

Д. Сакрисон

Лекции
об аналоговой
связи



Van Nostrand Reinhold
Notes on System Sciences
General Editor: G. L. Turin

Notes on
ANALOG COMMUNICATION

by
D. J. SAKRISON
University of California, Berkeley

VAN NOSTRAND REINHOLD COMPANY
New York • Cincinnati • Toronto • London • Melbourne
1970

Д. Сакрисон

Лекции об аналоговой связи

Перевод с английского

В. Я. ЛЕВИТА и Р. З. ХАСЬМИНСКОГО

Под редакцией

Р. З. ХАСЬМИНСКОГО

ИЗДАТЕЛЬСТВО «МИР» Москва 1974



Книга посвящена вопросам теории передачи непрерывных сообщений и связанным с ними проблемам оценивания неизвестного параметра непрерывного случайного процесса. Особый интерес представляет обсуждение вопроса о границах снизу дисперсии оценки неизвестного параметра, основанной на наблюдении течения процесса. Наряду с известными из математической статистики границами Рао — Крамера, Баранкина и т. д. приводятся и принципиально другие границы, вытекающие из информационных соображений: так называемая граница плотной упаковки и граница, основанная на понятии ϵ -энтропии.

Кроме того, в книге на современном уровне излагаются известные результаты Котельникова о потенциальной помехоустойчивости при импульсной модуляции; обсуждается, в частности, пороговый эффект, а также система фазовой автоподстройки частоты.

Книга интересна студентам, аспирантам, инженерам и научным сотрудникам, специализирующимся в различных областях радиотехники, кибернетики и прикладной математики.

Редакция литературы по математическим наукам

ПРЕДИСЛОВИЕ РЕДАКТОРА ПЕРЕВОДА

Эта книга естественным образом дополняет изданные в 1972 г. в русском переводе «Лекции о цифровой связи» Дж. Турина. Если книга Турина посвящена оптимальным методам передачи по каналам связи случайных величин, принимающих конечное число значений (что с математической точки зрения приводит к задаче проверки гипотез), то предлагаемая вниманию читателя книга посвящена передаче параметров и случайных величин, возможными значениями которых служит целый интервал числовой оси, а также передаче случайных процессов. С точки зрения математической статистики это задачи теории оценивания. Поэтому большое место в книге занимает изучение свойств различных оценок, построенных по наблюдениям. Наряду с хорошо известными границами снизу дисперсии оценок типа неравенства Рао — Крамера приводятся и некоторые новые границы, основанные на соображениях теории информации.

Автор старается обходиться по возможности сравнительно простым математическим аппаратом. Это приводит к тому, что многие места в книге не удовлетворяют канонам математической строгости. Однако этот недостаток искупается доступностью изложения и тем, что автор всегда подробно останавливается на идеях, лежащих в основе рассмотрений. К тому же в большей части книги математик будет в состоянии сам уточнить рассуждения, сделав их строгими.

К сожалению, последнее замечание не относится к гл. 4 и 5. Однако эти главы, быть может, будут

наиболее интересны математику, так как они содержат краткое, но яркое изложение некоторых важных результатов теории связи (например, теорию потенциальной помехоустойчивости В. А. Котельникова), не имеющих пока четкого математического обоснования. Математик может рассматривать значительную часть материала этих глав как интересные и актуальные задачи, которые ждут своего решения.

Книга написана с большим педагогическим мастерством; во многих случаях автору удается, начав с «азов», довести изложение до современного состояния проблемы.

В русском переводе устранены некоторые опечатки; в ряде случаев ссылки на американские учебники заменены ссылками на источники, более доступные советскому читателю. Добавлено несколько примечаний.

Можно с уверенностью сказать, что прекрасно написанная и богатая содержанием книга профессора Сакрисона будет интересна и полезна широкому кругу математиков и инженеров.

Р. Хасьминский

*С надеждой на дальнейшее
развитие сотрудничества между
советскими и американскими
учеными.*

ПРЕДИСЛОВИЕ К РУССКОМУ ИЗДАНИЮ

Мне очень приятно и лестно, что мои советские коллеги сочли уместным перевод этих лекций на русский язык, и я благодарен Б. Левиту и Р. Хасьминскому за их перевод и редактирование.

Эта книга предназначена для студентов, специализирующихся в электро- и радиотехнике, и по замыслу должна охватывать большую часть основных понятий, относящихся к приему и передаче непрерывной информации. Довольно полное описание рассматриваемых в ней вопросов дано в гл. 1, и поэтому их обсуждение здесь излишне.

Представляется интересным только указать, как сейчас, по прошествии пяти лет с тех пор, когда книга писалась по-английски, распределяются усилия исследователей и каким представляется будущее отдельных отраслей.

Проблемы, обсуждаемые в гл. 2, посвящены оценке неизвестных параметров и составляют классический раздел математической статистики. Исследования в этой области сейчас уже не столь интенсивны.

В гл. 3 и 4 описаны применения общей теории к случаю, когда наблюдение представляет собой выборочную функцию случайного процесса. В этой области остались открытыми еще некоторые интересные, но трудные проблемы, относящиеся к оптимальному выбору вида передаваемых сигналов и их зависимости от параметров, несущих информацию. Новые результаты здесь появ-

ляются медленно, что объясняется трудностью задач. По-видимому, разработка недорогих устройств, которые дадут возможность осуществлять сложные процессы обработки цифровой информации, обусловит развитие непрерывной импульсной модуляции, и это позволит полностью перейти к импульсно-кодовой модуляции (ИКМ). Мне кажется, что содержание гл. 4 в наибольшей мере способствует восприятию идей, лежащих в основе этой проблемы. Здесь уместно также отметить непосредственное влияние на материал гл. 4 классической работы В. А. Котельникова.

Глава 5, посвященная угловой модуляции, имеет важные практические приложения, однако возможности для дальнейших теоретических исследований здесь кажутся ограниченными. В гл. 6 на основе понятия функции скорости при заданном искажении (ϵ -энтропии) описан теоретический подход к проблеме кодирования источников. Я считаю, что из всех вопросов, рассматриваемых в этих лекциях, именно материал гл. 6 открывает наибольшие возможности для исследователей. В частности, представляется важным вычисление функции скорости при заданном искажении для таких распределений источников и критериев искажения, которые отвечали бы реальным интересующим нас каналам, а также построение практических методов универсального кодирования, т. е. методов, приводящих к близким к оптимальным характеристикам для широкого класса распределений источника.

Необходимые при чтении этих лекций математические знания отвечают обычной подготовке электро- и радиоинженера. Некоторые более тонкие математические понятия, такие, как производная Радона — Никодима, используются неформально, на интуитивном уровне (хотя по существу корректно). Я надеюсь, что перевод этих лекций окажется интересным и полезным для советских исследователей в области теории информации и связи.

Д. Дж. Сакрисон

ОТ АВТОРА

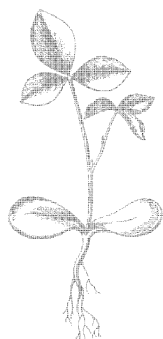
Обычно предисловия пишутся для преподавателей, которым данная книга может оказаться полезной, и в них делаются попытки описать основные идеи и методы, а также приводятся соображения, связанные с тематикой и отбором материала. Для тех же читателей, которые должны учиться по книге, аналогичные вопросы обычно излагаются в первой главе.

Настоящие лекции адресованы читателям, которые имеют подготовку в объеме технических вузов. Такие читатели, как правило, заинтересованы в актуальности рассматриваемых вопросов и глубине их изложения в той же мере, как и преподаватели. Поэтому обсуждение содержания этой книги мы отложим до первой главы.

Здесь мне хочется поблагодарить всех тех лиц, сотрудничество с которыми оставило след в этой книге, особенно своих аспирантов. Я весьма признателен М. Закаи за критический просмотр текста — его замечания позволили увидеть ряд вопросов, которые я упустил из виду или недостаточно ясно представлял, а также дали толчок к пересмотру ряда мест, казавшихся мне несомненными. Я обязан также Дж. Турину за его замечания и лестное для меня предложение опубликовать книгу в этой серии.

Д. Дж. Сакрисон

Кембридж,
Англия



Глава 1

ВВЕДЕНИЕ

Предметом настоящих лекций является проблема передачи (в том случае когда в нашем распоряжении имеется канал связи) и получения (например, с помощью радиолокационных импульсов) информации при наличии случайных шумов или помех. И в том, и в другом случае мы будем рассматривать источники информации, у которых в отличие от цифровых источников сообщения или параметры, несущие информацию, имеют непрерывный (аналоговый) характер. Темы, охватываемые различными разделами этой книги, кажутся весьма далекими, но тем не менее они посвящены одной центральной проблеме: как с помощью сигналов, доступных наблюдению, кодировать или модулировать параметры, несущие информацию, и каким образом, а также с какой точностью можно оценивать эти параметры по принятым сигналам при наличии шумов.

В гл. 2 рассматриваются основные понятия теории оценивания. Основываясь на наблюдаемой случайной величине Z , распределение которой зависит от некоторого параметра α , мы изучим различные «хорошие», или практически целесообразные, методы построения оценок величины α по наблюдению Z и получим некоторые границы точности, присущей таким процедурам оценивания. Для простоты мы будем исследовать случай, когда наблюдение Z и параметр α являются конечномерными векторами, однако изложение построено так, что нетрудно получить дальнейшие обобщения.

В гл. 3 задача оценивания рассматривается в ситуации, когда наблюдается случайный процесс

$$Z(t) = s(t, \alpha) + \mathcal{N}(t), \quad (1.1)$$

где $s(t, \alpha)$ — сигнал, несущий информацию, а $\mathcal{N}(t)$ — гауссовский шум. Здесь интересующий нас параметр α

по-прежнему представляет собой конечномерный вектор. Применительно к этой ситуации уточняются понятия и методы гл. 2. Особое внимание мы уделим задаче радиолокационного измерения дальности, чтобы пояснить понятие функции неопределенности и привести пример использования неравенства Крамера — Рао.

Глава 4 также посвящена проблеме оценки параметра α по наблюдению вида (1.1), однако в этой главе внимание концентрируется не на построении методов оценивания, а на том, как следует кодировать или модулировать параметр α с помощью сигнала $s(t, \alpha)$. В ситуации радиолокационного обнаружения характер модуляции не зависит от воли инженера-связиста, поскольку дальность и скорость цели определяют соответственно задержку и сдвиг частоты возвращаемого сигнала. Однако при построении систем импульсной модуляции инженер, проектирующий линию связи, свободен в выборе вида сигналов $s(t, \alpha)$, зависящих от параметра α . Вопрос о внутренней точности, присущей задаче оценивания параметра α при оптимальной модуляции, т. е. при оптимальном выборе семейства сигналов $s(t, \alpha)$, как раз и изучается в гл. 4. В частности, подробно обсуждается вопрос о том, почему с помощью нелинейной модуляции можно получить выигрыш в отношении сигнал/шум на выходе канала в обмен на расширение полосы частот и почему такой обмен сопровождается пороговым эффектом.

В гл. 5 кратко описывается угловая модуляция, которая на практике применяется в тех случаях, когда модулируются не параметры или их последовательности, а временные сигналы. В этой главе мы ограничимся нахождением структуры оптимального метода оценивания или оптимального приемника.

Вплоть до гл. 6 предмет настоящей книги рассматривается с точки зрения теоретика- или инженера-связиста: приняв определенную модель искажений (аддитивный гауссовский шум), мы уделяем внимание в основном тому, каким образом изменение вида модуляции влияет на качество передачи. В гл. 6 излагается подход к передаче сообщений в рамках теории информации. Точнее, мы рассмотрим канал связи с произвольным ви-

дом искажений и источник сообщений, порождающий последовательность независимых выходных символов, имеющих одинаковое, но произвольное распределение. Мы будем считать, что канал связи характеризуется своей теоретико-информационной пропускной способностью. На основе некоторого критерия точности мы припишем числовую характеристику величине расхождения между символом, поступившим на выход источника сообщений, и сообщением, вырабатываемым приемником. Это естественным образом приводит к следующему основному вопросу: каково минимальное расхождение, допускаемое передачей по заданному каналу, если источник сообщений связан с приемником посредством произвольного кодирующего прибора, заданного канала и произвольного декодирующего устройства? Ответ на этот вопрос дает теорема Шеннона о кодировании источников, использующая понятие границы скорости передачи при заданном искажении; эта теорема достаточно подробно излагается в гл. 6.

