MCMC For Solving The Multitask Co-Learning Problem

Ricardo Menchaca-Mendez

University of California, Santa Cruz

June-2007
Statement of the Multitask Co-Learning Problem

Introduction to bounding mixing time of Markov chain
- Sampling uniformly from $n$ dimensional hypercube.
- Sampling matchings in a graph.

Towards bounding the mixing time of the expert chain.

Experimental results
Problem Definition

- In each trial $t = 1, 2, \ldots$ the master algorithm has to make a prediction $\hat{y} = (\hat{y}_1(t), \hat{y}_2(t), \ldots \hat{y}_K(t))$.
- The master algorithm is given access to a set of $N$ experts which make a prediction $h_{i,k}(t)$ on each task.
- At the end of the trial the algorithm receives the real output $y_k(t)$:
  - the master algorithm suffers a loss of $l(t) = \sum_k l(\hat{y}_k(t), y_k(t))$.
  - expert $i$ suffers a loss of $l_{i,k}(t) = l(h_{i,k}(t), y_k(t))$ per task.
Goal

- We want to perform better than a master algorithm predicting independently on each task.
- To measure the performance we consider the total loss of the master algorithm $L(T) = \sum_{t=1}^{T} l(t)$, relative to the cumulative loss of the best set of experts of size $m$:

$$L^*(T) = \min_{S} \sum_{k=1}^{K} \min_{i} \{L_{i,k}(T) | k \in S\}$$

- MCMC method achieve loss bound already

$$L(T) \leq aL^*(T) + b(K \log m + m \log \frac{N}{m})$$

but there is no bound on the mixing time of the chain.
Algorithm 1: Metropolis

Data: \( x \in \Omega \), Number of iterations \( R \).

Result: Sample from \( P(\cdot) \).

for \( i = 1 \ldots R \) do

  with probability \( \frac{1}{2} \) let \( y = x \); otherwise;
  let \( y' = \text{RandomLocalMove}(x) \);

  with probability \( \min \{ 1, \frac{P(y')}{P(x)} \} \) let \( y := y' \);
  set \( x := y \);

end

Consider a finite, irreducible, and ergodic markov chain with transition matrix \( P \). If there is distribution \( \pi(\cdot) \) satisfying

\[
\pi(x) P_{x \rightarrow y} = \pi(y) P_{y \rightarrow x} = Q(x, y)
\]

then \( \pi(\cdot) \) is the stationary distribution corresponding to \( P \).
Goal: Sampling from a specific distribution $P(x)$ with $x \in \Omega$. (Usually, $\Omega$ is exponentially big)

Applications

- Statistical Physics.
- Approximate Counting.
- Combinatorial Optimization.
Define a set $\Gamma$ of canonical paths. i.e. a path between every pair of states $x, y$ in the markov chain.

Show that the probability flow

$$\sum_{\gamma_{xy} \in e} \pi(x) \pi(y) |\gamma_{xy}| \frac{Q(e)}{\Phi}$$

across every edge $e$ is upper bounded by a polynomial in $n$ (the natural dimension of the problem)
Definition (Congestion)

\[ \rho = \rho(\Gamma) = \max_e \frac{1}{Q(e)} \sum_{\gamma_{xy} \ni e} \pi(x) \pi(y) |\gamma_{xy}| \]

Theorem (Sinclair 92)

Let \( M \) be a finite, reversible, ergodic Markov Chain with loop probabilities \( P(x, x) \geq 1/2 \) for all states \( x \). Let \( \Gamma \) be a set of canonical paths with maximum edge loading \( \rho(\Gamma) \). Then the mixing time of \( M \) satisfies

\[ \tau_x(\epsilon) \leq \rho(\ln \pi(x)^{-1} + \ln \epsilon^{-1}) \] (1)
Sampling uniformly from the boolean hypercube \([5]\)

**Input:** Number of iterations \(R\).

**Output:** \(x_R\), Sample from \(\pi(x)\).

Start with \(x_1 = \{0, 0, ..., 0\}\);

For\(\{i = 1..R\}\)\{

With probability \(\frac{1}{2}\) let \(x = x_i\), otherwise

select a component \(i\) u.a.r. and set

\[
x = (x_1, 0, ..., 1 - x_1,i, x_1,i+1, ..., x_1,n)
\]

\(x_{i+1} = x\)

\}
Define a set $\Gamma$ of canonical paths, i.e. a path between every pair of states $x, y$.

