Local Likelihood Modeling of Temporal Text Streams

Guy Lebanon and Yang Zhao

Statistics and ECE

Purdue University

Concept Drift

- Data $z = (x, y)$ is generated by a time-varying distribution

$$z^{(t)} \sim p_t \qquad t \in I \subset \mathbb{R}$$

  Assumption: Drift $t \mapsto p_t$ is continuous (or $\{p_t : t \in I\}$ is a continuous curve in the simplex)

- Learning task: approximate the drift $\{\hat{p}_t : t \in I\}$ based on data $z^{(t_1)}, \ldots, z^{(t_n)}$.
  - Generative: $\hat{p}_t(x, y) = \hat{p}_t(x|y)\hat{p}_t(y)$
  - Discriminative: $\hat{p}_t(x, y) = \hat{p}_t(y|x)\hat{p}_t(x)$

- News/Blog stories (x: story, y: category, t publication time, $t \mapsto p_t$: war, hunger, elections etc.)

Concept Drift

- Data $z = (x, y)$ is generated by a time-varying distribution

$$z^{(t)} \sim p_t \qquad t \in I \subset \mathbb{R}$$

  Assumption: Drift $t \mapsto p_t$ is continuous (or $\{p_t : t \in I\}$ is a continuous curve in the simplex)
- Learning task: approximate the drift $\{\hat{p}_t : t \in I\}$ based on data $z^{(t_1)}, \ldots, z^{(t_n)}$.
  - Generative: $\hat{p}_t(x, y) = \hat{p}_t(x|y)\hat{p}_t(y)$
  - Discriminative: $\hat{p}_t(x, y) = \hat{p}_t(y|x)\hat{p}_t(x)$
- News/Blog stories (x: story, y: category, t publication time, $t \mapsto p_t$: war, hunger, elections etc.)

Concept Drift

- Data $z = (x, y)$ is generated by a time-varying distribution

$$z^{(t)} \sim p_t \qquad t \in I \subset \mathbb{R}$$

  Assumption: Drift $t \mapsto p_t$ is continuous (or $\{p_t : t \in I\}$ is a continuous curve in the simplex)
- Learning task: approximate the drift $\{\hat{p}_t : t \in I\}$ based on data $z^{(t_1)}, \ldots, z^{(t_n)}$.
  - Generative: $\hat{p}_t(x, y) = \hat{p}_t(x|y)\hat{p}_t(y)$
  - Discriminative: $\hat{p}_t(x, y) = \hat{p}_t(y|x)\hat{p}_t(x)$
- News/Blog stories (x: story, y: category, t publication time, $t \mapsto p_t$: war, hunger, elections etc.)

Concept Drift

- Data $z = (x, y)$ is generated by a time-varying distribution

$$z^{(t)} \sim p_t \qquad t \in I \subset \mathbb{R}$$

  Assumption: Drift $t \mapsto p_t$ is continuous (or $\{p_t : t \in I\}$ is a continuous curve in the simplex)
- Learning task: approximate the drift $\{\hat{p}_t : t \in I\}$ based on data $z^{(t_1)}, \ldots, z^{(t_n)}$.
  - Generative: $\hat{p}_t(x, y) = \hat{p}_t(x|y)\hat{p}_t(y)$
  - Discriminative: $\hat{p}_t(x, y) = \hat{p}_t(y|x)\hat{p}_t(x)$
- News/Blog stories (x: story, y: category, t publication time, $t \mapsto p_t$: war, hunger, elections etc.)

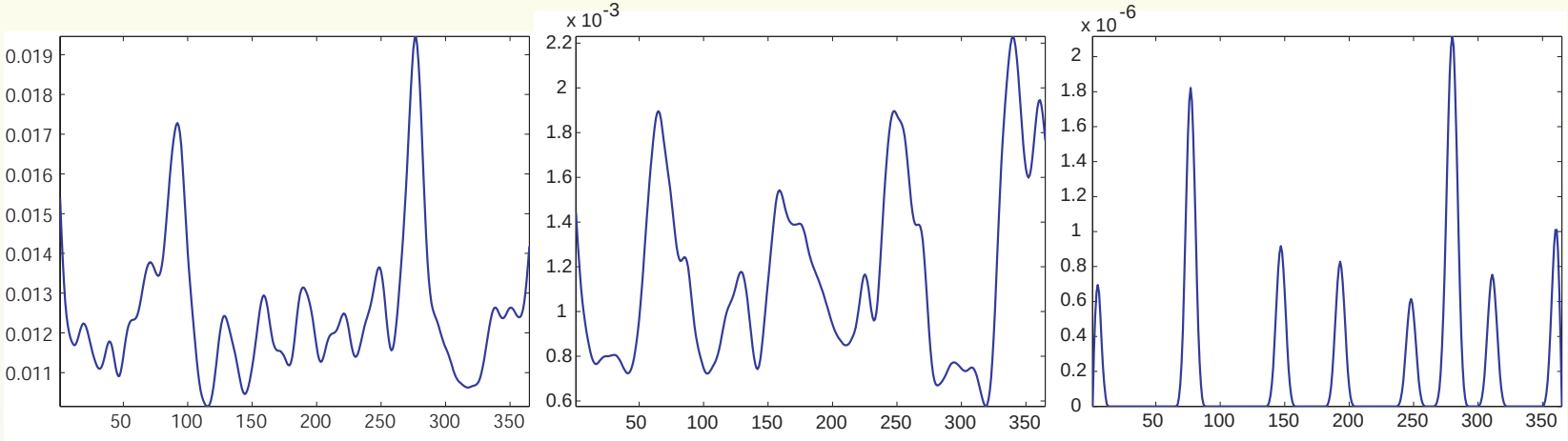
Concept Drift

- Data $z = (x, y)$ is generated by a time-varying distribution

$$z^{(t)} \sim p_t \qquad t \in I \subset \mathbb{R}$$

  Assumption: Drift $t \mapsto p_t$ is continuous (or $\{p_t : t \in I\}$ is a continuous curve in the simplex)
- Learning task: approximate the drift $\{\hat{p}_t : t \in I\}$ based on data $z^{(t_1)}, \ldots, z^{(t_n)}$.
  - Generative: $\hat{p}_t(x, y) = \hat{p}_t(x|y)\hat{p}_t(y)$
  - Discriminative: $\hat{p}_t(x, y) = \hat{p}_t(y|x)\hat{p}_t(x)$
- News/Blog stories (x: story, y: category, t publication time, $t \mapsto p_t$: war, hunger, elections etc.)

