Chapter 5

Comparing Information II: Cost of Information

So far we have considered decision problems in which the signal informing the agent's decision is given exogenously. In many economic applications, agents can acquire information at a cost and thereby control the signal that they observe. The full problem the agent faces is often specified as

$$\max_{\sigma:\Theta\to\Delta(S)} \int_{\Delta(\Theta)} \max_{a\in A} \mathbb{E}_q[u(a,\theta)]d\tau_\sigma(q) - \text{cost of acquiring } \sigma$$

where τ_σ denotes the distribution over posterior beliefs induced by signal σ.

This chapter discusses how to model the cost of information, and is divided into two sections. Section 5.2 considers *prior-dependent* cost functions that are a function both of the agent's prior $p \in \Delta(\Theta)$ and of the signal $\sigma : \Theta \to \Delta(S)$. Section 5.3 considers *prior-independent* cost functions that depend only the signal σ. The former are often interpreted as costs of information processing while the latter are often associated with a physical or exogenous cost of producing information. Both approaches draw from information theory, and we review relevant concepts in Section 5.1.

Two useful benchmarks to keep in mind are the following.

Example 5.1 (Binary). The unknown state θ is equally likely to take the value 0 or 1, and the agent chooses an action $a \in \{0,1\}$ with payoff $u(a,\theta) = \mathbb{1}(a=\theta)$. This action is based on the signal

$$\begin{array}{ccc} & s=0 & s=1 \\ \theta=0 & \varphi & 1-\varphi \\ \theta=1 & 1-\varphi & \varphi \end{array}$$

where the agent chooses $\varphi \in [0,1]$.

Example 5.2 (Gaussian). An agent chooses an action $a \in \mathbb{R}$ and receives the payoff $-(a-\theta)^2$, where $\theta \sim \mathcal{N}(\mu, \sigma_\theta^2)$ is an unknown state. This action is based on a signal $X = \theta + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, and the signal noise σ_ε^2 is chosen by the agent.

5.1 Information Theoretic Preliminaries

This section reviews the definitions of entropy and KL divergence.

5.1.1 Entropy

First assume a finite set of states Θ with $n \equiv |\Theta|$, and consider beliefs $p = (p_1, \ldots, p_n)$ defined over this set.

Definition 5.1 (Shannon (1948)). *Let* $\Theta = \{\theta_1, \ldots, \theta_n\}$ *for any* $n < \infty$. *The* entropy *of belief* $p \in \Delta(\Theta)$ *is*

$$H(p) = -\sum_{\theta \in \Theta} p(\theta) \ln(p(\theta)) = \mathbb{E}_{\theta \sim p}[-\ln(p(\theta))]$$

where $0 \ln 0 = 0$.

Remark 5.1. Entropy is also sometimes defined as a function of the random variable rather than its distribution, i.e., $H(\theta) = \mathbb{E}[-\ln(p(\theta))]$.

Entropy is a quantification of uncertainty in a distribution. The higher the entropy of the distribution, the more information is contained in the realization of a random variable it governs. (Entropy is also often interpreted as the "surprise factor" of the outcome.)

Example 5.3. Suppose $\Theta = \{\theta_1, \theta_2\}$. The entropy of any belief $(q, 1-q)$ is

$$H(q) = -q \ln(q) - (1-q) \ln(1-q). \tag{5.1}$$

This curve is depicted in Figure 5.1 below. It is concave, minimized at the two degenerate distributions $(0, 1)$ and $(1, 0)$, and maximized at the uniform distribution $(1/2, 1/2)$.

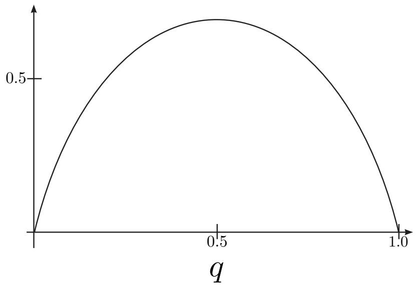

Figure 5.1: Plot of the entropy of the distribution $(q, 1-q)$ as q varies in $[0, 1]$.

Several key properties of entropy are collected below.

Property 1 (Maximal Value). $H(p) \leq H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ for every $n < \infty$ and $p \in \Delta(\{\theta_1, \ldots, \theta_n\})$; that is, entropy is maximized at the uniform distribution.

Property 2 (Probability Zero States). $H(p) = H(p_1, \dots, p_n, 0)$ for every $n < \infty$ and $p \in \Delta(\{\theta_1, \dots, \theta_n\})$; that is, entropy is unchanged by an expansion of the state space to include probability-zero outcomes.

Property 3 (Continuity). H is continuous with respect to all of its arguments.

Property 4 (Chain Rule). Suppose $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} = \{x_1, \dots, x_n\}$ and $\mathcal{Y} = \{y_1, \dots, y_m\}$, where the joint distribution of (X, Y) is denoted p, the marginal distribution of X is p_X, and the conditional distribution of Y given X is $p_{Y|X}$. Then

$$H(p) = H(p_X) + \sum_{i=1}^{n} p_X(x_i) H(p_{Y|X=x_i})$$

or more simply

$$H(X, Y) = H(X) + H(Y \mid X)$$

where $H(X, Y) \equiv H(p)$ is the entropy of the joint distribution, $H(X) \equiv H(p_X)$ is the entropy of of the marginal distribution of X, and

$$H(Y \mid X) \equiv \sum_{i=1}^{n} p_X(x_i) H(p_{Y|X=x_i})$$

is the expected entropy of the conditional distribution of Y given X, also known as the *conditional entropy* of Y given X.

Remark 5.2. In the special case where X and Y are independent, Property 4 implies $H(X, Y) = H(X) + H(Y)$.

Property 5 (Nonnegativity). $H(p) \geq 0$ for all distributions p.

Property 6 (Degenerate Distributions). $H(p) = 0$ for all degenerate distributions p.

Property 7 (Concavity). H is concave.

Property 8 (Relabelling of States). $H(p_1, \dots, p_n) = H(p_{\pi(1)}, \dots, p_{\pi(n)})$ for any bijection π from $\{1, \dots, n\}$ to itself; that is, entropy is invariant to a relabelling of states.

Property 9 (Information Reduces Uncertainty). $H(Y \mid X) \leq H(Y)$ with equality if and only if X and Y are independent; that is, conditioning on information reduces expected entropy.

Properties 1-4 constitute a set of necessary and sufficient conditions for the form of H given in (5.1), up to rescaling.

