An Asymptotic Analysis of Estimators: Generative, Discriminative, Pseudolikelihood

ICML 2008 Helsinki, Finland

July 6, 2008

Percy Liang Michael I. Jordan

UC Berkeley
Goal: structured prediction

\[ x \Rightarrow y = (y_1, \ldots, y_n) \]

We focus on probabilistic models of \( x \) and \( y \)
Goal: structured prediction

\[ x \Rightarrow y = (y_1, \ldots, y_n) \]

We focus on probabilistic models of \( x \) and \( y \)

Many approaches

- Discriminative (logistic regression, conditional random fields)
- Generative (Naive Bayes, Bayesian networks, HMMs)
Goal: structured prediction

\[ x \Rightarrow y = (y_1, \ldots, y_n) \]

We focus on probabilistic models of \( x \) and \( y \)

Many approaches

Discriminative (logistic regression, conditional random fields)
Generative (Naive Bayes, Bayesian networks, HMMs)
Pseudolikelihood [Besag, 1975]
Goal: structured prediction

\[ x \Rightarrow y = (y_1, \ldots, y_n) \]

We focus on probabilistic models of \( x \) and \( y \)

Many approaches

- Discriminative (logistic regression, conditional random fields)
- Generative (Naive Bayes, Bayesian networks, HMMs)
- Pseudolikelihood [Besag, 1975]
- Composite likelihood [Lindsay, 1988]
- Multi-conditional learning [McCallum, et al., 2006]
- Piecewise training [Sutton & McCallum, 2005]
- Variational relaxations [Wainwright, 2006]
- Agreement-based learning [Liang, et al., 2008]

...how to choose among these approaches?
Goal: structured prediction

\[ x \Rightarrow y = (y_1, \ldots, y_n) \]

We focus on probabilistic models of \( x \) and \( y \)

Many approaches

Discriminative (logistic regression, conditional random fields)
Generative (Naive Bayes, Bayesian networks, HMMs)
Pseudolikelihood [Besag, 1975]
Composite likelihood [Lindsay, 1988]
Multi-conditional learning [McCallum, et al., 2006]
Piecewise training [Sutton & McCallum, 2005]
Variational relaxations [Wainwright, 2006]
Agreement-based learning [Liang, et al., 2008]

...how to choose among these approaches?

Our work:

• Put first three in a unified composite likelihood framework
• Compare their statistical properties theoretically
Goal: structured prediction

\[ x \Rightarrow y = (y_1, \ldots, y_n) \]

We focus on probabilistic models of \( x \) and \( y \)

Many approaches

- **Discriminative** (logistic regression, conditional random fields)
- **Generative** (Naive Bayes, Bayesian networks, HMMs)
- **Pseudolikelihood** [Besag, 1975]
- Composite likelihood [Lindsay, 1988]
- Multi-conditional learning [McCallum, et al., 2006]
- Piecewise training [Sutton & McCallum, 2005]
- Variational relaxations [Wainwright, 2006]
- Agreement-based learning [Liang, et al., 2008]

…how to choose among these approaches?

Our work:

- Put first three in a unified **composite likelihood** framework
- Compare their statistical properties theoretically


Existing intuitions:

- **Discriminative**: lower bias
- **Generative**: lower variance

[Ng & Jordan, 2002; Bouchard & Triggs, 2004]
Existing intuitions:

- **Discriminative**: lower bias
- **Generative**: lower variance
  
  [Ng & Jordan, 2002; Bouchard & Triggs, 2004]

- **Pseudolikelihood**: slower statistical convergence
  
  [Besag, 1975]
Existing intuitions:

- **Discriminative**: lower bias
  - **Generative**: lower variance
  [Ng & Jordan, 2002; Bouchard & Triggs, 2004]

- **Pseudolikelihood**: slower statistical convergence
  [Besag, 1975]

Our general result:

Derive the (excess) risk of composite likelihood estimators
Existing intuitions:

- **Discriminative**: lower bias
  - **Generative**: lower variance
  [Ng & Jordan, 2002; Bouchard & Triggs, 2004]

- **Pseudolikelihood**: slower statistical convergence
  [Besag, 1975]

Our general result:

Derive the (excess) risk of composite likelihood estimators

Specific conclusions:

If the model is well-specified:

\[ \text{Risk( generative) } < \text{Risk( discriminative) } < \text{Risk( pseudolikelihood) } \]
Existing intuitions:

- **Discriminative**: lower bias
  - **Generative**: lower variance
  [Ng & Jordan, 2002; Bouchard & Triggs, 2004]
- **Pseudolikelihood**: slower statistical convergence
  [Besag, 1975]

