2022

Lecture 13: Self-normalized Processes and Linear Bandits

Week 13

Lecturer: Tianyu Wang

Makeup for Regression Trees first.

1 Regression Trees

Recall in classification trees/decision tree classifiers, the model is greedily trained using entropy as the impurity measure.

For regression trees, the entropy is replaced with variance. I'll draw an example.

1.1 Random Forest

Consider a dataset $\{(x_i, y_i)\}_{i=1}^n$. A random forest $\widehat{f}$ fits K $i.i.d.$ regression trees (or classification trees) in the following way.

- For $k = 1, 2, \cdots, K$, draw an $i.i.d.$ random dataset of size M from $\{(x_i, y_i)\}_{i=1}^n$ (usually repetition is allowed). On this "new" dataset, fit a regression tree $\widehat{f}_k$.
- The random forest is an average of the deccision trees: $\widehat{f} = \frac{1}{K}\sum_{k=1}^K \widehat{f}_k$.

1.2 Variance reduction

At any x, the variance of the prediction of the random forest model is

$$
\begin{aligned}
& \mathbb{E}\left[\left(\widehat{f}(x) - \mathbb{E}\left[\widehat{f}(x)\right]\right)^2\right] \\
= & \mathbb{E}\left[\left(\frac{1}{K}\sum_{k=1}^K \widehat{f}_k(x) - \frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\widehat{f}_k(x)\right]\right)^2\right] \\
= & \frac{1}{K^2}\sum_{k=1}^K \mathbb{E}\left[\left(\widehat{f}_k(x) - \mathbb{E}\left[\widehat{f}_k(x)\right]\right)^2\right] + \frac{1}{K^2}\sum_{i,j:i\neq j}^K \underbrace{\mathbb{E}\left[\left(\widehat{f}_i(x) - \mathbb{E}\left[\widehat{f}_i(x)\right]\right)\left(\widehat{f}_j(x) - \mathbb{E}\left[\widehat{f}_j(x)\right]\right)\right]}_{\text{covariance}}.
\end{aligned}
$$

If all $\widehat{f}_k$ are trained in an $i.i.d.$ way, the coorelation of two models are small, thus the variance of the random forest is usually smaller than the variance of each tree in the forest.

2 Self-normalized Processes

Theorem 2.1 (Self-Normalized Bound for Vector-Valued Martingales)**.** *Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a sequence of filtered σ-algebra. Let $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that η_t is $\mathcal{F}_t$-measurable and η_t is conditionally R-sub-Gaussian for some $R > 0$, i.e.*

$$\mathbb{E}\left[e^{\lambda \eta_t} | \mathcal{F}_{t-1}\right] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right), \qquad \forall \lambda \in \mathbb{R}.$$

Let $\{X_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is $\mathcal{F}_{t-1}$-measurable. Assume that V is a $d \times d$ positive definite matrix. For any positive integer t, define

$$\overline{V}_t = V + \sum_{s=1}^{t} X_s X_s^\top \qquad\qquad S_t = \sum_{s=1}^{t} \eta_s X_s$$

Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any constant $t \in \mathbb{N}$,

$$\|S_t\|_{\overline{V}_t^{-1}}^2 \leq 2R^2 \log\left(\frac{\det(\overline{V}_t)^{1/2} \det(V)^{-1/2}}{\delta}\right).$$

In the above theorem, one can replace the constant t with a stopping time τ with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$.

Lemma 2.2. *Let $z \in \mathbb{R}^d$ be arbitrary and consider for any $t \geq 0$*

$$M_t^z = \exp\left(\sum_{s=1}^{t}\left(\frac{\eta_s z^\top X_s}{R} - \frac{1}{2}\left(z^\top X_s\right)^2\right)\right).$$

Let τ be a stopping time (A random variable such that $\{\tau = t\}$ is measurable by $\mathcal{F}_t$ for all $t \geq 0$.) with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. Then

$$\mathbb{E}\left[M_t^z\right] \leq 1 \qquad \textit{and} \qquad \mathbb{E}\left[M_{\min\{\tau,t\}}^z\right] \leq 1,$$

for any $t \in \mathbb{N}$.

