Stochastic Approximation for Canonical Correlation Analysis

Raman Arora, Teodor V. Marinov, Poorya Mianjy

Department of Computer Science, Johns Hopkins University, Baltimore

Presented by Liqun Chen

Nov 3rd, 2017
1 Introduction
   • Background
   • Challenges
   • Notations
   • Problem Definition

2 Methods
   • Matrix Stochastic Gradient
   • Matrix Exponentiated Gradient

3 Experiments
Outline

1 Introduction
   - Background
   - Challenges
   - Notations
   - Problem Definition

2 Methods
   - Matrix Stochastic Gradient
   - Matrix Exponentiated Gradient

3 Experiments
Canonical Correlation Analysis (CCA) is a statistical technique for finding linear components of two sets of random variables that are maximally correlated. CCA is often posed as a dimensionality reduction problem about a fixed dataset of \( n \) paired data points. In this paper, they take a stochastic optimization view of CCA. They pose CCA as the following stochastic optimization problem: given a pair of random vectors \((x, y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}\), with some (unknown) joint distribution \( D \), find matrices \( U \in \mathbb{R}^{d_x \times k} \) and \( V \in \mathbb{R}^{d_y \times k} \) that solve:

\[
\text{maximize } \mathbb{E}_{x,y}[x^T U V^T y] \quad \text{subject to} \quad U^T \mathbb{E}_x[x x^T] U = V^T \mathbb{E}_y[y y^T] V = I_k.
\]
the goal is to find a “good” enough subspace in terms of capturing correlation rather than a “true” subspace (Domain adaptation)

- the stochastic optimization view motivates stochastic approximation algorithms that can easily scale to very large datasets.
Challenges

- non-convex
- it is a unconstraint optimization problem, since there is no stochastic constraint on $U$ and $V$, e.g., for $k = 1$, we can consider the objective
  \[
  \rho(u^T x, v^T y) = \mathbb{E}_{x,y} [u^T x y^T v] / \left( \sqrt{\mathbb{E}_x [u^T x x^T u]} \sqrt{\mathbb{E}_y [v^T y y^T v]} \right).
  \]
- CCA problem given in (1) is non-learnable, i.e. there exists distributions on which it may incur small empirical error on the training sample but arbitrary large generalization error.
Notations

- For any matrix $X$, spectral norm, nuclear norm, and Frobenius norm are represented by $\|X\|_2$, $\|X\|_F$, and $\|X\|_*$ respectively.

- The standard inner-product between the two is given as $\langle X, Y \rangle = \text{Tr}(X^\top Y)$.

- Let singular value decomposition (SVD) of arbitrary $X$ be given as $X = U\Sigma V^\top$.

- Let auto-covariance matrices $C_{xx} = \mathbb{E}_x[xx^\top]$, $C_{yy} = \mathbb{E}_y[yy^\top]$, and cross-covariance matrix $C_{xy} = \mathbb{E}_{(x,y)}[xy^\top]$.

- Denote the regularized auto-covariance matrices by $C_x = C_{xx} + r_x I$ and $C_y = C_{yy} + r_y I$. 
Problem Definition

The regularized version of Problem 1 can be formulated as follows. With this notation, re-write the regularized CCA problem as follows.

\[
\max_{\tilde{U} \in \mathbb{R}^{d_x \times k}, \tilde{V} \in \mathbb{R}^{d_y \times k}} \text{Tr} \left( \tilde{U}^\top \mathbb{E}[x y^\top] \tilde{V} \right)
\]
subject to \( \tilde{U}^\top \mathbb{E}[x x^\top + r_x I] \tilde{U} = I, \ \tilde{V}^\top \mathbb{E}[y y^\top + r_y I] \tilde{V} = I \) \quad (2)

Since \( C_x^{1/2} \) and \( C_y^{1/2} \) are positive definite matrices one can make the simple change of variables \( U = C_x^{1/2} \tilde{U} \) and \( V = C_y^{1/2} \tilde{V} \) to get the equivalent problem

\[
\max_{U \in \mathbb{R}^{d_x \times k}, V \in \mathbb{R}^{d_y \times k}} \text{Tr} \left( U^\top \mathbb{E}[x x^\top + r_x I]^{-\frac{1}{2}} \mathbb{E}[x y^\top] \mathbb{E}[y y^\top + r_y I]^{-\frac{1}{2}} V \right)
\]
subject to \( U^\top U = I, \ V^\top V = I \) \quad (3)
Let $\Phi \in \mathbb{R}^{d_x \times k}$ and $\Psi \in \mathbb{R}^{d_y \times k}$ denote the top-$k$ left and right singular vectors of $C_x^{-\frac{1}{2}} C_{xy} C_y^{-\frac{1}{2}}$. Then, the optimum of Problem 2 is achieved at $\tilde{U} = C_x^{-\frac{1}{2}} \Phi$ and $\tilde{V} = C_y^{-\frac{1}{2}} \Psi$. An equivalent definition for the regularized empirical problem in the 1-dimensional setting is

