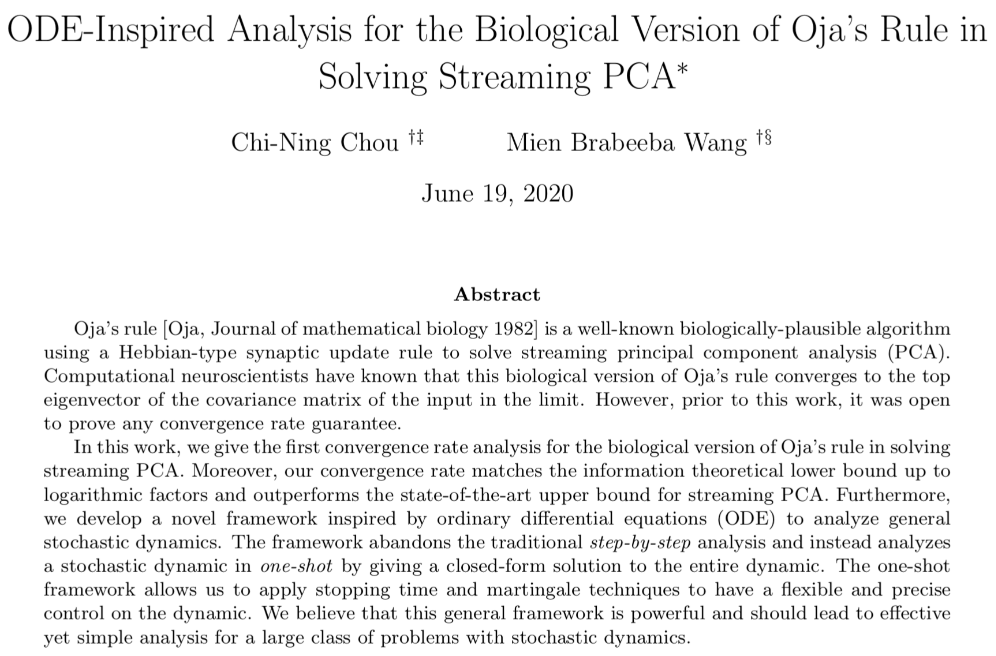
OJA'S RULE : A NEW ANALYSIS FOR STREAMING PCA

CHI-NING CHOU & MIEN BRABEEBA WANG , COLT 2020

SEPTEMBER 2, 2020 | Unsupervised learning reading group | Geelon So



ODE-Inspired Analysis for the Biological Version of Oja's Rule in Solving Streaming PCA*

Chi-Ning Chou [†‡] Mien Brabeeba Wang [†§]

June 19, 2020

Abstract

Oja's rule [Oja, Journal of mathematical biology 1982] is a well-known biologically-plausible algorithm using a Hebbian-type synaptic update rule to solve streaming principal component analysis (PCA). Computational neuroscientists have known that this biological version of Oja's rule converges to the top eigenvector of the covariance matrix of the input in the limit. However, prior to this work, it was open to prove any convergence rate guarantee.

In this work, we give the first convergence rate analysis for the biological version of Oja's rule in solving streaming PCA. Moreover, our convergence rate matches the information theoretical lower bound up to logarithmic factors and outperforms the state-of-the-art upper bound for streaming PCA. Furthermore, we develop a novel framework inspired by ordinary differential equations (ODE) to analyze general stochastic dynamics. The framework abandons the traditional *step-by-step* analysis and instead analyzes a stochastic dynamic in *one-shot* by giving a closed-form solution to the entire dynamic. The one-shot framework allows us to apply stopping time and martingale techniques to have a flexible and precise control on the dynamic. We believe that this general framework is powerful and should lead to effective yet simple analysis for a large class of problems with stochastic dynamics.

PROBLEM (streaming PCA).

