# Learning with Gradually Changing Distributions by Global Trend Filtering and Local Model Adjustment

**Peng Zhao**
zhaop@lamda.nju.edu.cn
January 3, 2019

## Abstract

if there is abstract...

## 1 Introduction

In real-life applications, data are often accumulated with time, and the underlying data distributions are inherently changing in nature. The evolving nature, if simply ignored, will largely do harm to the performance of the learning system. Therefore, it is quite crucial to develop the approaches which are able to adapt the evolving environments, that is, to accommodate the data whose distribution may change during the collection procedure.

As shown in the survey papers [Tsymbal, 2004; Gama *et al.*, 2014], the data distribution change can be classified into two different types: abrupt change and gradual change. The former one refers to the situations when the data distribution suddenly changes. The latter one essentially considers the scenarios when the environment is dynamically evolving, and thus the data distribution evolves and changes gradually. We argue that in many real-world applications, the data distribution usually changes in a gradually evolving manner. For instance,

Therefore, we focus on the gradual change case in this paper. Specifically, we borrow the idea from the filed of "trend filtering" [Kim *et al.*, 2009; Wang *et al.*, 2016] to explicitly model the global trend of the data stream, and utilize the new data batch to locally adjust the model based on the learned trend.

We propose ATF (Adapting evolving data stream by Trend Filtering) to alleviate and adapt the

## 2 Related Work

There are several work considering the theoretical foundations of learning with gradually changing distributions [Bartlett, 1992].

## 3 The Proposed Approach

### 3.1 Problem Statement and Notations

We first introduce the notations and define the problem of learning from the data stream, where the data are coming in a batch mode. Specifically, let $S_t = \{(\mathbf{x}_1^t, y_1^t), \ldots, (\mathbf{x}_{m_t}^t, y_{m_t}^t)\}$ be a set of data collected at time $t$, in which $m_t$ denotes the number of data collected at time $t$, that is, $m_t = |S_t|$. Besides, $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^d$ represents the $d$-dimension feature vector and $y \in \mathcal{Y}$ is the label. Our approach is able to be applied in multi-class classification, we focus on the binary classification for simplicity, namely, $\mathcal{Y} = \{-1, +1\}$. Our goal is to learn the classifier $h : \mathbb{R}^d \to \mathcal{Y}$, which is able to make predictions for the newly coming data batch $S_{t+1}$.

Throughout the paper, we will let $\|\cdot\|$ denote the $\ell_2$-norm, that is, for a $d$-dimensional vector $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\| = \sum_{i=1}^{d} x_i$.

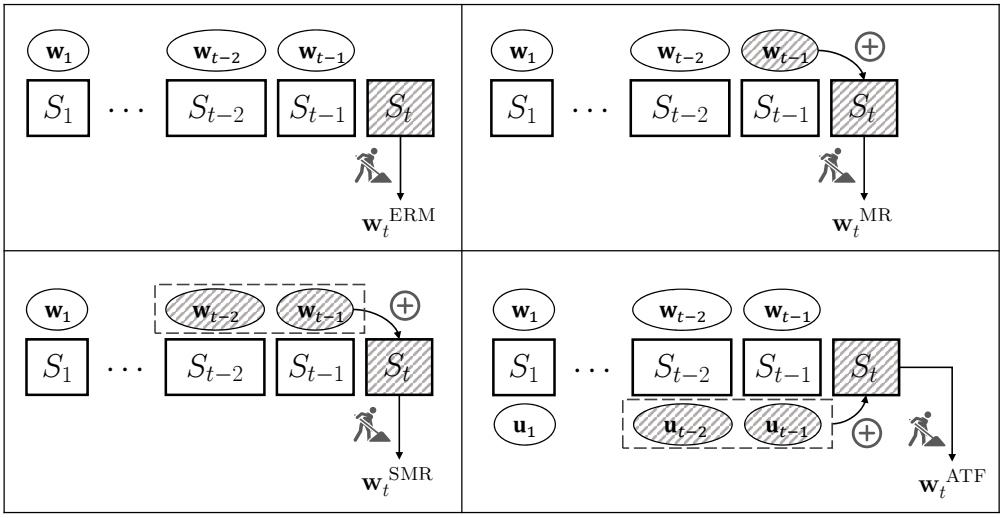


Figure 1: Illustration of the main mechanisms of three competitors: ERM, MR and SMR, which are variants of ATF. The symbol $S_t$ denotes the data collected at time $t$; $\mathbf{w}_t$ and $\mathbf{u}_t$ denote the model and tread learned at time $t$, respectively.

