Adaptive Passive-Aggressive Framework for Online Regression with Side Information

Runhao Shi, Jiaxi Ying, and Daniel P. Palomar

{rshiaf, jx,ying}@connect.ust.hk, palomar@ust.hk
The Hong Kong University of Science and Technology

December, 2024

Passive-Aggressive (PA) method [Crammer et al., 2006]

- A popular online algorithm used for regression problems involving streaming data.
- Update parameters in a passive-aggressive manner based on whether the error exceeds a predefined threshold:

$$\widehat{\mathbf{w}}_{t+1} = \arg\min_{\mathbf{w}\in\mathbb{R}^N} \frac{1}{2}||\mathbf{w}-\mathbf{w}_t||_2^2 \qquad \text{subject to} \quad \ell_\varepsilon(\mathbf{w};(\mathbf{x}_t, y_t)) = 0,$$

where ℓ_ε is the ε-insensitive hinge loss function defined as follows:

$$\ell_\varepsilon\left(\mathbf{w};\left(\mathbf{x},y\right)\right) = \begin{cases} 0 & \left|\mathbf{w}^\top\mathbf{x}-y\right| \le \varepsilon, \\ \left|\mathbf{w}^\top\mathbf{x}-y\right| - \varepsilon & \text{otherwise.} \end{cases}$$

Motivation

1. **Challenge 1**: PA struggles with determining the optimal threshold ε.
2. **Challenge 2**: PA cannot adapt well to side information, limiting its potential performance.

Reformulate PA

We regard the weight selected by PA as a function of ε as follows:

$$\widehat{\mathbf{w}}_{t+1}(\varepsilon) = \begin{cases} \mathbf{w}_t & |\mathbf{w}_t^\top \mathbf{x}_t - y_t| \le \varepsilon, \\ \mathbf{w}_t + \operatorname{sign}\left[y_t - \mathbf{w}_t^\top \mathbf{x}_t\right] \tau_t \mathbf{x}_t & \text{otherwise,} \end{cases}$$

where

$$\tau_t = \begin{cases} \left(\left|\mathbf{w}_t^\top \mathbf{x}_t - y_t\right| - \varepsilon\right) / ||\mathbf{x}_t||_2^2 & \text{(PA)} \\ \left(\left|\mathbf{w}_t^\top \mathbf{x}_t - y_t\right| - \varepsilon\right) / \left(||\mathbf{x}_t||_2^2 + \frac{1}{2C}\right) & \text{(PA-II).} \end{cases}$$

For the regression problem with constraints $\mathcal{W}$, the final weight is:

$$\mathbf{w}_{t+1}(\varepsilon) = \arg\min_{\mathbf{w} \in \mathcal{W}} ||\mathbf{w} - \widehat{\mathbf{w}}_{t+1}(\varepsilon)||_2^2. \quad (1)$$

Methodology

Passive-Aggressive with Side information (PAS) framework

- **Weight selection**: PAS integrates the side performance $h_t(\cdot)$ into the weight selection:

$$\mathbf{w}_{t+1}(\varepsilon) = \arg\min_{\mathbf{w}\in\mathcal{W}} \left(h_t(\mathbf{w}) + \frac{1}{2\lambda} ||\mathbf{w} - \widehat{\mathbf{w}}_{t+1}(\varepsilon)||_2^2 \right) = \text{prox}_{\lambda h_t}\left(\widehat{\mathbf{w}}_{t+1}(\varepsilon)\right),$$

  considering the trade-off between tracking accuracy and side performance.
- **Loss function**: The infimum defined by $\mathbf{w}_{t+1}(\varepsilon)$ is essentially the Moreau Envelope, which we define as the loss function with respect to ε:

$$f_t(\varepsilon) = \inf_{\mathbf{w}\in\mathcal{W}} \left[h_t(\mathbf{w}) + \frac{1}{2\lambda} ||\mathbf{w} - \widehat{\mathbf{w}}_{t+1}(\varepsilon)||_2^2 \right] = M_{\lambda h_t}\left(\widehat{\mathbf{w}}_{t+1}(\varepsilon)\right).$$

Methodology

Assumption 1. *The feasible domain $\mathscr{D}$ of the parameter ε is bounded with $\mathscr{D} = [\nu, D]$.*

Assumption 2. *The subderivatives of $f_t(\varepsilon)$ is bounded, such that $\sup_{\varepsilon \in \mathscr{D}, t \in [T]} |\partial f_t(\varepsilon)| \leq G$.*

Adaptive PAS (APAS)