Proposition 14 (Khinchin (1957)). *Let $H(p_1, \dots, p_n)$ be a function defined for any $n \in \mathbb{Z}_+$ and for all values $p_1, \dots, p_n$ satisfying $p_i \geq 0$ for each $i = 1, \dots, n$ and $\sum_{i=1}^n p_i = 1$. Then H satisfies Properties 1-4 if and only if*

$$H(p_1, \dots, p_n) = -\lambda \sum_{i=1}^{n} p_i \ln(p_i)$$

for some constant $\lambda > 0$.[1]

Properties 5, 6, and 8 are immediate from the functional form of entropy. Property 7 (concavity) follows because $-x \log(x)$ is concave, and the sum of concave functions is concave. (In fact, the same argument shows that entropy is *strictly* concave, so Property 1 can be strengthened to the statement that the uniform distribution is the unique maximum.) The following exercise asks you to prove that entropy satisfies Property 9.

Exercise 5.1 (G). *Suppose $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with $|\mathcal{X}| = n$ and $|\mathcal{Y}| = m$, where p_X and p_Y denote the marginal distributions of X and Y, and $p_{Y|X}$ denotes the conditional distribution of Y given X. Let $H(Y) \equiv H(p_Y)$ be the entropy of of the marginal distribution of Y, and $H(Y \mid X) \equiv \sum_{i=1}^n p_X(x_i) H(p_{Y|X=x_i})$ be the conditional entropy of Y given X. Prove that $H(Y \mid X) \leq H(Y)$.*

Shannon (1948) defines a continuous version of entropy.

Definition 5.2. *The entropy of probability density p on $\Theta \subseteq \mathbb{R}$ is*

$$H(p) = -\int_{\theta \in \Theta} p(\theta) \ln(p(\theta)) d\theta$$

Example 5.4. Recall that the normally distributed variable $\theta \sim \mathcal{N}(\mu, \sigma^2)$ has density $p(\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\theta-\mu}{\sigma}\right)^2}$. The entropy of this distribution is

$$\begin{aligned} \mathbb{E}\left[-\ln\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\theta-\mu}{\sigma}\right)^2}\right)\right] &= -\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \frac{1}{2\sigma^2}\mathbb{E}\left[(\theta-\mu)^2\right] \\ &= \frac{1}{2}\ln\left(2\pi\sigma^2\right) + \frac{1}{2} \qquad (5.2) \end{aligned}$$

using in the second equality that $\mathbb{E}[(\theta - \mu)^2] = \sigma^2$. So entropy and variance order normal distributions in the same way.

5.1.2 Kullback-Liebler Divergence

The *Kullback-Liebler Divergence (KL divergence)*, also known as *relative entropy*, quantifies how different two distributions are.

[1] Recalling that $\log_b(x) = \frac{\log_a(x)}{\log_a(b)}$ for any two bases $a, b > 0$, changing the logarithm to a different basis simply rescales the measure. Choice of base 2 and of base e are both common.

Definition 5.3 (KL-Divergence). *Let $\Theta = \{\theta_1, \ldots, \theta_n\}$ for any $n < \infty$, and let $p, q \in \Delta(\Theta)$. Then the KL divergence from q to p is*

$$D(p\|q) = \sum_{\theta \in \Theta} p(\theta) \ln\left(\frac{p(\theta)}{q(\theta)}\right) = \mathbb{E}_{\theta \sim p}\left[\ln\left(\frac{p(\theta)}{q(\theta)}\right)\right]$$

where $0 \ln 0 = 0$*.*

Example 5.5 (Binary). Let $\Theta = \{\theta_1, \theta_2\}$ with $(p, 1-p)$ and $(q, 1-q)$ be two distributions on this set. Then

$$D(p\|q) = p \ln\left(\frac{p}{q}\right) + (1-p)\ln\left(\frac{1-p}{1-q}\right).$$

Intuitively, larger log likelihood ratios $\ln\left(\frac{p}{q}\right)$ and $\ln\left(\frac{1-p}{1-q}\right)$ reflect distributions that are more different. KL divergence aggregates these log likelihood ratios by weighting them with respect to their probabilities under a reference distribution, which is chosen to be either of p or q.

Example 5.6 (Gaussian). Let p and q denote two Gaussian densities with common variance σ and different means μ_p and μ_q. Then

$$\begin{aligned} D(p\|q) &= \mathbb{E}_{\theta \sim p}\left[\ln\left(\frac{e^{-\frac{1}{2}\left(\frac{\theta-\mu_p}{\sigma}\right)^2}}{e^{-\frac{1}{2}\left(\frac{\theta-\mu_q}{\sigma}\right)^2}}\right)\right] \\ &= \frac{\mu_q^2 - \mu_p^2}{2\sigma^2} - \frac{\mu_q - \mu_p}{\sigma^2} \cdot \mathbb{E}_{\theta \sim p}(\theta) = \frac{(\mu_q - \mu_p)^2}{2\sigma^2} \end{aligned}$$

So as we might expect, the further the two means, the larger the KL divergence between the two distributions.

KL divergence is not in general symmetric (with Example 5.6 being a notable exception) and hence it is not a metric. Other key properties of the KL divergence include:

Property 10 (Nonnegativity). $D(p\|q) \geq 0$ for all $p, q \in \Delta(\Theta)$, with equality if and only if $p = q$.

To prove this, observe that

$$\begin{aligned} -D(p\|q) &= \mathbb{E}_{\theta \sim p}\left[\ln\left(\frac{q(\theta)}{p(\theta)}\right)\right] \\ &\leq \ln\left(\mathbb{E}_{\theta \sim p}\left[\frac{q(\theta)}{p(\theta)}\right]\right) && \text{by Jensen's inequality} \\ &= \ln(1) = 0 && \text{since } \textstyle\sum_{\theta \in \Theta} p(\theta)\left(\frac{q(\theta)}{p(\theta)}\right) = 1 \end{aligned}$$

Property 11 (Additivity for Independent Distributions). Suppose $p_1 \in \Delta(\mathcal{X}_1)$ and $p_2 \in \Delta(\mathcal{X}_2)$ are independent distributions, with $p(x_1, x_2) = p_1(x_1)p_2(x_2)$. Likewise suppose $q_1 \in \Delta(\mathcal{X}_1)$ and $q_2 \in \Delta(\mathcal{X}_2)$ are independent distributions with $q(x_1, x_2) = q(x_1)q(x_2)$. Then

$$D(p\|q) = D(p_1\|q_1) + D(p_2\|q_2),$$

i.e., KL divergence is additive for independent distributions.

