

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

$$x \Rightarrow y = (y_1, \dots, y_\ell)$$

We focus on probabilistic models of x and y

Goal: structured prediction

$$x \Rightarrow y = (y_1, \dots, y_\ell)$$

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

$$x \Rightarrow y = (y_1, \dots, y_\ell)$$

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

$$x \Rightarrow y = (y_1, \dots, y_\ell)$$

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

$$x \Rightarrow y = (y_1, \dots, y_\ell)$$

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**:

Risk(generative) < **Risk**(discriminative) < **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**:

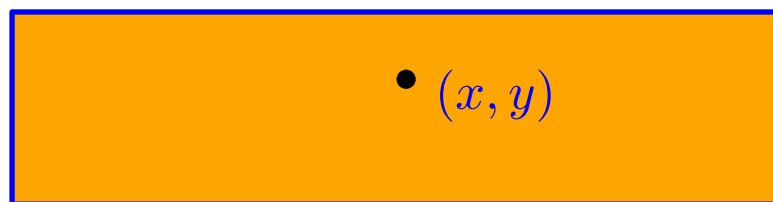
Risk(discriminative) < **Risk**(pseudolikelihood), **Risk**(generative)

Model-based estimators and neighborhoods

Generative: $\hat{\theta}_g = \operatorname{argmax}_{\theta} \hat{\mathbb{E}} \log p_{\theta}(x, y)$

Model-based estimators and neighborhoods

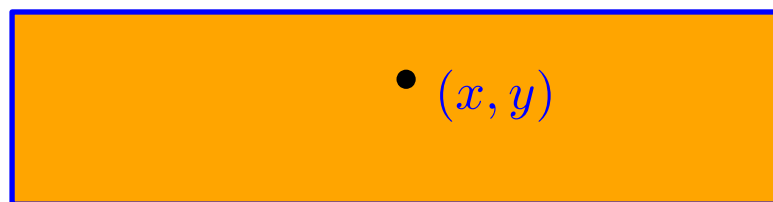
Generative: $\hat{\theta}_g = \operatorname{argmax}_{\theta} \hat{\mathbb{E}} \log p_{\theta}(x, y)$



 = $\{(*, *)\}$

Model-based estimators and neighborhoods

Generative: $\hat{\theta}_g = \operatorname{argmax}_{\theta} \hat{\mathbb{E}} \log p_{\theta}(x, y)$



 = $\{(*, *)\}$

Discriminative: $\hat{\theta}_d = \operatorname{argmax}_{\theta} \hat{\mathbb{E}} \log p_{\theta}(y | x)$

Model-based estimators and neighborhoods

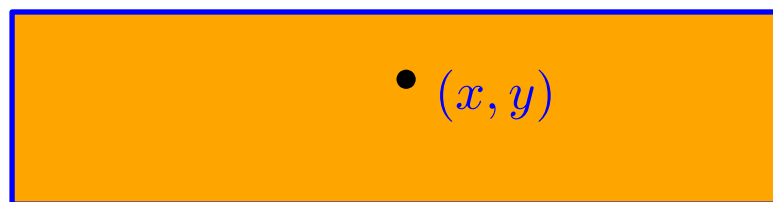
Generative: $\hat{\theta}_g = \operatorname{argmax}_{\theta} \hat{\mathbb{E}} \log p_{\theta}(x, y)$



Discriminative: $\hat{\theta}_d = \operatorname{argmax}_{\theta} \hat{\mathbb{E}} [\log p_{\theta}(x, y) - \log p_{\theta}(x)]$

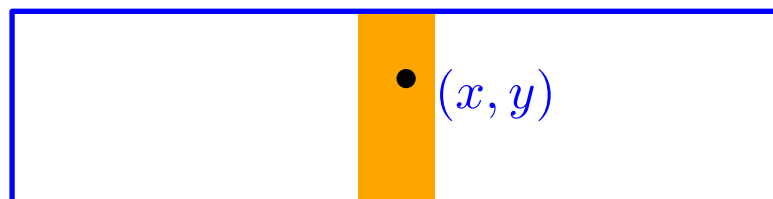
Model-based estimators and neighborhoods


Generative: $\hat{\theta}_g = \operatorname{argmax}_{\theta} \hat{\mathbb{E}} \log p_{\theta}(x, y)$



 = $\{(*, *)\}$

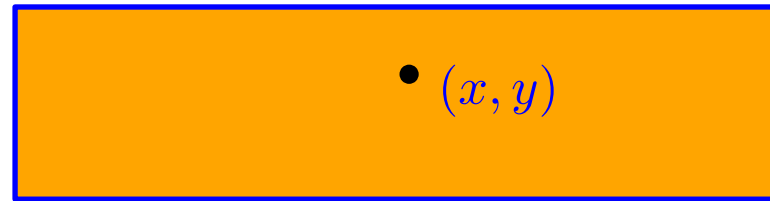
Discriminative: $\hat{\theta}_d = \operatorname{argmax}_{\theta} \hat{\mathbb{E}} [\log p_{\theta}(x, y) - \log p_{\theta}(x)]$



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

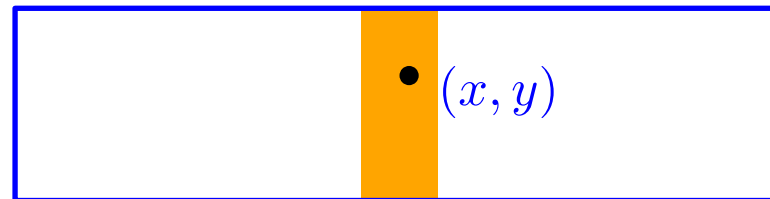
Model-based estimators and neighborhoods


Generative: $\hat{\theta}_g = \operatorname{argmax}_{\theta} \hat{\mathbb{E}} \log p_{\theta}(x, y)$



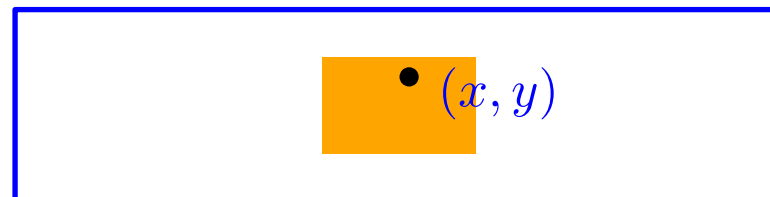
 = $\{(*, *)\}$


Discriminative: $\hat{\theta}_d = \operatorname{argmax}_{\theta} \hat{\mathbb{E}} [\log p_{\theta}(x, y) - \log p_{\theta}(x)]$



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

More generally: $\hat{\theta} = \operatorname{argmax}_{\theta} \hat{\mathbb{E}} [\log p_{\theta}(x, y) - \log p_{\theta}(r(x, y))]$



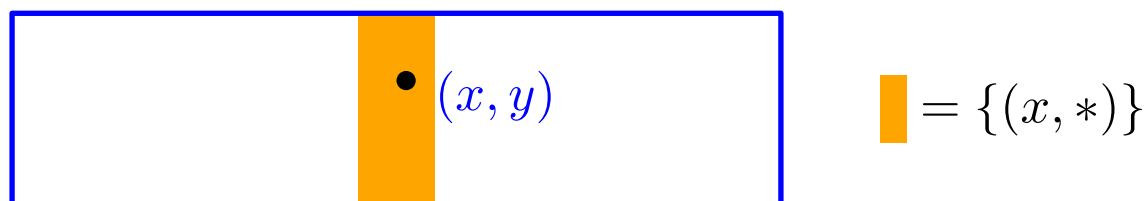
 = $r(x, y)$

Model-based estimators and neighborhoods

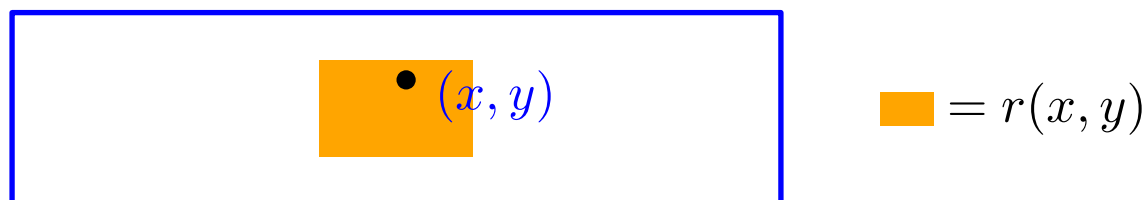
Generative: $\hat{\theta}_g = \operatorname{argmax}_{\theta} \hat{\mathbb{E}} \log p_{\theta}(x, y)$



Discriminative: $\hat{\theta}_d = \operatorname{argmax}_{\theta} \hat{\mathbb{E}} [\log p_{\theta}(x, y) - \log p_{\theta}(x)]$



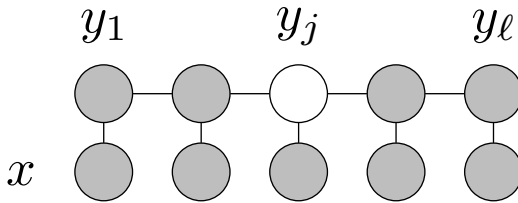
More generally: $\hat{\theta} = \operatorname{argmax}_{\theta} \hat{\mathbb{E}} [\log p_{\theta}(x, y) - \log p_{\theta}(r(x, y))]$