Show that the probability flow $\left( \sum_{\gamma_{xy} \in e} \pi(x) \pi(y) |\gamma_{xy}| \right) / Q(e)$ across every edge $e$ is upper bounded by a polynomial in $n$ (the dimension of the hypercube).

Plug it in Eq (1) and get a bound for the mixing time.
Let $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ be two arbitrarily states in $\Omega$. Define the canonical path $\gamma_{xy}$ as the path $x \sim w_1 \sim w_2 \sim ... \sim w_n = y$ where $w_i = (y_1, ..., y_i, x_{i+1}, ..., x_n)$.

**Example**

$x = (1, 0, 1, 1, 0), \; y = (0, 1, 1, 0, 1)$

$\gamma_{xy} \equiv x = (1, 0, 1, 1, 0) \sim (0, 0, 1, 1, 0) \sim (0, 1, 1, 1, 0) \sim (0, 1, 1, 1, 0) \\
\sim (0, 1, 1, 0, 0) \sim (0, 0, 1, 0, 1) = y$
Computing Congestion (Counting)

Pick an arbitrarily edge

\[ e = (w, w') = ((w_1, \ldots, w_i, \ldots, w_n), (w_1, \ldots, w'_i, \ldots, w_n)) \]

How many canonical paths use edge \( e \)?

\[ e = (w, w') = ((w_1, \ldots w_{i-1}, w_i = x_i, \ldots, w_n = x_n), (w_1 = y_1, \ldots, w'_i = y_i, w_{i+1}, \ldots, w_n)) \]

There are \( i - 1 \) different choices for \( x \); and there are \( n - i \) choices for \( y \). Thus, there are \( 2^{i-1}2^{n-i} = 2^{n-1} \) paths using \( e \)

\[
\frac{1}{Q(e)} \sum_{\gamma_{xy} \exists e} \pi(x)\pi(y) |\gamma_{xy}| \leq 2n2^n \cdot \frac{n2^{n-1}}{2^n2^n} \leq n^2 \\
\tau_x(\epsilon) \leq n^2(n \ln 2 + \ln \epsilon^{-1})
\]
Define an injective function. Let $cp(e) = \{(x, y) | \gamma_{xy} \ni e\}$ be the set of all canonical paths that use edge $e$. Define $\eta_e : cp(e) \rightarrow \Omega$ such that $\pi(x)\pi(y) \leq f(n)\pi(w)\pi(cp(e))$ for $f(n) = O(n^k)$.

$$\eta_e((x, y)) = (x_1, \ldots, x_{i-1}, w_i, y_{i+1}, \ldots, y_n)$$

$\eta_e$ is injective since we can recover $(x, y)$ unambiguously.

Notice that in this particular case $\pi(x)\pi(y) = \pi(w)\pi(cp(e))$ i.e.

$$\frac{1}{Q(e)} \sum_{\gamma_{xy} \ni e} \pi(x)\pi(y) |\gamma_{xy}| \leq \frac{2n}{\pi(w)}\pi(w)n \sum_{\gamma_{xy} \ni e} \pi(cp(e)) \leq 2n^2$$
A matching of a graph $G = (V, E)$ is a collection $M \subseteq E$ of vertex-disjoint edges.

We want to sample from

$$\pi(M) = \frac{\lambda^{|M|}}{\sum_M \lambda^{|M|}}$$
**Sampling Algorithm**

**Input:** $\lambda$, Number of iterations $R$.

**Output:** $M_R$, Sample from $\pi(M)$.

Start with $M_1 = \{\}$;

For $\{i = 1..R\}$

- With probability $\frac{1}{2}$ let $M = M_i$, otherwise
- select an edge $e = (u, v)$ u.a.r. and set

\[
M' = \begin{cases} 
M_i - e & \text{if } e \in M_i \\
M_i + e & \text{if both } u \text{ and } v \text{ are unmatched in } M_i \\
M_i + e - e' & \text{if exactly one of } u \text{ and } v \text{ is matched in } M_i \\
M_i & \text{otherwise}
\end{cases}
\]

With probability $\min\{1, \frac{\pi(M')}{\pi(M_i)}\}$ set $M = M'$

$M_{i+1} = M$

\[
\]

