Supplementary Material — Proof of the Theorems

Chi Jin
Key Laboratory of Machine Perception, MOE
School of Physics
Peking University
chijin06@gmail.com

Liwei Wang
Key Laboratory of Machine Perception, MOE
School of EECS
Peking University
wanglw@cis.pku.edu.cn

Here we give proof sketches of the theorems and propositions as well as some technical discussions.

Proof of Theorem 3.1. The proof is an application of the PAC-Bayes theorem (Theorem 2.1) and a refinement of the proof of Theorem 2.2.

First observe that when considering distributions of homogeneous linear classifiers $c_{\mathbf{w}}$ in $\mathbb{R}^d$, we only need to restrict ourselves in distributions of $\mathbf{w}$ on the $(d-1)$-dimensional unit sphere S^{d-1}. For any probability distribution π of vectors in $\mathbb{R}^d$, let π_p denote the corresponding probability distribution on S^{d-1} by projecting π from $\mathbb{R}^d$ to S^{d-1}.

Choose the prior distribution P of classifiers $c_{\mathbf{w}} = \mathrm{sgn}(< \mathbf{w}, \cdot >)$ corresponding to $\mathbf{w} \sim \mathcal{N}_p(0, \mathbf{I})$, i.e., the uniform distribution on S^{d-1}. Let the posterior distribution $Q(\mu, \hat{\mathbf{w}})$ be defined as in Theorem 3.1. It is obvious that $Q(\mu, \hat{\mathbf{w}})$ of $c_{\mathbf{w}}$ corresponds to the distribution of $\mathbf{w} \sim \mathcal{N}_p(\mu\hat{\mathbf{w}}, I)$. Thus to finish the proof we only need to show

$$\mathrm{KL}(\mathcal{N}_p(\mu\hat{\mathbf{w}}, \mathbf{I})||\mathcal{N}_p(0, \mathbf{I})) \leq \frac{d}{2}\ln(1 + \frac{\mu^2}{d}). \tag{1}$$

Observe that for all $\sigma > 0$, we have

$$\begin{aligned} \mathrm{KL}(\mathcal{N}_p(\mu\hat{\mathbf{w}}, \mathbf{I})||\mathcal{N}_p(0, \mathbf{I})) &= \mathrm{KL}(\mathcal{N}_p(\mu\hat{\mathbf{w}}, \mathbf{I})||\mathcal{N}_p(0, \sigma^2\mathbf{I})) \\ &\leq \mathrm{KL}(\mathcal{N}(\mu\hat{\mathbf{w}}, \mathbf{I})||\mathcal{N}(0, \sigma^2\mathbf{I})). \end{aligned}$$

The last inequality holds according to the chain rule of the KL divergence [1]. Taking $\sigma^2 = 1 + \frac{\mu^2}{d}$ completes the proof. □

It is worth pointing out that (1) is almost a tight upper bound. Thus the dimensionality d involved is intrinsic. Note that $\frac{d}{2}\ln(1 + \frac{\mu^2}{d}) \sim d\ln\mu$ as $\mu \to \infty$. In fact we can show $\mathrm{KL}(\mathcal{N}_p(\mu\hat{\mathbf{w}}, \mathbf{I})||\mathcal{N}_p(0, \mathbf{I})) \sim (d-1)\ln\mu$.

To see this, let $P = \mathcal{N}_p(0, \mathbf{I}))$, and $Q = \mathcal{N}_p(\mu\hat{\mathbf{w}}, \mathbf{I})$. Since P is the uniform distribution on S^{d-1}, we have $\mathrm{KL}(Q||P) = \ln\frac{2\pi^{d/2}}{\Gamma(d/2)} - h(Q)$, where $h(Q)$ is the differential entropy of Q. So we only need to show $-h(Q) \sim (d-1)\ln\mu$. Let $\hat{\mathbf{v}} \in S^{d-1}$, and let $\cos\alpha =< \hat{\mathbf{v}}, \hat{\mathbf{w}} >$. Let $q(\hat{\mathbf{v}})$ be the density of Q. We have

$$\begin{aligned} q(\hat{\mathbf{v}}) &= \int_0^\infty \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\left(r^2 + \mu^2 - 2r\mu\cos\alpha\right)\right) r^{d-1} dr \\ &= \frac{\exp(-\frac{\mu^2\sin^2\alpha}{2})}{(2\pi)^{d/2}} \int_0^\infty \exp\left(\frac{1}{2}(r - \mu\cos\alpha)^2\right) r^{d-1} dr. \end{aligned}$$