### 3.2 Naïve Trails and Benchmark Approaches

In Figure 1, we present the illustration of the main mechanisms of three competitors: ERM, MR and SMR, which are variants of ATF.

### 3.3 ATF Approach

We denote the overall model at time stamp $t$ as $\mathbf{w}_t$, and decompose it into the following form,

$$\mathbf{w}_t \approx \mathbf{u}_t + \mathbf{v}_t, \tag{1}$$

where we can view $\mathbf{u}_t$ as the global trend, $\mathbf{v}_t$ as the local model, and $\boldsymbol{\epsilon}_t$ as the noise term. Essentially, (1) implies a decomposition of the final prediction, which consists of three terms, i.e., global trend prediction, local model prediction and the noise. That is,

$$\mathbf{w}_t^{\mathrm{T}}\Phi(\mathbf{x}) = \mathbf{u}_t^{\mathrm{T}}\Phi(\mathbf{x}) + \mathbf{v}_t^{\mathrm{T}}\Phi(\mathbf{x}) + \epsilon_t \tag{2}$$

Therefore, it is desired to not only learn the overall model, but also consider to learn the global trend. We propose ATF approach, with the following general objective function, to learn the final model and the global trend jointly,

$$(\mathbf{w}_t, \mathbf{u}_t) = \arg\min_{\mathbf{w},\mathbf{u}} \Big\{ \sum_{i=1}^{m_t} \ell(\mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}_t^{(i)}), y_t^{(i)}) + \lambda_1 \Omega(\mathbf{w}, \mathbf{u}) + \lambda_2 \mathcal{S}(\mathbf{u}, \mathbf{u}_{t-1}, \mathbf{u}_{t-2}) \Big\}. \tag{3}$$

In the objective function (3), $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is the loss function, and $\Phi : \mathcal{X} \to \mathbb{H}$ be a feature mapping associated with a specified kernel $K$. Besides, $\Omega(\cdot, \cdot)$ is the *adjustment* regularizer, which essentially depicts the bias of the local model $\mathbf{w}$ with respect to the global trend $\mathbf{u}$. Besides, $\mathcal{S}(\mathbf{u}, \mathbf{u}_{t-1}, \mathbf{u}_{t-2})$ is a *trend* regularizer, enforcing the global trend $\mathbf{u}$ is smooth.

In the following, we choose the trend regularizer as $\mathcal{S}(\mathbf{u}, \mathbf{u}_{t-1}, \mathbf{u}_{t-2}) = \|\mathbf{u} + \mathbf{u}_{t-2} - 2\mathbf{u}_{t-1}\|^2$, and choose the adjustment regularizer as the squared $\ell_2$ norm, that is, $\Omega(\mathbf{w}, \mathbf{u}) = \|\mathbf{w} - \mathbf{u}\|^2$. Then, we can specify (3) in the following form,

$$(\mathbf{w}_t, \mathbf{u}_t) = \arg\min_{\mathbf{w},\mathbf{u}} \Big\{ \sum_{i=1}^{m_t} \ell(\mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}_t^{(i)}), y_t^{(i)}) + \lambda_1 \|\mathbf{w} - \mathbf{u}\|^2 + \lambda_2 \|\mathbf{u} + \mathbf{u}_{t-2} - 2\mathbf{u}_{t-1}\|^2 \Big\}. \tag{4}$$

## 4 Optimization

In this section, we first develop optimization method based on first order gradient descent to address the target optimization problem (4). Then, we provide an explanation of the target function from the perspective of alternative optimization.

To simplify the notations, we omitted the subscript $t$, specifically, using $m$ instead of $m_t$, and $\mathbf{w}$ ($\mathbf{u}$) instead of $\mathbf{w}_t$ ($\mathbf{u}_t$), besides, we use $\mathbf{x}_i$ ($y_i$) instead of $\mathbf{x}_t^{(i)}$ ($y_t^{(i)}$), with a slight abuse of notations. The simplified version of target optimization problem becomes,

$$(\mathbf{w}^*, \mathbf{u}^*) = \arg\min_{\mathbf{w},\mathbf{u}} \Big\{ \sum_{i=1}^{m} \ell(\mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}_i), y_i) + \lambda_1 \|\mathbf{w} - \mathbf{u}\|^2 + \lambda_2 \|\mathbf{u} + \mathbf{u}_{t-2} - 2\mathbf{u}_{t-1}\|^2 \Big\}. \quad (5)$$

Throughout the paper, we will adopt the linear mapping and squared hinge loss for simplicity, that is, $\Phi(\mathbf{x}) = \mathbf{x}$ and $\ell(\widehat{y}, y) = \max\{0, (1 - \widehat{y}y)\}^2$.