Proof. Let

$$D_t^z := \exp\left(\frac{\eta_t z^\top X_t}{R} - \frac{1}{2}\left(z^\top X_t\right)^2\right).$$

Since η_t is R-sub-Gaussian, we have

$$\begin{aligned}
\mathbb{E}\left[D_t^z | \mathcal{F}_{t-1}\right] &\leq \mathbb{E}\left[\exp\left(\frac{\eta_t z^\top X_t}{R} - \frac{1}{2}\left(z^\top X_t\right)^2\right) | \mathcal{F}_{t-1}\right] \\
&\leq \exp\left(\frac{\left(z^\top X_t\right)^2}{R^2}\frac{R^2}{2}\right) \exp\left(-\frac{1}{2}\left(z^\top X_t\right)^2\right) \\
&\leq 1.
\end{aligned}$$

Thus we have

$$\mathbb{E}\left[M_t^z|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[M_{t-1}^z D_t^z|\mathcal{F}_{t-1}\right] = M_{t-1}^z \mathbb{E}\left[D_t^z|\mathcal{F}_{t-1}\right] \leq M_{t-1}^z,$$

which inductively proves that $\mathbb{E}\left[M_t^z\right] \leq 1$ for any $t \in \mathbb{N}$. Thus we have $\mathbb{E}\left[M_{\min\{\tau,t\}}^z\right] \leq 1$ for any $t \in \mathbb{N}$.

□

Proof of Theorem 2.1 (taken from Abbasi-Yadkori, Pál, Szepesvári, 2011). Without loss of generality, assume that $R = 1$. Let

$$V_t = \sum_{s=1}^{t} X_s X_s^\top \qquad\qquad M_t^z = \exp\left(z^\top S_t - \frac{1}{2}\|z\|_{V_t}^2\right)$$

By Lemma 2.2, the expectation of M_t^z is not larger than one. Let Λ be a Gaussian random variable which is independent of all the other random variables, whose covariance is V^{-1}. Define

$$M_t = \mathbb{E}\left[M_t^\Lambda|\mathcal{F}_\infty\right],$$

where $\mathcal{F}_\infty = \bigcup_{t=0}^\infty \mathcal{F}_t$. Clearly, we still have $\mathbb{E}\left[M_t\right] = \mathbb{E}\left[\mathbb{E}\left[M_t^\Lambda|\Lambda\right]\right] \leq 1$ for any constant t.

Let f denote the density of λ and for a positive definite matrix V, and let

$$Z(P) = \sqrt{(2\pi)^d/\det(P)} = \int \exp\left(-\frac{1}{2}x^\top P x\right) dx.$$

For M_t, we have

$$\begin{aligned}
M_t &= \int_{\mathbb{R}^d} \exp\left(z^\top S_t - \frac{1}{2}\|z\|_{V_t}^2\right) f(z)\, dz \\
&= \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\left\|z - V_t^{-1}S_t\right\|_{V_t}^2 + \frac{1}{2}\|S_t\|_{V_t}^2\right) f(z)\, dz \\
&= \frac{1}{Z(V)} \exp\left(\frac{1}{2}\|S_t\|_{V_t}^2\right) \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\left\|z - V_t^{-1}S_t\right\|_{V_t}^2 - \frac{1}{2}\|z\|_V^2\right) dz.
\end{aligned}$$

Note that if P is positive semi-definite and Q is positive definite, it holds that

$$\|x - z\|_P^2 + \|x\|_Q^2 = \|x - (P+Q)^{-1}Pz\|_{P+Q} + \|z\|_P^2 - \|Pz\|_{P+Q}^2.$$

Thus we have

$$\left\|z - V_t^{-1}S_t\right\|_{V_t}^2 + \|z\|_V^2 = \left\|z - \left(V_t + V\right)^{-1} S_t\right\|_{V+V_t}^2 + \|S_t\|_{V_t}^2 - \|S_t\|_{(V+V_t)^{-1}}^2,$$

which gives

$$
\begin{aligned}
M_t &= \frac{1}{Z(V)} \exp\left(\frac{1}{2}\|S_t\|^2_{V_t}\right) \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\left\|z - V_t^{-1}S_t\right\|^2_{V_t} - \frac{1}{2}\|z\|^2_V\right) dz \\
&= \frac{1}{Z(V)} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\left\|z - (V+V_t)^{-1}S_t\right\|^2_{V+V_t} + \frac{1}{2}\|S_t\|^2_{(V+V_t)^{-1}}\right) dz \\
&= \frac{Z(V+V_t)}{Z(V)} \exp\left(\frac{1}{2}\|S_t\|^2_{(V+V_t)^{-1}}\right) \\
&= \left(\frac{\det(V)}{\det(V+V_t)}\right)^{1/2} \exp\left(\frac{1}{2}\|S_t\|^2_{(V+V_t)^{-1}}\right)
\end{aligned}
$$

