EXERCISE · SLIDE 9 · QUESTION

Entropy

Which binary source is hardest to predict?

Compare A: p=(1,0), B: p=(3/4,1/4), C: p=(1/2,1/2).

A. A has the largest entropy

B. B has the largest entropy

C. C has the largest entropy

Goodfellow et al., Deep Learning (2016), §3.13

EXERCISE · SLIDE 9 · SOLUTION 1/2

Entropy

Which binary source is hardest to predict?

Solution 1 / 2

Compute each expected surprise

H is entropy in bits. A, B, C are the given binary distributions.

1. $H(A) = 0$. A single certain outcome has no surprise.
2. $H(B) = -(3/4)\log_2(3/4) - (1/4)\log_2(1/4) \approx 0.811$.
3. $H(C) = -(1/2)(-1) - (1/2)(-1) = 1$.

Goodfellow et al., Deep Learning (2016), §3.13

EXERCISE · SLIDE 9 · SOLUTION 2/2

Entropy

Which binary source is hardest to predict?

Solution 2 / 2

Compare the distributions

The entropy values are 0, 0.811, and 1 bit.

---

1. Choose C. Its two outcomes are equally plausible.
2. A is fully predictable. B retains a more likely outcome.
3. Maximum binary entropy occurs at equal probabilities, not certainty.

Goodfellow et al., Deep Learning (2016), §3.13

EXERCISE · SLIDE 9 · CHECK THE RESULT

Entropy

Which binary source is hardest to predict?

Compare A: p=(1,0), B: p=(3/4,1/4), C: p=(1/2,1/2).

A. A has the largest entropy

B. B has the largest entropy

C. C has the largest entropy ✓

Goodfellow et al., Deep Learning (2016), §3.13

EXERCISE · SLIDE 14 · QUESTION

KL divergence

What changes when source and model swap?

Now p=(1/2,1/2), q=(3/4,1/4). Find KL in bits.

A About 0.189

B About 0.208

C Exactly zero

Goodfellow et al., Deep Learning (2016), §3.13

EXERCISE · SLIDE 14 · SOLUTION 1/2

KL divergence

What changes when source and model swap?

Solution 1 / 2

Average the model log penalties

p=(1/2,1/2) generates data; q=(3/4,1/4) scores it.

1. $H(p, q) = -(1/2)\log_2(3/4) - (1/2)\log_2(1/4)$.
2. $H(p, q) = 0.207519 + 1 = 1.207519$ bits.
3. $H(p) = 1$ bit because the source is fair.

Goodfellow et al., Deep Learning (2016), §3.13

EXERCISE · SLIDE 14 · SOLUTION 2/2

KL divergence

What changes when source and model swap?

Solution 2 / 2

Subtract source entropy

KL is the excess cross entropy for the given order.

---

1. $D_{\mathrm{KL}}(p\|q) = 1.207519 - 1 = 0.207519$ bits.
2. Choose B. A is the value for the reversed source/model roles.
3. C would require the two distributions to agree.

Goodfellow et al., Deep Learning (2016), §3.13

EXERCISE · SLIDE 14 · CHECK THE RESULT

KL divergence

What changes when source and model swap?

Now p=(1/2,1/2), q=(3/4,1/4). Find KL in bits.

A About 0.189

B About 0.208 ✓

C Exactly zero

Goodfellow et al., Deep Learning (2016), §3.13

EXERCISE · SLIDE 16 · QUESTION

KL divergence

Can a model forbid an event that occurs?

p=(0.9,0.1), q=(1,0). The second outcome is possible under p.

A KL(p||q) is infinite

B KL(p||q) is zero

C The zero-probability term is ignored

Goodfellow et al., Deep Learning (2016), §3.13

EXERCISE · SLIDE 16 · SOLUTION 1/2

KL divergence

Can a model forbid an event that occurs?

Solution 1 / 2

Check the support condition

p assigns probability 0.1 to outcome 2; q assigns it zero.

1. The second term is $0.1 \log(0.1/0) = +\infty$.
2. It has positive source weight, so it cannot be omitted.
3. The convention $0 \log 0 = 0$ applies to zero source weight, not zero model probability.

Goodfellow et al., Deep Learning (2016), §3.13

EXERCISE · SLIDE 16 · SOLUTION 2/2

KL divergence

Can a model forbid an event that occurs?

Solution 2 / 2

Interpret the failure

The data source can produce an event that the model declares impossible.

1. Choose A. The average log penalty is infinite.
2. B would require matching distributions. C confuses source and model zeros.
3. For finite KL, q must be positive wherever p is positive.

Goodfellow et al., Deep Learning (2016), §3.13

EXERCISE · SLIDE 16 · CHECK THE RESULT

KL divergence

Can a model forbid an event that occurs?

p=(0.9,0.1), q=(1,0). The second outcome is possible under p.