Просмотрев оглавление или перелистав книгу, читатель заметит, что одним вопросам уделено достаточно много места, тогда как другие изложены весьма кратко. В соответствии с замыслом этой книги, рассчитанной на достаточно подготовленного читателя, при определении объема изложения различных тем мы исходили из двух критериев. Во-первых, мы старались вести изложение так, чтобы можно было вводить новые понятия и методы, не вдаваясь в частные результаты ради них самих. Во-вторых, мы избегали детального рассмотрения тех вопросов, которые достаточно подробно освещаются в имеющихся учебниках.

Несколько заключительных замечаний: в дальнейшем прописные буквы всюду используются для обозначения случайных величин, а строчные — для их значений. Так, мы будем говорить о вероятности того, что случайная величина Z меньше данного значения z . Скалярное произведение двух векторов \mathbf{u} и \mathbf{v} , принадлежащих n -мерному евклидову пространству, мы будем обозначать

$$(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^n u_i v_i.$$

Глава 2

ОЦЕНКА ПАРАМЕТРОВ

В настоящей главе мы рассмотрим следующую задачу. Имеется некоторое наблюдение z , статистически связанное в силу физической природы с некоторым ненаблюдаемым параметром α , значение которого мы хотим оценить. Известный пример такого рода — задача радиолокационного обнаружения, где z представляет собой отраженный от цели сигнал с наложенным шумом, по которому необходимо оценить дальность или скорость цели.

Начнем с обозначений. Предположим, что наблюдение является векторной случайной величиной \mathbf{Z} , допустимые частные значения которой будут обозначаться через \mathbf{z} . Впоследствии мы рассмотрим общую ситуацию, в которой наблюдение будет выборочной функцией случайного процесса; пока же будем считать, что вектор \mathbf{z} конечномерен.

Параметр, подлежащий оценке, будем обозначать через α , если это число (например, дальность цели), и через $\mathbf{\alpha}$, если это вектор (например, двумерный вектор дальности и скорости). Мы не будем проводить различия между случайной величиной α и значением, принятым ею, попросту по той причине, что прописные греческие буквы не слишком употребительны; в то же время строчные греческие буквы довольно широко используются для обозначения оцениваемых параметров.

Обозначим через $f_{\alpha}(\mathbf{z})$ условную плотность распределения величины \mathbf{Z} при заданном значении α , а через $f(\alpha)$ — плотность априорного распределения параметра α , если оно определено.

Наша цель заключается в построении и исследовании свойств оценок величины α , которые мы будем обозначать через $\hat{\alpha}$. Заметим во избежание путаницы, что

термин *оценка* используется для обозначения двух различных, хотя и связанных между собой объектов:

- 1) $\hat{\alpha}$ есть функция, ставящая в соответствие каждому наблюдаемому значению z некоторое значение $\hat{\alpha} = \hat{\alpha}(z)$ параметра α ; с этой точки зрения построение оценки состоит в указании вида функциональной зависимости;
- 2) выбранная функция порождает некоторую случайную величину $\hat{\alpha} = \hat{\alpha}(Z)$; качество оценки (как функции) определяется близостью случайной величины $\hat{\alpha}$ к оцениваемому значению α .

В дальнейшем, желая подчеркнуть, что речь идет о той роли, которую оценка $\hat{\alpha}$ играет как функция (в противоположность ее роли в значении случайной величины), мы будем иногда говорить о *методе оценивания*.

При изучении задачи оценивания естественным образом выделяются следующие два случая: ситуации, когда параметр α является случайной величиной, которой некоторым разумным образом можно приписать плотность априорного распределения, и ситуации, когда говорить об α как о случайной величине с некоторым априорным распределением не имеет смысла.

2.1. ОЦЕНКА СЛУЧАЙНЫХ ПАРАМЕТРОВ

Байесовские риски и оценки

Поскольку наша цель состоит в построении оценок $\hat{\alpha}(\cdot)$, для которых случайная величина $\hat{\alpha}(Z)$ близка к параметру α , мы должны прежде всего выбрать меру такой «близости». Начнем с рассмотрения функций потерь, приписывающих размер потерь ошибкам $\alpha - \hat{\alpha}$. Несколько примеров таких функций потерь изображены на рис. 2.1.

Функция потерь $c(\cdot)$ называется *симметричной*, если

$$c(e) = c(-e), \quad (2.1)$$

и *выпуклой*, если для произвольных e_1, e_2

$$c[\theta e_1 + (1 - \theta) e_2] \leq \theta c(e_1) + (1 - \theta) c(e_2), \quad (2.2)$$

где $0 \leq \theta \leq 1$.

Рис. 2.1, в графически иллюстрирует понятие выпуклости. Отметим, что функция

$$c_p(e) = |e|^p$$

выпукла при любом $p \geq 1$.

Определим *средний риск*, или *среднюю величину потерь*, как математическое ожидание величины $c(\alpha - \hat{\alpha})$:

$$\mathcal{E}_c = \mathbf{E} \{c(\alpha - \hat{\alpha})\} = \int_{-\infty}^{\infty} dz f(z) \int_{-\infty}^{\infty} d\alpha c[\alpha - \hat{\alpha}(z)] f(\alpha | z), \quad (2.3)$$

где совместная плотность $f(\alpha, z)$ представлена в виде произведения двух сомножителей $f(\alpha | z)f(z)$. Величину

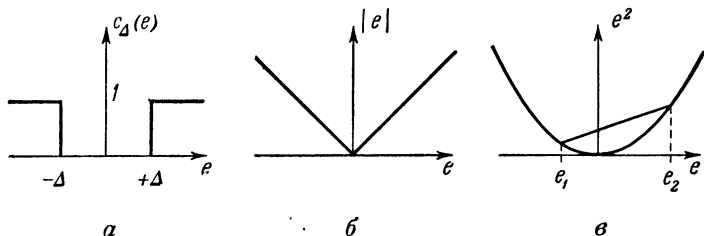


Рис. 2.1. Примеры весовых функций ошибок.

(2.3) иногда называют также *байесовским риском* или *байесовскими потерями*. Индекс c указывает на зависимость этой величины от функции потерь. В случае $c(e) = |e|^p$ мы будем использовать обозначение \mathcal{E}_p ; если же $c(e) = c_\Delta(e)$ (см. рис. 2.1, а), то средний риск будем обозначать \mathcal{E}_Δ . Внутренний интеграл в (2.3) иногда называют *условным риском* и обозначают $\mathcal{E}_c(z)$. В дальнейшем, не ограничивая общности, мы будем считать, что рассматриваемые нами функции потерь неотрицательны, т. е. $c(\cdot) \geq 0$.

Оптимальные оценки, минимизирующие байесовский риск, называют *байесовскими оценками*. Аналитическое выражение для байесовских оценок можно найти лишь в сравнительно немногих случаях, примеры которых мы сейчас рассмотрим. Сначала сделаем следующее общее замечание: поскольку плотности вероятностей и функции потерь, входящие в (2.3), неотрицательны, то средний

риск \mathcal{E}_c будет минимальным в том случае, когда оценка $\hat{\alpha}(\cdot)$ минимизирует условный риск $\mathcal{E}_c(\mathbf{z})$ при каждом \mathbf{z} .

Пример 2.1. Среднеквадратическая ошибка. В этом примере $c(e) = e^2$ и условный риск

$$\mathcal{E}_2(\mathbf{z}) = \int_{-\infty}^{\infty} [\alpha - \hat{\alpha}(\mathbf{z})]^2 f(\alpha | \mathbf{z}) d\alpha \quad (2.4)$$

минимален, если в качестве $\hat{\alpha}(\mathbf{z})$ выбрано *условное математическое ожидание*

$$\hat{\alpha}(\mathbf{z}) = E\{\alpha | \mathbf{z}\} = \int_{-\infty}^{\infty} \alpha f(\alpha | \mathbf{z}) d\alpha. \quad (2.5)$$

Пример 2.2. Средняя по модулю ошибка. Для функции потерь $c(e) = |e|$ условный риск

$$\mathcal{E}_1(\mathbf{z}) = \int_{-\infty}^{\infty} |\alpha - \hat{\alpha}(\mathbf{z})| f(\alpha | \mathbf{z}) d\alpha \quad (2.6)$$

минимален, если в качестве $\hat{\alpha}$ выбрано любое решение уравнения

$$\int_{-\infty}^{\hat{\alpha}(\mathbf{z})} f(\alpha | \mathbf{z}) d\alpha = \int_{\hat{\alpha}(\mathbf{z})}^{\infty} f(\alpha | \mathbf{z}) d\alpha. \quad (2.7)$$

Это решение, называемое *условной медианой*, однозначно определено, кроме случая, когда $f(\alpha | \mathbf{z}) = 0$ на некотором интервале, содержащем одно из решений $\hat{\alpha}(\mathbf{z})$. В этом случае в качестве $\hat{\alpha}(\mathbf{z})$ можно выбрать любое число из связного интервала, на котором $f(\alpha | \mathbf{z}) = 0$.

Задача 2.1. Доказать, что соотношения (2.5) и (2.7) определяют соответствующие байесовские оценки для \mathcal{E}_2 и \mathcal{E}_1 .

Пример 2.3. Прямоугольная функция потерь. Для функции потерь, изображенной на рис. 2.1, а, минимизация условного риска

$$\mathcal{E}_\Delta(\mathbf{z}) = \int_{|\alpha - \hat{\alpha}(\mathbf{z})| > \Delta} f(\alpha | \mathbf{z}) d\alpha, \quad (2.8)$$

очевидно, эквивалентна выбору значения $\hat{\alpha}(\mathbf{z})$, максимизирующего величину

$$1 - \mathcal{E}_{\Delta}(\mathbf{z}) = \mathbf{P}\{|\alpha - \hat{\alpha}| < \Delta | \mathbf{z}\} = \int_{\hat{\alpha}(\mathbf{z}) - \Delta}^{\hat{\alpha}(\mathbf{z}) + \Delta} f(\alpha | \mathbf{z}) d\alpha. \quad (2.9)$$

В общем случае нельзя указать явного способа построения $\hat{\alpha}$. Однако в следующих двух конкретных ситуациях, представляющих практический интерес, для нахождения $\hat{\alpha}$ существует простое правило. Предположим, что

- i) $f(\alpha | \mathbf{z})$ как функция от α *унимодальна* (т. е. имеет единственный максимум) и симметрична относительно своей моды

либо

- ii) $f(\alpha | \mathbf{z})$ как функция от α мало меняется на интервалах длины 2Δ .

В первом случае оценка $\hat{\alpha}(\mathbf{z})$ совпадает с условной модой, т. е. со значением α , для которого величина $f(\alpha | \mathbf{z})$ максимальна; во втором случае совпадение приближенное. Условную моду часто называют *оценкой максимума апостериорной вероятности* (МAB).

Устойчивость оценок по отношению к функциям потерь

Задача нахождения оптимальных (байесовских) оценок сводится, как мы видели, к отысканию правила, быть может приближенного, позволяющего сопоставить с каждым значением \mathbf{z} величину $\hat{\alpha}(\mathbf{z})$. В большинстве случаев, отличных от рассмотренных нами выше, это весьма трудно. Мы приведем сейчас две теоремы, интересные тем, что они показывают, что при определенных условиях оптимальная оценка не зависит от выбора функции потерь. Практически это означает, что для интересующей нас задачи байесовскую оценку можно найти путем определения байесовской оценки для другой, аналитически более простой функции потерь.

Теорема 2.1. Пусть $\bar{\alpha}(\mathbf{z})$ — условное математическое ожидание величины α при заданном значении $\mathbf{Z} = \mathbf{z}$, определяемое правой частью равенства (2.5). Если при всех \mathbf{z} функция $f(\alpha | \mathbf{z})$ симметрична относительно $\bar{\alpha}(\mathbf{z})$, то для любой симметричной выпуклой функции потерь $c(\cdot)$ байесовский риск \mathcal{E}_c минимален при $\hat{\alpha}(\mathbf{z}) = \bar{\alpha}(\mathbf{z})$.

Доказательство этой теоремы см. в работе Ван Триса [1, п. 2.4.1]. Отметим, что при выполнении условий теоремы 2.1 все три величины — условное математическое ожидание, условная мода, условная медиана — равны между собой. В следующей теореме мы ослабим ограничения на функцию потерь за счет дополнительного сужения класса функций $f(\alpha|z)$.