Concept Drift in $p_t(x|y)$



- Smoothed relative frequencies of `million`, `common`, and `Handelsgesellschaft` as a function of time (365 days) for the most popular category of RCV1.

Local Likelihood

- When the drift ($p_t(x|y)$ or $p_t(y|x)$)) is substantial, stationary likelihood methods are unlikely to produce an accurate model.
- Often, there is little or no knowledge regarding the nature of the drift making parametric drift modeling difficult.
- A useful non-parametric alternative that can be applied with no assumptions about p_t is local likelihood
  - guaranteed convergence $\hat{p}_t \longrightarrow p_t$ assuming smoothness of $t \mapsto p$ and $n \longrightarrow \infty$
  - works in both generative and discriminative setups
  - asymptotic expansion of mse enables theoretical analysis of performance and optimal bandwidth.

Local Likelihood

- When the drift ($p_t(x|y)$ or $p_t(y|x)$)) is substantial, stationary likelihood methods are unlikely to produce an accurate model.
- Often, there is little or no knowledge regarding the nature of the drift making parametric drift modeling difficult.
- A useful non-parametric alternative that can be applied with no assumptions about p_t is local likelihood
  - guaranteed convergence $\hat{p}_t \longrightarrow p_t$ assuming smoothness of $t \mapsto p$ and $n \longrightarrow \infty$
  - works in both generative and discriminative setups
  - asymptotic expansion of mse enables theoretical analysis of performance and optimal bandwidth.

Local Likelihood

- When the drift ($p_t(x|y)$ or $p_t(y|x)$)) is substantial, stationary likelihood methods are unlikely to produce an accurate model.
- Often, there is little or no knowledge regarding the nature of the drift making parametric drift modeling difficult.
- A useful non-parametric alternative that can be applied with no assumptions about p_t is local likelihood
  - guaranteed convergence $\hat{p}_t \longrightarrow p_t$ assuming smoothness of $t \mapsto p$ and $n \longrightarrow \infty$
  - works in both generative and discriminative setups
  - asymptotic expansion of mse enables theoretical analysis of performance and optimal bandwidth.

Local Likelihood

- When the drift ($p_t(x|y)$ or $p_t(y|x)$)) is substantial, stationary likelihood methods are unlikely to produce an accurate model.
- Often, there is little or no knowledge regarding the nature of the drift making parametric drift modeling difficult.
- A useful non-parametric alternative that can be applied with no assumptions about p_t is local likelihood
  - guaranteed convergence $\hat{p}_t \longrightarrow p_t$ assuming smoothness of $t \mapsto p$ and $n \longrightarrow \infty$
  - works in both generative and discriminative setups
  - asymptotic expansion of mse enables theoretical analysis of performance and optimal bandwidth.

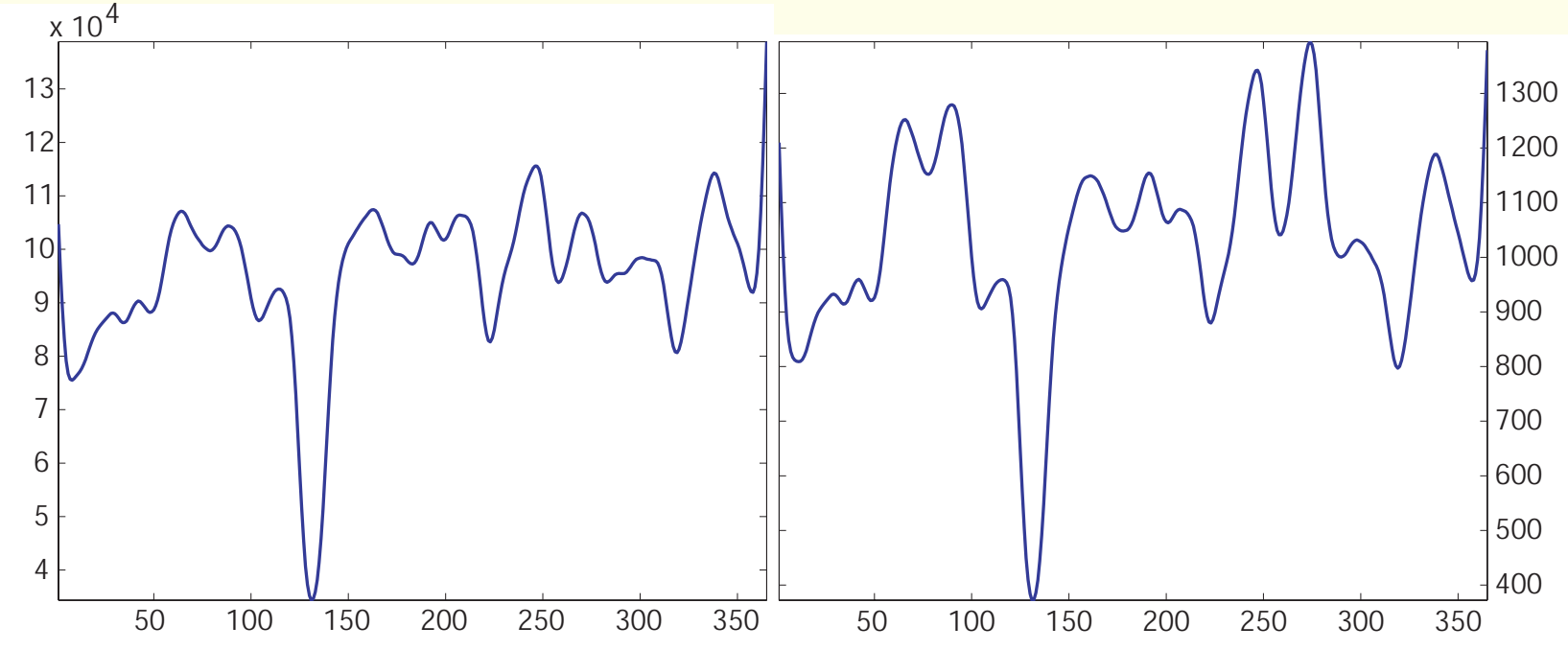
Local Likelihood

- When the drift ($p_t(x|y)$ or $p_t(y|x)$)) is substantial, stationary likelihood methods are unlikely to produce an accurate model.
- Often, there is little or no knowledge regarding the nature of the drift making parametric drift modeling difficult.
- A useful non-parametric alternative that can be applied with no assumptions about p_t is local likelihood
  - guaranteed convergence $\hat{p}_t \longrightarrow p_t$ assuming smoothness of $t \longmapsto p$ and $n \longrightarrow \infty$
  - works in both generative and discriminative setups
  - asymptotic expansion of mse enables theoretical analysis of performance and optimal bandwidth.