Let

$$I_n(t) = \frac{1}{\sqrt{2\pi}} \int_0^\infty \exp\left(\frac{1}{2}(r-t)^2\right) r^{d-1} dr.$$

Integration by parts yields a recursive formula

$$I_n(t) = tI_{n-1}(t) + (n-2)I_{n-2}(t).$$

Also we have $I_1(t) = \Phi(t)$, and $I_2(t) = \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} + t\Phi(t)$. Some calculation yields

$$I_d(t) = \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} f_d(t) + \Phi(t) g_d(t),$$

where $f_d(t)$ and $g_d(t)$ are polynomials of t with $(d-2)$ and $(d-1)$ degree both with the leading coefficient being 1. Thus

$$q(\hat{\mathbf{v}}) = \frac{e^{-\frac{\mu^2}{2}}}{(2\pi)^{d/2}} f_d(\mu\cos\alpha) + \frac{\exp(-\frac{\mu^2\sin^2\alpha}{2})}{(2\pi)^{\frac{d-1}{2}}} \Phi(\mu\cos\alpha) g_d(\mu\cos\alpha).$$

When μ is sufficiently large, the first term in above formula is clearly negligible. For the second term, we only need to consider $\alpha \leq \mu^{-1/2}$, since otherwise the term is negligible. Thus

$$\begin{aligned}
\int_{S^{d-1}} q(\hat{\mathbf{v}}) \ln q(\hat{\mathbf{v}}) d\Omega &\sim \int_{S^{d-1}} q(\hat{\mathbf{v}}) \ln \left(\frac{\exp(-\frac{\mu^2\sin^2\alpha}{2})}{(2\pi)^{\frac{d-1}{2}}} \Phi(\mu\cos\alpha)(\mu\cos\alpha)^{d-1} \right) d\Omega \\
&\sim \ln \frac{\mu^{d-1}}{(2\pi)^{\frac{d-1}{2}}} + \int_{S^{d-1}, \alpha \leq \mu^{-1/2}} \frac{\exp(-\frac{\mu^2\alpha^2}{2})}{(2\pi)^{\frac{d-1}{2}}} \mu^{d-1} \left(-\frac{\mu^2\alpha^2}{2} \right) d\Omega.
\end{aligned}$$

Some calculations show that

$$-\frac{d-1}{2} \leq \int_{S^{d-1}, \alpha \leq \mu^{-1/2}} \frac{\exp(-\frac{\mu^2\alpha^2}{2})}{(2\pi)^{\frac{d-1}{2}}} \mu^{d-1} \left(-\frac{\mu^2\alpha^2}{2} \right) d\Omega \leq 0.$$

We obtain the results.

Proof of Proposition 3.2. Obvious since $\frac{d}{2}\ln\left(1+\frac{\mu^2}{d}\right) < \frac{\mu^2}{2}$ for any $d < \infty$ and $\mu > 0$; and as $d \to \infty$, $\frac{d}{2}\ln\left(1+\frac{\mu^2}{d}\right) \to \frac{\mu^2}{2}$. □

Proof of Corollary 3.3. We will show that for every $\epsilon > 0$ and every $\delta \geq 2e^{-2n\epsilon^2}$, with probability $1-\delta$

$$er_{\mathcal{D}}(c_{\mathbf{w}}) \leq er_{\mathcal{S}}(c_{\mathbf{w}}) + \sqrt{\frac{d\ln\left(1+\left(\frac{2n}{d}\right)\right) + \frac{1}{2}\ln\frac{2(n+1)}{\delta}}{n}} + 4\epsilon \tag{2}$$

holds simultaneously for all homogeneous linear classifiers $c_{\mathbf{w}}$ with $\mathbf{w} \in \mathbb{R}^d$ satisfying

$$P_{\mathcal{D}}\left(\left| y \cdot \frac{<\mathbf{w}, \mathbf{x}>}{\|\mathbf{w}\|\|\mathbf{x}\|} \right| \leq \frac{td^{3/2}}{n^2} \right) \leq \epsilon, \tag{3}$$

where $t = \frac{1}{4}\overline{\Phi}^{-1}(\epsilon)$. Setting $\epsilon = \frac{1}{4}\left(\frac{d+\ln n}{n}\right)^{1/2}$ yields the result (assuming $n > 5$).

