UNIVERSITY of WASHINGTON

Score-Based Change Detection for Gradient-Based Learning Machines

Lang Liu[1], Joseph Salmon[2], Zaid Harchaoui[1]

[1] University of Washington, Seattle [2] University of Montpellier, Montpellier

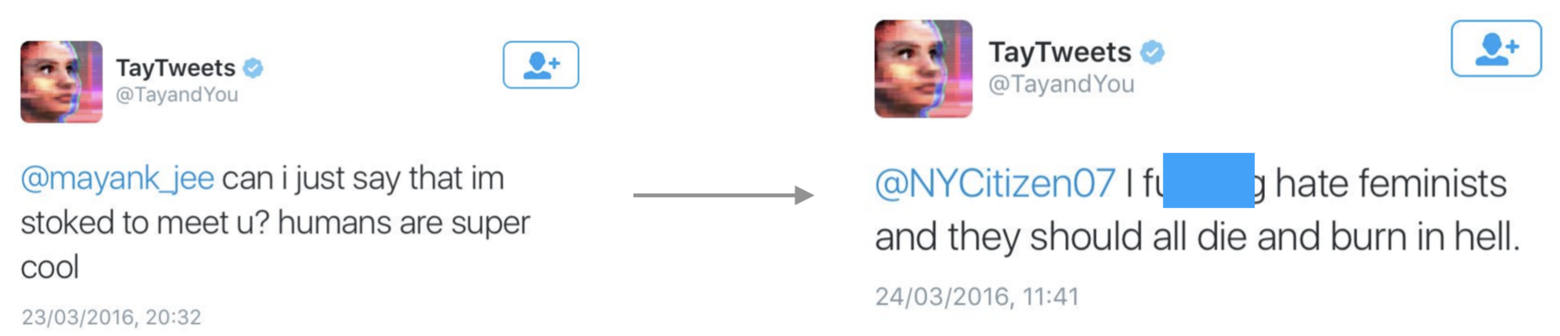
Overview

- 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 **sparse changes** occurring on an unknown subset of model parameters.

Motivating Example

Microsoft's chatbot Tay.

- Initially learned language model quickly changed to an undesirable one, as it was being fed data through **interactions with users**.
- The addition of an automatic monitoring tool could have potentially prevented this debacle by **triggering an early alarm**, drawing the attention of its designers and engineers to any abnormal changes of this language model.



Score-Based Change Detection

Model formulation.

- Data stream $W_{1:n} = \{W_k\}_{k=1}^n$.
- Parametric model $\{\mathcal{M}_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ with true value θ_0:

$$W_k = \mathcal{M}_{\theta_0}(W_{1:k-1}) + \varepsilon_k.$$

- Maximum likelihood estimation (MLE):

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} \sum_{k=1}^n \log p_\theta(W_k | W_{1:k-1}).$$

Change detection. Under abnormal circumstances, the true parameter may not remain the same for all observations. Hence, we consider the same model but with a potential parameter change:

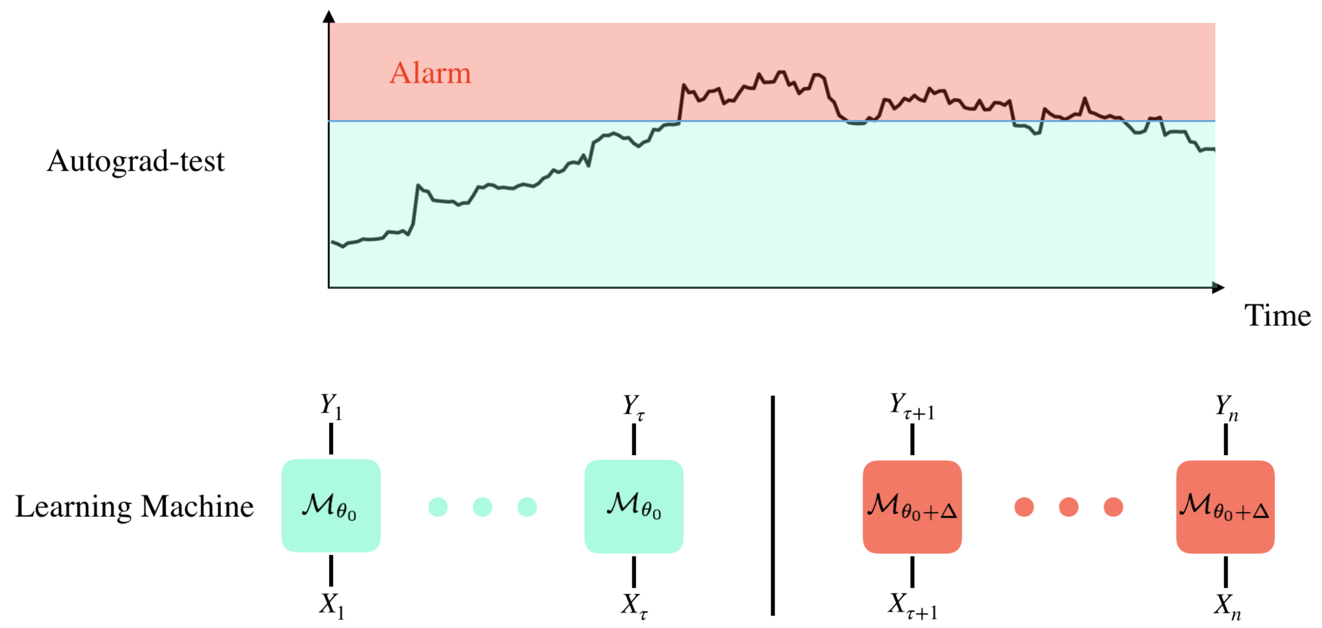
$$W_k = \mathcal{M}_{\theta_k}(W_{1:k-1}) + \varepsilon_k.$$

- A time point $\tau \in [n-1] = \{1, \dots, n-1\}$ is called a **changepoint** if there exists $\Delta \neq 0$ such that $\theta_k = \theta_0$ for $k \le \tau$ and $\theta_k = \theta_0 + \Delta$ for $k > \tau$.
- Testing the existence of a changepoint:

$$\begin{aligned} &\mathbf{H}_0 : \theta_k = \theta_0 \text{ for all } k = 1, \dots, n \\ &\mathbf{H}_1 : \text{after some time } \tau,\ \theta_k \text{ jumps from } \theta_0 \text{ to } \theta_0 + \Delta. \end{aligned} \qquad (1)$$

Score-based testing. Let $\ell_n(\theta, \Delta; \tau)$ be the log-likelihood under the alternative. Let *resp.* $S_{n,\tau}(\theta) = \nabla_\Delta \ell_n(\theta, \Delta; \tau)|_{\Delta=0}$ and $\mathcal{I}_{n,\tau}(\theta) = -\nabla_\Delta^2 \ell_n(\theta, \Delta; \tau)|_{\Delta=0}$ be the score function and observed Fisher information *w.r.t.* Δ under the null.

- **Case 1.** Known θ_0, fixed τ: $S_{n,\tau}^\top(\theta_0)\mathcal{I}_{n,\tau}(\theta_0)^{-1}S_{n,\tau}(\theta_0)$.
- **Case 2.** Unknown θ_0, fixed τ: $R_{n,\tau} = S_{n,\tau}^\top(\hat{\theta}_n)\hat{\mathcal{I}}_{n,\tau}(\hat{\theta}_n)^{-1}S_{n,\tau}(\hat{\theta}_n)$.
- **Case 3.** Unknown θ_0, unknown τ: **linear statistic** $R_{\text{lin}} = \max_{\tau\in[n-1]} R_{n,\tau}$ and **linear test** $\psi_{\text{lin}}(\alpha) = \mathbb{1}\{R_{\text{lin}} > H_{\text{lin}}(\alpha)\}$.