- GIVEN: μ distribution over unit sphere $S^{n-1} \subset \mathbb{R}^n$
  - let $\Sigma = \mathbb{E}_{x \sim \mu}[xx^T]$ covariance
    - eigenvectors $v_1, \dots, v_n$
    - eigenvalues $\lambda_1 > \lambda_2 \geq \dots \geq \lambda_n \geq 0$
- STREAM: $x_1, x_2, \dots, x_T \overset{iid.}{\sim} \mu$
- RETURN: $w \in \mathbb{R}^n$ s.t. $\frac{\langle w, v_1 \rangle^2}{\|w\|^2} > 1 - \varepsilon$.

PERSPECTIVE (PCA).

Find $w \in \mathbb{R}^n$ maximizing: $\mathbb{E}_x \left[\frac{\langle w, x \rangle^2}{\|w\|_2^2} \right]$.

Pf.

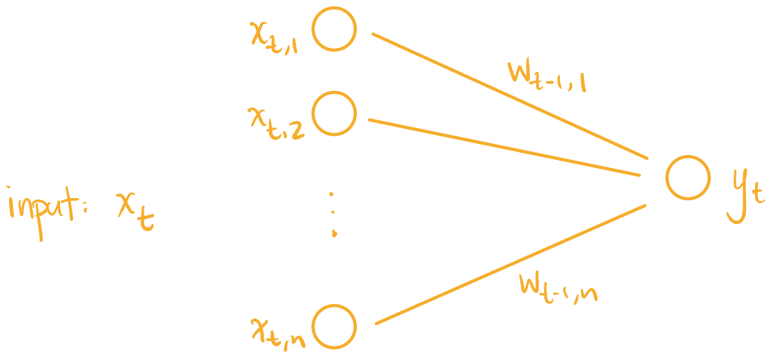
$$\mathbb{E}_x \left[\frac{\langle w, x \rangle^2}{\|w\|_2^2} \right] = \frac{\mathbb{E}_x [\mathrm{tr}(x^T w w^T x)]}{\|w\|_2^2}$$

$$= \frac{\mathbb{E}_x [\mathrm{tr}(xx^T ww^T)]}{\|w\|_2^2}$$

$$= \frac{\mathrm{tr}\left(\mathbb{E}_x [xx^T] \, ww^T \right)}{\|w\|_2^2} = \frac{w^T \Sigma w}{\|w\|_2^2}.$$

This is maximized when $w = v_1$, the first eigenvector. □

QUESTION. Can we solve streaming PCA using a neural architecture?



idea: update weights $w_{t,1}, \dots, w_{t,n}$ using the stream $(x_t)_{t=1}^{T}$ to max y.

OJA'S RULE: a local update rule

```
- initialize $w_0 \in S^{n-1}$ arbitrarily
- for each $x_t$ in $x_1, \dots, x_T$
    - compute $y_t = w_{t-1}^T x_t$
    - update weight
          $w_t \leftarrow w_{t-1} + \eta_t y_t (x_t - y_t w_{t-1})$
```

ASIDE:

$$\frac{d}{dw}\left(\langle w,x\rangle^2 - \lambda(\|w\|^2 - 1)\right)$$
$$= 2\langle w,x\rangle x - 2\lambda w$$
$$\propto yx - \lambda w$$

INTUITION for OJA'S RULE

1. if v is an eigenvector, then v is a fixed point in expectation.

Pf. suffices to show that

$$\mathbb{E}\left[(x^T v)(x - (x^T v)v)\right] = \mathbb{E}\left[(xx^T)v - tr(xx^T vv^T)v\right]$$
$$= \Sigma v - tr(\Sigma vv^T)v$$
$$= \lambda v - \lambda v = 0. \quad \square$$

in fact v_1 is a stable fixed point.