This property follows from straightforward algebra:

$$\begin{aligned}
D(p\|q) &= \sum_{x_1\in\mathcal{X}_1}\sum_{x_2\in\mathcal{X}_2} p(x_1,x_2)\ln\left(\frac{p(x_1,x_2)}{q(x_1,x_2)}\right)\\
&= \sum_{x_1\in\mathcal{X}_1}\sum_{x_2\in\mathcal{X}_2} p_1(x_1)p_2(x_2)\ln\left(\frac{p_1(x_1)p_2(x_2)}{q_1(x_1)q_2(x_2)}\right)\\
&= \sum_{x_2\in\mathcal{X}_2} p_2(x_2)\left(\sum_{x_1\in\mathcal{X}_1} p_1(x_1)\ln\left(\frac{p_1(x_1)}{q_1(x_1)}\right)\right)\\
&\quad + \sum_{x_1\in\mathcal{X}_1} p_1(x_1)\left(\sum_{x_2\in\mathcal{X}_2} p_2(x_2)\ln\left(\frac{p_2(x_2)}{q_2(x_2)}\right)\right) = D(p_1\|q_1)+D(p_2\|q_2)
\end{aligned}$$

where independence is invoked in the second equality.

Property 12 (Convexity). D is convex: For any two pairs (p,q) and (p',q'), and any $\alpha\in[0,1]$, we have

$$D\left(\alpha p+(1-\alpha)p'\|\alpha q+(1-\alpha)q'\right)\le \alpha D(p\|q)+(1-\alpha)D(p'\|q')$$

Exercise 5.2 (G). *Prove the above property using the following fact:*

Fact 5.1 (Log-Sum Inequality). *Let $a_1,\dots a_n$ and $b_1,\dots b_n$ be nonnegative real numbers. Then*

$$\sum_{i=1}^n a_i\ln\left(\frac{a_i}{b_i}\right)\ge\left(\sum_{i=1}^n a_i\right)\ln\left(\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}\right).$$

There is a close relationship between KL divergence and entropy. First, the entropy of a distribution $p\in\Delta(\Theta)$ with $n\equiv|\Theta|<\infty$ can be rewritten directly in terms of KL divergence:

$$H(p)=\ln n - D(p\|U)$$

where U denotes the uniform distribution on Θ. Thus, the larger the KL divergence from the uniform distribution to p, the lower the entropy of p. This is proved by observing that

$$\begin{aligned}
\ln n - D(p\|U) &= \ln n - \sum_{\theta\in\Theta} p(\theta)\ln\left(\frac{p(\theta)}{1/n}\right)\\
&= \sum_{\theta\in\Theta} p(\theta)(\ln n - \ln(np(\theta))) \qquad \text{since } \textstyle\sum_{\theta\in\Theta} p(\theta)=1\\
&= -\sum_{\theta\in\Theta} p(\theta)\ln(p(\theta)) = H(p)
\end{aligned}$$

Remark 5.3. Together with Property 10, the above relationship implies that entropy is maximized at the uniform distribution (Property 1).

KL divergence cannot be rewritten directly in terms of entropy, although

$$D(p\|q)=-\sum_{\theta\in\Theta} p(\theta)\ln(q(\theta))-H(p)$$

where $-\sum_{\theta\in\Theta} p(\theta)\ln(q(\theta))$ is the *cross-entropy* of distribution q relative to p.

5.2 Prior-Dependent Costs

Returning to the question of how to model the cost function, we begin with *prior-dependent* cost functions. Dependence on the prior belief means that the cost of absorbing the information content of a signal varies with what the agent already knows. This feature may be justified if we view the cost of information as an information processing or cognitive cost: For example, processing a news article about a proposed tax change may be relatively easy for someone who already understands this tax change well, but cognitively taxing for someone who does not.

It will be convenient to represent signals as distributions over posterior beliefs, as in Section 2.2.2. Following Definition 2.2, we use $\mathcal{T}(p)$ to denote the set of Bayes plausible distributions given prior p, and we further define

$$\mathcal{S} = \{(p, \tau) : p \in \Delta(\Theta), \tau \in \mathcal{T}(p)\}$$

to be the domain of prior beliefs and Bayes plausible distributions. The cost functions in this section will take the form $C : \mathcal{S} \to \mathbb{R}$.

5.2.1 Uniform Posterior Separability

One popular class of cost functions are those that are *uniformly posterior separable.*

Definition 5.4 (Caplin and Dean (2013); Caplin, Dean and Leahy (2022)). *The cost function $C : \mathcal{S} \to \mathbb{R}$ is* uniformly posterior separable *(henceforth UPS) if there is a strictly concave function Φ such that*

$$C(p, \tau) = \Phi(p) - \mathbb{E}_{q\sim\tau}[\Phi(q)] \quad \forall (p, \tau) \in \mathcal{S}.$$

We can interpret this cost of information as the expected reduction of uncertainty, where $\Phi : \Delta(\Theta) \to \mathbb{R}$ measures how uncertain the belief is.

Remark 5.4. The cost of "no information" is zero, since $\Phi(p) - \mathbb{E}_{q\sim\delta_p}[\Phi(q)] = \Phi(p) - \Phi(p) = 0$ (with δ_p denoting the degenerate distribution at the prior p).

Remark 5.5. Concavity of Φ guarantees that uncertainty decreases in expectation when more information is received. Together with Bayes plausibility of τ, this further implies that UPS cost functions are everywhere positive:

$$\begin{aligned}
\Phi(p) - \mathbb{E}_{q\sim\tau}[\Phi(q)] &\geq \Phi(p) - \Phi(\mathbb{E}_{q\sim\tau}[q]) && \text{by Jensen's inequality} \\
&= \Phi(p) - \Phi(p) && \text{by Bayes plausibility of } \tau \\
&= 0
\end{aligned}$$

Remark 5.6. UPS cost functions are consistent with the Blackwell order. That is, let σ and σ' be arbitrary signals where σ Blackwell dominates σ'. Fix any prior p, and let τ_σ and $\tau_{\sigma'}$ denote the distributions over posteriors that are induced by σ and σ'. Then for any UPS cost function C, we have $C(p, \tau_\sigma) \geq C(p, \tau_{\sigma'})$ since

$$C(p, \tau) = \int (\Phi(p) - \Phi(q)) d\tau(q)$$

where $\Phi(p) - \Phi(q)$ is convex in q, and τ_σ dominates $\tau_{\sigma'}$ in the convex order (see the characterization of the Blackwell order in Section 4.3.2).

The leading specification of C is the expected reduction of the entropy of the agent's belief.

Example 5.7 (Entropy Reduction). Let H be the entropy function given in Definition 5.1. Then define

$$C_{\text{Ent}}(p,\tau) = H(p) - \mathbb{E}_{q\sim\tau}[H(q)] \quad \forall (p,\tau) \in \mathcal{S} \tag{5.3}$$

to be the expected reduction in the entropy of the agent's belief.