$r(x, y)$ is subset of input-output space we want to contrast

Composite likelihood estimators

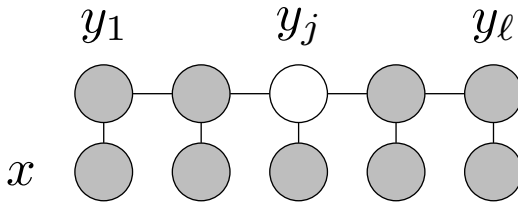
Discriminative pseudolikelihood:



$$\hat{\theta}_p = \operatorname{argmax}_{\theta} \hat{\mathbb{E}} \left[\sum_{j=1}^{\ell} \log p(y_j \mid x, y \setminus \{y_j\}) \right]$$

Composite likelihood estimators

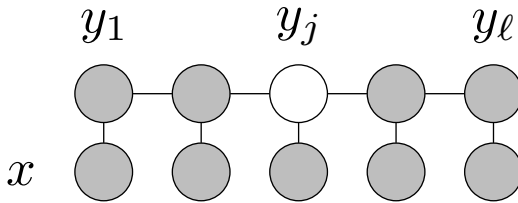
Discriminative pseudolikelihood:



$$\hat{\theta}_p = \operatorname{argmax}_{\theta} \sum_{j=1}^{\ell} \hat{\mathbb{E}}[\log p(y_j \mid x, y \setminus \{y_j\})]$$

Composite likelihood estimators

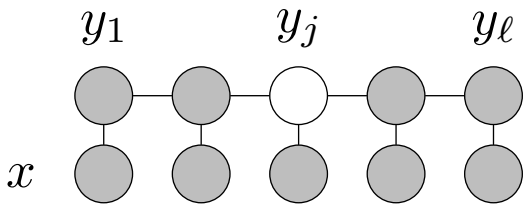
Discriminative pseudolikelihood:



$$\hat{\theta}_p = \operatorname{argmax}_{\theta} \sum_{j=1}^{\ell} \hat{\mathbb{E}}[\log p(x, y) - \log p(x, y \setminus \{y_j\})]$$

Composite likelihood estimators

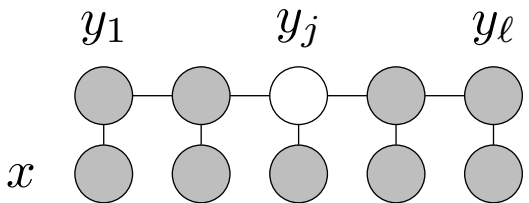
Discriminative pseudolikelihood:



$$\hat{\theta}_p = \operatorname{argmax}_{\theta} \sum_{j=1}^{\ell} \hat{\mathbb{E}}[\log p(x, y) - \log p(\underbrace{x, y \setminus \{y_j\}}_{r_j(x, y)})]$$

Composite likelihood estimators

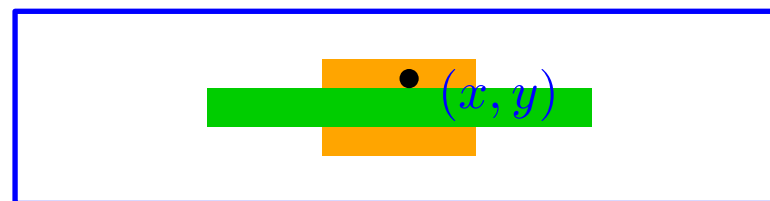
Discriminative pseudolikelihood:





$$\hat{\theta}_p = \operatorname{argmax}_{\theta} \sum_{j=1}^{\ell} \hat{\mathbb{E}}[\log p(x, y) - \log p(\underbrace{x, y \setminus \{y_j\}}_{r_j(x, y)})]$$

General composite likelihood:

$$\hat{\theta} = \operatorname{argmax}_{\theta} \sum_j w_j \hat{\mathbb{E}}[\log p_{\theta}(x, y) - \log p_{\theta}(r_j(x, y))]$$



 = $r_1(x, y)$
 = $r_2(x, y)$

Review of exponential families

$$\log p_{\theta}(x, y \mid r(x, y)) =$$
$$\underbrace{\phi(x, y)}_{\text{features}} \cdot \underbrace{\theta}_{\text{parameters}} - \underbrace{\log \sum_{(x', y') \in r(x, y)} \exp\{\phi(x', y')^{\top} \theta\}}_{\text{log-partition function}}$$

Review of exponential families

$$\log p_{\theta}(x, y \mid r(x, y)) = \underbrace{\phi(x, y)}_{\text{features}} \cdot \underbrace{\theta}_{\text{parameters}} - \underbrace{\log \sum_{(x', y') \in r(x, y)} \exp\{\phi(x', y')^{\top} \theta\}}_{\text{log-partition function}}$$

Moment-generating properties:

Mean:

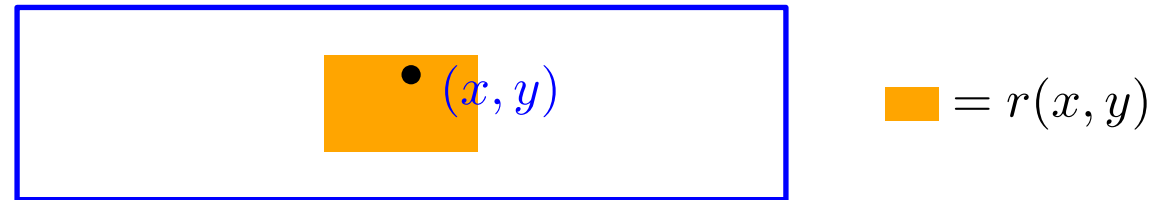
$$\nabla \log p_{\theta}(x, y \mid r(x, y)) = \phi - \mathbb{E}_{\theta}[\phi \mid r]$$

Variance:

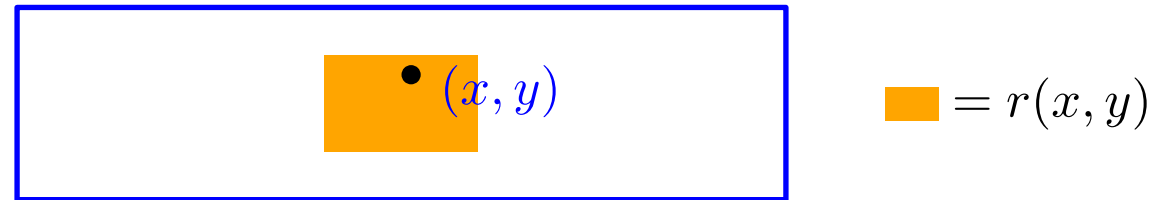
$$\nabla^2 \log p_{\theta}(x, y \mid r(x, y)) = -\text{var}_{\theta}[\phi \mid r]$$

Derivatives are useful for asymptotic Taylor expansions

Sketch of arguments for comparing estimators



Sketch of arguments for comparing estimators

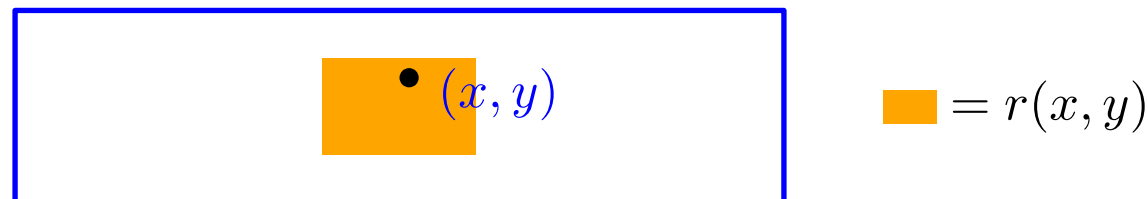


Intuition:

Grow $r \Rightarrow$ model more about data

\Rightarrow data tells us more about parameters

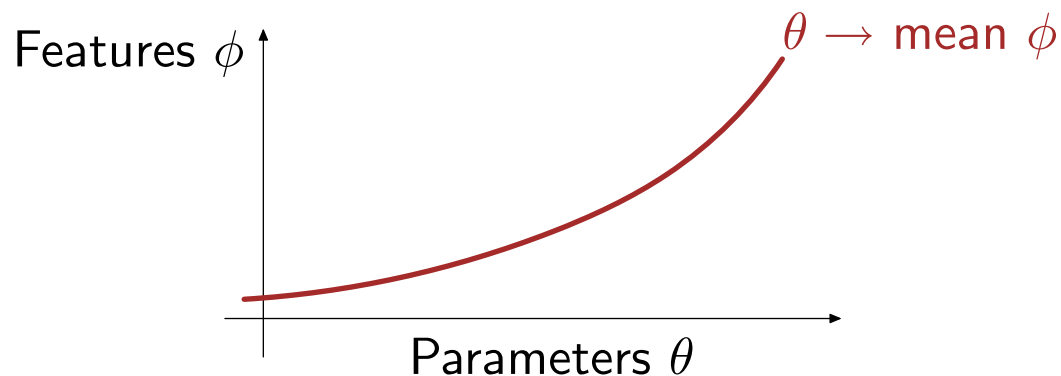
Sketch of arguments for comparing estimators



