Lecture 6: February 14

Lecturer: Alessandro Rinaldo *Scribes: Beomjo Park*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

6.1 Notes on Regression

Let $Y_1, \cdots, Y_n$ be independent response variable in $\mathbb{R}$. In particular,

$$Y_i = \mathsf{E}[Y_i] + \epsilon_i, \qquad \epsilon_i \in \mathrm{SG}(\sigma^2)$$

Suppose we observed n covariates $(x_1, \cdots, x_n) \subseteq \mathbb{R}^d$; fixed. Further assume that the first coordinate of each x_i be 1. Consider a following *model*:

$$Y_i = f(X_i) + \eta_i, \qquad f : \mathbb{R}^d \to \mathbb{R}, \qquad \mathsf{E}(\eta_i) = 0$$

The model specification need not be linear, however we often choose $f(x_i) = x_i^\top \gamma,\ \gamma \in \mathbb{R}^d$ for the brevity. Under the squared error loss, our goal is to solve the following problem:

$$\min_\gamma \sum_{i=1}^n \mathsf{E}(y_i - x_i^\top \gamma)^2$$

Let β be the minimizer of the problem. Assuming $\Sigma_n = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ be full rank,

$$\beta = \Sigma_n^{-1} \frac{1}{n} \sum_{i=1}^n x_i \mu_i = (X^\top X)^{-1} X^\top \boldsymbol{\mu}$$

where $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^\top$, $\boldsymbol{X} = (x_1, \cdots, x_n)^\top \subseteq \mathbb{R}^{n\times d}$, $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)^\top$ and $\mu_i := \mathsf{E}[Y_i]$. With empirical risk minimization, we estimate β with $\hat\beta = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$. Note that the excess risk $R(\gamma)$ becomes

$$R(\gamma) = \mathsf{E}\left[\frac{1}{n}\sum_{i=1}^n (Y_i - x_i^\top \gamma)^2 - \frac{1}{n}\sum_{i=1}^n (Y_i - x_i^\top \beta)^2\right]$$

See [HKZ12] for the probabilistic bound on $R(\gamma)$ and details.

6.2 Concentration of L-Lipschitz function of Gaussian vectors

Recall from *Mill's ratio*,

$$X \sim \mathrm{N}(\mu, \sigma^2) \implies \mathsf{P}(|X-\mu| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right), \qquad t > 0$$

Definition 6.1 (Lipschitz Condition) *$f : \mathbb{R}^d \to \mathbb{R}$ satisfies L-Lipschitz condition with respect to Euclidean norm, if*

$$\exists L \in \mathbb{R} \text{ s.t. } |f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L\|\boldsymbol{x}-\boldsymbol{y}\|, \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$$

Theorem 6.2 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz. Suppose $\boldsymbol{Z} := (Z_1, \cdots, Z_d)^\top \sim \mathrm{N}_d(\mathbf{0}, \sigma^2 I_d)$. Then*

$$\mathsf{P}\left(|f(\boldsymbol{Z}) - \mathsf{E}[f(\boldsymbol{Z})]| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sigma^2 L^2}\right), \qquad t > 0$$

That is, $f(\boldsymbol{Z}) \in \mathrm{SG}\left(\sigma^2 L^2\right)$

The take-home message is that the concentration of any L-Lipschitz function of isotropic Gaussian random vector is like a scalar Gaussian variable with variance $\sigma^2 L^2$ **independent to the dimension** d.

Remark. Historically, $\mathsf{E}(\cdot)$ was $\mathsf{M}(\cdot)$ where M denotes a median operator. See [BLM13] for detail.

Remark. Theorem 6.2 holds even when $d \to \infty$, which essentially is an isotropic Gaussian Process.

Corollary 6.3 *Let $\boldsymbol{X} = (X_1, \cdots, X_d)$ be i.i.d. random variables with $X_i \sim \mathrm{Unif}[0,1]$. Suppose $f : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz. Then,*

$$\mathsf{P}\left(|f(\boldsymbol{X}) - \mathsf{E}[f(\boldsymbol{X})]| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2L_\alpha^2}\right)$$

where L_α depends on L.