Local Likelihood

- When the drift ($p_t(x|y)$ or $p_t(y|x)$)) is substantial, stationary likelihood methods are unlikely to produce an accurate model.
- Often, there is little or no knowledge regarding the nature of the drift making parametric drift modeling difficult.
- A useful non-parametric alternative that can be applied with no assumptions about p_t is local likelihood
  - guaranteed convergence $\hat{p}_t \longrightarrow p_t$ assuming smoothness of $t \longmapsto p$ and $n \longrightarrow \infty$
  - works in both generative and discriminative setups
  - asymptotic expansion of mse enables theoretical analysis of performance and optimal bandwidth.

Sampling Density



Generative Process: $t \sim g(t)$

$(x, y) \sim p_t$

Document length does not depend on t. Total number of words per day (left) and documents per day (right) for the most popular category in RCV1.

Global and Local Models

- Extreme global model: ignore the temporal drift and use all of the training samples regardless of their time stamp.

  Single global model $\hat{p}$ representing the simplicial flow $\{p_t : t \in I\}$ as a degenerate curve

  Low variance but high bias

- Extreme local model: model p_t using only data sampled from time t

  Decomposes the concept drift into a sequence of disconnected estimation problems

  Low bias but high variance

- Sliding Window: model p_t using data sampled from neighboring time points $[t - \epsilon, t + \epsilon]$

Global and Local Models

- Extreme global model: ignore the temporal drift and use all of the training samples regardless of their time stamp.

  Single global model $\hat{p}$ representing the simplicial flow $\{p_t : t \in I\}$ as a degenerate curve

  Low variance but high bias
- Extreme local model: model p_t using only data sampled from time t

  Decomposes the concept drift into a sequence of disconnected estimation problems

  Low bias but high variance
- Sliding Window: model p_t using data sampled from neighboring time points $[t - \epsilon, t + \epsilon]$

Global and Local Models

- Extreme global model: ignore the temporal drift and use all of the training samples regardless of their time stamp.

  Single global model $\hat{p}$ representing the simplicial flow $\{p_t : t \in I\}$ as a degenerate curve

  Low variance but high bias

- Extreme local model: model p_t using only data sampled from time t

  Decomposes the concept drift into a sequence of disconnected estimation problems

  Low bias but high variance

- Sliding Window: model p_t using data sampled from neighboring time points $[t - \epsilon, t + \epsilon]$

Global and Local Models

- Extreme global model: ignore the temporal drift and use all of the training samples regardless of their time stamp.

  Single global model $\hat{p}$ representing the simplicial flow $\{p_t : t \in I\}$ as a degenerate curve

  Low variance but high bias

- Extreme local model: model p_t using only data sampled from time t

  Decomposes the concept drift into a sequence of disconnected estimation problems

  Low bias but high variance

- Sliding Window: model p_t using data sampled from neighboring time points $[t - \epsilon, t + \epsilon]$

Global and Local Models

- Extreme global model: ignore the temporal drift and use all of the training samples regardless of their time stamp.

  Single global model $\hat{p}$ representing the simplicial flow $\{p_t : t \in I\}$ as a degenerate curve

  Low variance but high bias

- Extreme local model: model p_t using only data sampled from time t

  Decomposes the concept drift into a sequence of disconnected estimation problems

  Low bias but high variance

- Sliding Window: model p_t using data sampled from neighboring time points $[t - \epsilon, t + \epsilon]$

Global and Local Models

- Extreme global model: ignore the temporal drift and use all of the training samples regardless of their time stamp.

  Single global model $\hat{p}$ representing the simplicial flow $\{p_t : t \in I\}$ as a degenerate curve

  Low variance but high bias
- Extreme local model: model p_t using only data sampled from time t

  Decomposes the concept drift into a sequence of disconnected estimation problems

  Low bias but high variance
- Sliding Window: model p_t using data sampled from neighboring time points $[t - \epsilon, t + \epsilon]$

Global and Local Models

- Extreme global model: ignore the temporal drift and use all of the training samples regardless of their time stamp.

  Single global model $\hat{p}$ representing the simplicial flow $\{p_t : t \in I\}$ as a degenerate curve

  Low variance but high bias
- Extreme local model: model p_t using only data sampled from time t

  Decomposes the concept drift into a sequence of disconnected estimation problems

  Low bias but high variance
- Sliding Window: model p_t using data sampled from neighboring time points $[t - \epsilon, t + \epsilon]$

Local Likelihood

- Generalizes the sliding window approach
- Intuition: Due to similarity between p_t and $p_{t+\epsilon}$, it makes sense to estimate p_t using samples from neighboring time points $t+\epsilon$. However, the contribution of $t+\epsilon$ should decrease as ϵ increases.
- Locally constant likelihood (standard extension to local linear/polynomial likelihood)

$$\ell_t(\theta|D) = \sum_{\tau \in I'} K_h(t-\tau) \sum_{j=1}^{N_\tau} \log p(x_{\tau j}; \theta).$$

- Maximum local likelihood estimator at each t: $\hat{p}_{\theta_t} = \arg\max_\theta \ell_t(\theta|D)$

Local Likelihood

- Generalizes the sliding window approach
- Intuition: Due to similarity between p_t and $p_{t+\epsilon}$, it makes sense to estimate p_t using samples from neighboring time points $t+\epsilon$. However, the contribution of $t+\epsilon$ should decrease as ϵ increases.
- Locally constant likelihood (standard extension to local linear/polynomial likelihood)

$$\ell_t(\theta|D) = \sum_{\tau \in I'} K_h(t-\tau) \sum_{j=1}^{N_\tau} \log p(x_{\tau j}; \theta).$$

