Convergence Rates of λ

October 26, 2016

Let's bound $\|\hat{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\lambda}}\|$ instead.
This will actually get rid of the geometric mean term in Theorems 1 and 3.
We will suppose that the data is generated from the model:

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

where ϵ are independent, sub-gaussian errors. The penalized regression models are

$$\hat{g}(\cdot|\boldsymbol{\lambda}) = \arg\min_{g \in \mathcal{G}} L_T(g|\boldsymbol{\lambda})$$

Let the model class after fitting on the training data be

$$\mathcal{G}(T) = \{\hat{g}(\cdot|\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \Lambda\}$$

The selected penalty parameters are

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in \Lambda} \|y - \hat{g}(\cdot|\boldsymbol{\lambda})\|_V^2$$

Convergence of $\hat{\lambda}$ to $\tilde{\lambda}$

Suppose that if $\|\epsilon\|_T \le 2\sigma$, then $\mathcal{G}(T)$ satisfies the entropy condition

$$\int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V) du \le \psi_T(R)$$

Furthermore, suppose that

$$\frac{\psi_T(u)}{u^2}$$

is nonincreasing wrt to u for all $u > 0$.
Let $L_V^*(\boldsymbol{\lambda}) = \|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|_V^2$ be the true validation loss and let $\tilde{\boldsymbol{\lambda}}$ be the global minimizer of $L_V^*(\boldsymbol{\lambda})$.

$$\tilde{\boldsymbol{\lambda}} = \arg\min_{\lambda} L_V^*(\boldsymbol{\lambda})$$

Let $\tilde{\boldsymbol{\lambda}}_{gen}$ be the global minimizer of the generalization error.

$$\tilde{\boldsymbol{\lambda}}_{gen} = \arg\min_{\lambda} E_V\left[L_V^*(\boldsymbol{\lambda})\right] = \arg\min \|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|^2$$

- **Local strong convexity assumption:** Suppose that there is a neighborhood $N(\tilde{\boldsymbol{\lambda}}_{gen})$ around $\tilde{\boldsymbol{\lambda}}_{gen}$ such that the true validation loss is smooth in $\boldsymbol{\lambda}$ for all $\boldsymbol{\lambda} \in N\left(\tilde{\boldsymbol{\lambda}}_{gen}\right)$ and for all $\boldsymbol{\lambda} \in N(\tilde{\boldsymbol{\lambda}})$, the true validation loss is m-strongly convex in $\boldsymbol{\lambda}$ for some $m > 0$:

  $$\nabla_\lambda^2 L_V^*(\boldsymbol{\lambda}) = \nabla_\lambda^2 \|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|_V^2 \succeq mI$$

  Important: m cannot shrink in n_T and n_V

- **Lipschitz assumption:** Let us also assume that for all $\boldsymbol{\lambda} \in N(\tilde{\boldsymbol{\lambda}}_{gen})$, the fitted functions are locally K-Lipschitz:
$$\left\|\hat{g}(\cdot|\boldsymbol{\lambda}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V \le K\|\boldsymbol{\lambda} - \tilde{\boldsymbol{\lambda}}\|$$
Important: K cannot grow in n_V, n_T (I don't even think this holds for ridge regression though...)

If δ is chosen such that

$$\sqrt{n_V}\delta^2 \quad \ge \quad 2C\left[\psi_T\left(2\delta\right) \vee \left(2\delta\right)\right]$$

then we have

$$Pr\left(\|\hat{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\lambda}}\| \ge \delta \wedge \|\epsilon\|_V \le 2\sigma \wedge \|\epsilon\|_T \le 2\sigma\right) \quad \le \quad c\exp\left(-\frac{n_V\delta^2 m^2}{c^2K^2}\right)$$

for some constant c.