- **A** KL(p||q) is infinite ✓
- **B** KL(p||q) is zero
- **C** The zero-probability term is ignored

Goodfellow et al., Deep Learning (2016), §3.13

EXERCISE · SLIDE 24 · QUESTION

KL fitting

An uneven mixture shifts both fitted moments

$p=0.75\,\mathcal{N}(-2,1)+0.25\,\mathcal{N}(2,1)$. Fit one Gaussian by forward KL.

A. Mean -1, variance 1

B. Mean -1, variance 4

C. Mean 0, variance 5

Goodfellow et al., Deep Learning (2016), §§3.8,3.9.3,3.13; course derivation

EXERCISE · SLIDE 24 · SOLUTION 1/2

KL fitting

An uneven mixture shifts both fitted moments

Solution 1 / 2

Compute moments of the target

Component means are -2 and 2; each component variance is 1.

1. $\mu = 0.75(-2) + 0.25(2) = -1.$
2. $\mathbb{E}[X^2] = 0.75(1+4) + 0.25(1+4) = 5.$
3. $v = \mathbb{E}[X^2] - \mu^2 = 5 - 1 = 4.$

Goodfellow et al., Deep Learning (2016), §§3.8,3.9.3,3.13; course derivation

EXERCISE · SLIDE 24 · SOLUTION 2/2

KL fitting

An uneven mixture shifts both fitted moments

Solution 2 / 2

Use the Gaussian-family result

The fitted Gaussian has adjustable mean m and variance s^2.

1. Forward-KL minimization gives $m = \mu = -1$ and $s^2 = v = 4$. Choose B.
2. A omits between-component spread. C ignores the unequal weights.
3. The standard deviation is 2, distinct from variance 4.

Goodfellow et al., Deep Learning (2016), §§3.8,3.9.3,3.13; course derivation

EXERCISE · SLIDE 24 · CHECK THE RESULT

KL fitting

An uneven mixture shifts both fitted moments

$p=0.75\,\mathcal{N}(-2,1)+0.25\,\mathcal{N}(2,1)$. Fit one Gaussian by forward KL.

A. Mean -1, variance 1

B. Mean -1, variance 4 ✓

C. Mean 0, variance 5

Goodfellow et al., Deep Learning (2016), §§3.8,3.9.3,3.13; course derivation

EXERCISE · SLIDE 26 · QUESTION

KL fitting

Does reverse KL always select one mode?

The candidate family contains the exact bimodal target p.

A Yes, one mode is always optimal

B No, q=p minimizes both directions

C Neither direction can reach zero

Goodfellow et al., Deep Learning (2016), §3.13

EXERCISE · SLIDE 26 · SOLUTION 1/2

KL fitting

Does reverse KL always select one mode?

Solution 1 / 2

Apply nonnegativity and equality

The exact target p is available as a candidate q.

1. $D(p\|p) = 0$ in either optimization.
2. Every KL divergence is nonnegative, so zero is a global minimum.
3. The exact bimodal target keeps both modes.

Goodfellow et al., Deep Learning (2016), §3.13

EXERCISE · SLIDE 26 · SOLUTION 2/2

KL fitting

Does reverse KL always select one mode?

Solution 2 / 2

State the scope of the heuristic

A model family may or may not be able to represent the target.

1. Choose B. Mode selection can arise from a restricted family and target geometry.
2. A turns a tendency into a universal rule. C overlooks the exact target candidate.
3. Always name the family before predicting the fitted shape.

Goodfellow et al., Deep Learning (2016), §3.13

EXERCISE · SLIDE 26 · CHECK THE RESULT

KL fitting

Does reverse KL always select one mode?

The candidate family contains the exact bimodal target p.

A Yes, one mode is always optimal

B No, q=p minimizes both directions ✓

C Neither direction can reach zero

Goodfellow et al., Deep Learning (2016), §3.13

EXERCISE · SLIDE 32 · QUESTION

Mutual information

A noisier sensor leaves more uncertainty

X is fair. Y flips X independently with probability e=0.25. Find I(X;Y).

A. About 0.811 bits

B. About 0.189 bits

C. Exactly 0.75 bits

Shannon (1948), §§6,9–12; original course examples

EXERCISE · SLIDE 32 · SOLUTION 1/2

Mutual information

A noisier sensor leaves more uncertainty

Solution 1 / 2

Compute remaining uncertainty

The binary state is fair, with independent flip probability e=0.25.

1. $H(X) = 1$ bit.
2. $H(X \mid Y) = h_2(0.25) = -(1/4)\log_2(1/4) - (3/4)\log_2(3/4)$.
3. $H(X \mid Y) \approx 0.811278$ bits.

Shannon (1948), §§6,9–12; original course examples

Mutual information

A noisier sensor leaves more uncertainty

Solution 2 / 2

Subtract from prior entropy

Mutual information is the expected reduction in entropy.