2. $\|w_t\| \rightsquigarrow 1$

Pf. the update :

$$w_t \leftarrow w_{t-1} + \overbrace{\eta_t y_t (x_t - y_t w_{t-1})}^{\xi}$$

which direction is the update w.r.t. ω_{t-1}?

$$\langle w_{t-1} , \eta_t y_t (x_t - y_t \omega_{t-1})\rangle$$

$$= \left\langle w_{t-1} , \eta_t y_t^2 \left(\frac{x_t}{\langle x_t, w_{t-1}\rangle} - w_{t-1} \right)\right\rangle$$

$$= \eta_t y_t^2 \cdot \left(\frac{\langle x_t , w_{t-1}\rangle}{\langle x_t , w_{t-1}\rangle} - \| w_t \|^2 \right)$$

$$= \eta_t y_t^2 \cdot (1 - \|w\|^2).$$

$\rightsquigarrow$ if $\|w\| > 1 \implies \langle w, \xi \rangle < 1$

$\|w\| < 1 \implies \langle w, \xi \rangle > 1$

□

If η_t's are small enough, w_t's will always be around unit:

> Lemma. If $\eta \in (0, \frac{1}{10})$ and if $\forall t$, $\eta_t \le \eta$, then
>
> $$1 - 10\eta \le \|w_t\|_2^2 \le 1 + 10\eta .$$

3. Oja's rule is derived from a Taylor expansion of the power method for finding v_1:

POWER METHOD:

```
- choose $w_0 \in S^{n-1}$
- for $t = 1, 2, \dots$
    $w_t \leftarrow \frac{(I + \eta_t x_t x_t^T) w_{t-1}}{\|(I + \eta_t x_t x_t^T) w_{t-1}\|}$
```

MAIN RESULT.

THEOREM. Let $\gamma = \lambda_1 - \lambda_2$ be the gap between two largest eigenvalues.

- Local convergence: if $\frac{\langle w_0, v_1 \rangle^2}{\|w_0\|^2} = \Omega(1)$, then

$$\Pr\left[\frac{\langle w_T, v_1 \rangle^2}{\|w_T\|^2} < 1 - \varepsilon \right] < \delta$$

where $\eta = \tilde{\Theta}\left(\frac{\varepsilon \cdot \gamma}{\lambda_1}\right)$ and $T = \tilde{\Theta}\left(\frac{\lambda_1}{\varepsilon \cdot \gamma^2}\right)$

- Global convergence: if $w_0 \in S^{n-1}$ uniform, same guarantee, but replace ε with $\varepsilon \wedge \delta^2 := \min\{\varepsilon, \delta^2\}$

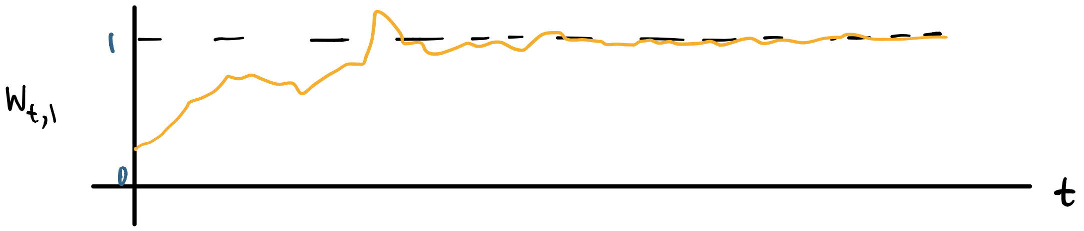
ANALYSIS.

REMARK. WLOG, assume $\Sigma = \lambda_1 e_1 + \lambda_2 e_2 + \cdots + \lambda e_n$

where $\{e_i\}_{i=1}^{n}$ is the standard basis

Then, the goal of streaming PCA is:

$$\frac{(W_{t,1})^2}{\| W_{t,1} \|^2} \longrightarrow 1.$$



if $W_{t,1}$ were a continuous stochastic process, we want $W_{t,1} \longrightarrow 1$.

NOTATION. Let $X_1, X_2, \ldots \overset{iid}{\sim} \mu$ and

$$W_t := W_t(X_t, W_{t-1})$$

random variables in Oja's rule.

