Supplementary Materials to *"Semi-Supervised Learning: the Case When Unlabeled Data is Equally Useful"*

A Derivation in the proof of Lemma 6

Now we prove the claim that $Q(y^*|z^n, x')$ is lower bounded by some constant independent of n. In the following, we will explicitly write y^* as $y^*_{\theta_0}$ where the subscript denotes that the optimal estimator is obtained when the true parameter is θ_0. To show this, we first observe that the optimal (Bayes) estimator $y^*_{\theta_0}$ with the $0-1$ loss is given by

$$y^*_{\theta_0} = \text{argmax}_y p_{\theta_0}(y|x').$$

Now we examine the term

$$\begin{aligned}
Q(y^*_{\theta_0}|z^n, x') &= \frac{Q(y^*_{\theta_0}, z^n, x')}{Q(z^n, x')d\theta} = \frac{\int p_\theta(y^*_{\theta_0}, z^n, x')q(\theta)}{\int p(z^n, x')q(\theta)d\theta} \\
&= \frac{\int p_\theta(y^*_{\theta_0}|z^n, x')p_\theta(z^n, x')d\theta}{\int p_\theta(z^n, x')d\theta} \\
&= \frac{\int p_\theta(y^*_{\theta_0}|x')p_\theta(z^n, x')d\theta}{\int p_\theta(z^n, x')d\theta}
\end{aligned}$$

as $y^*_{\theta_0}$ does not depend on z^n. Let y^*_θ denote the optimal estimator with respect to the parameter θ. Also notice that we can always write

$$p_\theta(y^*_{\theta_0}|x') = Kp_\theta(y^*_\theta|x')$$

for some $K > 0$ due to the assumption that $p_\theta(y^*_{\theta_0}|x') \neq 0$. Moreover, it holds that $p_\theta(y^*_\theta|x') > C$ for some nonnegative constant C because y^*_θ is one maximizer of $p_\theta(y|x')$. We can continue as

$$\begin{aligned}
Q(y^*_{\theta_0}|z^n, x') &= \frac{K\int p_\theta(y^*_\theta|x')p_\theta(z^n, x')d\theta}{\int p_\theta(z^n, x')d\theta} \\
&> \frac{KC\int p_\theta(z^n, x')d\theta}{\int p_\theta(z^n, x')d\theta} = KC
\end{aligned}$$

which proves the claimed result.

B Proof of Lemma 4

Our proof will largely follow the strategy used in [Clarke and Barron, 1990], [Clarke, 1989]. The main idea is to approximate the density ratio $p(Z^n, \tilde{X}^m|\theta)/Q(Z^n, \tilde{X}^m)$ around θ using Laplace's method, and control the decay rate of the remaining terms. The definitions of various sets in the proof differ slightly from [Clarke and Barron, 1990] to suit our purpose. As our proof is long but follows closely to the above two references, we will highlight the different parts and refer to the original proof for repetitive steps.

We use $p(Z^n, \tilde{X}^m|\theta)$ to denote the likelihood defined as

$$p(Z^n, \tilde{X}^m|\theta) := \prod_{i=1}^n p_\theta(X_i, Y_i) \prod_{j=1}^m p_\theta(\tilde{X}_j).$$

Define the (unnormalized) score function as

$$\begin{aligned}
l_{XY}(\theta) &:= \nabla \log p(Z^n|\theta) \\
l_X(\theta) &:= \nabla \log p(\tilde{X}^n|\theta),
\end{aligned}$$

and the (unnormalized) empirical information matrix

$$\begin{aligned}
I^*_{XY}(\theta) &:= -[\partial^2(\log p(Z^n|\theta))/\partial\theta_j\partial\theta_k]_{j,k=1,\dots,d} \\
I^*_X(\theta) &:= -[\partial^2(\log p(\tilde{X}^m|\theta))/\partial\theta_j\partial\theta_k]_{j,k=1,\dots,d}
\end{aligned}$$