1. $I(X;Y) = 1 - 0.811278 \approx 0.188722$ bits. Choose B.
2. A is remaining uncertainty. C is the accuracy, not information.
3. More flips move the report toward independence when e approaches 0.5.

Shannon (1948), §§6,9–12; original course examples

EXERCISE · SLIDE 32 · CHECK THE RESULT

Mutual information

A noisier sensor leaves more uncertainty

X is fair. Y flips X independently with probability e=0.25. Find I(X;Y).

A. About 0.811 bits

B. About 0.189 bits ✓

C. Exactly 0.75 bits

Shannon (1948), §§6,9–12; original course examples

EXERCISE · SLIDE 34 · QUESTION

Mutual information

How much does the square reveal?

X is uniform on $\{-1,0,1\}$; $Y=X^2$. $H(Y|X)=0$.

A. $I(X;Y)=0$ because covariance is zero

B. $I(X;Y)=\log_2 3$

C. $I(X;Y)=h_2(1/3)\approx 0.918$ bits

Shannon (1948), §§6,9–12; original course examples

EXERCISE · SLIDE 34 · SOLUTION 1/2

Mutual information

How much does the square reveal?

Solution 1 / 2

Find the distribution of the output

Y=0 when X=0; Y=1 when X=−1 or 1.

1. $P(Y=0) = 1/3$, $P(Y=1) = 2/3$.
2. $H(Y) = -(1/3)\log_2(1/3) - (2/3)\log_2(2/3) \approx 0.918296$.
3. $H(Y \mid X) = 0$ because Y is determined by X.

Shannon (1948), §§6,9–12; original course examples

EXERCISE · SLIDE 34 · SOLUTION 2/2

Mutual information

How much does the square reveal?

Solution 2 / 2

Use the information identity

Mutual information is symmetric, so use the simpler output entropy.

1. $I(X;Y) = H(Y) - H(Y \mid X) = H(Y)$. Choose C.
2. A confuses zero correlation with independence.
3. B is H(X); observing the square does not reveal the sign when Y=1.

Shannon (1948), §§6,9–12; original course examples

EXERCISE · SLIDE 34 · CHECK THE RESULT

Mutual information

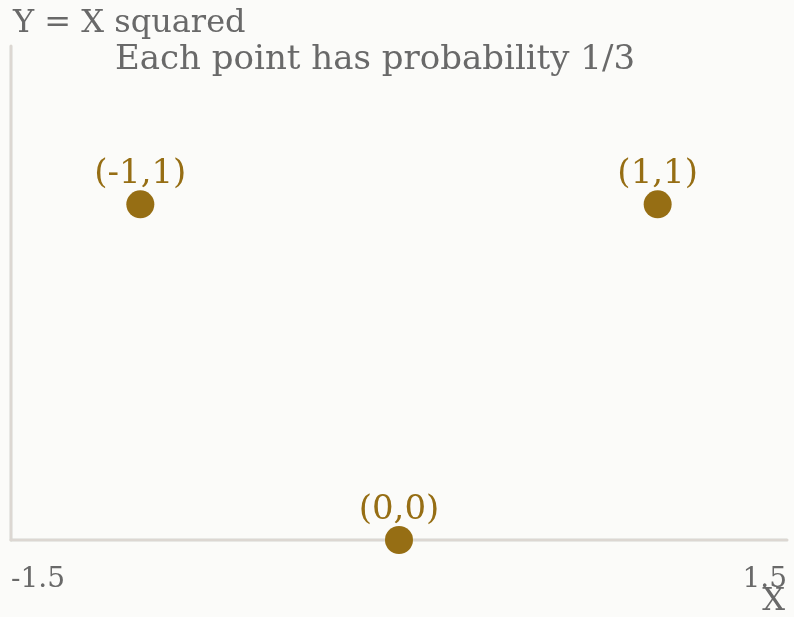
How much does the square reveal?

X is uniform on $\{-1,0,1\}$; $Y=X^2$. $H(Y|X)=0$.

A. $I(X;Y)=0$ because covariance is zero

B. $I(X;Y)=\log_2 3$

C. $I(X;Y)=h_2(1/3)\approx 0.918$ bits ✓



Shannon (1948), §§6,9–12; original course examples

EXERCISE · SLIDE 35 · QUESTION

Mutual information

Compare a direct and an inverted sensor

X is a fair binary state. Sensor A gives $Y=X$. Sensor B gives $Y=1-X$.

A. A has more mutual information

B. Both provide 1 bit

C. B provides zero information

Shannon (1948), §§6,9–12; original course examples

EXERCISE · SLIDE 35 · SOLUTION 1/2

Mutual information

Compare a direct and an inverted sensor

Solution 1 / 2

Determine each conditional uncertainty

X is binary and both sensor mappings are deterministic and invertible.