### 4.1 Nesterov's Accelerated Gradient Descent

Since the objective function is jointly convex and differentiable in terms of both model term $\mathbf{w}$ and the trend term $\mathbf{u}$. Therefore, we will adopt *Nesterov's accelerated gradient descent* [Nesterov, 2004; Bubeck, 2015] to solve this problem.

To simplify the presentation, we introduce the notation $\mathbf{s} = [\mathbf{w}^{\mathrm{T}}, \mathbf{u}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{2d}$ as the auxiliary variable, and immediately, we have

$$\mathbf{w} = \mathbf{s}^{\mathrm{T}} \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \end{bmatrix} \text{ and } \mathbf{u} = \mathbf{s}^{\mathrm{T}} \begin{bmatrix} \mathbf{0} \\ \mathbf{1} \end{bmatrix},$$

where $\mathbf{0}, \mathbf{1} \in \mathbb{R}^d$ are the $d$-dimensional zero and one column vector, respectively. Therefore, (4) is equivalent to the following optimization with respect to $\mathbf{s}$:

$$\mathbf{s}^* = \arg\min_{\mathbf{s}} \{ \sum_{i=1}^{m} \ell( \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \end{bmatrix}^{\mathrm{T}} \mathbf{s}, \mathbf{x}_i, y_i) + \lambda_1 \| \begin{bmatrix} -\mathbf{1} \\ \mathbf{1} \end{bmatrix}^{\mathrm{T}} \mathbf{s}\|^2 + \lambda_2 \| \begin{bmatrix} \mathbf{0} \\ \mathbf{1} \end{bmatrix}^{\mathrm{T}} \mathbf{s} + \mathbf{u}_{t-2} - 2\mathbf{u}_{t-1}\|^2 \}. \quad (6)$$

**Theorem 1.** *Assume the loss function $\ell : \mathcal{W} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ is convex, ATF algorithm will finally convergent to the global minimum of* (4)*. Moreover, let $\mathbf{w}_t^k$ and $\mathbf{u}_t^k$ be results returned in the $k$-th iteration, and $\mathbf{w}_t^*$ and $\mathbf{u}_t^*$ be the global minimum, then it reaches $\mathcal{L}(\mathbf{w}_t^k, \mathbf{u}_t^k) - \mathcal{L}(\mathbf{w}_t^*, \mathbf{u}_t^*) \leq \epsilon$ after $O(1/\sqrt{\epsilon})$ iterations.*

*Proof.* The objective function is convex with respect to $\mathbf{s}$ (a.k.a., jointly in terms of variables $\mathbf{w}$ and $\mathbf{u}$). Therefore, we are able to utilize the celebrated Nesterov's accelerated method [Nesterov, 2004].
<<< Peng: Cheng, please add the convergence analysis of the Nesterov's accelerated method (or Momentum) here.
Of course you can write an analysis independent with notations in your own note, however, please be sure to make your notations consistent with the contexts in the draft. >>> □

### 4.2 An Explanation from Alternative Optimization

Instead of using the first-order gradient based approach to directly optimize the concatenation variable $\mathbf{s}$, we can adopt the alternative optimization to solve $\mathbf{w}$ and $\mathbf{u}$ in problem (5). That is, we can update the trend $\mathbf{u}$ with a fixed model $\mathbf{w}$, and then update the model $\mathbf{w}$ with a fixed trend $\mathbf{u}$, until convergence.

This alternative optimization approach turns out that it presents us a new and insightful explanation of the original objective function. From the perspective of alternative optimization, we can view the objective function as an alternative optimization over the global trend $\mathbf{u}$ and the local model $\mathbf{w}$. That is,

∗ **Phase I. Global Trend Filtering** (fix $\mathbf{w}$ to update $\mathbf{u}$): when the local model is fixed, the sub-optimization problem essentially learns a global trend which is constrained to be smooth and not much deviated from the local model.