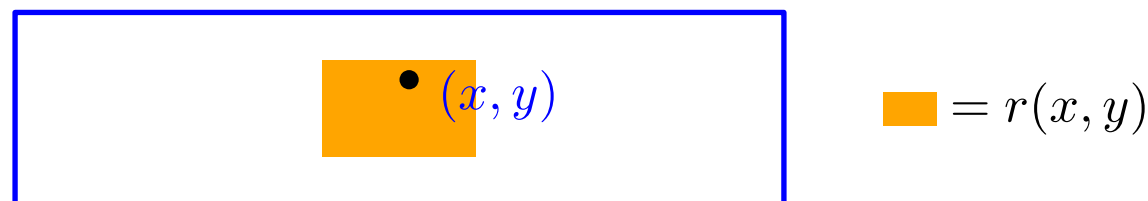
Intuition:

Grow $r \Rightarrow$ model more about data
 \Rightarrow data tells us more about parameters

For exponential families:



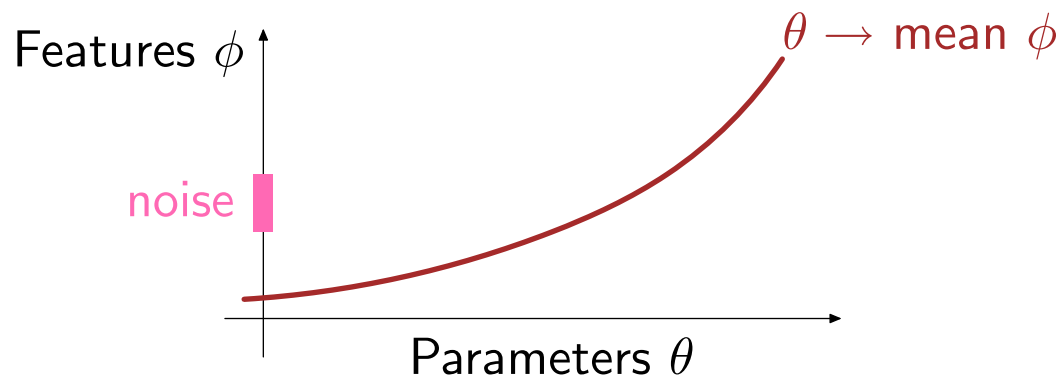
Sketch of arguments for comparing estimators



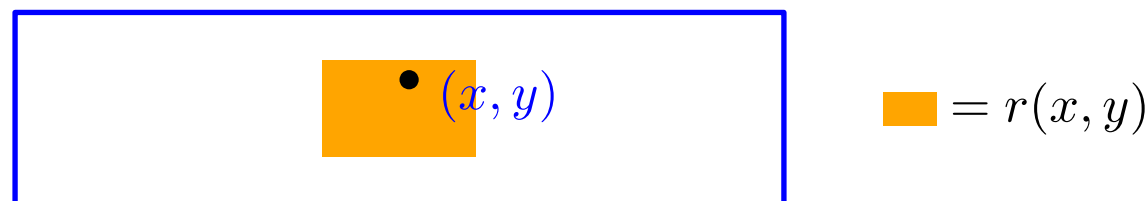
Intuition:

Grow $r \Rightarrow$ model more about data
 \Rightarrow data tells us more about parameters

For exponential families:



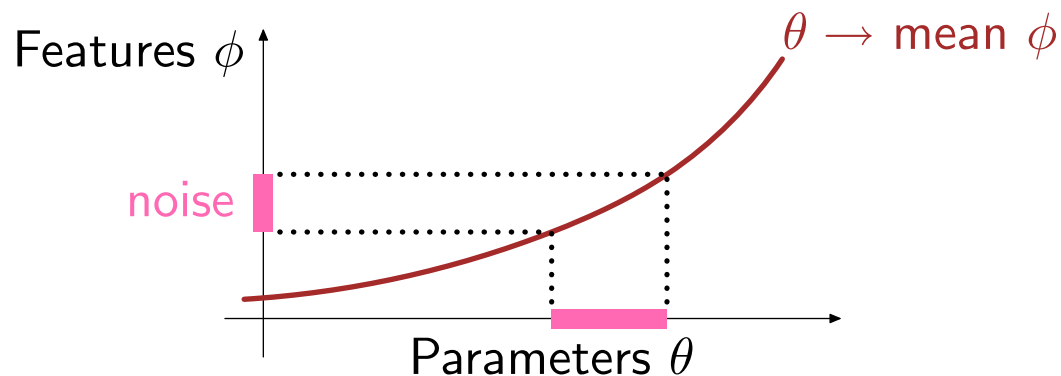
Sketch of arguments for comparing estimators



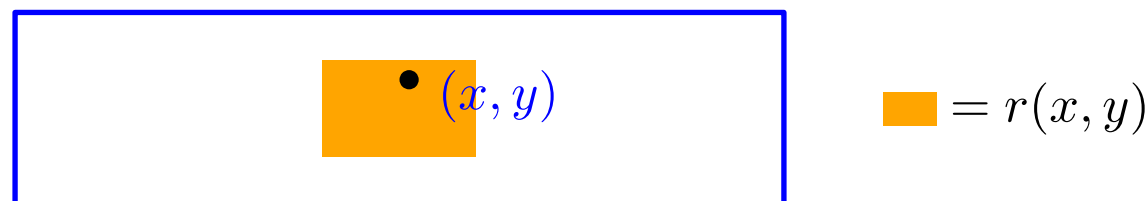
Intuition:

Grow $r \Rightarrow$ model more about data
 \Rightarrow data tells us more about parameters

For exponential families:



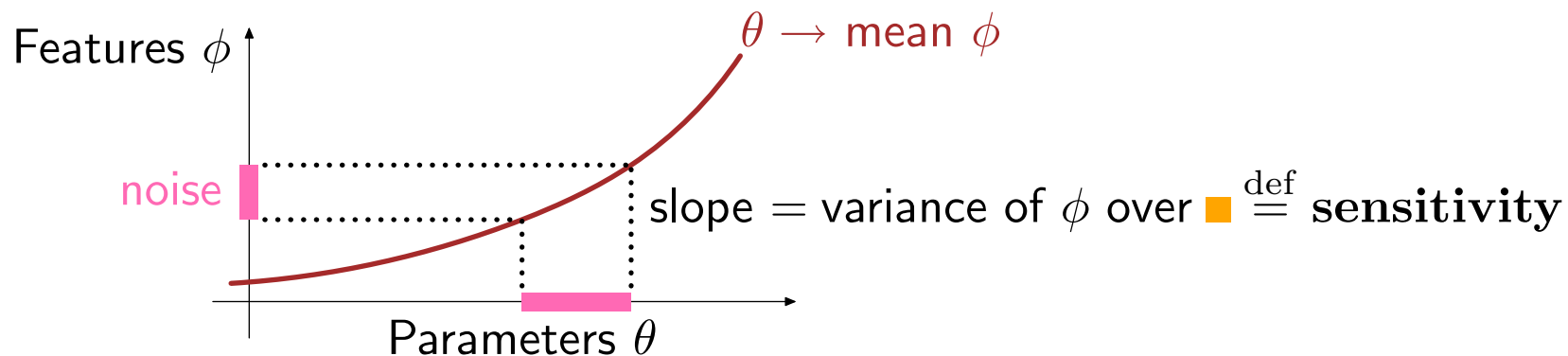
Sketch of arguments for comparing estimators



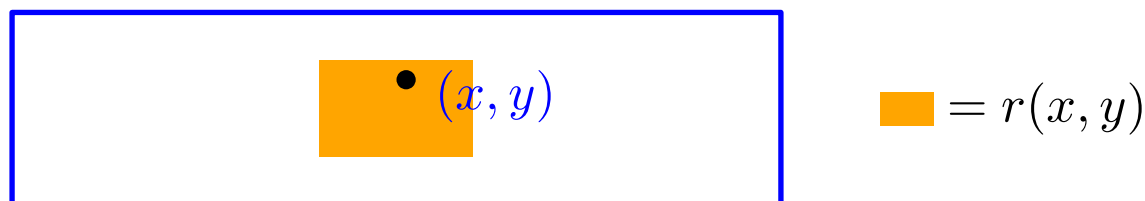
Intuition:

Grow $r \Rightarrow$ model more about data
 \Rightarrow data tells us more about parameters

For exponential families:



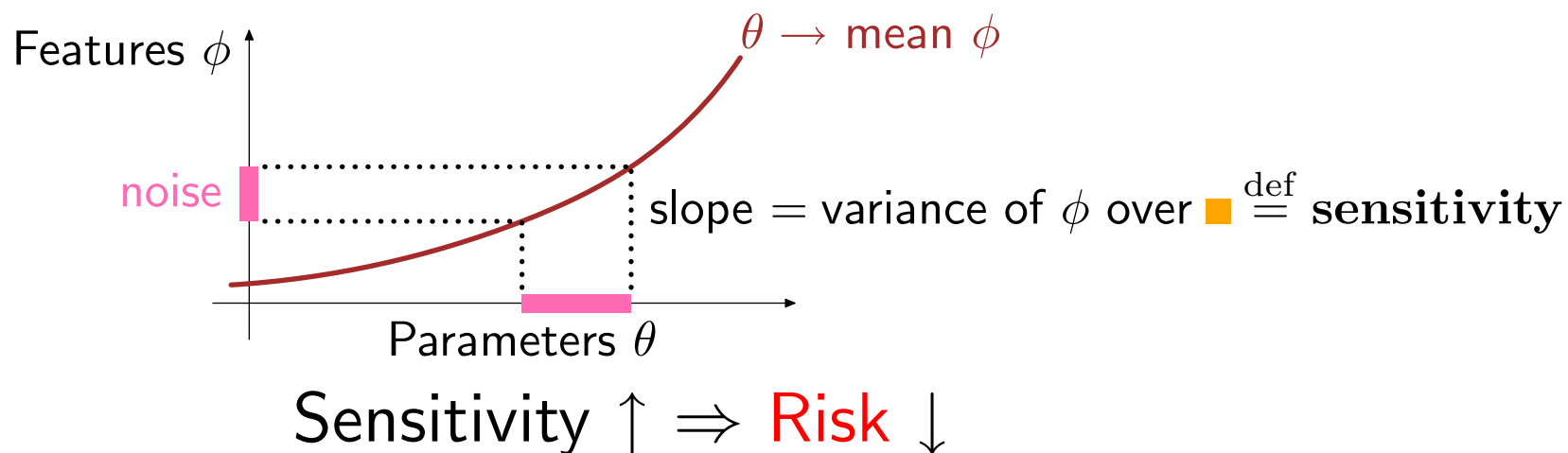
Sketch of arguments for comparing estimators