Set $\mu = \frac{4n^2}{d^{3/2}}$ in Theorem 3.1. Also let $Q(\mu, \hat{\mathbf{w}})$ be defined as in Theorem 3.1. By the simple fact that

$$kl(er_{\mathcal{S}}(Q)||er_{\mathcal{D}}(Q)) \geq 2\left(er_{\mathcal{S}}(Q) - er_{\mathcal{D}}(Q)\right)^2,$$

we obtain from Theorem 3.1 that with probability $1 - \frac{\delta}{2}$ for all $\hat{\mathbf{w}} \in \mathbb{R}^d$ with $\|\hat{\mathbf{w}}\| = 1$

$$er_{\mathcal{D}}(Q(\mu, \hat{\mathbf{w}})) \leq er_{\mathcal{S}}(Q(\mu, \hat{\mathbf{w}})) + \sqrt{\frac{d\left(\ln(1+\frac{2n}{d})\right) + \frac{1}{2}\ln\frac{2(n+1)}{\delta}}{n}}. \tag{4}$$

Let $\eta = \Phi^{-1}(\epsilon)$ and $z = \mu y \frac{<\hat{\mathbf{w}}, \mathbf{x}>}{\|\mathbf{x}\|}$, we have

$$\begin{aligned} er_{\mathcal{D}}(Q(\mu, \hat{\mathbf{w}})) &= E_{\mathcal{D}} \overline{\Phi}(z) \\ &= P_{\mathcal{D}}(z \leq \eta) \cdot E_{\mathcal{D}} \left(\overline{\Phi}(z) | z \leq \eta \right) + \\ &\quad P_{\mathcal{D}}(\eta < z \leq 0) \cdot E_{\mathcal{D}} \left(\overline{\Phi}(z) | \eta < z \leq 0 \right) + \\ &\quad P_{\mathcal{D}}(z > 0) \cdot E_{\mathcal{D}} \left(\overline{\Phi}(z) | z > 0 \right) \\ &\geq (er_{\mathcal{D}}(c_{\hat{\mathbf{w}}}) - \epsilon) \cdot (1 - \epsilon) \\ &\geq er_{\mathcal{D}}(c_{\hat{\mathbf{w}}}) - 2\epsilon. \end{aligned} \tag{5}$$

By the assumption of the theorem and the Chernoff bound, it is easy to see that with probability $1 - \frac{\delta}{2}$, where $\delta \geq 2e^{-2n\epsilon^2}$,

$$P_{\mathcal{S}} \left(\left| y \cdot \frac{< \hat{\mathbf{w}}, \mathbf{x} >}{\|\mathbf{x}\|} \right| \leq \frac{\overline{\Phi}^{-1}(\epsilon) d^{3/2}}{4n^2} \right) \leq 2\epsilon.$$

Similarly we can also show that

$$er_{\mathcal{S}}(Q(\mu, \hat{\mathbf{w}})) \leq er_{\mathcal{S}}(c_{\hat{\mathbf{w}}}) + 2\epsilon. \tag{6}$$

Combining (4), (5) and (6) with the union bound, the theorem follows. □

Proof of Proposition 3.4. First it is easy to check that $P_{\mathbf{w} \sim Q} \left(y \frac{<\mathbf{w}, \mathbf{x}>}{\|\mathbf{x}\|} \leq \theta \right) = \overline{\Phi} \left(\mu y \frac{<\hat{\mathbf{w}}, \mathbf{x}>}{\|\mathbf{x}\|} - \theta \right)$, where Q is the abbreviation of $Q(\mu, \hat{\mathbf{w}})$ defined in Theorem 3.1. Also observe that for every θ

$$I[t \leq 0] \leq \frac{\overline{\Phi}(t - \theta)}{\overline{\Phi}(-\theta)}.$$

