Online (and Distributed) Learning with Information Constraints

Ohad Shamir
Weizmann Institute of Science

Online Algorithms and Learning Workshop
Leiden, November 2014
Main Question

How well can we learn with information constraints on how we can interact with data?
Only part of the data is visible to the learning algorithm
Partial-Information Constraints

Only part of the data is visible to the learning algorithm.

**Examples**

- A few coordinates from each example
  - Multi-armed bandits + variants, attribute-efficient learning, missing features...
Partial-Information Constraints

Only part of the data is visible to the learning algorithm

Examples

- A few coordinates from each example
  - Multi-armed bandits + variants, attribute-efficient learning, missing features...

- A linear projection of each example
  - Bandit linear optimization
Partial-Information Constraints

Only part of the data is visible to the learning algorithm

Examples

- A few coordinates from each example
  - Multi-armed bandits + variants, attribute-efficient learning, missing features…
- A linear projection of each example
  - Bandit linear optimization
- Some other function of each example
  - Partial monitoring; bandit convex optimization…
Partial-Information Constraints

Only part of the data is visible to the learning algorithm

Examples

- A few coordinates from each example
  - Multi-armed bandits + variants, attribute-efficient learning, missing features...
- A linear projection of each example
  - Bandit linear optimization
- Some other function of each example
  - Partial monitoring; bandit convex optimization...

Price of partial information: Known in some cases, setting-dependent
Consider the squared loss or exp-concave losses $O(\log(T))$ regret possible, but known algorithms require $\Omega(d^2)$ memory [Vovk 1998; Azoury and Warmuth 2001; Hazan et al. 2007].

$O(\sqrt{T})$ regret possible with $O(d)$-memory algorithms (e.g. online gradient descent).

Is there a price to pay for linear-memory methods?
Memory Constraints

Example (Online Convex Optimization)

- Consider the squared loss or exp-concave losses
Example (Online Convex Optimization)

- Consider the squared loss or exp-concave losses
- $O(\log(T))$ regret possible, but known algorithms require $\Omega(d^2)$ memory [Vovk 1998; Azoury and Warmuth 2001; Hazan et al. 2007]
Example (Online Convex Optimization)

- Consider the squared loss or exp-concave losses
- $O(\log(T))$ regret possible, but known algorithms require $\Omega(d^2)$ memory [Vovk 1998; Azoury and Warmuth 2001; Hazan et al. 2007]
- $O(\sqrt{T})$ regret possible with $O(d)$-memory algorithms (e.g. online gradient descent)

Is there a price to pay for linear-memory methods?
Example (Online PCA)

- i.i.d. data stream \( x_1, \ldots, x_m \in \mathbb{R}^d \)
- Goal: Find direction/s with most variance
- Solution: Consider empirical covariance matrix \( \frac{1}{m} \sum_{i=1}^{m} x_i x_i^\top \)
- \( d \times d \) matrix: too large in high dimensions


Ohad Shamir
Learning with Information Constraints
Memory Constraints

Example (Online PCA)

- i.i.d. data stream $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^d$
- Goal: Find direction/s with most variance
- Solution: Consider empirical covariance matrix $\frac{1}{m} \sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^\top$
- $d \times d$ matrix: too large in high dimensions

Can we do online PCA with limited memory?


Ohad Shamir
Learning with Information Constraints
Communication Constraints

- Training examples partitioned in a distributed system
- Communication is (relatively) slow and expensive
- Can we learn with small communication complexity?
Main Questions

- Can we quantify – information theoretically – how such constraints affect our performance?
- Can we trade-off between data size and information constraints?
Related Work

- Problem-specific lower bounds (e.g. multi-armed bandits)
- Learning with communication constraints [e.g. Balcan et al. 2012, Zhang et al. 2013]: Non i.i.d. data or very small communication budget
- Communication/space bounds in theoretical computer science. But not learning problems and/or non-i.i.d. data
$\mathbf{Protocols}$

**Definition**

For $t = 1, \ldots, m$
- Receive $n$ i.i.d. instances $X_t$
- Compute $W_t = f_t(X_t, W_1, W_2, \ldots, W_{t-1})$, where $|W_t| \leq b$ bits
- Return $f(W_1, \ldots, W_m)$
\((b, n, m)\) Protocols