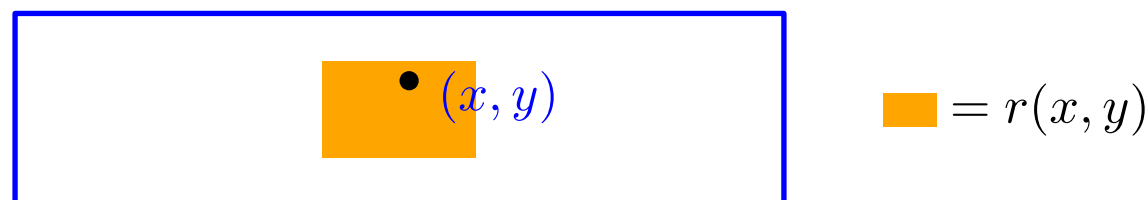
Intuition:

Grow $r \Rightarrow$ model more about data
 \Rightarrow data tells us more about parameters

For exponential families:



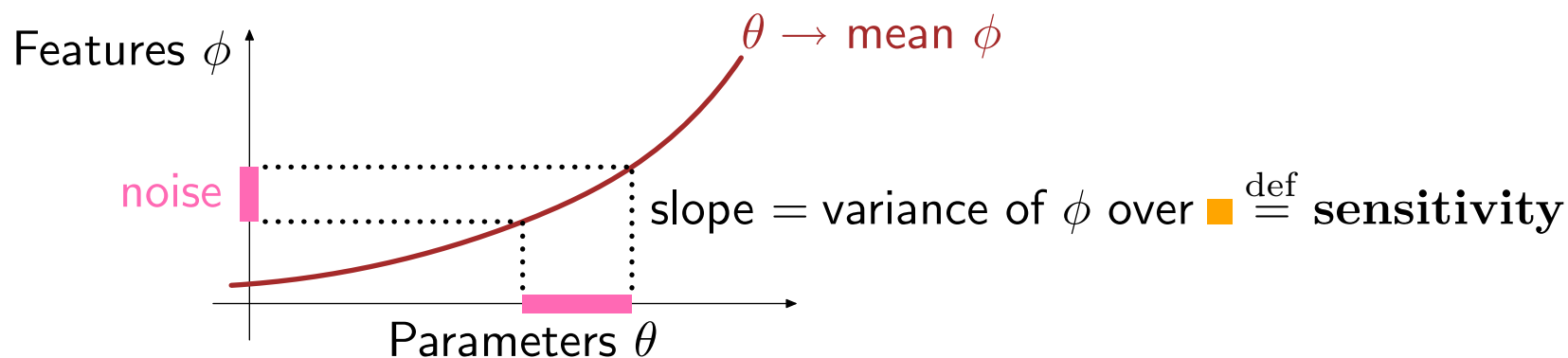
Sketch of arguments for comparing estimators



Intuition:

Grow $r \Rightarrow$ model more about data
 \Rightarrow data tells us more about parameters

For exponential families:



Sensitivity $\uparrow \Rightarrow$ Risk \downarrow

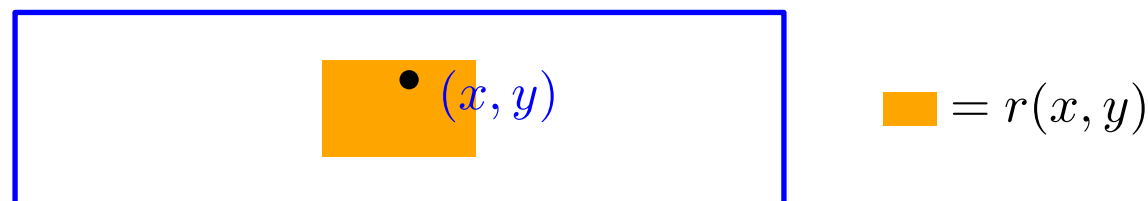
Generative

$\text{var}(\phi)$?

Discriminative

$\mathbb{E} \text{var}(\phi | X)$

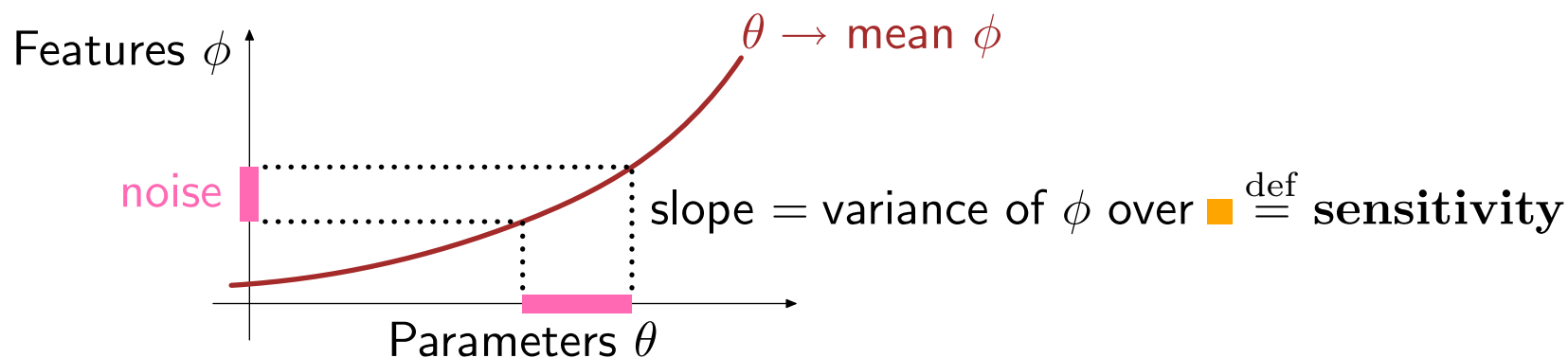
Sketch of arguments for comparing estimators



Intuition:

Grow $r \Rightarrow$ model more about data
 \Rightarrow data tells us more about parameters

For exponential families:



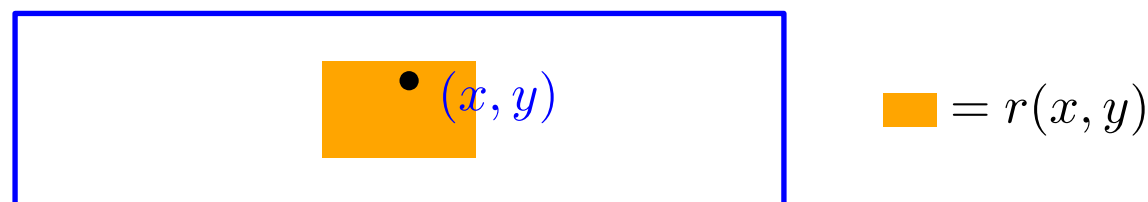
Sensitivity $\uparrow \Rightarrow$ Risk \downarrow

Generative

Discriminative

$$\text{var}(\phi) = \mathbb{E} \text{var}(\phi | X) + \text{var} \mathbb{E}(\phi | X)$$

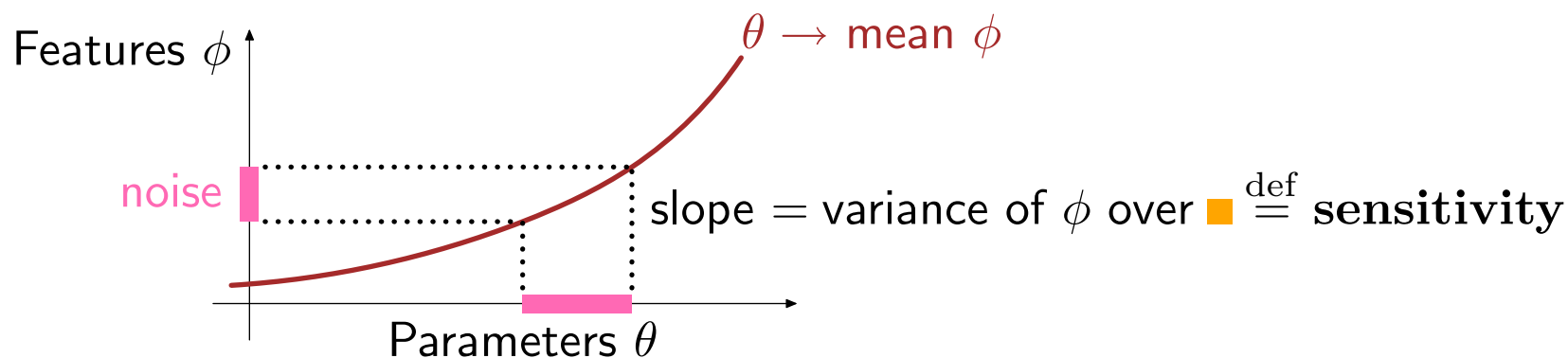
Sketch of arguments for comparing estimators