- let $F_t = \sigma(X_1, \ldots, X_t)$ be the σ-field generated by $X_1, \ldots, X_t$.
  ie $(F_t)_{t=1}^{\infty}$ is the natural filtration of $(X_t)_{t=1}^{\infty}$.

ASIDE. F_t contains all the information of the stochastic process $(X_t)_{t=1}^{\infty}$ up to time t.

- conditional expectation: $\mathbb{E}[\cdot \mid F_t]$ is the expectation given that you know the value of $X_1, \ldots, X_t$.

NAIVE CONVERGENCE ANALYSIS: WTS $\mathbb{E}[W_{t,1}] \longrightarrow \pm 1$

- one could potentially bound improvement:

$$\mathbb{E}[W_{t,1}] = \mathbb{E}\Big[\mathbb{E}\big[W_{t-1,1} + \eta_t (X_t^T W_{t-1}) X_{t,1} - \eta_t (X_t^T W_{t-1})^2 W_{t-1,1} \mid F_{t-1}\big]\Big]$$

$$= \mathbb{E}\Big[W_{t-1,1} + \eta_t \lambda_1 W_{t-1,1} - \eta_t \big(\textstyle\sum \lambda_i W_{t-1,i}^2\big) W_{t-1,1}\Big]$$

→ show improvement $\mathbb{E}[W_{t,1} - W_{t-1,1}]$ and prove concentration ; induct.

- difficulties in analysis
  - nonlinear update w.r.t. $W_{t-1,1}$
  - sensitivity to $W_{t-1,1}$ lost to expectation

- see Zeyuan Allen-Zhu and Yuanzhi Li. Fast efficient convergence for streaming k-PCA. 2017. for such analysis.

OVERVIEW of ANALYSIS

A. Continuous Intuition

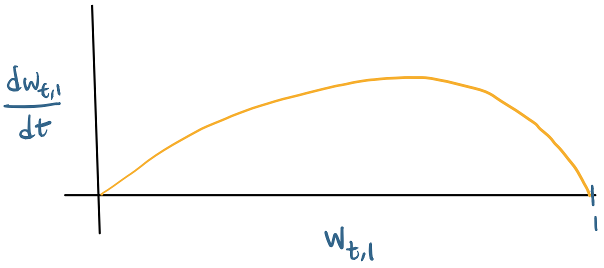
- if we let the learning rate $\eta \to 0$, what happens to the limiting process $(w_t)_{t\geq 0}$?

CLAIM. Continuous Oja's rule is deterministic, satisfying:

consider Brownian motion + SDE

$$dw = \left[\Sigma w_t - w_t^T \Sigma w_t w_t\right] dt .$$

$$\Longrightarrow \qquad \frac{dw_{t,1}}{dt} \geq (\lambda_1 - \lambda_2)\, w_{t,1} (1 - w_{t,1})^2 .$$



LINEARIZATION. if we have an ODE

linear term

$$\frac{dy(t)}{dt} = a y(t) + b(t)$$

noise term.

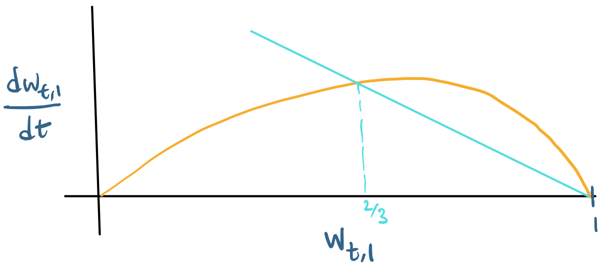
then

$$y(T) = e^{aT}\left(y(0) + \underbrace{\int_0^T e^{-at}\, b(t)\, dt}_{\text{if noise term small, then } y(0) \text{ dominates.}} \right) .$$

IDEA. if we knew that $w_{0,1} > 2/3$, then

$$\frac{dw_{t,1}}{dt} \geq \frac{2}{3}(\lambda_1 - \lambda_2)(1 - w_{t,1}) .$$

choose this as the linear term.



