Lecture 10
Page 1

CS 59800 - RL

Linear Regression.

Agenda:

- Ridge regression
- Concentration bound

Consider a linear model of

$$X_t = \langle A_t, \theta_* \rangle + \eta_t$$

X_t: scalar; η_t: scalar; $A_t \in R^d$; $\theta_* \in R^d$

and a sequence of $A_1, X_1, \dots, A_n, X_n$, with a filtration $\mathbb{F} = \{ \mathcal{F}_t \}$, such that, $\mathcal{F}_t = \sigma(A_1, X_1, \dots A_{t+1})$

Note that, X_t is $\mathcal{F}_t$ measurable.

For noise η_t, we have:

$$E[\exp(\alpha \eta_t) / \mathcal{F}_{t-1}] \leq \exp\left(\frac{\alpha^2}{2}\right)$$

for all $\alpha \in R$ and $t \in [n]$.

This is the 1-sub-Gaussian assumption on the noise.

The process is as follows:

- At each time step t, someone chooses A_t from a set D_t, and we observe X_t.

Given $A_1, X_1, \dots A_t, X_t$,

Can we estimate θ_*?

- How about solving a ridge regression for θ_* ?

→ at time t,

$$\min_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t} (X_s - \langle A_s, \theta \rangle)^2 + \|\theta\|_V^2$$

For positive Definite matrix V

$$\|\theta\|_V^2 = \theta^T V \theta$$

Let's define

- $S_t = \sum_{s=1}^{t} \eta_s A_s$
- $V_t = \sum_{s=1}^{t} A_s A_s^T$
- $V_t(V) = V + V_t = V_t(V)$

As you may have seen, we usually set $V = \lambda I$ for $\lambda > 0$

- $V_t(\lambda) = \lambda I + V_t$

Therefore, the minimizer of ridge regression problem is

$$\hat{\theta}_t = V_t(V)^{-1} \sum_{s=1}^{t} X_s A_s$$ (why?)

we are interested is how good is this $\hat{\theta}_t$ estimate.

To simplify the notation, we use $V = \lambda I$

Let's define $\bullet$ $M_t(\alpha) = \exp\left(\langle \alpha, S_t \rangle - \frac{1}{2} \|\alpha\|^2_{V_t}\right)$
for $\alpha \in \mathbb{R}^d$

Lemma: For any $\alpha \in \mathbb{R}^d$, the process $M_t(\alpha)$ is an $\mathbb{F}$-adapted supermartingale.

Proof:

– It is clear that $M_t(\alpha)$ is $\mathcal{F}_t$-measurable for all t by definition. We are left to show

$$E\left[M_t(\alpha) \mid \mathcal{F}_{t-1}\right] \leqslant M_{t-1}(\alpha) \quad \text{a.s.}$$

Let's expand $M_t(\alpha) \rightarrow$

$$E\left[M_t(\alpha) \mid \mathcal{F}_{t-1}\right] = E\left[\exp\left(\langle \alpha, S_t \rangle - \frac{1}{2}\|\alpha\|^2_{V_t}\right) \Big| \mathcal{F}_{t-1}\right]$$

$$= E\left[\exp\left(\langle \alpha, S_{t-1} \rangle - \frac{1}{2}\|\alpha\|^2_{V_{t-1}}\right) \exp\left(\eta_t \langle \alpha, A_t \rangle - \frac{1}{2}\|\alpha\|_{A_t A_t^T}\right) \Big| \mathcal{F}_{t-1}\right]$$

$$= M_{t-1}(\alpha) \underbrace{E\left[\exp \eta_t \langle \alpha, A_t \rangle - \frac{1}{2}\|\alpha\|_{A_t A_t^T} \Big| \mathcal{F}_{t-1}\right]}_{\leqslant 1 \text{ a.s.}}$$

$$\leqslant M_{t-1}(\alpha) \quad \text{a.s.}$$

↳ we concluded that $M_t(\alpha)$ is a supermartingale sequence.

Page 4

For a Gaussian measure h with covariance V and mean zero let's define:

$$\overline{M}_t = \int_{R^d} M_t(\alpha)\, dh(\alpha)$$

now using Radon - Nikodym derivative, we have:

$$\overline{M}_t = \frac{1}{(2\pi)^{d/2} \det(V^{-1})^{\frac{1}{2}}} \int_{R^d} \exp\left(\langle \alpha, S_t\rangle - \frac{1}{2}\|\alpha\|^2_{V_t} - \frac{1}{2}\|\alpha\|^2_{V}\right) d\alpha$$