**Definition**

- For \( t = 1 \ldots m \)
  - Receive \( n \) i.i.d. instances \( X^t \)
  - Compute \( W^t = f_t(X^t, W^1, W^2, \ldots, W^{t-1}) \), where \( |W^t| \leq b \) bits
- Return \( f(W^1 \ldots W^m) \)
Examples of \((b, n, m)\) Protocols

Bounded memory online algorithms

For \(t = 1, \ldots, m\)

Receive i.i.d. instance \(x_t\)

Compute \(W_t = f_t(x_t, W_{t-1})\), where \(|W_t| \leq b\) bits

Return \(f(W_m)\)
Examples of \((b, n, m)\) Protocols

**Bounded memory online algorithms**

- For \(t = 1 \ldots m\)
  - Receive i.i.d. instance \(x_t\)
  - Compute \(W^t = f_t(x_t, W^{t-1})\), where \(|W^t| \leq b\) bits
- Return \(f(W^m)\)

```
Ohad Shamir
Learning with Information Constraints
10/23
```
Examples of \((b, n, m)\) Protocols

### Learning from Expert Advice w/ partial information

- For \(t = 1 \ldots m\)
  - Reward vector \(x_t\) generated
  - Observe
    \[ W^t = f_t(x_t, W^1, W^2, \ldots, W^{t-1}), \]
    where \(|W^t| \leq b\) bits

E.g. for multiarmed bandits, \(W^t = x_{t,i_t}\) where coordinate \(i_t\) selected based on past observations
Examples of \((b, n, m)\) Protocols

Non-interactive distributed learning

Machine 1
\(X^1\)
\(W^1\)

Machine 2
\(X^2\)
\(W^2\)

\vdots

Machine m
\(X^m\)
\(W^m\)
Examples of \((b, n, m)\) Protocols

One-pass distributed learning

Machine 1
\[
\begin{align*}
X^1 & \\
\vdots & \\
W^1 & \\
\end{align*}
\]

Machine 2
\[
\begin{align*}
X^2 & \\
\vdots & \\
W^2 & \\
\end{align*}
\]

Machine \(m\)
\[
\begin{align*}
X^m & \\
\vdots & \\
W^m & \\
\end{align*}
\]
Hide-and-Seek Problem

Instances $x_t$ drawn i.i.d. from a product distribution on $\{-1, +1\}$

A single unknown coordinate $j^* \in \{1, \ldots, d\}$ is biased

$E[x_t] = \rho e^{j^*}$

Goal: Given sampled data, find $j^*$
Instances $x_t$ drawn i.i.d. from a product distribution on $\{-1, +1\}^d$

A single unknown coordinate $j^* \in \{1, \ldots, d\}$ is biased

- $E[x_t] = \rho \, e_{j^*}$

Goal: Given sampled data, find $j^*$
Result for \((b, 1, m)\) Protocols (e.g. bounded-memory)

Theorem

Given \(m\) i.i.d. instances; bias \(\rho\):

Can detect biased coordinate if \(m \gg \log(d)/\rho^2\)

Just return coordinate \(\tilde{J}\) with highest empirical mean

Any \((b, 1, m)\) protocol will fail if \(m \ll (d/b)/\rho^2\)

Tight up to log factors

Can be generalized to \(n > 1\)
Theorem

Given $m$ i.i.d. instances; bias $\rho$:

- Can detect biased coordinate if $m \gg \log(d)/\rho^2$
- Just return coordinate $\tilde{J}$ with highest empirical mean
- Any $(b, 1, m)$ protocol will fail if $m \ll (d/b)/\rho^2$
- Tight up to log factors

Can be generalized to $n > 1$
Result for \((b, 1, m)\) Protocols (e.g. bounded-memory)

**Theorem**

Given \(m\) i.i.d. instances; bias \(\rho\):

- **Can detect biased coordinate** if \(m \gg \frac{\log(d)}{\rho^2}\)
- *Just return coordinate \(\tilde{J}\) with highest empirical mean*
Result for \((b, 1, m)\) Protocols (e.g. bounded-memory)

**Theorem**

Given \(m\) i.i.d. instances; bias \(\rho\):