Let θ_0 denote the true parameter that generate the data $Z^n, \tilde{X}^m$. Define $N_\delta = \{\theta : \|\theta - \theta_0\| \leq \delta\}$. For convenience, the norm is defined as

$$\|\xi\|^2 = \xi^T(I_{XY}(\theta_0) + I_X(\theta_0))\xi.$$

For $0 < \epsilon < 1$ and $\delta > 0$, define

$$\begin{aligned}
A(\delta, \epsilon) := \Bigg\{ &\int_{N^c_\delta} p(Z^n, \tilde{X}^m|\theta)q(\theta)d\theta \\
&\leq \epsilon \int_{N_\delta} p(Z^n, \tilde{X}^m|\theta)q(\theta)d\theta \Bigg\}.
\end{aligned}$$

For convenience, we also define

$$I_{n,m} := nI_{XY}(\theta_0) + mI_X(\theta_0)$$

and

$$D(\theta_0) := (l_{XY}(\theta_0) + l_X(\theta_0))^T I^{-1}_{n,m}(l_{XY}(\theta_0) + l_X(\theta_0)).$$

Notice that

$$\begin{aligned}
\mathbb{E}\{D(\theta_0)\} = \mathbb{E}\{&\text{Tr}((I^{-1}_{n,m})(l_{XY}(\theta_0) + l_X(\theta_0))^T \\
&(l_{XY}(\theta_0) + l_X(\theta_0))\} \\
&= \text{Tr}(I^{-1}_{n,m}(nI_{XY}(\theta) + mI_{XY}(\theta))) = d
\end{aligned}$$

Lastly, define

$$\begin{aligned}
B(\delta, \epsilon) := \{&(1-\epsilon)(\theta-\theta_0)^T I_{n,m}(\theta-\theta_0) \\
&\leq (\theta-\theta_0)^T(I^*_{XY}(\theta') + I^*_X(\theta'))(\theta-\theta_0) \\
&\leq (1+\epsilon)(\theta-\theta_0)^T I_{n,m}(\theta-\theta_0) \\
&\text{for all } \theta, \theta' \in N_\delta\} \\
C(\delta) :=\{&D(\theta_0) \leq \min\{n, m\}\delta^2\}
\end{aligned}$$

and

$$\rho(\delta, \theta_0) := \sup_{\theta \in N_\delta} |\log \frac{q(\theta)}{q(\theta_0)}|.$$

In the sequel, we assume that both m, n increase in a way that either $m = \alpha n$ for some $\alpha > 0$, or $m = n^{1+\gamma}$ for some $\gamma > 0$. Following [Clarke and Barron, 1990], we have following upper and lower bounds on the density ratio.

Lemma 6 *Assume that the Condition 1 is satisfied, and $q(\theta)$ continuous at θ_0. Then on the set $A \cap B$, we have*

$$\frac{Q(Z^n, \tilde{X}^m)}{p(Z^n, \tilde{X}^m|\theta_0)} \leq (1+\epsilon)q(\theta_0)e^{\rho(\delta,\theta_0)}(2\pi)^{d/2} \cdot e^{1/(2(1-\epsilon))D(\theta_0)}|(1-\epsilon)I_{n,m}|^{-1/2}$$

On the set $B \cap C$, we have

$$\frac{Q(Z^n, \tilde{X}^m)}{p(Z^n, \tilde{X}^m|\theta_0)} \geq q(\theta_0)e^{-\rho(\delta,\theta_0)}(2\pi)^{d/2}e^{1/(2(1+\epsilon))D(\theta_0)} \cdot (1-2^{d/2}e^{-\epsilon^2\min\{n,m\}\delta^2/8})|(1+\epsilon)I_{n,m}|^{-1/2}$$

Proof: The proof of this lemma is very similar to the proof of [Clarke and Barron, 1990, Lemma 4.1], except for minor modifications to account for the different definition of the set $B(\delta, \epsilon)$ and $C(\delta)$. The main idea is to use Laplace's method to approximate the integration in $Q(Z^n, \tilde{X}^m)$ around the true parameter θ_0. We omit the details. □

