Consider the joint optimization problem on a training/validation split to find the best regularization parameter λ in Λ:

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \frac{1}{2}\|y - \hat{g}_\lambda\|_V^2$$

$$\hat{g}(\lambda) = \arg\min_{g \in \mathcal{G}} \frac{1}{2}\|y - g\|_T^2 + \lambda\left(P(g) + \frac{w}{2}\|g\|_T^2\right)$$

Let the range of Λ be from $[\lambda_{min}, \lambda_{max}]$. Both limits can grow and shrink at any polynomial rate, e.g. $\lambda_{max} = O_P(n^{\tau_{max}})$ and $\lambda_{min} = O_P(n^{-\tau_{min}})$.

Assumptions

- Suppose that there is some constants K, k s.t. $\|g\|_V \le K n^k P(g)$.
- Suppose that $P(\cdot)$ is a semi/psuedo-norm (satisfies the triangle inequality) and a continuous function.
- Suppose that $\mathcal{G}$ is a cone (potentially bounded). Let g be some fixed function in $\mathcal{G}$. Suppose for any fixed function g, we can write $\mathcal{G} = \{g + mh : m \in \mathbb{R}, h \in \mathcal{G}, P(h) = 1\}$.
- Suppose that $\frac{\partial}{\partial m}P(g+mh)$ and $\frac{\partial^2}{\partial m^2}P(g+mh)$ exist everywhere. Suppose that $\frac{\partial^2}{\partial m^2}P(g+mh) \ge 0$ (some functional version of convex).

Proof

Step 1: Find the entropy of the model class $\mathcal{G}_\lambda$

We show that the entropy $H(u, \mathcal{G}_\lambda, \|\cdot\|_V)$ of the class

$$\mathcal{G}_\lambda = \left\{ g_{\hat{\theta}(\lambda)} : \lambda \in \Lambda \right\}$$

is bounded at a near-parametric rate:

$$H\left(u, \mathcal{G}_\lambda, \|\cdot\|_V\right) \le \log\left(\frac{(1 + w^{1/2}KR)M}{uwK}\right) + (\tau_{max} + \tau_{min})\log n$$

We are interested in bounding

$$\|\hat{g}_\lambda - \hat{g}_{\lambda+\delta}\|_V$$

For a fixed g and h, consider the problem

$$m_\lambda = \arg\min_m \frac{1}{2}\|y - (g + mh)\|_T^2 + \lambda\left(P(g+mh) + \frac{w}{2}\|g + mh\|_T^2\right)$$

Taking the derivative wrt m, we have

$$-\langle h, y - (g + m_\lambda h)\rangle_T + \lambda\left(\frac{\partial}{\partial m}P(g + m_\lambda h) + w\langle h, g + m_\lambda h\rangle_T\right) = 0$$

Now take the implicit derivative wrt λ.

$$\frac{\partial m_\lambda}{\partial \lambda}\|h\|_T^2 + \frac{\partial}{\partial m}P(g + m_\lambda h) + w\langle h, g + m_\lambda h\rangle_T + \lambda\left(\frac{\partial^2}{\partial m^2}P(g + m_\lambda h) + w\|h\|_T^2\right)\frac{\partial m_\lambda}{\partial \lambda} = 0$$

Rearranging, we get

$$\frac{\partial m_\lambda}{\partial \lambda} = -\left(\|h\|_T^2 + \lambda\frac{\partial^2}{\partial m^2}P(g + m_\lambda h) + \lambda w\|h\|_T^2\right)^{-1}\left(\frac{\partial}{\partial m}P(g + m_\lambda h) + w\langle h, g + m_\lambda h\rangle_T\right)$$

Hence

$$\left\|\frac{\partial m_\lambda}{\partial \lambda}\right\| \leq K^{-2}n^{-2k}(\lambda w)^{-1}\left(\left|\frac{\partial}{\partial m}P(g+m_\lambda h)\right| + wKn^k\|g+m_\lambda h\|_T\right)$$

Let's bound the two terms on the RHS.
To bound the derivative, note that by the triangle inequality

$$|P(g+mh)-P(g)| \leq mP(h)$$

As $m \to 0$, assuming the derivative exists, we have

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

To bound $\|g+m_\lambda h\|_T$, note that by definition,

$$\frac{\lambda w}{2}\|g+m_\lambda h\|_T^2 \leq \frac{1}{2}\|y-g^*\|_T^2 + \lambda P(g^*) + \frac{\lambda w}{2}\|g^*\|^2$$

so with high probability

$$\|g+m_\lambda h\|_T \leq \sqrt{(\lambda w)^{-1}4\sigma^2 + w^{-1}P(g^*) + \|g^*\|^2}$$

Hence

$$\left\|\frac{\partial m_\lambda}{\partial \lambda}\right\| \leq K^{-2}n^{-2k}(\lambda w)^{-1}\left(1 + wKn^k\sqrt{(\lambda w)^{-1}4\sigma^2 + w^{-1}P(g^*) + \|g^*\|^2}\right)$$

Let's now bound $\|\hat{g}_\lambda - \hat{g}_{\lambda+\delta}\|_V$. For some h s.t. $P(h)=1$, we can write

$$\hat{g}_{\lambda+\delta} = \hat{g}_\lambda + m_{\lambda+\delta}h$$

By the mean value theorem, there is some $\alpha \in [0,1]$ and constant R that only depends on g^*, σ s.t.

$$\begin{aligned}
\|\hat{g}_\lambda - \hat{g}_{\lambda+\delta}\|_V &= m_{\lambda+\delta}\|h\|_V \\
&\leq O_p(K^{-2}n^{-2k})m_{\lambda+\delta} \\
&\leq O_p(K^{-2}n^{-2k})\delta\left|\left.\frac{\partial m_\lambda}{\partial \lambda}\right|_{\lambda=\lambda+\alpha\delta}\right| \\
&\leq O_p(K^{-2}n^{-2k})\delta(\lambda w)^{-1}\left(1 + wKn^k\sqrt{(\lambda w)^{-1}4\sigma^2 + w^{-1}P(g^*) + \|g^*\|^2}\right) \\
&\leq O_p(K^{-1}n^{-k})\delta\lambda_{min}^{-2}w^{-1/2}R \\
&\leq \delta O_p(n^{2\tau_{min}-k})K^{-1}w^{-1/2}R
\end{aligned}$$

Then the covering number is, for constants κ that depend linearly on $\tau_{min}, \tau_{max}, k$,

$$N\left(u, \mathcal{G}_\lambda, \|\cdot\|_V\right) \leq \frac{R}{uK\sqrt{w}}O_p(n^\kappa)$$

so the entropy is

$$H\left(u, \mathcal{G}_\lambda, \|\cdot\|_V\right) \leq \log\left(\frac{R}{uK\sqrt{w}}\right) + \kappa \log n$$

Step 2,3,4 should all carry through nicely