Erratum
Prediction and estimation consistency of sparse multi-class penalized optimal scoring

IRINA GAYNANOVA

Department of Statistics
Texas A&M University
MS 3143
College Station, TX 77843
E-mail: irinag@stat.tamu.edu

An error has been found in the proof of Theorem 3 in the bound (9) for $I_2$ term on page 305. The error is in incorrect statement of Lemma 8 from [2] due to differences in norm notation used across the papers. Nevertheless, the statement of Theorem 3 remains true with an adapted proof that we state here.

First, we prove the supplementary Lemma S.1 that bounds the $\ell_2$ norm of the iid sum of random vectors with sub-exponential marginals.

**Lemma S.1.** Let $\gamma_1, \ldots, \gamma_n \in \mathbb{R}^d$ be iid random vectors, $\mathbb{E}(\gamma_i) = 0$ and let $L = \sup_{\|x\|_2 = 1} \| \langle x, \gamma_1 \rangle \|_{\psi_1}$. Then for some constant $C > 0$ and any $t > 0$

$$\text{pr} \left( \| \frac{1}{n} \sum_{i=1}^n \gamma_i \|_2 \geq t \right) \leq 5^d 2 \exp \left[ -C n \min \left( \frac{t^2}{L^2}, \frac{t}{L} \right) \right].$$

**Proof.** Let $v_1, v_2, \ldots$ be the $\varepsilon$-net of a unit sphere $S^{d-1}$ with $\varepsilon = 0.5$. Let

$$v(\gamma) = \frac{1}{\| \frac{1}{n} \sum_{i=1}^n \gamma_i \|_2} \frac{1}{n} \sum_{i=1}^n \gamma_i.$$

Then there exists $j = j(\gamma)$ such that $\| v(\gamma) - v_j \|_2 \leq 0.5$. Therefore

$$\left\| \frac{1}{n} \sum_{i=1}^n \gamma_i \right\|_2 = \langle \frac{1}{n} \sum_{i=1}^n \frac{1}{\| \gamma_i \|_2} \gamma_i, \frac{1}{n} \sum_{i=1}^n \gamma_i \rangle = \langle v(\gamma), \frac{1}{n} \sum_{i=1}^n \gamma_i \rangle = \langle v_j, \frac{1}{n} \sum_{i=1}^n \gamma_i \rangle + \langle v(\gamma) - v_j, \frac{1}{n} \sum_{i=1}^n \gamma_i \rangle$$

$$\leq \left| \langle v_j, \frac{1}{n} \sum_{i=1}^n \gamma_i \rangle \right| + 0.5 \left\| \frac{1}{n} \sum_{i=1}^n \gamma_i \right\|_2.$$

From the above display

$$\left\| \frac{1}{n} \sum_{i=1}^n \gamma_i \right\|_2 \leq 2 \left| \langle v_j, \frac{1}{n} \sum_{i=1}^n \gamma_i \rangle \right|,$$

and therefore

$$\text{pr} \left( \| \frac{1}{n} \sum_{i=1}^n \gamma_i \|_2 \geq t \right) \leq \text{pr} \left( \left| \langle v_j, \frac{1}{n} \sum_{i=1}^n \gamma_i \rangle \right| \geq t/2 \right) \leq \text{pr} \left( \bigcup_i \left\{ \left| \langle v_i, \frac{1}{n} \sum_{i=1}^n \gamma_i \rangle \right| \geq t/2 \right\} \right)$$

$$\leq 5^d \max_i \text{pr} \left( \left| \langle v_i, \frac{1}{n} \sum_{i=1}^n \gamma_i \rangle \right| \geq t/2 \right).$$
where we used the upper bound $5^d$ on the covering number of the unit sphere with $\varepsilon = 0.5$ [3, Lemma 5.2].

For a fixed $\nu \in S^{d-1}$, consider

$$\Pr \left( \left\langle \nu, \frac{1}{n} \sum_{i=1}^{n} \gamma_i \right\rangle \geq t/2 \right) = \Pr \left( \left\langle \frac{1}{n} \sum_{i=1}^{n} (\nu, \gamma_i) \right\rangle \geq t/2 \right) = \Pr \left( \left\langle \frac{1}{n} \sum_{i=1}^{n} \zeta_i \right\rangle \geq t/2 \right),$$

where $\zeta_i := \langle \nu, \gamma_i \rangle$ are iid with $E(\zeta_i) = 0$ and $\|\zeta_i\|_{\psi_1} \leq L$. By Bernstein’s inequality [3, Corollary 5.17]

$$\Pr \left( \left\langle \frac{1}{n} \sum_{i=1}^{n} \zeta_i \right\rangle \geq t/2 \right) \leq 2 \exp \left[ -C n \min \left( \frac{t^2}{L^2}, \frac{t}{L} \right) \right]$$

for some constant $C > 0$. Combining the above displays gives

$$\Pr \left( \|\frac{1}{n} \sum_{i=1}^{n} \gamma_i\|_2 \geq t \right) \leq 5^d \exp \left[ -C n \min \left( \frac{t^2}{L^2}, \frac{t}{L} \right) \right].$$

Now we restate the Theorem 3 with the corresponding assumptions, and provide the full updated proof for completeness. The proof differs from the original one in the bound for $I_2$ term.