∗ **Phase II. Local Model Adjustment** (fix $\mathbf{u}$ to update $\mathbf{w}$): when the global trend is fixed, the sub-optimization problem aims to learn a local model by taking the global trend as the basis.

**Global Trend Filtering.** When updating the trend term $\mathbf{u}$ with fixed model $\mathbf{w}^{(k)}$, the sub-optimization problem becomes,

$$\mathbf{u}^{(k)} = \arg\min_{\mathbf{u}} \lambda_1 \|\mathbf{w}^{(k)} - \mathbf{u}\|^2 + \lambda_2 \|\mathbf{u} + \mathbf{u}_{t-2} - 2\mathbf{u}_{t-1}\|^2. \tag{7}$$

Essentially, the global trend $\mathbf{u}_t$ is constrained to, on one hand, be smooth in terms of the previous trend terms (i.e., $\mathbf{u}_{t-1}$ and $\mathbf{u}_{t-2}$); and on the other hand, have a small error with the overall model (i.e., $\mathbf{w}^{(k)}$). And fortunately, the problem (7) admits a close-form solution as follows,

$$\mathbf{u}^{(k)} = \frac{\lambda_1}{\lambda_1 + \lambda_2}\mathbf{w}^{(k)} + \frac{\lambda_2}{\lambda_1 + \lambda_2}(2\mathbf{u}_{t-1} - \mathbf{u}_{t-2}).$$

**Local Model Adjustment.** When updating the model $\mathbf{w}$ with a fixed trend term $\mathbf{u}^{(k)}$, the sub-optimization problem becomes,

$$\mathbf{w}^{(k+1)} = \arg\min_{\mathbf{w}} \sum_{i=1}^{m_t} \ell(\mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}_i), y_i) + \lambda_1 \|\mathbf{w} - \mathbf{u}^{(k)}\|^2. \tag{8}$$

The objective function (8) consists of two terms, the first one is the sum of loss over the data in the current chunk (that is, empirical risk), and the second one is a biased regularizer with the global trend $\mathbf{u}^{(k)}$. The latter one essentially depicts the idea of *model reuse*. Specifically, it is desired to adjust the local model on the basis of the global trend rather than the cold start.

Since the loss function and the regularizer are both convex and continuous in (8), the overall objective function is strongly convex with respect to the model $\mathbf{w}$. Thus, we can also adopt Nesterov's accelerated gradient descent to solve this sub-problem.

For the local model adjustment phase, we have the following theoretical analysis which gives the generalization error bound for the returned model $\widehat{\mathbf{w}}_t$ on the current data distribution.

To present our theoretical results, we need first give some formal definitions and notations. Considering the scenario in the $t$-th time, let $S_t$ be the set of $m_t$ data collected and $m_t = |S_t|$, where the underlying data distribution is denoted as $\mathcal{D}_t$. For any hypothesis $\mathbf{w} \in \mathcal{W}$, we denote the associated risk as $R(\mathbf{w})$ and the empirical risk as $\widehat{R}(\mathbf{w})$,

$$R(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_t}[\ell(\mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}), y)], \quad \widehat{R}(\mathbf{w}) = \sum_{i=1}^{m_t} \ell(\mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}_i), y_i).$$

**Theorem 2.** *Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite symmetric (PDS) kernel with $r^2 = \sup_{\mathbf{x}\in\mathcal{X}} K(\mathbf{x}, \mathbf{x})$. Let $\Phi : \mathcal{X} \to \mathbb{H}$ be a feature mapping associated with kernel $K$ and let the hypothesis set $\mathcal{W} = \{\mathbf{x} \to \mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}) : \|\mathbf{w}\|_{\mathbb{H}} \le \Lambda\}$ for some $\Lambda \ge 0$. Meanwhile, assume that we adopt the squared hinge loss and the regularizer $\Omega : \mathcal{H} \to \mathbb{R}_+$ is $\sigma$-strongly convex with respect to a norm $\|\cdot\|$. Let $\widehat{\mathbf{w}}_t$ be the model returned by the local model adjustment of ATF algorithm, based on the trend term $\mathbf{u}_t$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, we have*

$$R(\widehat{\mathbf{w}}_t) - \widehat{R}(\widehat{\mathbf{w}}_t) \le 2((r\Lambda + 1)^2 + 1)\sqrt{\frac{\epsilon_1}{m_t}} + 3\sqrt{\frac{\epsilon_2 \log(1/\delta)}{m_t}} + \frac{3(r\Lambda + 1)^2 \log(1/\delta)}{4m_t}, \tag{9}$$