1. A reveals X directly; for B, recover $X = 1 - Y$.
2. For either report, $H(X \mid Y) = 0$.
3. With fair X, $H(X) = 1$, so $I(X;Y) = 1$ bit.

Shannon (1948), §§6,9–12; original course examples

EXERCISE · SLIDE 35 · SOLUTION 2/2

Mutual information

Compare a direct and an inverted sensor

Solution 2 / 2

Separate information from an uncorrected decision

A known inverted report can be decoded before choosing an action.

1. Choose B. Mutual information measures dependence, not raw classification accuracy.
2. A and C treat inversion as missing information, though it is reversible.
3. Calibration tells the robot how to interpret the report.

Shannon (1948), §§6,9–12; original course examples

EXERCISE · SLIDE 35 · CHECK THE RESULT

Mutual information

Compare a direct and an inverted sensor

X is a fair binary state. Sensor A gives $Y=X$. Sensor B gives $Y=1-X$.

A. A has more mutual information

B. Both provide 1 bit ✓

C. B provides zero information

Shannon (1948), §§6,9–12; original course examples

EXERCISE · SLIDE 40 · QUESTION

Monte Carlo estimation

What does four times as much data buy?

Independent draws have the same finite variance. Increase n from 100 to 400.

A. Standard error becomes one quarter

B. Standard error becomes one half

C. Every realized estimate is closer

UC Berkeley CS70 (2017), Note18: mean, variance, concentration

EXERCISE · SLIDE 40 · SOLUTION 1/2

Monte Carlo estimation

What does four times as much data buy?

Solution 1 / 2

Use the square-root law

The one-sample standard deviation σ is unchanged.

1. $\mathrm{SE}_{100} = \sigma/10$.
2. $\mathrm{SE}_{400} = \sigma/20$.
3. $\mathrm{SE}_{400} \,/\, \mathrm{SE}_{100} = 1/2$.

UC Berkeley CS70 (2017), Note18: mean, variance, concentration

Monte Carlo estimation

What does four times as much data buy?

Solution 2 / 2

Interpret the guarantee

Standard error describes the distribution of estimates over repeated datasets.

1. Choose B. Variance becomes one quarter, but standard error becomes one half.
2. C is not guaranteed: an individual larger dataset can give a worse error.
3. The sampling uncertainty shrinks in distribution, not monotonically on every path.

UC Berkeley CS70 (2017), Note18: mean, variance, concentration

EXERCISE · SLIDE 40 · CHECK THE RESULT

Monte Carlo estimation

What does four times as much data buy?

Independent draws have the same finite variance. Increase n from 100 to 400.

A Standard error becomes one quarter

B Standard error becomes one half ✓

C Every realized estimate is closer

[UC Berkeley CS70 (2017), Note18: mean, variance, concentration]

EXERCISE · SLIDE 43 · QUESTION

Monte Carlo estimation

A tighter tolerance changes sample cost

Use $p(1-p) \le 1/4$ and the variance bound. Keep failure probability ≤ 0.05; reduce error tolerance from 0.1 to 0.05.

A. 250 trials suffice by this bound

B. 1000 trials suffice by this bound

C. 2000 trials suffice by this bound

[UC Berkeley CS70 (2017), Note18: mean, variance, concentration]

EXERCISE · SLIDE 43 · SOLUTION 1/2

Monte Carlo estimation

A tighter tolerance changes sample cost

Solution 1 / 2

Substitute the new tolerance

n is fixed; ε=0.05 and failure bound δ=0.05.

1. $P(|\hat{p} - p| \geq \varepsilon) \leq 1/(4n\varepsilon^2)$.
2. $n \geq 1/[4(0.05)^2(0.05)] = 2000$.
3. Halving ε multiplies the sufficient n by four.

UC Berkeley CS70 (2017), Note18: mean, variance, concentration

Monte Carlo estimation

A tighter tolerance changes sample cost

Solution 2 / 2

Read the statement precisely

This is a conservative sufficient count under independent Bernoulli trials.

1. Choose C. A moves in the wrong direction; B ignores the square on ε.
2. The result controls probability of error at least 0.05.
3. It does not assert that each estimate is exactly within tolerance.

UC Berkeley CS70 (2017), Note18: mean, variance, concentration

EXERCISE · SLIDE 43 · CHECK THE RESULT

Monte Carlo estimation

A tighter tolerance changes sample cost

Use $p(1-p) \le 1/4$ and the variance bound. Keep failure probability ≤ 0.05; reduce error tolerance from 0.1 to 0.05.

A 250 trials suffice by this bound

B 1000 trials suffice by this bound

C 2000 trials suffice by this bound ✓

[UC Berkeley CS70 (2017), Note18: mean, variance, concentration]

EXERCISE · SLIDE 48 · QUESTION

Stopping rules

Is the stopped success fraction unbiased?

