Definitions

We find the best model for y over function class $\mathcal{G}$. Presume $g^* \in \mathcal{G}$ is the true model and

$$y = g^*(X) + \epsilon$$

Given a training set T , We define the fitted models

$$\hat{g}_\lambda = \|y - g\|_T^2 + \lambda^2 I^v(g)$$

Given a validation set T , let the CV-fitted model be

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

Assumptions

Suppose we have sub-Gaussian errors ϵ for constants K and σ_0^2:

$$\max_{i=1:n} K^2 \left(E \left[\exp(|\epsilon_i|^2 K^2) - 1 \right] \right) \leq \sigma_0^2$$

Suppose $v > 2\alpha/(2+\alpha)$.

Suppose that the entropy of the class $\mathcal{G}'$ is

$$H \left(\delta, \mathcal{G}' = \left\{ \frac{g - g^*}{I(g) + I(g^*)} : g \in \mathcal{G}, I(g) + I(g^*) > 0 \right\}, P_n \right) \quad \leq \quad \tilde{A}\delta^{-\alpha}$$

Suppose for all $\lambda \in \Lambda$, $I^v(\hat{g}_\lambda)$ is upper bounded by $\|\hat{g}_\lambda\|_n^2 = \frac{1}{n}\sum_{i=1}^n \hat{g}_\lambda(x_i)$. See Lemma 1 below for the specific assumption. This assumption includes Ridge, Lasso, Generalized Lasso, Group Lasso, and the Sobelov norm. In the example below, I show that the Sobelov norm also satisfies these conditions.

Result 1:

For now, we will suppose $P_n = \{X_i\}_{i=1}^n$ are the same between the validation and training set.

Also, suppose the penalty normalizes the empirical norm such that:

$$\sup_{g \in \mathcal{G}} \frac{\|g - g^*\|_n}{I(g) + I(g^*)} \leq R < \infty$$

Suppose for all $\lambda \in \Lambda$, $I^v(\hat{g}_\lambda)$ is upper bounded by its L_2-norm with some constant M and M_0 such that

$$I^v(\hat{g}_\lambda) \leq M\|\hat{g}_\lambda\|_n^2 + M_0$$

Then

$$\|\hat{g}_{\hat{\lambda}} - g^*\|_n = O_p(n^{-1/(2+\alpha)}) \left(M^{\alpha/v(2+\alpha)} \|g^*\|_n^{\alpha/2v(2+\alpha)} \vee I^{2\alpha/(2+\alpha)}(g^*) \right)$$

Proof

Let $\tilde{\lambda}$ be the optimal λ under the given assumptions, as specified by Van de geer. From the definition of $\hat{\lambda}$, we get the following basic inequality

$$\begin{aligned}
\|g^* - \hat{g}_{\hat{\lambda}}\|_V^2 \quad &\leq \quad \|g^* - \hat{g}_{\tilde{\lambda}}\|_V^2 + 2(\epsilon, \hat{g}_{\hat{\lambda}} - \hat{g}_{\tilde{\lambda}})_V \\
&\leq \quad \|g^* - \hat{g}_{\tilde{\lambda}}\|_V^2 + 2(\epsilon, \hat{g}_{\hat{\lambda}} - g^*)_V + 2(\epsilon, g^* - \hat{g}_{\tilde{\lambda}})_V \\
&\leq \quad \|g^* - \hat{g}_{\tilde{\lambda}}\|_V^2 + 2\left|(\epsilon, \hat{g}_{\hat{\lambda}} - g^*)_V\right| + 2\left|(\epsilon, g^* - \hat{g}_{\tilde{\lambda}})_V\right|
\end{aligned}$$

By considering the largest term on the RHS, we have following three cases.

Case 1: $\|g^* - \hat{g}_{\tilde{\lambda}}\|_V^2$ is the largest

Since we have assumed that the validation and training set are equal, then $\|g^* - \hat{g}_{\tilde{\lambda}}\|_V$ converges at the optimal rate $O_p(n^{-1/(2+\alpha)})$.

