Subgauss r.v. & Rates of learning

(1) CGFs:

Recall that we've not had PAC-learnability yet because the convergence statement isn't enough: the rate is slow

Note for hac. We consider $\{ I_{(-n,n)} =: f_n \}_{n \in \mathbb{R}'}$ which's uniform bdd by 1. And then we can do some better.

Def: X. $\mathbb{R}'$-r.v. $\mathcal{I}_X(\alpha) := \mathbb{E}(e^{\alpha X}) : \mathbb{R}' \to \overline{\mathbb{R}}^{>0}$

And cumulant generating func. (CGF)

$\Psi_X(\alpha) = \log(\mathcal{I}_X(\alpha))$.

Rmk: They can tell us sth. about exp. rate of tails of r.v. X.

Lem. If $\mathcal{I}_X(\alpha)$ is finite on interval I. Then $\Psi_X(\alpha)$ is convex on I.

Pf: Set $\mathbb{P}_\alpha(A) := \int I_A e^{\alpha X} / \mathbb{E}(e^{\alpha X}) \, d\mathbb{P}$.

$\mathbb{P}_\alpha$ is p.m. And then we have:

$$k^2\psi_X(\alpha)/k\alpha^2 = \mathbb{E}_{\mathbb{P}_\alpha}(X^2) - \mathbb{E}_{\mathbb{P}_\alpha}(X)^2$$
$$= \mathrm{Var}_{\mathbb{P}_\alpha}(X) \geq 0.$$

(2) Bound:

Def: $\psi : \mathbb{R}_+ \to \mathbb{R} \cup \{+\infty\}$. s.t. $\psi(\alpha)$ is finite at least one $\alpha \in \mathbb{R}^+$. Its Legendre transfer is $\psi^* : \mathbb{R}_+ \to \mathbb{R} \cup \{+\infty\}$ defined by $\psi^*(\varepsilon) = \sup\limits_{\alpha \in \mathbb{R}^+} \{\alpha\varepsilon - \psi(\alpha)\}$.

Rmk. ψ convex $\Rightarrow$ ψ^* convex.

Lem. (Chernov bdd)

X is $\mathbb{R}^1$-r.v. st. $\mathcal{L}_X(\alpha)$ is finite on $[0, \alpha^*)$ and $\psi_X(\alpha)$ is its CGF. Then we have:

$\mathbb{P}(X \geq \varepsilon) \leq \exp(-\psi_X^*(\varepsilon))$

Pf: Apply Chebyshev inequi.

e.g. For $X \sim N(0, \sigma^2)$. We have $\psi_X(\alpha) = \frac{1}{2}\sigma^2\alpha^2$. So: $\psi_X^*(\varepsilon) = \varepsilon^2/2\sigma^2$ $\Rightarrow$

$\mathbb{P}(X \geq \varepsilon) \leq \exp(-\varepsilon^2/2\sigma^2)$

Def: A centered $\mathbb{R}^1$-valued r.v. X with CGF $\psi_X(\tau)$ is called subgaussian with var. $\sigma^2 > 0$ if $\psi_X(\alpha) \leq \frac{1}{2}\sigma^2\alpha^2$. $\forall \tau > 0$.

Rmk: i) Chernov bdd for $N(0,\sigma^2)$ above still works for subgaussian r.v.

ii) Its tail at worst as thick as normal distribution.

Thm. For $\{X_k\} \overset{i.i.d.}{\sim}$ subgaussian r.v. with var. σ^2. Then: $\overline{X}_n = \frac{1}{n}\sum_1^n X_k$ is subgaussian with var. σ^2/n. So $\mathbb{P}(\overline{X}_n \geq \varepsilon) \leq e^{-\frac{n}{2}\frac{\varepsilon^2}{\sigma^2}}$

Rmk: If $-X_k$ is also subgaussian with σ^2

$\Rightarrow \mathbb{P}(|\overline{X}_n| \geq \varepsilon) \leq 2\exp(-\frac{n}{2}\frac{\varepsilon^2}{\sigma^2})$.