*where $\epsilon_1 = \frac{4r^2\Lambda^2}{\sigma\lambda_1} + \frac{2\Lambda^2\rho}{\sigma}$ and $\epsilon_2 = \frac{(r\Lambda+1)^2}{4}\left(R_u + 16((r\Lambda + 1)^2 + 1)\sqrt{\frac{\epsilon_1}{m_t}}\right)$. Moreover, $\rho = \sup_{\mathbf{u}\in\mathcal{H}}\|\mathbf{u}\|^2$, and is supposed in the order of $O(1/m)$.*

*To make the presentation simpler, we only keep terms with respect to $m_t$, $\sigma$, $\lambda_1$ and $R_u$, obtaining that*

$$R(\widehat{\mathbf{w}}_t) - \widehat{R}(\widehat{\mathbf{w}}_t) = O\left(\frac{\epsilon_1}{\sqrt{m_t}} + \frac{\epsilon_2}{m_t}\right), \tag{10}$$

*where $\epsilon_1 = \left(\sqrt{R_u} + \sqrt{\frac{R_u}{\sigma\lambda_1}} + \sqrt[4]{\frac{R_u}{\sigma\lambda_1 m_t}}\right)$ and $\epsilon_2 = \left(\sqrt{\frac{1}{\sigma}} + \sqrt[4]{\frac{1}{\sigma}}\right)$. Besides, $R_u = R(\mathbf{u}_t) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_t}[\ell(\mathbf{u}_t^{\mathrm{T}}\Phi(\mathbf{x}), y)]$, representing the risk of the trend term on current data distribution.*

In order to better present the theoretical results, we will only provide the proof sketch in the following and defer the detailed proof in Appendix A.

**Proof Sketch.** The intuitive idea is that compared with the traditional empirical risk minimization problem, our objective function has an additional term which aligns the model with the global trend. This term essentially shrink the search space of the hypothesis set, that is, the complexity of the "effective" hypothesis set is reduced due to the existence of the global trend term.

Technically, we adopt the Bernett's concentration inequality

**Remark 1.** *From Eq.* (10)*, we can see that the first term* $\frac{\epsilon_1}{\sqrt{m_t}}$ *is the leading term in general, and thus the generalization error is in the order of* $O(1/\sqrt{m_t})$*, which is the typical convergence rate with respect to the number of learning examples. Furthermore, when the learned trend behaves well enough on the current data distribution, that is, the corresponding risk* $R_u$ *is sufficiently small (*$R_u \to o(1)$*), then it is able to achieve a* fast rate *convergence* $O(1/m_t)$*, which could be much faster than the original* $O(1/\sqrt{m_t})$*. This provides the theoretical justifications on the effectiveness of leveraging the global trend as the basis when learning the local model.*

**Remark 2.** *Although the current result is carried on the square hinge loss and binary classification scenario, the analysis can be also applied to many other loss functions satisfying certain conditions and extended to the multi-class classification. Actually, the essential analysis does not require the convexity, which means that we can even deal with* non-convex *loss function in the generalization error analysis (though the optimization is hard). It turns out that it is sufficient for loss function being bounded and Lipschitz continuous.*

## 5 Experiments

In this section, we provide the empirical performance on both synthetic and real-world datasets to validate the effectiveness of our approach. Moreover, we also report the parameter studies.

### 5.1 Synthetic Datasets

We compare the proposed approach $\text{ATF}_1$ and $\text{ATF}_2$ with TIX [Forman, 2006], Learn$^{++}$.NSE [Elwell and Polikar, 2011], and DTEL [Sun *et al.*, 2018].

### 5.2 Real-World Datasets

Table 1: Basic statistics of concept drift datasets, along with the information of data chunk.

| Dataset | #instance | #dim | #class | #chunk | Chunk Size |
|---|---|---|---|---|---|
| Usenet-1 | 1,500 | 100 | 2 | | |
| Usenet-2 | 1,500 | 100 | 2 | | |
| Luxembourg | 1,900 | 32 | 2 | | |
| Spam | 9,324 | 500 | 2 | | |
| Email | 1,500 | 913 | 2 | | |
| Weather | 18,159 | 8 | 2 | | |
| GasSensor | 4,450 | 129 | 6 | | |
| Powersupply | 29,928 | 2 | 2 | | |
| Electricity | 45,312 | 8 | 2 | | |
| Covertype | 581,012 | 54 | 2 | | |

### 5.3 Parameter Studies

In our approach, there are two hyper-parameters, i.e., adjustment coefficient $\lambda_1$ and smoothness coefficient $\lambda_2$, which play an crucial role in trading off empirical risk minimization, global trend learning and local model adjustments. In this paragraph, we will report the empirical performance of our approach with different coefficient values.