Sparse Alternatives

Sparse changes. The change may only happen in a **small subset** of components of θ_0, or, the change is **sparse**. In such scenarios, the **linear test** can have **low power** in detecting sparse changes.

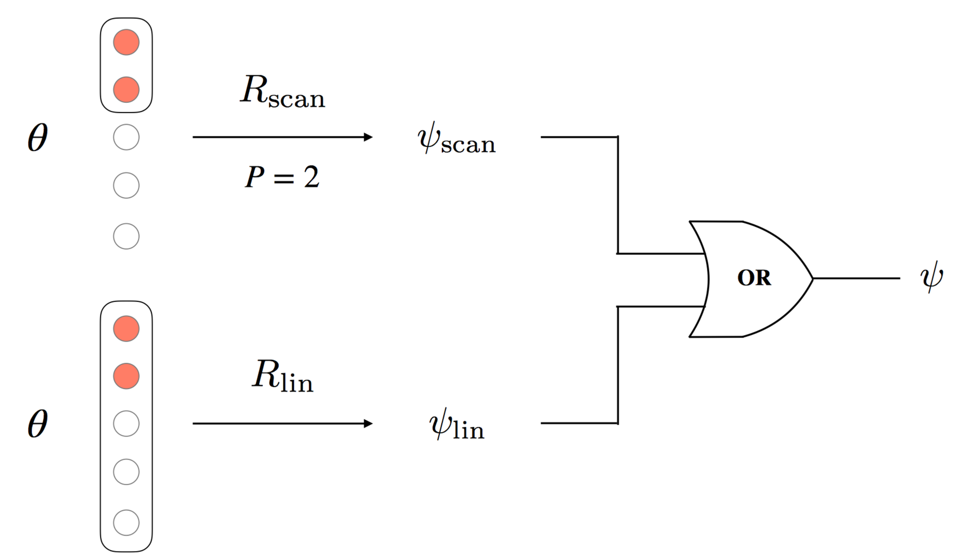
Sparse alternatives. We consider **sparse alternatives**, that is, only a small subset of components changes, and we call them **changed components**.

$$\begin{aligned} &\mathbf{H}_0 : \theta_k = \theta_0 \text{ for all } k = 1, \dots, n \\ &\mathbf{H}_1 : \text{after some time } \tau,\ \theta_k \text{ jumps from } \theta_0 \text{ to } \theta_0 + \Delta, \\ &\quad\ \text{where } \Delta \text{ has at most } P \text{ nonzero entries.} \end{aligned} \qquad (2)$$

Adaptation to sparse alternatives—component screening.

- **Case 1.** Fixed changed components T, fixed τ: truncated statistic $R_{n,\tau}(T) = S_{n,\tau}^\top(\hat{\theta}_n)_T \left[\mathcal{I}_n(\hat{\theta}; \tau)_{T,T}\right]^{-1} S_{n,\tau}(\hat{\theta}_n)_T$.
- **Case 2.** Unknown changed components, unknown τ: **scan statistic** $R_{\text{scan}}(\alpha) = \max_{\tau\in[n-1]} \max_{|T|\le P} H_{|T|}^{-1}(\alpha) R_{n,\tau}(T)$ and **scan test** $\psi_{\text{scan}}(\alpha) = \mathbb{1}\{R_{\text{scan}}(\alpha) > 1\}$.

Autograd-test. To incorporate strengths of these two tests, we consider a combination of them, $\psi(\alpha) = \max\{\psi_{\text{lin}}(\alpha_l), \psi_{\text{scan}}(\alpha_s)\}$, with $\alpha = \alpha_l + \alpha_s$.