---

RMM (UCSC)  
MCMC and Co-Learning  
June-2007  
15 / 41
Markov Chain

Type 0 transition (Deletion)
Markov Chain

Type 0 transition (Deletion)
Type 1 transition (Swap)
Markov Chain

Type 1 transition (Swap)
Markov Chain

Type 2 transition (Insertion)
Markov Chain

Type 2 transition (Insertion)
Canonical paths

Path from $x$ to $y$: unwind $x \oplus y$ ($\oplus$ denotes symmetric difference)
Canonical paths

$x$: $P_1 \quad P_2 \quad \ldots \quad P_{i-1} \quad P_i \quad P_{i+1} \quad \ldots \quad P_m$

$y$

$M$

$M'$

$RMM$ (UCSC)

MCMC and Co-Learning
Injective Function

\[ \eta_t(X, Y) = \begin{cases} 
X \oplus Y \oplus (M \cup M') - e_{XYt} & \text{if } t \text{ is type 1 and current path is a cycle} \\
X \oplus Y \oplus (M \cup M') & \text{otherwise}
\end{cases} \]
Injective Function

\[ X \oplus Y = \begin{cases} 
\eta_t(X, Y) \oplus (M \cup M') + e_{XYt} & \text{if } t \text{ is type 1 and current path is} \\
\eta_t(X, Y) \oplus (M \cup M') & \text{otherwise}
\end{cases} \]

Recover \( X \) and \( Y \) by using the global ordering of the paths.
Congestion Bound

\[ \frac{\pi(X) \pi(Y)}{\pi(M) \pi(\eta_t(X, Y))} \leq \lambda \]

\[ \frac{1}{Q(e)} \sum_{\gamma_{xy} \ni e} \pi(x) \pi(y) \mid \gamma_{xy} \mid \leq |E| \max\{1, \frac{\pi(M)}{\pi(M')}\} (2n)\lambda \sum_{\gamma_{xy} \ni e} \pi(cp(e)) \]

\[ \leq 2n |E| \lambda^2 \]

\[ x \oplus y \]

\[ M \cup M' \]

\[ \eta_t(X, Y) \]
The goal is to sample a subset of experts $X$ of size $m$, according to the following distribution

$$
\pi(X) = \prod_{k=1}^{K} \sum_{x \in X} w_{k,x}
$$

Algorithm 2: Sampling a subset of $m$ experts

Data: Matrix of current weights $W(t)$, Number of iterations $R$.
Result: $S_R$ (Sample from $\pi(X)$).

Pick an arbitrarily subset $S_0$;

for $i = 1..R$ do

Choose u.a.r. an element $x \in [n] \setminus S_i$ and $y \in S_i$. Let $S'_i = S_i \cup x \setminus y$;

With probability $\min \{1, \frac{\pi(S'_i)}{\pi(S)}\}$, set $S_{i+1} = S'_i$, otherwise $S_{i+1} = S_i$

end
Canonical Paths

- Define an arbitrarily ordering over all pairs of nodes.
- Let the canonical path between the $X$ and $Y$ be $X \sim S_1 \sim \ldots \sim S_m = Y$ where
- Hard to define weight preserving injections, since $\pi(X)$ does not factor.

$$
\pi(X) = \prod_{k=1}^{K} \sum_{x \in X} w_{k,x}
$$
We need to bound

\[
\frac{\pi(X)\pi(Y)}{\pi(W)\pi(\eta_e((X, Y)))} \leq ?
\]

Example

\(X = \{x, y, z\}, \ Y = \{a, b, c\}, \ W = \{x, y, c\}\) and \(\eta_e((X, Y)) = \{a, b, z\}\)

\[
(x_1 + y_1 + z_1)(x_2 + y_2 + z_2)(a_1 + b_1 + c_1)(a_2 + b_2 + c_2) \\
\neq (x_1 + y_1 + c_1)(x_2 + y_2 + c_2)(a_1 + b_1 + z_1)(a_2 + b_2 + z_2)
\]

\(x_i = y_i = z_i = 0\)

\(a_i = b_i = c_i = 1\)
Sample a mapping $M : [k] \rightarrow [n]$ according to the distribution

$$
\pi(M) = \prod_{k=1}^{K} w_{k,M(k)}
$$

We can prove a bound for this chain.

$$
\rho \leq 2nm^2 \max \left\{ \frac{w_{k,h}}{w_{k,h'}} \right\}
$$

However this chain is equivalent to predict without the knowledge that $m$ experts perform well on all tasks.
Mappings

- Include sample in the average only if the mapping $M$ involves only $m$ experts.
- Not work because the total probability weight on mapping with more than $m$ experts can be exponentially bigger.
Penalize mappings that use many experts.