Thus we have

$$\begin{aligned} er_{\mathcal{D}}(c_{\hat{\mathbf{w}}}) &= E_{\mathcal{D}} I \left[y \frac{< \hat{\mathbf{w}}, \mathbf{x} >}{\|\mathbf{x}\|} \leq 0 \right] \\ &\leq E_{\mathcal{D}} \frac{\overline{\Phi} \left(\mu y \frac{<\hat{\mathbf{w}}, \mathbf{x}>}{\|\mathbf{x}\|} - \theta \right)}{\overline{\Phi}(-\theta)} \\ &= \frac{1}{\Phi(\theta)} E_{\mathbf{w} \sim Q} P_{\mathcal{D}} \left(y \frac{< \mathbf{w}, \mathbf{x} >}{\|\mathbf{x}\|} \leq \theta \right) \\ &= \frac{er_{\mathcal{D}, \theta}(Q(\mu, \hat{\mathbf{w}}))}{\Phi(\theta)} \end{aligned}$$

□

Proof of Proposition 3.5. Let $\epsilon = er_{\mathcal{D}, \theta}(Q)$. We only need to show

$$\epsilon + \overline{\Phi}(\theta) - \frac{\epsilon}{\Phi(\theta)} \geq 0. \tag{7}$$

Note that $1 - \Phi(\theta) = \overline{\Phi}(\theta)$. The LHS of (7) equals to $\overline{\Phi}(\theta) \left[1 - \frac{\epsilon}{\Phi(\theta)} \right]$.

Finally, observe that if $\epsilon + \overline{\Phi}(\theta) < 1$, then $\epsilon < \Phi(\theta)$. The proposition follows. □

Proof of Lemma 3.6. Due to the symmetry of Gaussian distribution $\mathcal{N}(\mu \hat{\mathbf{w}}, I)$, simple analysis shows that $P_{\mathbf{w} \sim \mathcal{N}(\mu \hat{\mathbf{w}}, I)} \left(y \frac{<\mathbf{w}, \mathbf{x}>}{\|\mathbf{w}\| \|\mathbf{x}\|} \leq \theta \right)$ is only a function of $\frac{<\hat{\mathbf{w}}, y\mathbf{x}>}{\|\mathbf{x}\|}$, θ, and μ. We denote this function as $F(\mu, \frac{<\hat{\mathbf{w}}, y\mathbf{x}>}{\|\mathbf{x}\|}, \theta)$.

A slight modification of the proof of Proposition 3.4 yields

$$er_{\mathcal{D}}(c_{\hat{\mathbf{w}}}) \leq \frac{er^{\mathbf{N}}_{\mathcal{D},\theta}\left(Q(\mu, \hat{\mathbf{w}})\right)}{F(\mu, 0, \theta)}. \tag{8}$$

Let $\hat{\mathbf{u}}, \hat{\mathbf{v}}$ to be two unit vectors satisfying $< \hat{\mathbf{w}}, \hat{\mathbf{u}} >= 0$ and $\hat{\mathbf{v}} = \sqrt{1-\theta^2}\hat{\mathbf{u}} - \theta\hat{\mathbf{w}}$. It's not difficult to show that for an arbitrary vector $\mathbf{w}$:

$$< \mathbf{w}, \hat{\mathbf{v}} >\leq 0 \Rightarrow \frac{< \mathbf{w}, \hat{\mathbf{u}} >}{\|\mathbf{w}\|} \leq \theta$$

Thus we have:

$$\begin{aligned} F(\mu, 0, \theta) &= P_{\mathbf{w}\sim\mathcal{N}(\mu\hat{\mathbf{w}}, I)}\left(\frac{< \mathbf{w}, \hat{\mathbf{u}} >}{\|\mathbf{w}\|} \leq \theta\right) \\ &\geq P_{\mathbf{w}\sim\mathcal{N}(\mu\hat{\mathbf{w}}, I)}\left(< \mathbf{w}, \hat{\mathbf{v}} >\leq 0\right) \\ &= \overline{\Phi}\left(-\mu\theta\right) = \Phi\left(\mu\theta\right) \end{aligned} \tag{9}$$

Combining (8) and (9) finishes the proof. □

Proof of Proposition 3.7. Immediate. □

References

[1] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons, Inc., New York, USA, 1991.