$$\min_{u \in \mathbb{R}^{d_x}, v \in \mathbb{R}^{d_y}} \frac{1}{2t} \|u^\top X - v^\top Y\|_2^2 + r_x \|u\|_2^2 + r_y \|v\|_2^2$$

subject to $u^\top C_{x,t} u = I$, $v^\top C_{y,t} v = I$
Outline

1 Introduction
   - Background
   - Challenges
   - Notations
   - Problem Definition

2 Methods
   - Matrix Stochastic Gradient
   - Matrix Exponentiated Gradient

3 Experiments
Proposed Method (1)

Problem 3 is not a convex optimization problem – not only is the objective non-convex, the constraint set of orthogonal matrices is also non-convex. Re-parametrize Problem 3 via the following variable substitute: $M = UV^\top$. Furthermore, they take the convex hull of the constraint:

$$\max_{M \in \mathbb{R}^{dx \times dy}} \quad \langle M, C_{x}^{-\frac{1}{2}} C_{xy} C_{y}^{-\frac{1}{2}} \rangle$$

subject to $\|M\|_2 \leq 1, \|M\|_* \leq k$. 

(5)
Proposed Method (2)

- the gradient of the CCA objective w.r.t. $M$ is $g := C^{-\frac{1}{2}}_x C_{xy} C^{-\frac{1}{2}}_y$, however, the marginal distribution for $x$ and $y$ is unknown.

- use an “inexact” first order oracle

\[
\partial_t := W_{x,t} x_t y_t W_{y,t} \approx g_t,
\]

- the choice of $W_{x,t} := C^{-\frac{1}{2}}_{x,t}$ and $W_{y,t} := C^{-\frac{1}{2}}_{y,t}$ being the empirical estimates of the whitening transform matrices.

- denote the true whitening transforms by $W_x := \mathbb{E}[(xx^\top + r_x I)^{-\frac{1}{2}}]$ and $W_y := \mathbb{E}[(yy^\top + r_y I)^{-\frac{1}{2}}]$. 
Proposed Method (3)

Design stochastic approximation algorithms for solving Problem 5; In each update, minimize the instantaneous loss measured on a single data point while trying to stay close to the previous iterate.

- In each iteration minimize the sum $\Delta(M_t, M) + \eta \mathbb{E}_{\mathcal{D}} [\ell(M, g_t)]$, over all feasible $M$
- $\Delta(\cdot, \cdot)$ is a divergence function between the successive iterates
- $\ell$ is the instantaneous loss
- $\eta$ is the learning-rate parameter controlling the tradeoff between loss and divergence
- the divergence function is defined in terms of a potential function
Matrix Stochastic Gradient

stochastic mirror descent with the choice of potential function as the Frobenius norm:

\[ M_t = \mathcal{P}_F(M_{t-1} + \eta_t \partial_t) \]

**Algorithm 1** Matrix Stochastic Gradient (MSG)

**Input:** \( M_0, \{(x_t, y_t)\}_{t=1}^T, \eta \)

**Output:** \( \bar{M} \)

1. for \( t = 1, \ldots, T \) do
2. \[ C_{x,t} \leftarrow \frac{t-1}{t} C_{x,t-1} + \frac{1}{t} x_t x_t^T, \quad W_{x,t} \leftarrow C_{x,t}^{-\frac{1}{2}} \]
3. \[ C_{y,t} \leftarrow \frac{t-1}{t} C_{y,t-1} + \frac{1}{t} y_t y_t^T, \quad W_{y,t} \leftarrow C_{y,t}^{-\frac{1}{2}} \]
4. \[ \partial_t \leftarrow W_{x,t} x_t y_t W_{y,t} \]
5. \[ M_t \leftarrow \mathcal{P}_F(M_t + \eta \partial_t) \]
6. end for
7. \[ \bar{M} = \frac{1}{T} \sum_{t=1}^T M_{t-1} \]
8. \( \tilde{M} = \text{rounding}(\bar{M}) \)
Convergence of Matrix Stochastic Gradient

Lemma

Assume that for all \((x_t, y_t) \sim \mathcal{D}\) we have \(\max\{\|x_t\|^2, \|y_t\|^2\} \leq B\). There exists a constant \(\kappa\) independent of \(t\) such that for all \(t\) the following holds in expectation:

\[
\mathbb{E}_\mathcal{D}[\|E_t\|_2] \leq \frac{\kappa}{\sqrt{t}}
\]

Theorem

After \(T\) iterations of MSG with step size \(\eta = \frac{2\sqrt{k}}{G\sqrt{T}}\), and starting at \(M^{(1)} = 0\),

\[
\mathbb{E}[\langle M_* - \tilde{M}, C^{-\frac{1}{2}} x C^{-\frac{1}{2}} y \rangle] \leq \frac{2\sqrt{kG} + 2k\kappa}{\sqrt{T}}
\]

where \(\kappa\) is a universal constant given by lemma 1, \(M_*\) is the optimum of (5), expectation is with respect to the i.i.d. samples and rounding, and \(G\) is given by Lemma in Appendix such that \(\|\partial_t\|_F \leq G\) for all the iterates \(t = 1, \cdots, T\).
the set is that of $d_k$-dimensional (paired) subspaces and the multiplicative algorithm is an instance of matrix exponentiated gradient (MEG) update. Let $d := d_x + d_y$.