Initially proposed as an information cost in Sims (2003), this cost function is a cornerstone of the rational inattention literature (Caplin and Dean, 2013; Caplin, Dean and Leahy, 2015; Hebert and Woodford, 2021*a*; Hébert and La'O, 2022). Various conceptual foundations for entropic costs and uniformly posterior separable cost functions (as well as the broader class of posterior separable cost functions discussed in Section 5.2.3) can be found in Caplin and Dean (2013), Matějka and McKay (2015), Morris and Strack (2019), Hebert and Woodford (2021*b*), Bloedel and Zhong (2021), and Denti (2022) among others.

Example 5.8. In the setting of Example 5.1, we have

$$C_{\text{Ent}}(p,\tau_\varphi) = -\ln\left(\frac{1}{2}\right) + (\varphi\ln(\varphi) + (1-\varphi)\ln(1-\varphi))$$

where τ_φ denotes the distribution over posterior beliefs induced by the signal indexed to φ. The cost of the signal is largest when $\varphi \in \{0,1\}$ (corresponding to a fully revealing signal) and smallest when $\varphi = 1/2$ (corresponding to an uninformative signal).

Besides entropy, another natural choice of Φ is variance.

Example 5.9 (Variance Reduction). Let

$$C_{\text{Var}}(p,\tau) = \text{Var}(p) - \mathbb{E}_{q\sim\tau}[\text{Var}(q)] \tag{5.4}$$

be the expected reduction in the variance of the agent's belief.

Exercise 5.3 (G). *Prove that variance is strictly concave, so C_{Var} is a UPS cost function.*

Example 5.10. Consider the setting of Example 5.2 (where we use $\tau_{\sigma_\varepsilon^2}$ to denote the distribution over posterior beliefs induced by observing the signal $X = \theta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0,\sigma_\varepsilon^2)$). Applying (5.2),

$$\begin{aligned} C_{Ent}(p,\tau_{\sigma_\varepsilon^2}) &= \left(\frac{1}{2}\ln(2\pi\sigma_\theta^2) + \frac{1}{2}\right) - \left(\frac{1}{2}\ln\left(2\pi\left(\frac{\sigma_\theta^2\sigma_\varepsilon^2}{\sigma_\theta^2+\sigma_\varepsilon^2}\right)\right) + \frac{1}{2}\right) \\ &= \frac{1}{2}\ln\left(\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_\varepsilon^2}\right) \end{aligned}$$

while

$$C_{Var}(p, \tau_{\sigma_\varepsilon^2}) = \sigma_\theta^2 - \frac{\sigma_\theta^2 \sigma_\varepsilon^2}{\sigma_\theta^2 + \sigma_\varepsilon^2} = \frac{\sigma_\theta^4}{\sigma_\theta^2 + \sigma_\varepsilon^2}.$$

For every fixed prior variance σ_θ^2, both cost functions are strictly decreasing in the noise variance σ_ε^2, and thus correspond to different cardinal representations of the same ordering over signals. One interesting contrast is that $C_{Ent}(p, \sigma_\varepsilon^2) \to \infty$ as $\sigma_\varepsilon^2 \to 0$, while $C_{Var}(p, \sigma_\varepsilon^2) \to \sigma^2$. That is, the cost of information using C_{Var} is bounded above by the agent's prior uncertainty, while entropy cost is unbounded.

5.2.2 Decision-Theoretic Foundations

The function Φ is interpreted in the previous section as a "pure" measure of uncertainty, without reference to why this uncertainty matters. Parallel to Section 4.2's assessment of the value of information using decision problems, Frankel and Kamenica (2019) microfound the function Φ as measuring the instrumental loss of uncertainty for a specific decision problem.

Definition 5.5. *For any belief $q \in \Delta(\Theta)$ and decision problem $\mathcal{D} = (A, u)$, let*

$$\Phi_{\mathcal{D}}(q) = \mathbb{E}_q \left[\max_{a \in A} u(a, \theta) \right] - \max_{a \in A} \mathbb{E}_q \left[u(a, \theta) \right].$$

The first term of this expression is the agent's best expected payoff when conditioning his action directly on the realized state (which is random and distributed according to the agent's belief q). The second term is the best expected payoff that the agent with belief q can achieve given no additional information on which to condition his action. Thus $\Phi_{\mathcal{D}}$ quantifies the agent's payoff loss from not knowing a state which is distributed according to q.

Definition 5.6 (Frankel and Kamenica (2019)). *Say that $\Phi : \Delta(\Theta) \to \mathbb{R}$ is* valid *if there is a decision problem $\mathcal{D}$ such that $\Phi = \Phi_{\mathcal{D}}$.*

Any function Φ that is concave and takes value zero at degenerate distributions (i.e., satisfies Properties 6 and 7) can be microfounded using a decision problem in this way.

Proposition 15 (Frankel and Kamenica (2019))**.** *$\Phi : \Delta(\Theta) \to \mathbb{R}$ is valid if and only if it satisfies Properties 6 and 7.*

This result follows from the subsequent lemma, which is of independent interest.

Lemma 1. *Let Θ be a finite set. Then every convex function $V : \Delta(\Theta) \to \mathbb{R}$ can be represented as*

$$V(q) = \sup_{a \in A} \mathbb{E}_q[u(a, \theta)] \quad \forall q \in \Delta(\Theta) \tag{5.5}$$

for some decision problem (A, u), where A is a set (not necessarily finite) and u is a map $u : \Theta \times A \to [-\infty, +\infty]$.

The key points in the proof of this lemma are that $\mathbb{E}_q(u(a,\theta))$ is affine in q, and that every convex function is the supremum of affine functions lying below it. We'll prove this lemma assuming that V is continuous and has a nonvertical supporting hyperplane at every point $q \in \Delta(\Theta)$, leaving the completion of the proof when these assumptions fail as Exercise 5.4.[2]

Proof. Our approach is to construct a set of actions indexed to beliefs, $A = \{a_q : q \in \Delta^n\}$, and to construct a utility function such that each action a_q is optimal at the belief q. To do this, define a family of affine functions $(U_{a_q})_{q \in \Delta^n}$, where each $U_{a_q} : \Delta^n \to \mathbb{R}$ is a supporting hyperplane of the epigraph of V at q, as depicted below in Figure 5.2.[3]