Recall that $D(p(X^n, Y^n, \tilde{X}^m|\theta)||Q(X^n, Y^n, \tilde{X}^m)) = \mathbb{E}\left\{\log \frac{p(X^n,Y^n,\tilde{X}^m|\theta)}{Q(X^n,Y^n,\tilde{X}^m)}\right\}$. Given the above bounds, we now can define the reminder term Re as follows.

$$Re := \log \frac{p(Z^n, \tilde{X}^m|\theta_0)}{Q(Z^n, \tilde{X}^m)} - \left(\frac{d}{2}\log\frac{1}{2\pi} + \log\frac{1}{q(\theta_0)} + \frac{1}{2}\log|I_{n,m}| - D(\theta_0)/2\right).$$

It is clear that Lemma 4 is established if we show the expectation of Re converges to 0 with an appropriate rate, which we will do next.

Equipped with Lemma 6, and using the same argument as in [Clarke and Barron, 1990, pp.464] (see also [Clarke and Barron, 1994]), we can show the following upper bound and lower bounds on $\mathbb{E}\{Re\}$:

$$\begin{aligned}
&\mathbb{E}\{Re\} \\
&\geq -\log(1+\epsilon) - \rho(\delta,\theta_0) - \frac{\epsilon}{2(1-\epsilon)}d + \frac{d}{2}\log(1-\epsilon) \\
&+ \mathbb{P}\{(A\cap B)^c\}(\log\mathbb{P}\{(A\cap B)^c\} + \frac{d}{2}\log\frac{1}{2\pi}) \\
&- \mathbb{P}\{(A\cap B)^c\}\log\frac{\sqrt{|I_{n,m}|}}{q(\theta_0)} \qquad (9)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\{Re\} \leq \rho(\delta,\theta_0) + \frac{\epsilon}{2(1+\epsilon)}d + \frac{d}{2}\log(1+\epsilon) \\
&- \log(1-2^{d/2}e^{-\epsilon^2\min\{m,n\}\delta^2/8}) + \mathbb{E}\left\{D(\theta_0)\mathbf{1}_{(B\cap C)^c}\right\} \\
&+ \mathbb{P}\{(B\cap C)^c\}\left(\frac{d}{2}\log\frac{1}{2\pi} + |\log\int_{N_\delta} q(\theta)d\theta| \qquad (10)\right. \\
&\left.+ \log\frac{\sqrt{|I_{n,m}|}}{q(\theta_0)}\right) \\
&+ n\mathbb{P}\{(B\cap C)^c\}\mathbb{E}\{f(Z)\} + m\mathbb{P}\{(B\cap C)^c\}\mathbb{E}\left\{f(\tilde{X})\right\} \\
&+ (n\mathbb{P}\{(B\cap C)^c\})^{\frac{1}{2}}\mathbb{E}\{f^2(Z)\}^{\frac{1}{2}} \\
&+ (m\mathbb{P}\{(B\cap C)^c\})^{\frac{1}{2}}\mathbb{E}\left\{f^2(\tilde{X})\right\}^{\frac{1}{2}} \qquad (11)
\end{aligned}$$

where $f(\cdot) := \sup_{\theta',\theta''\in N_\delta}(\theta'-\theta_0)^T\nabla\log p(\cdot|\theta'')$

The following lemmas (Lemma 7, 8, 9) show that the probability that $(Z^n, \tilde{X}^m)$ belongs to each of the set A^c, B^c and C^c is smaller than $O(e^{-\min\{m,n\}\rho})$ for some $\rho > 0$. We also show in Lemma 7 and 8 that we can take $\epsilon = e^{-\max\{m,n\}r}$ for some $r > 0$. Moreover, as we can choose the prior distribution $q(\theta)$ to our liking (cf. Lemma 1), we will choose $q(\theta)$ to be the uniform distribution over Λ, so that $\rho(\delta, \theta_0) = 0$. So the first four terms in the lower bound (9) scales as $O(\epsilon) = O(e^{-\min\{m,n\}}) = o(1/\max\{m,n\})$ for large m and n. Notice that $|I_{n,m}|$ scales as $\log\max\{m,n\}$, so the last two terms in (9) scale as $O(e^{-\min\{m,n\}}\max\{m,n\})$ which is also $o(1/\max\{m,n\})$ for large m and n.