Consider the self-adjoint dilation of the matrix $E_D [g_t] C := E_D \begin{pmatrix} 0 & g_t \\ g_t^T & 0 \end{pmatrix}$. Assume that $E_D [g_t]$ has no repeated singular values and its SVD is given by $E_D [g_t] = U \Sigma V^T$ then the eigen-decomposition of $C$ is given by

$$C = \frac{1}{2} \begin{pmatrix} U & U \\ V & -V \end{pmatrix} \begin{pmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{pmatrix} \begin{pmatrix} U & U \\ V & -V \end{pmatrix}.$$  

Then re-formulate regularized CCA as

$$\max_{W \in \mathbb{R}^{d \times k}} \quad \text{Tr} \left( WW^T C \right)$$

subject to $W^T W = I$.

(7)
Problem 7 is not a convex optimization, but it admits the following convex relaxation by setting \( M = \frac{1}{k} WW^\top \):

\[
\begin{align*}
\max_{M \in \mathbb{R}^{d \times d}} & \quad \text{Tr} (MC) \\
\text{subject to} & \quad M \succeq 0, \|M\|_2 \leq \frac{1}{k}, \text{Tr} (M) = 1
\end{align*}
\]  

(8)

Update \( M_{t-1} \) by solving

\[
\begin{align*}
M_t & := \arg\min_{M \in \mathbb{R}^{d \times d}} \quad \Delta (M, M_{t-1}) - \eta \text{Tr} (MC_t) \\
\text{subject to} & \quad M \succeq 0, \|M\|_2 \leq \frac{1}{k}, \text{Tr} (M) = 1
\end{align*}
\]  

(9)

where \( C_t \) is the self-adjoint dilation of \( g_t \) and \( \Delta (x, y) \) is the quantum relative entropy between \( x \) and \( y \). Setting the Lagrangian to 0 and solving:

\[
\hat{M}_t = \frac{\exp (\log (M_{t-1}) + \eta C_t)}{\text{Tr} (\exp (\log (M_{t-1}) + \eta C_t))}, \quad M_t = \mathcal{P} \left( \hat{M}_t \right)
\]  

(10)

where \( \mathcal{P} \) denotes the projection onto the convex set of constraints in (9). In practice they replace the self-adjoint dilation of \( g_t \) by self-adjoint dilation of \( \partial_t \) which is denoted by \( \tilde{C}_t \).
Algorithm 2 MEG for CCA

Input: $M_0$, $\{ (x_t, y_t) \}_{t=1}^T$, $\eta$

Output: $M$

for $t = 1$ to $T$ do

$C_{x,t} \leftarrow \frac{t-1}{t} C_{x,t-1} + \frac{1}{t} x_t x_t^\top$, $W_{x,t} \leftarrow C_{x,t}^{-\frac{1}{2}}$

$C_{y,t} \leftarrow \frac{t-1}{t} C_{y,t-1} + \frac{1}{t} y_t y_t^\top$, $W_{y,t} \leftarrow C_{y,t}^{-\frac{1}{2}}$

$\tilde{C}_t \leftarrow \begin{pmatrix} 0 & \partial_t \\ \partial_t^\top & 0 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} W_{x,t} x_t \\ W_{y,t} y_t \end{pmatrix} \begin{pmatrix} W_{x,t} x_t \\ W_{y,t} y_t \end{pmatrix}^\top - \frac{1}{2} \begin{pmatrix} W_{x,t} x_t \\ -W_{y,t} y_t \end{pmatrix} \begin{pmatrix} W_{x,t} x_t \\ -W_{y,t} y_t \end{pmatrix}^\top$

$\hat{M}_t \leftarrow \frac{\exp(\log(M_{t-1}) + \eta C_t)}{\text{Tr}(\exp(\log(M_{t-1}) + \eta C_t))}$

$M_t \leftarrow \mathcal{D} \left( \hat{M}_t \right)$

end for

$\tilde{M} = \frac{1}{T} \sum_{t=1}^T M_{t-1}$

$\hat{M} = \text{rounding} \left( \tilde{M} \right)$
Outline

1 Introduction
   • Background
   • Challenges
   • Notations
   • Problem Definition

2 Methods
   • Matrix Stochastic Gradient
   • Matrix Exponentiated Gradient

3 Experiments
Figure 1: Comparisons of CCA-Lin, CCA-ALS, MSG, and MEG for CCA optimization on the synthetic dataset, in terms of the objective value as a function of iteration (top) and as a function of CPU runtime (bottom).
Mediamill

Figure 2: Comparisons of CCA-Lin, CCA-ALS, MSG, and MEG for CCA optimization on the MediaMill dataset, in terms of the objective value as a function of iteration (top) and as a function of CPU runtime (bottom).