Our general result:

Derive the *(excess) risk* of composite likelihood estimators

Specific conclusions:

If the model is well-specified:

Risk(generative) < Risk(discriminative) < Risk(pseudolikelihood)

If the model is misspecified:

Risk(discriminative) < Risk(pseudolikelihood), Risk(generative)
Model-based estimators and neighborhoods

Generative: $\hat{\theta}_g = \arg\max_{\theta} \hat{E} \log p_\theta(x, y)$
Model-based estimators and neighborhoods

Generative: $\hat{\theta}_g = \arg\max_{\theta} \hat{E} \log p_\theta(x, y)$

$\bullet (x, y) = \{(\ast, \ast)\}$
Model-based estimators and neighborhoods

Generative:  
\[ \hat{\theta}_g = \arg\max_{\theta} \mathbb{E} \log p_{\theta}(x, y) \]

\[ (x, y) = \{(\ast, \ast)\} \]

Discriminative:  
\[ \hat{\theta}_d = \arg\max_{\theta} \mathbb{E} \log p_{\theta}(y | x) \]

\[ = \{(\ast, \ast)\} \]
Model-based estimators and neighborhoods

Generative: \( \hat{\theta}_g = \arg\max_{\theta} \hat{\mathbb{E}} \log p_{\theta}(x, y) \)

\[ (x, y) = \{(*, *) \} \]

Discriminative: \( \hat{\theta}_d = \arg\max_{\theta} \hat{\mathbb{E}}[\log p_{\theta}(x, y) - \log p_{\theta}(x)] \)
Model-based estimators and neighborhoods

Generative: \( \hat{\theta}_g = \arg\max_{\theta} \hat{\mathbb{E}} \log p_{\theta}(x, y) \)

\( (x, y) = \{(*, *)\} \)

Discriminative: \( \hat{\theta}_d = \arg\max_{\theta} \hat{\mathbb{E}}[\log p_{\theta}(x, y) - \log p_{\theta}(x)] \)

\( (x, y) = \{(x, *)\} \)
Model-based estimators and neighborhoods

Generative: \( \hat{\theta}_g = \arg\max_{\theta} \mathbb{E} \log p_{\theta}(x, y) \)

\( (x, y) \) = \{(\ast, \ast)\}

Discriminative: \( \hat{\theta}_d = \arg\max_{\theta} \mathbb{E}[\log p_{\theta}(x, y) - \log p_{\theta}(x)] \)

\( (x, y) \) = \{(x, \ast)\}

More generally: \( \hat{\theta} = \arg\max_{\theta} \mathbb{E}[\log p_{\theta}(x, y) - \log p_{\theta}(r(x, y))] \)

\( (x, y) \) = \( r(x, y) \)
Model-based estimators and neighborhoods

Generative: \( \hat{\theta}_g = \arg\max_{\theta} \hat{E} \log p_\theta(x, y) \)

\[ (x, y) = \{(*, *) \} \]

Discriminative: \( \hat{\theta}_d = \arg\max_{\theta} \hat{E} [\log p_\theta(x, y) - \log p_\theta(x)] \)

\[ (x, y) = \{(x, *)\} \]

More generally: \( \hat{\theta} = \arg\max_{\theta} \hat{E} [\log p_\theta(x, y) - \log p_\theta(r(x, y))] \)

\[ (x, y) = r(x, y) \]

\( r(x, y) \) is subset of input-output space we want to contrast
Model-based estimators and neighborhoods

Generative: \( \hat{\theta}_g = \arg\max_{\theta} \hat{E} \log p_{\theta}(x, y) \)

Discriminative: \( \hat{\theta}_d = \arg\max_{\theta} \hat{E} [\log p_{\theta}(x, y) - \log p_{\theta}(x)] \)

More generally: \( \hat{\theta} = \arg\max_{\theta} \hat{E} [\log p_{\theta}(x, y) - \log p_{\theta}(r(x, y))] \)

\( r(x, y) \) is subset of input-output space we want to contrast
Composite likelihood estimators

Discriminative pseudolikelihood:

\[ \hat{\theta}_p = \arg\max_{\theta} \mathbb{E} \left[ \sum_{j=1}^{\ell} \log p(y_j | x, y \setminus \{y_j\}) \right] \]
Composite likelihood estimators

Discriminative pseudolikelihood:

\[ \hat{\theta}_p = \arg\max_{\theta} \sum_{j=1}^{\ell} \hat{E}[\log p(y_j \mid x, y\{y_j\})] \]
Composite likelihood estimators

