Guest lecture: Adaptive Online learning and distribution Shift (Continuous)

Setting: Nature Chooses $P_1, P_2, \cdots, P_n$

1. $(x_t, y_t) \sim P_t$, x_t is revealed to learner
2. learner generates $\hat{y}_t = f_{\theta_t}(x_t) = \theta_t^T x_t$
   chooses θ_t
3. Incurs loss $(\hat{y}_t - y_t)^2$. $\ell((x_t, y_t), \hat{y}_t)$

No assumption on $P_1 \cdots P_n$

$\theta_t^* = \arg\min \mathbb{E}_{(x,y)\sim P_t} (x^T\theta - y)^2$

Example 0: $P_1 = P_2 = \cdots = P_n$ iid setting

Example (Covariance Shift): $P_t(x,y) = \underbrace{P_t(x)} \cdot P(y|x)$

Example (Concept Shift): $P_t(x,y) = P(x) \cdot \underbrace{P_t(y|x)}$

Example (Label shift) $P_t(x,y) = \underbrace{P_t(y)} \cdot P(x|y)$

No regret Online learning:

1. (x_t, y_t) is chosen by nature. (adversarial setting)
2. $\sim \theta_t \to \hat{y}_t$
3. $\sim$ loss$(\hat{y}_t - y_t)^2$

Static Regret:

$$\underbrace{\sum_{t=1}^{n} (x_t^T\theta_t - y_t)^2}_{\text{your performance}} - \underbrace{\sum_{t=1}^{n} (x_t^T u - y_t)^2}_{u = \theta^* = \arg\min_\theta \sum_{t=1}^{n} (x_t^T\theta - y_t)^2} = \underbrace{o(n)}_{O(\log n)}$$

Static Comparator

Alg 1: Vovk-Azoury-~~Warmuth~~ Warmuth VAW Forecaster

Alg 2: Online Newton Step

$\ell_t(\theta) = (x^T\theta - y)^2$

$\nabla^2 \ell_t(\theta) = 2xx^T \succeq 0$

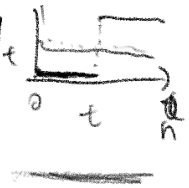
$\succeq \alpha \cdot \nabla \ell(\theta) \nabla \ell(\theta)^T$

exp concave

It doesn't work well with non-stationarity.

Example: $x_t = 1 \ \forall t$, $y_t = 0$ for $t = 1, 2, \cdots \frac{n}{2}$

$y_t = 1$ for $t = \frac{n}{2}+1, \cdots, n$

y_t 

$$\underbrace{\min_{\theta} \sum_{t=1}^{n} (\theta - y_t)^2}_{= \frac{1}{4} n} \qquad \theta^* = \frac{1}{2}$$

$$\sum_{t=2}^{n} (\theta_t - y_t)^2 \le \frac{1}{4} n + O(\log n)$$

Dynamic Regret (Universal Dynamic Regret)

$$\underset{(u_{1:n})}{\text{DynReg}} := \sum_{t=1}^{n} (x_t^T \theta_t - y_t)^2 - \sum_{t=1}^{n} (x_t^T u_t - y_t)$$

$u_1 \cdots u_n =: u_{1:n}$ is a seq of Comparators

(*) Sublinear Dyn Regret is impossible.

Example $x_t = 1 \ \forall t$, $y_t \sim \mathrm{Ber}(0.5)$ i.i.d., $\boxed{u_t = y_t}$

$$\mathbb{E}[\text{Dynamic Regret}] \ge \Omega(n)$$

$$\mathbb{E}\Big[\sum_t (\hat{y}_t - y_t)^2 - \underbrace{(u_t - y_t)^2}_{=0}\Big] \ge \sum_{t=1}^{n} \mathbb{E}[(0.5 - y_t)^2] = \frac{1}{4} n$$

What can we still do?

