Errata of "Lower and Upper Bounds on the Generalization of Stochastic Exponentially Concave Optimization"

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

$$|X_i^2| < R \rightarrow |X_i| \leq R$$
$$\sqrt{2}R\sqrt{\log\frac{2t+1}{\delta^2}} \rightarrow R\sqrt{\log\frac{2t+1}{\delta^2}}$$

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

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

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))**.** *Let's consider a discrete set $\mathcal{V}$ and each element $\mathbf{v} \in \mathcal{V}$ leads to a vector $\theta_{\mathbf{v}} \in \Theta$ that results in a distribution $P \in \mathcal{P}$. Given a function $\rho_{\mathcal{V}} : \mathcal{V} \times \mathcal{V} \mapsto \mathbb{R}$ and a scalar t, we define the separation function*

$$\delta(t) := \sup\left\{\delta | \rho(\theta_{\mathbf{v}}, \theta_{\mathbf{w}}) \geq \delta \textit{ for all } \mathbf{v}, \mathbf{w} \in \mathcal{V} \textit{ such that } \rho_{\mathcal{V}}(\mathbf{v}, \mathbf{w}) > t\right\}.$$

We assume the canonical estimation setting: nature chooses a vector $V \in \mathcal{V}$ uniformly at random, and conditioned on this choice $V = \mathbf{v}$, a sample X^n of size n is drawn i.i.d. from the distribution $P \in \mathcal{P}$ with parameter $\theta_{\mathbf{v}}$. Then, we have

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

where $N_t^{\max} = \max_{\mathbf{v} \in \mathcal{V}}\{\text{card}\{\mathbf{v}' \in \mathcal{V} | \rho_{\mathcal{V}}(\mathbf{v}, \mathbf{v}') \leq t\}\}$.

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

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

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

$$\mathcal{L}(\widehat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_*) = \mathrm{E}\left[(X^\top \widehat{\mathbf{w}} - X^\top \mathbf{w}_*)^2\right] = (\widehat{\mathbf{w}} - \mathbf{w}_*)^\top \mathrm{E}[XX^\top](\widehat{\mathbf{w}} - \mathbf{w}_*) = \|\widehat{\mathbf{w}} - \mathbf{w}_*\|_C^2$$

where we define $C = \mathrm{E}[XX^\top]$. Then, the semi-metric is naturally defined as

$$\rho(\mathbf{w}, \mathbf{w}') = \|\mathbf{w} - \mathbf{w}'\|_C$$

and $\Phi(z) = z^2$. Let $\mathcal{P}_{X,Y}$ be a family of joint distributions of X and Y. Using these notations, the minimax risk for the generalization error bound becomes

$$\mathfrak{M}_T\left(\theta(\mathcal{P}_{X,Y}), \Phi \circ \rho\right) = \inf_{\widehat{\mathbf{w}}} \sup_{P \in \mathcal{P}_{X,Y}} \mathrm{E}_P\left[\left\|\widehat{\mathbf{w}}((X,Y)^T) - \mathbf{w}(P)\right\|_C\right]$$

where $\mathbf{w}(P)$ is used to represent the parameter vector for distribution P, $(X,Y)^T = \{(X_1, Y_1), \ldots, (X_T, Y_T)\}$ are T samples drawn from P and $\widehat{\mathbf{w}}(\cdot)$ is a measurable function of $(X,Y)^T$.

To utilize Lemma 1, we introduce a discrete set $\mathcal{V} = \{\mathbf{v} \in \{-1, 0, 1\}^d \mid \|\mathbf{v}\|_0 = c_1 d\}$ for some constant $c_1 < 1$, define $\mathbf{w}_{\mathbf{v}} = \varepsilon \mathbf{v}$ for $\varepsilon > 0$, and assume $\mathbf{w}_* \in \{\varepsilon \mathbf{v} : \mathbf{v} \in \mathcal{V}\} \subseteq \mathcal{W}$. In our analysis, we set $t = c_2 d$ with $c_2 < c_1$, and define $\rho_{\mathcal{V}}(\mathbf{v}, \mathbf{w}) = \|\mathbf{v} - \mathbf{w}\|_0$. Then, we lower bound the separation function $\delta(\cdot)$ by

$$\begin{aligned}
\delta\left(c_2 d\right) =& \sup\left\{\delta \,|\varepsilon\|\mathbf{v} - \mathbf{w}\|_C \geq \delta \text{ for all } \mathbf{v}, \mathbf{w} \in \mathcal{V} \text{ such that } \|\mathbf{v} - \mathbf{w}\|_0 > c_2 d\right\} \\
=& \min\left\{\varepsilon\|\mathbf{v} - \mathbf{w}\|_C \mid \text{for all } \mathbf{v}, \mathbf{w} \in \mathcal{V} \text{ such that } \|\mathbf{v} - \mathbf{w}\|_0 > c_2 d\right\} \\
\geq& \min\left\{\varepsilon\|\mathbf{z}\|_C \,\middle|\, \text{for all } \mathbf{z} \in \{-2, -1, 0, +1, +2\}^d \text{ such that } c_2 d < \|\mathbf{z}\|_0 \leq 2c_1 d\right\} \\
\geq& \varepsilon\sqrt{c_2 d}\,\underbrace{\min\left\{\|\mathbf{z}\|_C \mid \text{for all } \|\mathbf{z}\|_2 \geq 1, \|\mathbf{z}\|_0 \leq 2c_1 d\right\}}_{:=\mu}
\end{aligned}$$

