Statistical Inference from Dependent Observations

Constantinos Daskalakis
EECS and CSAIL, MIT
Supervised Learning Setting

Given:

• Training set \((x_i, y_i)_{i=1}^n\) of examples
  - \(x_i\): feature (aka covariate) vector of example \(i\)
  - \(y_i\): label (aka response) of example \(i\)

• Assumption: \((x_i, y_i)_{i=1}^n \sim_{iid} D\) where \(D\) is unknown

• Hypothesis class \(\mathcal{H} = \{h_\theta \mid \theta \in \Theta\}\) of (randomized) responses
  - e.g. \(\mathcal{H} = \{\langle \theta, \cdot \rangle \mid \|\theta\|_2 \leq 1\}\),
  - e.g. 2 \(\mathcal{H} = \{\text{some class of Deep Neural Networks}\}\)
  - generally, \(h_\theta\) might be randomized

• Loss function: \(\ell : Y \times Y \rightarrow \mathbb{R}\)

Goal: Select \(\theta\) to minimize expected loss: \(\min_\theta (\mathbb{E}_{(x,y)} \mathbb{E}_{D} [\ell(h_\theta(x), y)])\)

Goal 2: In realizable setting (i.e. when, under \(D\), \(y \sim h_{\theta^*}(x)\)), estimate \(\theta^*\)
E.g. Linear Regression

**Input:** $n$ feature vectors $x_i \in \mathbb{R}^d$ and responses $y_i \in \mathbb{R}, i = 1, \ldots, n$

**Assumption:** for all $i$, $y_i$ sampled as follows:

- $y_i = \theta^\top x_i + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 1)$

**Goal:** Infer $\theta$
E.g.2 Logistic Regression

**Input:** $n$ feature vectors $x_i \in \mathbb{R}^d$ and binary responses $y_i \in \{\pm 1\}$

**Assumption:** for all $i$, $y_i$ sampled as follows:

- $\Pr[y_i = \sigma_i] = \frac{1}{1 + e^{-2\theta x_i \sigma_i}}$

**Goal:** Infer $\theta$
Maximum Likelihood Estimator (MLE)

In the standard linear and logistic regression settings, under mild assumptions about the design matrix $X$, whose rows are the covariate vectors, MLE is strongly consistent.

MLE estimator $\hat{\theta}$ satisfies: $\|\hat{\theta} - \theta\|_2 = O \left( \frac{\sqrt{d}}{\sqrt{n}} \right)$
Beyond Realizable Setting: Learnability

- Recall:
  - **Assumption:** \( \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \sim_{iid} D \), where \( D \) unknown
  - **Goal:** choose \( h \in \mathcal{H} \) to minimize expected loss under some loss function \( \ell \), i.e.
    \[
    \min_{h \in \mathcal{H}} \mathbb{E}_{(x, y) \sim D} [\ell(h(x), y)]
    \]
    Loss\(_D\)\(_h\)

- Let \( \hat{h} \in \mathcal{H} \) be the **Empirical Risk Minimizer**, i.e.
  \[
  \hat{h} \in \arg \min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i} \ell(h(x_i), y_i)
  \]

- Then:
  \[
  \text{Loss}_D(\hat{h}) \leq \inf_{h \in \mathcal{H}} \text{Loss}_D(h) + O\left(\sqrt{VC(\mathcal{H})}/n\right), \text{ for Boolean } \mathcal{H}, \text{ 0-1 loss } \ell
  \]
  \[
  \text{Loss}_D(\hat{h}) \leq \inf_{h \in \mathcal{H}} \text{Loss}_D(h) + O(\lambda \mathcal{R}_n(\mathcal{H})), \text{ for general } \mathcal{H}, \text{ } \lambda\text{-Lipschitz } \ell
  \]
  \[
  \vdots
  \]
Supervised Learning Setting on Steroids

But what do these results really mean?
Broader Perspective

- **ML methods commonly operate under stringent assumptions**
  - train set: independent and identically distributed (i.i.d.) samples
  - test set: i.i.d. samples from same distribution as training

- **Goal:** *relax standard assumptions* to accommodate two important challenges
  - (i) censored/truncated samples and (ii) dependent samples
  - Censoring/truncation $\Leftarrow$ systematic missing of data
    $\Rightarrow$ train set $\neq$ test set
  - Data Dependencies $\Leftarrow$ peer effects, spatial, temporal dependencies
    $\Rightarrow$ no apparent source for independence
Today’s Topic: Statistical Inference from Dependent Observations
Why dependent?