Intuition:

Grow $r \Rightarrow$ model more about data
 \Rightarrow data tells us more about parameters

For exponential families:



Sensitivity $\uparrow \Rightarrow$ Risk \downarrow

Generative

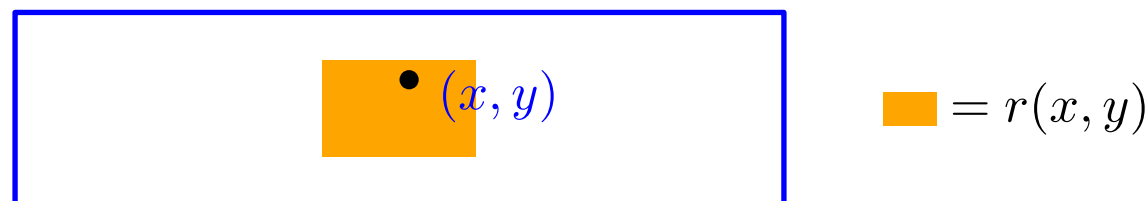
$$\text{var}(\phi)$$

\succ

Discriminative

$$\mathbb{E} \text{var}(\phi | X)$$

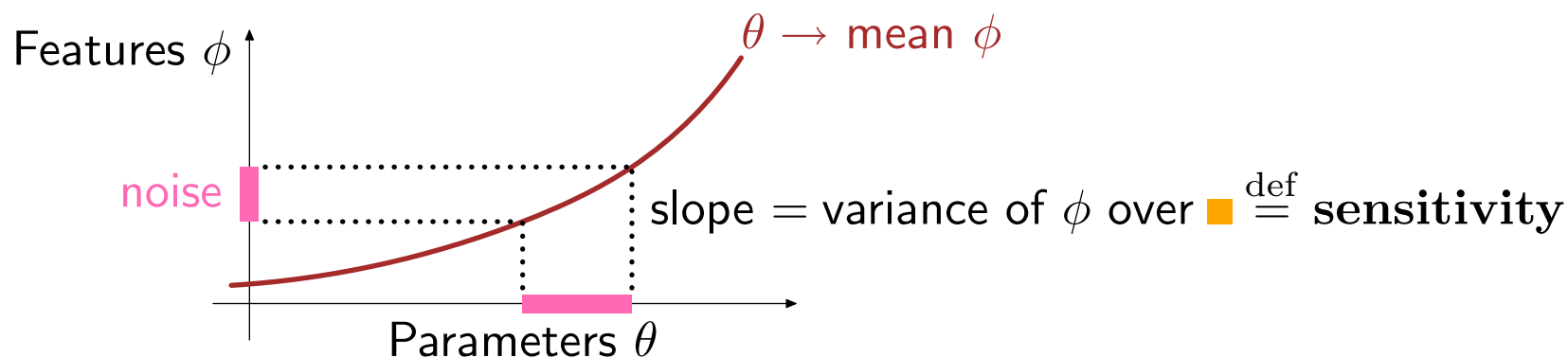
Sketch of arguments for comparing estimators



Intuition:

Grow $r \Rightarrow$ model more about data
 \Rightarrow data tells us more about parameters

For exponential families:



Sensitivity $\uparrow \Rightarrow$ Risk \downarrow

Generative Discriminative

$\text{var}(\phi)$ $\mathbb{E} \text{var}(\phi | X)$

Risk(generative) $<$ Risk(discriminative)

Overview of asymptotic analysis

How accurately can we estimate the parameters?

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

Σ : asymptotic variance of parameters
 n : number of training examples

Overview of asymptotic analysis

How accurately can we estimate the parameters?

$$\text{ParameterError} = O\left(\frac{\Sigma}{\sqrt{n}}\right) \quad \begin{array}{l} \Sigma: \text{asymptotic variance of parameters} \\ n: \text{number of training examples} \end{array}$$

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?

$$\text{ParameterError} = O\left(\frac{\Sigma}{\sqrt{n}}\right) \quad \begin{array}{l} \Sigma: \text{asymptotic variance of parameters} \\ n: \text{number of training examples} \end{array}$$

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?

$$\text{ParameterError} = O\left(\frac{\Sigma}{\sqrt{n}}\right) \quad \begin{array}{l} \Sigma: \text{asymptotic variance of parameters} \\ n: \text{number of training examples} \end{array}$$

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 Σ

Overview of asymptotic analysis

How accurately can we estimate the parameters?

$$\text{ParameterError} = O\left(\frac{\Sigma}{\sqrt{n}}\right) \quad \begin{array}{l} \Sigma: \text{asymptotic variance of parameters} \\ n: \text{number of training examples} \end{array}$$

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 Σ

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 Σ s of different estimators...

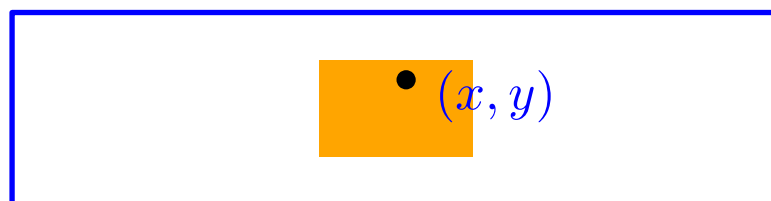
Well-specified case

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

Thus, sufficient to just compare Σ s of different estimators...

Estimator:

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



■ = $r(x, y)$

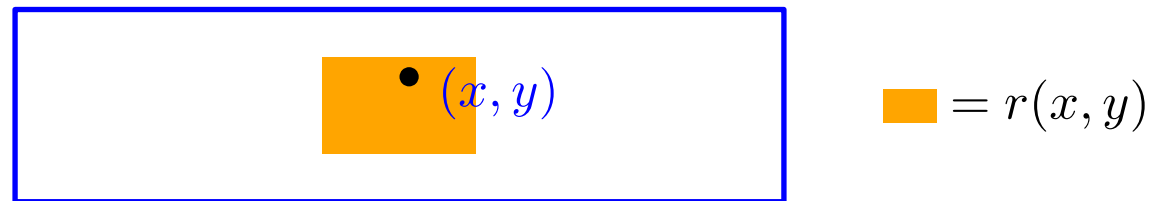
Well-specified case

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

Thus, sufficient to just compare Σ s of different estimators...

Estimator:

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



Asymptotic variance:

$\Sigma = \Gamma^{-1}$, where $\Gamma = \mathbb{E} \operatorname{var}(\phi \mid r)$ is the sensitivity

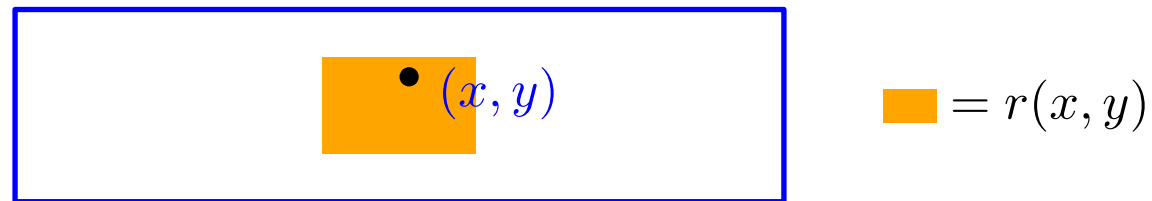
Well-specified case

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

Thus, sufficient to just compare Σ s of different estimators...

Estimator:

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



Asymptotic variance:

$$\Sigma = \Gamma^{-1}, \text{ where } \Gamma = \mathbb{E} \operatorname{var}(\phi \mid r) \text{ is the sensitivity}$$

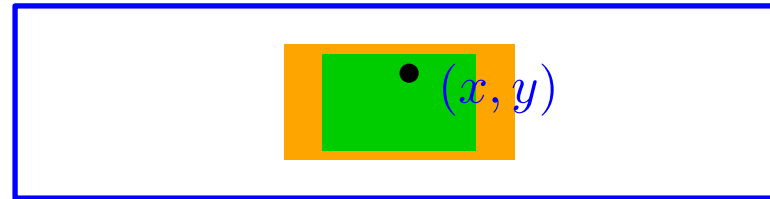
Proof:

By Taylor expansion and moment-generating properties.