- Maximum local likelihood estimator at each t: $\hat{p}_{\theta_t} = \arg\max_\theta \ell_t(\theta|D)$

Local Likelihood

- Generalizes the sliding window approach
- Intuition: Due to similarity between p_t and $p_{t+\epsilon}$, it makes sense to estimate p_t using samples from neighboring time points $t+\epsilon$. However, the contribution of $t+\epsilon$ should decrease as ϵ increases.
- Locally constant likelihood (standard extension to local linear/polynomial likelihood)

$$\ell_t(\theta|D) = \sum_{\tau \in I'} K_h(t-\tau) \sum_{j=1}^{N_\tau} \log p(x_{\tau j}; \theta).$$

- Maximum local likelihood estimator at each t: $\hat{p}_{\theta_t} = \arg\max_\theta \ell_t(\theta|D)$

Local Likelihood

- Generalizes the sliding window approach
- Intuition: Due to similarity between p_t and $p_{t+\epsilon}$, it makes sense to estimate p_t using samples from neighboring time points $t+\epsilon$. However, the contribution of $t+\epsilon$ should decrease as ϵ increases.
- Locally constant likelihood (standard extension to local linear/polynomial likelihood)

$$\ell_t(\theta|D) = \sum_{\tau \in I'} K_h(t-\tau) \sum_{j=1}^{N_\tau} \log p(x_{\tau j}; \theta).$$

- Maximum local likelihood estimator at each t: $\hat{p}_{\theta_t} = \arg\max_\theta \ell_t(\theta|D)$

Smoothing Kernel

- $K_h(r)$ is a normalized density concentrated around 0 and parameterized by a scale parameter $h > 0$ reflecting its spread and satisfying $K_h(r) = h^{-1}K(r/h)$
- K has bounded support and $\int u^r K(u)\, du < \infty$ for $r \leq 2$.
- Three popular kernel choices

$$K(r) = (1 - |r|^3)^3 \cdot 1_{\{|r|<1\}} \quad \text{tricube} \tag{1}$$

$$K(r) = (1 - |r|) \cdot 1_{\{|r|<1\}} \quad \text{triangular} \tag{2}$$

$$K(r) = 2^{-1} \cdot 1_{\{|r|<1\}} \quad \text{uniform (sliding window)} \tag{3}$$

- $K(r) = 0$ for $r \leq 0$: predict present given past (online). Otherwise predict present given past and future (offline)

Smoothing Kernel

- $K_h(r)$ is a normalized density concentrated around 0 and parameterized by a scale parameter $h > 0$ reflecting its spread and satisfying $K_h(r) = h^{-1}K(r/h)$
- K has bounded support and $\int u^r K(u)\, du < \infty$ for $r \leq 2$.
- Three popular kernel choices

$$K(r) = (1 - |r|^3)^3 \cdot 1_{\{|r|<1\}} \quad \text{tricube} \tag{1}$$

$$K(r) = (1 - |r|) \cdot 1_{\{|r|<1\}} \quad \text{triangular} \tag{2}$$

$$K(r) = 2^{-1} \cdot 1_{\{|r|<1\}} \quad \text{uniform (sliding window)} \tag{3}$$

- $K(r) = 0$ for $r \leq 0$: predict present given past (online). Otherwise predict present given past and future (offline)

Smoothing Kernel

- $K_h(r)$ is a normalized density concentrated around 0 and parameterized by a scale parameter $h > 0$ reflecting its spread and satisfying $K_h(r) = h^{-1}K(r/h)$
- K has bounded support and $\int u^r K(u)\, du < \infty$ for $r \leq 2$.
- Three popular kernel choices

$$K(r) = (1 - |r|^3)^3 \cdot 1_{\{|r|<1\}} \quad \text{tricube} \tag{1}$$

$$K(r) = (1 - |r|) \cdot 1_{\{|r|<1\}} \quad \text{triangular} \tag{2}$$

$$K(r) = 2^{-1} \cdot 1_{\{|r|<1\}} \quad \text{uniform (sliding window)} \tag{3}$$

- $K(r) = 0$ for $r \leq 0$: predict present given past (online). Otherwise predict present given past and future (offline)

Smoothing Kernel

- $K_h(r)$ is a normalized density concentrated around 0 and parameterized by a scale parameter $h > 0$ reflecting its spread and satisfying $K_h(r) = h^{-1}K(r/h)$
- K has bounded support and $\int u^r K(u)\, du < \infty$ for $r \leq 2$.
- Three popular kernel choices

$$K(r) = (1 - |r|^3)^3 \cdot 1_{\{|r|<1\}} \quad \text{tricube} \tag{1}$$

$$K(r) = (1 - |r|) \cdot 1_{\{|r|<1\}} \quad \text{triangular} \tag{2}$$

$$K(r) = 2^{-1} \cdot 1_{\{|r|<1\}} \quad \text{uniform (sliding window)} \tag{3}$$

- $K(r) = 0$ for $r \leq 0$: predict present given past (online). Otherwise predict present given past and future (offline)

Local Likelihood for n-Grams

- For n-gram model, local likelihood maximizer has closed form

$$n = 1 \qquad [\hat{\theta}_t]_w = \frac{\sum_{\tau \in I} K_h(t-\tau) \sum_{j=1}^{N_\tau} c(w, x_{\tau j})}{\sum_{\tau \in I} K_h(t-\tau) |x_\tau|}.$$

- Normalized linear combination of word counts (different from linear combination of relative frequencies)

Local Likelihood for n-Grams

- For n-gram model, local likelihood maximizer has closed form

$$n = 1 \qquad [\hat{\theta}_t]_w = \frac{\sum_{\tau \in I} K_h(t-\tau) \sum_{j=1}^{N_\tau} c(w, x_{\tau j})}{\sum_{\tau \in I} K_h(t-\tau) |x_\tau|}.$$

- Normalized linear combination of word counts (different from linear combination of relative frequencies)

Precise Bias Variance Analysis

Proposition

The bias vector and variance matrix of $\hat{\theta}_t$ in the 1-gram or multinomial case are

$$bias(\hat{\theta}_t) = \frac{\sum_{\tau \in I} K_h(t-\tau)|x_\tau| \, (\theta_\tau - \theta_t)}{\sum_{\tau \in I} K_h(t-\tau)|x_\tau|}$$

$$Var(\hat{\theta}_t) = \frac{\sum_{\tau \in I} K_h^2(t-\tau)|x_\tau| \, (diag(\theta_\tau) - \theta_\tau \theta_\tau^\top)}{\left(\sum_{\tau \in I} K_h(t-\tau)|x_\tau|\right)^2}$$

where $diag(z)$ is the diagonal matrix $[diag(z)]_{ij} = \delta_{ij} z_i$.