Since $\mathbb{E}\left[M_t\right] \leq 1$, we obtain

$$
\begin{aligned}
\mathbb{P}\left(\|S_t\|^2_{(V+V)^{-1}} \geq 2\log\frac{\det(V+V_t)^{1/2}}{\delta\det(V)^{1/2}}\right) &= \mathbb{P}\left(\left(\frac{\det(V)}{\det(V+V_t)}\right)^{1/2} \exp\left(\frac{1}{2}\|S_t\|^2_{(V+V)^{-1}}\right) > 1/\delta\right) \\
&\leq \delta\mathbb{E}\left[\left(\frac{\det(V)}{\det(V+V_t)}\right)^{1/2} \exp\left(\frac{1}{2}\|S_t\|^2_{(V+V)^{-1}}\right)\right] \\
&\qquad\qquad\qquad \text{(by Markov inequality.)} \\
&\leq \delta.
\end{aligned}
$$

□

Theorem 2.3. *Let $\{X_t\}_{t=1}^{\infty}$ be a sequence in $\mathbb{R}^d$, V a $d\times d$ positive definite matrix and define $V_t = V + \sum_{s=1}^{t} X_s X_s^\top$. Then it holds that*

$$
\log\frac{\det \overline{V}_n}{\det V} \leq \sum_{t=1}^{n} \|X_t\|^2_{\overline{V}^{-1}_{t-1}}.
$$

Further, if $\|X_t\|_2 \leq L$ for all t, then

$$
\begin{aligned}
\sum_{t=1}^{n} \min\left\{1, \|X_t\|^2_{\overline{V}^{-1}_{t-1}}\right\} &\leq 2\left(\log\det\overline{V}_2 - \log\det V\right) \\
&\leq 2\left(d\log\left(\left(\textit{trace}(V) + nL^2\right)/d\right) - \log\det V\right)
\end{aligned}
$$

In addition, if $\lambda_{\min}(V) \geq \max(1, L^2)$, then

$$
\sum_{t=1}^{n} \|X_t\|^2_{\overline{V}^{-1}_{t-1}} \leq 2\log\frac{\det\overline{V}_n}{\det V}.
$$

Proof. Since $\det(I + xx^\top) = 1 + \|x\|^2$, we have

$$\begin{aligned}
\det(\overline{V}_n) &= \det(\overline{V}_{n-1} + X_n X_n^\top) \\
&= \det\left(\overline{V}_{n-1}\right) \det\left(I + \overline{V}_{n-1}^{-1/2} X_n X_n^\top \overline{V}_{n-1}^{-1/2}\right) \\
&= \det\left(\overline{V}_{n-1}\right) \left(1 + \|X_n\|^2_{\overline{V}_{n-1}^{-1}}\right) \\
&= \det(V) \prod_{t=1}^{n} \left(1 + \|X_t\|^2_{\overline{V}_{t-1}^{-1}}\right).
\end{aligned}$$

Since $\log(1 + x) \le x$, we have

$$\log\det(\overline{V}_n) = \log\det(V) + \sum_{t=1}^{n} \log\left(1 + \|X_t\|^2_{\overline{V}_{t-1}^{-1}}\right) \le \log\det(V) + \sum_{t=1}^{n} \|X_t\|^2_{\overline{V}_{t-1}^{-1}}.$$

Since $x \le 2\log(1 + x)$ for all $x \in [0, 1]$, we have

$$\sum_{t=1}^{n} \min\left\{1, \|X_t\|^2_{\overline{V}_{t-1}^{-1}}\right\} \le \sum_{t=1}^{n} \log\left(1 + \|X_t\|^2_{\overline{V}_{t-1}^{-1}}\right) = 2\log\det(\overline{V}_n) - 2\log\det(V).$$

The trace of $\overline{V}_n$ is bounded by $trace(V) + nL^2$ if $\|X_t\|_2 \le L$ for all t. Hence,

$$\det(\overline{V}_n) \le \left(\frac{trace(\overline{V}_n)}{d}\right)^d \le \left(\frac{trace(V) + nL^2}{d}\right)^d.$$

Notice that $\|X_t\|^2_{\overline{V}_{t-1}^{-1}} \le \left(\lambda_{\min}(\overline{V}_{t-1})\right)^{-1} \|X_t\| \le \frac{L^2}{\lambda_{\min}(V)}$. Hence, if $\lambda_{\min}(V) \ge \max(1, L^2)$, we have $\|X_t\|^2_{\overline{V}_{t-1}^{-1}} \le 1$, and thus

$$\log\frac{\det\overline{V}_n}{\det V} \le \sum_{t=1}^{n} \|X_t\|^2_{\overline{V}_{t-1}^{-1}} \le 2\log\frac{\det\overline{V}_n}{\det V}.$$

$\square$

2.1 Connection to machine learning

Residual and Confidence in Regression

Consider a dataset $\{(x_i, y_i)\}_{i=1}^t$ governed by the linear model:

$$y = \theta^\top x + \eta,$$

where η is an independent sub-Gaussian noise.