Case 2: $\left|(\epsilon, g^* - \hat{g}_{\tilde{\lambda}})_V\right|$ is the largest

In this case, since ϵ_V is independent of $\hat{g}_{\tilde{\lambda}}$, then by Cauchy Schwarz,

$$\begin{aligned} \left|(\epsilon, g^* - \hat{g}_{\tilde{\lambda}})_V\right| &\leq \|\epsilon_V\|\|g^* - \hat{g}_{\tilde{\lambda}}\|_V \\ &\leq O_p\left(n^{-1/2}\right)\|g^* - \hat{g}_{\tilde{\lambda}}\|_V \end{aligned}$$

Hence $\left|(\epsilon, g^* - \hat{g}_{\tilde{\lambda}})_V\right|$ will shrink a bit faster than the optimal rate at a rate of $O_p(n^{-(\frac{1}{2+\alpha}+\frac{1}{2})})$.

Case 3: $\left|(\epsilon, g^* - \hat{g}_{\hat{\lambda}})_V\right|$ is the largest.

By the assumptions given, Vandegeer (10.6) gives us that

$$\sup_{g\in\mathcal{G}} \frac{\left|(\epsilon, g - g*)_n\right|}{\|g - g*\|_n^{1-\alpha/2}(I(g^*) + I(g))^{\alpha/2}} = O_p(n^{-1/2})$$

Hence

$$\left|(\epsilon, g^* - \hat{g}_{\hat{\lambda}})_V\right| \leq O_p(n^{-1/2})\|\hat{g}_{\hat{\lambda}} - g*\|_n^{1-\alpha/2}(I(g^*) + I(\hat{g}_{\hat{\lambda}}))^{\alpha/2}$$

If $I(g^*) \geq I(g_{\hat{\lambda}})$, then

$$\|g^* - \hat{g}_{\hat{\lambda}}\|_V \leq O_p(n^{-1/(2+\alpha)})I(g^*)^{\alpha/(2+\alpha)}$$

Otherwise, we have

$$\|\hat{g}_{\hat{\lambda}} - g*\|_n^{1+\alpha/2} \leq O_p(n^{-1/2})I(\hat{g}_{\hat{\lambda}})^{\alpha/2}$$

By Lemma 1 below, using the assumption that the penalty of $\hat{g}_\lambda$ is bounded above by its $L_2(P_n)$ norm, we have that

$$\|g^* - \hat{g}_{\hat{\lambda}}\|_n \leq O_p(n^{-1/(2+\alpha)})M^{\alpha/v(2+\alpha)}\|g^*\|_n^{\alpha/2v(2+\alpha)}$$

Result 2:

Now suppose that the training and validation set are independently sampled, so the values X_i are not necessarily the same. We suppose the training and validation sets are both of size n.

Suppose the penalty normalizes the empirical norm as follows:

$$\sup_{g\in\mathcal{G}} \frac{\|g - g^*\|_T}{I(g) + I(g^*)} \leq R < \infty, \ \sup_{g\in\mathcal{G}} \frac{\|g - g^*\|_V}{I(g) + I(g^*)} \leq R < \infty$$

Suppose for all $\lambda \in \Lambda$, $I^v(\hat{g}_\lambda)$ is upper bounded by its L_2-norm with constants M and M_0:

$$I^v(\hat{g}_\lambda) \leq M\left(\|\hat{g}_\lambda\|_T^2 + \|\hat{g}_\lambda\|_V^2\right) + M_0 = M\|\hat{g}_\lambda\|_{2n}^2 + M_0$$

Then for any $\xi > 0$,

$$\|\hat{g}_{\hat{\lambda}} - g^*\|_V = O_p(n^{-1/(2+\alpha+\xi)})\left(M^{\alpha/v(2+\alpha)}\|g^*\|_{2n}^{\alpha/2v(2+\alpha)} \vee I(g^*)\right)$$

Hence the convergence rate is infinitely close to the optimal convergence rate $O_p(n^{-1/(2+\alpha)})$ attained by $\tilde{\lambda}$, but it doesn't actually attain it.