Теорема 2.2. Если функция $f(\alpha|z)$ унимодальна, симметрична и для любого z

$$\lim_{\alpha \rightarrow \infty} c(\alpha) f(\alpha|z) = 0, \quad (2.10)$$

то для любой симметричной функции потерь, монотонной в том смысле, что

$$c(e') \geq c(e) \text{ при } e' \geq e \geq 0, \quad (2.11)$$

байесовский риск \mathcal{E}_c минимален при $\hat{\alpha}(z) = \bar{\alpha}(z)$.

Доказательство (см. Витерби [2, стр. 376]). Так как наши рассуждения применимы к условной плотности $f(\alpha|z)$ при произвольном значении z , мы для удобства опустим переменную z . Полагая $\alpha^* = \alpha - \bar{\alpha}$, запишем условный риск при каком-либо значении оценки $\hat{\alpha}$ в виде

$$\begin{aligned} \mathcal{E}_c(\hat{\alpha}) &= \int_{-\infty}^{\infty} c(\alpha - \hat{\alpha}) f(\alpha) d\alpha = \int_{-\infty}^{\infty} c(\alpha^* - \hat{\alpha} + \bar{\alpha}) f^*(\alpha^*) d\alpha^* = \\ &= \int_{-\infty}^{\infty} c(w) f^*(w + \hat{\alpha} - \bar{\alpha}) dw, \end{aligned}$$

где через $f^*(\alpha^*) = f(\alpha^* + \bar{\alpha})$ обозначена центрированная условная плотность. Используя симметрию функций c и f^* , находим

$$\begin{aligned} \mathcal{E}_c(\hat{\alpha}) - \mathcal{E}_c(\bar{\alpha}) &= -2 \int_0^{\infty} c(w) f^*(w) dw + \\ &+ \int_0^{\infty} c(w) f^*(w + \hat{\alpha} - \bar{\alpha}) dw + \int_{-\infty}^0 c(w) f^*(w + \hat{\alpha} - \bar{\alpha}) dw = \\ &= \int_0^{\infty} c(w) [f^*(w + \hat{\alpha} - \bar{\alpha}) + f^*(w - \hat{\alpha} + \bar{\alpha}) - 2f^*(w)] dw. \quad (2.12) \end{aligned}$$

Обозначим

$$v = c(w),$$

$$du = [f^*(w + \hat{\alpha} - \bar{\alpha}) + f^*(w - \hat{\alpha} + \bar{\alpha}) - 2f^*(w)] dw$$

и проинтегрируем выражение справа по частям. Считая, что $\hat{\alpha} > \bar{\alpha}$, получаем при всех $w \geq 0$

$$\begin{aligned} u(w) &= - \int_{-\bar{\alpha}}^0 f^*(y) dy - \int_0^w f^*(y) dy + \\ &\quad + \int_{-(\hat{\alpha}-\bar{\alpha})}^{-(\hat{\alpha}-\bar{\alpha})+w} f^*(y) dy + \int_{\hat{\alpha}-\bar{\alpha}}^{\hat{\alpha}-\bar{\alpha}+w} f^*(y) dy = \\ &= - \int_{-\bar{\alpha}}^w f^*(y) dy + \int_{\hat{\alpha}-\bar{\alpha}-w}^{\hat{\alpha}-\bar{\alpha}+w} f^*(y) dy = \\ &= - \int_{-w+(\hat{\alpha}-\bar{\alpha})}^{-w} f^*(y) dy + \int_w^{w+(\hat{\alpha}-\bar{\alpha})} f^*(y) dy = \\ &= \int_0^{\hat{\alpha}-\bar{\alpha}} [f^*(y+w) - f^*(y-w)] dy \leq 0. \end{aligned} \quad (2.13)$$

Смысл промежуточных выкладок в (2.13) проще всего уяснить, изобразив графически промежутки интегрирования; последнее неравенство следует из монотонности функции f .

Объединяя (2.10), (2.12) и (2.13), получаем

$$\mathcal{E}_c(\hat{\alpha}) - \mathcal{E}_c(\bar{\alpha}) = \lim_{x \rightarrow \infty} c(w) u(w) \Big|_0^x - \int_0^\infty dw \frac{dc(w)}{dw} u(w) \geq 0, \quad (2.14)$$

так как $dc(w)/dw \geq 0$ в силу монотонности функции c ¹⁾. Заметим, что из условия (2.10) и вида функции $u(w)$

¹⁾ Если функция $c(\cdot)$ недифференцируема, то интеграл в (2.14) можно рассматривать как интеграл Стильтьеса. Тогда из монотонности функции $c(w)$ и неравенства $u(w) \leq 0$ следует, что этот интеграл неположителен. При этом формула интегрирования по частям по-прежнему верна.

в (2.13) вытекает, что $\lim_{\omega \rightarrow \infty} c(\omega) u(\omega) = 0$. В силу симметрии те же рассуждения применимы и при $\hat{\alpha} - \alpha < 0$.

Задача 2.2. Пусть α — скалярная случайная величина, а \mathbf{Z} обозначает n -мерный случайный вектор, причем совместное распределение α и \mathbf{Z} гауссовское. Предположим, что α и \mathbf{Z} имеют нулевое математическое ожидание и матрицу ковариаций

$$R_{\alpha\mathbf{Z}} = \begin{matrix} & Z_1 & \dots & Z_n & \alpha \\ \begin{matrix} Z_1 \\ \vdots \\ Z_n \\ \alpha \end{matrix} & \begin{bmatrix} & & & \\ & R_{\mathbf{Z}} & & \mathbf{r}] \\ & & & \\ \alpha & \underline{\mathbf{r}} & & \sigma_{\alpha}^2 \end{bmatrix} \end{matrix},$$

где $R_{\mathbf{Z}}$ есть $(n \times n)$ -матрица, $\underline{\mathbf{r}}$ — вектор-строка, $\mathbf{r}]$ — вектор-столбец, σ_{α}^2 — скаляр.

Обозначим через Q матрицу, обратную к $R_{\alpha\mathbf{Z}}$:

$$Q = \begin{matrix} & Z_1 & \dots & Z_n & \alpha \\ \begin{matrix} Z_1 \\ \vdots \\ Z_n \\ \alpha \end{matrix} & \begin{bmatrix} & & & \\ & Q_{\mathbf{Z}} & & \mathbf{q}] \\ & & & \\ \alpha & \underline{\mathbf{q}} & & q_{\alpha} \end{bmatrix} \end{matrix},$$

где $Q_{\mathbf{Z}}$ есть $(n \times n)$ -матрица.

Показать, что $E\{\alpha|\mathbf{Z}\}$ — линейная функция от \mathbf{Z} . При умелом использовании матричных обозначений и вида матрицы Q эта задача не требует громоздких вычислений.

2.2. ОЦЕНИВАНИЕ НЕСЛУЧАЙНЫХ ПАРАМЕТРОВ

Существуют различные ситуации (такие, как измерение радиолокатором скорости вращения планеты), когда нет смысла предполагать, что параметр α выбирается

случайным образом из некоторого множества, на котором можно разумным способом задать плотность априорного распределения $f(\alpha)$. Этот случай, который мы исследуем в настоящем разделе, логически не столь прост, как предыдущий. Трудность состоит в том, что нелегко непосредственно определить, что мы будем понимать под наилучшей оценкой. Чтобы обойти эту трудность, выберем следующий путь. Сначала, опираясь на интуитивные соображения, определим некоторую оценку специального вида — так называемую *оценку максимального правдоподобия*. Затем получим границу качества оценивания для *произвольной* оценки α . Наконец, рассмотрим ряд «хороших» свойств, которыми могут обладать те или иные оценки, и покажем, что если существует оценка, обладающая этими свойствами, то она совпадает с оценкой максимального правдоподобия.

Оценка максимального правдоподобия

Чтобы объяснить, почему мы уделяем оценке максимального правдоподобия особое внимание, заметим следующее.

i) Как мы видели, при некоторых слабых ограничениях на плотность $f(\alpha|z)$ для широкого класса функций потерь оптимальной оценкой является оценка максимума апостериорной вероятности (МАВ); другими словами, это оценка, принимающая то значение α , при котором величина

$$f(\alpha|z) = \frac{f(z|\alpha)f(\alpha)}{f(z)} \quad (2.15)$$

максимальна.

ii) Поскольку $f(z)$ в последнем равенстве не зависит от α , оценка МАВ максимизирует $f(z|\alpha)f(\alpha)$. Если же предположить, что наблюдение Z заметно увеличивает надежность оценивания α , то функция $f(z|\alpha)f(\alpha)$ должна иметь гораздо более выраженный пик вблизи своей моды, чем $f(\alpha)$. Следовательно, в этом случае максимум (относительно α) функции $f(z|\alpha)f(\alpha)$ расположен вблизи максимума функции $f(z|\alpha)$.

Таким образом, в качестве оценки разумно выбрать то значение α , которое максимизирует $f(z|\alpha)$; эту оценку

мы и будем называть оценкой максимального правдоподобия (МП). Желая подчеркнуть, что параметр α здесь не предполагается случайным, мы будем для условной плотности $f(\mathbf{z}|\alpha)$ использовать в дальнейшем обозначение $f_\alpha(\mathbf{z})$. Это функция, которая при каждом фиксированном α является вероятностной плотностью по \mathbf{z} . Рассматриваемая же как функция от α при фиксированном \mathbf{z} , она часто называется *функцией правдоподобия*.

В нашем рассуждении мы лишь для простоты считали, что α — скаляр; все замечания, относящиеся к оценке максимального правдоподобия, в равной мере применимы и к векторному параметру α .

Свойства оценок

Определим свойства рассматриваемых статистических оценок. Для ясности изложения будем различать следующие значения α :

α' — фактическое (истинное) значение параметра,

α — произвольное значение параметра,

$\hat{\alpha}$ — значение оценки.

Параметр α является векторным, однако его размерность не связана с размерностью вектора \mathbf{z} . Мы будем обозначать через $E_\alpha \{ \}$ математическое ожидание величины, стоящей в фигурных скобках, относительно распределения вероятностей $dP_\alpha(\mathbf{z}) = f_\alpha(\mathbf{z}) d\mathbf{z}$.

Будем называть оценку $\hat{\alpha}$ *несмещенной*, если

$$E_\alpha \{ \hat{\alpha}(\mathbf{Z}) \} = \alpha \quad \text{для всех } \alpha \in \mathcal{A}, \quad (2.16)$$

где \mathcal{A} — множество тех значений α , которые считаются априори возможными.

Пусть \mathbf{Z}_k , $k = 1, 2, \dots$, — последовательность независимых одинаково распределенных случайных величин. Предположим, что задана последовательность оценок

$$\hat{\alpha}_n(\mathbf{z}_1, \dots, \mathbf{z}_n), \quad n = 1, 2, \dots$$

Будем называть оценку $\hat{\alpha}_n$ *состоятельной*, если $\hat{\alpha}_n(\mathbf{z}_1, \dots, \mathbf{z}_n)$ сходится по вероятности к α' при $n \rightarrow \infty$.

Если существует такая функция $\tilde{\alpha}(\mathbf{z})$, что $f_{\alpha}(\mathbf{z})$ можно представить в виде

$$f_{\alpha}(\mathbf{z}) = g(\tilde{\alpha}, \alpha) w(\mathbf{z}), \quad w(\mathbf{z}) \geq 0, \quad (2.17)$$

то $\tilde{\alpha}(\mathbf{z})$ называется *достаточной статистикой* для оценки параметра α . Важность этого понятия объясняет

Теорема 2.3. Если $\tilde{\alpha}(\mathbf{z})$ — достаточная статистика для параметра α , то при любом выборе функции потерь байесовская оценка зависит лишь от $\alpha(\mathbf{z})$.