Observations \((x_i, y_i)_i\) are commonly collected on some *spatial* domain, some *temporal* domain, or on a *social network*. As such, they are not independent, but intricately dependent

e.g. spin glasses
• neighboring particles influence each other

e.g. social networks
decisions/opinions of nodes are influenced by the decisions of their neighbors (*peer effects*)
Statistical Physics and Machine Learning

- Spin Systems [Ising’25]
- MRFs, Bayesian Networks, Boltzmann Machines
  - Probability Theory
  - MCMC
  - Machine Learning
  - Computer Vision
  - Game Theory
  - Computational Biology
  - Causality
  - …
Peer Effects on Social Networks

Several studies of peer effects, in applications such as:
- criminal networks [Glaeser et al’96]
- welfare participation [Bertrand et al’00]
- school achievement [Sacerdote’01]
- participation in Retirement Plans [Duflo-Saez’03]
- obesity [Trogdon et al’08, Christakis-Fowler’13]

AddHealth Dataset:
- National Longitudinal Study of Adolescent Health
- National study of students in grades 7-12
- friendship networks, personal and school life, age, gender, race, socio-economic background, health,…

MicroEconomics:
- Behavior/Opinion Dynamics e.g.
  [Schelling’78], [Ellison’93], [Young’93, ’01],
  [Montanari-Saberi’10],…

Econometrics:
- Disentangling individual from network effects [Manski’93], [Bramouille-Djebarri-Fortin’09], [Li-Levina-Zhu’16],…
Menu

• Motivation
• Part I: Regression w/ dependent observations
• Part II: Statistical Learning Theory w/ dependent observations
• Conclusions
Menu

• Motivation
• Part I: Regression w/ dependent observations
• Part II: Statistical Learning Theory w/ dependent observations
• Conclusions
Standard Linear Regression

**Input:** $n$ feature vectors $x_i \in \mathbb{R}^d$ and responses $y_i \in \mathbb{R}$, $i = 1, \ldots, n$

**Assumption:** for all $i$, $y_i$ sampled as follows:

- $y_i = \theta^\top x_i + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 1)$
- $y_1, \ldots, y_n$ independent

**Goal:** Infer $\theta$
Standard Logistic Regression

**Input:** $n$ feature vectors $x_i \in \mathbb{R}^d$ and binary responses $y_i \in \{\pm 1\}$

**Assumption:** for all $i$, $y_i$ sampled as follows:

- $\Pr[y_i = \sigma_i] = \frac{1}{1 + e^{-2\theta^\top x_i \sigma_i}}$
- $y_1, \ldots, y_n$ independent

**Goal:** Infer $\theta$
In the standard linear and logistic regression settings, under mild assumptions about the design matrix $X$, whose rows are the covariate vectors, MLE is strongly consistent.

MLE estimator $\hat{\theta}$ satisfies: $\|\hat{\theta} - \theta\|_2 = O\left(\frac{\sqrt{d}}{\sqrt{n}}\right)$
Standard Linear Regression

**Input:** \( n \) feature vectors \( x_i \in \mathbb{R}^d \) and responses \( y_i \in \mathbb{R}, \ i = 1, \ldots, n \)

**Assumption:** for all \( i \), \( y_i \) sampled as follows:

**Goal:** Infer \( \theta \)
Linear Regression w/ Dependent Samples

**Input:** $n$ feature vectors $x_i \in \mathbb{R}^d$ and responses $y_i \in \mathbb{R}$, $i = 1, \ldots, n$

**Assumption:** for all $i$, $y_i$ sampled as follows conditioning on $y_{-i}$:

- $y_i = \theta^\top x_i + \beta \sum_{j \neq i} A_{ij} (y_j - \theta^\top x_j) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 1)$
- $y_1, \ldots, y_n$ independent

