First order expansion of convex regularized estimators

Pierre C Bellec and Arun Kuchibhotla
Rutgers University University of Pennsylvania

What is a First Order Expansion?

- Classical estimators for regression are defined as

$$\hat{\beta} := \operatorname{argmin}_{\beta\in\mathbb{R}^p} \; n^{-1}\sum_{i=1}^n \ell(Y_i, X_i^\top\beta),$$

for a loss function $\ell(\cdot,\cdot)$.

- Under differentiability of ℓ (in second argument), it can be proved for a matrix K that

$$\hat{\beta} = \beta^* + \tfrac{1}{n}\sum_{i=1}^n K^{-1}X_i\ell'(Y_i, X_i^\top\beta^*) + O_p\left(\tfrac{1}{n}\right). \quad (1)$$

This is a **first order expansion** of $\hat{\beta}$.

- In addition to proving $\hat{\beta}$ is "close" to β^*, it also gives a precise characterization of error $\hat{\beta}-\beta^*$.
- In the context of convex regularized estimators defined as

$$\hat{\beta} := \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \; n^{-1}\sum_{i=1}^n \ell(Y_i, X_i^\top\beta) + h(\beta),$$

for any convex regularizer $h(\cdot)$, many existing results show that $\hat{\beta}$ is close to β^*.

What is a first order expansion for convex regularized estimators?

Intuitive Expansion

- For unregularized estimators (with $h\equiv 0$), the expansion $\eta = \beta^* + n^{-1}\sum_{i=1}^n K^{-1}\ell'(Y_i, X_i^\top\beta^*)$ of $\hat{\beta}$ can be thought of as minimizing (over $\beta\in\mathbb{R}^p$)

$$F_n(\beta^*)+(\beta-\beta^*)^\top F_n'(\beta^*)+\frac{1}{2}\|\beta-\beta^*\|_K^2+h(\beta), \quad (2)$$

where $h\equiv 0$ and $F_n(\beta) := n^{-1}\sum_{i=1}^n \ell(Y_i, X_i^\top\beta)$.

- If K is approximately the second derivative of $F_n(\beta)$, then the objective of η is a quadratic approximation of the objective of $\hat{\beta}$.
- For convex and non-zero regularizers $h(\cdot)$, we can define the first order expansion as η minimizing (2).

Result for Regularized Linear Regression

- Consider $(X_i, Y_i)\in\mathbb{R}^p\times\mathbb{R}$ i.i.d random vectors

$$Y_i = X_i^\top\beta^* + \varepsilon_i, \quad \varepsilon_i \text{ independent of } X_i,$$

the penalized estimator $\hat{\beta}$ and its approximation η

$$\hat{\beta} := \operatorname{argmin}_{\beta\in\mathbb{R}^p} \; n^{-1}\sum_{i=1}^n (Y_i - X_i^\top\beta)^2 + 2h(\beta),$$

$$\eta := \operatorname{argmin}_{\beta\in\mathbb{R}^p} \; \|\Sigma(\beta-\beta^*)\|^2 - \tfrac{2}{n}\varepsilon^T\mathsf{X}(\beta-\beta^*) + 2h(\beta)$$

where X has rows $X_1,\dots,X_n$. Here η has simplified form thanks to to $K := \mathbb{E}[\nabla_2 F_n(\beta^*)] = \Sigma$ for squared loss $\ell(y,u) = (y-u)^2/2$.

- $h(\beta) = \lambda\|\beta\|_1$ (Lasso) or $h(\beta) = \lambda\sum_{k=1}^G \|\beta_{G_k}\|$ for M groups of size $d = p/M$ (Group-Lasso).
Then for $\Sigma = \mathbb{E}[XX^\top]$, under Restricted Eigenvalue (RE) or bounded condition number for Σ:

	Lasso	Group-Lasso
Tuning $\lambda \gtrsim$	$[\frac{2}{n}\log\frac{p}{s}]^{\frac12}$	$[d+\frac{2}{n}\log\frac{p}{s}]^{\frac12}$
Minimax r_n	$\|\hat{\beta}-\beta^*\| \lesssim r_n$ $r_n = [\frac{2s}{n}\log\frac{p}{s}]^{\frac12}$	$\|\hat{\beta}-\beta^*\| \lesssim r_n$ $[\frac{d}{n}+\frac{s}{n}\log\frac{M}{s}]^{\frac12}$
$RE(s,\Sigma)\ge C$	$\|\eta-\hat{\beta}\| \lesssim r_n^{3/2}$	$\|\eta-\hat{\beta}\| \lesssim r_n^{3/2}$
$\phi_{cond}(\Sigma)\le C$	$\|\eta-\hat{\beta}\| \lesssim r_n^{2}$	$\|\eta-\hat{\beta}\| \lesssim r_n^{2}$

- Approximation $\|\hat{\beta}-\eta\|$ negligible compared to r_n.
- Similar results obtainable for Slope, Nuclear norm,...