Instead use the following stationary distribution:

$$\pi(M) = \frac{1}{\zeta c(M) - m} \prod_{k=1}^{K} w_{k,M(k)}$$

where $c(M)$ is the number of experts used by mapping $M$ i.e. penalize mappings involving more than $m$ experts.
Let $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$ be two arbitrarily states in $\Omega$ where $x_i, y_i \in [n]$. Define the canonical path $\gamma_{xy}$ as the path $X \sim M_1 \sim M_2 \sim \ldots \sim M_n = Y$ where $M_i = (y_1, \ldots, y_i, x_{i+1}, \ldots, x_n)$. 

![Diagram of canonical paths](image)
Canonical Paths

Pick an arbitrarily edge
\( e = (W, W') = ((w_1, \ldots, w_i, \ldots, w_n), (w_1', \ldots, w_i', \ldots, w_n)) \). Let
\( cp(e) = \{(X, Y)|\gamma_{XY} \ni e, X \in \Omega, c(Y) = m\} \) be the set of all canonical paths that use edge \( e \). Define \( \eta_e : cp(e) \rightarrow \Omega \) as follows:

\[
\eta_e((X, Y)) = (x_1, \ldots, x_{i-1}, w_i, y_{i+1}, \ldots, y_n)
\]

\( \eta_e \) is injective since we can recover \((X, Y)\) unambiguously. However the best bound I could find for \( \pi(Y)\pi(Y) \) in this case is

\[
\frac{\pi(X)\pi(Y)}{\pi(W)\pi(\eta_e((X, Y)))} = \frac{\left(\frac{1}{\xi c(X)} \prod_{k=1}^{K} w_k, x(k)\right) \left(\prod_{k=1}^{K} w_k, y(k)\right)}{\left(\frac{1}{\xi c(W)} \prod_{k=1}^{K} w_k, w(k)\right) \left(\frac{1}{\xi c(\eta)} \prod_{k=1}^{K} w_k, \eta(k)\right)}
\]

\[
= \frac{\bar{\zeta}c(W) \bar{\zeta}c(\eta)}{\bar{\zeta}c(X)} \leq \bar{\zeta}^m
\]
Canonical Paths

\[
\frac{1}{Q(e)} \sum_{\gamma_{xy} \ni e} \pi(x) \pi(y) | \gamma_{xy} | \leq \frac{2nm^2 \xi^{m-1}}{\min\{\pi(W), \pi(W')\}} \pi(W) \sum_{\gamma_{xy} \ni e} \pi(cp(e))
\]

\[
\leq \frac{2nm^2 \xi^m \pi(W)}{\min\{\pi(W), \pi(W')\}}
\]

\[
= 2nm^2 \xi^m \max\{1, \frac{\pi(W)}{\pi(W')}\}
\]

\[
\leq 2nm^2 \xi^{m+1} \max\{\frac{W_{k,h}}{W_{k,h'}}\}
\]
The true output $y_k(t)$ for all tasks $1 \leq k \leq K$ and all steps $1 \leq t \leq T$ is zero.

The outputs $h_{i,k}(t)$ of each expert $i$ on each task $k$ at every step $t$ are generated by first picking a subset $S$ of experts of size $m \ll K$ at random. This set $S$ is the set of experts that perform well on all the tasks.

For each task $k$ pick an expert $s$ at random from the set $S$ and set $h_{s,k}(t) = X$, where $X$ is a uniform random variable in the interval $(0, 1/2\sqrt{1/10})$, and for all other experts $i \neq s$ set $h_{i,k}(t) = Y$ where $Y$ is a uniform random variable in the interval $(0, y)$.
Results

![Graphs showing time and master loss for y=0.75 and y=0.5](image)

- For y=0.75:
  - Blue line: Mappings
  - Green line: WMA
  - Red line: Subsets

- For y=0.5:
  - Blue line: Mappings
  - Green line: WMA
  - Red line: Subsets

Legend:
- Mappings
- WMA
- Subsets
Results

![Graph 1](image1)

![Graph 2](image2)
Random Linear Separators