For the upper bound in (11), by choosing $q(\theta)$ to be the uniform distribution, the first four terms scales as $O(e^{-\min\{m,n\}})$ as in the lower bound. Using the same argument as in [Clarke and Barron, 1994, pp. 51], $\mathbb{E}\left\{D(\theta_0)\mathbf{1}_{(B\cap C)^c}\right\}$ can be upper bounded using Hölder's inequality by $O(\mathbb{P}\{(B\cap C)^c\}^{s/(1+s)})$ for some $s > 0$. Furthermore, we can make $|f|$ a very small constant by choosing δ sufficiently small. So it is easy to see that the rest terms in (11) are of the order $O(e^{-\min\{m,n\}s/(1+s)}) + O(\sqrt{e^{-\min\{m,n\}}\max\{m,n\}})$ which also scales as $o(1/\max\{m,n\})$ for large m and n. In the following, we conclude the proof by showing that the probability of the set A^c, B^c, C^c is upper bounded by an exponentially fast decaying term.

Lemma 7 (Probability of A^c) *Assume Condition 2 holds so that for all $\theta \in N_\delta$, the (normalized) Renyi divergence of order $1+\lambda$*

$$\int p(x|\theta_0)^{1+\lambda}p(x|\theta)^{-\lambda}dx, \int p(x,y|\theta_0)^{1+\lambda}p(x,y|\theta)^{-\lambda}dxdy$$

are bounded for some $\lambda > 0$ small enough. Let $n' = \max\{n,m\}$. Then for δ sufficiently small, there is an

$r > 0$ and $\rho > 0$ so that

$$\mathbb{P}\left\{(Z^n, \tilde{X}^m) \in A^c(\delta, e^{-n'r})\right\} = O(e^{-\min\{m,n\}\rho})$$

Proof: For simplicity, we use T to denote $(Z^n, \tilde{X}^m)$ in the proof. For any given $r' > 0$, define the event

$$U = \left\{e^{-n'r'}p(T|\theta_0) < \int_{N_\delta} q(\theta)p(T|\theta)d\theta\right\}.$$

We have

$$\begin{aligned}
&\mathbb{P}\left\{A^c(\delta, e^{-n'r})\right\} \\
&= \mathbb{P}\left\{\int_{N_\delta} p(T|\theta)q(\theta)d\theta < e^{n'r}\int_{N_\delta^c} p(T|\theta)q(\theta)d\theta\right\} \\
&\leq \mathbb{P}\Bigg\{U \cap \Big\{\int_{N_\delta} p(T|\theta)q(\theta)d\theta \\
&< e^{n'r}\int_{N_\delta^c} p(T|\theta)q(\theta)d\theta\Big\}\Bigg\} + \mathbb{P}\left\{U^c\right\} \\
&\leq \mathbb{P}\left\{p(T|\theta_0) < e^{n'(r+r')}\int_{N^c} q(\theta)p(T|\theta)d\theta\right\} \\
&+ \mathbb{P}\left\{e^{nr'}\int_{N_\delta} p(T|\theta)q(\theta)d\theta < p(T|\theta_0)\right\} \qquad (12)
\end{aligned}$$

by intersecting with U and U^c.

