0.0.1 Lemma 0

Consider any empirical distributions T and D.

Consider the function class

$$\hat{\mathcal{G}}(T,\epsilon_T) = \left\{ \hat{g}_\lambda(\cdot|T,\epsilon_T) = \arg\min_{g\in\mathcal{G}} \frac{1}{2}\|y-g\|_T^2 + \lambda\left(P^v(g) + \frac{w}{2}\|g\|_D^2\right) : \lambda \in \Lambda \right\}$$

Suppose the penalty function P is a semi-norm, smooth, and convex. Suppose for all h, $\|h\|_D \leq O_p(n^p)P(h)$.

Suppose $v \geq 1$.

Suppose $\lambda_{min} = O_P(n^{-\tau_{min}})$ and $\lambda_{max} = O_P(n^{\tau_{max}})$.

Then the entropy bound is

$$H\left(d, \hat{\mathcal{G}}(T,\epsilon_T), \|\cdot\|_D\right) \leq 2\log\left(\frac{1}{d}\right) + \kappa \log n + \log\left[2v\left(\|\epsilon\|_T^2 + P^v(g^*) + \frac{w}{2}\|g^*\|_D^2 + G\right)/(Cw)\right]$$

where κ, C only depend on $\tau_{min}, \tau_{max}, v, u, p$.

(Notation: κ, C, c are constants that only depend on $\tau_{min}, \tau_{max}, u, p, v$.)

Proof

Let

$$\delta(d) = n^{-c} d^2 w v^{-1}\left(\|\epsilon\|_T^2 + P^v(g^*) + \frac{w}{2}\|g^*\|_D^2 + G\right)^{-1} C/2$$

(c, C are defined below).

We will show that the following set $\Omega_{\delta(d)}$ forms a d-cover set for $\hat{\mathcal{G}}(T,\epsilon_T)$:

$$\Omega_{\delta(d)} = \left\{ \hat{g}_{\delta_i}(\cdot|T) : \delta_i = i\delta(d) + \lambda_{min} \text{ for } i = 0, ..., \left\lceil \frac{\lambda_{max} - \lambda_{min}}{\delta(d)} \right\rceil \right\}$$

Consider any $\lambda \in [\lambda_{min}, \lambda_{max}]$ and suppose $\delta_i < \lambda < \delta_{i+1}$. Let $h = \hat{g}_{\delta_i}(\cdot|T) - \hat{g}_\lambda(\cdot|T)$. Suppose $\|h\|_D > d$ for contradiction.

Consider the one-dimensional problem with any λ_0

$$\hat{m}_h(\lambda_0) = \arg\min_m \frac{1}{2}\|y - (\hat{g}_{\delta_i} + mh)\|_T^2 + \lambda_0\left(P^v(\hat{g}_{\delta_i} + mh) + \frac{w}{2}\|\hat{g}_{\delta_i} + mh\|_D^2\right)$$

Clearly $\hat{m}_h(\delta_i) = 0$ and $\hat{m}_h(\lambda) = 1$. Also, by the mean-value theorem, there is some $\alpha \in (\delta_i, \lambda)$ s.t

$$\hat{m}_h(\lambda) = (\lambda - \delta_i)\left|\frac{\partial}{\partial\lambda_0}\hat{m}_h(\lambda_0)\right|_{\lambda_0=\alpha} \leq \delta\left|\frac{\partial}{\partial\lambda_0}\hat{m}_h(\lambda_0)\right|_{\lambda_0=\alpha}$$

To get $\frac{\partial}{\partial\lambda_0}\hat{m}_h(\lambda_0)$, we take lots of derivatives.

Taking the derivative of the criterion wrt m, we get

$$-\langle h, y - (\hat{g}_{\delta_i} + mh)\rangle_T + \lambda_0\left(\frac{\partial}{\partial m}P^v(\hat{g}_{\delta_i} + mh) + w\langle h, \hat{g}_{\delta_i} + mh\rangle_D\right)\Bigg|_{m=\hat{m}_h(\lambda_0)} = 0$$

By implicit differentiation wrt λ_0, we have

$$\frac{\partial}{\partial\lambda_0}\hat{m}_h(\lambda_0) \quad = \quad -\left(\|h\|_T^2 + \lambda_0\frac{\partial^2}{\partial m^2}P^v\left(\hat{g}_{\delta_i} + mh\right) + \lambda_0 w\|h\|_D^2\right)^{-1}\left(\frac{\partial}{\partial m}P^v(\hat{g}_{\delta_i} + mh) + w\langle h, \hat{g}_{\delta_i} + mh\rangle_D\right)\Bigg|_{m=\hat{m}_\lambda(\lambda_0)}$$

To bound $\left|\frac{\partial}{\partial\lambda_0}\hat{m}_h(\lambda_0)\right|$, we bound each multiplicand.

1st multiplicand: Since penalty P is convex (regardless of the direction of h),

$$\begin{aligned}\left|\|h\|_T^2 + \lambda_0 \frac{\partial^2}{\partial m^2} P^v\left(\hat{g}_{\delta_i} + mh\right) + \lambda_0 w \|h\|_D^2\right|^{-1} &\leq \lambda_0^{-1} w^{-1} \|h\|_D^{-2} \\ &\leq n^{\tau_{min}} w^{-1} d^{-2}\end{aligned}$$

2nd multiplicand:
We first bound

$$\left|\frac{\partial}{\partial m} P^v(\hat{g}_{\delta_i} + mh)\right| = \left|v P^{v-1}(\hat{g}_{\delta_i} + mh) \frac{\partial}{\partial m} P(\hat{g}_{\delta_i} + mh)\right|$$