Proof: We follow the same proof structure of going thru the three cases, modifying the proofs as appropriate:

Case 1: $\|g^* - \hat{g}_{\tilde{\lambda}}\|_V^2$ is the largest
By Lemma 2, we have

$$\big|\|g^* - \hat{g}_{\tilde{\lambda}}\|_T - \|g^* - \hat{g}_{\tilde{\lambda}}\|_V\big| \le O_p(n^{-1/(2+\alpha+\xi)})\left(I(\hat{g}_{\tilde{\lambda}}) + I(g^*)\right)$$

Therefore, for any $\xi > 0$, we have

$$\begin{aligned}\|g^* - \hat{g}_{\hat{\lambda}}\|_V &\le O_p(1)\|g^* - \hat{g}_{\tilde{\lambda}}\|_V \\ &\le O_p(1)\|g^* - \hat{g}_{\tilde{\lambda}}\|_T + O_p(n^{-1/(2+\alpha+\xi)})I(g^*)\end{aligned}$$

Case 2: $\big|(\epsilon, g^* - \hat{g}_{\tilde{\lambda}})_V\big|$ is the largest
The same proof still holds.
Case 3: $\big|(\epsilon, g^* - \hat{g}_{\hat{\lambda}})_V\big|$ is the largest.
Again, we have by Van de geer (10.6),

$$\big|(\epsilon, g^* - \hat{g}_{\hat{\lambda}})_V\big| \le O_p(n^{-1/2})\|\hat{g}_{\hat{\lambda}} - g*\|_V^{1-\alpha/2}(I(g^*) + I(\hat{g}_{\hat{\lambda}}))^{\alpha/2}$$

If $I(g^*) \ge I(g_{\hat{\lambda}})$ is true, then result is clearly attained.
Otherwise, we have

$$\|\hat{g}_{\hat{\lambda}} - g*\|_V^{1+\alpha/2} \le O_p(n^{-1/2})I(\hat{g}_{\hat{\lambda}})^{\alpha/2}$$

By Lemma 1 below, since the penalty is bounded above by the $L_2(P_n)$ norm, it follows that

$$\|g^* - \hat{g}_{\hat{\lambda}}\|_V \le O_p(n^{-1/(2+\alpha)})M^{\alpha/v(2+\alpha)}\|g^*\|_{2n}^{\alpha/2v(2+\alpha)}$$

Lemmas

Lemma 1:

Suppose for all $\lambda \in \Lambda$, the penalty function $I^v(g_\lambda)$ is upper-bounded by $\|g_\lambda\|_n^2 = \frac{1}{n}\sum_{i=1}^n g_\lambda^2(x_i)$ with constants M_0 and M:

$$I^v(g_\lambda) \le M\|g_\lambda\|_n^2 + M_0$$

Suppose there is some function $g^* \in \mathcal{G}$ such that

$$\|g^* - g_\lambda\|_n^{1+\alpha/2} \le O_p(n^{-1/2})I^{\alpha/2}(g_\lambda)$$

then for sufficiently large n,

$$\|g^* - g_\lambda\|_n \le O_p(n^{-1/(2+\alpha)})M^{\alpha/v(2+\alpha)}\|g^*\|_n^{\alpha/2v(2+\alpha)}$$

Proof:
From the assumption that $I^v(g_\lambda)$ is upper-bounded by $\|g_\lambda\|_n^2$,

$$\|g^* - g_\lambda\|_n^{1+\alpha/2} \quad \le \quad O_p(n^{-1/2})\left(M\|g_\lambda\|_n^2 + M_0\right)^{\alpha/2v}$$

