Chapter 7: Minimax lower bounds

1. Minimax risk
2. Reduction to hypothesis testing
3. Le Cam's method: binary testing
4. Interlude: Deviations between probability distributions
5. Fano's method for multi-hypothesis testing

Minimax risk: the definition

Setting:

- Family of probability distributions $\mathcal{P}$ on some space $\mathcal{X}$.
- Surjective Function $\theta\colon \mathcal{P} \to \Theta$, modeling the parameter to be estimated
- An estimator is a function $\hat{\theta}\colon \mathcal{X} \to \Theta$.
- A semi-metric[1] $\rho(\cdot,\cdot)$ on Θ

Goal: Establish a lower-bound on the minimax-risk:

$$\mathcal{M}(\theta;\rho) := \inf_{\hat{\theta}} \sup_{P\in\mathcal{P}} \mathop{\mathbb{E}}_{z\sim P}[\rho(\hat{\theta}(z), \theta(P))].$$

Examples: Estimating the mean, median, mode, density, variance, ...

It useful to rescale ρ by an increasing function Φ, yielding the minimax-risk:

$$\mathcal{M}(\theta;\Phi\circ\rho) := \inf_{\hat{\theta}} \sup_{P\in\mathcal{P}} \mathop{\mathbb{E}}_{z\sim P}[\Phi(\rho(\hat{\theta}(z), \theta(P)))].$$

[1] A semi-metric ρ satisfies all the assumptions of a metric except distinct θ and θ' may satisfy $\rho(\theta,\theta') = 0$.

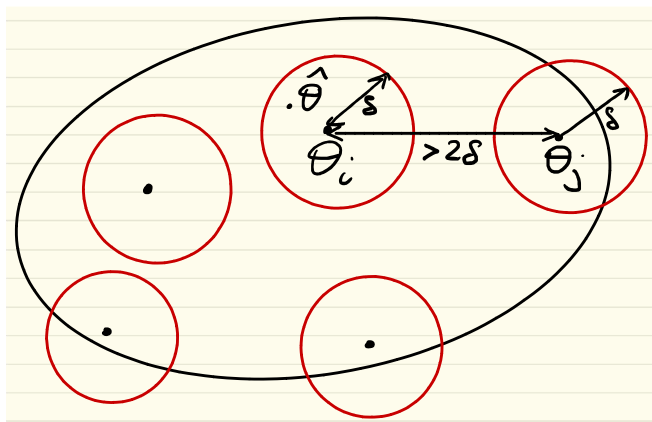
Reduction to hypothesis testing

Lower-bounds on $\mathcal{M}(\theta; \Phi \circ \rho)$ are obtained by reducing to hypothesis testing.

Step 1 (discretize): Let $\{\theta_1, \ldots, \theta_m\} \subset \Theta$ be a 2δ-separated set, meaning

$$\rho(\theta_i, \theta_j) \geq 2\delta \qquad \forall i \neq j.$$

For each j, choose any P_j satisfying $\theta(P_j) = \theta_j$.



Step 2 (mixture): Let J be uniformly sampled from $\{1, \ldots, m\}$ and let Z have distribution P_J.

Step 3 (testing): The goal of hypothesis testing is to determine the index J from the observation Z. This is done with a testing function $\psi \colon \mathcal{X} \to [m]$, which is judged by the mislabeling error $Pr[\psi(Z) \neq J]$.

Key observation: any estimator $\hat{\theta}$ defines a testing function

$$\psi(z) := \arg\min_{j \in [m]} \rho(\theta_j, \hat{\theta}(z)).$$

The following follows immediately from 2δ-separation.

Lemma (Correct testing)

Equality $\psi = J$ holds in the event $E := \{\rho(\hat{\theta}, \theta_J) < \delta\}$ and therefore

$$Pr[\Psi(Z) \neq J] \leq Pr[\rho(\hat{\theta}, \theta_J) \geq \delta].$$

With this lemma, we can reduce the task of establishing minimax lower bounds to hypothesis testing.