B. DISCRETE ANALOGUE

LINEARIZATION. Rewrite Oja's update in the following form:

$$W_t \gtrsim \underbrace{a \cdot W_{t-1}}_{\text{linear term}} + \underbrace{N_t}_{\text{noise term}}$$

$$\Rightarrow \quad W_T \gtrsim a^T \cdot \left(W_0 + \sum_{t=1}^{T} a^{-t} N_t \right)$$

where $\underbrace{\left(\sum_{i=1}^{t} a^{-i} \cdot N_i \right)_{t=1}^{\infty}}_{(M_t)_{t=1}^{\infty}}$ forms an $(\mathcal{F}_t)_{t=1}^{\infty}$ - martingale.

or 'almost' one.

MARTINGALE ANALYSIS.

Apply martingale inequalities to show that $\left| \sum_{i=1}^{t} a^{-i} \cdot N_i \right|$ is small.

- DIFFICULTY: N_t is correlated with W_{t-1}; hard to prove a useful bounded deviation.

  IDEA. Consider the stopped process where τ is the stopping time $\{M_t > \alpha\}$.

  $$\Rightarrow (M_{t \wedge \tau})_{t=1}^{\infty} \text{ amenable to Freedman's inequality:}$$

  $$\Pr\left[\max_{0 \le t \le T} |M_{t \wedge \tau} - M_0| \gtrsim \varepsilon \right] \lesssim \delta .$$

- DIFFICULTY: how to extend concentration from $(M_{t \wedge \tau})_{t=1}^{\infty}$ to $(M_t)_{t=1}^{\infty}$?

  IDEA. for this stochastic process, choose stopping time criterion so that if the original martingale doesn't deviate too much, then the stopped process doesn't stop:

  $$\Pr\left[\tau > t \,\middle|\, \sup_{1 \le t' \le t} |M_{t'} - M_0| < \varepsilon \right] = 1$$

ANALYSIS.

Let $\tilde{W}_t = W_{t,1} - 1$. Goal: $\tilde{W}_T > -\varepsilon$.

Step 1: linearization

Lemma. For any $t \in \mathbb{N}_0$ and $\eta \in (0,1)$,

$$\tilde{W}_t \geq a \cdot \tilde{W}_{t-1} + N_t$$

for a properly chosen N_t.

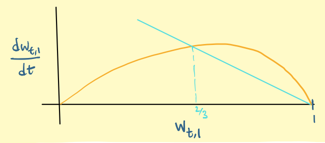
where $a = 1 - \frac{2}{3}(\lambda_1 - \lambda_2)\eta$.

RECALL: continuous analysis to choose a.

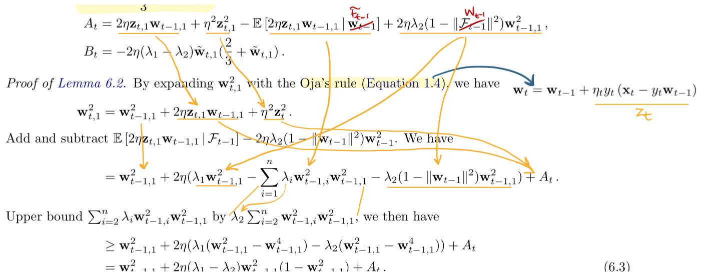
IDEA. if we knew that $W_{0,1} > 2/3$, then

$$\frac{dW_{t,1}}{dt} \geq \frac{2}{3}(\lambda_1 - \lambda_2)(1 - W_{t,1}).$$

choose this as the linear term.