**Parameters:**
- coefficient vector $\theta$, inverse temperature $\beta$ (unknown)
- Interaction matrix $A$ (known)

**Goal:** Infer $\theta, \beta$
Standard Logistic Regression

**Input:** $n$ feature vectors $x_i \in \mathbb{R}^d$ and binary responses $y_i \in \{\pm 1\}$

**Assumption:** for all $i$, $y_i$ sampled as follows:

- $\Pr[y_i = \sigma_i] = \frac{1}{1 + e^{-2\theta^\top x_i \sigma_i}}$

- $y_1, \ldots, y_n$ independent

**Goal:** Infer $\theta$
Logistic Regression w/ Dependent Samples (today’s focus)

**Input:** \( n \) feature vectors \( x_i \in \mathbb{R}^d \) and binary responses \( y_i \in \{\pm 1\} \)

**Assumption:** for all \( i \), \( y_i \) sampled as follows **conditioning on** \( y_{-i} \):

\[
\text{Pr}[y_i = \sigma_i] = \frac{1}{1 + e^{-2(\mathbf{\theta}^\top x_i + \beta \sum_{j \neq i} A_{ij} y_j)\sigma_i}}
\]

- \( y_1, \ldots, y_n \) independent

**Parameters:**
- coefficient vector \( \mathbf{\theta} \), inverse temperature \( \beta \) (unknown)
- Interaction matrix \( A \) (known)

**Goal:** Infer \( \theta, \beta \)
Logistic Regression w/ Dependent Samples (today’s focus)

**Input:** $n$ feature vectors $x_i \in \mathbb{R}^d$ and binary responses $y_i \in \{\pm 1\}$

**Assumption:** $y_1, \ldots, y_n$ are sampled jointly according to measure $\Pr[y = \sigma] = \frac{\exp\left(\sum_{i=1}^{n} (\theta^T x_i)\sigma_i + \beta \sigma^T A \sigma\right)}{Z}$

**Parameters:**
- coefficient vector $\theta$, inverse temperature $\beta$ (unknown)
- Interaction matrix $A$ (known)

**Goal:** Infer $\theta, \beta$
- Challenge: one sample, likelihood contains $Z$
  i.e. lack of LLN, hard to compute MLE

For $\beta = 0$ equivalent to Logistic regression
Main Result for Logistic Regression from Dependent Samples

**Theorem.** Make the standard assumptions that $\theta_0$ and all $x_i$'s are bounded in $\ell_2$, and that $\lambda_{\text{min}}\left(\frac{1}{n}X^TX\right)$ is bounded away from zero, where $X$ is the matrix whose rows are the feature vectors.

Suppose additionally that $\beta_0$ and $\|A\|_{\infty}$ are bounded by absolute constants (independent of $n$), and that $\|A\|_F^2 = \Omega(n)$.

In time $O(nd \log n)$ an estimate $(\hat{\theta}, \hat{\beta})$ can be computed, which is $O\left(\sqrt{\frac{d}{n}}\right)$ consistent. In particular, $\left\| (\hat{\theta}, \hat{\beta}) - (\theta_0, \beta_0) \right\|_2 \leq O\left(\sqrt{\frac{d}{n}}\right)$ with probability 99.9%.

N.B. We show similar results for linear regression with peer effects.
Menu

• Motivation
• Part I: Regression w/ dependent observations
  • Proof Ideas
• Part II: Statistical Learning Theory w/ dependent observations
• Conclusions
Menu

• Motivation
• Part I: Regression w/ dependent observations
  • Proof Ideas
• Part II: Statistical Learning Theory w/ dependent observations
• Conclusions
MLE instead of MLE

- **Likelihood** involves $Z = Z_{\theta, \beta}$ (the partition function), so is non-trivial to compute.
- [Besag’75,…,Chatterjee’07] studies the maximum pseudolikelihood estimator (MPLE)

\[
PL(\beta, \theta) := \left( \prod_i \Pr_{\beta, \theta, A}[y_i|y_{-i}] \right)^{1/n} = \left( \prod_i \frac{1}{1 + \exp(-2(\theta^T x_i + \beta A_i^T y_i))} \right)^{1/n}
\]