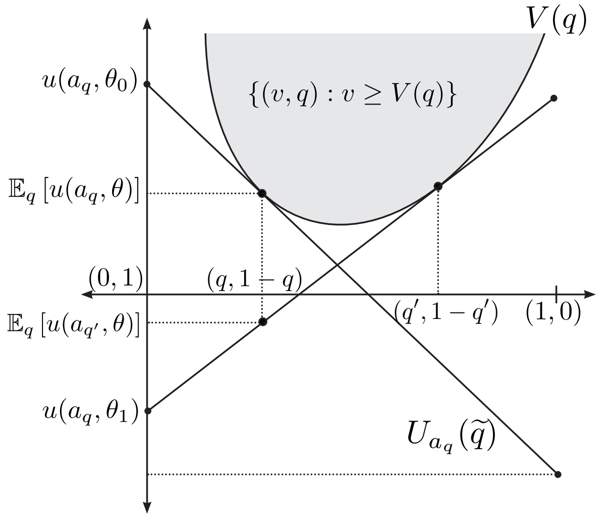
$V(q)$
$u(a_q, \theta_0)$
$\{(v,q) : v \geq V(q)\}$
$\mathbb{E}_q\left[u(a_q,\theta)\right]$
$(0,1)$
$(q, 1-q)$
$(q', 1-q')$
$(1,0)$
$\mathbb{E}_q\left[u(a_{q'},\theta)\right]$
$u(a_q, \theta_1)$
$U_{a_q}(\widetilde{q})$

Figure 5.2: Example construction for a binary state space $\Theta = \{\theta_0, \theta_1\}$. The action a_q is optimal at belief q; that is, for every other belief q' we have $\mathbb{E}_q[u(a_q,\theta)] \geq \mathbb{E}_q[u(a_{q'},\theta)]$, as depicted here.

Since V is continuous and convex, it can be represented on its domain as the supremum of all affine functions lying below it. Since each U_{a_q} is affine and lies below V, we have that

$$V(q) \geq U_a(q) \quad \forall a \in A, q \in \Delta^n.$$

Moreover (by definition) U_{a_q} supports V at q, so $U_{a_q}(q) = V(q)$. This implies that

$$U_{a_q}(q) = \max_{a \in A} U_a(q) \quad \forall q \in \Delta(\Theta) \tag{5.6}$$

We now need to express U_{a_q} as an expected utility function. Since each belief q' is a convex combination of the degenerate beliefs $(\delta_\theta)_{\theta \in \Theta}$ (with weights given

[2] Under these assumptions, the supremum in (5.5) can be replaced with maximum, as the following proof demonstrates.

[3] Recall that the epigraph of V is $\{(q,v) : v \geq V(q)\}$, the set of points lying on or above V.

by $q'(\theta)$), and U_{a_q} is affine, it follows that

$$U_{a_q}(q') = \sum_{\theta\in\Theta} q'(\theta)U_{a_q}(\delta_\theta) \quad \forall q' \in \Delta(\Theta) \tag{5.7}$$

Now define the utility function $u : \mathbb{R}^n \to \mathbb{R}$ to satisfy $u(a,\theta) = U_a(\delta_\theta)$ for every $a \in A$ and $\theta \in \Theta$. Then from (5.7),

$$U_{a_q}(q') = \sum_{\theta\in\Theta} q'(\theta)u(a_q,\theta)$$

and so (5.6) implies that

$$\mathbb{E}_q[u(a_q,\theta)] \geq \mathbb{E}_q[u(a,\theta)]$$

for every $q \in \Delta(\Theta)$ and $a \in A$. Thus each action a_q is optimal at belief q, and achieves the expected utility $U_{a_q}(q) = V(q)$ as desired. ■

Exercise 5.4 (G). *Complete the proof by showing that the statement of Lemma 1 continues to hold when V is discontinuous and/or there exists a belief q at which every supporting hyperplane of V is vertical.*

Hint 1. *Observe that vertical supporting hyperplanes can only exist on the boundary of $\Delta(\Theta)$, and that discontinuities can only occur at degenerate beliefs.*

We'll now use this lemma to prove Proposition 15.

Proof. Suppose Φ satisfies Assumptions 6 and 7. Then $-\Phi$ is convex, so by Lemma 1, there is a set of actions A and a utility function $u : A \times \Theta \to \mathbb{R}$ such that

$$-\Phi(q) = \max_{a\in A} \mathbb{E}_q[u(a,\theta)] \quad \forall q \in \Delta(\Theta). \tag{5.8}$$

We need to verify that

$$\Phi(q) = \mathbb{E}_q\left[\max_{a\in A} u(a,\theta)\right] - \max_{a\in A} \mathbb{E}_q[u(a,\theta)] \tag{5.9}$$

for every $q \in \Delta(\Theta)$. Again index the states by $\theta_1, \ldots, \theta_n$ (where $n \equiv |\Theta|$), and define δ_{θ_i} to be the belief that is degenerate at state θ_i. Then for any $\theta_i \in \Theta$

$$\begin{aligned}
\max_{a\in A} u(a,\theta_i) &= \max_{a\in A} \mathbb{E}_{\delta_{\theta_i}}[u(a,\theta)] \\
&= -\Phi(\delta_{\theta_i}) && \text{by (5.8)} \\
&= 0 && \text{by Assumption 6}
\end{aligned}$$

Thus also $\mathbb{E}_q\left[\max_{a\in A} u(a,\theta)\right] = 0$ for any belief q, which together with (5.8) implies that (5.9) reduces to $\Phi(a) = 0 - (-\Phi(a))$ and is thus true.

In the other direction,

$$\Phi(\delta_\theta) = \max_{a\in A} u(a,\theta) - \max_{a\in A} u(a,\theta) = 0 \quad \forall \theta \in \Theta$$

implying Property 6. Concavity of Φ (Property 7) follows by construction of Φ since $\mathbb{E}_q\left[\max_{a\in A} u(a,\theta)\right]$ is affine while $\sup_{a\in A} \mathbb{E}_q[u(a,\theta)]$ is a pointwise supremum of affine functions, and thus convex. ■

By Proposition 15, the two example cost functions from the previous section, C_{Ent} and C_{Var}, can be microfounded using decision problems. These decision problems are given below.

Example 5.11 (Microfoundation for Entropy Cost). Set $A = \Delta(\Theta)$ and $u(a,\theta) = \ln(a(\theta))$, where $\ln 0 = -\infty$. Then the cost of uncertainty is

$$\Phi_{\mathcal{D}}(q) = \mathbb{E}_q\left[\max_a\left[\ln(a(\theta))\right]\right] - \max_a \mathbb{E}_q\left[\ln(a(\theta))\right] = H(q).$$

Example 5.12 (Microfoundation for Variance Cost). Set $A = \Theta \subseteq \mathbb{R}$ and $u(a,\theta) = -(a-\theta)^2$. Then

$$\Phi_{\mathcal{D}}(q) = \mathbb{E}_q\left[\max_a\left[-(a-\theta)^2\right]\right] - \max_a \mathbb{E}_q\left[-(a-\theta)^2\right] = Var_q(\theta)$$

5.2.3 Posterior Separability

A weaker requirement than uniform posterior separability is that the cost of τ can be written in a way that is separable in the realized posteriors.