Proof sketch.



$$A_t = 2\eta \mathbf{z}_{t,1}\mathbf{w}_{t-1,1} + \eta^2 \mathbf{z}_{t,1}^2 - \mathbb{E}\left[2\eta \mathbf{z}_{t,1}\mathbf{w}_{t-1,1} \,|\, \overset{\mathcal{F}_{t-1}}{\cancel{\mathbf{w}_{t-1}}}\right] + 2\eta\lambda_2(1 - \|\overset{\mathbf{w}_{t-1}}{\cancel{\mathcal{F}_{t-1}}}\|^2)\mathbf{w}_{t-1,1}^2,$$

$$B_t = -2\eta(\lambda_1 - \lambda_2)\tilde{\mathbf{w}}_{t,1}(\frac{2}{3} + \tilde{\mathbf{w}}_{t,1}).$$

Proof of Lemma 6.2. By expanding $\mathbf{w}_{t,1}^2$ with the Oja's rule (Equation 1.4), we have

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \underbrace{(\mathbf{x}_t - y_t \mathbf{w}_{t-1})}_{z_t}$$

$$\mathbf{w}_{t,1}^2 = \mathbf{w}_{t-1,1}^2 + 2\eta \mathbf{z}_{t,1}\mathbf{w}_{t-1,1} + \eta^2 \mathbf{z}_t^2.$$

Add and subtract $\mathbb{E}[2\eta \mathbf{z}_{t,1}\mathbf{w}_{t-1,1} \,|\, \mathcal{F}_{t-1}] - 2\eta\lambda_2(1 - \|\mathbf{w}_{t-1}\|^2)\mathbf{w}_{t-1}^2$. We have

$$= \mathbf{w}_{t-1,1}^2 + 2\eta(\lambda_1 \mathbf{w}_{t-1,1}^2 - \sum_{i=1}^{n} \lambda_i \mathbf{w}_{t-1,i}^2 \mathbf{w}_{t-1,1}^2 - \lambda_2(1 - \|\mathbf{w}_{t-1}\|^2)\mathbf{w}_{t-1,1}^2) + A_t.$$

Upper bound $\sum_{i=2}^{n} \lambda_i \mathbf{w}_{t-1,i}^2 \mathbf{w}_{t-1,1}^2$ by $\lambda_2 \sum_{i=2}^{n} \mathbf{w}_{t-1,i}^2 \mathbf{w}_{t-1,1}^2$, we then have

$$\geq \mathbf{w}_{t-1,1}^2 + 2\eta(\lambda_1(\mathbf{w}_{t-1,1}^2 - \mathbf{w}_{t-1,1}^4) - \lambda_2(\mathbf{w}_{t-1,1}^2 - \mathbf{w}_{t-1,1}^4)) + A_t$$

$$= \mathbf{w}_{t-1,1}^2 + 2\eta(\lambda_1 - \lambda_2)\mathbf{w}_{t-1,1}^2(1 - \mathbf{w}_{t-1,1}^2) + A_t. \qquad (6.3)$$

Algebra. □

Corollary. For any $t_0 \in \mathbb{N}_0$, $t \in \mathbb{N}$, $\eta \in (0,1)$,

$$\tilde{W}_{t_0+t} \geq a^t \cdot \left(\tilde{W}_{t_0} + \sum_{i=t_0+1}^{t_0+t} a^{t_0-i} \cdot N_i\right).$$

Step 2: controlling the noise

(recall $\tilde{W}_t \geq a \cdot \tilde{W}_{t-1} + N_t$) ← problem if N_t gets too negative

Lemma. Let N_t as above. $\exists A_t, B_t$ s.t. $\forall t \in \mathbb{N}$,

- $N_t = A_t + B_t$
- Bounded difference: ← A_t has bounded deviations

  → $|A_t| = O\left(\eta |\tilde{W}_{t-1}| + \eta |\tilde{W}_{t-1}|^{1/2} + \eta^{3/2}\right)$ a.s.

  → if $\tilde{W}_{t-1} \geq -\frac{2}{3}$, then $B_t \geq -O(\eta^2)$ a.s. ← B_t won't become too negative

- Conditional expectation:

  → $\mathbb{E}[A_t \mid \mathcal{F}_{t-1}] = O(\eta^2 \lambda_1)$ ← A_t's conditional expectation is almost zero

- Conditional variance:

  → $\mathrm{Var}[A_t \mid \mathcal{F}_{t-1}] = O\left(\eta^2 \lambda_1 \left(|\tilde{W}_{t-1}|^2 + |\tilde{W}_{t-1}| + \eta\right)\right)$.