Completed paths: 1, 01, 00. Their probabilities are 1/2,1/4,1/4 and estimates 1,1/2,0.

A. Its expectation is 1/2

B. Its expectation is 5/8

C. Its expectation is 1

Sheffield, MIT 18.175 (2016), Lecture 14, pp.34–35; original example

Stopping rules

Is the stopped success fraction unbiased?

Solution 1 / 2

Average over completed experiments

The true success probability is p=1/2.

1. $\mathbb{E}[\hat{p}_T] = (1/2)(1) + (1/4)(1/2) + (1/4)(0).$
2. $\mathbb{E}[\hat{p}_T] = 1/2 + 1/8 = 5/8.$
3. The bias is $5/8 - 1/2 = 1/8.$

Sheffield, MIT 18.175 (2016), Lecture 14, pp.34–35; original example

EXERCISE · SLIDE 48 · SOLUTION 2/2

Stopping rules

Is the stopped success fraction unbiased?

Solution 2 / 2

Locate the broken fixed-n argument

The stopping rule favors a short experiment when success occurs early.

1. Choose B. The rule is fixed, but T depends on outcomes.
2. A incorrectly imports the fixed-n unbiasedness proof. C ignores failed first trials.
3. A random denominator correlated with data cannot be pulled outside expectation.

Sheffield, MIT 18.175 (2016), Lecture 14, pp.34–35; original example

EXERCISE · SLIDE 48 · CHECK THE RESULT

Stopping rules

Is the stopped success fraction unbiased?

Completed paths: 1, 01, 00. Their probabilities are 1/2,1/4,1/4 and estimates 1,1/2,0.

A Its expectation is 1/2

B Its expectation is 5/8 ✓

C Its expectation is 1

Sheffield, MIT 18.175 (2016), Lecture 14, pp.34–35; original example

EXERCISE · SLIDE 49 · QUESTION

Stopping rules

Reverse the stopping preference

Stop after the first failure; otherwise take a second fair trial. Report success fraction.

A Expected estimate 3/8

B Expected estimate 1/2

C Expected estimate 5/8

Sheffield, MIT 18.175 (2016), Lecture 14, pp.34–35; original example

EXERCISE · SLIDE 49 · SOLUTION 1/2

Stopping rules

Reverse the stopping preference

Solution 1 / 2

List the completed paths

Each trial is independent Bernoulli(1/2); stop after failure or after two trials.

1. Path 0: probability 1/2, estimate 0.
2. Path 10: probability 1/4, estimate 1/2.
3. Path 11: probability 1/4, estimate 1.

Sheffield, MIT 18.175 (2016), Lecture 14, pp.34–35; original example

EXERCISE · SLIDE 49 · SOLUTION 2/2

Stopping rules

Reverse the stopping preference

Solution 2 / 2

Compute the reported average

Weight each stopped estimate by its completed-path probability.

1. $\mathbb{E}[\hat{p}_T] = (1/2)0 + (1/4)(1/2) + (1/4)1 = 3/8.$
2. Choose A. The downward bias is $3/8 - 1/2 = -1/8.$
3. B assumes fixed n. C belongs to stopping after first success.

Sheffield, MIT 18.175 (2016), Lecture 14, pp.34–35; original example

EXERCISE · SLIDE 49 · CHECK THE RESULT

Stopping rules

Reverse the stopping preference

Stop after the first failure; otherwise take a second fair trial. Report success fraction.

- **A** Expected estimate 3/8 ✓
- **B** Expected estimate 1/2
- **C** Expected estimate 5/8

Sheffield, MIT 18.175 (2016), Lecture 14, pp.34–35; original example

EXERCISE · SLIDE 52 · QUESTION

Monte Carlo estimation

How many draws to see a rare route at least once?

A route has probability 0.1 in each independent draw. Find the smallest fixed integer n with detection probability at least 0.95.

$$P(\text{no rare route}) = (0.9)^n$$

n fixed independent sample count

A At least 10 draws

B At least 29 draws

C At least 95 draws

UC Berkeley CS70 (2017), Note18: mean, variance, concentration

EXERCISE · SLIDE 52 · SOLUTION 1/2

Monte Carlo estimation

How many draws to see a rare route at least once?

Solution 1 / 2

Turn detection into its complement

Each independent draw misses the rare route with probability 0.9.

1. $P(\text{seen at least once}) = 1 - 0.9^n$.
2. $1 - 0.9^n \geq 0.95$ means $0.9^n \leq 0.05$.
3. $n \log 0.9 \leq \log 0.05$.

UC Berkeley CS70 (2017), Note18: mean, variance, concentration

Monte Carlo estimation

How many draws to see a rare route at least once?