Theorem (Reduction to testing)

$$\mathcal{M}(\theta; \Phi \circ \rho) \geq \Phi(\delta) \cdot \inf_{\psi} Pr[\psi(Z) \neq J].$$

Remark: Typically, we will choose $\delta^\star$ such that $Pr[\psi(Z) \neq J] \geq \frac{1}{2}\}$ and then

$$\mathcal{M}(\theta; \Phi \circ \rho) \geq \frac{\Phi(\delta^\star)}{2}.$$

Proof

Fix an estimator $\hat{\theta}$. For any $P \in \mathcal{P}$ define $\theta_P = \theta(P)$. Markov's inequality gives

$$\begin{aligned}\mathop{\mathbb{E}}_{P}\left[\Phi(\rho(\hat{\theta}, \theta_P))\right] &\geq \Phi(\delta) \cdot P[\Phi(\rho(\hat{\theta}, \theta_P)) \geq \Phi(\delta))] \\ &\geq \Phi(\delta) \cdot P[\rho(\hat{\theta}, \theta_P) \geq \delta)].\end{aligned}$$

Next, since the supremum is greater than the average we have

$$\sup_{P \in \mathcal{P}} P[\rho(\hat{\theta}, \theta_P) \geq \delta)] \geq \frac{1}{m} \sum_{j=1}^{m} P_j[\rho(\hat{\theta}, \theta_j) \geq \delta)] = Pr[\rho(\hat{\theta}, \theta_J) \geq \delta].$$

Applying Lemma (correct testing) for ψ induced by $\hat{\theta}$ completes the proof. $\square$

Le Cam's method: binary testing

Surprisingly, one may obtain interesting lower-bounds even for a binary packing $\{\theta_1, \theta_2\}$. In this setting, we must lower bound

$$Pr[\psi(Z) \neq J] = \frac{1}{2} P_1[\psi \neq 1] + \frac{1}{2} P_2[\psi \neq 2].$$

Note that there is a one-to-one correspondence between ψ and measurable partitions (A, A^c) of Θ. Therefore

$$\begin{aligned}
\inf_{\psi} Pr[\psi(Z) \neq J] &= \inf_{A} \frac{1}{2} P_1[A] + \frac{1}{2} P_2[A^c] \\
&= \frac{1}{2}(1 - \sup_{A}\{P_1[A] - P_2[A]\}) \\
&\geq \frac{1}{2}(1 - \|P_1 - P_2\|_{\mathrm{TV}}).
\end{aligned}$$

The right-hand-side measures the similarity between P_1 and P_2.

Interlude: controlling the total variation (TV) distance

Let P and Q be two probability distributions with densities p and q with respect some base measure ν.

- Total Variation (TV) distance

$$\|P - Q\|_{\mathrm{TV}} = \sup_A |P(A) - Q(A)|$$

- Kullback-Leibler (KL) divergence

$$D(P||Q) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) \; \nu(dx)$$

- Squared Hellinger distance

$$H^2(P||Q) = \int \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 \; \nu(dx)$$

Basic properties of the three distances

The TV norm between P and Q is related to the L_1-norm between p and q.

Lemma (TV and L_1 norm)

$$\|P - Q\|_{\mathrm{TV}} = \frac{1}{2}\|p - q\|_{L_1}$$

Proof

Define

$$I_+ := \{x : p(x) \geq q(x)\} \qquad \text{and} \qquad I_- = \{x : p(x) < q(x)\}.$$

We claim $\int_{I+} |p-q| = \int_{I-} |p-q|$. Indeed, this follows from the computation

$$0 = \int p - \int q = \int_{I+} (p-q) - \int_{I-} (q-p).$$

Next, observe $\int |p-q| = 2\int_{I+} |p-q| = 2\int_{I-} |p-q|$. Consequently

$$\|P-Q\|_{\mathrm{TV}} \geq |P(I^+) - Q(I^+)| = \int_{I+} |p-q| = \frac{1}{2}\|P-Q\|_{L_1}.$$

Conversely, for any measurable A, we have

$$\begin{aligned}
|P(A) - Q(A)| &= \left| \int_{A\cap I^+} (p-q) - \int_{A\cap I^-} (q-p) \right| \\
&\leq \max\left\{ \int_{A\cap I^+} (p-q), \int_{A\cap I^-} (q-p) \right\} \\
&\leq \max\left\{ \int_{I^+} |p-q|, \int_{I^-} |q-p| \right\} \leq \frac{1}{2} \int |p-q|,
\end{aligned}$$

where the first inequality uses the identity $|a-b| \leq \max(a,b)$ for all $a, b \geq 0$.