Discriminative pseudolikelihood:

\[
\hat{\theta}_p = \arg\max_\theta \sum_{j=1}^\ell \hat{E}[\log p(x, y) - \log p(x, y \setminus \{y_j\})]
\]
Composite likelihood estimators

Discriminative pseudolikelihood:

\[
\hat{\theta}_p = \arg\max_{\theta} \sum_{j=1}^{\ell} \hat{E}[\log p(x, y) - \log p(x, y \setminus \{y_j\})]
\]

\[
\hat{\theta}_p = \arg\max_{\theta} \sum_{j=1}^{\ell} \hat{E}[\log p(x, y) - \log p(x, y \setminus \{y_j\})]
\]

\[
\hat{\theta}_p = \arg\max_{\theta} \sum_{j=1}^{\ell} \hat{E}[\log p(x, y) - \log p(x, y \setminus \{y_j\})]
\]
Composite likelihood estimators

Discriminative pseudolikelihood:

\[
\hat{\theta}_p = \arg\max_\theta \sum_{j=1}^\ell \hat{E}[\log p(x, y) - \log p(x, y\{y_j\})]
\]

General composite likelihood:

\[
\hat{\theta} = \arg\max_\theta \sum_j w_j \hat{E}[\log p_\theta(x, y) - \log p_\theta(r_j(x, y))]
\]
Composite likelihood estimators

Discriminative pseudolikelihood:

\[
\hat{\theta}_p = \arg\max_{\theta} \sum_{j=1}^{\ell} \hat{E}[\log p(x, y) - \log p(x, y \setminus \{y_j\})]
\]

General composite likelihood:

\[
\hat{\theta} = \arg\max_{\theta} \sum_j w_j \hat{E}[\log p_\theta(x, y) - \log p_\theta(r_j(x, y))]
\]

\( \hat{E}[(x, y)] = r_1(x, y) \)

\( = r_2(x, y) \)
Review of exponential families

\[
\log p_{\theta}(x, y \mid r(x, y)) = \phi(x, y) \cdot \theta - \log \sum_{(x', y') \in r(x, y)} \exp\{\phi(x', y')^\top \theta\}
\]

features parameters log-partition function
Review of exponential families

\[ \log p_\theta(x, y \mid r(x, y)) = \phi(x, y) \cdot \theta - \log \sum_{(x', y') \in r(x, y)} \exp\{\phi(x', y')^\top \theta\} \]

Moment-generating properties:

Mean:

\[ \nabla \log p_\theta(x, y \mid r(x, y)) = \phi - \mathbb{E}_\theta[\phi \mid r] \]

Variance:

\[ \nabla^2 \log p_\theta(x, y \mid r(x, y)) = -\text{var}_\theta[\phi \mid r] \]

Derivatives are useful for asymptotic Taylor expansions
Sketch of arguments for comparing estimators

\( (x, y) = r(x, y) \)
Sketch of arguments for comparing estimators

\[(x, y) = r(x, y)\]

Intuition:

Grow \(r\) \(\Rightarrow\) model more about data
\(\Rightarrow\) data tells us more about parameters
Sketch of arguments for comparing estimators

\[(x, y) = r(x, y)\]

Intuition:
Grow \(r\) \(\Rightarrow\) model more about data
\(\Rightarrow\) data tells us more about parameters

For exponential families:

\[\theta \rightarrow \text{mean } \phi\]
Sketch of arguments for comparing estimators

\[(x, y) = r(x, y)\]

Intuition:
- Grow \( r \) ⇒ model more about data
- \( \Rightarrow \) data tells us more about parameters

For exponential families:
- Parameters \( \theta \)
- Features \( \phi \)
- \( \theta \rightarrow \text{mean} \ \phi \)

Noise
Sketch of arguments for comparing estimators

\[(x, y) = r(x, y)\]

Intuition:
Grow \( r \) \( \Rightarrow \) model more about data
\( \Rightarrow \) data tells us more about parameters

For exponential families:
Parameters \( \theta \)
Features \( \phi \)
\( \theta \rightarrow \) mean \( \phi \)

\( = r(x, y) \)
Sketch of arguments for comparing estimators

\((x, y) = r(x, y)\)

**Intuition:**

Grow \(r\) \(\Rightarrow\) model more about data
\(\Rightarrow\) data tells us more about parameters

For exponential families:

![Diagram showing relationship between parameters and features](image)