If $M_0 > \|g_\lambda\|_n^2$, then the result immediately follows.
Otherwise, if $M_0 \le \|g_\lambda\|_n^2$, then

$$\begin{aligned}\|g^* - g_\lambda\|_n^{1+\alpha/2} &\leq O_p(n^{-1/2}) M^{\alpha/2v} \|g_\lambda\|_n^{\alpha/v} \\ &\leq O_p(n^{-1/2}) M^{\alpha/2v} \left(\|g_\lambda - g^*\|_n + \|g^*\|_n\right)^{\alpha/v}\end{aligned}$$

Case 1: $\|g_\lambda - g^*\|_n \leq \|g^*\|_n$
The result immediately follows.
Case 2: $\|g_\lambda - g^*\|_n > \|g^*\|_n$
We show for sufficiently large n, this case will not occur. Suppose this case occurs. Then

$$\|g^* - g_\lambda\|_n^{1+\alpha/2} \leq O_p(n^{-1/2}) M^{\alpha/v(2+\alpha)} \|g_\lambda - g^*\|_n^{\alpha/v}$$

Rearranging, we have that

$$\|g^* - g_\lambda\|_n^{1+\alpha/2-\alpha/v} \leq O_p(n^{-1/2}) M^{\alpha/v(2+\alpha)}$$

Since the LHS exponent is $1 + \alpha/2 - \alpha/v > 0$, $\|g^* - g_\lambda\|_n$ decreases with n. With sufficiently large n, we can ensure that only Case 1 occurs. (Check this statement!!!)

Note: I believe we can often provide a good estimate of M for the entire class $\mathcal{G}$, which means that we can always estimate the sample size needed to ensure this case never occurs. That is, I believe we can often estimate M s.t.

$$I^v(g) \leq M\|g\|_n^2 + M_0 \forall g \in \mathcal{G}$$

Lemma 2:

Let $P_{n'}$ and $P_{n''}$ be empirical distributions over $\{X_i'\}_{i=1}^n, \{X_i''\}_{i=1}^n$. Let $P_{2n} = \frac{1}{2}\left(P_{n'} + P_{n''}\right)$. Suppose $|X_i| \leq R_X < \infty$.

Let $\mathcal{G}' = \left\{\frac{g - g^*}{I(g) + I(g^*)} : g \in \mathcal{G}, I(g) + I(g^*) > 0\right\}$. Suppose

$$\sup_{f \in \mathcal{G}'} \|f\|_{P_{2n}} \leq R < \infty$$

and

$$H\left(\delta, \mathcal{G}', P_{n'}\right) \leq \tilde{A}\delta^{-\alpha},\ H\left(\delta, \mathcal{G}', P_{n''}\right) \leq \tilde{A}\delta^{-\alpha}$$

Then
BAD :(

$$Pr\left(\sup_{g \in \mathcal{G}} \frac{\left|\|g^* - g\|_{P_n} - \|g^* - g\|_{P_{n''}}\right|}{I(g^*) + I(g)} \geq 6\delta\right) \leq 2\exp\left(2\tilde{A}\delta^{-\alpha} - \frac{4\delta^2}{\left(R + \sqrt{2}\delta\right)^2}\right)$$

Proof: The proof is very similar to that in Pollard 1984 (page 32), so some details below are omitted.

First note that for any function f and h, we have

$$\|f\|_{P_{n'}} - \|h\|_{P_{n'}} \leq \|f - h\|_{P_{n'}} \leq \sqrt{2}\|f - h\|_{P_{2n}}$$

Similarly for $P_{n''}$.

