Bandit PCA

Thursday, September 24, 2020 7:24 PM

Interaction protocol

For $t = 1, \ldots, T$

- learner picks $w_t \in \mathbb{R}^d$, $\|w\|_2 = 1$, in randomized fashion
- environment picks $L_t \in \mathbb{R}^{d \times d}$, $\|L_t\|_2 \leq 1$
- learner observes <u>loss</u> $\langle w_t w_t^T, L_t \rangle = tr(w_t w_t^T L_t)$

<u>Note</u>: - L_t is loss matrix

- allowed to depend on entire history (except w_t).
- thus, adaptive adversaries allowed.

<u>Goal</u>: minimize regret,

$$R_T = \max_{u : \|u\|=1} \sum_{t=1}^{T} \mathbb{E}\left[\langle w_t w_t^T - u u^T, L_t \rangle\right]$$

↳ over randomization of learner

<u>Algorithm</u> <u>Online mirror descent</u>

Param: $\eta > 0$ (learning rate), $\gamma \in [0,1]$ (exploration rate)

Init: $W_1 = \frac{1}{d} I_d$

for $t = 1, \ldots, T$

$$W_t \overset{EVD}{=} \sum_{i=1}^{d} \mu_i u_i u_i^T$$

$$\underline{\lambda} = (1-\gamma)\underline{\mu} + \gamma(1/d, \ldots, 1/d) \rightarrow \text{controls exploration, exploitation trade-off}$$

$$\tilde{L}_t = \underline{sample}(\lambda, \{\mu_i\}_{i=1}^{d})$$

↳ 2 possible sub-routines

$$W_{t+1} = (W_t^{-1} + \eta \tilde{L}_t + \beta I)^{-1} \text{ with } \beta \text{ s.t. } tr(W_{t+1}) = 1$$

$$W_{t+1} = \left(W_t^{-1} + \eta \tilde{L}_t + \beta I\right)^{-1} \text{ with } \beta \text{ s.t. } \operatorname{tr}(W_{t+1}) = 1$$

<u>Note</u> : $W \triangleq (1-\gamma) W_t + \frac{\gamma}{d} I_d$ here

A. Dense sampling

Sample $(\underline{\lambda}, \{u_i\}_{i=1}^d)$

$B \sim \text{bern}(1/2)$

to est. diag. elements

if $B = 1$

$I \sim \lambda$, $w_t \leftarrow u_I$

- sample one eig vect s.t. $p(I=i) = \lambda_i$

$E[w_t w_t^T] = E\left[\sum_{i=1}^d \mathbb{1}_{\{I=i\}} u_i u_i^T\right] = \sum \lambda_i u_i u_i^T = W$

to estimate off-diag elements

else

$s \in \{-1,1\}^d$ iid uniformly

$w_t \leftarrow \sum_i s_i \sqrt{\lambda_i}\, u_i$

$E[w_t w_t^T] = E_s\left[\sum_{i,j} s_i s_j \sqrt{\lambda_i \lambda_j}\, u_i u_j^T\right] = \sum_{i,j} \delta_{ij} \sqrt{\lambda_i \lambda_j}\, u_i u_j^T = W$

play w_t, observe $\langle w_t w_t^T, L_t \rangle = l_t$

$\langle u_i u_j^T, L_t \rangle = \operatorname{tr}(u_i u_j^T L_t) = u_j^T L_t u_i$

if $B = 1$

$\tilde{L}_t \leftarrow 2 l_t W_t^{-1/2} w_t w_t^T W_t^{-1/2}$

- intuitevly, observe we can estimate $\approx u_i^T L_t u_j$

else

$\tilde{L}_t \leftarrow l_t \left(W_t^{-1} w_t w_t^T W_t^{-1} - W_t^{-1}\right)$

lemma1 $\rightarrow E[\tilde{L}_t] = L_t$

return $\tilde{L}_t$

<u>Bounding regret</u> $\left(R_T = \max_{u:\|u\|=1} \sum_{t=1}^T E\left[\langle w_t w_t^T - u u^T, L_t \rangle\right]\right)$