- **Can detect biased coordinate** if \(m \gg \log(d)/\rho^2\)
  - *Just return coordinate* \(\tilde{J}\) *with highest empirical mean*
- **Any \((b, 1, m)\) protocol will fail** if \(m \ll (d/b)/\rho^2\)
Result for \((b, 1, m)\) Protocols (e.g. bounded-memory)

**Theorem**

Given \(m\) i.i.d. instances; bias \(\rho\):

- **Can detect biased coordinate** if \(m \gg \log(d)/\rho^2\)

  - Just return coordinate \(\tilde{J}\) with highest empirical mean

- **Any \((b, 1, m)\) protocol will fail** if \(m \ll (d/b)/\rho^2\)

- **Tight up to log factors**

- Can be generalized to \(n > 1\)
As long as $W_t$ doesn’t “know” $j^*$, can only convey $b/d$ bits on $x_t$, $j^*$ in expectation

Standard data processing inequality:

$W_t$ conveys $\leq \min\{\rho/2, b/d\}$ bits on $j^*$

We prove an information contraction inequality:

$W_t$ conveys $\leq \rho/2 b/d$ bits on $j^*$
As long as $W^t$ doesn’t "know" $j^*$, can only convey $b/d$ bits on $x_{t,j^*}$ in expectation.
• As long as $W^t$ doesn’t ”know” $j^*$, can only convey $b/d$ bits on $x_{t,j^*}$ in expectation
• $x_{t,j^*}$ conveys at most $\rho^2$ bits on $j^*$
As long as $W^t$ doesn’t ”know” $j^*$, can only convey $b/d$ bits on $x_{t,j^*}$ in expectation.

- $x_{t,j^*}$ conveys at most $\rho^2$ bits on $j^*$.

- Standard data processing inequality: $W^t$ conveys $\leq \min\{\rho^2, b/d\}$ bits on $j^*$.
As long as $W^t$ doesn’t “know” $j^*$, can only convey $\frac{b}{d}$ bits on $x_{t,j^*}$ in expectation

$x_{t,j^*}$ conveys at most $\rho^2$ bits on $j^*$

Standard data processing inequality: $W^t$ conveys $\leq \min\{\rho^2, \frac{b}{d}\}$ bits on $j^*$

We prove an information contraction inequality: $W^t$ conveys $\leq \rho^2 \frac{b}{d}$ bits on $j^*$
Proof Idea

- \( W^t \) conveys \( \rho^2 b/d \) bits on \( j^* \)
- After \( m \) rounds: \( m\rho^2 b/d \) bits of information on \( j^* \).
Proof Idea

- $W^t$ conveys $\rho^2 b/d$ bits on $j^*$
- After $m$ rounds: $m\rho^2 b/d$ bits of information on $j^*$.
- $\Rightarrow$ Insufficient to detect if $m \ll (d/b)/\rho^2$
Application: Online learning with partial information

- Reward vectors $x_1, \ldots, x_T$ chosen from $\{0, 1\}^d$.
- Some bits of each $x_t$ are sequentially observed.

**Theorem:**
Expected regret $\Omega(\sqrt{dbT})$

Result is generic – holds regardless of what are those $b$ bits:
- Chosen coordinate $x_t, i_t$ (Multi-armed bandits)
- Some other coordinate $x_t, j_t$; Some subset of coordinates $x_t, j_1, x_t, j_2, \ldots, x_t, j_k$ (semi-bandit feedback; prediction with limited advice; bandits with side-information; attribute-efficient learning...);
- Linear projection of $x_t$ (bandit linear optimization);
- Using some bounded-width feedback matrix (partial monitoring)...

- Even if algorithm can choose $b$ bits based on $x_t$!
Application: Online learning with partial information

- \( T \) reward vectors \( \mathbf{x}_1, \ldots, \mathbf{x}_T \) chosen from \( \{0, 1\}^d \).
- **Some** \( b \) bits of each \( \mathbf{x}_t \) are sequentially observed.
Application: Online learning with partial information

- $T$ reward vectors $\mathbf{x}_1, \ldots, \mathbf{x}_T$ chosen from $\{0, 1\}^d$.
- **Some** $b$ bits of each $\mathbf{x}_t$ are sequentially observed.