- Features: \(\phi\)
- Parameters: \(\theta\)
- Noise
- \(\theta \rightarrow \text{mean } \phi\)
- Slope = variance of \(\phi\) over \(\text{def sensitivity}\)
Sketch of arguments for comparing estimators

\[(x, y) = r(x, y)\]

Intuition:

Grow \( r \) \( \Rightarrow \) model more about data
\( \Rightarrow \) data tells us more about parameters

For exponential families:

\( \theta \rightarrow \) mean \( \phi \)

\( \text{slope} = \text{variance of } \phi \) over \( \text{def } \) sensitivity

Sensitivity \( \uparrow \) \( \Rightarrow \) Risk \( \downarrow \)
Sketch of arguments for comparing estimators

\( (x, y) = r(x, y) \)

Intuition:
Grow \( r \) \( \Rightarrow \) model more about data
\( \Rightarrow \) data tells us more about parameters

For exponential families:

- \( \theta \rightarrow \text{mean } \phi \)
- slope = variance of \( \phi \) over \( \text{def } = \text{sensitivity} \)

Sensitivity \( \uparrow \) \( \Rightarrow \) Risk \( \downarrow \)

Generative
\( \text{var}(\phi) \)  

Discriminative
\( \mathbb{E} \text{var}(\phi | X) \)
Sketch of arguments for comparing estimators

\((x, y) = r(x, y)\)

Intuition:

\[ \text{Grow } r \Rightarrow \text{model more about data} \Rightarrow \text{data tells us more about parameters} \]

For exponential families:

\[ \text{Sensitivity } \uparrow \Rightarrow \text{Risk } \downarrow \]

\[ \text{Generative} \quad \text{Discriminative} \]

\[ \text{var}(\phi) = \mathbb{E} \text{var}(\phi \mid X) + \text{var} \mathbb{E}(\phi \mid X) \]
Sketch of arguments for comparing estimators

\[(x, y) = r(x, y)\]

Intuition:
Grow \(r\) ⇒ model more about data
⇒ data tells us more about parameters

For exponential families:

\[\theta \rightarrow \text{mean } \phi\]

Sensitivity ↑⇒ Risk ↓

Generative

\[\text{Sensitivity} \uparrow \Rightarrow \text{Risk} \downarrow\]

Discriminative

\[\text{var}(\phi) > \mathbb{E} \text{var}(\phi | X)\]
Sketch of arguments for comparing estimators

Intuition:
Grow \( r \) ⇒ model more about data
⇒ data tells us more about parameters

For exponential families:

\[ \text{Sensitivity} \uparrow \Rightarrow \text{Risk} \downarrow \]

Generative  \( \text{Risk(generative)} < \text{Risk(discriminative)} \)

\[ \text{var}(\phi) \geq \mathbb{E} \text{var}(\phi | X) \]
Sketch of arguments for comparing estimators

\[(x, y) = r(x, y)\]

Intuition:

Grow \( r \) \( \Rightarrow \) model more about data
\( \Rightarrow \) data tells us more about parameters

For exponential families:

Features \( \phi \)

\[ \theta \rightarrow \text{mean } \phi \]

Sensitivity \( \uparrow \) \( \Rightarrow \) Risk \( \downarrow \)

Generative

\[ \text{var}(\phi) \geq \mathbb{E} \text{var}(\phi | X) \]

Risk(generative) < Risk(discriminative)

Discriminative

\[ \text{def } \text{sensitivity} \]

noise
Overview of asymptotic analysis

How accurately can we estimate the parameters?

Parameter Error = \( O \left( \frac{\Sigma}{\sqrt{n}} \right) \)

\( \Sigma \): asymptotic variance of parameters
\( n \): number of training examples
Overview of asymptotic analysis

How accurately can we estimate the parameters?

ParameterError \(= O\left(\frac{\Sigma}{\sqrt{n}}\right)\)

\(\Sigma\): asymptotic variance of parameters

\(n\): number of training examples

How fast can we drive the excess risk (expected log-loss) to 0?

In general, get normal rate:

\(\text{Risk} = O\left(\frac{\Sigma}{\sqrt{n}}\right)\)
Overview of asymptotic analysis

How accurately can we estimate the parameters?

\[ \text{ParameterError} = O \left( \frac{\Sigma}{\sqrt{n}} \right) \]

\( \Sigma \): asymptotic variance of parameters
\( n \): number of training examples

How fast can we drive the excess risk (expected log-loss) to 0?