By definition of $\hat{g}_{\delta_i} + \hat{m}_h(\lambda_0)h$ and $\hat{g}_{\delta_i}$,

$$\begin{aligned}\lambda_0 P^v(\hat{g}_{\delta_i} + \hat{m}_h(\lambda_0)h) &\leq \frac{1}{2}\|y - \hat{g}_{\delta_i}\|_T^2 + \lambda_0 \left(P^v(\hat{g}_{\delta_i}) + \frac{w}{2}\|\hat{g}_{\delta_i}\|_D^2\right) \\ &\leq \frac{1}{2}\|y - g^*\|_T^2 + \delta_i \left(P^v(g^*) + \frac{w}{2}\|g^*\|_D^2\right) + (\lambda_0 - \delta_i)\left(P^v(\hat{g}_{\delta_i}) + \frac{w}{2}\|\hat{g}_{\delta_i}\|_D^2\right)\end{aligned}$$

We know that

$$P^v(\hat{g}_{\delta_i}) + \frac{w}{2}\|\hat{g}_{\delta_i}\|_D^2 \leq \frac{1}{2\delta_i}\|y - g^*\|_T^2 + P^v(g^*) + \frac{w}{2}\|g^*\|_D^2$$

Hence

$$\begin{aligned}P^{v-1}(\hat{g}_{\delta_i} + \hat{m}_h(\lambda_0)h) &\leq \left(\frac{1}{2\delta_i}\|\epsilon\|_T^2 + P^v(g^*) + \frac{w}{2}\|g^*\|_D^2\right)^{(v-1)/v} \\ &\leq \left(\frac{n^{\tau_{min}}}{2}\|\epsilon\|_T^2 + P^v(g^*) + \frac{w}{2}\|g^*\|_D^2\right)^{(v-1)/v}\end{aligned}$$

Note that since P is a semi-norm, then

$$|P(\hat{g}_{\delta_i} + mh) - P(\hat{g}_{\delta_i})| \leq |m| P(h)$$

Therefore as we take $m \to 0$, we have

$$\left|\frac{\partial}{\partial m} P(\hat{g}_{\delta_i} + mh)\right| \leq P(h)$$

Since P is a semi-norm,

$$P(h) = P(\hat{g}_{\delta_i} - \hat{g}_{\lambda_0}) \leq P(\hat{g}_{\delta_i}) + P(\hat{g}_{\lambda_0})$$

We bound the penalties $P(\hat{g}_{\delta_i})$ and $P(\hat{g}_{\lambda_0})$ by the same logic as above. Hence we know that

$$P(h) \leq 2\left(\frac{n^{\tau_{min}}}{2}\|\epsilon\|_T^2 + P^v(g^*) + \frac{w}{2}\|g^*\|_D^2\right)^{1/v}$$

Now we bound $|w\langle h, \hat{g}_{\delta_i} + mh\rangle_D|$.
By Cauchy Schwarz and the assumption that $\sup_{g \in \mathcal{G}} \|g\| \leq G$, we have

$$\begin{aligned}|w\langle h, \hat{g}_{\delta_i} + mh\rangle| &\leq w\|h\|\|\hat{g}_{\delta_i} + mh\| \\ &\leq w n^p P(h) G\end{aligned}$$

Combining the above bounds, we have

$$
\begin{aligned}
& \left|\frac{\partial}{\partial \lambda_0}\hat{m}_h(\lambda_0)\right| \\
\leq\ & n^{\tau_{min}} w^{-1} d^{-2} \left(2v\left(\frac{n^{\tau_{min}}}{2}\|\epsilon\|_T^2 + P^v(g^*) + \frac{w}{2}\|g^*\|_D^2\right)^{(v-1)/v} + wn^p G\right)\left(\frac{n^{\tau_{min}}}{2}\|\epsilon\|_T^2 + P^v(g^*) + \frac{w}{2}\|g^*\|_D^2\right)^{1/v} \\
\leq\ & C d^{-2} n^c w^{-1} v \left(\|\epsilon\|_T^2 + P^v(g^*) + \frac{w}{2}\|g^*\|_D^2 + G\right)
\end{aligned}
$$

Hence by the MVT, we have found that

$$\hat{m}_h(\lambda) \leq 1/2$$

which is a contradiction.

Therefore $\Omega_{\delta(d)}$ forms a d-cover set. The d-covering number is

$$
\begin{aligned}
N\left(d, \hat{\mathcal{G}}(T, \epsilon_T), \|\cdot\|_D\right) &\leq \left\lceil \frac{\lambda_{max} - \lambda_{min}}{\delta(d)} \right\rceil \\
&= 2n^{\kappa} v \left(\frac{\|\epsilon\|_T^2 + P^v(g^*) + \frac{w}{2}\|g^*\|_D^2 + G}{wCd^2}\right)
\end{aligned}
$$

and the entropy is

$$H\left(d, \hat{\mathcal{G}}(T, \epsilon_T), \|\cdot\|_D\right) \leq 2\log\left(\frac{1}{d}\right) + \kappa \log n + \log\left[2v\left(\|\epsilon\|_T^2 + P^v(g^*) + \frac{w}{2}\|g^*\|_D^2 + G\right)/(Cw)\right]$$

Note that this also bounds the entropy for any metric norm calculated using a subset $D_0 \subseteq D$. Since

$$\|f\|_D \geq \sqrt{\frac{n_{D_0}}{n}}\|f\|_{D_0}$$

we have

$$H\left(d, \hat{\mathcal{G}}(T, \epsilon_T), \|\cdot\|_{D_0}\right) \leq H\left(\sqrt{\frac{n_{D_0}}{n}} d, \hat{\mathcal{G}}(T, \epsilon_T), \|\cdot\|_{D_0}\right)$$