Def: $TV(u_{1:n}) = \sum_{t=2}^{n} \|u_t - u_{t-1}\|_1$

we can parametrize Regret with $TV(u_{1:n})$

poly(d) dependence omitted

(Zinkevich, 2003) $\sqrt{n}\,(1 + TV(u_{1:n}))$

(Zhang, Zhou, 2016) $\sqrt{n(1 + TV(u_{1:n}))}$ ← optimal for Convex-Lipschitz

(Baby and W., 2021) $n^{\frac{1}{3}} (TV(u_{1:n}))^{\frac{2}{3}} + \log n$. optimal for exp-concave losses

---

other description of $u_{1:n}$

changes. $\ell_0\text{-}TV_{(u_{1:n})} = \sum_{t=2}^{n} \mathbb{1}(u_{t-1} \ne u_t)$

$\ell_2\text{-}TV(u_{1:n}) = \sum_{t=2}^{n} \|u_{t-1} - u_t\|_2$ "dimension-free"

In fact Baby and W.

$$\text{D. Regret} \le \log n \cdot (1 + \#\text{ of changes}) \wedge \left(n^{\frac{1}{3}} TV(u_{1:n})^{\frac{2}{3}} + \log n\right)$$

in fact.

$\overbrace{\qquad\qquad}^{P}$

$\underline{3\log n} + n^{\frac{1}{3}} TV^{\frac{2}{3}}$ (sequence, excluding change points)

- nonparametric regression

$$\ell_t(\theta) = (f_0(x_t) + \varepsilon_t - \theta_t)^2 \qquad \boxed{x_t = 0, \tfrac{1}{n}, \tfrac{2}{n}, \dots, \tfrac{n-1}{n}, 1}$$

$$\underbrace{\mathbb{E}\left[\sum_{t=1}^{n} (f_0(x_t) - \theta_t)^2\right]}_{MSE} = \mathbb{E}\left[\sum_{t=1}^{n} ((f_0(x_t) + \varepsilon_t - \theta_t)^2)\right] - \underbrace{\mathbb{E}\left[\sum_{t=1}^{n} \varepsilon_t^2\right]}_{n\sigma^2}$$

Optimal Rate for Holder class / Sobolev class, Besov classes, TV class. $u_t = f_0(x_t)$

- Apply to non-stochastic control (LQR) Baby and W. (NeurIPS 22)

$\times$ Optimal Regret

Apply to nonstationary supervised learning

$$(x_t, y_t) \sim P_t, \quad \theta_t^* := \mathbb{E}_{(x,y)\sim P_t}[(y - x^T\theta)^2]$$

"Covariate shift" $y_t \sim N(\theta_0^T x_t, \sigma^2) \Rightarrow \theta_t^* \equiv \theta_0$

online learner gives $O(\log n)$ static Regret

"Concept shift" $x_t \sim P, \; y_t \sim N(\theta_t^{*T} x_t, \sigma^2)$

(quiz: concept + covariate shift) $u_t = \theta_t^*$, Dyn. Regret $(\theta_1^* \dots \theta_n^*) \le O(n^{\frac{1}{3}} TV(\theta_{1:n}^*)^{\frac{2}{3}} + \log n)$

"label shift"

$$\theta_t^* = \underbrace{\mathbb{E}_{y\sim P_t(\cdot)} \mathbb{E}_{x\sim P(x|y)}[\ell((x,y),\theta)]}_{L_t(\theta)}$$

Assume L_t is λ-strongly convex in $\|\cdot\|$

$$\boxed{\text{Lem from Convex Opt: } \|\theta_t^* - \theta_{t-1}^*\| \le \|\nabla_\theta (L_t - L_{t-1})(\theta_{t-1}^*)\|_*}$$

$$L_t(\theta) = \sum_{y\in Y} P_t(y) \cdot E_x[\ell((x,y),\theta)] = \langle P_t, \mathbb{E}_x[\ell(x,\cdot)] \rangle$$

$$\nabla_\theta (L_t - L_{t-1}) = \langle P_t - P_{t-1}, \nabla_\theta \mathbb{E}_x[\ell(x,\cdot),\theta)] \rangle$$

$$\|\nabla(L_t - L_{t-1})(\theta_t^*)\| < \|P_t - P_{t-1}\|_1 \cdot \max_y \|\nabla_\theta E_x[\ell((x,y),\theta_t^*)]\|_*$$

$\leq G$

$$TV(\theta^*_{1:n}) \leq \underbrace{\sum_t \|P_t - P_{t-1}\|_1}_{\text{cumulative distshift}} \cdot G$$

$$\boxed{D\text{ Regret} = O\left(n^{\frac{1}{3}} \cdot TV(P_{1:n})^{\frac{2}{3}} \cdot G^{\frac{2}{3}} \cdot \log n\right)}$$

What if $\lambda = 0$?