<u>Theorem 3</u> : let $\eta \leq \frac{1}{2d}$, $\gamma = 0$. Then,

$$R_T \leq \frac{d \log T}{\eta} + \eta (d^2+1) \sum_{t=1}^T E[l_t^2] + 2$$

{ not so necessary for first read I think

<u>Corollary</u> : let $\eta = \min\left\{\sqrt{\frac{\log T}{dT}}, \frac{1}{2d}\right\}$, $\gamma = 0$

$$R_T \leq O\left(d^{3/2} \sqrt{T \log T}\right)$$

T^*

$$R_T \leq O\left(d^{3/2}\sqrt{T\log T}\right)$$

- if assume that $L_t \succeq 0$, $\min_{\|u\|_2 = 1} \sum_t \operatorname{tr}(uu^T L_t) \leq L_T^*$

↳ approx. says if $\forall t$, $\|L_t\| = O_T(1)$ then it gives better bnd

then with $\eta = \min\left\{\sqrt{\frac{\log T}{d L_T^*}}, \frac{1}{4d^2}\right\}$

$$R_T = O\left(d^{3/2}\sqrt{L_T^* \log T} + d^3 \log T\right)$$

B. Sparse sampling

Sample $(\lambda, \{u_i\}_{i=1}^d)$

draw $I, J \sim \lambda$

if $I = J$

$w_t \leftarrow u_I$

else

$s \in \{-1, 1\}$ uniformly

$w_t \leftarrow \frac{1}{\sqrt{2}}(u_I + s u_J)$

Same as in dense case, $E[w_t w_t^T] = W$

little more work, but can show $E[w_t w_t^T] = W$

play w_t, observe $l_t = \langle w_t w_t^T, L_t \rangle$

if $I = J$

$$\tilde{L}_t = \frac{l_t}{\lambda_I^2} u_I u_I^T$$

else

$$\tilde{L}_t = \frac{s\, l_t}{2\lambda_I \lambda_J}\left(u_I u_J^T + u_J u_I^T\right)$$

lemma 2 shows that $E[\tilde{L}_t] = L_t$

Bounding regret

Thm 6: $\eta \leq 1/2d$, $\gamma = \eta d$.

$$R_T \leq \frac{d \log T}{\eta} + 2\eta d + 2 + 8\eta d \sum_t E\left[\|L_t\|_F^2\right]$$

$E[\|L_t\|^2] \leq r$ if $\eta = \min\left\{\sqrt{\frac{\log T}{rT}}, \frac{1}{2d}\right\}$

Cor : let $\frac{1}{T}\sum E\left[\|L_t\|_F^2\right] \leq r$. if $\eta = \min\left\{\sqrt{\frac{\log T}{rT}}, \frac{1}{2d}\right\}$

$\gamma = \eta d$. then

$$R_T = O\left(d\sqrt{rT \log T}\right)$$

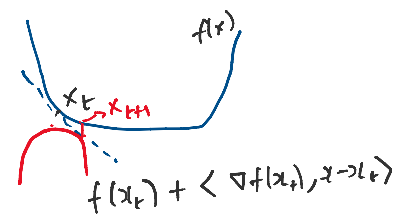
note : since $\|L_t\|_2 \leq 1$, $\|L_t\|_F \leq \text{rank}(L_t)$

$\Rightarrow$ for $L_t = x_t x_t^T$ (online PCA), $R_T \leq O\left(d\sqrt{T\log T}\right)$

Understanding the algorithm (Online mirror descent)

(aside) mirror descent ! (from Yuxin's slides)

proximal viewpoint of projected G.D



$$- \; x_{t+1} = \arg\min_{x \in C}\left\{\underbrace{f(x_t) + \langle \nabla f(x_t), x - x_t\rangle}_{\text{linear approx}} + \underbrace{\frac{1}{2\eta_t}\|x - x_t\|_2^2}_{\text{prox}}\right\}$$

- the proximal term is based on geometry of problem
- e.g. if C is probability simplex, better idea to use
    - KL divergence, T.V. distance, ...

mirror descent

$$x_{t+1} = \arg\min_{x \in C}\left\{f(x_t) + \langle \nabla f(x_t), x - x_t\rangle + \underbrace{\frac{1}{\eta_t} D_\psi(x, x_t)}_{\text{Bregman divergence}}\right\}$$

$$D_\psi(x,y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x-y \rangle$$

essentially the difference in first order approximation!