Furthermore, this completely removes the geometric term since we also have that

$$\begin{aligned} Pr\left(\left\|\hat{g}(\cdot|\boldsymbol{\lambda}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V \ge \frac{\delta}{K} \wedge \|\epsilon\|_V \le 2\sigma \wedge \|\epsilon\|_T \le 2\sigma\right) &\le Pr\left(\|\hat{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\lambda}}\| \ge \delta \wedge \|\epsilon\|_V \le 2\sigma \wedge \|\epsilon\|_T \le 2\sigma\right) \\ &\le c\exp\left(-\frac{n_V\delta^2 m^2}{c^2K^2}\right) \end{aligned}$$

Proof

Let $\hat{\boldsymbol{\lambda}}$ be the global minimizer of the validation loss. Therefore

$$\left\|y - \hat{g}(\cdot|\hat{\boldsymbol{\lambda}})\right\|_V^2 \le \left\|y - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V^2$$

Recall the basic inequality

$$L_V^*(\hat{\boldsymbol{\lambda}}) - L_V^*(\tilde{\boldsymbol{\lambda}}) = \left\|g^* - \hat{g}(\cdot|\hat{\boldsymbol{\lambda}})\right\|_V^2 - \left\|g^* - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V^2 \le 2\left\langle \epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\rangle_V$$

Suppose that $\hat{\boldsymbol{\lambda}} \in N(\tilde{\boldsymbol{\lambda}}_{gen})$. Using the mean value theorem and the local strong convexity assumption, there is some $\alpha \in (0,1)$ such that

$$\begin{aligned} L_V^*(\boldsymbol{\lambda}) - L_V^*(\tilde{\boldsymbol{\lambda}}) &= \left(\hat{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\lambda}}\right)^\top \nabla_\lambda^2 L_V^*(\boldsymbol{\lambda})\big|_{\boldsymbol{\lambda}=\alpha\tilde{\boldsymbol{\lambda}}+(1-\alpha)\hat{\boldsymbol{\lambda}}}\left(\hat{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\lambda}}\right) \\ &\ge \left(\min_\alpha \nabla_\lambda^2 L_V^*(\boldsymbol{\lambda})\big|_{\boldsymbol{\lambda}=\alpha\tilde{\boldsymbol{\lambda}}+(1-\alpha)\hat{\boldsymbol{\lambda}}}\right)\|\hat{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\lambda}}\|_2^2 \\ &\ge m\|\hat{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\lambda}}\|_2^2 \end{aligned}$$

Therefore we get
$$m\|\boldsymbol{\lambda} - \tilde{\boldsymbol{\lambda}}\|_2^2 \le 2\left\langle \epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\rangle_V$$

Anyhow, if we assume local strong convexity and the Lipschitz condition, we can proceed with a peeling argument

$$\begin{aligned} Pr\left(\|\hat{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\lambda}}\| \ge \delta\right) &= \sum_{s=0}^\infty Pr\left(2^s\delta \le \|\hat{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\lambda}}\| \le 2^{s+1}\delta\right) \\ &= \sum_{s=0}^\infty Pr\left(\|\hat{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\lambda}}\|^2 \ge 2^{2s}\delta^2 \wedge \|\hat{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\lambda}}\| \le 2^{s+1}\delta\right) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{s=0}^{\infty} Pr\left(2\left\langle \epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\rangle_V \geq m 2^{2s}\delta^2 \wedge \left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V \leq 2^{s+1}\delta K\right) \\
&\leq \sum_{s=0}^{\infty} Pr\left(\sup_{\boldsymbol{\lambda},\boldsymbol{\lambda}' \in N(\tilde{\boldsymbol{\lambda}}_{gen}): \|\hat{g}(\cdot|\lambda) - \hat{g}(\cdot|\lambda')\|_V \leq 2^{s+1}\delta K} \left\langle \epsilon, \hat{g}(\cdot|\boldsymbol{\lambda}) - \hat{g}(\cdot|\boldsymbol{\lambda}')\right\rangle_V \geq m 2^{2s-1}\delta^2\right)
\end{aligned}
$$

To apply the lemma based on vandegeer corollary 8.3 (see below), we must check all the conditions are satisfied.