Proof: *Use inverse CDF method.* ■

Maxima of Gaussians

Lemma 6.4 *Let $\boldsymbol{Y} := (Y_1, \cdots, Y_d)^\top \sim \mathrm{N}_d(\mathbf{0}, \Sigma),\ \Sigma \in S_d^+$*

$$\mathsf{P}\left(\left|\max_i Y_i - \mathsf{E}(\max_i Y_i)\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sigma_{\max}^2}\right), \qquad t > 0$$

where $\sigma_{\max}^2 = \max_i \Sigma_{ii}$

Proof: Let $\boldsymbol{Y} = \Sigma^{1/2}\boldsymbol{z}$ where $\Sigma^{1/2}$ is a square root of Σ, i.e, $\Sigma = \Sigma^{1/2}(\Sigma^{1/2})^\top$ and $\boldsymbol{z} \sim \mathrm{N}_d(\mathbf{0}, I_d)$.

Claim: $f : \boldsymbol{x} \in \mathbb{R}^d \mapsto \max_i(\Sigma^{1/2}\boldsymbol{x})_i$ is L-Lipschitz with $L = \max_i \sqrt{\Sigma_{ii}}$

Fix coordinate i and let $\Sigma_i^{1/2}$ be the ith row of $\Sigma^{1/2}$. For all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$

$$\begin{aligned}\left|(\Sigma^{1/2}\boldsymbol{x})_i - (\Sigma^{1/2}\boldsymbol{y})_i\right| = \left|\Sigma_i^{1/2}(\boldsymbol{x}-\boldsymbol{y})\right| &\leq \sqrt{\Sigma_i^{1/2}\left(\Sigma_i^{1/2}\right)^\top} \cdot \|\boldsymbol{x}-\boldsymbol{y}\| \qquad (\because \text{Cauchy-Schwartz Ineq.}) \\ &= \sqrt{\Sigma_{ii}} \cdot \|\boldsymbol{x}-\boldsymbol{y}\|\end{aligned}$$

■

Theorem 6.5 (Talagrand) *Let $\boldsymbol{X} = (X_1, \cdots, X_d)^\top$ be independent random variable with $X_i \in [0,1], \forall i$. Suppose $f : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz with respect to Euclidean norm and f be convex function. Then,*

$$\mathsf{P}\left(|f(\boldsymbol{X}) - \mathsf{E}[f(\boldsymbol{X})]| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2L^2}\right), \qquad t > 0$$

As in Theorem 6.2, E can be replaced to M. Theorem 6.5 is a corollary of *Convex Distance Inequality* [T95]. See [BLM13] for detail.

Theorem 6.6 (Convex Distance Inequality) *For any subset $A \subseteq \mathcal{X}^n$*

$$\mathsf{P}(X \in A)\mathsf{P}\left(\sup_{\alpha \in [0,\infty)^n : \|\alpha\|=1} d_\alpha(X, A) \geq t\right) \leq \exp(-t^2/4)$$

where $d_\alpha(x, A) = \inf_{y \in A} \sum_{i : x_i \neq y_i} |\alpha_i|$ s.t. $\alpha \in [0,\infty)^n$; weighted Hamming distance.

6.3 Prelude to the Next topic

Until now, we consider index set $\mathcal{I}$ to be finite, i.e., $|\mathcal{I}| < \infty$. However, often we need to bound either in probability or in expectation of the form $\max_{i \in \mathcal{I}} X_i$ or $\max_{i \in \mathcal{I}} |X_i|$ where $|\mathcal{I}| = \infty$ under the assumption $X_i \in \mathrm{SG}(\cdot)$ or $\mathrm{SE}(\cdot)$. One strategy is to approximate $\max_{i \in \mathcal{I}} X_i$ with finite covers of a ball where the maximum within the ball be bounded by that within the covers.

References

[HKZ12] Hsu, D. and Kakade, S. M. and Zhang, T. (2012) *A tail inequality for quadratic forms of subgaussian random vectors*, Electron. Commun. Prob., **17**, No.52 pp. 1–6.

[BLM13] Boucheron, S., Lugosi, G., and Massart, P. (2013) *Concentration inequalities: A nonasymptotic theory of independence.* Oxford university press.

[T95] Talagrand, M. (1995) *Concentration of measure and isoperimetric inequalities in product spaces.* Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques, 81(1), 73-205.