Доказательство. Используя (2.17), представим байесовский риск в виде

$$\begin{aligned} \mathcal{E}_c &= \int d\mathbf{z} \int d\alpha f(\alpha) f_{\alpha}(\mathbf{z}) c(\alpha - \hat{\alpha}) = \\ &= \int d\mathbf{z} w(\mathbf{z}) \int d\alpha g[\tilde{\alpha}(\mathbf{z}), \alpha] c(\alpha - \hat{\alpha}). \end{aligned} \quad (2.18)$$

Поскольку $w(\mathbf{z}) \geq 0$, минимальное значение \mathcal{E}_c достигается тогда, когда внутренний интеграл минимален при каждом значении \mathbf{z} . Но так как этот интеграл зависит лишь от значения, принятого величиной $\tilde{\alpha}$, а не от самих значений \mathbf{z} , то любая оценка $\hat{\alpha}$, минимизирующая (2.18), зависит лишь от $\tilde{\alpha}$. Заметим, что в этом рассуждении существенную роль играет неотрицательность функции w .

Практическое значение достаточных статистик состоит в том, что $\tilde{\alpha}$ может иметь меньшую размерность, чем \mathbf{z} , так что замена наблюдения \mathbf{z} статистикой $\tilde{\alpha}$ значительно упрощает обработку наблюдений и вычисление оценок.

Пример 2.4. Пусть Z_1, \dots, Z_n — последовательность независимых гауссовских случайных величин с из-

вестной дисперсией σ^2 и неизвестным средним α . Тогда

$$\begin{aligned} f_{\alpha}(z) &= \prod_{k=1}^n \frac{1}{\sqrt{2\pi} \sigma} \exp \left[-\frac{(z_k - \alpha)^2}{2\sigma^2} \right] = \\ &= \left(\frac{1}{\sqrt{2\pi} \sigma} \right)^n \exp \left[\frac{-\sum_{k=1}^n z_k^2}{2\sigma^2} \right] \exp \left[\frac{-n\alpha^2 + 2\alpha \sum_{k=1}^n z_k}{2\sigma^2} \right]. \end{aligned} \quad (2.19)$$

Здесь

$$\begin{aligned} w(z) &= \left(\frac{1}{\sqrt{2\pi} \sigma} \right)^n \exp \left[\frac{-\sum_{k=1}^n z_k^2}{2\sigma^2} \right], \\ g(\tilde{\alpha}, \alpha) &= \exp \left[\frac{-n\alpha^2 + 2\alpha \sum_{k=1}^n z_k}{2\sigma^2} \right]. \end{aligned}$$

Таким образом, $\sum_{k=1}^n z_k$ (а также $(1/n) \sum_{k=1}^n z_k$) служит достаточной статистикой для оценивания параметра α .

Задача 2.3. Пусть распределение имеет тот же вид, что и выше, со средним, равным нулю, и неизвестной дисперсией σ^2 . Найти достаточную статистику для σ^2 . Является ли эта величина достаточной статистикой для стандартного отклонения σ ?

Отношение правдоподобия

Прежде чем приступить к выводу неравенства Крамера — Рао, рассмотрим *функцию отношения правдоподобия*. Это позволит сформулировать результаты, которые окажутся справедливыми и тогда, когда вместо конечномерного наблюдения \mathbf{z} мы будем иметь дело с выборочными функциями случайных процессов.

Снова обозначим через \mathcal{A} множество значений α , предполагаемых априори возможными. Пусть существует такое фиксированное значение α_0 , что

$$f_{\alpha_0}(\mathbf{z}) > 0 \quad (2.20)$$

для тех \mathbf{z} , для которых $f_{\alpha}(\mathbf{z}) > 0$ хотя бы при одном $\alpha \in \mathcal{A}$. Определим отношение правдоподобия равенством ¹⁾

$$\Lambda_{\alpha}(\mathbf{z}) = \frac{f_{\alpha}(\mathbf{z})}{f_{\alpha_0}(\mathbf{z})}. \quad (2.21)$$

Отметим, что условие (2.20) необходимо для того, чтобы функция $\Lambda_{\alpha}(\mathbf{z})$ была определена при каждом $\alpha \in \mathcal{A}$ для всех \mathbf{z} -событий, имеющих положительную вероятность. При этом в качестве α_0 можно взять любое значение, удовлетворяющее условию (2.20); на практике выбор α_0 определяется лишь удобством вычислений. Читатель заметит также, что в случае конечномерных наблюдений проводимые ниже вычисления, основанные на Λ_{α} , можно с таким же успехом провести с использованием лишь f_{α} .

Укажем теперь некоторые свойства отношения правдоподобия. Прежде всего для любого наблюдения \mathbf{z} оценку максимального правдоподобия параметра α можно определить как значение α , максимизирующее $\Lambda_{\alpha}(\mathbf{z})$. Далее, для любой случайной величины $G = g(\mathbf{Z})$, зависящей от \mathbf{Z} ,

$$\begin{aligned} E_{\alpha}\{G\} &= \int d\mathbf{z} g(\mathbf{z}) f_{\alpha}(\mathbf{z}) = \\ &= \int d\mathbf{z} g(\mathbf{z}) \Lambda_{\alpha}(\mathbf{z}) f_{\alpha_0}(\mathbf{z}) = E_{\alpha_0}\{G \Lambda_{\alpha}\}, \end{aligned} \quad (2.22)$$

¹⁾ В том случае, когда распределение вероятностей $P_{\alpha}(\mathbf{z})$ нельзя описать функцией плотности, α_0 выбирается таким образом, чтобы соотношение $P_{\alpha_0}(B) \neq 0$ выполнялось для всех событий B , для которых $P_{\alpha}(B) \neq 0$ хотя бы для некоторого $\alpha \in \mathcal{A}$. Тогда в качестве $\Lambda_{\alpha}(\mathbf{z})$ можно взять производную Радона — Никодима [3, стр. 140] вероятностной меры $P_{\alpha}(\mathbf{z})$ относительно меры $P_{\alpha_0}(\mathbf{z})$. Хотя повсюду в этой главе наши рассуждения опираются на существование вероятностной плотности, все окончательные и промежуточные результаты справедливы и в общем случае. Если же нельзя указать такое значение α_0 , что выполнено сформулированное выше условие, то при осуществлении события B по наблюдению \mathbf{Z} можно безошибочно определить, какое из двух значений параметра передавалось. Такой случай принято называть *сингулярным*. Как мы видим, сингулярная математическая модель не является адекватной для большинства реальных физических задач.

где под Λ_α понимается случайная величина $\Lambda_\alpha(\mathbf{Z})$. Заметим, что

$$1 = \int d\mathbf{z} f_\alpha(\mathbf{z}) = \int d\mathbf{z} \Lambda_\alpha(\mathbf{z}) f_{\alpha_0}(\mathbf{z}) = E_{\alpha_0} \{ \Lambda_\alpha \}. \quad (2.23)$$

Возьмем частные производные по α_j от обеих частей этого равенства и изменим порядок дифференцирования и взятия математического ожидания; получим

$$\begin{aligned} 0 &= \int d\mathbf{z} \left[\frac{\partial}{\partial \alpha_j} \Lambda_\alpha(\mathbf{z}) \right] f_{\alpha_0}(\mathbf{z}) = \\ &= \int d\mathbf{z} \left[\frac{\partial}{\partial \alpha_j} \ln \Lambda_\alpha(\mathbf{z}) \right] \Lambda_\alpha(\mathbf{z}) f_{\alpha_0}(\mathbf{z}) = \\ &= E_\alpha \left\{ \frac{\partial}{\partial \alpha_j} \ln \Lambda_\alpha(\mathbf{z}) \right\}. \end{aligned} \quad (2.24)$$

Если указанное изменение порядка допустимо¹⁾ и, следовательно, выполняется равенство (2.24), то функцию Λ_α называют *регулярной* относительно ее частных производных по α первого порядка. Если же правую часть равенства (2.24) можно вторично продифференцировать по α_k , изменив порядок дифференцирования и взятия математического ожидания, то функцию Λ_α называют *регулярной* относительно ее вторых частных производных по α .

Задача 2.4. Показать, что если функция Λ_α регулярна относительно вторых частных производных, то

$$E_\alpha \left\{ \frac{\partial}{\partial \alpha_j} \ln \Lambda_\alpha \frac{\partial}{\partial \alpha_k} \ln \Lambda_\alpha \right\} = - E_\alpha \left\{ \frac{\partial^2}{\partial \alpha_j \partial \alpha_k} \ln \Lambda_\alpha \right\}. \quad (2.25)$$

Левая часть равенства (2.25) представляет собой ковариацию двух случайных величин, являющуюся функцией параметра α . Обозначим эту величину через $b_{jk}(\alpha)$ и положим $b_{jk} = b_{jk}(\alpha')$. Матрицу ковариаций, отвечающую $b_{jk}(\alpha)$, обозначим через $B(\alpha)$, а обратную к ней — через $G(\alpha) = B^{-1}(\alpha)$. Отметим, что хотя

¹⁾ По теореме о мажорируемой сходимости [3, стр. 135] это имеет место, если частные производные $\partial \Lambda_\alpha / \partial \alpha_j$ ограничены абсолютно интегрируемой (по мере $P_{\alpha_0}(\mathbf{z})$) функцией в некоторой последовательности точек α , сходящейся к интересующему нас значению параметра.

полученные в этом пункте результаты и не имеют непосредственной статистической интерпретации, однако, как мы увидим, они окажутся полезными в дальнейшем.

Неравенство Крамера — Рао

Найдем абсолютную границу эффективности, которую можно будет использовать как эталон для сравнения различных оценок. При этом мы ограничимся лишь квадратичными функциями потерь и, не предполагая наличия априорного распределения параметра α , рассмотрим условный байесовский риск при $\alpha = \alpha'$.

Естественной функцией потерь для векторного параметра α могла бы служить величина $\|\hat{\alpha} - \alpha'\|^2$. Однако в ряде случаев такая мера близости не адекватна поставленной задаче, поскольку нас может интересовать не только оценка самого параметра α , но и некоторых функций от α . Например, если α — радиолокационные координаты цели, то нас могут интересовать ее декартовы координаты. При этом, если ошибки малы (например, ошибка в измерении дальности в процентном отношении мала по сравнению с дальностью цели), для вычисления ошибки в преобразованных координатах можно воспользоваться разложением с точностью до первого порядка малости. Таким образом, в одной из преобразованных координат ошибка будет иметь вид

$$e = \sum_{j=1}^n c_j (\alpha'_j - \hat{\alpha}_j) = (c, \alpha' - \hat{\alpha}), \quad (2.26)$$

и, таким образом, наша задача состоит в исследовании условных рисков вида

$$\begin{aligned} E_{\alpha'} \{e^2\} &= E_{\alpha'} \left\{ \sum_{j,k=1}^n c_j c_k (\alpha'_j - \hat{\alpha}_j) (\alpha'_k - \hat{\alpha}_k) \right\} = \\ &= \sum_{j,k=1}^n c_j c_k E_{\alpha'} \{(\alpha'_j - \hat{\alpha}_j) (\alpha'_k - \hat{\alpha}_k)\} = (c, R_{\alpha} c), \end{aligned} \quad (2.27)$$

где R_{α} — корреляционная матрица ошибок в α -координатах,

Мы хотим найти минимальное значение, которое может принимать этот условный риск для любого значения s , какова бы ни была несмещенная оценка $\hat{\alpha}$. Сначала поясним, почему удобно потребовать выполнение условия несмещенности. Дело в том, что поскольку истинное значение α' параметра α нам неизвестно, мы должны были бы найти границу величины $E_{\alpha'}\{e^2\}$ для некоторых значений α' , которые кажутся нам наиболее характерными или вероятными. Заметим, однако, что для каждого отдельного значения α' можно получить $E_{\alpha'}\{e^2\} = 0$, полагая попросту $\hat{\alpha}(z) \equiv \alpha'$. Естественно, нам хотелось бы избавиться от таких патологических оценок. Точнее, мы хотели бы, чтобы принимаемые функцией $\hat{\alpha}(z)$ значения отражали зависимость от α' при *любых* значениях α' . Одним из удобных и содержательных условий такой зависимости является требование несмещенности оценки $\hat{\alpha}(z)$:

$$E_{\alpha}\{\hat{\alpha}(Z)\} = \alpha \quad \text{при всех } \alpha \in \mathcal{A}.$$

В ряде задач несмещенных оценок может не быть. В этом случае граница качества оценок должна также учитывать возможную величину смещения; см. ниже замечание 2.3.