Solution 2 / 2

Solve and round in the correct direction

log 0.9 is negative, so dividing reverses the inequality.

1. $n \geq \log 0.05 / \log 0.9 \approx 28.433$. Thus n≥29. Choose B.
2. A mistakes expected waiting count for a 95% detection guarantee.
3. C is sufficient but is not the minimum among fixed integer counts.

UC Berkeley CS70 (2017), Note18: mean, variance, concentration

EXERCISE · SLIDE 52 · CHECK THE RESULT

Monte Carlo estimation

How many draws to see a rare route at least once?

A route has probability 0.1 in each independent draw. Find the smallest fixed integer n with detection probability at least 0.95.

$$P(\text{no rare route}) = (0.9)^n$$

n fixed independent sample count

A At least 10 draws

B At least 29 draws ✓

C At least 95 draws

UC Berkeley CS70 (2017), Note18: mean, variance, concentration

EXERCISE · SLIDE 55 · QUESTION

Density and modes

Does the largest observed count prove the true mode?

Three categories have counts 5,3,2 in ten independent observations.

A. The first category is the empirical mode

B. The first category is certainly the population mode

C. The sample mean defines the mode

Wasserman, 36-708 Density Estimation, §§1,3–4

EXERCISE · SLIDE 55 · SOLUTION 1/2

Density and modes

Does the largest observed count prove the true mode?

Solution 1 / 2

Distinguish observed and population quantities

Counts are 5,3,2 and sample size n=10.

1. Empirical probabilities are $(0.5, 0.3, 0.2)$.
2. The largest observed frequency belongs to the first category.
3. Finite sampling error can change the rank of the observed frequencies.

Wasserman, 36-708 Density Estimation, §§1,3–4

EXERCISE · SLIDE 55 · SOLUTION 2/2

Density and modes

Does the largest observed count prove the true mode?

Solution 2 / 2

Choose only what the data establish

A population mode maximizes true probabilities, not necessarily sample frequencies.

1. Choose A. B makes a certainty claim that ten observations do not support.
2. C confuses an average of labels with the most frequent outcome.
3. Repeat the experiment to see how the empirical mode can vary.

Wasserman, 36-708 Density Estimation, §§1,3–4

EXERCISE · SLIDE 55 · CHECK THE RESULT

Density and modes

Does the largest observed count prove the true mode?

Three categories have counts 5,3,2 in ten independent observations.

- **A** The first category is the empirical mode ✓
- **B** The first category is certainly the population mode
- **C** The sample mean defines the mode

Wasserman, 36-708 Density Estimation, §§1,3–4

EXERCISE · SLIDE 59 · QUESTION

Histograms

Compare bins with different widths

n=7. Bin A: count 4,width 2. Bin B: count 3,width 1. Which has higher histogram density?

A Bin A

B Bin B

C They have equal density

Wasserman, 36-708 Density Estimation, §3; original example

Histograms

Compare bins with different widths

Solution 1 / 2

Normalize each count by its own width

n=7, N_A=4, h_A=2, N_B=3, h_B=1.

1. $\hat{p}_A = 4/(7 \cdot 2) = 2/7$.
2. $\hat{p}_B = 3/(7 \cdot 1) = 3/7$.
3. Although A has more total mass, B concentrates mass into less width.

Wasserman, 36-708 Density Estimation, §3; original example

EXERCISE · SLIDE 59 · SOLUTION 2/2

Histograms

Compare bins with different widths

Solution 2 / 2

Check the areas

Each density times its bin width returns its probability mass.

1. $2(2/7) + 1(3/7) = 1$.
2. Choose B. A compares counts rather than density; C ignores widths.
3. Density and probability mass answer different questions.

Wasserman, 36-708 Density Estimation, §3; original example

EXERCISE · SLIDE 59 · CHECK THE RESULT

Histograms

Compare bins with different widths

n=7. Bin A: count 4,width 2. Bin B: count 3,width 1. Which has higher histogram density?

A Bin A

B Bin B ✓

C They have equal density

Wasserman, 36-708 Density Estimation, §3; original example

EXERCISE · SLIDE 63 · QUESTION

Kernel density estimation

What does a larger bandwidth change?

Use the same samples and a normalized Gaussian kernel. Increase h.

A. Wider bumps, same total area

B. More total probability mass

C. Narrower bumps and more sharp peaks

[Wasserman, 36-708 Density Estimation, §4; original example]

EXERCISE · SLIDE 63 · SOLUTION 1/2

Kernel density estimation

What does a larger bandwidth change?

Solution 1 / 2

Follow the scale factors

The bump is $K((x-x_i)/h)/h$ with $h>0$.

1. Increasing h spreads the bump horizontally.
2. The factor 1/h lowers its height so its area stays 1.
3. The mean of n such bumps remains a normalized density.