Deviations of products

The main issue with the TV distance is that it is difficult to compute for product distributions P^n and Q^n. The KL-divergence and square Hellinger behave much nicer. You will prove the following for homework.

Lemma

Let $(P_1, \ldots, P_n)$ and $(Q_1, \ldots, Q_n)$ be probability distributions and let $P^{1:n}$ and $Q^{1:n}$ be the product measures. Then

$$D(P^{1:n}||Q^{1:n}) = \sum_{i=1}^{n} D(P_i||Q_i)$$

$$\frac{1}{2}H^2(P^{1:n}||Q^{1:n}) = 1 - \prod_{i=1}^{n}(1 - \frac{1}{2}H^2(P_i||Q_i))$$

In particular, of $P_i = P_1$ and $Q_i = Q_1$ for each i, then

$$D(P^{1:n}||Q^{1:n}) = nD(P_1||Q_1)$$

$$\frac{1}{2}H^2(P^{1:n}||Q^{1:n}) = 1 - (1 - \frac{1}{2}H^2(P_1||Q_1))^n \leq \frac{1}{2}nH^2(P_1||Q_1)$$

Pinsker's and Le Cam's inequalities

Thus we may try to bound the TV distance by the KL divergence and/or the square Hellinger distance.

Theorem (Pinsker)

$$\|P - Q\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2} D(P||Q)}$$

Theorem (Le Cam's inequality)

$$\|P - Q\|_{\mathrm{TV}} \leq H(P||Q) \cdot \sqrt{1 - \frac{H^2(P||Q)}{4}}$$

Proof of Pinsker's inequality (due to John M. Pollard)

We will use to basic facts. First, the inequality

$$(1+t)\log(1+t) - t \geq \frac{1}{2} \cdot \frac{t^2}{1+t/3} \qquad \forall t.$$

This can be verified by elementary calculus (do it!) Secondly, for any random variable X and a nonnegative random variable Y the Cauchy-Schwarz inequality gives

$$(\mathbb{E}|X|)^2 = \left(\mathbb{E}\frac{|X|}{\sqrt{Y}}\sqrt{Y}\right)^2 \leq \mathbb{E}\left[\frac{X^2}{Y}\right] \cdot \mathbb{E}Y.$$

Now setting $r(x) = \frac{p(x)}{q(x)} - 1$, we compute

$$\begin{aligned} D(P||Q) &= \mathbb{E}_Q[(1+r(x))\log(1+r(x)) - r(x)] \\ &\geq \frac{1}{2}\mathbb{E}_Q\left[\frac{r(x)^2}{1+r(x)/3}\right] \\ &\geq \frac{1}{2}\frac{(\mathbb{E}_Q|r(x)|)^2}{\mathbb{E}_Q(1+r(x)/3)} = \frac{1}{2}(\mathbb{E}_Q|r(x)|)^2 = \frac{1}{2}\left(\int |p-q|\right)^2, \end{aligned}$$

as claimed. $\square$

Proof of Le Cam's inequality

The Cauchy-Schwarz inequality gives

$$\begin{aligned}
2\|P-Q\|_{\mathrm{TV}} = \int |p-q| &= \int |\sqrt{p}-\sqrt{q}|(\sqrt{p}+\sqrt{q}) \\
&\leq \sqrt{\int (\sqrt{p}-\sqrt{q})^2} \cdot \sqrt{\int (\sqrt{p}+\sqrt{q})^2} \\
&= H(P\|Q) \cdot \sqrt{2 + 2\int \sqrt{p}\sqrt{q}}
\end{aligned}$$

Taking into account

$$\begin{aligned}
2\int \sqrt{p}\sqrt{q} &= -\int \left(\left(\sqrt{p}-\sqrt{q}\right)^2 - p - q \right) \\
&= 2 - H^2(P\|Q),
\end{aligned}$$

completes the proof. $\square$

Lower bounds for estimating the mean of a 1D Gaussian

For a fixed variance σ^2, set $P_\theta = N(\theta, \sigma^2)$ and define

$$\mathcal{P} = \{P_\theta^n : \theta \in \mathbb{R}\}.$$

Let us lower-bound the minimax risk:

$$\mathcal{R}_2 := \inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}} \mathop{\mathbb{E}}_{P_\theta} [(\hat{\theta} - \theta)^2].$$