Connection between bias and variance and $\theta_t, \dot{\theta}_t, \lambda, K, h, g$ etc. not very informative

Asymptotic Bias Variance Analysis

Proposition

Assuming (i) θ_t, g are smooth in t, (ii) $h \to 0, hn \to \infty$, (iii) $g > 0$ in a neighborhood of t, (iv) $\mu_{kl}(K) = \int u^k K^l(u)\, du < \infty$, and (v) document lengths have expectation λ indep. of t, we have in the offline case

$$bias(\hat{\theta}_t|I) = h^2 \mu_{21}(K) \left(\dot{\theta}_t \frac{g'(t)}{g(t)} + \frac{1}{2} \ddot{\theta}_t \right) + o_P(h^2)$$

$$Var(\hat{\theta}_t|I) = \frac{\mu_{02}(K)}{(nh)g(t)\lambda} (diag(\theta_t) - \theta_t \theta_t^\top) + o_P((nh)^{-1})$$

and in the online case

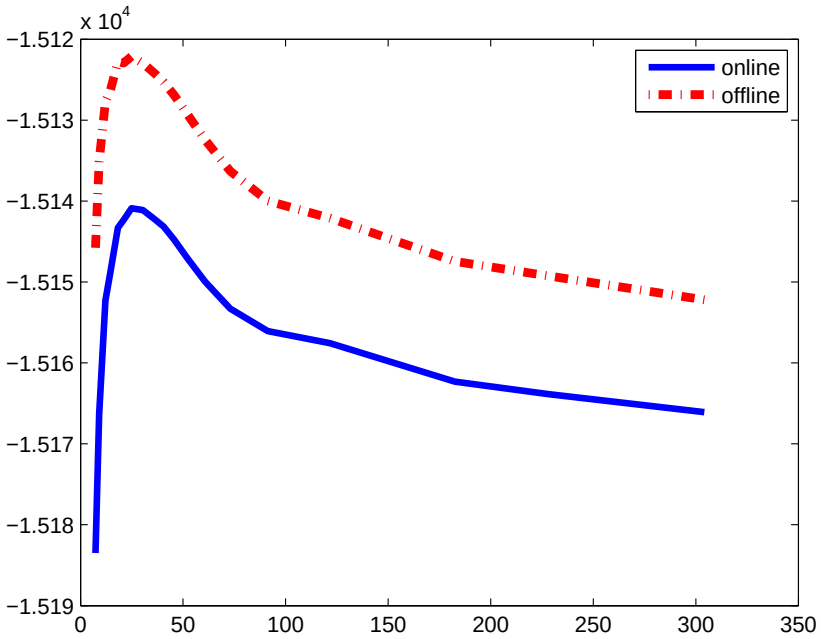
$$bias(\hat{\theta}_t|I) = h\mu_{11}(K)\dot{\theta}_t + o_P(h)$$

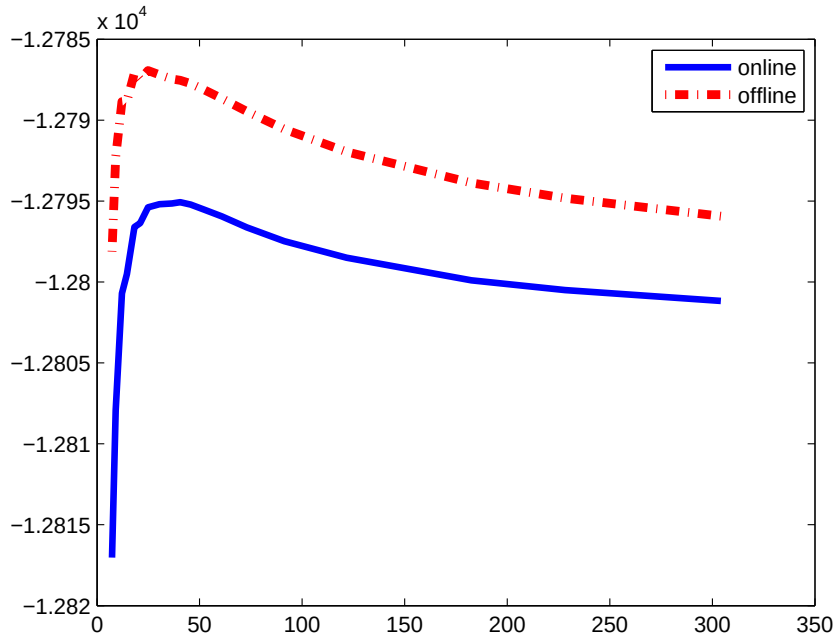
$$Var(\hat{\theta}_t|I) = \left(\frac{\mu_{02}(K)}{nhg(t)\lambda} + \frac{\mu_{12}(K)g'(t)}{ng^2(t)\lambda} \right) (diag(\theta_t) - \theta_t \theta_t^\top)$$
$$+ \frac{\mu_{12}(K)}{n\lambda g(t)} (diag(\dot{\theta}_t) - \dot{\theta}_t \theta_t^\top - \theta_t \dot{\theta}^\top) + o_P((nh)^{-1})$$

- Bias and variance converge to 0 as $h \to 0, nh \to \infty$ in both offline and online setting.
- Bias of online kernels converge at a linear rate rather the offline quadratic rate
- Asymptotic bias and variance reveal performance gain associated with slower drift $\dot{\theta}_t$ and $g'(t)$ and with more (represented by n) and longer (represented by λ) documents.

