Ensembling discounted VAW experts with a VAW meta-learner for adaptive online linear regression

G.A. Karapetyants (joint work with D.B. Rokhlin)

Institute of Mathematics, Mechanics and Computer Sciences
of the Southern Federal University

The 10th International Conference on Stochastic Methods (ICSM-10)
Divnomorskoe, June 01-06, 2025

Online linear regression

In online linear regression, at each round $t = 1, \dots, T$:

1. The learner receives $x_t \in \mathbb{R}^d$.
2. The learner predicts $\hat{y}_t = \langle w_t, x_t \rangle$ using $w_t \in \mathbb{R}^d$.
3. The true target $y_t \in \mathbb{R}$ is revealed.
4. The learner incurs squared loss $\ell_t(w_t) = \frac{1}{2}(y_t - \hat{y}_t)^2$.
5. The learner updates w_t to w_{t+1}.

Dynamic regret against a comparator sequence $\boldsymbol{u} = (u_1, \dots, u_T)$ is

$$R_T(\boldsymbol{u}) = \sum_{t=1}^{T} \ell_t(w_t) - \sum_{t=1}^{T} \ell_t(u_t).$$

Vovk-Azoury-Warmuth algorithm

In the VAW algorithm[1] the weight w_t is allowed to depend on the feature mapping x_t, indicating that features x_t are available at time t before predicting the label y_t:

$$w_t = \operatorname*{argmin}_{w\in\mathbb{R}^d} \left\{ \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{2}\sum_{i=1}^{t-1}(\langle(x_i, w\rangle - y_i)^2 + \frac{1}{2}\langle x_t, w\rangle^2 \right\}.$$

Explicitly,

$$w_t = S_t^{-1}\sum_{i=1}^{t-1} y_i x_i, \quad S_t = \lambda I_d + \sum_{i=1}^{t} x_i x_i^\top. \tag{1}$$

Moreover, $S_t^\top$ can be computed recursively by the Sherman-Morrison formula:

$$S_t^{-1} = S_{t-1}^{-1} - \frac{S_{t-1}^{-1}x_t(S_{t-1}^{-1}x_t)^T}{1 + x_t^T S_{t-1}^{-1}x_t}, \quad S_0^{-1} = \lambda^{-1}I_d. \tag{2}$$

[1] Cesa-Bianchi and Lugosi 2006, Section 11.8.

Regret bound for the VAW algorithm

In the sequel, we will assume that

$$\|x_t\|_2 \leq a, \quad |y_t| \leq Y.$$

The static regret

$$R_T(u) = \frac{1}{2}\sum_{t=1}^{T}(\langle x_t, w_t\rangle - y_t)^2 - \frac{1}{2}\sum_{t=1}^{T}(\langle x_t, u\rangle - y_t)^2$$

of the VAW algorithm satisfies the bound[2]

$$R_T(u) \leq \frac{\lambda}{2}\|u\|_2^2 + \frac{dY^2}{2}\ln\left(1 + \frac{a^2 T}{\lambda d}\right). \tag{3}$$

[2] Cesa-Bianchi and Lugosi 2006.

Discounted VAW algorithm

Let $\gamma \in (0,1]$, $\lambda > 0$, $\tilde{y}_1 = 0$, and $\tilde{y}_t \in [-\widetilde{Y}, \widetilde{Y}]$ for $t > 1$. Define

$$h_t(w) = \frac{1}{2}(\tilde{y}_t - \langle x_t, w\rangle)^2, \quad \ell_0(w) = \frac{\lambda}{2}\|w\|_2^2.$$

Recursively define

$$\Sigma_t = x_t x_t^\top + \gamma \Sigma_{t-1}, \quad \Sigma_0 = \lambda I.$$

Set $w_1 = 0$. Discounted VAW (DVAW) algorithm[3]:

$$w_t = \arg\min_{w \in \mathbb{R}^d} \left\{ h_t(w) + \gamma \sum_{s=0}^{t-1} \gamma^{t-1-s} \ell_s(w) \right\}. \tag{4}$$

Explicitly,

$$w_t = \Sigma_t^{-1} \left[\tilde{y}_t x_t + \gamma \sum_{s=1}^{t-1} \gamma^{t-1-s} y_s x_s \right]. \tag{5}$$

[3] Jacobsen and Cutkosky 2024.