Let $\{h_j\}_{j=1}^N$ be the $\sqrt{2}\delta$-cover for $\mathcal{G}'$ (where $N = N(\sqrt{2}\delta, \mathcal{G}', P_{2n})$). Let h_j be the closest function (in terms of $\|\cdot\|_{P_{2n}}$) to any $f \in \mathcal{G}'$.

$$\begin{aligned}\|f\|_{P_{n'}} - \|f\|_{P_{n''}} &\leq \|f - h_j\|_{P_{n'}} + \left|\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}\right| + \|f - h_j\|_{P_{n''}} \\ &\leq 4\delta + \left|\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}\right|\end{aligned}$$

Then

$$\begin{aligned} Pr\left(\sup_{g\in\mathcal{G}} \frac{\big|\|g^*-g\|_{P_n} - \|g^*-g\|_{P_{n''}}\big|}{I(g^*)+I(g)} \geq 6\delta\right) &\leq Pr\left(\sup_{j\in 1:N} \big|\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}\big| \geq 2\delta\right) \\ &\leq N \max_{j\in 1:N} Pr\left(\big|\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}\big| \geq 2\delta\right) \end{aligned}$$

Now note that

$$\begin{aligned} \big|\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}\big| &= \frac{\left|\|h_j\|^2_{P_{n'}} - \|h_j\|^2_{P_{n''}}\right|}{\|h_j\|_{P_{n'}} + \|h_j\|_{P_{n''}}} \\ &\leq \frac{\left|\|h_j\|^2_{P_{n'}} - \|h_j\|^2_{P_{n''}}\right|}{\sqrt{2}\|h_j\|_{P_{2n}}} \end{aligned}$$

By Hoeffding's inequality,

$$\begin{aligned} Pr\left(\big|\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}\big| \geq 2\delta\right) &\leq Pr\left(\left|\|h_j\|^2_{P_{n'}} - \|h_j\|^2_{P_{n''}}\right| \geq 2\sqrt{2}\delta\|h_j\|_{P_{2n}}\right) \\ &= Pr\left(\left|\sum_{i=1}^n W_i\left(h_j^2(x_i') - h_j^2(x_i'')\right)\right| \geq 2\sqrt{2}n\delta\|h_j\|_{P_{2n}}\right) \\ &\leq 2\exp\left(-\frac{16\delta^2 n^2 \|h_j\|^2_{P_{2n}}}{4\sum_{i=1}^n \left(h_j^2(x_i') - h_j^2(x_i'')\right)^2}\right) \end{aligned}$$

Since

$$\begin{aligned} \sum_{i=1}^n \left(h_j^2(x_i') - h_j^2(x_i'')\right)^2 &\leq \sum_{i=1}^n h_j^4(x_i') + h_j^4(x_i'') \\ &\leq n^2\|h_j\|^4_{P_{2n}} \\ &\leq n^2\|h_j\|^2_{P_{2n}}\left(\|f\|_{P_{2n}} + \|f-h_j\|_{P_{2n}}\right)^4 \\ &\leq n^2\|h_j\|^2_{P_{2n}}\left(R+\sqrt{2}\delta\right)^2 \end{aligned}$$

Hence

$$Pr\left(\big|\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}\big| \geq 2\delta\right) \leq 2\exp\left(-\frac{4\delta^2}{\left(R+\sqrt{2}\delta\right)^2}\right)$$

Since

$$N(\sqrt{2}\delta, \mathcal{G}', P_{2n}) \leq N(\delta, \mathcal{G}', P_{n''}) + N(\delta, \mathcal{G}', P_{n''})$$

then

$$Pr\left(\sup_{g\in\mathcal{G}} \frac{\big|\|g^*-g\|_{P_n} - \|g^*-g\|_{P_{n''}}\big|}{I(g^*)+I(g)} \geq 6\delta\right) \leq 2\exp\left(2\tilde{A}\delta^{-\alpha} - \frac{4\delta^2}{\left(R+\sqrt{2}\delta\right)^2}\right)$$

Using shorthand, we can write that for any $\xi > 0$,

$$\sup_{g\in\mathcal{G}} \frac{\big|\|g^*-g\|_{P_n} - \|g^*-g\|_{P_{n''}}\big|}{I(g^*)+I(g)} = O_p(n^{-1/(2+\alpha+\xi)})$$