  ↑ A_t has bounded conditional variance

Proof sketch.

Recall:

If η_t's are small enough, w_t's will always be around unit:

> **Lemma.** If $\eta \in (0, \frac{1}{10})$ and if $\forall t$, $\eta_t \leq \eta$, then
>
> $$1 - 10\eta \leq \|W_t\|_2^2 \leq 1 + 10\eta.$$

- bounded difference: Cauchy-Schwarz, algebra, and ↷ .
- conditional expectation/variance: expand definitions.

□

Step 3: martingale analysis setup

Recall:

Lemma (Doob's inequality). Let $(M_t)_{t=0}^{\infty}$ be a martingale.

bounded deviation

Let $T \in \mathbb{N}$, and let $\varepsilon, C \geq 0$. If for $t \in [T]$, $|M_t - M_{t-1}| \leq C$ a.s.,

$$\Pr\left[\sup_{0 \leq t \leq T} |M_t - M_0| \geq \varepsilon\right] \leq \exp\left(-\Omega\left(\frac{\varepsilon^2}{C^2 T}\right)\right).$$

Lemma (Generalized Freedman's). Let $(M_t)_{t=0}^{\infty}$ be a stochastic process. Let $T \in \mathbb{N}$, and let $\varepsilon, C, \mu_t, \sigma_t^2 \geq 0$. If $\forall t \in [T]$,

- bounded deviation: $|M_t - M_{t-1}| \leq C$ a.s.
- small conditional expectation: $\left|\mathbb{E}\left[M_t - M_{t-1} \mid \mathcal{F}_{t-1}\right]\right| \leq \mu_t$
- small conditional variance: $\mathrm{Var}\left[M_t \mid \mathcal{F}_{t-1}\right] \leq \sigma_t^2$

then:

$$\Pr\left[\sup_{0 \leq t \leq T} |M_t - M_0| \geq \varepsilon + \sum_{t=0}^{T} \mu_t\right] \leq \exp\left(-\Omega\left(\frac{\varepsilon^2}{\sum \sigma_t^2 + \varepsilon C}\right)\right).$$

Recall:

$$\tilde{W}_{t_0+t} \geq a^t \cdot \left(\tilde{W}_{t_0} + \sum_{i=t_0+1}^{t_0+t} a^{t_0-i} \cdot N_i\right)$$

$$\geq a^t \cdot \left(\tilde{W}_{t_0} + \sum_{i=t_0+1}^{t_0+t} a^{t_0-i} \cdot (A_i + B_i)\right)$$

not too negative

bounded deviation depending on $\tilde{W}_{t-1}$
bounded conditional expectation & variance

→ Goal: analyze stochastic process of $\left(\sum_{i=1}^{t} a^{-i} A_i\right)_{t=0}^{\infty}$ to show that the accumulated noise never gets too negative.

→ we don't quite satisfy the bounded deviation criterion in Freedman's...

Step 4: analyzing stopped process

Let $M_t = \sum_{i=1}^{t} a^{-i} A_i$. Let τ be the stopping time $\{ \tilde{W}_t < \alpha \}$, where $\alpha \in (-\frac{2}{3}, 0)$. Let $(M_{t \wedge \tau})_{t=0}^{\infty}$ be the stopped process.

<u>Lemma</u>. Given conditions on $\varepsilon, \delta, \eta, a, \alpha$,

$$\Pr\left[\min_{1 \le t \le T} M_{t \wedge \tau} \le \frac{\alpha}{2} \right] < \delta.$$

the stopped process will not become too negative w.h.p.