Приступим теперь к выводу границы для величины условного риска в (2.27) при произвольной несмещенной оценке. Для такой оценки при любых s и α имеем

$$\begin{aligned} 0 &= E_{\alpha} \left\{ \sum_{j=1}^n c_j (\hat{\alpha}_j - \alpha_j) \right\} = \\ &= \sum_{j=1}^n c_j \int dz (\hat{\alpha}_j - \alpha_j) \Lambda_{\alpha}(z) f_{\alpha_0}(z). \end{aligned} \quad (2.28)$$

Предположим, что функция Λ_{α} регулярна относительно частных производных первого порядка. Дифференцируя обе части последнего равенства по α_i и меняя

порядок дифференцирования и интегрирования, после перегруппировки членов получаем

$$\begin{aligned} c_i &= \sum_{j=1}^n c_j \int dz (\hat{\alpha}_j - \alpha_j) \left[\frac{\partial}{\partial \alpha_i} \ln \Lambda_\alpha(\alpha) \right] f_\alpha(z) = \\ &= \sum_{j=1}^n c_j E_\alpha \left\{ (\hat{\alpha}_j - \alpha_j) \frac{\partial}{\partial \alpha_i} \ln \Lambda_\alpha \right\}. \end{aligned}$$

Умножим обе части этого равенства на произвольные постоянные d_i , $i = 1, \dots, n$, и просуммируем по i . Тогда

$$(c, d) = E_\alpha \left\{ (c, \hat{\alpha} - \alpha) \sum_{i=1}^n d_i \frac{\partial}{\partial \alpha_i} \ln \Lambda_\alpha \right\}.$$

Применим к правой части полученного равенства неравенство Шварца:

$$\begin{aligned} (c, d)^2 &\leq E_\alpha \{ (c, \hat{\alpha} - \alpha)^2 \} E_\alpha \left\{ \sum_{i,j=1}^n d_i d_j \frac{\partial \ln \Lambda_\alpha}{\partial \alpha_i} \frac{\partial \ln \Lambda_\alpha}{\partial \alpha_j} \right\} = \\ &= E_\alpha \{ (c, \hat{\alpha} - \alpha)^2 \} \sum_{i,j=1}^n d_i d_j E_\alpha \left\{ \frac{\partial \ln \Lambda_\alpha}{\partial \alpha_i} \frac{\partial \ln \Lambda_\alpha}{\partial \alpha_j} \right\} = \\ &= E_\alpha \{ (c, \hat{\alpha} - \alpha)^2 \} (d, B(\alpha) d). \end{aligned} \quad (2.29)$$

Рассмотрим неравенство (2.29) при $\alpha = \alpha'$, $B(\alpha') = B$. Предположим, что матрица B положительно определена, и положим $d = Gc = B^{-1}c$; тогда

$$(c, Gc)^2 \leq E_{\alpha'} \{ (c, \hat{\alpha} - \alpha')^2 \} (Gc, BB^{-1}c).$$

Заметим, что матрица G положительно определена (поскольку такова матрица B) и симметрична. Разделив обе части последнего неравенства на (c, Gc) , окончательно находим, что

$$E_{\alpha'} \{ e^2 \} = E_{\alpha'} \{ (c, \alpha' - \hat{\alpha})^2 \} = (c, R_\alpha c) \geq (c, Gc). \quad (2.30)$$

Мы получили хорошо известное *неравенство Крамера — Рао*, описывающее абсолютную нижнюю границу величины риска для любой несмещенной оценки при

произвольном преобразовании координат \mathbf{c} . Его вывод потребовал по существу лишь положительной определенности матрицы B .

В дальнейшем нам понадобятся обозначения

$$\mathbf{S}_\alpha = \text{grad}_\alpha \ln \Lambda_\alpha(\mathbf{Z}), \quad (2.31)$$

$$S_{i,\alpha} = \frac{\partial}{\partial \alpha_i} \ln \Lambda_\alpha(\mathbf{Z}), \quad i = 1, \dots, n. \quad (2.32)$$

Задача 2.5. Показать, что матрица B положительно определена, за исключением случая, когда случайные величины $S_{i,\alpha}$ линейно зависимы при $\alpha = \alpha'$ (т. е. одну из этих величин можно представить со среднеквадратической погрешностью, равной нулю, в виде линейной комбинации остальных величин).

Приведем ряд замечаний, связанных с применением неравенства Крамера — Рао.

Замечание 2.1. Обычно нахождение матрицы B упрощается при вычислении правой части равенства (2.25).

Замечание 2.2. В одномерном случае, когда α — скаляр, неравенство Крамера — Рао принимает вид

$$\begin{aligned} E_{\alpha'} \{(\alpha' - \hat{\alpha})^2\} &\geqslant \\ &\geqslant \left[E_{\alpha'} \left\{ \left(\frac{\partial \ln \Lambda_\alpha}{\partial \alpha} \right)^2 \right\} \right]^{-1} = - \left[E_{\alpha'} \left\{ \frac{\partial^2 \ln \Lambda_\alpha}{\partial \alpha^2} \right\} \right]^{-1}. \end{aligned} \quad (2.33)$$

Замечание 2.3. В случае смещенной оценки, т. е. когда

$$E_\alpha \{\hat{\alpha}\} = \alpha + \mathbf{b}(\alpha), \quad (2.34)$$

неравенство Крамера — Рао можно модифицировать с учетом влияния величины смещения. Мы предлагаем читателю ознакомиться с этим обобщением неравенства Крамера — Рао по книге Уилкса [4, разд. 12.2] либо в качестве упражнения разобраться в нем самому. При этом следует начать с замены левой части в (2.28) на $\mathbf{b}(\alpha)$.

Замечание 2.4. Пусть $\hat{\alpha}_N$ — оценка параметра α , построенная по N независимым одинаково распределен-

ным наблюдениям Z_1, \dots, Z_N , и пусть G — матрица, отвечающая наблюдению Z_h . Тогда

$$E_{\alpha'} \{(\hat{\alpha}_N - \alpha')^2\} \geq N^{-1}(c, Gc). \quad (2.35)$$

Задача 2.6. Получить неравенство (2.35).

Замечание 2.5. Рассмотрим случай обращения неравенства Крамера — Рао в равенство. В неравенстве Шварца (2.29) равенство достигается, когда

$$k(\alpha)(c, \hat{\alpha} - \alpha) = (d, S_\alpha),$$

где $k(\alpha)$ — «постоянная», которая может зависеть от α , но не зависит от z . Производя подстановку $d = Gc$, применявшуюся нами при выводе неравенства Крамера — Рао, получаем

$$k(\alpha)(c, \hat{\alpha} - \alpha) = (Gc, S_\alpha) = (c, GS_\alpha). \quad (2.36)$$

Это равенство является условием обращения неравенства Крамера — Рао в равенство при фиксированном выборе c . Если же неравенство Крамера — Рао обращается в равенство при любом c , то должно выполняться соотношение

$$k(\alpha)(\hat{\alpha} - \alpha) = GS_\alpha. \quad (2.37)$$

Если в неравенстве Крамера — Рао достигается равенство при $\alpha = \alpha'$, так что

$$R_\alpha = G, \quad (2.38)$$

то оценку $\hat{\alpha}$ называют *эффективной* (в точке α'). Если же $\hat{\alpha}_N$ — оценка, построенная на основании N независимых одинаково распределенных наблюдений величины Z , и G — матрица, отвечающая функции $\Lambda_\alpha(Z)$ для отдельного наблюдения Z , то оценку $\hat{\alpha}_N$ называют *асимптотически эффективной*, если

$$\lim_{N \rightarrow \infty} [NR_{\hat{\alpha}_N} - G] = 0, \quad (2.39)$$

где 0 — нулевая матрица.

Задача 2.7. Пусть $Z_m = g(m) + \mathcal{N}_m$, $m = 1, 2, \dots$, n , — n наблюдений, где \mathcal{N}_m — независимые гауссов-

ские случайные величины с нулевым средним и дисперсией σ^2 .

а) Пусть

$$g(m) = \sum_{k=1}^M \alpha_k g_k(m), \quad M \leq n,$$

где $g_k(m)$ — известные функции от m , а α_k — неизвестные параметры.

- (i) Какой вид имеет оценка максимального правдоподобия вектора α ?
- (ii) Является ли эта оценка несмещенной? Эффективной? Асимптотически эффективной?
- (iii) Какой вид имеет неравенство Крамера — Рао для дисперсии ошибки?

б) Пусть $g(m) = m^\alpha$, где α — неизвестный параметр, $\alpha > 0$.

- (i) Выписать уравнение для нахождения оценки максимального правдоподобия параметра α .
- (ii) Является ли эта оценка эффективной?
- (iii) Допустим, что A — случайная величина с экспоненциальным распределением. Выяснить, трудно ли получить в этом случае выражение для $E\{\alpha|Z\}$.

Пример 2.5. Рассмотрим более подробно частный случай предыдущей задачи. Пусть

$$Z = g(\alpha) + \mathcal{N},$$

где для простоты будем считать Z , \mathcal{N} и α скалярными величинами. Пусть \mathcal{N} — гауссовская случайная величина с нулевым средним и дисперсией σ^2 . Тогда

$$S_\alpha = \frac{\partial}{\partial \alpha} \ln \Lambda_\alpha(z) = \frac{\partial}{\partial \alpha} \ln f_\alpha(z) = \sigma^{-2} [z - g(\alpha)] g'(\alpha).$$

Чтобы оценка $\hat{\alpha}(z)$ параметра α была эффективной, она должна иметь вид

$$\hat{\alpha}(z) = z \frac{g'(\alpha)}{k(\alpha) \sigma^2} + \alpha - \frac{g(\alpha) g'(\alpha)}{k(\alpha) \sigma^2}. \quad (2.40)$$

Но функция $\hat{\alpha}(z)$ может зависеть лишь от наблюдения z и не должна зависеть от параметра α . Для того чтобы

правая часть равенства (2.40) не зависела от α , должны выполняться равенства

$$\begin{aligned}\frac{g'(\alpha)}{k(\alpha)\sigma^2} &= c_1, \\ \alpha - \frac{g(\alpha)g'(\alpha)}{k(\alpha)\sigma^2} &= c_2,\end{aligned}$$

откуда следует, что

$$g(\alpha) = \frac{\alpha - c_2}{c_1}. \quad (2.41)$$

Таким образом, если функция $g(\alpha)$ нелинейна, ни один метод оценивания не позволяет достичь границы Крамера—Рао для среднеквадратической ошибки.

Полученное условие весьма ограничительно. Поэтому гораздо важнее выяснить, в каких условиях неравенство Крамера—Рао дает все же приближенную границу для качества оценок. Рассмотрим с этой целью рис. 2.2. Заметим, что для того, чтобы в неравенстве Шварца (2.29) выполнялось приближенное равенство, соотношение (2.40) должно приближенно выполняться для тех значений z , которым плотность вероятности $f_\alpha(z)$ приписывает основную вероятностную массу (на рис. 2.2 это интервал (z_1, z_2)). Следовательно, равенство (2.41) должно выполняться приближенно лишь для соответствующих значений α , т. е. в интервале (α_1, α_2) на рис. 2.2. Длина интервала (z_1, z_2) , а значит, и (α_1, α_2) , непосредственно зависит от дисперсии σ^2 величины \mathcal{N} .

Итак, если функция $g(\alpha)$ гладкая, то при достаточно малых σ она приближенно линейна на интервале (α_1, α_2) , и потому граница Крамера—Рао дает в этом случае хорошее приближение достижимого качества оценки $\hat{\alpha}$. С другой стороны, если σ настолько велико, что функцию $g(\alpha)$ уже нельзя считать линейной на интервале (α_1, α_2) , то граница Крамера—Рао будет существенно меньше фактически достижимой среднеквадратической ошибки.

Сделанные замечания обобщаются на случай, когда α и Z — векторы, а также когда наблюдение Z и гауссовский шум \mathcal{N} являются случайными процессами. Лишь

в случае, когда отношение сигнала к шуму, входящему в наблюдение, достаточно велико (над некоторой «пороговой» областью), так что система приближенно линейна, граница Крамера — Рао оказывается полезной. Мы вернемся к этому вопросу в гл. 4.