Definition 5.7 (Caplin and Dean (2013); Caplin, Dean and Leahy (2022)). *The cost function $C : S \to \mathbb{R}$ is* posterior separable *if*

$$C(p,\tau) = \mathbb{E}[\Phi_p(q)]$$

for some family of convex functions $(\Phi_p)_{p\in\Delta(\Theta)}$ where each $\Phi_p : \Delta(\Theta) \to \mathbb{R}$ is everywhere weakly positive, and $\Phi_p(p) = 0$ for every p.

Remark 5.7. When the cost function is posterior separable but not uniformly posterior separable, the cost of acquiring two signals in sequence may depend on the order in which these signals are acquired. This is not true for for UPS cost functions (Frankel and Kamenica, 2019; Bloedel and Zhong, 2021).

When the cost function is posterior separable, then the agent's payoff from choosing signal $\sigma : \Theta \to \Delta(S)$ and strategy $\alpha : S \to \Delta(A)$ is

$$\int_{\Delta(\Theta)}\int_{a\in A} \alpha(a \mid q)\mathbb{E}_q[u(a,\theta)]d\tau_\sigma(q) - C(p,\tau_\sigma),$$

and can be rewritten as

$$\int_{\Delta(\Theta)}\int_{a\in A} \alpha(a \mid q)\left(\mathbb{E}_q[u(a,\theta)] - \Phi_p(q)\right)d\tau_\sigma(q)$$

where the concave function $\mathbb{E}_q[u(a,\theta)] - \Phi_p(q)$ is the "net utility" of action a under posterior q. So maximizing the value function is equivalent to maximizing the expected net utility over all Bayes-plausible distributions and strategies, which is an optimization problem that can be solved using standard methods. This tractability is a part of the appeal of this family of cost functions.

A closely related concept appears in Frankel and Kamenica (2019), where $\Phi_p(q)$ is interpreted as the amount of information in news that moves an agent's belief from p to q. Frankel and Kamenica (2019) define the pair (Φ_p, Φ) as *coupled* if $\mathbb{E}[\Phi_p(q)] = \mathbb{E}[\Phi(p) - \Phi(q)]$, in which case the cost function is not only posterior separable but also uniformly posterior separable.

That uniform posterior separability is strictly stronger than posterior separability is nearly immediate, except for the requirement in the definition of posterior separable cost functions that $\Phi_p(q)$ is everywhere positive. We cannot therefore simply convert a UPS cost function $C(p,\tau) = \Phi(p) - \mathbb{E}_{q\sim\tau}[\Phi(q)]$ into a posterior separable cost function $C(p,\tau) = \mathbb{E}[\Phi_p(q)]$ by setting $\Phi_p(q) \equiv \Phi(p) - \Phi(q)$, as this quantity may be negative for some posterior beliefs q. The correct construction is instead to choose Φ_p to be a *Bregman divergence* of Φ (Frankel and Kamenica, 2019; Caplin, Dean and Leahy, 2022).

Definition 5.8. *Let* $\Phi : \Delta(\Theta) \to \mathbb{R}$ *be a concave function. A* supergradient *of* Φ *at* $p \in \Delta(\Theta)$ *is any vector* $\nabla\Phi(p)$ *such that*

$$\Phi(p) + \nabla\Phi(p) \cdot (q - p) \geq \Phi(q)$$

for every $q \in \Delta(\Theta)$.

Remark 5.8. When Φ is concave, then a supergradient $\nabla\Phi(q)$ exists for every q. When Φ is smooth at q, then $\nabla\Phi(q)$ is unique and equal to $\Phi'(q)$.

Definition 5.9. *Let* $\Phi : \Delta(\Theta) \to \mathbb{R}$ *be a concave function. A* Bregman divergence *of* Φ *is any map* $D_\Phi : \Delta(\Theta) \times \Delta(\Theta) \to \mathbb{R}$ *satisfying*

$$D_\Phi(p,q) = \Phi(p) - \Phi(q) + \nabla\Phi(p) \cdot (q - p) \quad \forall (p,q) \in \Delta(\Theta) \times \Delta(\Theta)$$

where $\nabla\Phi(q)$ *is a supergradient of* Φ *at* q.

This is the difference between the value of Φ at q and the value of the first-order Taylor expansion of Φ around p evaluated at point q.

Setting $\Phi_p(q) = D_\Phi(p,q)$ from Definition 5.9, we have

$$\Phi_p(q) = (\Phi(p) + \nabla\Phi(p) \cdot (q - p)) - \Phi(q) \geq 0 \quad \forall q \in \Delta(\Theta)$$

since $\nabla\Phi(p)$ is a supergradient of Φ, and also

$$\begin{aligned}\mathbb{E}_{q\sim\tau}[\Phi_p(q)] &= \mathbb{E}_{q\sim\tau}[\Phi(p) - \Phi(q) + \nabla\Phi(p) \cdot (q - p)] \\ &= \Phi(p) - \mathbb{E}_{q\sim\tau}[\Phi(q)]\end{aligned}$$

using in the second inequality that $\mathbb{E}_{q\sim\tau}(q - p) = 0$. The relationship between Φ_p and Φ is depicted in Figure 5.3.

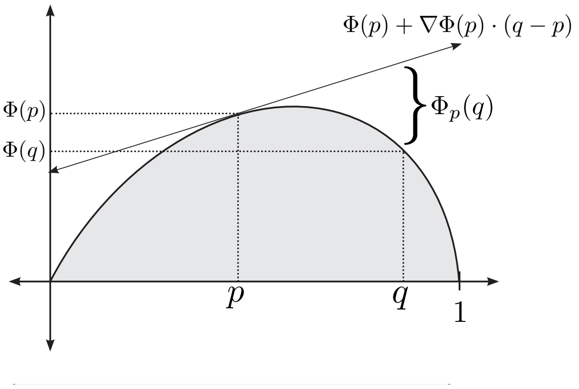

Figure 5.3: Relationship between Φ_p and Φ.

Example 5.13. Consider entropy cost $C_{\text{Ent}}(p, \tau) = H(p) - \mathbb{E}_{q\sim\tau}[H(q)]$. The Bregman divergence of entropy is KL divergence (Bregman, 1967), so

$$C_{\text{Ent}}(p, \tau) = H(p) - \mathbb{E}_{q\sim\tau}[H(q)] = \mathbb{E}[D(p\|q)].$$

Thus we can view the cost of a signal that generates the distribution of beliefs τ either as the expected reduction in the entropy of the agent's belief, or as the expected KL divergence from the agent's prior to the realized posterior belief.