Application: Exact Risk Identity

For squared error loss with arbitrary proper convex penalty $h(\cdot)$, if $X_1,\dots,X_n$ are i.i.d standard normal random vectors in $\mathbb{R}^p$, then we get

$$\frac{\|\hat{\beta}-\beta^*\|_2}{\left(\mathbb{E}[\|\operatorname{prox}_h(\beta^*+n^{-1/2}\sigma^*Z)-\beta^*\|_2^2]\right)^{1/2}} = 1+o_p(1),$$

whenever $s\log(ep/s)/n\to\infty$ and $s/p\to 0$ (Lasso). Here $\operatorname{prox}_h(\cdot)$ is the proximal operator and Z is a standard Gaussian random vector.

Result for Regularized Logistic Regression

- Suppose $(X_i, Y_i)\in\mathbb{R}^p\times\{0,1\}$ are i.i.d $\mathbb{P}(Y_i=1|X_i) = \frac{1}{1+\exp(-X_i^\top\beta^*)}$ (logistic model).
- Loss $\ell(y,u) = yu - \log(1+e^u)$ and estimator $\hat{\beta} := \operatorname{argmin}_{\beta\in\mathbb{R}^p} \; \frac{1}{n}\sum_{i=1}^n \ell(Y_i, X_i^\top\beta) + h(\beta)$.
- Due to the non-constancy of second derivative of $\beta\mapsto\ell(y, x^\top\beta)$, there is an extra term in $\|\eta-\hat{\beta}\|$

	Lasso	Group-Lasso
Tuning $\lambda \gtrsim$	$[\frac{2}{n}\log\frac{p}{s}]^{\frac12}$	$[d+\frac{2}{n}\log\frac{p}{s}]^{\frac12}$
Minimax r_n	$\|\hat{\beta}-\beta^*\| \lesssim r_n$ $r_n = [\frac{2s}{n}\log\frac{p}{s}]^{\frac12}$	$\|\hat{\beta}-\beta^*\| \lesssim r_n$ $[\frac{d}{n}+\frac{s}{n}\log\frac{M}{s}]^{\frac12}$
$RSC(s)\ge C$	$\|\eta-\hat{\beta}\| \lesssim r_n^{3/2}(1+r_n^3\sqrt{n})$	
$\phi_{cond}(K)\le C$	$\|\eta-\hat{\beta}\| \lesssim r_n^{2}(1+r_n^3\sqrt{n})$	

Application: Confidence intervals for $a^T\beta^*$

Linear regression $y = \mathsf{X}\beta^* + \varepsilon$, $y = (Y_1,\dots,Y_n)$
- X is the design matrix with rows $X_1,\dots,X_n$,
- $a\in\mathbb{R}^p$ direciton of interest, with $\|\Sigma^{-1/2}a\|_2 = 1$.

De-biased estimate

$$\hat{\theta} = a^T\hat{\beta} + n^{-1}a^T\Sigma^{-1}\mathsf{X}^T(y-\mathsf{X}\hat{\beta})$$

If $X_1,\dots,X_n$ are iid $N(0,\Sigma)$ independent of $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)\sim N(0, I_n)$ then, provided $r_n^2 n^{1/3}\to 0$,

$$\sqrt{n}(\hat{\theta} - a^T\beta^*) \to^d N(0,1).$$

	Lasso	Group-Lasso
Tuning $\lambda \gtrsim$	$[\frac{2}{n}\log\frac{p}{s}]^{\frac12}$	$[d+\frac{2}{n}\log\frac{p}{s}]^{\frac12}$
Minimax r_n	$r_n = [\frac{2s}{n}\log\frac{p}{s}]^{\frac12}$	$[\frac{d}{n}+\frac{s}{n}\log\frac{M}{s}]^{\frac12}$
$r_n^2 n^{1/3}\to 0$	$\frac{s\log(p/s)}{n^{2/3}}\to 0$	$\frac{sd+s\log\frac{M}{s}}{n^{2/3}}\to 0$

Goes beyond the $s\lesssim\sqrt{n}$ (Lasso) or $sd+s\log(M/s)\lesssim\sqrt{n}$ (Group-L) requirement of previous studies.

Full paper and supplement (ArXiv):