- LogPL does not contain $Z$ and is concave. Is MPLE consistent?
  - [Chatterjee’07]: yes, when $\beta > 0, \theta = 0$
  - [BM’18,GM’18]: yes, when $\beta > 0, \theta \neq 0 \in \mathbb{R}, x_i = 1$, for all $i$
- General case?
Problem: Given:
• \( \tilde{x} = (x_1, ..., x_n) \) and
• \( \tilde{y} = (y_1, ..., y_n) \in \{\pm 1\}^n \) sampled as
  \[
  \Pr[\tilde{y} = \sigma] = \frac{\exp(\sum_i (\theta^T x_i) \sigma_i + \beta \sigma^T A \sigma)}{z}
  \]
Infer \( \theta, \beta \)

\[
(\theta_t, \beta_t) = t(\theta_0, \beta_0) + (1 - t)(\hat{\theta}, \hat{\beta}) \text{ where } (\theta_0, \beta_0) \text{ true and } (\hat{\theta}, \hat{\beta}) \text{ MLE.}
\]
Analysis

Problem: Given:

- $\tilde{x} = (x_1, ..., x_n)$ and
- $\tilde{y} = (y_1, ..., y_n) \in \{\pm 1\}^n$ sampled as

\[
\Pr[\tilde{y} = \sigma] = \frac{\exp(\sum_i (\theta^T x_i) \sigma_i + \beta \sigma^T A \sigma)}{z}
\]

Infer $\theta, \beta$

\[
(\theta_t, \beta_t) = t(\theta_0, \beta_0) + (1 - t)(\hat{\theta}, \hat{\beta}) \text{ where } (\theta_0, \beta_0) \text{ true and } (\hat{\theta}, \hat{\beta}) \text{ MPLE.}
\]

\[
g(t) := (\theta - \hat{\theta}, \beta - \hat{\beta})^T \nabla \log PL(\theta_t, \beta_t), \quad g'(t) = (\theta - \hat{\theta}, \beta - \hat{\beta})^T \nabla^2 \log PL(\theta_t, \beta_t)(\theta - \hat{\theta}, \beta - \hat{\beta}).
\]

\[
\left\Vert (\theta - \hat{\theta}, \beta - \hat{\beta}) \right\Vert_2 \cdot \left\Vert \nabla \log PL(\theta_0, \beta_0) \right\Vert_2 \geq |g(1) - g(0)| \geq \left| \int_0^1 g'(t) dt \right| \geq \min_{(\theta, \beta)} \lambda_{\min} \left( -\nabla^2 \log PL(\theta, \beta) \right) \left\Vert (\theta - \hat{\theta}, \beta - \hat{\beta}) \right\Vert_2^2.
\]

Concentration: gradient of log-pseudolikelihood at truth should be small;
- show $\left\Vert \nabla \log PL(\theta_0, \beta_0) \right\Vert_2$ is $O(d/n)$ with probability at least 99%

Anti-concentration: Constant lower bound on $\min_{(\theta, \beta)} \lambda_{\min} \left( -\nabla^2 \log PL(\theta, \beta) \right)$ w.pr. $\geq 99%$
Menu

• Motivation
• Part I: Regression w/ dependent observations
  • Proof Ideas
• Part II: Statistical Learning Theory w/ dependent observations
• Conclusions
Menu

• Motivation
• Part I: Regression w/ dependent observations
  • Proof Ideas
• Part II: Statistical Learning Theory w/ dependent observations
• Conclusions
Supervised Learning

Given:

• Training set \((x_i, y_i)_{i=1}^n\) of examples

• Hypothesis class \(\mathcal{H} = \{h_\theta \mid \theta \in \Theta\}\) of (randomized) responses

• Loss function: \(\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}\)

Assumption: \((x_i, y_i)_{i=1}^n \sim_{iid} D\) where \(D\) is unknown

Goal: Select \(\theta\) to minimize expected loss: \(\min_\theta \mathbb{E}_{(x,y) \sim D} [\ell(h_\theta(x), y)]\)

Goal 2: In realizable setting (i.e. when, under \(D\), \(y \sim h_{\theta^*}(x)\)), estimate \(\theta^*\)
Supervised Learning w/ Dependent Observations

Given:
- Training set \((x_i, y_i)_{i=1}^n\) of examples
- Hypothesis class \(\mathcal{H} = \{h_\theta \mid \theta \in \Theta\}\) of (randomized) responses
- Loss function: \(\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}\)

Assumption: \((x_i, y_i)_{i=1}^n \sim_{\text{ua}} D\) where \(D\) is unknown

Goal: Select \(\theta\) to minimize expected loss: 
\[
\min_{\theta} \mathbb{E}_{(x,y) \sim D} [\ell(h_\theta(x), y)]
\]

Goal 2: In realizable setting (i.e. when, under \(D\), \(y \sim h_{\theta^*}(x)\)), estimate \(\theta^*\)
Supervised Learning w/ Dependent Observations

Given:

- Training set \((x_i, y_i)_{i=1}^{n}\) of examples
- Hypothesis class \(\mathcal{H} = \{h_\theta \mid \theta \in \Theta\}\) of (randomized) responses
- Loss function: \(\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}\)

Assumption': a joint distribution \(D\) samples training examples and unknown test sample \textit{jointly}, i.e.