Pf sketch. Application of Freedman's inequality and lemma from Step 2 controlling the noise. □

Step 5: pulling out stopping time with chaining condition

Main idea: we've shown that a concentration bound holds for a stopped process. If we can show the process never stopped, then bound holds for original process.

<u>Lemma</u>. Let $(M_t)_{t=0}^{\infty}$ be a stochastic process and τ a stopping time. Let $T \in \mathbb{N}$. If we have:

a) $$\Pr\left[\max_{1\le t\le T} M_{t\wedge\tau} \geq \varepsilon\right] < \delta$$

b) $$\forall t \in [T], \quad \Pr\left[\tau > t \;\middle|\; \max_{1\le i\le t} M_i < \varepsilon\right] = 1$$

Then: $$\Pr\left[\max_{1\le t\le T} M_t \geq \varepsilon\right] < \delta .$$

Need to show condition (b):

$$\Pr\left[\tau > t \;\middle|\; \underline{\min_{1\le i\le t} \sum_{j=1}^{i} a^{-j}\cdot\left(A_j + B_j\right) \geq \frac{\alpha}{2}}\right] = 1$$

Recall: $\boxed{\tilde{W}_t \geq a^t \cdot \left(\tilde{W}_0 + \sum_{i=1}^{t} a^{-i}\cdot(A_i + B_i)\right)}$

the right conditions $\Rightarrow$ $\tilde{W}_t \geq \alpha$, so the stopping time

$\tau > t$ a.s.

$\{\tilde{W}_t < \alpha\}$

$\Longrightarrow$ the noise term never gets too negative

<u>Lemma</u>. With the right conditions,

$$\Pr\left[\min_{1\le t\le T} \sum_{i=1}^{t} a^{-i}\cdot N_i \leq \frac{\alpha}{2}\right] < \delta .$$

MAIN TAKEAWAYS.

1. We've sketched out the local convergence proof:

   if w_0 starts out close to the PCA solution v_1

   then, it achieves convergence rate $\tilde{O}\left(\frac{\lambda_1 \log \frac{1}{\varepsilon}}{\varepsilon(\lambda_1 - \lambda_2)^2}\right)$.

   a) split the dynamics into a linear term + noise term

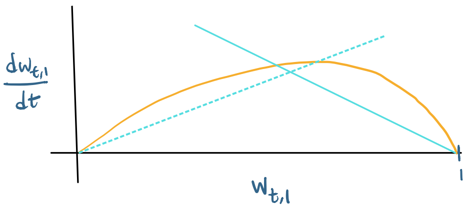
      - noise term reduced by choosing smaller learning rate η
      - convergence rate of linear term increases with larger learning rate η

   b) martingale analysis on noise term

      - bounded deviation achieved via stopped process
      - analysis for stopped process extended to original process by choosing stopped process with suitable chaining condition (i.e. process doesn't stop if nothing bad happened in original).

2. Analysis for global convergence:

   - linearize at other point
   - need to handle crossing between regimes



APPENDIX.

In the following, let $(X_t)_{t=0}^{\infty}$ be a stochastic process, and $(\mathcal{F}_t)_{t=0}^{\infty}$ be its natural filtration.

DEFINITION (Martingale). $(X_t)_{t=0}^{\infty}$ is a martingale if for each $t \in \mathbb{N}$,

$$\mathbb{E}[M_t - M_{t-1} \mid \mathcal{F}_{t-1}] = 0.$$

DEFINITION (Stopping time). An $\mathbb{N}_0$-valued random variable τ is a stopping time for $(X_t)_{t=0}^{\infty}$ if for all $t \in \mathbb{N}_0$,

$$\{\tau = t\} \subset \mathcal{F}_t.$$

DEFINITION (Stopped process). If τ is a stopping time for $(X_t)_{t=0}^{\infty}$, then the stochastic process $(X_{t \wedge \tau})_{t=0}^{\infty}$ is the stopped process, where

$$X_{t \wedge \tau} = \begin{cases} X_t & t \leq \tau \\ X_\tau & t > \tau \end{cases}.$$