Другого рода трудность, связанная с применением неравенства Крамера — Рао, состоит в том, что в нем не учитывается априорная информация о параметре α ,

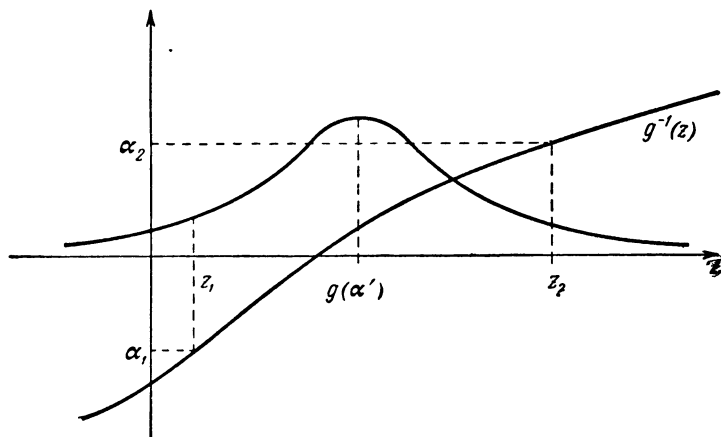


Рис. 2.2. Изображение области, в которой функция $g(\alpha)$ должна приближаться к линейной.

которой мы могли бы располагать, так что граница в этом случае оказывается завышенной. Другими словами, здесь мы имеем дело со слишком «пессимистическим» прогнозом в отличие от предыдущего случая, когда полученная граница оказалась слишком «оптимистической».

Рассмотрим, в частности, случай $Z = \alpha + \mathcal{N}$, где \mathcal{N} — гауссовская случайная величина с нулевым средним и дисперсией σ^2 . При этом оценка $\hat{\alpha}(z) = z$ несмещенная и ее дисперсия совпадает с границей Крамера — Рао. Однако если априори известно, что α принадлежит некоторому интервалу (a, b) , то качество оценки можно улучшить, урезая функцию $\hat{\alpha}(z) = z$ так, чтобы

она принадлежала тому же интервалу. Если же известно, что α — случайная величина с дисперсией S , то лучшую по сравнению с $\hat{\alpha}(z) = z$, но опять-таки смещенную оценку можно получить, взяв

$$\hat{\alpha}'(z) = \frac{S}{S + \sigma^2} z.$$

Ниже мы опишем вкратце получение модифицированной границы Крамера — Рао в ситуации, когда α — случайная величина с известным априорным распределением.

Важность неравенства Крамера — Рао объясняется тем, что в случае, когда можно пренебречь краевыми эффектами (длина интервала (a, b) существенно больше границы Крамера — Рао), легко оценить свойства системы в линейной области, *не прибегая* к исследованию вопроса о том, как построить хорошую оценку.

Дальнейший путь исследования состоит в нахождении более точных границ, дающих правильное представление об оптимальных свойствах оценок в нелинейной области (т. е. при малых значениях отношения сигнала к шуму). Такого рода граница найдена Баранкиным [5], однако, хотя в некоторых случаях с ее помощью были получены, по-видимому, довольно точные результаты, возможность использования границы Баранкина в ряде задач кажется проблематичной. Поясним кратко суть дела, в которой не все отдают себе отчет.

Граница Баранкина (частным случаем которой является граница Крамера — Рао) представляет собой наименьшее значение $E_{\alpha}\{(\hat{\alpha} - \alpha')^2\}$ для любой несмещенной оценки $\hat{\alpha}$. Эта граница точна в том смысле, что существует оценка с произвольно малым смещением, качество которой сколь угодно близко к данной границе при $\alpha = \alpha'$. Но эта оценка может оказаться неустойчивой, т. е. качество ее в близкой точке будет значительно хуже (см. [6]). Поскольку мы хотим построить оценку, используемую независимо от значения, принятого α , то граница Баранкина может оказаться намного меньше среднеквадратической ошибки, достижимой с помощью оценок, пригодных в некоторой области значений α .

Более полезные границы для задач, рассматриваемых нами в гл. 3 и 4, найдены Зивом и Закаи [7] с помощью концепций теории обнаружения. Мы вернемся к этому вопросу в гл. 3 и 4.

Свойства оценки максимального правдоподобия

Теперь мы можем обосновать сделанное выше заявление о том, что если существует оценка, обладающая теми или иными «хорошими» свойствами, то она совпадает с оценкой максимального правдоподобия.

Теорема 2.4. *Если эффективная оценка существует, то ею является оценка максимального правдоподобия.*

Доказательство. Согласно (2.36), эффективная оценка должна удовлетворять соотношению

$$k(\alpha)[\hat{\alpha}_{\text{eff}}(z) - \alpha] = G S_{\alpha}(z), \quad (2.37')$$

где

$$S_{\alpha}(z) = \text{grad}_{\alpha} \ln \Lambda_{\alpha}(z). \quad (2.31')$$

Соотношение (2.37') эквивалентно равенству

$$S_{\alpha}(z) = k(\alpha) B[\hat{\alpha}_{\text{eff}}(z) - \alpha]. \quad (2.42)$$

Рассмотрим оценку максимального правдоподобия $\hat{\alpha}_{\text{МП}}$, т. е. такое значение параметра α , которое максимизирует величину $\Lambda_{\alpha}(z)$ (или $\ln \Lambda_{\alpha}(z)$). Если же существует $\text{grad}_{\alpha} \ln \Lambda_{\alpha}(z)$, то оценка максимального правдоподобия — это значение α , для которого

$$\text{grad}_{\alpha} \ln \Lambda_{\alpha}(z) = S_{\alpha}(z) = 0,$$

и, следовательно, в силу (2.42) $\hat{\alpha}_{\text{МП}}$ совпадает со значением α , для которого

$$B[\hat{\alpha}_{\text{eff}}(z) - \alpha] = 0.$$

Таким образом, поскольку матрица B положительно определена (условие, наложенное при выводе неравенства Крамера — Рао), то $\alpha_{\text{МП}}$ определяется равенством $\alpha_{\text{МП}} = \hat{\alpha}_{\text{eff}}(z)$, и теорема доказана

Теорема 2.5. Если существует достаточная статистика $\tilde{\alpha}$, то оценка максимального правдоподобия зависит лишь от $\tilde{\alpha}$.

Задача 2.8. Доказать теорему 2.5.

Теорема 2.6. Пусть $\mathbf{Z}_1, \mathbf{Z}_2, \dots$ — последовательность независимых одинаково распределенных случайных величин, для которых отношение правдоподобия $\Lambda_{\alpha}(\mathbf{Z})$ регулярно относительно своих первых и вторых частных производных по α , и пусть ковариационная матрица B вектора \mathbf{S}_{α} положительно определена. Обозначим через $\hat{\alpha}_N$ оценку МП, построенную по наблюдениям $\mathbf{Z}_1, \dots, \mathbf{Z}_N$. Если оценка $\hat{\alpha}_N$ однозначно определена для всех N , больших некоторого N_0 , то оценка МП асимптотически нормальна со средним α и ковариационной матрицей $R_{\alpha} = N^{-1}G$ и, следовательно, асимптотически эффективна.

Доказательство этой теоремы можно найти в книге Уилкса [4, разд. 12.7].

Задача 2.9. Пусть \mathbf{Z} и α — соответственно m -мерный и n -мерный векторы, а матрица G и B определены, как выше. Показать, что для любой несмещенной оценки $\hat{\alpha}$

$$E \{(\alpha'_i - \hat{\alpha}_i)^2\} \geq g_{ii} \geq (b_{ii})^{-1}, \quad i = 1, \dots, n. \quad (2.43)$$

Задача 2.10. Пусть $\mathbf{Z} = \mathbf{g}(\alpha) + \mathcal{N}$, где \mathbf{Z} , $\mathbf{g}(\alpha)$ и \mathcal{N} обозначают m -мерные векторы, причем вектор \mathcal{N} — гауссовский с нулевым средним и ковариационной матрицей R . Пусть α будет n -мерным параметром. Предположим, что существует такое значение α_0 , что $\mathbf{g}(\alpha_0) = \mathbf{0}$.

- Найти выражение для Λ_{α} .
- Найти выражение для $\mathbf{S}_{\alpha} = \text{grad}_{\alpha} \Lambda_{\alpha}(\mathbf{z})$.
- Найти вид элементов b_{jh} ковариационной матрицы B вектора \mathbf{S} .
- Не вычисляя матрицу $G = B^{-1}$, выразить через ее элементы границу снизу величины

$$E \{ \|\alpha' - \hat{\alpha}\|^2 \}$$

для произвольной несмещенной оценки $\hat{\alpha}$,

д) Рассмотреть возможные упрощения в задаче в) в случае некоррелированного шума, т. е. при $R = N_0 I$, где I — единичная $(m \times m)$ -матрица.

Неравенство Крамера — Рао для случайных параметров

Интересно найти неравенство, аналогичное неравенству Крамера — Рао, для случая, когда известно априорное распределение параметра α . Такое неравенство действительно можно вывести, и мы сейчас этим займемся. Наши рассуждения следуют разд. 2.4 книги Ван Три-са [1].

Обозначим, как и в (2.34), смещение оценки через $\mathbf{b}(\alpha)$. Мы не будем здесь предполагать, что оценка параметра α является несмещенной, а лишь потребуем выполнение следующего условия:

$$\lim_{\alpha_i \rightarrow \pm \infty} f(\alpha) \mathbf{b}(\alpha) = 0, \quad i = 1, \dots, n. \quad (2.44)$$

Рассмотрим скалярное произведение вектора $\mathbf{b}(\alpha)$ и произвольного вектора \mathbf{c} ; умножая эту величину на $f(\alpha)$ и дифференцируя по α_i , получаем

$$\begin{aligned} \frac{\partial}{\partial \alpha_i} [f(\alpha) (\mathbf{c}, \mathbf{b}(\alpha))] &= \frac{\partial}{\partial \alpha_i} \sum_{j=1}^n c_j \int dz (\hat{\alpha}_j - \alpha_j) f(\alpha) \Lambda_\alpha(z) f_{\alpha_0}(z) = \\ &= -c_i f(\alpha) \int dz \Lambda_\alpha(z) f_{\alpha_0}(z) + \\ &+ \sum_{j=1}^n c_j \int dz (\hat{\alpha}_j - \alpha_j) \frac{\partial}{\partial \alpha_i} [f(\alpha) \Lambda_\alpha(z)] f_{\alpha_0}(z) = \\ &= -c_i f(\alpha) + \sum_{j=1}^n c_j \int dz (\hat{\alpha}_j - \alpha_j) \frac{\partial}{\partial \alpha_i} \{\ln [f(\alpha) \Lambda_\alpha(z)]\} f(\alpha) f_{\alpha_0}(z). \end{aligned} \quad (2.45)$$

Введем обозначения:

$$\mathbf{S}_{p, \alpha} = \text{grad}_\alpha \ln f(\alpha), \quad \mathbf{S}_{0, \alpha} = \text{grad}_\alpha \ln \Lambda_\alpha(z), \quad (2.46)$$

$$\mathbf{S}_{T, \alpha} = \mathbf{S}_{p, \alpha} + \mathbf{S}_{0, \alpha} = \text{grad}_\alpha \{\ln [f(\alpha) \Lambda_\alpha(z)]\}, \quad (2.47)$$

$$\mathbf{B}_p = \text{cov } \mathbf{S}_{p, \alpha}, \quad \mathbf{B}_0 = \text{cov } \mathbf{S}_{0, \alpha}, \quad (2.48)$$

где математическое ожидание берется относительно совместного распределения с плотностью $f(\alpha)f_\alpha(z)$. В качестве упражнения предлагаем читателю проверить, что

$$B_T = \text{cov } S_{T,\alpha} = B_p + B_0. \quad (2.49)$$

Отметим, что B_0 — математическое ожидание матрицы $B(\alpha)$, фигурирующей в неравенстве Крамера — Рао, по мере с плотностью $f(\alpha)$.

Возвращаясь к (2.45), заметим, что в силу (2.44) интегрирование этого равенства по α_i от $-\infty$ до ∞ дает в левой части 0. Таким образом,

$$0 = -c_i + \sum_{j=1}^n c_j \int d\alpha_i \int dz (\hat{\alpha}_j - \alpha_j) \times \\ \times \frac{\partial}{\partial \alpha_i} \{\ln [f(\alpha) \Lambda_\alpha(z)]\} f(\alpha) f_\alpha(z), \quad i = 1, \dots, n. \quad (2.50)$$

Умножая обе части последнего равенства на d_i и суммируя по i , получаем

$$(c, d) = E \{(c, \hat{\alpha} - \alpha)(d, S_T)\}.$$

Дальнейшие рассуждения, аналогичные использованным при выводе неравенства Крамера — Рао, приводят к неравенству

$$E \{(c, \hat{\alpha} - \alpha)^2\} \geq (c, B_T^{-1}c), \quad (2.51)$$

где B_T вычисляется по формуле (2.49). Нетрудно показать, что B_p выражается следующим образом через вторые производные:

$$[B_p]_{ij} = -E \left\{ \frac{\partial^2}{\partial \alpha_i \partial \alpha_j} \ln f(\alpha) \right\}. \quad (2.52)$$

Задача 2.11. Вывести (2.49) и (2.52).