Wasserman, 36-708 Density Estimation, §4; original example

EXERCISE · SLIDE 63 · SOLUTION 2/2

Kernel density estimation

What does a larger bandwidth change?

Solution 2 / 2

Interpret the estimator change

A larger bandwidth averages over a wider neighborhood.

1. Choose A. Nearby peaks may merge through more smoothing.
2. B violates normalization. C describes smaller rather than larger h.
3. More smoothing can conceal genuine structure.

Wasserman, 36-708 Density Estimation, §4; original example

EXERCISE · SLIDE 63 · CHECK THE RESULT

Kernel density estimation

What does a larger bandwidth change?

Use the same samples and a normalized Gaussian kernel. Increase h.

- **A** Wider bumps, same total area ✓
- **B** More total probability mass
- **C** Narrower bumps and more sharp peaks

[Wasserman, 36-708 Density Estimation, §4; original example]

EXERCISE · SLIDE 65 · QUESTION

Density and modes

Choose an action from two observed groups

Actions form two separated groups around -2 and 2. A robot needs one plausible action.

A. Always execute the mean 0

B. Model the distribution and check the sampled action

C. Declare every observed value a true mode

Wasserman, 36-708 Density Estimation, §§1,3–4

EXERCISE · SLIDE 65 · SOLUTION 1/2

Density and modes

Choose an action from two observed groups

Solution 1 / 2

Match the method to the goal

The goal is a plausible action, not a minimum squared-error summary.

1. The mean 0 summarizes symmetric data but can fall between the groups.
2. A distribution model can represent both groups and generate alternatives.
3. Generated actions still need task-specific feasibility checks.

Wasserman, 36-708 Density Estimation, §§1,3–4

EXERCISE · SLIDE 65 · SOLUTION 2/2

Density and modes

Choose an action from two observed groups

Solution 2 / 2

Keep the statistical limits visible

A finite sample and a chosen smoothing level do not reveal exact population modes.

1. Choose B. A can average incompatible behaviors.
2. C treats sampling noise as exact density structure.
3. Diffusion and flow models will represent action distributions in later lectures.

Wasserman, 36-708 Density Estimation, §§1,3–4

EXERCISE · SLIDE 65 · CHECK THE RESULT

Density and modes

Choose an action from two observed groups

Actions form two separated groups around -2 and 2. A robot needs one plausible action.

A. Always execute the mean 0

B. Model the distribution and check the sampled action ✓

C. Declare every observed value a true mode

Wasserman, 36-708 Density Estimation, §§1,3–4

EXERCISE · SLIDE 69 · QUESTION

Stable probability computation

Which change preserves the distribution?

For finite logits z, compare two transformations before softmax.

A $z \to z + 1000 \cdot 1$

B $z \to 2z$

C $z \to -z$

Blanchard, Higham & Higham (2021), §§1,4

EXERCISE · SLIDE 69 · SOLUTION 1/2

Stable probability computation

Which change preserves the distribution?

Solution 1 / 2

Factor the shared exponential

1 is the all-ones vector. Every logit receives the same added constant.

1. $e^{z_i+1000} = e^{1000}e^{z_i}$.
2. The common factor cancels between numerator and denominator.
3. Therefore softmax(z+1000·1)=softmax(z) in exact arithmetic.

Blanchard, Higham & Higham (2021), §§1,4

EXERCISE · SLIDE 69 · SOLUTION 2/2

Stable probability computation

Which change preserves the distribution?

Solution 2 / 2

Compare the other transformations

Changing score differences changes the distribution in general.

1. Choose A. Multiplying by 2 magnifies differences and usually sharpens probabilities.
2. Negating reverses score order.
3. Use a stable shifted evaluation rather than directly forming e^{1000}.

Blanchard, Higham & Higham (2021), §§1,4

EXERCISE · SLIDE 69 · CHECK THE RESULT

Stable probability computation

Which change preserves the distribution?

For finite logits z, compare two transformations before softmax.

- **A** $z \to z + 1000 \cdot 1$ ✓
- **B** $z \to 2z$
- **C** $z \to -z$

Blanchard, Higham & Higham (2021), §§1,4

EXERCISE · SLIDE 72 · QUESTION

Stable probability computation

Compute the observed-class log loss

$z=(1000,999)$. The observed class is the second. Compute loss in nats.

A. About 0.313

B. About 1.313

C. Infinite because e^{1000} overflows

Blanchard, Higham & Higham (2021), §§1,4

EXERCISE · SLIDE 72 · SOLUTION 1/2

Stable probability computation

Compute the observed-class log loss

Solution 1 / 2

Shift before computing the log probability

c=1000, shifted logits are(0,−1), true index y=2.

1. $\ell = \log(1 + e^{-1}) - (-1)$.
2. $\log(1 + e^{-1}) \approx 0.313262$.
3. $\ell \approx 1.313262$ nats.