- Bias and variance converge to 0 as $h \to 0, nh \to \infty$ in both offline and online setting.
- Bias of online kernels converge at a linear rate rather the offline quadratic rate
- Asymptotic bias and variance reveal performance gain associated with slower drift $\dot{\theta}_t$ and $g'(t)$ and with more (represented by n) and longer (represented by λ) documents.

- Bias and variance converge to 0 as $h \to 0, nh \to \infty$ in both offline and online setting.
- Bias of online kernels converge at a linear rate rather the offline quadratic rate
- Asymptotic bias and variance reveal performance gain associated with slower drift $\dot{\theta}_t$ and $g'(t)$ and with more (represented by n) and longer (represented by λ) documents.





Test log-likelihood as a function of the triangular kernel's bandwidth for the two largest RCV1 categories (CCAT (left) and GCAT (right)). In all four cases, the optimal bandwidth seems to be a support of 25 days for the online kernels and 50 days for the offline kernels.

Bandwidth Selection

An important question is how to select h

- Minimizing leading term of asymptotic expansion of mse

$$\hat{h}_t^5 = \frac{\mu_{02}(K)\mathrm{tr}(\mathrm{diag}(\theta_t) - \theta_t\theta_t^\top)}{4n\lambda\mu_{21}^2(K)\sum_j \left([\dot{\theta}_t]_j\, g'(t)/\sqrt{g(t)} + \sqrt{g(t)}[\ddot{\theta}_t]_j/2\right)^2}$$

- As expected, the optimal bandwidth decreases as $n, \lambda, \|\dot{\theta}_t\|, \|\ddot{\theta}_t\|$ increases. In these cases the variance decreases and bias either increases or stays constant.
- In practice, $\dot{\theta}_t, \ddot{\theta}_t$ may vary significantly with time which leads to the conclusion that a single bandwidth selection for all t may not perform adequately.

Bandwidth Selection

An important question is how to select h

- Minimizing leading term of asymptotic expansion of mse

$$\hat{h}_t^5 = \frac{\mu_{02}(K)\mathrm{tr}(\mathrm{diag}(\theta_t) - \theta_t\theta_t^\top)}{4n\lambda\mu_{21}^2(K)\sum_j \left([\dot{\theta}_t]_j\, g'(t)/\sqrt{g(t)} + \sqrt{g(t)}[\ddot{\theta}_t]_j/2\right)^2}$$

- As expected, the optimal bandwidth decreases as $n, \lambda, \|\dot{\theta}_t\|, \|\ddot{\theta}_t\|$ increases. In these cases the variance decreases and bias either increases or stays constant.
- In practice, $\dot{\theta}_t, \ddot{\theta}_t$ may vary significantly with time which leads to the conclusion that a single bandwidth selection for all t may not perform adequately.

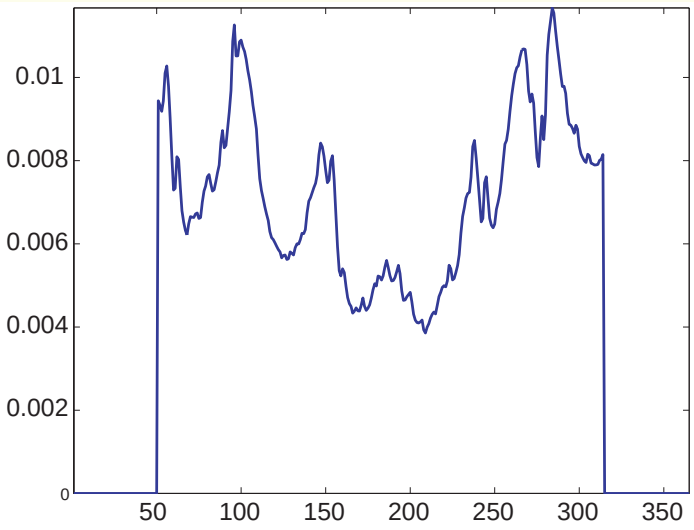
Bandwidth Selection

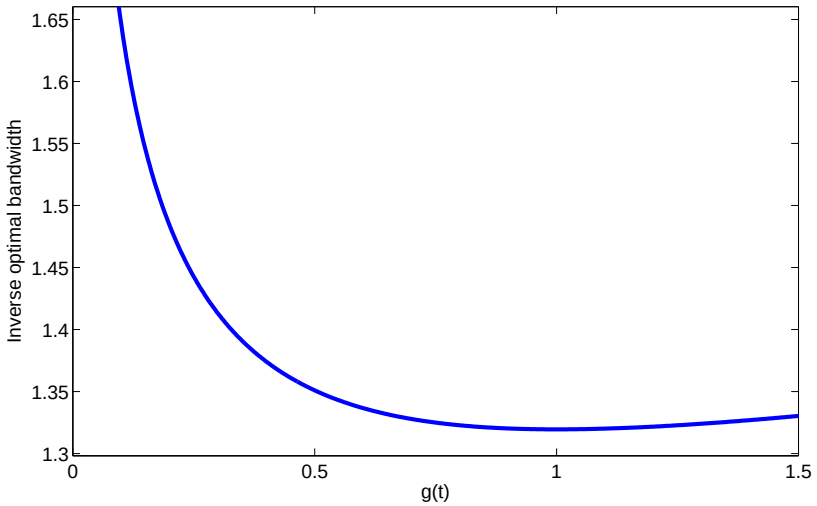
An important question is how to select h