In general, get normal rate:

\[ \text{Risk} = O \left( \frac{\Sigma}{\sqrt{n}} \right) \]

But if some condition is satisfied, get fast rate:

\[ \text{Risk} = O \left( \frac{\Sigma}{n} \right) \]
Overview of asymptotic analysis

How accurately can we estimate the parameters?
ParameterError = $O \left( \frac{\Sigma}{\sqrt{n}} \right)$

$\Sigma$: asymptotic variance of parameters

$n$: number of training examples

How fast can we drive the excess risk (expected log-loss) to 0?

In general, get normal rate:

$\text{Risk} = O \left( \frac{\Sigma}{\sqrt{n}} \right)$

But if some condition is satisfied, get fast rate:

$\text{Risk} = O \left( \frac{\Sigma}{n} \right)$

Issues:

• $O(n^{-1/2})$ or $O(n^{-1})$?

• Compare $\Sigma$
Overview of asymptotic analysis

How accurately can we estimate the parameters?

ParameterError = $O \left( \frac{\Sigma}{\sqrt{n}} \right)$

$\Sigma$: asymptotic variance of parameters

$n$: number of training examples

How fast can we drive the excess risk (expected log-loss) to 0?

In general, get normal rate:

$$\text{Risk} = O \left( \frac{\Sigma}{\sqrt{n}} \right)$$

But if some condition is satisfied, get fast rate:

$$\text{Risk} = O \left( \frac{\Sigma}{n} \right)$$

Issues:

• $O(n^{-\frac{1}{2}})$ or $O(n^{-1})$?

• Compare $\Sigma$

Agenda:

1. Well-specified, one component
2. Well-specified, multiple components
3. Misspecified
Well-specified case

Risk = \( O \left( \frac{\Sigma}{n} \right) \) for all consistent estimators

Thus, sufficient to just compare \( \Sigma \)'s of different estimators...
Well-specified case

Risk = \( O \left( \frac{\Sigma}{n} \right) \) for all consistent estimators
Thus, sufficient to just compare \( \Sigma \)s of different estimators...

Estimator:

\[
\hat{\theta} = \arg\max_{\theta} \mathbb{E}[\log p_{\theta}(x, y) - \log p_{\theta}(r(x, y))]
\]
Well-specified case

Risk = \( O\left(\frac{\Sigma}{n}\right) \) for all consistent estimators
Thus, sufficient to just compare \( \Sigma \)s of different estimators...

Estimator:
\[
\hat{\theta} = \text{argmax}_\theta \mathbb{E}[\log p_\theta(x, y) - \log p_\theta(r(x, y))]
\]

Asymptotic variance:
\[
\Sigma = \Gamma^{-1}, \text{ where } \Gamma = \mathbb{E} \text{ var}(\phi | r) \text{ is the sensitivity}
\]
Well-specified case

Risk = \( O \left( \frac{\Sigma}{n} \right) \) for all consistent estimators
Thus, sufficient to just compare \( \Sigma \)s of different estimators...

Estimator:
\[
\hat{\theta} = \arg\max_{\theta} \mathbb{E}[\log p_{\theta}(x, y) - \log p_{\theta}(r(x, y))]
\]

Asymptotic variance:
\[
\Sigma = \Gamma^{-1}, \text{ where } \Gamma = \mathbb{E} \text{var}(\phi \mid r) \text{ is the sensitivity}
\]

Proof:
By Taylor expansion and moment-generating properties.
Well-specified case: comparing two estimators

Two estimators:

\[ \hat{\theta}_j = \arg\max_{\theta} \hat{E}\left[ \log p_\theta(x, y) - \log p_\theta(r_j(x, y)) \right] \] for \( j = 1, 2 \)

\[ (x, y) = r_1(x, y) \]
\[ = r_2(x, y) \]
Well-specified case: comparing two estimators

Two estimators:

\[ \hat{\theta}_j = \arg\max_{\theta} \hat{E}[\log p_\theta(x, y) − \log p_\theta(r_j(x, y))] \quad \text{for } j = 1, 2 \]

Comparison theorem:

If model is well-specified and

\[ r_1(x, y) \supset r_2(x, y) \]

Then

\[ \text{Risk}(\hat{\theta}_1) \leq \text{Risk}(\hat{\theta}_2) \]
Well-specified case: comparing two estimators

Two estimators:
\[ \hat{\theta}_j = \arg\max_{\theta} \mathbb{E}[\log p_{\theta}(x, y) - \log p_{\theta}(r_j(x, y))] \quad \text{for } j = 1, 2 \]

Comparison theorem:
If model is well-specified and
\[ r_1(x, y) \supset r_2(x, y) \]
Then
\[ \text{Risk}(\hat{\theta}_1) \leq \text{Risk}(\hat{\theta}_2) \]