Let us use the Le Cam's two point estimate for P_0^n and $P_{2\delta}^n$ where $\delta > 0$ will be specified shortly. Then we know

$$\mathcal{R}_2 \geq \delta^2 \cdot \left(\frac{1}{2}(1 - \|P_0^n - P_{2\delta}^n\|_{\mathrm{TV}}) \right).$$

Pinsker's inequality gives $\|P_0^n - P_{2\delta}^n\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2} D(P_0^n || P_{2\delta}^n)}$ and algebra shows

$$D(P_0^n || P_{2\delta}^n) = n D(P_0 || P_{2\delta}) = n \frac{(2\delta)^2}{2\sigma^2} = \frac{2n\delta^2}{\sigma^2}.$$

Choosing $\delta = \sqrt{\frac{\sigma^2}{4n}}$ gives

$$\boxed{\mathcal{R}_2 \geq \frac{\sigma^2}{8n}}.$$

The sample mean achieves this lower-bound up to a constant.

Lower bounds for estimating the CDF of a 1D Gaussian

For $\theta \in \mathbb{R}$, set $P_\theta = N(\theta, 1)$ and define

$$\mathcal{P} = \{P_\theta^n : \theta \in \mathbb{R}\}.$$

Let F_θ be the CDF of P_θ. Let us lower-bound the minimax risk:

$$\mathcal{R} := \inf_{\hat{F}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_{P_\theta}[\|\hat{F} - F_\theta\|_\infty].$$

For any $\theta > 0$, we have

$$F_0(0) - F_\theta(0) = \frac{1}{\sqrt{2\pi}} \int_0^\theta e^{-t^2/2} dt \geq \frac{\theta}{\sqrt{2\pi}} e^{-\theta^2/2}.$$

Therefore we may set 2δ to be equal to the right-hand side and then $\|F_0 - F_\theta\|_\infty \geq 2\delta$. Le Cam's two point estimate for P_0^n and P_θ^n implies

$$\mathcal{R} \geq \delta \cdot \left(\frac{1}{2}(1 - \|P_0^n - P_\theta^n\|_{\mathrm{TV}}) \right).$$

Pinsker's inequality gives $\|P_0^n - P_\theta^n\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2} D(P_0^n \| P_\theta^n)} = \sqrt{\frac{n\theta^2}{4}}$. Setting $\theta = \frac{1}{\sqrt{n}}$ and noting $\delta \geq \frac{1}{\sqrt{8\pi e n}}$, we deduce

$$\boxed{\mathcal{R} \geq \frac{1}{8\sqrt{2\pi e}} \cdot \frac{1}{\sqrt{n}}}.$$

The empirical CDF matches this lower bound (recall DKW inequality).

Lower bounds for estimating a shifted uniform distribution

For any $\theta \in \mathbb{R}$, set P_θ be uniformly distributed on $(\theta, \theta+1)$ and define

$$\mathcal{P} = \{P_\theta^n : \theta \in \mathbb{R}\}.$$

Let us lower-bound the minimax risk:

$$\mathcal{R}_2 := \sup_{\hat{\theta}} \sup_{\theta \in \mathbb{R}} \mathop{\mathbb{E}}_{P_\theta} [(\hat{\theta} - \theta)^2].$$

Let us again use the Le Cam's two point estimate for P_0^n and $P_{2\delta}^n$. Then

$$\mathcal{R}_2 \geq \delta^2 \cdot \left(\frac{1}{2}(1 - \|P_0^n - P_{2\delta}^n\|_{\mathrm{TV}}) \right).$$

We can not use Pinsker's inequality because $D(P_\theta, P_{\theta'}) = \infty$ whenever $\theta \neq \theta'$ (why?). Let us compute the Hellinger distance instead. We may assume with loss of generality $\theta' > \theta$. It is easy to show that if $\theta' \in (\theta, \theta+1]$, then $H^2(P_\theta || P_{\theta'}) = 2|\theta - \theta'|$. Therefore as long as $2\delta < 1$, we have

$$H^2(P_0^n || P_{2\delta}^n) \leq n H^2(P_0 || P_{2\delta}) = 4n\delta.$$

Le Cam implies $\|P_0^{1:n} - P_{2\delta}^{1:n}\|_{\mathrm{TV}} \leq 2\sqrt{n\delta}\sqrt{1 - n\delta}$. With $\delta = \frac{1}{8n}$ get

$$\boxed{\mathcal{R}_2 \geq \frac{c}{n^2}}$$

for a constant $c > 0$. This rate is matched by $\hat{\theta}(z) = \min\{z_1, \ldots, z_n\}$ (HW).