**Assumption 1** (Class probabilities). $\Pr(x_i \in C_k) = \pi_k$ for $k = 1, \ldots, K$ with $0 < \pi_{\min} \leq \pi_k \leq \pi_{\max} < 1$.

**Assumption 2** (Normality). $x_i | x_i \in C_k \sim \mathcal{N}(\mu_k, \Sigma_W)$ for all $k = 1, \ldots, K$ with $\mu = \sum_{k=1}^{K} \pi_k \mu_k = 0$.

**Assumption 3** (Sample size). $\log p = o(n)$

Throughout, we use $\sigma_j^2$ to denote the diagonal elements of within-class covariance matrix $\Sigma_W$, and define

$$\tau := \max_{j=1, \ldots, p} \sqrt{\frac{\sigma_j^2 + \max_k \mu_{kj}^2}{}}.$$

**Theorem 3.** Let $\lambda_0 = C \tau \sqrt{\frac{(K-1) \log(pn^{-1})}{n}}$ for some $\eta \in (0, 1)$ and constant $C > 0$. Under Assumptions 1–3

$$\Pr \left( \frac{1}{n} \|X^T E\|_{2, \infty} \leq \lambda_0 \right) \geq 1 - \eta.$$

**Proof of Theorem 3.** Consider

$$\frac{1}{n} \|X^T E\|_{2, \infty} = \left\| \frac{1}{n} X^T Y - \frac{1}{n} X^T X\Sigma^{-1} \Delta \right\|_{2, \infty, 2} \leq \left\| \frac{1}{n} X^T Y - \Delta \right\|_{2, \infty, 2} + \left\| \Delta - \frac{1}{n} X^T X\Sigma^{-1} \Delta \right\|_{2, \infty, 2}.$$

Consider $I_1$. From Lemma 4, with probability at least $1 - \eta$ for some constant $C > 0$

$$I_1 = \left\| \frac{1}{n} X^T Y - \Delta \right\|_{2, \infty, 2} \leq C \max_j \sigma_j \sqrt{\frac{(K-1) \log(pn^{-1})}{n}}.$$

Consider

$$I_2 = \left\| \Delta - \frac{1}{n} X^T X\Sigma^{-1} \Delta \right\|_{2, \infty, 2} = \left\| \frac{1}{n} \sum_{i=1}^{n} \left( \Delta - x_i x_i^\top \Sigma^{-1} \Delta \right) \right\|_{2, \infty, 2}$$

$$= \max_j \left\| \frac{1}{n} \sum_{i=1}^{n} \left( \mathbb{E}(x_i x_i^\top \Sigma^{-1} \Delta) - x_i x_i^\top \Sigma^{-1} \Delta \right) \right\|_{2, \infty, 2}$$

$$= \max_j \left\| \frac{1}{n} \sum_{i=1}^{n} \left( \mathbb{E}(m_{ij}) - m_{ij} \right) \right\|_{2, \infty, 2} \leq 2 \left( \sum_{i=1}^{n} \left\| \mathbb{E}(m_{ij}) - m_{ij} \right\|_{2, \infty, 2} \right).$$

(1)
where $m_{ij}^\top := \Delta^\top \Sigma_T^{-1} x_i x_{ij} \in \mathbb{R}^{K-1}$.

We will next show that $\sup_{v: \|v\|_2 = 1} \|\langle m_{ij}, v \rangle\|_{\psi_1} \leq C\tau$ for some constant $C > 0$, that is $m_{ij}$ has sub-exponential marginals. Let $v$ be a fixed vector such that $\|v\|_2 = 1$. Applying Cauchy-Schwartz inequality together with [3, Lemma 5.14] leads to

$$\|\langle m_{ij}, v \rangle\|_{\psi_1} = \|x_{ij} v^\top \Delta^\top \Sigma_T^{-1} x_i\|_{\psi_1} \leq 2\|x_{ij}\|_{\psi_2} \|v^\top \Delta^\top \Sigma_T^{-1} x_i\|_{\psi_2}. \quad (2)$$

From Lemma 3, $\|x_{ij}\|_{\psi_2} \leq C\tau$.