We choose δ such that

$$
\begin{aligned}
\frac{\sqrt{n_V}}{8} &\geq \frac{C}{4\delta^2}\left[\psi_T(2\delta) \vee (2\delta)\right] \\
&\geq \frac{C}{2^{2s+2}\delta^2}\left[\psi_T\left(2^{s+1}\delta\right) \vee \left(2^{s+1}\delta\right)\right]
\end{aligned}
$$

where the second line follows from the assumption that $\psi_T(u)/u^2$ is nonincreasing wrt u. Hence we have satisfied the condition in corollary 8.3. So for all $s = 0, 1, ...$ since

$$
\sqrt{n_V} 2^{2s-1}\delta^2 \geq C\left[\psi_T\left(2^{s+1}\delta\right) \vee \left(2^{s+1}\delta\right)\right]
$$

we have

$$
Pr\left(\sup_{\boldsymbol{\lambda},\boldsymbol{\lambda}' \in N(\tilde{\boldsymbol{\lambda}}_{gen}): \|\hat{g}(\cdot|\lambda) - \hat{g}(\cdot|\lambda')\|_V \leq 2^{s+1}\delta K} \left\langle \epsilon, \hat{g}(\cdot|\boldsymbol{\lambda}) - \hat{g}(\cdot|\boldsymbol{\lambda}')\right\rangle_V \geq m 2^{2s-1}\delta^2 \wedge \|\epsilon\|_V \leq 2\sigma \wedge \|\epsilon\|_T \leq 2\sigma\right) \leq \exp\left(-n_V \frac{2^{4s\cdots}}{4C^2 (2\cdots}\right.
$$

Hence we have

$$
\begin{aligned}
Pr\left(\|\hat{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\lambda}}\| \geq \delta \wedge \|\epsilon\|_V \leq 2\sigma \wedge \|\epsilon\|_T \leq 2\sigma\right) &\leq C\sum_{s=0}^{\infty} \exp\left(-n_V \frac{2^{4s-2}\delta^4 m^2}{4C^2\left(2^{s+1}\delta\right)^2 K^2}\right) \\
&\leq c\exp\left(-\frac{n_V\delta^2}{c^2}\right)
\end{aligned}
$$

for some constant c.

Example

If $\mathcal{G}$ is a parametric family, we must choose

$$
\delta \geq R\left(n_V^{-1/2}\right)
$$

for some constant R, so we have an asymptotic convergence rate for

$$
Pr\left(\|\hat{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\lambda}}\| \geq R n_V^{-1/2} \wedge \|\epsilon\|_V \leq 2\sigma \wedge \|\epsilon\|_T \leq 2\sigma\right) \leq c\exp\left(-\frac{R^2 m^2}{c^2 K^2}\right)
$$

Jean's questions

- Locally m-strongly convex where m doesn't shrink with n_T or n_V seems like a strong assumption.
- Fitted functions are locally K-Lipschitz where K doesn't change with n_T seems like a strong assumption too.
  - Counterexample: In ridge regression, if p grows with n, I believe the Lipschitz constant is on the order of λ_{min}^{-2}. But if λ_{min} is shrinking with n, then K is changing with n.

- We need to ensure that $N(\tilde{\boldsymbol{\lambda}}_{gen})$ contains $\tilde{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\lambda}}$ with high probability - can we ensure this? Thoughts:
  - We might be able to bootstrap results to show that $N(\tilde{\boldsymbol{\lambda}}_{gen})$ contains $\tilde{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\lambda}}$ with high probability.
  - We know that the difference

    $$\|\hat{g}_{\hat{\lambda}} - g^*\|_V - \|\hat{g}_{\tilde{\lambda}} - g^*\|_V \leq \delta$$

    with "high probability". So if the global minimizer of $\|\hat{g}_{\tilde{\lambda}} - g^*\|_V$ is more than δ smaller than all other local minimas of $L_V^*(\boldsymbol{\lambda})$, then $\hat{\lambda}$ will be located in the same region as $\tilde{\lambda}$. By definition, this region must be quasi-convex.
  - If we can show that this region is $N(\tilde{\boldsymbol{\lambda}}_{gen})$ and it is strongly convex, we'd be done. The problem is if this region has crazy behavior (e.g. the loss is very flat wrt λ)