5.3 Prior-Independent Costs

We now turn to cost functions that do not depend on the agent's prior belief. If the cost of information is exogenous to the agent—for example, a price determined within a market, or a physical cost of producing information—then we may expect the cost of acquiring information to be the same for all consumers regardless of their beliefs or expertise in the area, and thus prior independent.

One common cost specification is the following.

Example 5.14. In the setting of Example 5.2, let

$$C(\sigma_\varepsilon^2) = \frac{\kappa}{\sigma_\varepsilon^2} \tag{5.10}$$

Then the cost of the signal scales linearly with the precision of the signal, $1/\sigma_\varepsilon^2$. This formulation of the cost is especially sensible if we interpret θ as an unknown population parameter (for instance, the average height in a population) and the signal as a sample of individuals from this population. Modeling each observation as $X_i = \theta + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ independent of θ and independent across agents, the conditional distribution of θ given the sample $(X_1, \ldots, X_n)$ is the same as the conditional distribution of θ given the signal $X = \theta + \delta$, $\delta \sim \mathcal{N}(0, \sigma^2/n)$ (see Exercise 2.10). So (5.10) corresponds to a fixed cost of κ/σ^2 for each individual in the sample. This cost function is used in Wald's classic model of sequential sampling (Wald, 1945; Arrow, Blackwell and Girshick, 1949), and is a common modeling choice in continuous-time sequential sampling problems where the signal corresponds to observation of a Brownian motion (Fudenberg, Strack and Strzalecki, 2018; Liang, Mu and Syrgkanis, 2022).

We now present a generalization of the above cost function due to Pomatto, Strack and Tamuz (2020). Let Θ be a finite set and S be a set of signal realizations equipped with σ-algebra Σ, with $\Delta(S)$ denoting the set of measurable probability distributions on S. A signal is a mapping $\sigma : \Theta \to \Delta(S)$, and we use $\sigma_\theta \equiv \sigma(\cdot \mid \theta) \in \Delta(S)$ to denote the conditional distribution over signal realizations when the state is θ.

Definition 5.10. *The log-likelihood ratio between states θ and θ' at signal realization s is*

$$\ell^{\sigma}_{\theta,\theta'}(s) = \ln\left(\frac{d\sigma_\theta(s)}{d\sigma_{\theta'}(s)}\right)$$

Definition 5.11. *For any state $\theta \in \Theta$ and map $\alpha : \Theta \to \mathbb{N}$, define*

$$M^{\sigma}_{\theta}(\alpha) = \int_S \left| \prod_{\theta' \neq \theta} \left(\ell^{\sigma}_{\theta,\theta'}(s)\right)^{\alpha(\theta')} \right| d\sigma_\theta$$

Assumption 2. *The expectation $M^{\sigma}_{\theta}(\alpha)$ is finite for every θ and every $\alpha : \Theta \to \mathbb{N}$.*

This assumption says that the log-likelihood ratios have finite moments, ruling out for example the signal structure

$$\begin{array}{ccc} & s_1 & s_2 \\ \theta_1 & 0 & 1 \\ \theta_2 & \frac{1}{2} & \frac{1}{2} \end{array}$$

where the signal realization s_1 is perfectly revealing of the state θ_2.

Let $\mathcal{E}$ be the class of all signals satisfying Assumption 2. An *information cost function* is any map $C : \mathcal{E} \to [0, \infty)$. Pomatto, Strack and Tamuz (2020) propose four axioms that such a cost function should further satisfy.

Axiom 1 (Consistency with the Blackwell order)**.** *If σ dominates σ' in the Blackwell order, then $C(\sigma) \geq C(\sigma')$.*

That is, more informative signals are more costly to acquire.

Definition 5.12 (Combining Independent Signals). *For any two signals $\sigma : \Theta \to \Delta(S)$ and $\sigma' : \Theta \to \Delta(S')$, let $\sigma \otimes \sigma'$ denote the product signal*

$$\sigma \otimes \sigma' : \Theta \to \Delta(S \times S')$$

where $\sigma \otimes \sigma'(s, s' \mid \theta) = \sigma(s \mid \theta)\sigma(s' \mid \theta)$.

Axiom 2 (Additivity with respect to independent experiments). *For any two signals σ and σ', $C(\sigma \otimes \sigma') = C(\sigma) + C(\sigma')$.*

That is, the cost of acquiring two (conditionally) independent signals is equal to the sum of their costs. This axiom imposes a constant marginal cost on information similar to the one used to motivate Example 5.14.

Definition 5.13 (Diluting Signals). *For any signal σ, the α-*dilution *of σ, denoted $\alpha \cdot \sigma$, is a signal where with probability α the realization of σ is observed, and otherwise a completely uninformative signal is observed. Formally, $\alpha \cdot \sigma$ is a map from Θ to $S \cup \{\emptyset\}$ where the signal outcome $\emptyset$ has a constant $1 - \alpha$ probability at every state $\theta \in \Theta$, and the remaining probability is assigned to S in proportion to σ.*

Axiom 3 (Linearity in the "dilution" of the experiment). $C(\alpha \cdot \sigma) = \alpha \cdot C(\sigma)$ *for every signal σ and weight $\alpha \in [0, 1]$.*

That is, the cost of a signal is linear in the probability that it generates information.

Remark 5.9. Every posterior separable cost function $C(p, \tau) = \mathbb{E}_{q \sim \tau}[\Phi_p(q)]$ satisfies Axiom 3. To see this, observe that the distribution over posterior beliefs given the diluted signal $\alpha \cdot \sigma$, denoted $\tau_{\alpha \cdot \sigma}$, is the convex combination that puts weight α on the distribution τ_σ generated by σ, and weight $1 - \alpha$ on the prior. So

$$\begin{aligned} C(p, \tau_{\alpha \cdot \sigma}) &= \mathbb{E}_{q \sim \alpha \tau_\sigma + (1-\alpha)\delta_p}[\Phi_p(q)] \\ &= \alpha \mathbb{E}_{q \sim \tau_\sigma}[\Phi_p(q)] + (1 - \alpha)\Phi_p(p) \\ &= \alpha \cdot C(p, \tau_\sigma) \end{aligned}$$

where the second equality uses that C is affine in τ and the third uses that $\Phi_p(p) = 0$ in the definition of a posterior separable cost function.

The final axiom imposes continuity of the cost function with respect to a nonstandard (pseudo)-metric given below.[4]

Definition 5.14. *Given an upper bound $N \geq 1$, define*

$$d_N(\sigma, \sigma') = \max_{\theta \in \Theta} d_{TV}(\sigma_\theta, \sigma'_\theta) + \max_{\theta \in \Theta} \max_{\alpha \in \{0, \ldots, N\}^n} |M^\sigma_\theta(\alpha) - M^{\sigma'}_\theta(\alpha)|$$

where d_{TV} denotes the total variation distance.