Let $\widehat{\theta}_t$ be the L_2-regularized least-squares estimate of θ with regularization parameter $\lambda > 0$:

$$\widehat{\theta}_t = \left(X_t^\top X_t + \lambda I\right)^{-1} X_t^\top Y_t$$

where X_t is the matrix whose rows are $X_i^\top$ and $Y_t = (y_1, \cdots, y_t)^\top$.

Then we have the following theorem.

Theorem 2.4 (Confidence Ellipsoid). *Let $\overline{V}_t = \lambda I + \sum_{s=1}^{t} x_s x_s^\top$ ($\lambda > 0$). Define $y_t = x_t^\top \theta + \eta_t$, and assume that $\|\theta\|_2 \leq S$. Then, for any $\delta > 0$, with probability at least $1-\delta$, for all $t \geq 0$, θ lies in the set*

$$C_t := \left\{ \theta \in \mathbb{R}^d : \left\|\widehat{\theta}_t - \theta\right\|_{\overline{V}_t} \leq R\sqrt{2\log\left(\frac{\det(\overline{V}_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{1/2} S \right\}.$$

Proof. For simplicity, let $\eta = (\eta_1, \eta_2, \cdots, \eta_t)^\top$, let $X = X_t$ and $Y = Y_t$. Since

$$\begin{aligned}
\widehat{\theta}_t &= \left(X^\top X + \lambda I\right)^{-1} X^\top (X\theta + \eta) \\
&= \left(X^\top X + \lambda I\right)^{-1} X^\top \eta + \left(X^\top X + \lambda I\right)^{-1} \left(X^\top X + \lambda I\right)\theta - \lambda \left(X^\top X + \lambda I\right)^{-1} \theta \\
&= \left(X^\top X + \lambda I\right)^{-1} X^\top \eta - \lambda \left(X^\top X + \lambda I\right)^{-1} \theta + \theta,
\end{aligned}$$

we get, for any $x \in \mathbb{R}^d$,

$$\begin{aligned}
x^\top \widehat{\theta}_t - x^\top \theta &= x \left(X^\top X + \lambda I\right)^{-1} X^\top \eta - x\lambda \left(X^\top X + \lambda I\right)^{-1} \theta \\
&= \left\langle x, X^\top \eta \right\rangle_{\overline{V}_t^{-1}} - \lambda \left\langle x, \theta \right\rangle_{\overline{V}_t^{-1}}
\end{aligned}$$

where $\overline{V}_t = X^\top X + \lambda I$. By the Cauchy-Schwarz inequality, we get

$$\begin{aligned}
\left| x^\top \widehat{\theta}_t - x^\top \theta \right| &\leq \|x\|_{\overline{V}_t^{-1}} \left(\left\|X^\top \eta\right\|_{\overline{V}_t^{-1}} + \lambda \|\theta\|_{\overline{V}_t^{-1}} \right) \\
&\leq \|x\|_{\overline{V}_t^{-1}} \left(\left\|X^\top \eta\right\|_{\overline{V}_t^{-1}} + \sqrt{\lambda} \|\theta\|_2 \right) \\
&\qquad \left(\text{since } \|\theta\|_{\overline{V}_t^{-1}} = \sqrt{\theta^\top \left(\lambda I + \textstyle\sum_{x=1}^{t} x_s^\top x_s\right)^{-1} \theta}.\right)
\end{aligned}$$

By Theorem 2.1, for any $\delta > 0$, with probability at least 1δ, $\forall t \geq 0$,

$$\left\|X^\top \eta\right\|_{\overline{V}_t^{-1}} = R\sqrt{2\log\left(\frac{\det(\overline{V}_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)}$$

Setting $x = \overline{V}_t(\widehat{\theta}_t - \theta)$, and using $\|\theta\|_2 \leq S$, we get

$$\left\|\widehat{\theta}_t - \theta\right\|_{\overline{V}_t}^2 \leq \left\|\overline{V}_t\left(\widehat{\theta}_t - \theta\right)\right\|_{\overline{V}_t^{-1}} \left(R\sqrt{2\log\left(\frac{\det(\overline{V}_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \sqrt{\lambda} S \right),$$

which concludes the proof since $\left\|\widehat{\theta}_t - \theta\right\|_{\overline{V}_t} = \left\|\overline{V}_t\left(\widehat{\theta}_t - \theta\right)\right\|_{\overline{V}_t^{-1}}$.