Blanchard, Higham & Higham (2021), §§1,4

EXERCISE · SLIDE 72 · SOLUTION 2/2

Stable probability computation

Compute the observed-class log loss

Solution 2 / 2

Interpret the result

The true class has lower logit and receives probability about 0.268941.

1. Choose B. A is the loss if the first class had occurred.
2. C describes an avoidable overflow in a naive computation, not the mathematical loss.
3. The model probability is positive, so the exact loss is finite.

Blanchard, Higham & Higham (2021), §§1,4

EXERCISE · SLIDE 72 · CHECK THE RESULT

Stable probability computation

Compute the observed-class log loss

z=(1000,999). The observed class is the second. Compute loss in nats.

A About 0.313

B About 1.313 ✓

C Infinite because e^{1000} overflows

Blanchard, Higham & Higham (2021), §§1,4

EXERCISE · SLIDE 74 · QUESTION

Stable probability computation

The loss gradient changes the right scores

q=(0.7,0.2,0.1). The true class is 2. Use gradient descent on logits.

$$\nabla_z \ell = q - e_2$$

`e_2` one-hot target (0,1,0)

`z` adjustable logits

A. Increase all logits equally

B. Increase class 2, decrease classes 1 and 3

C. Decrease class 2, increase classes 1 and 3

Blanchard, Higham & Higham (2021), §§1,4

EXERCISE · SLIDE 74 · SOLUTION 1/2

Stable probability computation

The loss gradient changes the right scores

Solution 1 / 2

Subtract the target

q=(0.7,0.2,0.1) and e_2=(0,1,0).

1. $\nabla_z \ell = (0.7, -0.8, 0.1)$.
2. Gradient descent uses $z_{new} = z - \eta \nabla_z \ell$ with η>0.
3. The class 2 component increases because its gradient is negative.

Blanchard, Higham & Higham (2021), §§1,4

EXERCISE · SLIDE 74 · SOLUTION 2/2

Stable probability computation

The loss gradient changes the right scores

Solution 2 / 2

Read all score changes

Classes 1 and 3 have positive gradients, so their logits decrease.

1. Choose B. C follows ascent instead of descent.
2. A is a shared shift and would leave probabilities unchanged.
3. The gradient components sum tozero:$0.7-0.8+0.1=0$.

Blanchard, Higham & Higham (2021), §§1,4

EXERCISE · SLIDE 74 · CHECK THE RESULT

Stable probability computation

The loss gradient changes the right scores

q=(0.7,0.2,0.1). The true class is 2. Use gradient descent on logits.

$$\nabla_z \ell = q - e_2$$

`e_2` one-hot target (0,1,0)

`z` adjustable logits

A Increase all logits equally

B Increase class 2, decrease classes 1 and 3 ✓

C Decrease class 2, increase classes 1 and 3

Blanchard, Higham & Higham (2021), §§1,4

EXERCISE · SLIDE 77 · QUESTION

Robot synthesis

Evaluate a changed route frequency

Source changes to p=(0.2,0.8); model stays q=(0.6,0.4). Collect a fresh fixed evaluation.

A. Entropy stays the same; cross entropy rises

B. Both quantities stay the same

C. Entropy must increase

Goodfellow et al., Deep Learning (2016), §3.13

EXERCISE · SLIDE 77 · SOLUTION 1/2

Robot synthesis

Evaluate a changed route frequency

Solution 1 / 2

Preserve the source/model roles

p=(0.2,0.8) generates routes; q=(0.6,0.4) is unchanged.

1. Swapping source labels leaves $H(p) \approx 0.721928$ bits.
2. $H(p, q) = -0.2 \log_2 0.6 - 0.8 \log_2 0.4 \approx 1.204936.$
3. $D_{\mathrm{KL}} \approx 1.204936 - 0.721928 = 0.483008$ bits.

Goodfellow et al., Deep Learning (2016), §3.13

EXERCISE · SLIDE 77 · SOLUTION 2/2

Robot synthesis

Evaluate a changed route frequency

Solution 2 / 2

Interpret the change

The model now assigns its smaller probability to the more common route.

1. Choose A. Source uncertainty is unchanged, but prediction quality worsens.
2. B ignores which route the model favors. C confuses uncertainty with mismatch.
3. Fresh fixed trials estimate the new source without mixing old and new regimes.

Goodfellow et al., Deep Learning (2016), §3.13

EXERCISE · SLIDE 77 · CHECK THE RESULT

Robot synthesis

Evaluate a changed route frequency

Source changes to p=(0.2,0.8); model stays q=(0.6,0.4). Collect a fresh fixed evaluation.

A Entropy stays the same; cross entropy rises ✓

B Both quantities stay the same

C Entropy must increase

Goodfellow et al., Deep Learning (2016), §3.13