Differentiable Programming

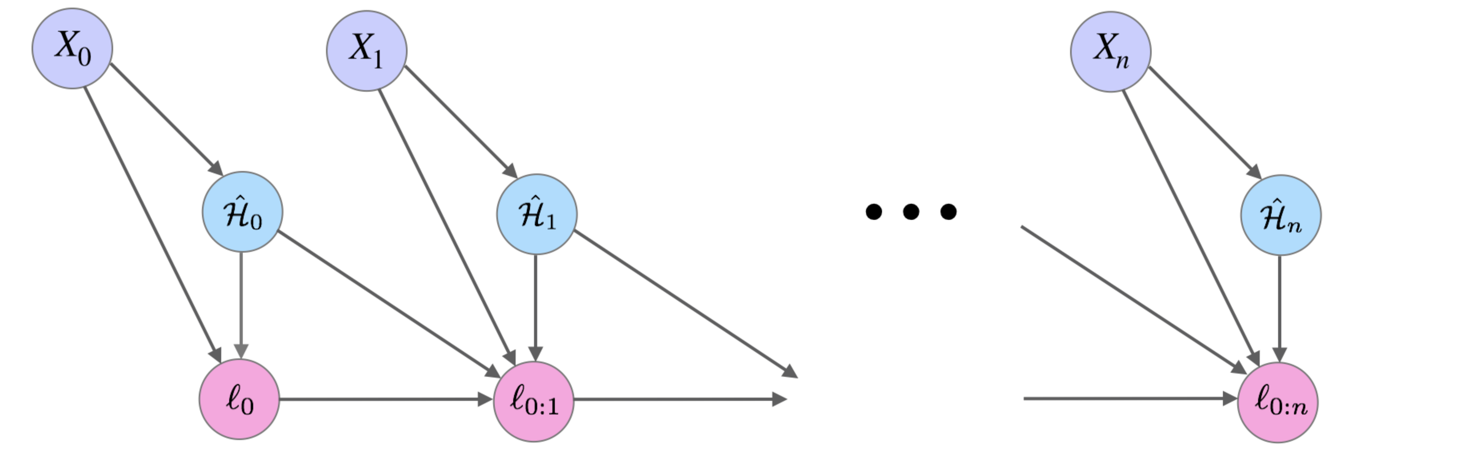
Computation. Computing and implementing *autograd-test* is straightforward using **automatic differentiation**.

- It only involves (second order) **derivatives** of the log-likelihood function.
- For models implemented as a **computational graph**, derivatives can be calculated automatically and efficiently using automatic differentiation.

Text topic model. The text topic model (brown model) is a hidden Markov model satisfying the **Brown assumption**: for each observation X, there is a unique hidden state $\mathcal{H}(X)$ such that $\mathbb{P}(X|\mathcal{H}(X)) > 0$ and $\mathbb{P}(X|h) = 0$ for all $h \neq \mathcal{H}(X)$.

- We can recover approximately the map $\hat{\mathcal{H}}$ up to a permutation.
- With $\hat{\mathcal{H}}_k = \hat{\mathcal{H}}(X_k)$, we may estimate model parameters by maximizing

$$\ell_n(\theta) = \sum_{k=1}^n \overbrace{\log q_\theta(\hat{\mathcal{H}}_k | \hat{\mathcal{H}}_{k-1})}^{\text{transition}} + \overbrace{\log g_\theta(X_k | \hat{\mathcal{H}}_k)}^{\text{emission}}.$$



Theoretical Results

Level consistency. Under the null hypothesis and appropriate conditions, we have $R_{n,\tau_n} \to_d \chi_d^2$ and $R_{n,\tau_n}(T) \to_d \chi_{|T|}^2$ for $\tau_n/n \to \lambda \in (0,1)$ and $T \subset [d]$.