- \((x_i, y_i)_{i=1}^{n} \cup (x, y) \sim D\)

Goal: select \(\theta\) to minimize: \(\min_{\theta} \mathbb{E}[\ell(h_\theta(x), y)]\)
Main result

Learnability is possible when joint distribution $D$ over training set and test set satisfies *Dobrushin's condition*. 
Dobrushin’s condition

• Given a joint probability vector $\vec{Z} = (Z_1, ..., Z_m)$ define the **influence of variable $i$ to variable $j$**:
  • “worst effect of $Z_i$ to the conditional distribution of $Z_j$ given $Z_{-i-j}$”
  • $\text{Inf}_{i \to j} = \sup_{z_i,z_i',z_{-i-j}} d_{TV} (\Pr[Z_j \mid z_i,z_{-i-j}],\Pr[Z_j \mid z_i',z_{-i-j}])$

• **Dobrushin’s condition:**
  • “all nodes have limited total influence exerted on them”
  • For all $i$, $\sum_{j \neq i} \text{Inf}_{j \to i} < 1$.

**Implies:** Concentration of measure
Fast mixing of Gibbs sampler
Learnability under Dobrushin’s condition

• Suppose:
  • \{ (X_1, Y_1), ..., (X_n, Y_n), (X, Y) \} \sim D, where D is joint distribution
  • D: satisfies Dobrushin’s condition
  • D: has the same marginals

• Define \( \text{Loss}_D(h) = \mathbb{E}[\ell(h(x), y)] \)

• There exists a learning algorithm which, given \{ (X_1, Y_1), ..., (X_n, Y_n) \}, outputs \( \hat{h} \in \mathcal{H} \) such that

\[
\text{Loss}_D(\hat{h}) \leq \inf_{h \in \mathcal{H}} \text{Loss}_D(h) + \tilde{O}\left(\sqrt{\text{VC}(\mathcal{H})/n}\right) \quad \text{for Boolean } \mathcal{H}, \ 0-1 \text{ loss } \ell
\]

\[
\text{Loss}_D(\hat{h}) \leq \inf_{h \in \mathcal{H}} \text{Loss}_D(h) + \tilde{O}(\lambda \mathcal{G}_n(\mathcal{H})) \quad \text{for general } \mathcal{H}, \ \lambda\text{-Lipschitz } \ell
\]

(need stronger than Dobrushin condition for general \( \mathcal{H} \))

Essentially the same bounds as i.i.d
Menu

• Motivation
• Part I: Regression w/ dependent observations
  • Proof Ideas
• Part II: Statistical Learning Theory w/ dependent observations
• Conclusions
Menu

• Motivation
• Part I: Regression w/ dependent observations
  • Proof Ideas
• Part II: Statistical Learning Theory w/ dependent observations
• Conclusions
Goals of Broader Line of Work

Goal: *relax standard assumptions* to accommodate two important challenges

- (i) *censored/truncated samples* and (ii) *dependent samples*
- Censoring/truncation $\iff$ *systematic missing of data*
  $\implies$ train set $\neq$ test set
- Data Dependencies $\iff$ *peer effects, spatial, temporal dependencies*
  $\implies$ no apparent source for independence
Statistical Estimation from Dependent Observations

- Regression on a network (linear and logistic, $\sqrt{d/n}$ rates):
  
  Constantinos Daskalakis, Nishanth Dikkala, Ioannis Panageas: *Regression from Dependent Observations*.
  In the 51st Annual ACM Symposium on the Theory of Computing (STOC’19).

- Statistical Learning Theory Framework (learnability and generalization bounds):
  
  Yuval Dagan, Constantinos Daskalakis, Nishanth Dikkala, Siddhartha Jayanti: *Generalization and learning under Dobrushin’s condition*.
  In the 32nd Annual Conference on Learning Theory (COLT’19).

Thank you!