Regret bound for DVAW algorithm

For the dynamic regret the following bounds holds true[4]

$$R_T(\mathbf{u}) \leq \frac{\gamma\lambda}{2}\|u_1\|_2^2 + \frac{d}{2}\max_{1\leq t\leq T}\Delta_t^2 \ln\left(1 + \frac{\sum_{t=1}^{T}\gamma^{T-t}a^2}{\lambda m}\right)$$
$$+\gamma\sum_{t=1}^{T-1}[F_t^\gamma(u_{t+1}) - F_t^\gamma(u_t)] + \frac{d}{2}\ln(1/\gamma)\Delta_{1:T}^2$$

where $F_t^\gamma(w) = \gamma^t\frac{\lambda}{2}\|w\|_2^2 + \sum_{s=1}^{t}\gamma^{t-s}\ell_s(w)$,

$$\Delta_t^2 = (y_t - \tilde{y}_t)^2, \quad \Delta_{1:T}^2 = \sum_{t=1}^{T}(y_t - \tilde{y}_t)^2.$$

[4] Jacobsen and Cutkosky 2024.

Simplified bound

Assume that $\|u_t\|_2 \leq R$, and put

$$P_T(\boldsymbol{u}) = \sum_{t=1}^{T-1} \|u_{t+1} - u_t\|_2.$$

Then

$$R_T(\boldsymbol{u}) \leq \eta a(Y + aR)P_T(\boldsymbol{u}) + \frac{d}{2\eta}\Delta_{1:T}^2 + \lambda R P_T(\boldsymbol{u})$$
$$+ \frac{d}{2} \max_{1 \leq t \leq T} \Delta_t^2 \ln\left(1 + \frac{a^2 T}{\lambda d}\right) + \frac{\lambda}{2} R^2,$$

where $\eta = \frac{\gamma}{1-\gamma}$.

Optimize the simplified bound over η:

$$\eta^* = \sqrt{\frac{d\Delta_{1:T}^2}{2a(Y + aR)P_T(u)}},$$

and substitute optimal η:

$$R_T(\boldsymbol{u}) \leq \sqrt{2da(Y + aR)\Delta_{1:T}^2 P_T(\boldsymbol{u})} + \lambda R P_T(\boldsymbol{u})$$
$$+ \frac{d}{2} \max_{1 \leq t \leq T} \Delta_t^2 \ln\left(1 + \frac{a^2 T}{\lambda d}\right) + \frac{\lambda}{2} R^2.$$

Following[5], put

$$b > 1, \quad \eta_{\min} = 2d, \quad \eta_{\max} = dT,$$

$$\mathcal{S}_\eta = \{\eta_i = \eta_{\min} b^i \wedge \eta_{\max} : i \in \mathbb{Z}_+\},$$

$$\mathcal{S}_\gamma = \left\{\gamma_i = \frac{\eta_i}{1+\eta_i} : i \in \mathbb{Z}_+\right\} \cup \{0\}$$

Theorem

For DVAW forecasters $\mathcal{A}_{\gamma_k}$, $\gamma_k \in \mathcal{S}_\gamma$ take VAW as a meta-algorithm $\mathcal{A}$. Put $\lambda = 1/T$ for $\mathcal{A}_k$. Then

$$R_T^{\mathcal{A}}(\boldsymbol{u}) = O\Bigg((MY^2 + d(Y + \widetilde{Y})^2)\ln T + (1+b)\sqrt{da(Y + aR)P_T(\boldsymbol{u})\Delta_{1:T}^2}\Bigg).$$

Note that the set S_γ contains $M = O(\log_b(\eta_{\max}/\eta_{\min})) = O(\log_b T)$ elements $\gamma_0, \ldots, \gamma_{M-1}$.

[5] Jacobsen and Cutkosky 2024.