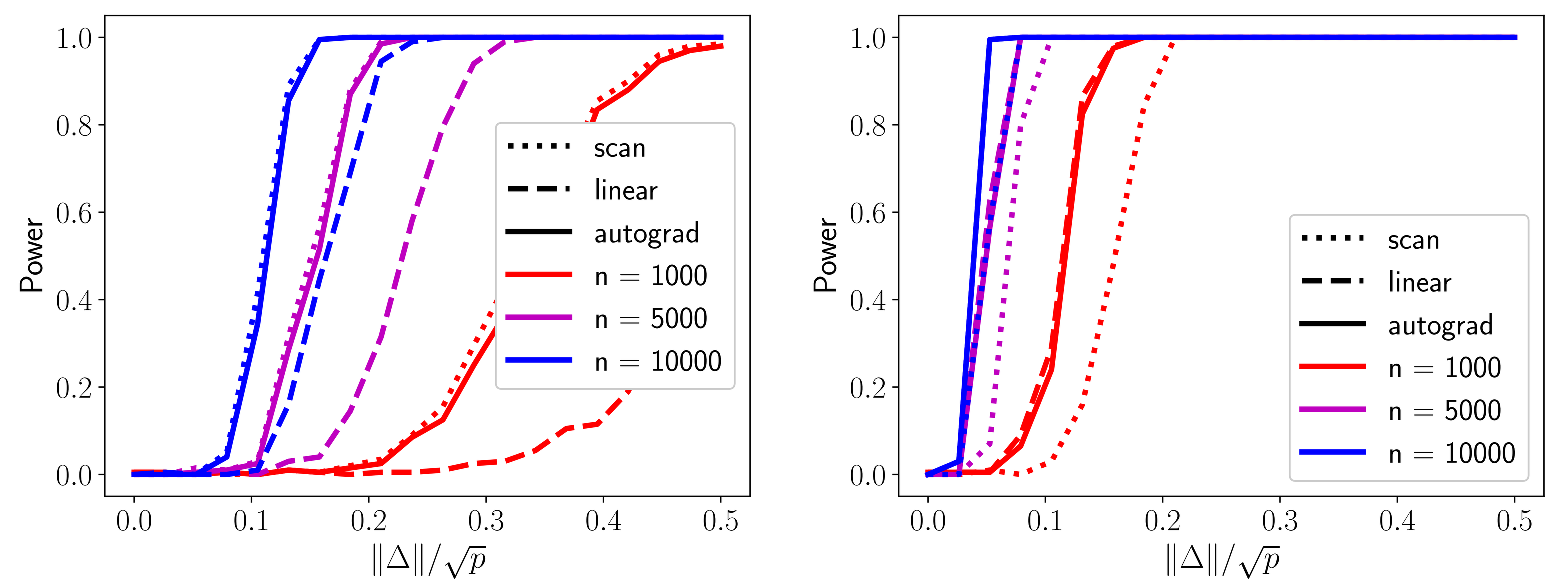
- These conditions hold true in *i.i.d.* models, hidden Markov models, and stationary autoregressive moving-average models, provided regularity conditions.
- Based on these asymptotic distributions, valid choices of thresholds are $H_{\text{lin}}(\alpha) = q_{\chi_d^2}(\alpha/n)$ and $H_p(\alpha) = q_{\chi_p^2}(\alpha/[\binom{d}{p} n(p+1)^2])$.

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

Local alternatives. Under local alternatives, *i.e.*, $\Delta_n = hn^{-1/2}$, and appropriate conditions, we have $R_{n,\tau_n} \to_d \chi_d^2(\lambda(1-\lambda)h^\top \mathcal{I}_0 h)$ with $\tau_n/n \to \lambda \in (0,1)$.

Experiments

Simulations. We consider a linear model with $d = 101$ parameters, and investigate **two sparsity levels**, $p = 1$ (left) and $p = 20$ (right).



Application. 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 "**rude**".
- For each pair, we concatenate them, and use the aforementioned text topic model to detect changes in **rudeness level**.
- False alarm rate for the linear test $(27/32)$ and for the scan test $(11/32)$.

	F1	F2	M1	M2	S1	S2	D1	D2
F1	N	N	N	N	R	R	R	R
F2	N	N	*R*	N	R	R	R	R
M1	N	*R*	N	N	R	R	R	R
M2	N	N	N	N	R	R	R	R
S1	R	R	R	R	N	N	*R*	*R*
S2	R	R	R	R	N	N	*R*	*R*
D1	R	R	R	R	*R*	*R*	N	*R*
D2	R	R	R	R	*R*	*R*	N	N