Fano's method for multi-hypothesis testing.

Recall the basic inequality:

$$\mathcal{M}(\theta; \Phi \circ \rho) \geq \Phi(\delta) \cdot \inf_{\psi} Pr[\psi(Z) \neq J].$$

Le Cam's method for binary testing used a binary 2δ separating set $\{\theta_1, \theta_2\}$ yielding the lower bound

$$\inf_{\psi} Pr[\psi(Z) \neq J] \geq \frac{1}{2}(1 - \|P_1 - P_2\|_{\mathrm{TV}}).$$

We next discuss Fano's method which provides a different lower-bound on $\inf_{\psi} Pr[\psi(Z) \neq J]$, which is valid for non-binary packings.

Fano's inequality

The main tool we will use is Fano's inequality, which we will prove later.

Theorem (Fano's inequality)

Consider a 2δ-separated set $\{\theta_1, \theta_2, \ldots, \theta_m\}$ and let J be uniform over $[m]$. Then for any testing function ψ we have

$$Pr[\psi(Z) \neq J] \geq 1 - \frac{\frac{1}{m}\sum_{j=1}^{m} D(P_j||P_J) + \log 2}{\log m}.$$

Typically, we choose a 2δ-separated set so that the right side is at least $1/2$. The main difficulty is in controlling $D(P_j||P_J)$. One upper bound we can use is

$$D(P_j||P_J) = \mathbb{E}_j \log\left(\frac{p_j}{\frac{1}{m}\sum_{i=1}^{m} p_i}\right) \leq \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_j \log\left(\frac{p_j}{p_i}\right) = \frac{1}{m}\sum_{i=1}^{m} D(P_j||P_i),$$

where the inequality follows from concavity of the log. Thus we deduce

$$Pr[\psi(Z) \neq J] \geq 1 - \frac{\frac{1}{m^2}\sum_{i,j=1}^{m} D(P_j||P_i) + \log 2}{\log m}.$$

Lower bounds for mean estimation of multivariate Gaussians

Suppose $d \geq 2$ and for a fixed variance σ^2, set $P_\theta = N(\theta, \sigma^2 I_d)$ and define

$$\mathcal{P} = \{P_\theta^n : \theta \in \mathbb{R}^d\}.$$

Let us lower-bound the minimax risk:

$$\mathcal{R}_2 := \inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}_{P_\theta}[\|\hat{\theta} - \theta\|_2^2].$$

Let us choose a 2δ-separated set $\{\theta_1, \ldots, \theta_m\}$ of the unit ball $r\mathbb{B}$ of radius r to be chosen. As we have seen, we may ensure $\log(m) \geq d \log(\frac{r}{2\delta})$. Then an easy computation gives $D(P_i || P_j) = \frac{\|\theta_i - \theta_j\|^2}{2\sigma^2} \leq \frac{2r^2}{\sigma^2}$. Therefore

$$Pr[\psi(Z) \neq J] \geq 1 - \frac{\frac{1}{m^2}\sum_{i,j=1}^m D(P_j^n || P_i^n) + \log 2}{\log m} \geq 1 - \frac{\frac{2r^2 n}{\sigma^2} + \log(2)}{d \log(\frac{r}{2\delta})}.$$

Setting $r^2 = \frac{\log(2) d\sigma^2}{8n}$ and $\delta = \frac{r}{4}$ makes the right-hand-side at least $1/4$ and therefore

$$\boxed{\mathcal{R}_2 \geq \frac{\log(2) d\sigma^2}{512n}}.$$

The sample mean matches this rate up to a constant.