Example 1: Sobelov norm

Consider the functions

$$\mathcal{G} = \left\{ g : [0,1] \mapsto \mathbb{R} : \int_0^1 g^{(m)}(z)^2 dz < \infty \right\}$$

Suppose x_i are all unique. Then the Sobelov norm for the class $\{\hat{g}_\lambda \in \mathcal{G} : \lambda \in \Lambda\}$ is bounded above by its $L_2(P_n)$ norm.

$$I^2(\hat{g}_\lambda) = \int_0^1 \left(\hat{g}_\lambda^{(m)}(z) \right)^2 dz \le M \|\hat{g}_\lambda\|_n^2 + M_0 \forall \lambda \in \Lambda$$

Proof:

Let $\tilde{g}$ satisfy $\tilde{g}(x_i) = y_i$ and have the smallest value for $\int_0^1 \left(\tilde{g}^{(m)}(z) \right)^2 dz$. This function $\tilde{g}$ should always exist.

Case 1: $\lambda \le 1/2$

By definition of $\hat{g}_\lambda$

$$\|y - \hat{g}_\lambda\|_n^2 + \lambda^2 I^2(\hat{g}_\lambda) \le \|y - (\tilde{g} - \lambda \hat{g}_\lambda)\|_n^2 + \lambda^2 I^2(\tilde{g} - \lambda \hat{g}_\lambda)$$

Note that

$$\begin{aligned} I^2(\tilde{g} - \lambda \hat{g}_\lambda) &= \int_0^1 \left(\tilde{g}^{(m)} - \lambda \hat{g}_\lambda^{(m)} \right)^2 dz \\ &= 2 \int_0^1 \max \left(\left| \tilde{g}^{(m)} \right|^2, \left| \lambda \hat{g}_\lambda^{(m)} \right|^2 \right) dz \\ &= 2 \left(\int_0^1 \left| \tilde{g}^{(m)} \right|^2 dz + \int_0^1 \left| \lambda \hat{g}_\lambda^{(m)} \right|^2 dz \right) \end{aligned}$$

Hence

$$\lambda^2 I^2(\hat{g}_\lambda) \le \lambda^2 \|\hat{g}_\lambda\|_n^2 + 2\lambda^2 I^2(\tilde{g}) + 2\lambda^4 I^2(\hat{g}_\lambda)$$

The following ineq follows, where the RHS is maximized when $\lambda = 1/2$

$$I^2(\hat{g}_\lambda) \le \frac{\lambda^2}{\lambda^2 - 2\lambda^4} \left(\|\hat{g}_\lambda\|_n^2 + 2I^2(\tilde{g}) \right) \le 2\|\hat{g}_\lambda\|_n^2 + 4I^2(\tilde{g})$$

Case 2: $\lambda > 1/2$

By definition of $\hat{g}_\lambda$

$$\|y - \hat{g}_\lambda\|_n^2 + \lambda^2 I^2(\hat{g}_\lambda) \le \|y\|_n^2$$

The RHS is maximized when $\lambda = 1/2$, so

$$I^2(\hat{g}_\lambda) \le 4\|y\|_n^2$$

Hence we have an upper bound for the Sobelov norm

$$I^2(\hat{g}_\lambda) \le 2\|\hat{g}_\lambda\|_n^2 + 4I^2(\tilde{g}) + 4\|y\|_n^2$$

Appendix

A cute lemma I found but never used: Supposing that $I^v(\hat{g}_\lambda)$ is continuous in λ, then given training data T,

$$\frac{\partial}{\partial \lambda} L_T(\hat{g}_\lambda, \lambda) \quad = \quad 2\lambda I^v(\hat{g}_\lambda)$$