- Minimizing leading term of asymptotic expansion of mse

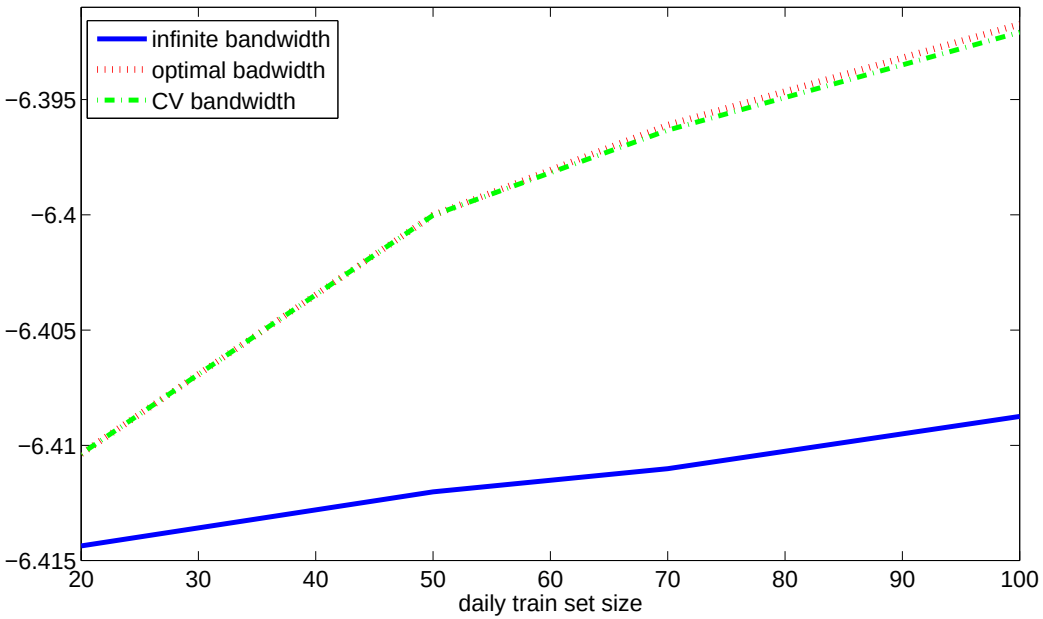
$$\hat{h}_t^5 = \frac{\mu_{02}(K)\mathrm{tr}(\mathrm{diag}(\theta_t) - \theta_t\theta_t^\top)}{4n\lambda\mu_{21}^2(K)\sum_j \left([\dot{\theta}_t]_j\, g'(t)/\sqrt{g(t)} + \sqrt{g(t)}[\ddot{\theta}_t]_j/2\right)^2}$$

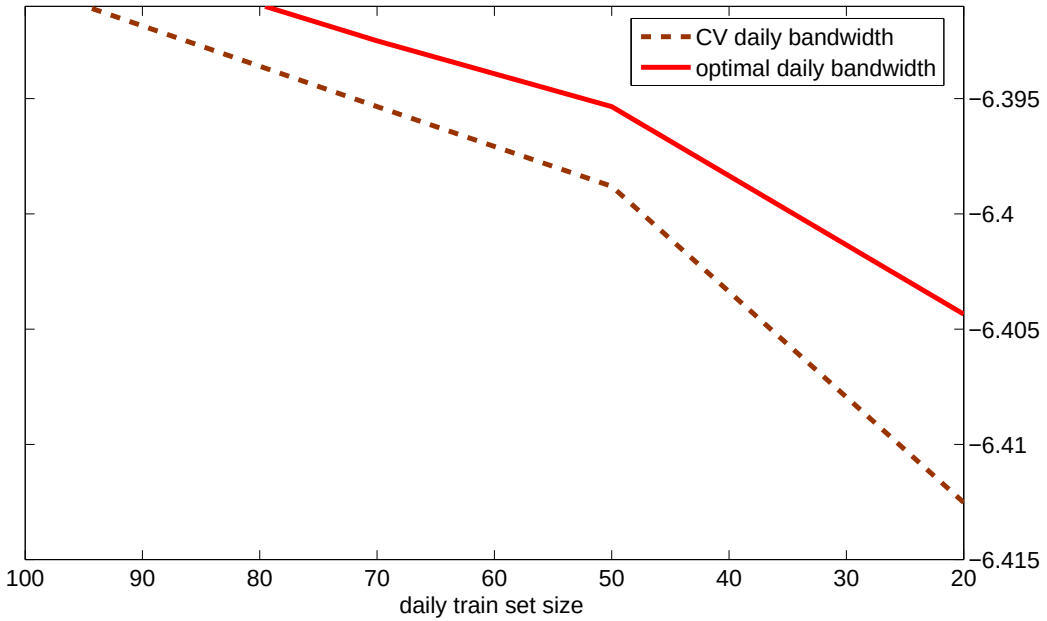
- As expected, the optimal bandwidth decreases as $n, \lambda, \|\dot{\theta}_t\|, \|\ddot{\theta}_t\|$ increases. In these cases the variance decreases and bias either increases or stays constant.
- In practice, $\dot{\theta}_t, \ddot{\theta}_t$ may vary significantly with time which leads to the conclusion that a single bandwidth selection for all t may not perform adequately.





Left: Estimated $\|\dot{\theta}_t\|$ as a function of t for RCV1 documents.
Right: Inverse of the optimal bandwidth as a function of $g(t)$.


infinite bandwidth
optimal badwidth
CV bandwidth
−6.395
−6.4
−6.405
−6.41
−6.415
20
30
40
50
60
70
80
90
100
daily train set size


CV daily bandwidth
optimal daily bandwidth
−6.395
−6.4
−6.405
−6.41
−6.415
100
90
80
70
60
50
40
30
20
daily train set size

Local Likelihood for Logistic Regression

- Local likelihood can also be applied to conditional modeling.

  In the case of logistic regression we have

$$\ell_t(\eta|D) = -\sum_{\tau \in I}^{n} K_h(t-\tau) \sum_{j=1}^{N_\tau} \log\left(1 + e^{-y_{\tau j}\langle x_{\tau j}, \eta\rangle}\right).$$

- Absence of closed form requires optimization for every t
- Convex problem leading to global optimum (for each t)

Local Likelihood for Logistic Regression

- Local likelihood can also be applied to conditional modeling. In the case of logistic regression we have

$$\ell_t(\eta|D) = -\sum_{\tau \in I}^{n} K_h(t-\tau) \sum_{j=1}^{N_\tau} \log\left(1 + e^{-y_{\tau j}\langle x_{\tau j}, \eta\rangle}\right).$$

- Absence of closed form requires optimization for every t
- Convex problem leading to global optimum (for each t)

Local Likelihood for Logistic Regression

- Local likelihood can also be applied to conditional modeling.
  In the case of logistic regression we have

$$\ell_t(\eta|D) = -\sum_{\tau \in I}^{n} K_h(t-\tau) \sum_{j=1}^{N_\tau} \log\left(1 + e^{-y_{\tau j}\langle x_{\tau j}, \eta\rangle}\right).$$