Proof:
\[ \Sigma_j = \mathbb{E} \text{var}(\phi \mid r_j)^{-1} \]
Well-specified case: comparing two estimators

Two estimators:

\[
\hat{\theta}_j = \arg\max_{\theta} \hat{E}[\log p_{\theta}(x, y) - \log p_{\theta}(r_j(x, y))] \quad \text{for } j = 1, 2
\]

Comparison theorem:

If model is well-specified and

\[r_1(x, y) \supset r_2(x, y)\]

Then

\[\text{Risk}(\hat{\theta}_1) \leq \text{Risk}(\hat{\theta}_2)\]

Proof:

\[\Sigma_j = \mathbb{E} \text{var}(\phi \mid r_j)^{-1} \quad \Sigma_1 \preceq \Sigma_2\]
Well-specified case: comparing two estimators

Two estimators:
\[ \hat{\theta}_j = \arg\max_{\theta} \hat{\mathbb{E}}[\log p_\theta(x, y) - \log p_\theta(r_j(x, y))] \] for \( j = 1, 2 \)

Comparison theorem:
If model is well-specified and \( r_1(x, y) \supset r_2(x, y) \)

Then
\[ \text{Risk}(\hat{\theta}_1) \leq \text{Risk}(\hat{\theta}_2) \]

Proof:
\[ \Sigma_j = \mathbb{E} \text{var}(\phi | r_j)^{-1} \quad \Sigma_1 \preceq \Sigma_2 \quad \text{Risk} = O \left( \frac{\Sigma_j}{n} \right) \]
Well-specified case: comparing two estimators

Two estimators:
\[
\hat{\theta}_j = \arg\max_{\theta} \hat{E}[\log p_{\theta}(x, y) - \log p_{\theta}(r_j(x, y))]
\]
for \( j = 1, 2 \)

Comparison theorem:
If model is well-specified and
\[
r_1(x, y) \supset r_2(x, y)
\]
Then
\[
\text{Risk}(\hat{\theta}_1) \leq \text{Risk}(\hat{\theta}_2)
\]

Proof:
\[
\Sigma_j = \mathbb{E} \text{var}(\phi \mid r_j)^{-1} \quad \Sigma_1 \preceq \Sigma_2 \quad \text{Risk} = O \left( \frac{\Sigma_j}{n} \right)
\]

Modeling more reduces error (when model is well-specified)
Multiple components

Asymptotic variance:

$$\Sigma = \Gamma^{-1} + \Gamma^{-1}C_c\Gamma^{-1}$$
Multiple components

Asymptotic variance:

\[ \Sigma = \Gamma^{-1} + \Gamma^{-1}C_c\Gamma^{-1} \]

\[ \Gamma = \sum_j w_j \mathbb{E} \text{var}(\phi | r_j) \text{ is the sensitivity} \]
Multiple components

Asymptotic variance:

$$\Sigma = \Gamma^{-1} + \Gamma^{-1} C_c \Gamma^{-1}$$

$$\Gamma = \sum_j w_j \mathbb{E} \text{var}(\phi \mid r_j)$$ is the sensitivity

$$C_c \succeq 0 :$$ correction due to multiple components
Multiple components

Asymptotic variance:

\[ \Sigma = \Gamma^{-1} + \Gamma^{-1} C_c \Gamma^{-1} \]

\[ \Gamma = \sum_j w_j \mathbb{E} \text{var}(\phi \mid r_j) \text{ is the sensitivity} \]

\[ C_c \succeq 0 : \text{correction due to multiple components} \]

Comparison theorem:

If the model is well-specified and

\( \hat{\theta}_1: \text{one component } r_1 \quad \hat{\theta}_2: \text{multiple components } \{r_{2,j}\} \)
Multiple components

Asymptotic variance:

\[ \Sigma = \Gamma^{-1} + \Gamma^{-1} C_c \Gamma^{-1} \]

\[ \Gamma = \sum_j w_j \mathbb{E} \text{var}(\phi | r_j) \] is the sensitivity

\[ C_c \succeq 0 : \] correction due to multiple components

Comparison theorem:

If the model is well-specified and

\[ \hat{\theta}_1: \text{one component } r_1 \quad \hat{\theta}_2: \text{multiple components } \{r_{2,j}\} \]

\[ r_1(x, y) \supset r_{2,j}(x, y) \text{ for all components } j \]
Multiple components

