\frac{2}{\eta}\right)\right)$ iterations with $\epsilon(T) = \frac{\eta^2 \rho_r^2}{128} \left(\frac{(2\rho_1/\rho_r) - 1}{(2\rho_1/\rho_r)^T - 1}\right)^2$, we have $\mathbf{u}_T^\top \mathbf{\Sigma}_{xx} \mathbf{u}_T = \mathbf{v}_T^\top \mathbf{\Sigma}_{yy} \mathbf{v}_T = 1$ and $\min\left((\mathbf{u}_T^\top \mathbf{\Sigma}_{xx} \mathbf{u}^*)^2,\ (\mathbf{v}_T^\top \mathbf{\Sigma}_{yy} \mathbf{v}^*)^2\right) \geq 1 - \eta$.

Shift-and-invert: inexact power iteration on $\begin{bmatrix} \lambda \mathbf{I} & -\mathbf{T} \\ -\mathbf{T}^\top & \lambda \mathbf{I} \end{bmatrix}^{-1}$

Eigenvalues: $\frac{1}{\lambda - \rho_1} \geq \cdots \geq \frac{1}{\lambda - \rho_r} \geq \cdots \geq \frac{1}{\lambda + \rho_r} \geq \cdots \geq \frac{1}{\lambda + \rho_1}$

with eigenvectors $\frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{\phi} \\ \boldsymbol{\psi} \end{bmatrix}, \ldots, \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{a}_r \\ \mathbf{b}_r \end{bmatrix}, \ldots, \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{a}_r \\ -\mathbf{b}_r \end{bmatrix}, \ldots, \frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{\phi} \\ -\boldsymbol{\psi} \end{bmatrix}$

For $\lambda = \rho_1 + c(\rho_1 - \rho_2)$ where $c = \mathcal{O}(1)$, relative eigengap is $1 + c$, a constant

Algorithm: For $t = 1, \ldots, T$

– Approximately solve $\begin{bmatrix} \tilde{\boldsymbol{\phi}}_t \\ \tilde{\boldsymbol{\psi}}_t \end{bmatrix} \leftarrow \begin{bmatrix} \lambda \mathbf{I} & -\mathbf{T} \\ -\mathbf{T}^\top & \lambda \mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\phi}_{t-1} \\ \boldsymbol{\psi}_{t-1} \end{bmatrix}$ or

$(\tilde{\mathbf{u}}_t, \tilde{\mathbf{v}}_t) \approx \arg\min_{\mathbf{u},\mathbf{v}} \frac{1}{2} \begin{bmatrix} \mathbf{u}^\top & \mathbf{v}^\top \end{bmatrix} \begin{bmatrix} \lambda \mathbf{\Sigma}_{xx} & -\mathbf{\Sigma}_{xy} \\ -\mathbf{\Sigma}_{xy}^\top & \lambda \mathbf{\Sigma}_{yy} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} - \mathbf{u}^\top \mathbf{\Sigma}_{xx} \mathbf{u}_{t-1} - \mathbf{v}^\top \mathbf{\Sigma}_{yy} \mathbf{v}_{t-1}$

– Intermediate normalization: $\begin{bmatrix} \mathbf{u}_t \\ \mathbf{v}_t \end{bmatrix} \leftarrow \sqrt{2} \begin{bmatrix} \tilde{\mathbf{u}}_t \\ \tilde{\mathbf{v}}_t \end{bmatrix} \Big/ \sqrt{\tilde{\mathbf{u}}_t^\top \mathbf{\Sigma}_{xx} \tilde{\mathbf{u}}_t + \tilde{\mathbf{v}}_t^\top \mathbf{\Sigma}_{yy} \tilde{\mathbf{v}}_t}$

Final normalization: $\hat{\mathbf{u}} \leftarrow \mathbf{u}_T / \sqrt{\mathbf{u}_T^\top \mathbf{\Sigma}_{xx} \mathbf{u}_T}, \qquad \hat{\mathbf{v}} \leftarrow \mathbf{v}_T / \sqrt{\mathbf{v}_T^\top \mathbf{\Sigma}_{yy} \mathbf{v}_T}$

Subproblem conditioning: $\mathcal{O}\left(\frac{\tilde{\kappa}'}{\rho_1 - \rho_2}\right)$ for NAG, $\mathcal{O}\left(\frac{\tilde{\kappa}}{\rho_1 - \rho_2}\right)$ for SVRG

Theorem: After $T = \mathcal{O}\left(\log\left(\frac{2}{\eta}\right)\right)$ iterations, we have $\hat{\mathbf{u}}^\top \mathbf{\Sigma}_{xx} \hat{\mathbf{u}} = \hat{\mathbf{v}}^\top \mathbf{\Sigma}_{yy} \hat{\mathbf{v}} = 1$, and $\min\left((\hat{\mathbf{u}}^\top \mathbf{\Sigma}_{xx} \mathbf{u}^*)^2,\ (\hat{\mathbf{v}}^\top \mathbf{\Sigma}_{yy} \mathbf{v}^*)^2\right) \geq 1 - \eta$.

There exists an efficient algorithm for locating λ.

