Err t o o r n pp r Boun s on th G n r l t on o to h st E pon nt ll Con v pt t on”

September 5, 2015

Abstract

We fix two typos in the statement of Theorem 4, and an error in Theorem 8. To be more clear, we rewrite the proof of the lower bound.

1 Statement of Theorem 4

$$|X_i^2| < R \to |X_i| \le R$$

$$\sqrt{2}R\sqrt{\log\frac{2t+1}{\delta^2}} \to R\sqrt{\log\frac{2t+1}{\delta^2}}$$

2 Proof of the Lower Bound

We now show that for square loss, which is a special case of exponentially concave functions, the minimax risk is $O(d/T)$. As a result, the online Newton step algorithm achieves the almost optimal excess risk bound. The proof of the lower bound is built upon the distance-based Fano inequality (Duchi and Wainwright, 2013).

Let $\mathcal{P}$ be a family of distributions on a sample space $\mathcal{X}$, and let $\theta : \mathcal{P} \mapsto \Theta$ be a function mapping $\mathcal{P}$ to some parameter space Θ. Given a set of n samples $X^n = \{X_1, \ldots, X_n\}$ drawn i.i.d. from a distribution $P \in \mathcal{P}$, let $\widehat{\theta}(X^n)$ be a measurable function of X^n, which is an estimate of the unknown quantity $\theta(P)$. Then, the minimax risk for the family $\mathcal{P}$ is given by

$$\mathfrak{M}_n\left(\theta(\mathcal{P}), \Phi \circ \rho\right) = \inf_{\widehat{\theta}} \sup_{P \in \mathbf{P}} \mathrm{E}_P\left[\Phi\left(\rho\left(\widehat{\theta}(X^n), \theta(P)\right)\right)\right]$$

where $\rho : \Theta \times \Theta \mapsto \mathbb{R}$ is a (semi)-metric on the parameter space, and $\Phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is a nondecreasing loss function. Our analysis is based on the following result from Duchi and Wainwright (2013).

Lemma 1 (Corollary 2 of Duchi and Wainwright (2013)). *s ons r s r s* $\mathcal{V}$ *n n* $\mathbf{v} \in \mathcal{V}$ *s o or* $\theta_{\mathbf{v}} \in \Theta$ *r s s n s r on* $P \in \mathcal{P}$ *G n un on* $\rho_{\mathbf{V}} : \mathcal{V} \times \mathcal{V} \mapsto \mathbb{R}$ *n s r t n s p r on un on*

$$\delta(t) := \sup\left\{\delta | \rho(\theta_{\mathbf{v}}, \theta_{\mathbf{w}}) \ge \delta \ \ or \ \ \mathbf{v}, \mathbf{w} \in \mathcal{V} \ su \ \ \rho_{\mathbf{V}}(\mathbf{v}, \mathbf{w}) > t\right\}.$$

ssu non s on s n n ur oos s or $V \in \mathcal{V}$ un or r n o n on on on s o $V = \mathbf{v}$ s p X^n o s n s r n ro s r on $P \in \mathcal{P}$ p r r $\theta_{\mathbf{v}}$ n

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi\left(\frac{\delta(t)}{2}\right)\left(1 - \frac{I(X^n; V) + \log 2}{\log|\mathcal{V}| - \log N_t^{\max}}\right), \quad \forall t$$

r $N_t^{\max} = \max_{\mathbf{v}\square\mathbf{V}}\{\mathrm{card}\{\mathbf{v}^{\square} \in \mathcal{V} | \rho_{\mathbf{V}}(\mathbf{v}, \mathbf{v}^{\square}) \leq t\}\}$

In our case, we are interested the generalization error bound $\mathcal{L}(\widehat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_{\square})$. For square loss, the stochastic optimization problem is given by

$$\min_{\mathbf{w}\square\mathbf{W}} \mathcal{L}(\mathbf{w}) = \mathrm{E}\left[(Y - X^{\square}\mathbf{w})^2\right]$$

where X is sampled from some underlying distribution P_X, and given $X = \mathbf{x}$ the response Y is sampled from an Gaussian distribution $\mathcal{N}(\mathbf{x}^{\square}\mathbf{w}_{\square}, 1)$, where $\mathbf{w}_{\square} \in \mathbb{R}^d$ is the parameter vector. Furthermore, we assume $\mathbf{w}_{\square} \in \mathcal{W}$. Then, it is easy to verify that the excess risk of a solution $\widehat{\mathbf{w}}$ is

$$\mathcal{L}(\widehat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_{\square}) = \mathrm{E}\Big[(X^{\square}\widehat{\mathbf{w}}$$

In addition, we have

$$I(V;(X,Y)^T) = TI(V;(X,Y))$$

and

$$\begin{aligned} &I(V;(X,Y)) = H(X,Y) - H(X,Y|V) \\ =&H(X) + H(Y|X) - H(X|V) - H(Y|X,V) = H(Y|X) - H(Y|X,V) \\ \le&\mathrm{E} \end{aligned}$$

References

John C. Duchi and Martin J. Wainwright. Distance-based and continuum fano inequalities with applications to statistical estimation. *ArXiv e-prints*, arXiv:1311.2669, 2013.

Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3):277–289, 2008.