**Theorem:** Expected regret $\Omega \left( \sqrt{\frac{d}{b} T} \right)$

Result is generic – holds regardless of what are those $b$ bits:
- Chosen coordinate $x_{t,i_t}$ (Multi-armed bandits)
Application: Online learning with partial information

- $T$ reward vectors $\mathbf{x}_1, \ldots, \mathbf{x}_T$ chosen from $\{0, 1\}^d$.
- Some $b$ bits of each $\mathbf{x}_t$ are sequentially observed.

**Theorem:** Expected regret $\Omega(\sqrt{\frac{d}{b} T})$

Result is generic – holds regardless of what are those $b$ bits:

- Chosen coordinate $x_{t,i_t}$ (Multi-armed bandits)
- Some other coordinate $x_{t,j_t}$; Some subset of coordinates $x_{t,j_1}, x_{t,j_2}, \ldots, x_{t,j_k}$ (semi-bandit feedback; prediction with limited advice; bandits with side-information; attribute-efficient learning...);
- Linear projection of $\mathbf{x}_t$ (bandit linear optimization); Using some bounded-width feedback matrix (partial monitoring)...
- Even if algorithm can choose $b$ bits based on $\mathbf{x}_t$!
Theorem

∃ stochastic/online optimization problems in $\mathbb{R}^d$, s.t.

- Possible to get $\tilde{O}(\sqrt{T})$ regret
- Any $b$-memory online algorithm (or $b$-communication distributed algorithm) has $\Omega(\sqrt{(d^2/b)T})$ regret

---

Caveat: Non-convex, but efficiently solvable
Theorem

\[ \exists \text{ stochastic/online optimization problems}^a \text{ in } \mathbb{R}^d, \text{ s.t.} \]

- Possible to get \( \tilde{O}(\sqrt{T}) \) regret
- Any \( b \)-memory online algorithm (or \( b \)-communication distributed algorithm) has \( \Omega(\sqrt{(d^2/b)T}) \) regret

---

\(^a\text{Caveat: Non-convex, but efficiently solvable}\)

- Conclusion: **Any online algorithm with \( o(d^2) \) memory is suboptimal** (e.g. gradient descent, mirror descent)
- **New Memory-sample and communication-sample trade-offs**: In a statistical setting, can get same average regret with more data, even with memory/communication constraints
Simplest Case - Detecting Correlations

Given i.i.d. sample $x_1, \ldots, x_m$, detect a single pair of correlated features.
Simplest Case - Detecting Correlations

Given i.i.d. sample $x_1, \ldots, x_m$, detect a single pair of correlated features

- Statistically optimal methods use empirical covariance matrix
  \[
  \frac{1}{m} \sum_{i=1}^{m} x_i x_i^\top
  \]
- Problem: Requires too much communication/memory when dimension $d$ and $m$ are large

- **We Show:** There are situations where no memory/communication-efficient method can be statistically optimal
**Theorem**

**Suppose:**
- $\mathbb{E}[xx^\top] = I_d + \tau (E_{i^*j^*} + E_{j^*i^*})$ for unknown $i^*, j^*$

**Given $m$ samples, $\forall i \neq j$,**
- $\Pr(|\tilde{x}_i x_j - \mathbb{E}[x_i x_j]| \geq \frac{\tau}{2}) \leq 2 \exp(-m\tau^2/6)$.

Then for $\tau = \tilde{\Theta}(1/d)$:
- $m \gg d^2$: Enough to return largest off-diagonal entry of empirical covariance matrix
- $m \ll \frac{d^4}{b}$: Any $b$-memory online algorithm / $(b, n, m)$ protocol for appropriate $n$ will fail for some such distribution
Summary

- **Generic framework** of information constraints in learning
- Same results applicable to
  - Different information constraints
    - (memory, communication, partial information...)
  - Different learning settings
    - (multi-armed bandits and variants, stochastic/online optimization, sparse PCA and covariance estimation...)

- **New resource trade-offs:**
  - More data - less memory
  - More data - less communication (in a natural regime)
Summary

- **Generic framework** of information constraints in learning

- Same results applicable to
  - Different **information constraints**
    (memory, communication, partial information...)
  - Different **learning settings**
    (multi-armed bandits and variants, stochastic/online optimization, sparse PCA and covariance estimation...)

- **New resource trade-offs:**
  - More data - less memory
  - More data - less communication (in a natural regime)

- Information **Trade-offs for other learning settings?**
More details: arXiv and NIPS 2014

THANKS!