Using Lemma 1, we have

$$\mathfrak{M}_T\left(\theta(\mathcal{P}_{X,Y}), \Phi \circ \rho\right) > c_2 d \varepsilon^2 \mu^2 \left(1 - \frac{I(V; (X,Y)^T) + \log 2}{\log|\mathcal{V}| - \log N_t^{\max}}\right).$$

In addition, we have

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

and

$$\begin{aligned}
&I(V;(X,Y)) = H(X,Y) - H(X,Y|V) \\
=&H(X) + H(Y|X) - H(X|V) - H(Y|X,V) = H(Y|X) - H(Y|X,V) \\
\leq&\mathrm{E}\left[\frac{1}{|\mathcal{V}|^2}\sum_{\mathbf{w}\in\mathcal{V}}\sum_{\mathbf{v}\in\mathcal{V}} D_{kl}\left(\mathcal{N}(\varepsilon X^\top \mathbf{w},1)\|\mathcal{N}(\varepsilon X^\top \mathbf{v},1)\right)\right] \\
=&\frac{\varepsilon^2}{2}\mathrm{E}\left[(V-W)^\top XX^\top (V-W)\right] = \frac{\varepsilon^2}{2}\mathrm{E}\left[\mathrm{tr}\left(XX^\top (V-W)(V-W)^\top\right)\right] = \varepsilon^2 c_1 \,\mathrm{tr}(C)
\end{aligned}$$

where V and W are two independent random variables that are uniformly distributed on $\mathcal{V}$, which implies $\mathrm{E}[VV^\top] = \mathrm{E}[WW^\top] = c_1 I$ and $\mathrm{E}[VW^\top] = 0$. Furthermore, it is easy to verify

$$\log|\mathcal{V}| - \log N_t^{\max} \geq c_3 d$$

for some constant $c_3 > 0$ when d is large enough and c_2 is small enough. Combining the above result, we have

$$\mathfrak{M}_T\left(\theta(\mathcal{P}_{X,Y}), \Phi\circ\rho\right) \geq c_2 d\varepsilon^2\mu^2\left(1 - \frac{T\varepsilon^2 c_1\,\mathrm{tr}(C)}{c_3 d}\right) = \frac{c_2 c_3 d}{4Tc_1}\cdot\frac{d\mu^2}{\mathrm{tr}(C)}$$

where we choose $\varepsilon^2 = \frac{c_3 d}{2Tc_1\,\mathrm{tr}(C)}$.

To show the minimax risk is of $O(d/T)$, we need to construct a matrix C such that $\mathrm{tr}(C) = O(d)$ and μ^2 is a sufficiently large constant. Furthermore, to ensure the optimization problem is exponential concave instead of strongly convex, C should be singular. The existence of such a matrix is guaranteed by the following theorem.

Theorem 1. *When c_1 is smaller enough, there exists a singular matrix C such that $\mathrm{tr}(C) = d$ and $\mu^2 \geq 1/2$.*

Proof. We prove this theorem by utilizing the uniform uncertainty principle of subgaussian matrices (Mendelson et al., 2008). Let $R \in \mathbb{R}^{d\times k}$ be a random matrix with R_{ij} sampled uniformly from $\{\pm 1\}$. Following Corollary 3.3 of Mendelson et al. (2008), with a probability at least $1-\exp(-ck)$

$$\mathbf{z}^\top \frac{RR^\top}{k}\mathbf{z} \geq \frac{1}{2}\|\mathbf{z}\|_2^2 \text{ for all } \|\mathbf{z}\|_0 \leq \frac{k}{c'\log d}$$

for some constant c, $c' > 0$. By choosing $C = \frac{RR^\top}{k}$ and $k = 2c'c_1 d\log d$, with a probability at least $1-\exp(-2cc'c_1 d\log d)$, we have

$$\mu = \min\left\{\|\mathbf{z}\|_C \mid \text{for all } \|\mathbf{z}\|_2 \geq 1, \|\mathbf{z}\|_0 \leq 2c_1 d\right\} \geq \frac{\sqrt{2}}{2}.$$

Since the success probability $1-\exp(-2cc'c_1 d\log d)$ is strictly greater than 0, there must exist such a matrix C. From our construction of R, it is easy to verify $\mathrm{tr}(C) = d$ and when $c_1 < 1/(2c'\log d)$, we have $k < d$ and thus C is singular. $\square$

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.