Score-Based Change Detection for Gradient-Based Learning Machines

Lang Liu¹, Joseph Salmon², Zaid Harchaoui¹

¹ University of Washington  ² University of Montpellier

Overview
- The widespread use of machine learning algorithms calls for automatic change detection algorithms to monitor their behavior over time.
- We present a generic change monitoring method based on quantities amenable to be computed efficiently whenever the model is implemented in a differentiable programming framework.
- This method is equipped with a scanning procedure, allowing it to detect small jumps occurring on an unknown subset of model parameters.

Motivating Example
Microsoft’s chatbot Tay
- A chatbot that started to deliver hate speech within one day after it was released on Twitter.
- Initially learned language model quickly changed to an undesirable one, as it was being fed data through interactions with users.
- This phenomenon is prevalent and known as neural toxic degeneration in natural language processing (e.g., Gehman et al. 2020).

A potential strategy to prevent such a degeneration is to equip the language model with an automatic monitoring tool, which can trigger an early alarm before the model actually produces toxic content.

Change Detection

Model formulation.
- Data stream \( W_n = |W_{n-k+1}| \).
- Parametric model \( \{M_n, \theta \in \Theta \subseteq \mathbb{R}^\ell \} \) with unknown true value \( \theta_0 \).
- Maximum likelihood estimation:
  \[ \hat{\theta}_n = \arg \max_{\theta \in \Theta} \sum_{i=1}^n \log p(W_i | W_{i-k+1}) \]

Change detection. Consider the changepoint model \( W_n = M_{\theta_0}(W_{n-k+1}) + c_k \).
- A time point \( \tau \in [n-1] = \{1, \ldots, n-1\} \) is called a changepoint if there exists \( \Delta > 0 \) such that \( \hat{\theta}_\tau = \theta_0 \) for \( \tau < \tau \) and \( \hat{\theta}_\tau = \theta_0 + \Delta \) for \( \tau > \tau \).
- Testing the existence of a changepoint: \( H_0: \theta_0 = \theta_0 \) for all \( k \leq \tau \), \( H_1: \theta_k \) jumps from \( \theta_0 \) to \( \theta_0 + \Delta \)

Hypothesis testing. Fix a significance level \( \alpha \).
- 1. Propose a test statistic \( R = R(W_{\tau+1}) \); the larger \( R \) is, the less likely \( H_0 \) is true.
- 2. Calibrate \( R \) by a threshold \( H = H(\alpha) \), leading to a test \( \psi = 1 \{ H^{-1}(R) > 1 \} \).
- 3. False alarm rate: \( \lim_{n \to \infty} \mathbb{P}(\psi = 1 | H_0) \leq \alpha \).
- 4. Detection power: \( \lim_{n \to \infty} \mathbb{P}(\psi = 1 | H_1) = 1 \).

Score-Based Change Detection

Score-based testing. Let \( f_\theta(\delta, \Delta, \tau) \) be the log-likelihood under the alternative.
- Score function \( S_{\theta_n} = \nabla f_\theta(\theta_n, \Delta, \tau) | \Delta \theta_{\tau} = \delta \).
- Fisher information \( I_{\theta_n} = -\nabla^2 f_\theta(\theta_n, \Delta, \tau) | \Delta \theta_{\tau} = 0 \).
- Fixed \( \tau : R_n = S_{\theta_n}^T I_{\theta_n}^{-1} S_{\theta_n} \) is “close” to 0 under the null.
- Unknown \( \tau : R_n = \max_{\tau \in [n-1]} H^{-1}(\alpha) R_n, \text{ and } \psi(\alpha) = 1 \{ R_n > 1 \} \).

Small jumps. The change may only happen in a small subset of components of \( \theta_0 \). In such scenarios, the linear test can have low power.

Component screening.
- Truncated statistic \( R_{\alpha}(T) = \{ S_{\theta_n}^T I_{\theta_n}^{-1} S_{\theta_n} | \tau \in [T] \} \).
- \( R_{\alpha} = \max_{\tau \in [n-1], \ell \in \ell} H^{-1}(\alpha) R_{\alpha}(T) \) and \( \psi(\alpha) = 1 \{ R_{\alpha} > 1 \} \).

Auto-test. \( \psi(\alpha) = \max \{ \psi_{\alpha_1}(\alpha_1), \psi_{\alpha_2}(\alpha_2) \} \), with \( \alpha = \alpha_1 + \alpha_2 \).

Differentiable Programming

Auto-test only involves inverse-Hessian-vector products of the log-likelihood.

Naive strategy. Compute the full Hessian by \( \text{AutoDiff} \).

AutoDiff-friendly strategy.
- Compute the gradient \( S \) by a forward pass and save its computational graph.
- Compute inverse-Hessian-vector products by the conjugate gradient algorithm.

Running time. A linear model with \( d = 1000 \) (left) and \( n = 10000 \) (right).

Consistency

Level consistency. Under the null hypothesis and appropriate conditions, we have \( R_n \to \chi^2_{\alpha} \) and \( R_{\alpha}(T) \to \chi^2_{\alpha} \) for \( n \to \lambda \in (0, 1) \) and \( T \to \delta \).
- These conditions hold true in i.i.d. models, hidden Markov models, and stationary autoregressive moving-average models, provided regularity conditions.
- Valid choices of thresholds are \( H(\alpha) = q_{\alpha}(\mathbb{E}[\frac{\chi^2_{\alpha}}{p+p+1}]) \).

Power consistency. Under fixed alternatives and appropriate conditions, the three proposed tests \( \psi(\alpha) \), \( \psi_{\alpha_1}(\alpha_1) \), \( \psi_{\alpha_2}(\alpha_2) \) with above thresholds are consistent in power.

Experiments

Synthetic data. Up: linear model with \( d = 101 \) parameters and two sparsity levels \( p = 1 \) (left) and \( p = 20 \) (right).
- Bottom: text topic model (Stratos et al. 2015) with \( p = 1 \) and two model sizes \( d = 21 \) (left) and \( d = 175 \) (right).

Real data. We collect subtitles of the first two seasons of four TV shows—Friends (F), Modern Family (M), the Sopranos (S), and Deadwood (D).
- The former two are viewed as polite and the latter two are viewed as toxic.
- For each pair, we concatenate them, and use the aforementioned text topic model to detect changes in toxicity.
- False alarm rate for the linear test \( 27/32 \) and for the scan test \( 11/32 \).

<table>
<thead>
<tr>
<th>F1</th>
<th>F2</th>
<th>M1</th>
<th>M2</th>
<th>S1</th>
<th>S2</th>
<th>D1</th>
<th>D2</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>R</td>
<td>N</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>N</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>N</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>N</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>R</td>
<td>N</td>
</tr>
<tr>
<td>D1</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>R</td>
<td>N</td>
</tr>
<tr>
<td>D2</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>R</td>
<td>N</td>
</tr>
</tbody>
</table>

Presented at ICASSP 2021. Copyright 2021 by the author(s).