Summary of running time

	Alternating least squares	Shift-and-invert ($\lambda > \rho_1$)
Power iterations	$\begin{bmatrix} \mathbf{0} & \mathbf{T} \\ \mathbf{T}^\top & \mathbf{0} \end{bmatrix}$	$\begin{bmatrix} \lambda \mathbf{I} & -\mathbf{T} \\ -\mathbf{T}^\top & \lambda \mathbf{I} \end{bmatrix}^{-1}$
Least squares	$\begin{bmatrix} \mathbf{\Sigma}_{xx} & \\ & \mathbf{\Sigma}_{yy} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{\Sigma}_{xy} \mathbf{v}_{t-1} \\ \mathbf{\Sigma}_{xy}^\top \mathbf{u}_{t-1} \end{bmatrix}$	$\begin{bmatrix} \lambda \mathbf{\Sigma}_{xx} & -\mathbf{\Sigma}_{xy} \\ -\mathbf{\Sigma}_{xy}^\top & \lambda \mathbf{\Sigma}_{yy} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{\Sigma}_{xx} \mathbf{u}_{t-1} \\ \mathbf{\Sigma}_{yy} \mathbf{v}_{t-1} \end{bmatrix}$
NAG	$d N \sqrt{\kappa'} \left(\frac{\rho_1^2}{\rho_1^2 - \rho_2^2}\right)^2 \cdot \log^2\left(\frac{1}{\eta}\right)$	$d N \sqrt{\tilde{\kappa}'} \sqrt{\frac{1}{\rho_1 - \rho_2}} \cdot \log^2\left(\frac{1}{\eta}\right)$
SVRG	$d\,(N + \kappa) \left(\frac{\rho_1^2}{\rho_1^2 - \rho_2^2}\right)^2 \cdot \log^2\left(\frac{1}{\eta}\right)$	$d \left(N + \left(\tilde{\kappa} \frac{1}{\rho_1 - \rho_2}\right)^2\right) \cdot \log^2\left(\frac{1}{\eta}\right)$
ASVRG	$d \sqrt{N} \sqrt{\kappa} \left(\frac{\rho_1^2}{\rho_1^2 - \rho_2^2}\right)^2 \cdot \log^2\left(\frac{1}{\eta}\right)$	$d N^{\frac{3}{4}} \sqrt{\tilde{\kappa}} \sqrt{\frac{1}{\rho_1 - \rho_2}} \cdot \log^2\left(\frac{1}{\eta}\right)$

Runtime (up to `polylog` factors) to find $(\mathbf{u}, \mathbf{v})$ with $\min\left((\mathbf{u}^\top \mathbf{\Sigma}_{xx} \mathbf{u}^*)^2,\ (\mathbf{v}^\top \mathbf{\Sigma}_{yy} \mathbf{v}^*)^2\right) \geq 1 - \eta$

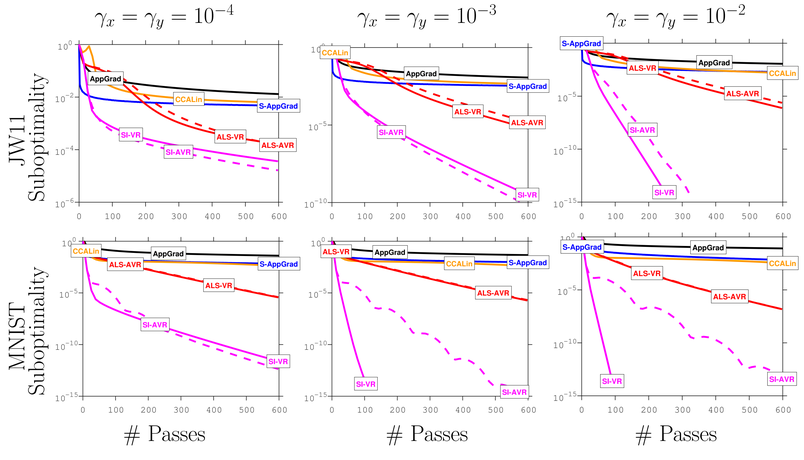
$$\kappa' := \max\left(\frac{\sigma_{\max}(\mathbf{\Sigma}_{xx})}{\sigma_{\min}(\mathbf{\Sigma}_{xx})}, \frac{\sigma_{\max}(\mathbf{\Sigma}_{yy})}{\sigma_{\min}(\mathbf{\Sigma}_{yy})}\right) \leq \kappa := \max\left(\frac{\max_i \|\mathbf{x}_i\|^2}{\sigma_{\min}(\mathbf{\Sigma}_{xx})}, \frac{\max_i \|\mathbf{y}_i\|^2}{\sigma_{\min}(\mathbf{\Sigma}_{yy})}\right)$$
$$\wedge \qquad\qquad \wedge$$
$$\tilde{\kappa}' := \frac{\max\left(\sigma_{\max}(\mathbf{\Sigma}_{xx}),\ \sigma_{\max}(\mathbf{\Sigma}_{yy})\right)}{\min\left(\sigma_{\min}(\mathbf{\Sigma}_{xx}),\ \sigma_{\min}(\mathbf{\Sigma}_{yy})\right)} \leq \tilde{\kappa} := \frac{\max_i \max\left(\|\mathbf{x}_i\|^2,\ \|\mathbf{y}_i\|^2\right)}{\min\left(\sigma_{\min}(\mathbf{\Sigma}_{xx}),\ \sigma_{\min}(\mathbf{\Sigma}_{yy})\right)}$$

Related Work

- **[Ma, Lu, and Foster 2015]**, `AppGrad`: one full gradient step on ALS subproblems, can only guarantee local convergence
- Parallel work: **[Ge et al, 2016]**, `CCALin`: ALS with subproblems solved by NAG, time complexity $\mathcal{O}\left(d N \sqrt{\kappa'} \frac{\rho_1^2}{\rho_1^2 - \rho_2^2} \cdot \log\left(\frac{1}{\eta}\right)\right)$

Experiments

Datasets	Description	d_x	d_y	N
JW11	Acoustic and articulation measurements	273	112	30,000
MNIST	Left and right halves of images	392	392	60,000



NIPS 2016, Barcelona, December 2016