Appendix

1 Proof of theorem 1

Proof. The proof follows from two lemmas. The first is directly from [1]; the second bounds the difference in densities in terms of the total variation distance.

Lemma 1.1.

$$\log \frac{a}{b} \leq \frac{|a-b|}{\min(a,b)} \qquad \forall \quad a, b \in \mathbb{R}_+.$$

Proof. See [1]. ■

Lemma 1.2.

$$p_1(z) - p_2(z) \leq \alpha ||F_1 - F_2||_{\text{TV}}.$$

Proof.

$$\begin{aligned} p_1(z) - p_2(z) &= \int_{\mathcal{Z}} p(z|x)[dF_1(x) - dF_2(x)] \\ &\leq (\sup_x p(z|x) - \inf_x p(z|x)) \int_{\mathcal{Z}} [dF_1(x) - dF_2(x)] \\ &= \sup_{x,x'} |p(z|x) - p(z|x')| ||F_1 - F_2||_{\text{TV}}. \end{aligned}$$

Now we inspect $\sup_{x,x'} |p(z|x) - p(z|x')|$. Recall $p(z|x) = \alpha \mathbb{1}\{z = x\} + (1-\alpha)g(z)$. It follows:

$$\begin{aligned} \sup_{x,x'} |p(z|x) - p(z|x')| &= \sup_{x,x'} [\alpha \mathbb{1}\{z = x\} + (1-\alpha)g(z) - \alpha \mathbb{1}\{z = x'\} - (1-\alpha)g(z)] \\ &= \alpha. \end{aligned}$$

■

We use Lemma 1.1 and 1.2 to derive the result.

$$
\begin{aligned}
&D_{\mathrm{KL}}(P_1||P_2)+D_{\mathrm{KL}}(P_2||P_1) \\
&\qquad = \int_{\mathcal{Z}} (p_1(z) - p_2(z)) \log \frac{p_1(z)}{p_2(z)} dz \\
&\qquad \leq \int_{\mathcal{Z}} |p_1(z) - p_2(z)| \frac{|p_1(z) - p_1(z)|}{\min(p_1(z), p_2(z))} dz \qquad \text{Lemma 1.1} \\
&\qquad \leq \alpha ||F_1 - F_2|| \int_{\mathcal{Z}} |p_1(z) - p_2(z)| \frac{1}{\min(p_1(z), p_2(z))} dz \qquad \text{Lemma 1.2} \\
&\qquad \leq \alpha \frac{||F_1 - F_2||_{\mathrm{TV}}}{(1-\alpha)g} \int_{\mathcal{Z}} |p_1(z) - p_2(z)| dz \\
&\qquad\quad \text{as } \min(p_1(z), p_2(z)) \geq \inf_x p(z|x) \geq (1-\alpha)g \\
&\qquad = \frac{\alpha^2}{(1-\alpha)} ||F_1 - F_2||_{\mathrm{TV}}^2 \operatorname{Vol}(\mathcal{Z}) \qquad \text{as } g = \frac{1}{\operatorname{Vol}(\mathcal{Z})}.
\end{aligned}
$$

□

1.1 Proof of corollary 1.1

Proof is direct by applying the upper bound on KL divergence stated in Theorem 1 to the usual form of the Le Cam bound presented in (2) in our main text.

1.2 Proof of corollary 1.2

We have a packing $\mathcal{V} \subset \Theta$ such that $||\theta_i - \theta_j||_2^2 \geq 2\delta$ for all $i \neq j$, and for some fixed τ,

$$D_{\mathrm{KL}}(F_{\theta_i}||F_{\theta_j}) \leq \tau\delta \quad \forall i, j.$$

Pinsker's inequality implies that

$$||F_i - F_j||_{\mathrm{TV}}^2 \leq \frac{1}{2}\tau\delta.$$

Combining this with Theorem 1 gives an upper bound on the KL divergence between the observed distributions:

$$\sum_i \sum_j D_{\mathrm{KL}}(F_{\theta_i}||F_{\theta_j}) \leq \frac{\alpha^2}{2(1-\alpha)} \tau\delta \operatorname{Vol}(\mathcal{Z}).$$

The result follows from using this in the upper bound on mutual information (4) and applying the usual Fano inequality (3) from our main text.