We first study the second term in (12) and show that it converges to zero exponentially. We follow the argument used in [Clarke, 1999]. Define $Q(T|N_\delta) = \int_{N_\delta} p(X|\theta)q(\theta|N_\delta)d\theta$ where $q(\theta|N_\delta) = q(\theta)/(\int_{N_\delta} q(\theta)d\theta)$. Define $\tilde{r} = r' - \frac{1}{n}\log\int_{N_\delta} q(\theta)d\theta$. Applying Jensen's inequality, we can upper bound the second term in (12) as

$$\begin{aligned}
&\mathbb{P}\left\{\log\frac{p(T|\theta_0)}{Q(T|N_\delta)} > n'\tilde{r}\right\} \\
&\leq \mathbb{P}\left\{\log p(T|\theta_0) - \int_{N_\delta}\log p(T|\theta)q(\theta|N_\delta)d\theta > n'\tilde{r}\right\} \\
&= \mathbb{P}\Bigg\{\int_{N_\delta}\log\frac{p(Z^n|\theta_0)}{p(Z^n|\theta_0)}q(\theta|N_\delta)d\theta \\
&+ \int_{N_\delta}\log\frac{p(\tilde{X}^m|\theta_0)}{p(\tilde{X}^m|\theta)}q(\theta|N_\delta)d\theta > n'\tilde{r}\Bigg\} \\
&= \mathbb{P}\left\{\sum_{i=1}^n g(Z_i) + \sum_{j=1}^m g(X_j) > n'\tilde{r}\right\} \\
&\leq \mathbb{P}\left\{\frac{1}{n'}\sum_{i=1}^n g(Z_i) > \tilde{r}/2\right\} + \mathbb{P}\left\{\frac{1}{n'}\sum_{i=1}^m g(X_j) > \tilde{r}/2\right\} \\
&\leq \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n g(Z_i) > \tilde{r}/2\right\} + \mathbb{P}\left\{\frac{1}{m}\sum_{i=1}^m g(X_j) > \tilde{r}/2\right\}
\end{aligned}$$

where we define

$$g(\cdot) := \int_{N_\delta}\log\frac{p(\cdot|\theta_0)}{p(\cdot|\theta)}q(\theta|N_\delta)d\theta.$$

Notice that the expectation of g is $\int_{N_\delta} D(p_{\theta_0}||p_\theta)w(\theta|N_\delta)d\theta$ is less than any fixed $\tilde{r}/2$ for δ sufficiently small. If it holds that for any θ in N_δ, moment generating functions $\int p(x|\theta_0)e^{\lambda g(x)}dx$ and $\int p(x,y|\theta_0)e^{\lambda g(x,y)}dxdy$ exist for some $\lambda \in I$ where I is an interval including 0, then using the standard Cramér-Chernoff method (see, e. g. [Boucheron et al., 2013]), both probabilities in the last inequality are upper bounded by terms in the order of $O(e^{-\rho n})$ and $O(e^{-\rho m})$ for some $\rho > 0$, respectively.

It can be shown that the existence of the moment generating function is guaranteed if Condition 2 holds. Indeed, applying Jensen's inequality gives

$$e^{\lambda g(x)} \leq \int\left(\frac{p(x|\theta_0)}{p(x|\theta)}\right)^\lambda q(\theta|N_\delta)d\theta.$$

Hence the moment generating function is bounded by

$$\int p(x|\theta_0)\left(\frac{p(x|\theta_0)}{p(x|\theta)}\right)^\lambda q(\theta|N_\delta)d\theta dx$$

which is upper bounded by the (unnormalized) Renyi divergence.

The first term in (12) can also be shown to be of the order of $O(e^{-\min\{n,m\}r''})$ for some $r'' > 0$. The proof is essentially the same as in [Clarke and Barron, 1990, Prop. 6.3] (see also [Clarke and Barron, 1994, pp. 49-50]), and is omitted here. □

Lemma 8 (Probability of B^c) *Assume that Condition 3 holds. Then for δ sufficiently small, there is a $\rho > 0$ such that*

$$\mathbb{P}\left\{(Z^n, \tilde{X}^m) \in B^c(\delta, \epsilon)\right\} = O(e^{-\min\{m,n\}\rho})$$