Well-specified case: comparing two estimators

Two estimators:

$$\hat{\theta}_j = \operatorname{argmax}_{\theta} \hat{\mathbb{E}}[\log p_{\theta}(x, y) - \log p_{\theta}(r_j(x, y))] \quad \text{for } j = 1, 2$$

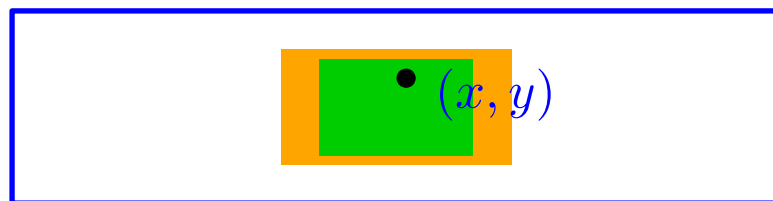


Orange square = $r_1(x, y)$
Green square = $r_2(x, y)$

Well-specified case: comparing two estimators

Two estimators:

$$\hat{\theta}_j = \operatorname{argmax}_{\theta} \hat{\mathbb{E}}[\log p_{\theta}(x, y) - \log p_{\theta}(r_j(x, y))] \quad \text{for } j = 1, 2$$



■ = $r_1(x, y)$
■ = $r_2(x, y)$

Comparison theorem:

If model is well-specified and

$$r_1(x, y) \supset r_2(x, y)$$

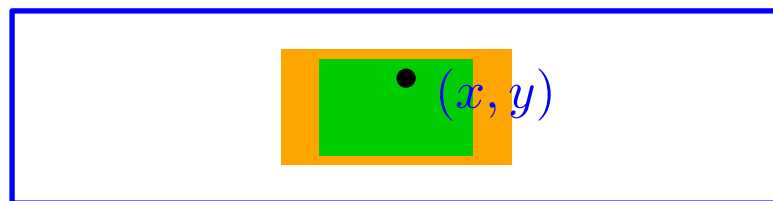
Then

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

Well-specified case: comparing two estimators

Two estimators:

$$\hat{\theta}_j = \operatorname{argmax}_{\theta} \hat{\mathbb{E}}[\log p_{\theta}(x, y) - \log p_{\theta}(r_j(x, y))] \quad \text{for } j = 1, 2$$



■ = $r_1(x, y)$
■ = $r_2(x, y)$

Comparison theorem:

If model is well-specified and

$$r_1(x, y) \supset r_2(x, y)$$

Then

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

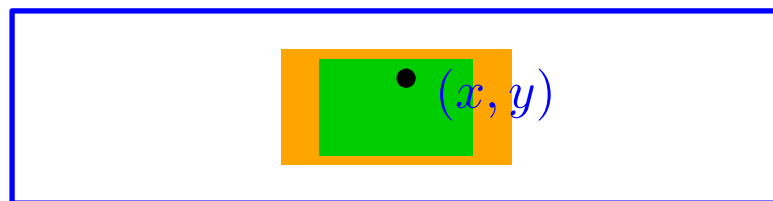
Proof:

$$\Sigma_j = \mathbb{E} \operatorname{var}(\phi \mid r_j)^{-1}$$

Well-specified case: comparing two estimators

Two estimators:

$$\hat{\theta}_j = \operatorname{argmax}_{\theta} \hat{\mathbb{E}}[\log p_{\theta}(x, y) - \log p_{\theta}(r_j(x, y))] \quad \text{for } j = 1, 2$$



Orange square = $r_1(x, y)$
Green square = $r_2(x, y)$

Comparison theorem:

If model is well-specified and

$$r_1(x, y) \supset r_2(x, y)$$

Then

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

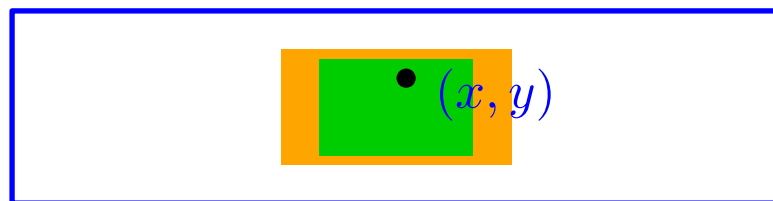
Proof:

$$\Sigma_j = \mathbb{E} \operatorname{var}(\phi \mid r_j)^{-1} \quad \Sigma_1 \preceq \Sigma_2$$

Well-specified case: comparing two estimators

Two estimators:

$$\hat{\theta}_j = \operatorname{argmax}_{\theta} \hat{\mathbb{E}}[\log p_{\theta}(x, y) - \log p_{\theta}(r_j(x, y))] \quad \text{for } j = 1, 2$$



Orange square = $r_1(x, y)$
Green square = $r_2(x, y)$

Comparison theorem:

If model is well-specified and

$$r_1(x, y) \supset r_2(x, y)$$

Then

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

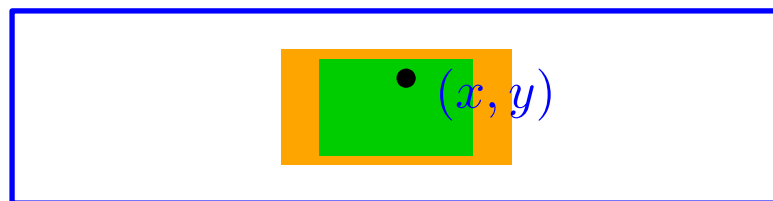
Proof:

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

Well-specified case: comparing two estimators

Two estimators:

$$\hat{\theta}_j = \operatorname{argmax}_{\theta} \hat{\mathbb{E}}[\log p_{\theta}(x, y) - \log p_{\theta}(r_j(x, y))] \quad \text{for } j = 1, 2$$



Orange square = $r_1(x, y)$
Green square = $r_2(x, y)$

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:

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

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:

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

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

Multiple components

Asymptotic variance:

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

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

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

Multiple components

Asymptotic variance:

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

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

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

Comparison theorem:

If the model is well-specified and

$\hat{\theta}_1$: one component r_1 $\hat{\theta}_2$: 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

$\hat{\theta}_1$: one component r_1 $\hat{\theta}_2$: multiple components $\{r_{2,j}\}$

$r_1(x, y) \supset r_{2,j}(x, y)$ 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

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

Comparison theorem:

If the model is well-specified and

$\hat{\theta}_1$: one component r_1 $\hat{\theta}_2$: multiple components $\{r_{2,j}\}$

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

Then

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

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

$\hat{\theta}_1$: one component r_1 $\hat{\theta}_2$: 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:

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

Misspecified case

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 **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 **discriminative estimator**, get fast rate:

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

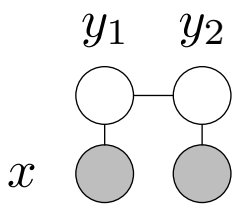
Corollary:

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

Key desirable property: training criterion = test criterion

Verifying the error rates empirically

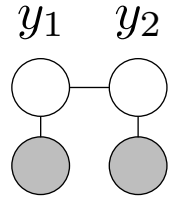
Setup:

Learn  from n training examples

Estimate (excess) risk from 10,000 trials

Verifying the error rates empirically

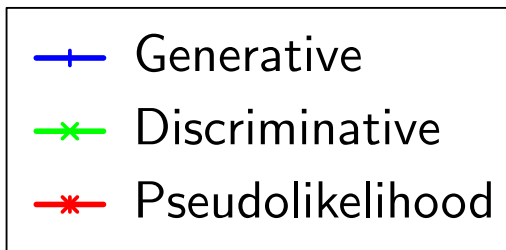
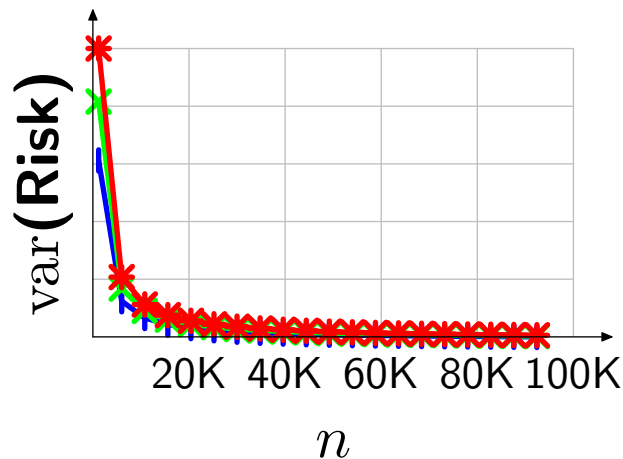
Setup:

Learn x  from n training examples

Estimate (excess) risk from 10,000 trials

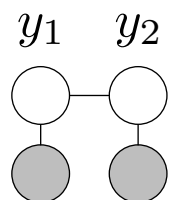
Well-specified

generate from 



Verifying the error rates empirically

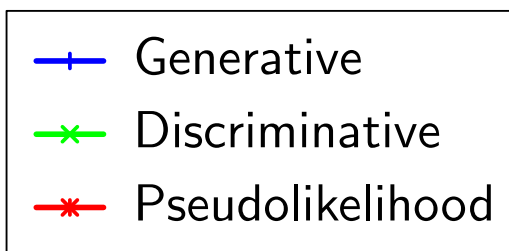
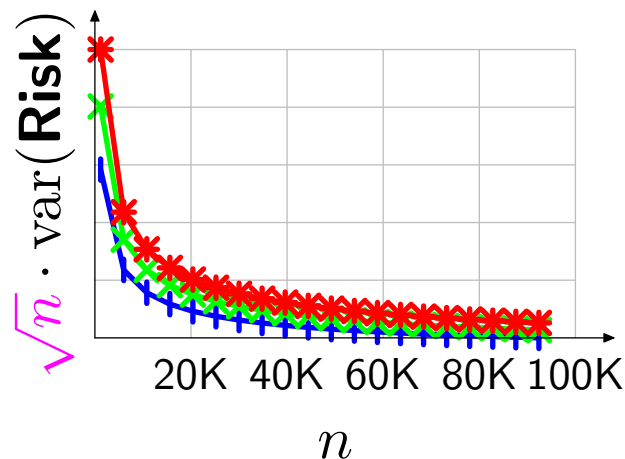
Setup:

Learn x  from n training examples

Estimate (excess) risk from 10,000 trials

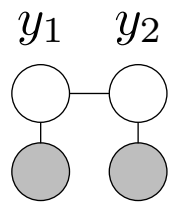
Well-specified

generate from 



Verifying the error rates empirically

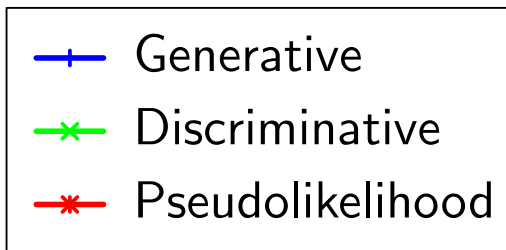
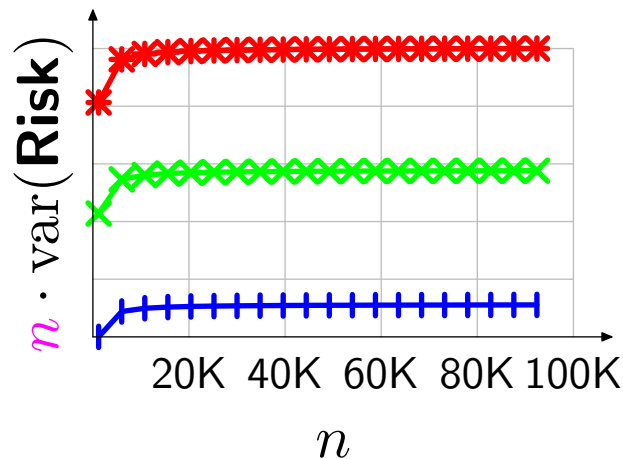
Setup:

Learn x  from n training examples

Estimate (excess) risk from 10,000 trials

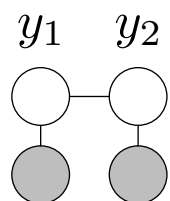
Well-specified

generate from 



Verifying the error rates empirically

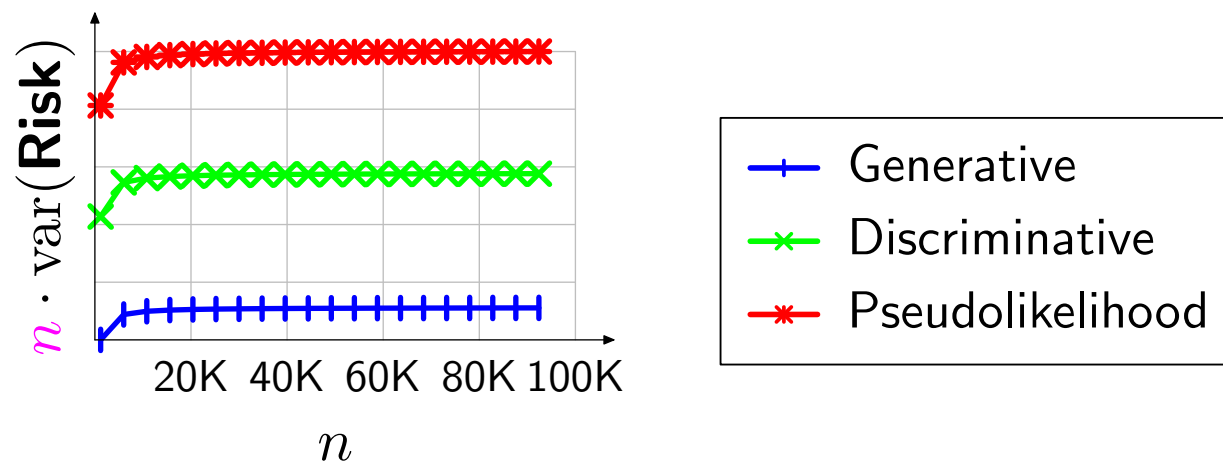
Setup:

Learn x  from n training examples

Estimate (excess) risk from 10,000 trials

Well-specified

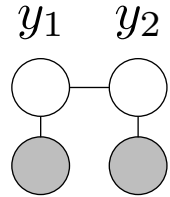
generate from 



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

Verifying the error rates empirically

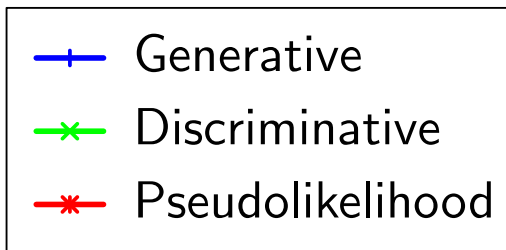
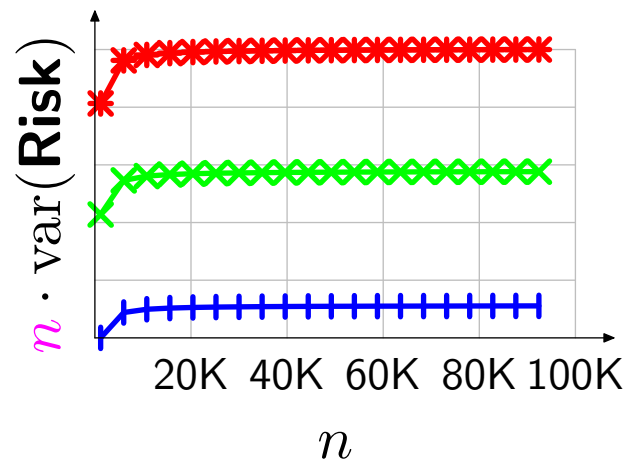
Setup:

Learn x  from n training examples

Estimate (excess) risk from 10,000 trials

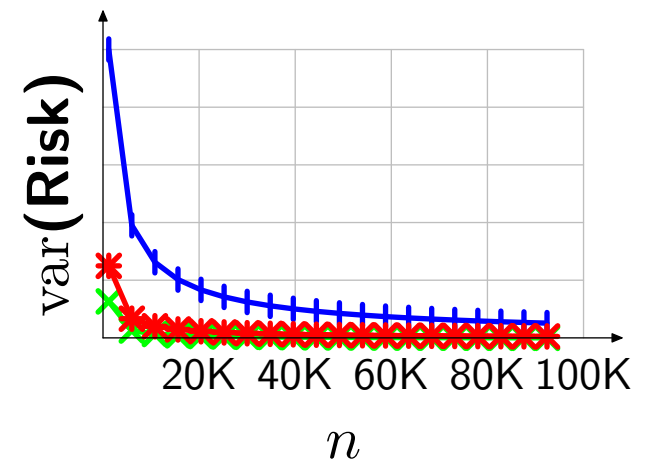
Well-specified

generate from 



Misspecified

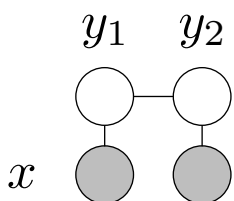
generate from 



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

Verifying the error rates empirically

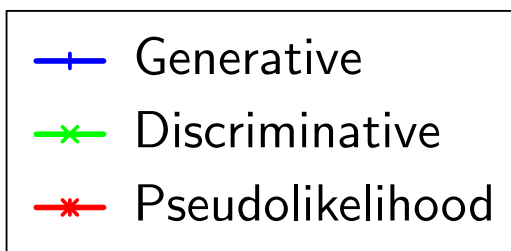
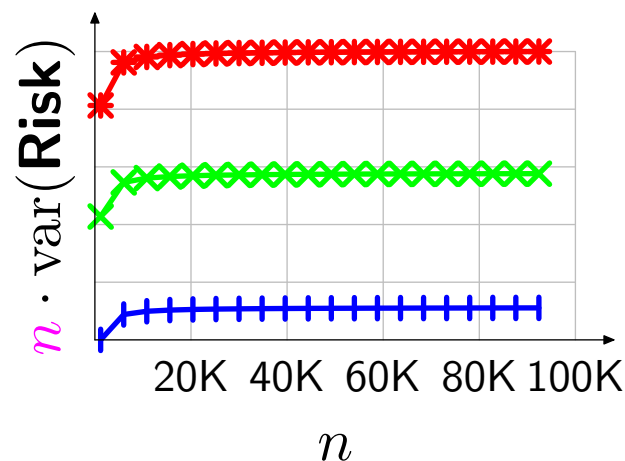
Setup:

Learn x  from n training examples

Estimate (excess) risk from 10,000 trials

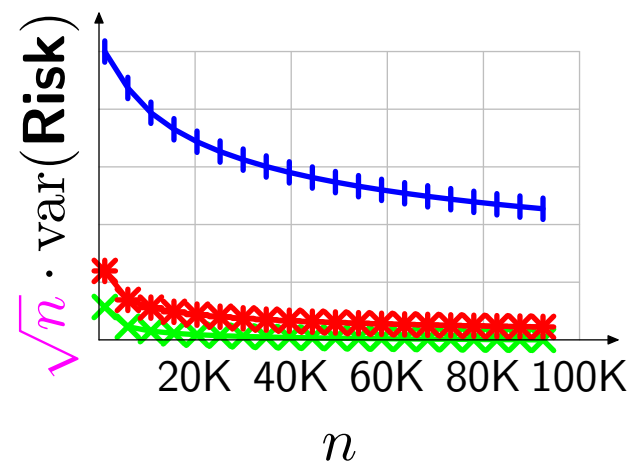
Well-specified

generate from 



Misspecified

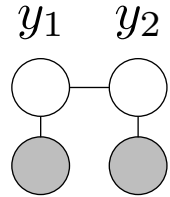
generate from 



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

Verifying the error rates empirically

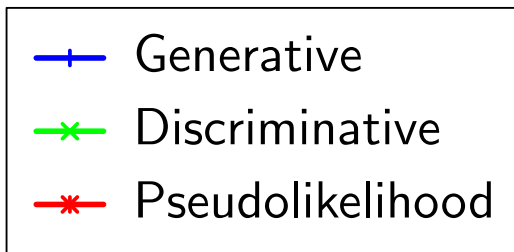
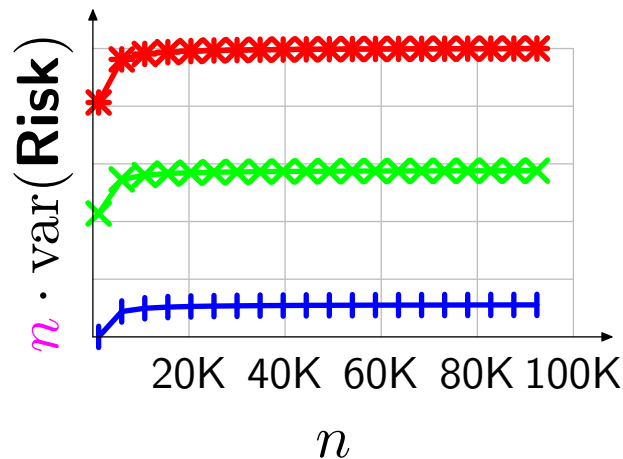
Setup:

Learn x  from n training examples

Estimate (excess) risk from 10,000 trials

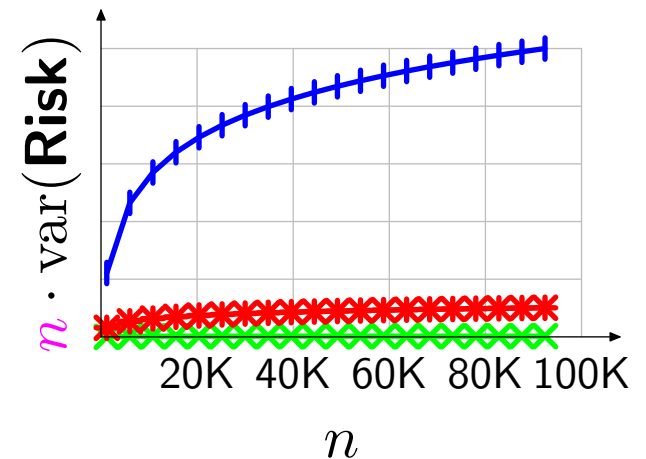
Well-specified

generate from 



Misspecified

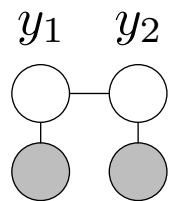
generate from 



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

Verifying the error rates empirically

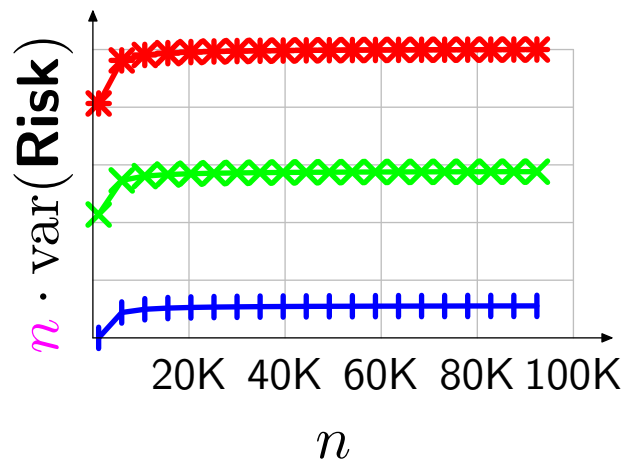
Setup:

Learn x  from n training examples

Estimate (excess) risk from 10,000 trials

Well-specified

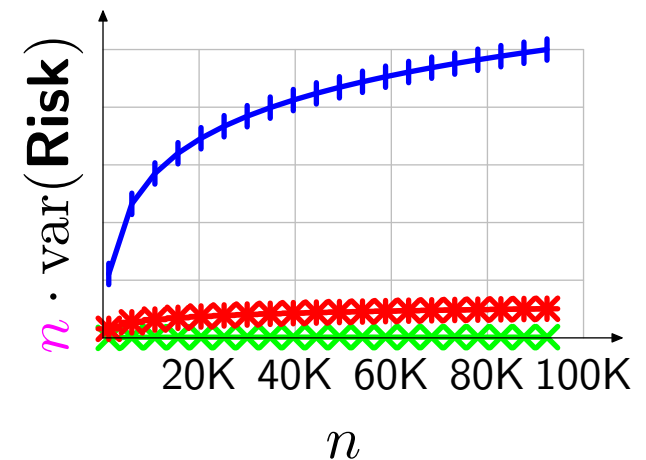
generate from 



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

Misspecified

generate from 



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

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

Application: part-of-speech tagging

Task:

y : Det — Noun — Verb — Det — Noun
 x : The cat ate a fish

Application: part-of-speech tagging

Task:

y : Det — Noun — Verb — Det — Noun
 x : The cat ate a fish

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

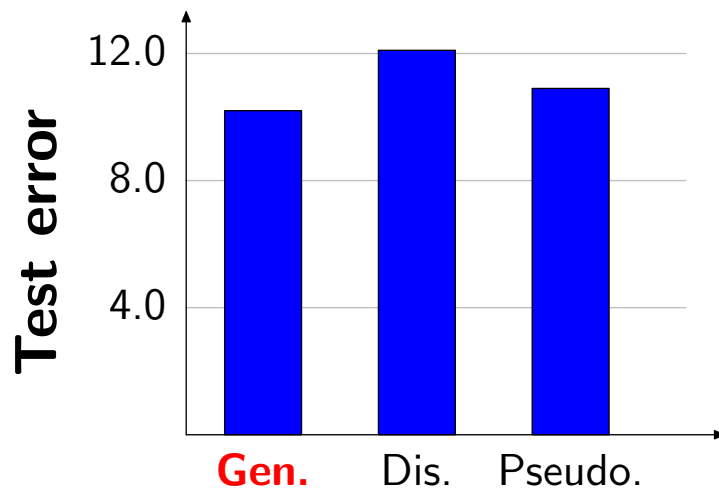
Application: part-of-speech tagging

Task:

y : Det — Noun — Verb — Det — Noun
 x : The cat ate a fish

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

Synthetic data (well-specified)



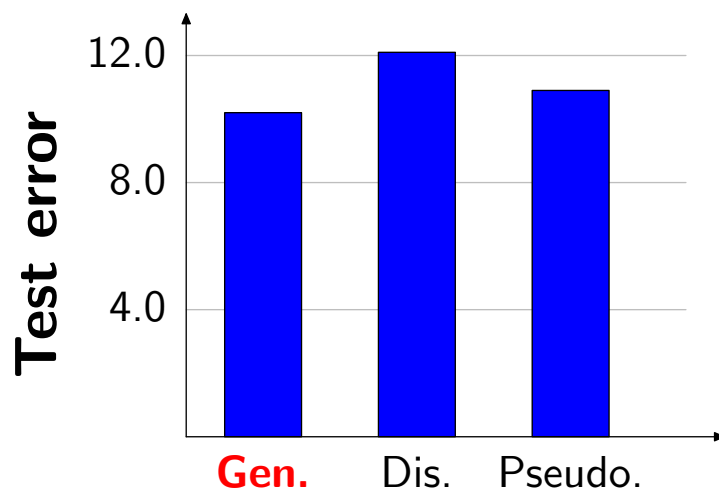
Application: part-of-speech tagging

Task:

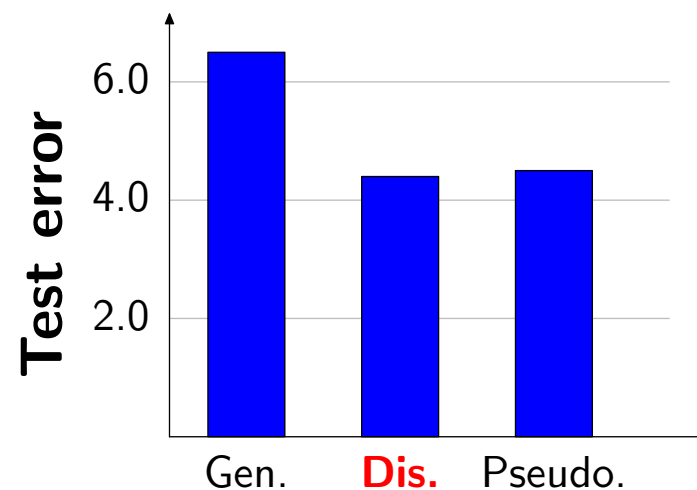
y : Det — Noun — Verb — Det — Noun
 x : The cat ate a fish

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

Synthetic data (well-specified)



Real data (misspecified)



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