- key point: choice of regularization and associated Bregman is crucial

Back to Bandit PCA:

- Consider differentiable, convex $R: \mathcal{S} \to \mathbb{R}$, $\mathcal{S} \to d \times d$ symmetric, PSD
- associated $D_R : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$

$$D_R(W \| W') = R(W) - R(W') - \langle \nabla R(W'), W - W' \rangle$$

- in Algo 1, initialize $W_1 = \frac{1}{d} I_d$

$$W_{t+1} = \underset{W \in \mathcal{W}}{\arg\min} \{ \eta \langle W, \tilde{L}_t \rangle + D_R(W \| W^t) \}$$

$\mathcal{W} \subset \mathcal{S}$ with $tr(W) = 1$

- choose $R(W) = -\log \det(W)$

another aside

$$\frac{\partial(-\log \det W)}{\partial W_{ij}} = \frac{-1}{\det(W)} \cdot \frac{\partial \det W}{\partial W_{ij}} = \frac{-1}{\det(W)} \cdot adj(W)_{ij} = -(W^{-1})_{ij}$$

then,
$$D_R(W \| U) = -\log\det(W) + \log\det(U) + \underbrace{\langle U^{-1}, W - U \rangle}_{= tr(U^{-1}W) - tr(I)}$$

$$= -\log\det(W) + \log\det U + tr(U^{-1}W) - d$$

$$= \log\det(UW^{-1}) + tr(U^{-1}W) - d$$

$$= tr(U^{-1}W) - \log\det(U^{-1}W) - d \quad \square$$

- Algo 1 iterates can be rewritten as

$$\tilde{W}_{t+1} = \arg\min_{W} \{ \eta \, tr(W \tilde{L}_t) + D_R(W \| W_t) \} \quad \text{(update)}$$

$$W_{t+1} = \arg\min_{W \in \mathcal{W}} D_R(W \| \tilde{W}_{t+1})$$

- $\tilde{W}_{t+1}$ satisfies (not verified yet)

$$\nabla R(\tilde{W}_{t+1}) = \nabla R(W_t) - \eta \tilde{L}_t$$

$$\Rightarrow \quad \tilde{W}_{t+1}^{-1} = W_t^{-1} + \eta \tilde{L}_t$$

$$\Rightarrow \quad \tilde{W}_{t+1} = (W_t^{-1} + \eta \tilde{L}_t)^{-1} = W_t^{1/2} (I + \eta W_t^{1/2} \tilde{L}_t W_t^{1/2})^{-1} W_t^{1/2}$$

<u>Lemma</u> 9: (from an older reference) for any $\eta > 0$, $\gamma \in [0,1]$

$$R_T \le \frac{d \log T}{\eta} + 2\gamma T + 2(1-\gamma) \underbrace{\sum_{t=1}^{T} E[\langle W_t - \tilde{W}_{t+1}, \tilde{L}_t \rangle]}_{T1}$$

let $B_t = W_t^{1/2} \tilde{L}_t W_t^{1/2}$, then

$$\tilde{W}_{t+1} = W_t^{1/2} (I + \eta B_t) W_t^{1/2} = W_t - \eta W_t^{1/2} B_t (I + \eta B_t)^{-1} W_t^{1/2}$$

Use

$$(I+A)^{-1} = I + (I+A)^{-1} - I$$
$$= I + (I+A)^{-1} - (I+A)(I+A)^{-1}$$
$$= I + (I - I - A)(I+A)^{-1}$$
$$= I - A(I+A)^{-1}$$

$$\langle W_t - W_{t+1}, L_t \rangle = \eta \, tr(W_t^{1/2} B_t (I + \eta B_t)^{-1} W_t^{1/2} \tilde{L}_t)$$

$- \eta \, tr(B_t (I + \eta B_t)^{-1} B_t)$

b_{ti} - eig vals of B_t

$$= \eta \sum_{t=1}^{d} \frac{b_{ti}^2}{1+\eta b_{ti}}$$

- these are bounded separately for dense, sparse cases
- lemma 10 bounds $T1_t \leq \eta(d^2+1)\, l_t^2$ (dense)
- lemma 11 bounds $T1_t \leq 8\eta d \, \|L_t\|_F^2$ (sparse)

Running time:
- Dense - $\tilde{O}(d^3)$ per trial
- sparse $\tilde{O}(d)$ per trial

Possible extensions

- what about non-square loss matrices?
- does looking at $w_t \in \mathbb{R}^{d \times r}$ make sense?
- can we look at sparse PCA?
- some other structure on w_t / L_t ...... ?
- missing data?? ie if we only see $tr(w_t w_t^T P_\Omega(L_t))$?
  or other places for $P_\Omega(\cdot)$ as well
  - check contaminated data paper by Nowak grp for "good" models.