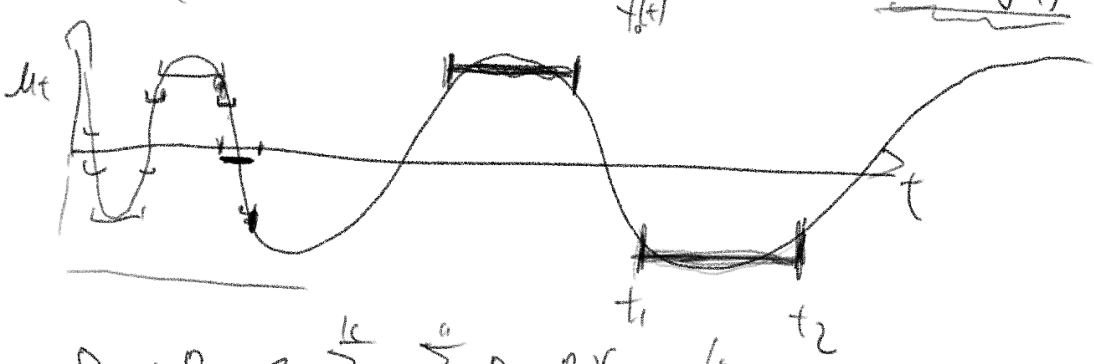
Compare with $\theta^*_{t,\lambda} = \arg\min \mathbb{E}_{y|x_t} \mathbb{E}_x \left(\ell(x,y,\theta)\right) + \lambda\gamma_\theta - \lambda\|\theta - \theta_t\|^2$

instantiate $u_t = \theta^*_t$ $\boxed{\theta^*_t \text{ is optimal for } P_t}$

$$\boxed{\underbrace{\sum_{t=1}^{n} \ell(\theta_t, (x_t, y_t))}_{\text{your learner}} \leq \min_{u_{1:n}} \left\{ \sum_t \ell(u_t, (x_t, y_t)) + \text{Regret}(u_{1:n}) \right\}}$$

Algorithmic Idea (Strongly Adaptive Online learner)

$\ell_t(\theta_t) = (y_t - \theta_t)^2$ $y_t \sim N(\mu_t, \sigma^2)$ $f(t)$ "Adaptive Regret" $\log n$



$$\text{Dynamic Regret} \leq \sum_{i=1}^{k} \sum_{t: t_{is} - t_{ie}} \underbrace{\text{DynamicRegret}(t_{is}:t_{ie})}_{\leq \log(n)} \log n$$

$k \simeq n^{\frac{1}{3}} TV(u_{1:n})^{\frac{2}{3}}$

$\boxed{C_i \geq \frac{B^2 \epsilon}{\sqrt{n_i}}}$ Bandelus

$TV(u_{t:t+1:n})$ length of interval i

invoke adaptive regret bound

$$C_n \geq \sum_{i=1}^{k} C_i \geq \sum_{i=1}^{k} \frac{B^2}{\sqrt{n_i}} = k \sum_{i=1}^{k} \frac{1}{k} \frac{B^2}{\sqrt{n_i}} \overset{\text{Jensen}}{\geq} k \frac{B^2}{\sqrt{\frac{n}{k}}}$$

"magical Rule"

$$k \le C n^{\frac{2}{3}} n^{\frac{1}{3}} B^{\frac{4}{3}}$$

$\frac{1}{\sqrt{x}}$ ~~Concave~~ Convex

$$= \frac{k}{k^{\frac{1}{2}}} \frac{B^2}{\sqrt{n}} = k^{\frac{3}{2}} \frac{B^2}{\sqrt{n}}$$

---

① Adaptive Restarting: "change point detector" checking "magical Rule"

② "1-edge" over many learner

$\sqrt{n}\, T(n,n)$