2 Proof of lemma 1

Proof.

$$
\begin{aligned}
&D_{\mathrm{KL}}(P_1||P_2)+D_{\mathrm{KL}}(P_2||P_1) \\
&\qquad = \int_{\mathcal{Z}} (p_1(z) - p_2(z)) \log \frac{p_1(z)}{p_2(z)} dz \\
&\qquad = \int_{\mathcal{Z}} |p_1(z) - p_2(z)|| \log \frac{p_1(z)}{p_2(z)}| dz \\
&\qquad \leq \int_{\mathcal{Z}} |p_1(z) - p_2(z)| \frac{|p_1(z) - p_1(z)|}{\min(p_1(z), p_2(z))} dz \qquad \text{Lemma 1.1} \\
&\qquad \leq ||P_1 - P_2||_{\mathrm{TV}}^2 \int_{\mathcal{Z}} \frac{1}{\min(p_1(z), p_2(z))} dz \\
&\qquad \leq \frac{||P_1 - P_2||_{\mathrm{TV}}^2}{\gamma} \operatorname{Vol}(\mathcal{Z}) \qquad \text{where} \gamma = \min[\min_z g_1(z), \min_z g_2(z)].
\end{aligned}
$$

We note that

$$||P_1 - P_2||_{\mathrm{TV}} = \frac{1}{2} \int_Z |\alpha(f_1(z) - f_2(z)) - (1-\alpha)(g_2(z) - g_1(z))| dz.$$

For $\alpha \leq 0.5$ we can choose $g_1(z) = \frac{\alpha f_2(z)}{(1-\alpha)} + c$, $g_2(z) = \frac{\alpha f_1(z)}{(1-\alpha)} + c$, such that $||P_1 - P_2||_{\mathrm{TV}} = 0$ and $\gamma > 0$. This choice results in a minimax risk lower bounded independent of the sample size by

$$\mathfrak{M}_n \geq L(\theta_1, \theta_2)(1 - \frac{\log 2}{\log N}).$$

Intuitively, with $\alpha < 0.5$ the attacker is free to inject points from a distribution G_ϕ from the same family $\mathcal{F}$ as F_θ but with different parameters, 'mimicing' a distribution from $\mathcal{F}$. This makes learning possible only up to permutation. □

3 Proof of theorem 2

Proof.

$$
\begin{aligned}
&D_{\mathrm{KL}}(P_1||P_2)+D_{\mathrm{KL}}(P_2||P_1) \\
&\qquad = \int_{\mathcal{Z}} |p_1(z) - p_2(z)|| \log \frac{\alpha f_1(z) + (1-\alpha) f_2(z)}{\alpha f_2(z) + (1-\alpha) f_1(z)}| dz \\
&\qquad \leq \int_{\mathcal{Z}} |p_1(z) - p_2(z)| \log \frac{\alpha}{1-\alpha} dz \\
&\qquad = 2||P_1 - P_2||_{\mathrm{TV}} \log \frac{\alpha}{1-\alpha} \\
&\qquad = (2\alpha - 1)||F_1 - F_2||_{\mathrm{TV}} \log(1 + \frac{2\alpha - 1}{1-\alpha}) \\
&\qquad \leq \frac{(2\alpha - 1)^2}{1-\alpha} ||F_1 - F_2||_{\mathrm{TV}}.
\end{aligned}
$$

□

3.1 Proof of corollary 2.1

Proof is direct by applying the upper bound on KL divergence in Theorem 2 to the usual form of the Le Cam bound presented in (2) in our main text.

References

(1) J. Duchi, M. Wainwright and M. Jordan, *arXiv preprint arXiv:1302.3203v4*, 2014.