### 5.4 The Effectiveness of Model Reuse and Smooth Trend Filtering

One of the baselines is to directly learn a model on the current data batch, and then use this model to predict on the future data batch.

**Competitor 1** (ERM)**.**

$$\mathbf{w}_t^{\mathrm{ERM}} = \arg\min_{\mathbf{w}} \Big\{ \sum_{i=1}^{m_t} \ell(\mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}_t^{(i)}), y_t^{(i)}) + \lambda\|\mathbf{w}\|^2 \Big\}. \tag{11}$$

To validate the effectiveness of smooth trend filtering, we compare our proposed approach with the following two competitors.

**Competitor 2** (Model Reuse)**.**

$$\mathbf{w}_t^{\mathrm{MR}} = \arg\min_{\mathbf{w}} \Big\{ \sum_{i=1}^{m_t} \ell(\mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}_t^{(i)}), y_t^{(i)}) + \lambda\|\mathbf{w} - \mathbf{w}_{t-1}\|^2 \Big\}. \tag{12}$$

**Competitor 3** (Smooth Model Reuse)**.**

$$\mathbf{w}_t^{\mathrm{SMR}} = \arg\min_{\mathbf{w}} \Big\{ \sum_{i=1}^{m_t} \ell(\mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}_t^{(i)}), y_t^{(i)}) + \lambda\|\mathbf{w} + \mathbf{w}_{t-2} - 2\mathbf{w}_{t-1}\|^2 \Big\}. \tag{13}$$

Table 2: Performance comparisons on various datasets with different variants of ATF.

| Dataset | DTEL | TIX | LEARN.NSE | ERM | MR | SMR | $\mathrm{ATF}_{\mathrm{alter}}$ | ATF |
|---|---|---|---|---|---|---|---|---|
| Usenet1 | — | — | — | 70.89 | 72.69 | — | 72.69 | 73.10 |
| Usenet2 | — | — | — | 74.07 | 75.65 | — | 75.65 | 75.72 |
| Weather | — | — | — | 77.00 | 76.80 | — | 78.43 | 77.53 |
| Electricity | — | — | — | 75.08 | 74.30 | — | 76.96 | 76.22 |
| Powersupply | — | — | — | 69.25 | 70.55 | — | 70.18 | 70.21 |
| — | — | — | — | — | — | — | — | — |
| — | — | — | — | — | — | — | — | — |
| — | — | — | — | — | — | — | — | — |

<<< Peng: Cheng, the final results should include both avg and var. For the streaming datasets, we cannot randomly split the data, which will violate the streaming nature of data. Therefore, we need to holdout a small batch of data, and sliding the consecutive indexes of data to format several subset copies. >>>

<<< Peng: As you need to add the var in the table, it would be better to split the table into two tables: one for the comparisons in the literature, and the other one for the three benchmark variants of ATF. >>>

## 6 Discussion

## 7 Conclusion

In this paper, we are consider the problem of learning from data stream with gradually changing distributions. We argue that it would be beneficial by first filtering the global trend in the data steam, and then make the local model adjustment by utilizing the newly coming batch of data. We develop this idea by formulating a simple and reasonable optimization problem, which can be solved by the Nesterov's accelerated gradient descent. Meanwhile, we provide an explanation from the perspective of alternative optimization, theoretically showing that the convergence rate of generalization error bound can be substantially improved when the learned global trend is sufficiently well-behaved. This provides the theoretical justifications on the reasonableness of our proposal. Finally, we conduct extensive experiments on both synthetic and real-world datasets to show the effectiveness of our proposed approach. Empirical studies are also devoted to validate the usefulness of two essential modules of our approach, namely, the global trend filtering and the local model adjustment mechanisms.