Two signals σ and σ' are close under this pseudo-metric if for every state θ, the induced distributions of log-likelihood ratios are close in total-variation distance and additionally have similar moments, for any vector of moments lower or equal to $(N, \ldots, N)$.

Axiom 4 (Continuity.). *The function C is uniformly continuous with respect to d_N.*

[4] This is a pseudometric rather than a metric, since $d_N(\sigma, \sigma')$ is equal to zero for $\sigma \neq \sigma'$ if they induce the same distribution over posterior beliefs.

REMARK 5.10. The topology of weak convergence of likelihood ratios and the topology of convergence of likelihood ratios in total variation distance are both more standard. But no cost function which satisfies Axioms 1-3 is continuous in these alternative topologies. To see this, let θ be the unknown bias of a coin, and let σ_n be the signal where with probability $1/n$ the outcome of n independent flips of this coin is observed, and otherwise no information is revealed. Axioms 1-3 imply that $C(\sigma_n) = C(\sigma_{n'})$ for all finite n, n'. But the likelihood ratios of these signals converge in the weak topology (and in the total variation topology) to those of the signal that produces no information, and thus a stronger form of Axiom 4 based on either of these alternative topologies would require these signals to all have zero cost.

Proposition 16. *The cost function $C : \mathcal{E} \to \mathbb{R}$ satisfies Axioms 1-4 if and only if there exists a unique collection of $\mathbb{R}_+$-valued parameters $(\beta_{\theta,\theta'})_{\theta,\theta' \in \Theta}$ such that*

$$C(\sigma) = \sum_{\theta,\theta' \in \Theta} \beta_{\theta,\theta'} \times \underbrace{\int_S \ln \frac{d\sigma_\theta(s)}{d\sigma_{\theta'}(s)} d\sigma_\theta(s)}_{\textit{KL-divergence from } \sigma(\cdot \mid \theta') \textit{ to } \sigma(\cdot \mid \theta)} \tag{5.11}$$

As discussed in Section 5.1.2, the KL-divergence from $\sigma(\cdot \mid \theta')$ to $\sigma(\cdot \mid \theta)$ is a measure of how different the distributions are. The larger this divergence is, the easier it is to reject the hypothesis that the state is θ' when it truly is θ.

REMARK 5.11. Axiom 4 can be dispensed with if $\Theta = \{\theta_0, \theta_1\}$, in which case Proposition 16 simplifies to the statement that C satisfies Axioms 1-3 if and only if there exist parameters $\beta_{01}, \beta_{10} \geq 0$ such that

$$C(\sigma) = \beta_{01} D(\sigma(\cdot \mid \theta_0) \| \sigma(\cdot \mid \theta_1)) + \beta_{10} D(\sigma(\cdot \mid \theta_1) \| \sigma(\cdot \mid \theta_0)).$$

A notable contrast with entropy cost is that this cost function permits differentiation between states.

EXAMPLE 5.15 (Pomatto, Strack and Tamuz (2020)). Suppose the unknown state θ is the US GDP per capita, and the agent holds a uniform prior over $\Theta = \{20,000, \ldots, 80,000\}$. Then under entropy cost C_{Ent}, it is equally costly to acquire the signal that reveals whether θ is above or below \$50,000, or the signal that reveals whether θ is even or odd.

The free parameters $\beta_{\theta,\theta'}$ in the representation in (5.11) reflect potentially different costs to distinguishing between different pairs of states. Specifically, we can interpret each $\beta_{\theta,\theta'}$ as the marginal cost of increasing the expected log-likelihood ratio of a signal with respect to states θ and θ' (when θ is the true state). Thus in Example 5.15, we may specify (for example) that it is easier to distinguish between states that are far apart than those that are nearby, i.e., if GDP is in fact 80,000 then it is easier to rule out that GDP is 20,000 than it is to rule out that it is 79,999. In the special case where no pair of states is a priori harder to distinguish than another, then all coefficients are equal to one another.

Example 5.16. Returning to the setting of Example 5.2, where we now use $C(\sigma_\varepsilon^2)$ to mean the cost of acquiring the signal $X = \theta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, we have

$$C(\sigma_\varepsilon^2) = \sum_{\theta,\theta' \in \Theta} \beta_{\theta,\theta'} \frac{(\theta - \theta')^2}{2\sigma_\varepsilon^2}.$$

This nests the precision of the signal $(1/\sigma_\varepsilon^2)$ as a special case when $\beta_{\theta,\theta'} = \frac{1}{(\theta-\theta')^2}$, with the interpretation that states that are closer (in squared distance) are harder to distinguish.

Remark 5.12. The class of cost functions identified in Proposition 16 does not presuppose that the agent is Bayesian and has a prior belief over the state space. But if the agent does have a prior p, then the cost of the signal that induces distribution τ over posterior beliefs can be restated as

$$\mathbb{E}_{q \sim \tau}[\Phi_p(q)] \tag{5.12}$$

where

$$\Phi_p(q) = \Phi(p) - \sum_{\theta,\theta'} \beta_{\theta,\theta'} \frac{q_\theta}{p_\theta} \ln \left(\frac{q_\theta}{q_{\theta'}} \right) \tag{5.13}$$

so this family of cost functions belongs to the class of posterior-separable cost functions (Definition 5.7), although not to the class of uniform posterior separable cost functions (Definition 5.4).[5]

Exercise 5.5 (G). *Verify that (5.12) is equivalent to the original representation in (5.11) when Φ is defined according to (5.13).*

Hint 2. *Recall from Section 2.2 that the prior p and posterior q at signal realization s are related by* $\log\left(\frac{q(\theta)}{q(\theta')}\right) = \log\left(\frac{p(\theta)}{p(\theta')}\right) + \log\left(\frac{d\sigma_\theta}{d\sigma_{\theta'}}(s)\right)$.

5.4 Additional Exercises

Exercise 5.6 (G). *Suppose $p, q \in \Delta(\mathcal{X} \times \mathcal{Y})$ with p_X and q_X denoting the marginal distributions on $\mathcal{X}$, and $p_{Y|X}$ and $q_{Y|X}$ denoting the respective conditional distributions. Prove that*

$$D(p\|q) = D(p_X\|q_X) + D(p_{Y|X}\|q_{Y|X}).$$

This is known as the chain rule for KL divergence.

Exercise 5.7 (G). *Prove that the entropy cost function in Definiton 5.3 fails Pomatto, Strack and Tamuz (2020)'s Axiom 2.*

[5] Pomatto, Strack and Tamuz (2020) show that a generalization of the representation in (5.11), which permits the parameters $\beta_{\theta,\theta'}$ to depend on the prior, can accommodate uniformly posterior separable cost functions.