$\square$

3 Linear Bandit

Now consider the following decision making process.

- For $t = 1, 2, \cdots, T$,
  - choose x_t from D, where D a compact set in $\mathbb{R}^d$;
  - observe a y_t, where $y_t = \theta^\top x_t + \eta_t$, $\theta \in \mathbb{R}^d$ is unknown and η_t is an independent R-sub-Gaussian noise.

Let x^* be the optimal choice in S. Performance measure: $Reg(T) = \sum_{t=1}^T \theta^\top x^* - \theta^\top x_t$. Algorithm for solving this problem:

- For $t = 1, 2, \cdots, T$,
  - Solve $(x_t, \widetilde{\theta}_t) \in \arg\max_{(x,\theta)\in D\times C_{t-1}} \theta^\top x$, where C_t is defined in Theorem 2.4, and C_0 is the unit ball by convention.
  - Play x_t, and observe y_t.

The above algorithm is called Optimism in Face of Uncertainty, which is essentially an Upper Confidence Bound algorithm.

Theorem 3.1. *With probability* $1 - \delta T$*, the regret for the above algorithm satisfies*

$$Reg(T) \leq \mathcal{O}\left(\sqrt{Td}\log\left(T/\delta\right)\right).$$

Choosing $\delta = \frac{1}{T^2}$ *gives a high probability bound.*

Proof. Let x_* be the optimal x. Decompose the regret at time t as follows:

$$\begin{aligned}
r_t &= \langle \theta, x_* \rangle - \langle \theta, x_t \rangle \\
&\leq \left\langle \widetilde{\theta}_t, x_t \right\rangle - \langle \theta, x_t \rangle && \text{(by Algoirthm design)} \\
&= \left\langle \widetilde{\theta}_t - \theta_*, x_t \right\rangle \\
&= \left\langle \widehat{\theta}_t - \theta_*, x_t \right\rangle + \left\langle \widetilde{\theta}_t - \widehat{\theta}_t, x_t \right\rangle \\
&\leq 2rad(C_{t-1})\|x_t\|_{\overline{V}_{t-1}^{-1}},
\end{aligned}$$

where $rad(C_t) := R\sqrt{2\log\left(\frac{\det(\overline{V}_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{1/2}S$. By Theorem 2.4, we can bound

$rad(C_{t-1})$. By Theorem 2.3 and boundedness of x_t and θ, we can bound the regret by

$$
\begin{aligned}
Reg(T) &= \sum_{t=1}^{T} r_t \\
&\leq \sqrt{T \sum_{t=1}^{T} r_t^2} && \text{(by Cauchy-Schwarz)} \\
&\leq \sqrt{T \sum_{t=1}^{T} \left(rad(C_{t-1})\right)^2 \|x_t\|_{\overline{V}_{t-1}^{-1}}} \\
&\leq \sqrt{T \sum_{t=1}^{T} \left(R\sqrt{2\log\left(\frac{\det(\overline{V}_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{1/2} S \right)^2 \|x_t\|^2_{\overline{V}_{t-1}^{-1}}} \\
&\leq \sqrt{T \left(R\sqrt{2\log\left(\frac{\det(\overline{V}_T)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{1/2} S \right)^2 \sum_{t=1}^{T} \|x_t\|^2_{\overline{V}_{t-1}^{-1}}} \\
&\leq \mathcal{O}\left(\sqrt{Td \log\left(T/\delta\right) \log \frac{\det \overline{V}_T}{\det(\lambda I)}} \right) \\
&\leq \mathcal{O}\left(\sqrt{Td} \log\left(T/\delta\right) \right)
\end{aligned}
$$

□

Acknowledgement

Mainly due to MJ Klass, TL Lai, VH Pena, QM Shao, and others. A reference in the machine learning is the "Improved Algorithms for Linear Stochastic Bandits" paper by Abbasi-Yadkori, Pál, Szepesvári.