Заключительное замечание

Мы определили некоторые основные понятия и сформулировали ряд результатов теории оценивания. Для более широкого изучения предлагаем читателю ознакомиться с книгой Уилкса [4, гл. 12] или Ван Триса [1, гл. 2].

СПИСОК ЛИТЕРАТУРЫ

1. Ван Трис Г. Л., Теория обнаружения, оценок и модуляции, т. 1, изд-во «Сов. радио», М., 1972.
2. Витерби Э. Д., Принципы когерентной связи, изд-во «Сов. радио», М., 1970.
3. Лоэв М., Теория вероятностей, ИЛ, М., 1962.
4. Уилкс С. С., Математическая статистика, изд-во «Наука», М., 1967.
5. Баранкин (Barankin E. W.), Locally best unbiased estimates, *Ann. Math. Statist.*, **20** (1949), 477—501.
6. Глейв (Glave F. W.), неопубликованное исследование качества различных оценок, используемых при время-импульсной модуляции.
7. Зив, Закаи (Ziv J., Zakai M.), Some lower bounds in signal parameter estimation, *IEEE Trans. Inform. Theory*, **IT-15** (May 1969), 386—391.

Глава 3

ОЦЕНКА ПАРАМЕТРОВ СИГНАЛА В АДДИТИВНОМ ГАУССОВСКОМ ШУМЕ

Применим идеи предыдущей главы к исследованию задачи модуляции, в которой интересующие нас параметры в процессе модуляции временного сигнала $s(t)$ порождают семейство сигналов $s(t, \alpha)$ $0 \leq t \leq T$. Наблюдение этих сигналов производится при наличии аддитивного гауссовского шума $Z_n(t)$, и, следовательно, результат наблюдения представляет собой выборочную функцию гауссовского случайного процесса

$$Z(t) = s(t, \alpha) + Z_n(t), \quad 0 \leq t \leq T, \quad (3.1)$$

длительностью T секунд.

Такая модель вполне приемлема для ряда реальных физических проблем, представляющих практический интерес, среди которых мы упомянем следующие.

1. *Измерение дальности радиолокатором.* В этом случае вектор α отвечает таким параметрам цели, как дальность, скорость, отражающая способность и т. д. Шум $Z_n(t)$ представляет собой собственный шум приемника, приведенный ко входу, или является результатом шумового воздействия основных излучателей. В данной задаче сигнал $s(t, \alpha)$ узкополосный.

2. *Сейсмические исследования.* Здесь сигнал $s(t, \alpha)$ представляет собой низкочастотные сейсмические колебания, возникающие в результате искусственного взрыва, а α обозначает параметры, характеризующие, например, глубину пласта и его проводящие свойства. В этом случае на возникновение шума оказывают влияние окружающий геофон, а также побочные источники, не интересующие исследователей.

3. *Импульсная модуляция.* Роль α здесь играет сообщение, представляющее собой случайный процесс, а в качестве $Z_n(t)$ снова выступают шумы в канале и собственный шум приемника, приведенный ко входу.

Сначала мы будем искать Λ_α применительно к общей модели вида (3.1). Затем сосредоточим внимание на задачах радиолокации и рассмотрим подробнее влияние выбора конкретного вида сигнала на качество приема. Задачам импульсной модуляции будет посвящена гл. 4.

При рассмотрении задач радиолокации нельзя не учитывать тот факт, что сигнал $s(t, \alpha)$ является узкополосным, поскольку именно на этом основано определение скорости движущегося объекта. Поэтому мы начнем с модели, специально приспособленной к случаю сигналов с ограничением на спектр. Обычно в радиолокации посылаемый сигнал имеет вид $\text{Re}\{s(t)e^{i\omega_0 t}\}$, где полоса частот модуляции $s(t)$ много меньше несущей частоты $f_0 = \omega_0/2\pi$. Представим принятый сигнал в виде

$$s(t, \alpha) = \text{Re}\{bs_\beta(t)e^{i\omega_0 t}\}, \quad 0 \leq t \leq T, \quad (3.2)$$

где n -вектор $\beta = (\beta_1, \dots, \beta_n)$ обозначает те параметры цели, которые могут представлять интерес, а именно дальность, скорость, отражающую способность, площадь рассеяния и т. п. Комплексный коэффициент усиления

$$b = Ae^{i\varphi} \quad (3.3)$$

указывает на влияние побочных факторов, также вызывающих изменение фазы и амплитуды, измерение которых, однако, нас не интересует; так, φ может указывать на сдвиг фазы между передатчиком и приемником. Таким образом, вектор α имеет вид $\alpha = (A, \varphi, \beta_1, \dots, \beta_n)$.

Будем предполагать, что шум, т. е. процесс $Z_n(t)$, является результатом прохождения широкополосного гауссовского процесса через высокочастотную (радио)-часть приемника. Соответственно в математической модели будем считать, что процесс $Z_n(t)$

- (i) гауссовский,
- (ii) имеет нулевое математическое ожидание,
- (iii) стационарный со спектральной плотностью, симметричной относительно ω_0 .

Прежде чем переходить к задачам оценивания, попытаемся описать такие узкополосные случайные процессы.

3.1. ПРЕДСТАВЛЕНИЕ УЗКОПОЛОСНЫХ СЛУЧАЙНЫХ ПРОЦЕССОВ

Преобразование Гильберта и комплексная огибающая

В этом разделе мы сформулируем некоторые полезные свойства комплексного представления узкополосных случайных процессов. Мы не будем подробно вдаваться в вопросы физической интерпретации, а также в детали обсуждаемых здесь свойств; читателю, интересующемуся более полным обсуждением этих вопросов, предлагаем ознакомиться с разд. 5.1 книги Сакрисона [1].

Для произвольной временной функции $z(t)$ определим ее *преобразование Гильберта* $\hat{z}(t)$ как сигнал, получающийся в результате воздействия на $z(t)$ линейной системы с передаточной функцией

$$H(f) = \begin{cases} -i, & f > 0, \\ 0, & f = 0, \\ +i, & f < 0. \end{cases} \quad (3.4)$$

Отметим, что знак $\hat{}$ имеет здесь другой смысл, нежели в случае оценок. Такое использование одного и того же символа для обозначения оценок и преобразований Гильберта в принципе может привести к путанице; однако в дальнейшем из текста всегда будет ясно, о чем идет речь. Заметим, что в обоих случаях использование этого знака общепринято.

Отметим простейшие следствия из данного нами определения ¹⁾:

- (i) $\hat{\hat{z}}(t) = -z(t)$,
- (ii) если $z(t)$ принимает действительные значения, то и $\hat{z}(t)$ принимает действительные значения.

¹⁾ Если преобразованию Фурье $Z(f)$ функции $z(t)$ отвечает импульс при $f = 0$, то $\hat{\hat{z}}$ отличается от $-z(t)$ постоянной, соответствующей этому импульсу.

Определим комплексную огибающую функции $z(t)$ формулой

$$z_+(t) = z(t) + i\dot{z}(t). \quad (3.5)$$

Если функция $z(t)$ допускает преобразование Фурье

$$Z(f) = \int_{-\infty}^{\infty} dt z(t) e^{-i2\pi ft},$$

то преобразованием Фурье функции $z_+(t)$ является

$$Z_+(f) = \begin{cases} 2Z(f), & f > 0, \\ Z(0), & f = 0, \\ 0, & f < 0. \end{cases} \quad (3.6)$$

Предположим, что функция $Z(f)$ отлична от нуля лишь в полосе $|f \pm f_0| < f_0$. Положим

$$\begin{aligned} v(t) &= z_+ e^{-i2\pi f_0 t}, \\ x(t) &= \operatorname{Re}\{v(t)\}, \quad y(t) = \operatorname{Im}\{v(t)\}, \\ a(t) &= +\sqrt{x^2(t) + y^2(t)} = |v(t)| = |z_+(t)|, \\ \theta(t) &= \arcsin[y(t)/a(t)] = \arccos[x(t)/a(t)]. \end{aligned} \quad (3.7)$$

Можно показать [1, разд. 5.1], что функция $z(t)$ представима в виде

$$\begin{aligned} z(t) &= \operatorname{Re}\{v(t) e^{i2\pi f_0 t}\} = x(t) \cos 2\pi f_0 t - y(t) \sin 2\pi f_0 t = \\ &= a(t) \cos[2\pi f_0 t + \theta(t)], \end{aligned} \quad (3.8)$$

где сигналы $x(t)$ и $y(t)$ низкочастотные, т. е. их преобразование Фурье отлично от нуля лишь при $|f| < f_0$. Это представление единственно, причем величины $x(t)$ и $y(t)$ можно получить, пропуская функцию $z(t)$, умноженную соответственно на $2 \cos 2\pi f_0 t$ и $2 \sin 2\pi f_0 t$, через фильтр с частотной характеристикой

$$H_{lp}(f) = \begin{cases} 1, & f < f_0, \\ 0, & f \geq f_0. \end{cases}$$

Задача 3.1. Проверить указанные свойства представления (3.8).

Рассмотрим теперь ситуацию, когда $Z(t)$ — случайный процесс. Обозначим через

$$\mathbf{E}\{X_{t+\tau}X_t\} = R_x(t+\tau, t),$$

$$\mathbf{E}\{X_{t+\tau}Y_t\} = R_{xy}(t+\tau, t)$$

соответственно ковариационную и взаимную ковариационную функции процессов $X(t)$ и $Y(t)$. В стационарном случае положим

$$\mathbf{E}\{X_{t+\tau}X_t\} = R_x(\tau), \quad \mathbf{E}\{X_{t+\tau}Y_t\} = R_{xy}(\tau),$$

а соответствующие спектральные плотности (преобразования Фурье функции R) обозначим через $S_x(f)$ и $S_{xy}(f)$.

Предположим, что $Z(t)$ — стационарный гауссовский случайный процесс с нулевым средним, спектральная плотность которого $S_z(f)$ отлична от нуля лишь при $|f \pm f_0| < f_0$. Опишем свойства процессов $X(t)$ и $Y(t)$ в представлении (3.8). Обозначим через $\hat{Z}(t)$ случайную функцию, являющуюся откликом на $Z(t)$ линейной системы с частотной характеристикой (3.4), и пусть

$$Z_+(t) = Z(t) + i\hat{Z}(t). \quad (3.9)$$

Процессы $Z(t)$ и $\hat{Z}(t)$ имеют нулевые математические ожидания, а их совместное распределение гауссовское (почему?). Отвечающие этим процессам спектральные плотности описываются соотношениями

$$\begin{aligned} S_{z\hat{z}}(f) &= -S_{\hat{z}z}(f) = (i \operatorname{sgn} f) S_z(f), \\ S_{\hat{z}}(f) &= |i \operatorname{sgn} f|^2 S_z(f) = \begin{cases} S_z(f), & f \neq 0, \\ 0, & f = 0, \end{cases} \end{aligned} \quad (3.10)$$

где

$$\operatorname{sgn} f = \begin{cases} +1, & f > 0, \\ 0, & f = 0, \\ -1, & f < 0. \end{cases} \quad (3.11)$$

Согласно (3.7), процессы $X(t)$ и $Y(t)$ определяются равенствами

$$\begin{aligned} X(t) &= \operatorname{Re}\{e^{-i2\pi f_0 t} Z_+(t)\} = Z(t) \cos 2\pi f_0 t + \hat{Z}(t) \sin 2\pi f_0 t, \\ Y(t) &= \operatorname{Im}\{e^{-i2\pi f_0 t} Z_+(t)\} = \hat{Z}(t) \cos 2\pi f_0 t - Z(t) \sin 2\pi f_0 t, \end{aligned} \quad (3.12)$$

из которых видно, что совместные распределения процессов $X(t)$ и $Y(t)$ также гауссовские, а отвечающие им средние равны нулю. Менее очевидно, что эти процессы стационарные; однако, отправляясь от равенств (3.12) и используя соотношения (3.10) для преобразования вида функций $R_x(t + \tau, t)$, $R_y(t + \tau, t)$ и $R_{xy}(t + \tau, t)$, получаем [1, разд. 5.1], что эти ковариационные функции зависят лишь от τ , а отвечающие им спектральные плотности имеют вид

$$\begin{aligned} S_x(f) = S_y(f) &= \frac{1}{2} \{S_z(f + f_0) + S_z(f - f_0) + \\ &+ [\operatorname{sgn}(\cdot) S_z(\cdot)] * [\delta(\cdot + f_0) - \delta(\cdot - f_0)]\}, \\ S_{xy}(f) = S_{yx}^*(f) &= \frac{i}{2} \{S_z(f + f_0) - S_z(f - f_0) + \\ &+ [\operatorname{sgn}(\cdot) S_z(\cdot)] * [\delta(\cdot + f_0) + \delta(\cdot - f_0)]\}. \end{aligned} \quad (3.13)$$

Полученные выражения, по-видимому, проще всего интерпретировать графически, зафиксировав некоторую спектральную плотность $S_z(f)$ шума с ограниченной полосой частот (см. рис. 3.1). Из рисунка видно, что

(i) плотности S_x , S_y и S_{xy} отвечают низкочастотным шумам, т. е. отличны от нуля лишь при $|f| < f_0$;

(ii) если функция $S_z(f)$ симметрична относительно f_0 при $f > 0$, то $S_{xy}(f) = 0$.