Now consider

$$v^\top \Delta^\top \Sigma_T^{-1} x_i = v^\top \Delta^\top \Sigma_T^{-1} \sum_{k=1}^K \mu_k \mathbb{1}\{x_i \in G_k\} + v^\top \Delta^\top \Sigma_T^{-1} \zeta_i = u_{i1} + u_{i2}.$$ 

Consider $u_{i1}$. Let $M = [\mu_1 \ldots \mu_k] \in \mathbb{R}^{p \times k}$, then

$$|u_{i1}| = |v^\top \Delta^\top \Sigma_T^{-1} \sum_{k=1}^K \mu_k \mathbb{1}\{x_i \in G_k\}| \leq \|v\|_2 \|\Delta^\top \Sigma_T^{-1}\|_2 \max_k \|\Sigma_T^{-1} \mu_k\|_2 \leq \sqrt{\|\Delta^\top \Sigma_T^{-1}\|_2 \|M^\top \Sigma_T^{-1} M\|_2},$$

where we used $v^\top v = 1$ in the last inequality. Let $\Pi = \text{diag}(\pi_1, \ldots, \pi_K)$, then $\Sigma_T = \Sigma_W + M\Pi M^\top$, and by Woodbury matrix identity

$$\|M \Sigma_T^{-1} M\|_2 = \|M^\top \Sigma_W^{-1} M\|_2 \Pi^{-1} + \|M^\top \Sigma_W^{-1} M\|_2 \leq C$$

for some constant $C > 0$, where the last inequality uses Assumption 1. Similarly,

$$\|\Delta^\top \Sigma_T^{-1} \Delta\|_2 = \|\Delta^\top \Sigma_W^{-1} \Delta + \Delta^\top \Sigma_W^{-1} \Delta^{-1}\|_2 \leq 1.$$

Therefore, $|u_{i1}|$ is sub-gaussian with constant $C_1$ independent of $v$, $\Sigma_W$ and $\mu_k$.

Consider $u_{i2}$. Since $\zeta_i \sim \mathcal{N}(0, \Sigma_W)$ and using $\Sigma_T = \Sigma_W + \Delta \Delta^\top$ [1, Proposition 2],

$$\text{Var}(u_{i2}) = v^\top \Delta^\top \Sigma_T^{-1} \Sigma_W \Sigma_T^{-1} \Delta v = v^\top \Delta^\top \Sigma_T^{-1} (\Sigma_T - \Delta \Delta^\top) \Sigma_T^{-1} \Delta v$$

$$= v^\top \Delta^\top \Sigma_T^{-1} (I - \Delta \Delta^\top) \Sigma_T^{-1} \Delta v = v^\top \Delta^\top \Sigma_T^{-1} \Delta v = v^\top \Delta^\top \Sigma_W^{-1} \Delta v$$

$$\leq \lambda_{\text{max}} \{ (I + \Delta^\top \Sigma_W^{-1} \Delta)^{-1} \Delta^\top \Sigma_W^{-1} \Delta (I + \Delta^\top \Sigma_W^{-1} \Delta)^{-1} \} \leq 1,$$

where we used $v^\top v = 1$. Therefore, $u_{i2}$ is sub-gaussian with constant one.

Combining the results of $u_{i1}$ and $u_{i2}$, it follows that $\|v^\top \Delta^\top \Sigma_T^{-1} x_i\|_{\psi_2} \leq C$ for some constant $C > 0$ independent of $v$, $\Sigma_W$ and $\mu_k$. From (2), it follows that $\sup_{v: \|v\|_2 = 1} \|\langle m_{ij}, v \rangle\|_{\psi_1} \leq C\tau$ for some constant $C > 0$, that is $m_{ij}$ has sub-exponential marginals. Furthermore,

$$\sup_{v: \|v\|_2 = 1} \|\langle m_{ij} - \mathbb{E}(m_{ij}), v \rangle\|_{\psi_1} \leq 2 \sup_{v: \|v\|_2 = 1} \|\langle m_{ij}, v \rangle\|_{\psi_1} \leq C_1\tau.$$

Applying Lemma S.1 to (1) together with union bound and Assumption 3 gives

$$I_2 = \max_j \left\| \frac{1}{n} \sum_{i=1}^n \{ \mathbb{E}(m_{ij}) - m_{ij} \} \right\|_2 \leq C_2\tau \sqrt{\frac{(K-1) \log(p)}{n}}$$

with probability at least $1 - \eta$ for some constant $C_2 > 0$.

Combining the results for $I_1$ and $I_2$ completes the proof. \qed
References

[1] Gaynanova, I., Booth, J. G. and Wells, M. T. (2016). Simultaneous sparse estimation of canonical vectors in the $p >> N$ setting. *Journal of the American Statistical Association* **111** 696–706.

[2] Obozinski, G., Wainwright, M. J. and Jordan, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *Annals of Statistics* **39** 1–47.

[3] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing* 210–268. Cambridge Univ. Press, Cambridge.