Pf: $\psi_{\overline{X}_n}(\alpha) = n\psi_X(\frac{\alpha}{n}) \leq n \cdot \frac{\sigma^2}{2}(\frac{\alpha}{n})^2$

Lem. (Hoeffding's inequi.)

$X: \Omega \to [a,b]$ is centered r.v. Then X is subgaussian with var. $(b-a)^2/4$.

Cor. $\mathbb{P}(|\overline{X}_n| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2/(b-a)^2}$ for $X_k \overset{i.i.d.}{\sim} X$.

Pf: Note $-X_k : \Omega \to [-a, -b]$. satisfies appl.

Pf: By Taylor: $\psi_X(\tau) = \psi_X(0) + \psi_X'(0)\tau + \frac{1}{2}\psi_X''(\tau^*)\tau^2$

$$\overset{E(X)=0}{=} \frac{1}{2}\psi_X''(\tau^*)\tau^2.$$

$\psi_X''(\tau^*) = \mathrm{Var}_{\mathbb{P}_{\tau^*}}(X)$ by Lem. in (1).

Lem. (Popoviciu's inequi. on Var.)

For $X : \Omega \to [a, b]$. r.v. Then:

$\mathrm{Var}(X) \leq (b-a)^2/4$.

Pf: $\mathbb{E}((b-X)(X-a))$

$= \mu(a+b) - ab - \mathbb{E}(X^2) \geq 0$. $(*)$

So: $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mu^2$

$\overset{(*)}{\leq} \mu(a+b) - ab - \mu^2$

$= (b-\mu)(\mu-a) \leq \left(\frac{b-a}{2}\right)^2$

$\Rightarrow \psi_X''(\tau^*) = \mathrm{Var}_{\mathbb{P}_{\tau^*}}(X) \leq (b-a)^2/4$.

Next, we can derive the new rate of PAC learning from empirical measures:

Lem. In i.i.d model. $\mathcal{J} = M_1^+(\mathbb{R}^1)$ is PAC-learnable by empirical dist. $\hat{\mu}_n = \frac{1}{n}\sum_1^n \delta_{x_i}$ w.r.t. d_{KS}. And $n(\varepsilon, \delta) = -\frac{8}{\varepsilon^2}\log\left(\frac{\varepsilon\delta}{8}\right)$.

Pf: Set $\Gamma_n^{(N)} := \max_{j \le N-1} |\cdots| \vee |\cdots|$ as before

$\hat{F}_n(x) = \frac{1}{n}\sum_1^n I(x_j \le x)$. Apply Hoeffding

inequi. on i.i.d $I\{x_j \le x\}$ and it also

works for $\hat{F}_n(x-) = \frac{1}{n}\sum_1^n I(x_j < x)$. So:

$\mathbb{P}(\Gamma_n^{(N)} \ge \varepsilon) = \mathbb{P}((\bigcup_1^{N-1} |\cdots| \ge \varepsilon) \cup (\bigcup_0^{N-1} |\cdots| \ge \varepsilon))$

$\le 2 \cdot 2 \cdot (N-1) \exp(-\frac{1}{2}\varepsilon^2 n)$

With $d_{KL}(\mu, \hat{\mu}_n) \le \Gamma_n^{(N)} + \frac{1}{N}$ which

we obtained before.

So $\mathbb{P}(d_{KL}(\mu, \hat{\mu}_n) > \varepsilon) \le 4(N-1) e^{-\frac{n}{2}(\varepsilon - \frac{1}{N})^2}$

let $N = \lfloor \frac{2}{\varepsilon} \rfloor$, we have:

$\mathbb{P}(d_{KL}(\mu, \hat{\mu}_n) > \varepsilon) \le \frac{8}{\varepsilon} e^{-\frac{1}{8}\varepsilon^2 n}$

Let RHS $= \delta \Rightarrow n(\varepsilon, \delta) = -\frac{8}{\varepsilon^2} \log \frac{\varepsilon\delta}{8}$.

Rmk: It works also for $\mathbb{R}^k$.