- APAS dynamically update the value of ε based on the designed loss function $f_t(\varepsilon)$.
- Under Assumptions 1 and 2, ε_{t+1} is updated as follows:

$$\varepsilon_{t+1} = \Pi_{\mathscr{D}} \left[\varepsilon_t - \eta_t \tilde{g}_t(\varepsilon_t) \right],$$

where $\Pi_{\mathscr{D}}[\varepsilon] = \min\{\max\{\varepsilon, \nu\}, D\}$, $\eta_t = \frac{\zeta_t \sqrt{D}}{G\sqrt{\nu t}}$, and $\zeta_t = \Pi_{\mathscr{D}}\left[|\mathbf{w}_t^\top \mathbf{x}_t - y_t|\right]$. Here, $\tilde{g}_t(\varepsilon)$ is:

$$\tilde{g}_t(\varepsilon) := \begin{cases} f_t'(\varepsilon) & \text{if } \varepsilon < \zeta_t, \\ \max\{0, \partial_- f_t(\zeta_t)\} & \text{otherwise.} \end{cases}$$

Adaptive PAS (APAS)

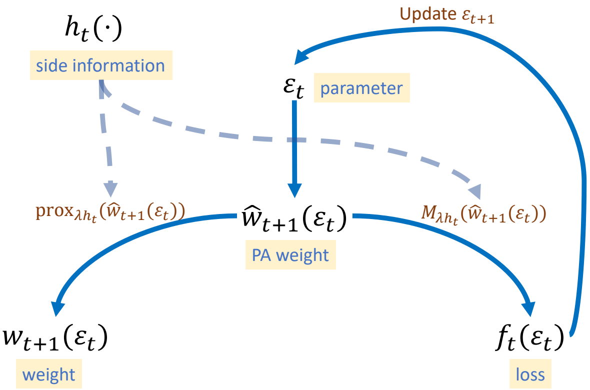


Figure 1: Adaptive learning scheme of APAS.

Regret Analysis of APAS

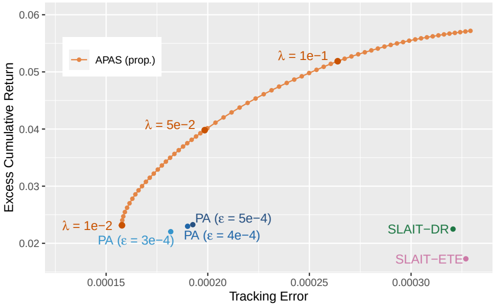
Theorem

Under Assumptions 1 and 2, the updating scheme of ε achieves the following regret bound for $T \geq 1$:

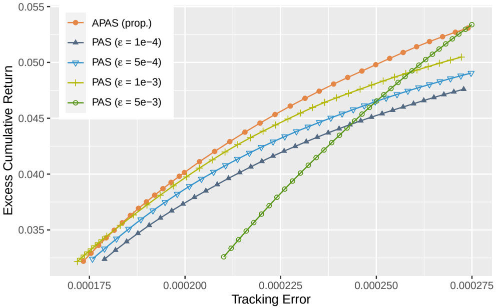
$$R_T = \sum_{t=1}^{T} f_t(\varepsilon_t) - \min_{\varepsilon \in \mathscr{D}} \sum_{t=1}^{T} f_t(\varepsilon) \leq 2\sqrt{\frac{D^3 G^2}{\nu}}\sqrt{T} = O(\sqrt{T}),$$

where D, ν, and G are constants defined in Assumptions 1 and 2.

Experiment - Synthetic Data



(a) Trade-off between tracking error and excess cumulative return of different methods.



(b) Ablation study: Comparison of PAS with fixed parameters and APAS.

Figure 2: Comparison of tracking error and excess cumulative return on the synthetic dataset.

Experiment - Real Market Data

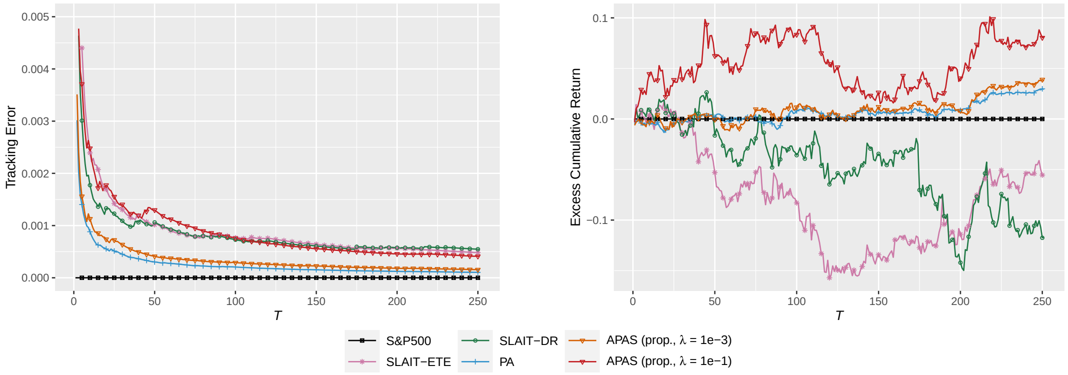


Figure 3: Tracking error and excess cumulative return over time T for different methods on S&P 500 dataset.

Thank you!

References I

[Crammer et al., 2006] Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. (2006).
Online passive-aggressive algorithms.
Journal of Machine Learning Research, 7(19):551–585.