Decompose the regret of the meta-algorithm $\mathcal{A}$ as

$$
\begin{aligned}
R_T^{\mathcal{A}}(\boldsymbol{u}) &= \frac{1}{2}\sum_{t=1}^{T}(\langle z_t, \alpha_t\rangle - y_t)^2 - \frac{1}{2}\sum_{t=1}^{T}(\langle z_t, e_k\rangle - y_t)^2 \\
&+ \frac{1}{2}\sum_{t=1}^{T}(z_{t,k} - y_t)^2 - \frac{1}{2}\sum_{t=1}^{T}(\langle u_t, x_t\rangle - y_t)^2) \\
&= \underbrace{R_T^{\mathcal{A}}(e_k)}_{\text{Meta-learner's regret w.r.t. expert } k} + \underbrace{R_T^{\mathcal{A}_{\gamma_k}}(\boldsymbol{u})}_{\text{Regret of expert k}}
\end{aligned}
$$

This is true for any $k \in \{0, \dots, M-1\}$, which can depend on $\eta_* = \eta_*(\boldsymbol{y}, \boldsymbol{u})$.

Idea of the proof 2/3: regret of the meta-learner

The meta-learner is the VAW forecater. Thus

$$R_T^{\mathcal{A}}(e_k) \leq \frac{\lambda}{2}\|e_k\|_2^2 + \frac{MY^2}{2} \ln\left(1 + \frac{1}{\lambda M}\sum_{t=1}^{T}\|z_t\|_2^2\right), \tag{6}$$

The cumulative squared norm of the predictions of the DVAW forecasters can be bounded as

$$\sum_{t=1}^{T}\|z_t\|_2^2 = \sum_{t=1}^{T}\sum_{k=0}^{M-1} z_{t,k}^2 = O(MT).$$

Idea of the proof 3/3: regrets of DVAW experts

Select k by the following rules:

(1) $k = 0$, if $\eta^* \leq \eta_{\min} = 2d$;

(2) take k such that $\eta_k \leq \eta^* \leq b\eta_k$, if $\eta_{\min} \leq \eta^* \leq \eta_{\max}$, where $\eta_k \in S_\eta$;

(3) $\eta_k = \eta_{\max} = dT$, if $\eta^* \geq \eta_{\max} = dT$.

Using the definition of η^*, it is possible to prove the bound:

$$R_T^{\mathcal{A}_{\gamma_k}}(\boldsymbol{u}) \leq (1+b)\sqrt{\frac{d}{2}a(Y+aR)P_T(\boldsymbol{u})\Delta^2_{1:T}} + \frac{1}{2}(Y+\widetilde{Y})^2 + \overline{\lambda}RP_T(\boldsymbol{u}) + \frac{d}{2}\max_{1\leq t\leq T}\Delta_t^2 \ln\left(1+\frac{a^2T}{\overline{\lambda}d}\right) + \frac{\overline{\lambda}}{2}R^2$$

It remains to put $\overline{\lambda} = 1/T$ and combine the bounds for the meta-learner and a DVAW expert learner. □

Experimental Setup

- **Goal:** Compare Meta-DVAW against standard VAW forecaster.
- Regularization $\lambda = 0.1$ for all VAW/DVAW components (base experts and meta-learner).
- Base DVAW experts use discount factors from the set: $\mathcal{G} = \{0.70, 0.85, 0.95, 1.00\}$.
- All base DVAW experts use hint $\tilde{y}_t = 0$.
- Evaluation Metric:

$$\text{MSE:} \quad \sum_{t=1}^{T} \frac{1}{2T} (y_t - \hat{y}_t)^2.$$

Artificial Datasets

$$T = 1000, d = 5, x_t \sim N(0, I),$$

$$y_t = \langle w_{true,t}, x_t \rangle + \varepsilon_t, \varepsilon_t \sim N(0, 0.2^2).$$

1. **Stationary:** $w_{true,t} = [1, -0.5, 0.2, -0.8, 1.2]^T$.
2. **Abrupt Drift:** $w_1 = [1, 1, 1, 1, 1]^T$; $w_2 = [-1, -1, -1, -1, -1]^T$; $w_3 = [1, -1, 1, -1, 1]^T$.
   $w_{true,t}$ switches from $w_1 \to w_2$ at $T/3$, then $w_2 \to w_3$ at $2T/3$.
3. **Gradual Drift (Random Walk):** $w_{true,0} = [0.5, \dots, 0.5]^T$;

$$w_{true,t} = w_{true,t-1} + v_t, \quad v_t \sim N(0, (0.05)^2 I).$$

1. **Gradual Drift (Sinusoidal):**

$$w_{true,t,j} = \sin(2\pi t/(100 + 50j)) + 0.5\cos(2\pi t/(150 + 30j)),$$

   for $j = 1, \dots, 5$.