Asymptotic variance:

\[ \Sigma = \Gamma^{-1} + \Gamma^{-1} C_c \Gamma^{-1} \]

\[ \Gamma = \sum_j w_j \mathbb{E} \text{var}(\phi | r_j) \] is the sensitivity

\[ C_c \succeq 0 : \text{correction due to multiple components} \]

Comparison theorem:

If the model is well-specified and

\( \hat{\theta}_1: \text{one component } r_1 \quad \hat{\theta}_2: \text{multiple components } \{ r_{2,j} \} \)

\[ r_1(x, y) \supset r_{2,j}(x, y) \text{ for all components } j \]

Then

\[ \text{Risk}(\hat{\theta}_1) \leq \text{Risk}(\hat{\theta}_2) \]
Multiple components

Asymptotic variance:
\[ \Sigma = \Gamma^{-1} + \Gamma^{-1} C_c \Gamma^{-1} \]
\[ \Gamma = \sum_j w_j \mathbb{E} \text{var}(\phi | r_j) \] is the sensitivity
\[ C_c \succeq 0 : \text{correction due to multiple components} \]

Comparison theorem:
If the model is well-specified and
\[ \hat{\theta}_1 : \text{one component } r_1 \quad \hat{\theta}_2 : \text{multiple components } \{r_{2,j}\} \]
\[ r_1(x, y) \supset r_{2,j}(x, y) \] for all components \( j \)
Then
\[ \text{Risk}(\hat{\theta}_1) \leq \text{Risk}(\hat{\theta}_2) \]

Note: does not apply if \( \hat{\theta}_1 \) has more than one component
Misspecified case

Result:
For any estimator in general, get normal rate:

\[ \text{Risk} = O\left(\frac{\Sigma}{\sqrt{n}}\right) \]
Result:

For any estimator in general, get normal rate:

\[ \text{Risk} = O \left( \frac{\Sigma}{\sqrt{n}} \right) \]

But for the discriminative estimator, get fast rate:

\[ \text{Risk} = O \left( \frac{\Sigma}{n} \right) \]
Misspecified case

Result:

For any estimator in general, get normal rate:

\[ \text{Risk} = O \left( \frac{\Sigma}{\sqrt{n}} \right) \]

But for the \textbf{discriminative estimator}, get fast rate:

\[ \text{Risk} = O \left( \frac{\Sigma}{n} \right) \]

Corollary:

\[ \text{Risk}(\text{discriminative}) < \text{Risk}(\text{pseudolikelihood}), \text{Risk}(\text{generative}) \]
Misspecified case

Result:

For any estimator in general, get normal rate:

\[ \text{Risk} = O \left( \frac{\Sigma}{\sqrt{n}} \right) \]

But for the \textbf{discriminative estimator}, get fast rate:

\[ \text{Risk} = O \left( \frac{\Sigma}{n} \right) \]

Corollary:

\[ \text{Risk} \text{(discriminative)} < \text{Risk} \text{(pseudolikelihood)}, \text{Risk} \text{(generative)} \]

\textbf{Key desirable property: training criterion = test criterion}
Verifying the error rates empirically

Setup:

Learn $x$ from $n$ training examples

Estimate (excess) risk from 10,000 trials
Verifying the error rates empirically

Setup:
Learn $x, y_1, y_2$ from $n$ training examples
Estimate (excess) risk from 10,000 trials

Well-specified

generate from

\[ n \]\n
\[ \text{var}(\text{Risk}) \]

\[ 20K \ 40K \ 60K \ 80K \ 100K \]

- Generative
- Discriminative
- Pseudolikelihood
Verifying the error rates empirically

Setup:

Learn \( y_1 y_2 \) from \( n \) training examples

Estimate (excess) risk from 10,000 trials

Well-specified

generate from

\[ \sqrt{n \cdot \text{var}(\text{Risk})} \]

\[ n = 20K, 40K, 60K, 80K, 100K \]

Diagram:

- Generative
- Discriminative
- Pseudolikelihood
Verifying the error rates empirically

Setup:

Learn \( x \) from \( n \) training examples

Estimate (excess) risk from 10,000 trials

Well-specified

generate from

\[ n \cdot \text{var}(\text{Risk}) \]

![Graph showing the relationship between n and var(Risk)]

- Blue: Generative
- Green: Discriminative
- Red: Pseudolikelihood
Verifying the error rates empirically

Setup:

Learn \( x \) from \( n \) training examples

Estimate (excess) risk from 10,000 trials

Well-specified

generate from

\[
\begin{align*}
\text{All: } O(n^{-1})
\end{align*}
\]
Verifying the error rates empirically

Setup:

Learn $x_{\cdot y_1 y_2}$ from $n$ training examples

Estimate (excess) risk from 10,000 trials

Well-specified

generate from

Misspecified

generate from

All: $O(n^{-1})$
Verifying the error rates empirically

Setup:

Learn $x, y_1, y_2$ from $n$ training examples

Estimate (excess) risk from 10,000 trials

Well-specified

generate from

Misspecified

generate from

All: $O(n^{-1})$
Verifying the error rates empirically

Setup:

Learn \( x \) from \( n \) training examples

Estimate (excess) risk from 10,000 trials

Well-specified

\[ \begin{align*}
\text{generate from} & \quad y_1 \\
\text{generate from} & \quad y_2
\end{align*} \]

Misspecified

\[ \begin{align*}
\text{generate from} & \quad y_1 \\
\text{generate from} & \quad y_2
\end{align*} \]

\[ n \cdot \text{var(Risk)} \]

\[ \begin{align*}
\text{20K} & \quad 40K \\
\text{60K} & \quad 80K \\
\text{100K} & \quad n
\end{align*} \]

\[ \text{Generative} \]

\[ \text{Discriminative} \]

\[ \text{Pseudolikelihood} \]

All: \( O(n^{-1}) \)
Verifying the error rates empirically

Setup:

Learn $x$ from $n$ training examples

Estimate (excess) risk from 10,000 trials

Well-specified

generate from

Misspecified

generate from

$n \cdot \text{var}(\text{Risk})$

$n$

20K 40K 60K 80K 100K

20K 40K 60K 80K 100K

$n$

All: $O(n^{-1})$

Fully dis.: $O(n^{-1})$

others: $O(n^{-\frac{1}{2}})$

Generative

Discriminative

Pseudolikelihood
Verifying the error rates empirically

Setup:

Learn $x \rightarrow y_1, y_2$ from $n$ training examples

Estimate (excess) risk from 10,000 trials

Well-specified

generate from

$\begin{array}{c}
\text{Generative} \\
\text{Discriminative} \\
\text{Pseudolikelihood}
\end{array}$

$\begin{array}{c}
20K \\
40K \\
60K \\
80K \\
100K
\end{array}$

$\begin{array}{c}
n \cdot \text{var}(\text{Risk}) \\
n
\end{array}$

All: $O(n^{-1})$

Misspecified

generate from

$\begin{array}{c}
\text{Fully dis.: } O(n^{-1}) \\
others: O(n^{-\frac{1}{2}})
\end{array}$
Application: part-of-speech tagging

Task:

\[ y: \text{Det} - \text{Noun} - \text{Verb} - \text{Det} - \text{Noun} \]

\[ x: \text{The} \quad \text{cat} \quad \text{ate} \quad \text{a} \quad \text{fish} \]
Application: part-of-speech tagging

Task:

\[ y: \text{Det–Noun–Verb–Det–Noun} \]
\[ x: \text{The cat ate a fish} \]

Data: Wall Street Journal news articles (40K sentences)
Application: part-of-speech tagging

Task:

\[ y: \text{Det Noun Verb Det Noun} \]

\[ x: \text{The cat ate a fish} \]

Data: Wall Street Journal news articles (40K sentences)

Synthetic data (well-specified)

Test error

\[ \begin{array}{ccc}
\text{Gen.} & 12.0 & 4.0 \\
\text{Dis.} & 8.0 & 4.0 \\
\text{Pseudo.} & 12.0 & 4.0 \\
\end{array} \]
Application: part-of-speech tagging

Task:

\[ y: \text{Det Noun Verb Det Noun} \]
\[ x: \text{The cat ate a fish} \]

Data: Wall Street Journal news articles (40K sentences)

Synthetic data (well-specified)  Real data (misspecified)

Test error

<table>
<thead>
<tr>
<th>Test error</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.0</td>
</tr>
<tr>
<td>8.0</td>
</tr>
<tr>
<td>4.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test error</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.0</td>
</tr>
<tr>
<td>4.0</td>
</tr>
<tr>
<td>2.0</td>
</tr>
</tbody>
</table>

Summary

Unifying composite likelihood framework for generative, discriminative, pseudolikelihood estimators
Summary

Unifying composite likelihood framework for generative, discriminative, pseudolikelihood estimators

Asymptotic statistics:
  a powerful tool for comparing estimators
Summary

Unifying composite likelihood framework for generative, discriminative, pseudolikelihood estimators

Asymptotic statistics:

a powerful tool for comparing estimators

General conclusions:

• Well-specified case: modeling more of data reduces error
• Desirable: training criterion = test criterion