## References


Peter L. Bartlett. Learning with a slowly changing distribution. In *Proceedings of the Fifth Annual ACM Conference on Computational Learning Theory, COLT*, pages 243–252, 1992.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

Ryan Elwell and Robi Polikar. Incremental learning of concept drift in nonstationary environments. *IEEE Transactions on Neural Networks*, 22(10):1517–1531, 2011.

George Forman. Tackling concept drift by temporal inductive transfer. In *SIGIR*, pages 252–259, 2006.

João Gama, Indre Zliobaite, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37, 2014.

Seung-Jean Kim, Kwangmoo Koh, Stephen P. Boyd, and Dimitry M. Gorinevsky. L1 trend filtering. *SIAM Review*, 51(2):339–360, 2009.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.

Yu Sun, Ke Tang, Zexuan Zhu, and Xin Yao. Concept drift adaptation by exploiting historical knowledge. *IEEE Transactions on Neural Networks and Learning Systems*, 29(10):4822–4832, 2018.

Alexey Tsymbal. The problem of concept drift: definitions and related work. *Technical Report, Computer Science Department, Trinity College Dublin*, 106, 2004.

Yu-Xiang Wang, James Sharpnack, Alexander J. Smola, and Ryan J. Tibshirani. Trend filtering on graphs. *Journal of Machine Learning Research*, 17:105:1–105:41, 2016.


## A Proof of Theorem 2

We first provide the formal definitions of the function properties, including strong convexity, Lipschitz continuity and smoothness.

**Definition 1** (Lipschitz Continuity)**.** *A function $f : \mathcal{K} \to \mathbb{R}$ is $L$-Lipschitz continuous w.r.t. a norm $\|\cdot\|$ over domain $\mathcal{K}$ if for all $\mathbf{x}, \mathbf{y} \in \mathcal{K}$, we have*

$$|f(\mathbf{y}) - f(\mathbf{x})| \leq L\|\mathbf{y} - \mathbf{x}\|.$$

**Definition 2** (Strong Convexity)**.** *A function $f : \mathcal{K} \to \mathbb{R}$ is $\lambda$-strongly convex w.r.t. a norm $\|\cdot\|$ if for all $\mathbf{x}, \mathbf{y} \in \mathcal{K}$ and for any $\alpha \in [0, 1]$, we have*

$$f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}) - \frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{x} - \mathbf{y}\|^2.$$

*A common and equivalent form for the differentiable case is,*

$$f(y) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^{\mathrm{T}}(y - \mathbf{x}) + \frac{\lambda}{2}\|y - \mathbf{x}\|^2. \tag{14}$$

**Definition 3** (Smoothness)**.** *A function $f : \mathcal{K} \to \mathbb{R}$ is $\sigma$-smooth w.r.t. a norm $\|\cdot\|$ if $f$ is everywhere differentiable and for all $\mathbf{x}, \mathbf{y} \in \mathcal{K}$, we have*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^{\mathrm{T}}(\mathbf{y} - \mathbf{x}) + \frac{\sigma}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

*The above condition is equivalent to a Lipschitz condition over the gradients,*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq \sigma\|\mathbf{x} - \mathbf{y}\|.$$

As we choose the squared hinge loss as the loss function, the following claim holds,

**Claim 1.** *For the feasible domain satisfying conditions in Theorem 2, the squared hinge loss* $\ell(y, \widehat{y}) = \max(0, 1 - y\widehat{y})^2$ *is* $M$*-bounded and* $L$*-Lipschitz function, with* $M = (r\Lambda + 1)^2$ *and* $L = 2(r\Lambda + 1)^2 + 2$.

For the squared hinge loss $\ell(y, \widehat{y}) = \max(1 - y\widehat{y})^2$, where $y$ is the true label and is taking value from $\{-1, +1\}$, while $\widehat{y}$ is the predictive label and $\widehat{y} \in \mathbb{R}$. Notice that $\widehat{y}_i = \mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}_i)$, and thus $\widehat{y}_i \leq \|\mathbf{w}\|_{\mathbb{H}}\sqrt{K(\mathbf{x}_i, \mathbf{x}_i)} \leq \Lambda r$. Therefore, the squared hinge loss admits

$$\ell(y, \widehat{y}) = \max(1 - y\widehat{y})^2 \leq 1 + y^2\widehat{y}^2 - 2y\widehat{y} \leq (\Lambda r + 1)^2.$$

Meanwhile, <<< Peng: here, we need to add the local Lipschitz continuity proof. >>>

*Proof.* □