2. **Changing Noise:** $w_{true,t}$ as in Stationary,

$$\sigma_{noise,t} = 0.1 \quad \text{for } t \leq T/2; \qquad \sigma_{noise,t} = 0.5 \quad \text{for} \quad t > T/2.$$

3. **Covariate Shift:** $w_{true,t}$ as in Stationary,

$$x_t \sim N(0, I) \quad \text{for } t \leq T/2; \quad x_t \sim N([1, 1, 0, 0, 0]^T, I) \quad \text{for } t > T/2.$$

Results: artificial datasets

Table: MSE averaged over 10 runs

Dataset Type	VAW	Meta-DVAW
Stationary Linear	**0.0402**	0.0480
Abrupt Drift Linear	2.2579	**0.3110**
Gradual Drift (RW)	1.0396	**0.1812**
Gradual Drift (Sin)	1.1525	**0.2869**
Changing Noise	**0.0639**	0.0719
Covariate Shift	**0.0440**	0.0529

- *Stationary, Changing Noise, Covariate Shift:* Standard VAW performs slightly better or comparably. Meta-DVAW's overhead is minimal.
- *Drifting $w_{true,t}$ (Abrupt, Gradual RW, Gradual Sine):* Meta-DVAW demonstrates substantially MSE loss.

Financial time series datasets

Table: MSE for daily log-return predictions

Dataset	VAW	Meta-DVAW	Trivial (Last Val)	Trivial (MA-5)
IBM	9.99e-05	**9.95e-05**	2.05e-04	1.19e-04
Microsoft	1.34e-04	**1.34e-04**	2.96e-04	1.63e-04
Google	**1.51e-03**	1.51e-03	3.07e-03	1.80e-03
S&P 500 ETF	6.08e-05	**6.06e-05**	1.33e-04	7.36e-05
NASDAQ 100 ETF	8.53e-05	**8.50e-05**	1.85e-04	1.03e-04

Feature vector x_t for financial experiments

The target variable y_t is the daily log return: $y_t = \ln(P_t/P_{t-1})$. The feature vector x_t is constructed from data up to day $t-1$ and includes:

- **Lagged Log Returns (5 features):** $y_{t-1}, y_{t-2}, \ldots, y_{t-5}$.
- **Lagged Volume % Change (3 features):** $\text{Vol\%Chg}_{t-1}, \ldots,$ $\text{Vol\%Chg}_{t-3}$ (if volume is valid).
- **MACD Histogram (1 feature):** Standard (12, 26, 9) periods.
- **RSI (1 feature):** 14-day Relative Strength Index.
- **Realized Volatility (2 features):** Std. dev. of log returns over past 10 days and 30 days.

The final feature vector x_t includes all available components from the above. Maximum dimension $d = 12$.

Gas sensor array drift dataset

http://archive.ics.uci.edu/ml/datasets/Gas+Sensor+Array+Drift+Dataset+at+Different+Concentrations

Table: Gas concentrations MSE

Target Gas	Std VAW MSE	Meta-DVAW MSE	Trivial (Last Val)	Trivial (MA-10)
Ethanol	1.74e+09	**3.99e+08**	5.38e+08	4.52e+08
Ethylene	5.04e+07	**3.92e+07**	1.05e+08	5.31e+07
Ammonia	3.35e+07	**2.94e+07**	**2.25e+07**	3.82e+07
Acetaldehyde	8.74e+08	**4.61e+08**	**1.55e+08**	1.99e+08
Acetone	8.47e+09	**2.20e+09**	1.61e+09	**1.32e+09**
Toluene	5.48e+08	**1.89e+08**	1.87e+08	**1.09e+08**

References I

Cesa-Bianchi, N. and G. Lugosi (2006). *Prediction, learning, and games*. Cambridge University Press.

Jacobsen, Andrew and Ashok Cutkosky (2024). "Online Linear Regression in Dynamic Environments via Discounting". In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov et al. Vol. 235. Proceedings of Machine Learning Research. PMLR, pp. 21083–21120. URL: https://proceedings.mlr.press/v235/jacobsen24a.html.