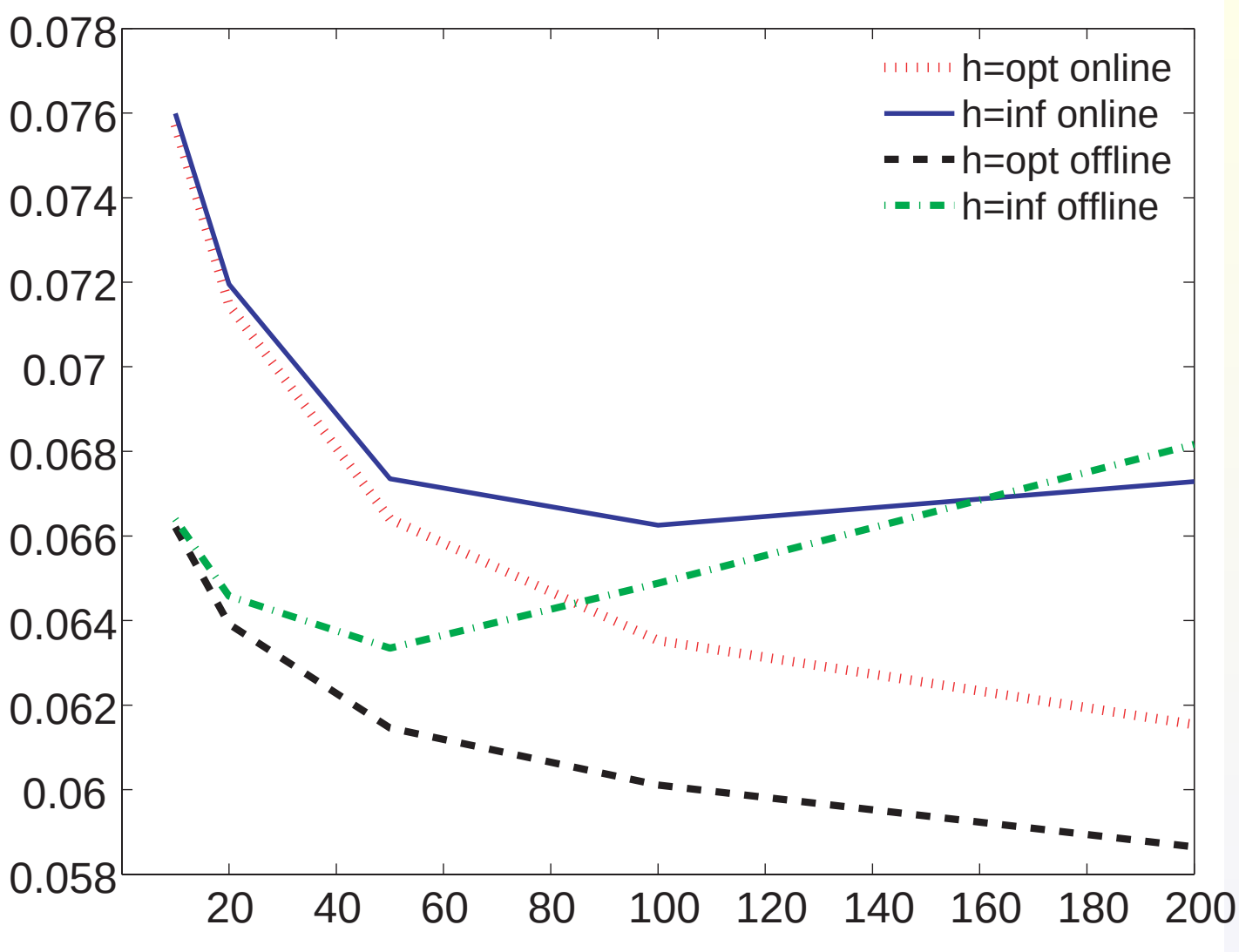
- Absence of closed form requires optimization for every t
- Convex problem leading to global optimum (for each t)

Local Likelihood for Logistic Regression

- Local likelihood can also be applied to conditional modeling.
  In the case of logistic regression we have

$$\ell_t(\eta|D) = -\sum_{\tau \in I}^{n} K_h(t-\tau) \sum_{j=1}^{N_\tau} \log\left(1 + e^{-y_{\tau j}\langle x_{\tau j}, \eta\rangle}\right).$$

- Absence of closed form requires optimization for every t
- Convex problem leading to global optimum (for each t)


0.078
0.076
0.074
0.072
0.07
0.068
0.066
0.064
0.062
0.06
0.058
20
40
60
80
100
120
140
160
180
200
h=opt online
h=inf online
h=opt offline
h=inf offline

Discussion

- Local likelihood techniques offer effective solution to the concept drift problem, in both the generative and discriminative settings.
- In the case of n-grams, closed form expressions leads to efficient computation
- Asymptotic expansions reveals complex dependencies between mse and $\dot{\theta}_t, \ddot{\theta}_t, n, \lambda, g(t), h$
- Experiments demonstrate that a single bandwidth is not appropriate for all times t and for all words.

Discussion

- Local likelihood techniques offer effective solution to the concept drift problem, in both the generative and discriminative settings.
- In the case of n-grams, closed form expressions leads to efficient computation
- Asymptotic expansions reveals complex dependencies between mse and $\dot{\theta}_t, \ddot{\theta}_t, n, \lambda, g(t), h$
- Experiments demonstrate that a single bandwidth is not appropriate for all times t and for all words.

Discussion

- Local likelihood techniques offer effective solution to the concept drift problem, in both the generative and discriminative settings.
- In the case of n-grams, closed form expressions leads to efficient computation
- Asymptotic expansions reveals complex dependencies between mse and $\dot{\theta}_t, \ddot{\theta}_t, n, \lambda, g(t), h$
- Experiments demonstrate that a single bandwidth is not appropriate for all times t and for all words.

Discussion

- Local likelihood techniques offer effective solution to the concept drift problem, in both the generative and discriminative settings.
- In the case of n-grams, closed form expressions leads to efficient computation
- Asymptotic expansions reveals complex dependencies between mse and $\dot{\theta}_t, \ddot{\theta}_t, n, \lambda, g(t), h$
- Experiments demonstrate that a single bandwidth is not appropriate for all times t and for all words.

Thank You!