Proof: Using the same argument as in [Clarke, 1989, pp. 42], the set $B(\delta, \epsilon)$ can be rewritten as

$$\left\{\left|\frac{\xi^T I_{m,n}^{-1/2}(I_{XY}^*(\theta') + I_X^*(\theta') - I_{n,m})I_{m,n}^{-1/2}\xi}{\xi^T\xi}\right| < \epsilon\right\},$$

where $\xi = I_{m,n}^{1/2}(\theta - \theta_0)$, and we can upper bound the probability of B^c by

$$
\begin{aligned}
&\mathbb{P}\left\{(Z^n, \tilde{X}^m) \in B^c(\delta, \epsilon)\right\} \\
&\le \sum_{j,k} \Bigg(\mathbb{P}\left\{ \sup_{|\theta_0 - \theta| < \delta} |\frac{1}{n}\sum_{i=1}^{n} I^*_{j,k}(\theta, i) - \frac{1}{n}\sum_{i=1}^{n} I^*_{j,k}(\theta_0, i)| > \frac{\epsilon}{4d} \right\} \\
&+ \mathbb{P}\left\{ |\frac{1}{n}\sum_{i=1}^{n} I^*_{j,k}(\theta_0, i) - I_{j,k}(\theta_0, i)| > \frac{\epsilon}{4d} \right\} \\
&+ \mathbb{P}\left\{ \sup_{|\theta_0 - \theta| < \delta} |\frac{1}{m}\sum_{\ell=1}^{m} \tilde{I}^*_{j,k}(\theta) - \frac{1}{m}\sum_{\ell=1}^{m} \tilde{I}^*_{j,k}(\theta_0)| > \frac{\epsilon}{4d} \right\} \\
&+ \mathbb{P}\left\{ |\frac{1}{m}\sum_{\ell=1}^{m} \tilde{I}^*_{j,k}(\theta_0) - \tilde{I}_{j,k}(\theta_0)| > \frac{\epsilon}{4d} \right\} \Bigg)
\end{aligned}
$$

where we use $I^*_{j,k}(\theta, i), \tilde{I}^*_{j,k}(\theta, \ell)$ to denote $-\frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(Z_i|\theta)$ and $-\frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(\tilde{X}_\ell|\theta)$ respectively, and use $I_{j,k}(\theta), \tilde{I}_{j,k}(\theta)$ to denote the j, k entry of $I_{XY}(\theta)$ and $I_X(\theta)$, respectively. Using the standard Cramér-Chernoff method to replace the Chebyshev inequality with Chernoff inequality (applicable because Condition 3 holds) for the steps in [Clarke, 1989, pp. 43], it is easy to show that the first two terms are upper bounded by $O(e^{-n\rho})$ and the last two terms are upper bounded by $O(e^{-m\rho})$ for some $\rho > 0$. □

Lemma 9 (Probability of C^c) *Assume that Condition 4 holds. Then for some $\rho > 0$, we have*

$$\mathbb{P}\left\{(Z^n, \tilde{X}^m) \in C^c(\delta)\right\} \le O(e^{-\min\{m,n\}\rho})$$

Proof: Define $l_i := \nabla \log p(Z_i|\theta)$ and $\tilde{l}_j = \nabla \log p(\tilde{X}_j|\theta)$. We rewrite $D(\theta_0)$ as

$$
\begin{aligned}
D(\theta_0) &= (\sum_{i=1}^{n} l_i + \sum_{j=1}^{n} \tilde{l}_j)^T I_{m,n}^{-1} (\sum_{i=1}^{n} l_i + \sum_{j=1}^{n} \tilde{l}_j) \\
&= \sum_{i=1}^{n} l_i^T I_{m,n}^{-1} l_i + \sum_{k \ne i} l_i^T I_{m,n}^{-1} l_k \\
&+ \sum_{j=1}^{n} \tilde{l}_j^T I_{m,n}^{-1} \tilde{l}_j + \sum_{k \ne j} \tilde{l}_j^T I_{m,n}^{-1} \tilde{l}_k
\end{aligned}
$$