Lower bounds for linear regression

Consider the regression observation model

$$y = X\theta^\star + g$$

where $X \in \mathbb{R}^{n\times d}$ is a fixed design matrix and $g \sim N(0, \sigma^2 I_n)$ is the noise. Equivalently, we observe $y \sim N(X\theta^\star, \sigma^2 I_n)$ Define the family of distributions

$$\mathcal{P} = \{N(v, \sigma^2 I_n) : v \in \mathrm{Range}(X)\}.$$

We aim to lower-bound the quantity

$$\mathcal{R}_2 := \inf_{\hat{\theta}} \sup_{\theta\in\mathbb{R}^d} \mathop{\mathbb{E}}_{P_\theta} \left[\frac{1}{n}\left\|X(\hat{\theta}-\theta)\right\|_2^2\right].$$

From the lower-bound on mean-estimation for Gaussians, we have

$$\boxed{\mathcal{R}_2 \geq \frac{\log(2)}{512}\frac{\sigma^2\cdot \mathrm{rank}(X)}{n}}.$$

This bound is achieved by the ordinary least squares estimator. Why doesn't the efficiency of the ridge estimator contradict this?

Towards a proof of Fano's inequality

We will need to introduce some notation from information theory.

Definition (Entropy)

Let Q be a probability distribution with density $q = \frac{dQ}{d\mu}$ with respect to some base measure μ. The Shannon entropy is the function

$$H(Q) = -\mathbb{E}_Q \log(q) = -\int q(x) \log(q(x)) d\mu(x).$$

If X is discrete with mass function $q(x) = Pr(X = x)$ then

$$H(X) = -\sum_{x \in \mathcal{X}} q(x) \log(q(x))$$

Conditional entropy

Definition (Conditional entropy)

Given a pair of random variables (X, Y) with conditional distribution $Q_{X|Y}$, the conditional entropy of $X|Y$ is given by

$$H(X|Y) = \mathbb{E}_Y[H(Q_{X|Y})],$$

If X and Y are discrete with joint mass function $p(x, y)$, then

$$\begin{aligned} H(X|Y) &= -\sum_y \sum_x \log(p(x \mid y))p(x \mid y)p(y) \\ &= -\sum_y \sum_x \log\left(\frac{p(x,y)}{p(y)}\right) p(x,y) \end{aligned}$$

Elementary properties: You will verify these for homework

$$\begin{aligned} H(X) &\leq \log(|\text{support}(X)|) \\ H(X|Y) &\leq H(X) && \text{[contractive]} \\ H(X,Y) &= H(Y) + H(X|Y) && \text{[chain rule]} \\ H(X,Y|Z) &= H(Y|Z) + H(X|Y,Z) && \text{[conditional chain rule]} \end{aligned}$$

Proof of Fano's inequality

Define the random binary random variable

$$V = 1_{[\psi(Z) \neq J]},$$

and let Z be distributed according to P_J. We will prove the following.

Lemma

$$H(V) + Pr[V = 1] \log(m - 1) \geq H(J|Z)$$

Chain rule plus a short computation (do it!) gives

$$H(J|Z) = \underbrace{H(J)}_{=\log(m)} - \underbrace{[H(Z) + H(J) - H(Z, J)]}_{=\frac{1}{m}\sum_{j=1}^{m} D(P_j||P_J)}$$

Since $H(V) \leq \log(2)$, we deduce

$$\log(2) + Pr[\psi(Z) \neq J] \log(m) \geq \log(m) - \frac{1}{m}\sum_{j=1}^{m} D(P_j||P_J),$$

which after rearranging is Fano's inequlity.

Proof of the lemma

We expand $H(V, J|Z)$ in two different ways

$$H(V, J|Z) = H(J|Z) + H(V|J, Z) = H(J|Z)$$
$$H(V, J|Z) = H(V|Z) + H(J|V, Z) \leq H(V) + H(J|V, Z),$$

where the first inequality holds because J is constant conditioned on J and Z. Nest, by definition of the conditional entropy we have

$$H(J|V, Z) = Pr(V = 1)H(J|Z, V = 1) + Pr(V = 0)H(J|Z, V = 0).$$

If $V = 0$, then $J = \psi(Z)$ and therefore $H(J|Z, V = 0) = 0$. On the other hand, if $V = 1$, then $J \neq \psi(Z)$ so that J conditioned $Z, [V = 1]$ can take at most $m - 1$ values and therefore $H(J|Z, V = 1) \leq \log(m - 1)$. We have shown

$$H(V, J|Z) \leq H(V) + Pr[\psi(Z) \neq 1] \log(m - 1),$$

which completes the proof. $\square$