Also, L_T is convex in λ.
Proof:
By definition,

$$L_T(\hat{g}_\lambda, \lambda) = \|y - \hat{g}_\lambda\|_T^2 + \lambda^2 I^v(\hat{g}_\lambda) \leq \|y - \hat{g}_{\lambda'}\|_T^2 + \lambda^2 I^v(\hat{g}_{\lambda'}) = L_T(\hat{g}_{\lambda'}, \lambda)$$

Then we can provide upper and lower bounds for $L_T(\hat{g}_{\lambda_2}, \lambda_2) - L_T(\hat{g}_{\lambda_1}, \lambda_1)$:

$$\begin{aligned} L_T(\hat{g}_{\lambda_2}, \lambda_2) - L_T(\hat{g}_{\lambda_1}, \lambda_1) &\leq L_T(\hat{g}_{\lambda_1}, \lambda_2) - L_T(\hat{g}_{\lambda_1}, \lambda_1) \\ &= \|y - \hat{g}_{\lambda_1}\|_T^2 + \lambda_2^2 I^v(\hat{g}_{\lambda_1}) - \|y - \hat{g}_{\lambda_1}\|_T^2 - \lambda_1^2 I^v(\hat{g}_{\lambda_1}) \\ &= (\lambda_2^2 - \lambda_1^2) I^v(\hat{g}_{\lambda_1}) \end{aligned}$$

$$\begin{aligned} L_T(\hat{g}_{\lambda_2}, \lambda_2) - L_T(\hat{g}_{\lambda_1}, \lambda_1) &\geq L_T(\hat{g}_{\lambda_2}, \lambda_2) - L_T(\hat{g}_{\lambda_2}, \lambda_1) \\ &= \|y - \hat{g}_{\lambda_2}\|_T^2 + \lambda_2^2 I^v(\hat{g}_{\lambda_2}) - \|y - \hat{g}_{\lambda_2}\|_T^2 - \lambda_1^2 I^v(\hat{g}_{\lambda_2}) \\ &= (\lambda_2^2 - \lambda_1^2) I^v(\hat{g}_{\lambda_2}) \end{aligned}$$

So suppose WLOG $\lambda_2 > \lambda_1$:

$$(\lambda_2 + \lambda_1) I^v(\hat{g}_{\lambda_2}) \leq \frac{L_T(\hat{g}_{\lambda_2}, \lambda_2) - L_T(\hat{g}_{\lambda_1}, \lambda_1)}{\lambda_2 - \lambda_1} \leq (\lambda_2 + \lambda_1) I^v(\hat{g}_{\lambda_1})$$

So as $\lambda_1 \to \lambda_2 = \lambda$, we have by the sandwich theorem,

$$\frac{\partial}{\partial \lambda} L_T(\hat{g}_\lambda, \lambda) = 2\lambda I^v(\hat{g}_\lambda)$$

Furthermore, given training data T

$$\frac{\partial}{\partial \lambda} L_T(\hat{g}_\lambda, \lambda) = \frac{\partial}{\partial \lambda} \|y - \hat{g}_\lambda\|_T^2 + 2\lambda I^v(\hat{g}_\lambda) + \lambda^2 \frac{\partial}{\partial \lambda} I^v(\hat{g}_\lambda)$$

then, combining this with the lemma, we have that

$$\frac{\partial}{\partial \lambda} \|y - \hat{g}_\lambda\|_T^2 = -\lambda^2 \frac{\partial}{\partial \lambda} I^v(\hat{g}_\lambda)$$

Finally, to see that L_T is convex in λ, note that

$$\frac{\partial^2}{\partial \lambda^2} L_T(\hat{g}_\lambda, \lambda) = 2I^v(\hat{g}_\lambda) + 2\lambda v I^{v-1}(\hat{g}_\lambda) \frac{\partial}{\partial \lambda} I(\hat{g}_\lambda) > 0$$

since $\frac{\partial}{\partial \lambda} I(\hat{g}_\lambda) > 0$.