Note that:

$$\|S_t\|^2_{\underbrace{(V+V_t)^{-1}}_{V_t(V)}} - \|\alpha - (V+V_t)^{-1} S_t\|^2_{V+V_t}$$

$$= 2\langle \alpha, S_t\rangle - \|\alpha\|^2_{V_t} - \|\alpha\|^2_{V} \quad \text{(why?)}$$

$$\Rightarrow \overline{M}_t = \frac{1}{(2\pi)^{d/2} \det(V^{-1})^{\frac{1}{2}}} \int_{R^d} \exp\left(\|S_t\|^2_{V_t(V)^{-1}} \times \frac{1}{2} - \frac{1}{2}\|\alpha - V_t(V)^{-1} S_t\|^2_{V_t(V)}\right) d\alpha$$

$$= \frac{1}{(2\pi)^{d/2} \det(V^{-1})^{\frac{1}{2}}} \underline{\exp\left(\frac{1}{2}\|S_t\|^2_{V_t(V)^{-1}}\right)} \int_{R^d} \exp\left(-\frac{1}{2}\|\alpha - V_t(V)^{-1} S_t\|^2_{V_t(V)}\right) d\alpha$$

$$= \frac{\det(V_t(V)^{-1})^{\frac{1}{2}}}{\det(V^{-1})^{\frac{1}{2}}} \exp\left(\frac{1}{2}\|S_t\|^2_{V_t(V)^{-1}}\right) \times \underbrace{\left(\frac{1}{(2\pi)^{d/2} \det\left((V_t(V))^{-1}\right)^{\frac{1}{2}}} \int_{R^d} \exp\left(-\frac{1}{2}\|\alpha - V_t(V)^{-1} S_t\|^2_{V_t(V)}\right) d\alpha \right)}_{\text{equal to } 1}$$

Page 5

Note that $\bar{M}_t$ is also supermartingale (why?)

Using the general form of maximal inequality, we have:

$$\mathbb{P}\left(\sup_t \bar{M}_t > \frac{1}{\delta}\right) \leqslant \delta$$

we know that $\left\{ t\in[n] ;\ \frac{\det\left(V_t(V)\right)^{\frac{1}{2}}}{\det(V^{-1})^{\frac{1}{2}}} \exp\left(\frac{1}{2}\|S_t\|_{V_t(V)^{-1}}\right) > \frac{1}{\delta} \right\}$

$$\subset \left\{ \sup_t \bar{M}_t > \frac{1}{\delta} \right\}$$

$$\hookrightarrow \mathbb{P}\left(t\in[n];\ \frac{\det(V_t(V)^{-1})^{\frac{1}{2}}}{\det(V^{-1})^{\frac{1}{2}}} \exp\left(\frac{1}{2}\|S_t\|_{V_t(V)^{-1}}\right) > \delta\right) \leqslant \delta$$

$$\hookrightarrow \mathbb{P}\left(t\in[n];\ \|S_t\|_{V_t(V)^{-1}} > 2\log\frac{1}{\delta} + \log\left(\frac{\det(V_t(V))}{\det(V)}\right)\right) \leqslant \delta$$

Theorem: For $\delta \in (0,1)$, with probability at least $1-\delta$, for $t\in[n]$, we have; for $V=\lambda I$

$$\|\hat{\theta}_t - \theta_*\|_{V_t(V)} \leqslant \sqrt{\beta_t(\delta)} : \sqrt{\lambda}\,\|\theta_*\| + \underbrace{\sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_t(V))}{\det(V)}\right)}}$$

Furthermore if $\|\theta_*\| \leqslant S$, define confidene interval/set

For t-1

$$C_t(\delta) = \left\{\theta\in\mathbb{R}^d :\ \|\hat{\theta}_{t-1} - \theta\|_{V_{t-1}(V)} \leqslant \sqrt{\lambda}\, S + (\ \ ) \right\}$$

Page 6]

Then $\mathbb{P}\left(\text{exists } t \in [n] ;\ \theta_* \notin C_t(\delta)\right) \leq \delta$

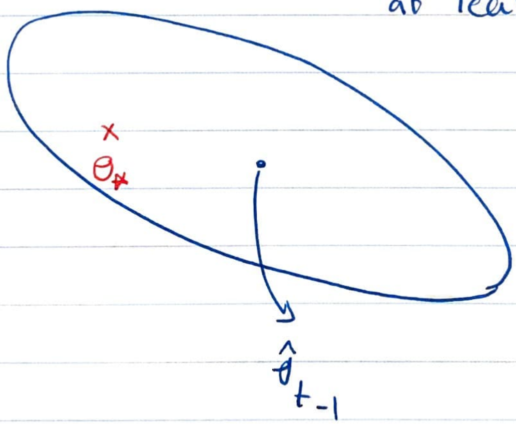
where $C_t(\delta) = \Big\{ \theta \in \mathbb{R}^d : \|\hat\theta_{t-1} - \theta\|_{V_{t-1}(V)}$

$$\leq \sqrt{\lambda}\, S + \sqrt{2 \log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_{t-1}(V))}{\det(V)}\right)} \Big\}$$

what does it mean?

$$\left\{ \theta \in \mathbb{R}^d ;\ \|\hat\theta_{t-1} - \theta\|^2_{V_{t-1}(\lambda)} \leq \beta_{t-1}(\delta) \right\}$$

is an elipse such that θ_* is in it, always, with propability at least $1-\delta$



Can we simplify $\sqrt{\beta_t(\delta)}$? So far, we did not use the fact that we want to set $V = \lambda I$. The resulbs holds for any positive definit V. For $V = \lambda I \rightarrow \det(V) = \lambda^d$