Then

$$
\begin{aligned}
&\mathbb{P}\left\{(Z^n, \tilde{X}^m) \in C^c(\delta)\right\} = \mathbb{P}\left\{D(\theta_0) > \min\{m,n\}\delta^2\right\} \\
&\le \mathbb{P}\left\{ \frac{1}{n}\sum_{i=1}^{n} l_i^T I_{m,n}^{-1} l_i > \frac{\min\{m,n\}\delta^2}{4n} \right\} \\
&+ \mathbb{P}\left\{ \frac{1}{n(n-1)} \sum_{k \ne i} l_i^T I_{m,n}^{-1} l_k > \frac{\min\{m,n\}\delta^2}{4n(n-1)} \right\} \\
&+ \mathbb{P}\left\{ \frac{1}{m}\sum_{j=1}^{n} \tilde{l}_j^T I_{m,n}^{-1} \tilde{l}_j > \frac{\min\{m,n\}\delta^2}{4m} \right\} \\
&+ \mathbb{P}\left\{ \frac{1}{m(m-1)} \sum_{k \ne j} \tilde{l}_j^T I_{m,n}^{-1} \tilde{l}_k > \frac{\min\{m,n\}\delta^2}{4m(m-1)} \right\}
\end{aligned}
\tag{13}
$$

We can show that each of the four terms has an exponentially fast decay. To see this notice that

$$
\begin{aligned}
&\mathbb{E}\left\{l_i^T I_{m,n}^{-1} l_i\right\} = \mathrm{Tr}(I_{m,n}^{-1} \mathbb{E}\left\{l_i^T l_i\right\}) \\
&\le \frac{1}{\min\{m,n\}} \mathrm{Tr}((I_{XY}(\theta) + I_X(\theta))^{-1} I_{XY}) \\
&\le \frac{1}{\min\{m,n\}} \mathrm{Tr}((I_{XY}(\theta) + I_X(\theta))^{-1} (I_{XY}(\theta) + I_X(\theta))) \\
&= \frac{d}{\min\{m,n\}}
\end{aligned}
$$

where the inequalities hold because $I_{XY}(\theta)$ and $I_X(\theta)$ are positive definite.

$$\mathbb{E}\left\{l_k^T I_{m,n}^{-1} l_i\right\} = \mathrm{Tr}(I_{m,n}^{-1} \mathbb{E}\left\{l_k^T l_i\right\}) = 0$$

as l_i and l_k are independent. Similarly, we also have

$$\mathbb{E}\left\{\tilde{l}_j^T I_{m,n}^{-1} \tilde{l}_j\right\} \le \frac{d}{\min\{m,n\}}$$

and $\mathbb{E}\left\{\tilde{l}_k^T I_{m,n}^{-1} \tilde{l}_k\right\} = 0$.

Assume Condition 4 holds, the Chernoff bound shows that the first term in (13) can be upper bounded by a term of the form $O(e^{-n\rho})$ for some $\rho > 0$ if it holds that

$$\frac{\min\{m,n\}\delta^2}{4n} > \frac{d}{\min\{m,n\}}$$

which always holds for large enough n for the cases $m = \alpha n$ or $m = n^{1+\gamma}$. Similarly, the second term in (13) can be upper bounded by an exponentially fast decaying term if $\frac{\min\{m,n\}\delta^2}{4n(n-1)} > 0$, which is always holds for $\delta > 0$. The same argument holds for the last two terms in (13), which can be upper bounded by a term of the order $O(e^{-m\rho})$ for some $\rho > 0$. □

In the above, we have given the proof of Lemma 4 when $m = \alpha n$ for some $\alpha > 0$, or $m = n^{1+\gamma}$ for some $\gamma > 0$. The case when $m = 0$ follows an almost identical proof except for minor details (in fact this case is even simpler and closer to the proof in [Clarke and Barron, 1990]), and we will not repeat it here.