Суммируя сказанное, приходим к представлению

$$Z(t) = \operatorname{Re} \{V(t) e^{i2\pi f_0 t}\} = X(t) \cos 2\pi f_0 t - Y(t) \sin 2\pi f_0 t, \quad (3.14)$$

причем, если процесс $Z(t)$

- (i) имеет нулевое среднее,
- (ii) гауссовский,
- (iii) стационарный с ограниченным спектром и спектральной плотностью, симметричной при $f > 0$ относительно f_0 ,

то процессы $X(t)$ и $Y(t)$

- (i) имеют нулевые средние,
- (ii) совместно гауссовские,
- (iii) совместно стационарные, некоррелированные (а следовательно, независимые в силу гауссовости) и имеют общую спектральную плотность в ограниченной полосе частот.

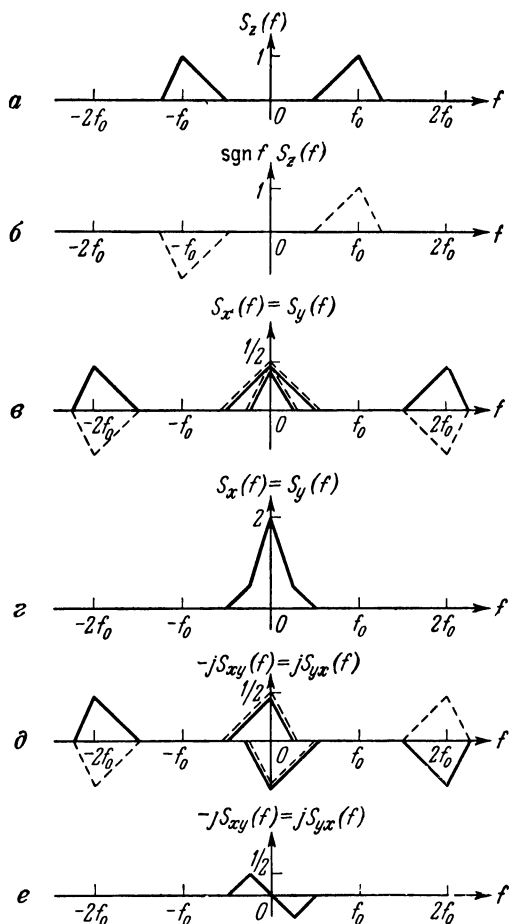


Рис. 3.1. Спектральные функции, используемые при нахождении $S_x(f)$ и $S_{xy}(f)$: *a*) $S_z(f)$, *б*) $S_z(f) \text{sgn } f$, *в*) компоненты функции $S_x(f)$, *г*) результирующая функция $S_x(f)$, *д*) компоненты функции $S_{xy}(f)$, *е*) результирующая функция $S_{xy}(f)$.

Перечисленные свойства эквивалентны тому, что $V(t) = X(t) + iY(t)$ является комплексным стационарным гауссовским процессом с нулевым средним и

$$\mathbf{E}\{V_{t+\tau}V_t^*\} = R_v(\tau) = 2R_x(\tau) = 2\mathcal{F}^{-1}\{S_x(f)\}, \quad (3.15a)$$

$$\mathbf{E}\{V_{t+\tau}V_t\} = 0, \quad (3.15b)$$

где \mathcal{F}^{-1} — обратное преобразование Фурье.

Разложение Карунена — Лозва для комплексных процессов

Рассмотрим разложение Карунена — Лозва комплекснозначного процесса $V(t)$ с указанными свойствами. В общем виде это разложение рассматривается в книгах Лозва [2, разд. 34.5], Давенпорта и Рута [3, разд. 6.4]. Здесь мы лишь сформулируем ряд основных фактов, относящихся к этому разложению.

Как обычно, введем в пространстве $L_2[0, T]$ комплекснозначных функций, определенных и интегрируемых на отрезке $[0, T]$, скалярное произведение

$$(f, g) = \int_0^T dt f(t) g^*(t). \quad (3.16)$$

Интегральный оператор, соответствующий ковариационной функции $R_v(t-s) = \mathbf{E}\{V_t V_s^*\}$, определяется формулой

$$[R_v f](t) = \int_0^T ds R_v(t-s) f(s), \quad 0 \leq t \leq T. \quad (3.17)$$

Подставляя в (3.17) равенство $R_v(s-t) = R_v^*(t-s)$, получаем, что оператор R_v симметричный, т. е.

$$(R_v f, g) = (f, R_v g). \quad (3.18)$$

Напомним, что оператор называется *положительно определенным*, если для любой функции φ , норма которой отлична от нуля, т. е.

$$\|\varphi\| = +(\varphi, \varphi)^{1/2} > 0,$$

выполняется неравенство

$$(\varphi, R_v \varphi) > 0. \quad (3.19)$$

Заметим, что если функция $R_v(\tau)$ ограничена и абсолютно интегрируема на $(-\infty, \infty)$, то оператор R_v положительно определен.

Задача 3.2. Доказать это утверждение. [Указание. Преобразуйте сначала двойной интеграл в выражении для $(\varphi, R_v \varphi)$ к обычному интегралу, содержащему спектральную плотность $S_v(f)$ и преобразование Фурье $\Phi(f) = \mathcal{F}[\varphi(t)]$. Далее, используя тот факт, что функция $\varphi(t)$ отлична от нуля лишь на отрезке $[0, T]$, покажите, что $\Phi(f)$ не может равняться нулю на невырожденном интервале, не будучи равным нулю тождественно.]

Рассмотрим теперь нормированные собственные функции уравнения

$$\lambda_k \varphi_k(t) = \int_0^T ds R_v(t-s) \varphi_k(s), \quad 0 \leq t \leq T, \quad (3.20)$$

или, в операторных обозначениях,

$$\lambda_k \varphi_k = R_v \varphi_k. \quad (3.20')$$

В дальнейшем мы будем предполагать, что функция $R_v(\tau)$ ограничена, непрерывна и абсолютно интегрируема (а следовательно, интегрируема в квадрате). При этих предположениях оператор R_v обладает следующими свойствами:

а) Собственные функции φ_k , отвечающие различным собственным значениям, ортогональны в силу симметричности оператора R_v , т. е. $(\varphi_k, \varphi_j) = 0$ при $\lambda_k \neq \lambda_j$; при этом собственные значения λ_k действительны.

б) В силу интегрируемости в квадрате функции $R_v(\cdot)$ существует не более счетного числа различных собственных значений, причем каждому собственному значению отвечает не более конечного числа линейно независимых собственных функций; эти функции можно выбрать попарно ортогональными, и впредь мы будем считать, что они выбраны именно таким образом.

в) Непрерывность функции $R_v(\cdot)$ влечет непрерывность функций φ_k .

г) Если оператор R_v положительно определен, то система функций φ_k полна в $L_2[0, T]$.

Читатель, интересующийся доказательствами этих утверждений, может познакомиться с ними по книге Рисса и Надя [4, разд. 97]. Мы видим, таким образом, что при соответствующих предположениях множество функций φ_k , связанных с уравнением (3.20), образует ортонормированный базис в $L_2[0, T]$. В дальнейшем мы будем считать, что соответствующие этой системе собственные значения λ_k расположены в порядке убывания: $\lambda_1 \geq \lambda_2 \geq \dots$.

Если функция $R_v(\tau)$ непрерывна, то процесс $V(t)$ можно разложить по ортонормированному базису $\{\varphi_k\}$:

$$V(t) = \lim_{N \rightarrow \infty} \sum_{k=1}^N V_k \varphi_k(t), \quad 0 \leq t \leq T, \quad (3.21)$$

где

$$V_k = (V, \varphi_k) = \int_0^T dt V(t) \varphi_k^*(t). \quad (3.22)$$

Равенство (3.21) понимается в том смысле, что при $N \rightarrow \infty$ разность $V(t) - \sum_{k=1}^N V_k \varphi_k(t)$ стремится в среднем-квадратическом к нулю для любого $0 \leq t \leq T$. Выбор функций φ_k , отвечающих уравнению (3.20), в качестве базиса разложения процесса $V(t)$ основан на том, что коэффициенты V_k и V_j^* , входящие в разложение, некоррелированы при $j \neq k$; в частности, если $E\{V_k\} = E\{V_j^*\} = 0$, то

$$E\{V_k V_j^*\} = \lambda_k \delta_{jk}. \quad (3.23)$$

Доказательство соотношений (3.21) и (3.23) можно найти в книге Лозва [2, разд. 34.5]; при этом доказательство (3.21) опирается на теорему Мерсера, с которой интересующийся читатель может познакомиться по книге Рисса и Надя [4, разд. 98].

Задача 3.3. Пусть $V(t)$ — гауссовский случайный процесс с нулевым средним, и пусть выполнено

соотношение (3.15b). Показать, что коэффициенты разложения V_1, V_2, \dots, V_M распределены с плотностью

$$f_{V_1 \dots V_M}(v_1, \dots, v_M) = \left[\prod_{k=1}^M (\pi \lambda_k)^{-1} \right] \exp \left[- \sum_{k=1}^M \frac{|v_k|^2}{\lambda_k} \right]. \quad (3.24)$$

3.2. ОТНОШЕНИЕ ПРАВДОПОДОБИЯ

Вернемся к проблеме оценки параметра α по наблюдаемому процессу

$$\begin{aligned} Z(t) &= s(t, \alpha) + Z_n(t) = \\ &= \operatorname{Re} \{ b s_\beta(t) e^{i\omega_0 t} \} + \operatorname{Re} \{ \mathcal{N}(t) e^{i\omega_0 t} \}, \quad 0 \leq t \leq T, \end{aligned} \quad (3.25)$$

где через $\mathcal{N}(t)$ обозначено комплексное низкочастотное представление процесса $Z_n(t)$. Заметим, что при наших предположениях о процессе $Z_n(t)$ коэффициенты в разложении $\mathcal{N}(t)$ имеют совместное распределение с плотностью вида (3.24).

Положим

$$W(t) = b s_\beta(t) + \mathcal{N}(t), \quad (3.26)$$

так что

$$Z(t) = \operatorname{Re} \{ W(t) e^{i\omega_0 t} \}.$$

Поскольку действительную и мнимую части $W(t)$ можно получить квадратурной синхронной демодуляцией процесса $Z(t)$ (демодуляцией посредством $\cos \omega_0 t$ и $\sin \omega_0 t$), наблюдение процесса $Z(t)$ эквивалентно наблюдению процесса $W(t)$. Поэтому мы найдем отношение правдоподобия для процесса $W(t)$.

Заметим, что построение отношения правдоподобия важно как для

- (i) нахождения оценок максимального правдоподобия параметров b и β по наблюдению $W(t)$,

так и для

- (ii) получения границы Крамера — Рао.

Учитывая полноту систему $\{\varphi_k\}$, представим $s_\beta(t)$ в виде

$$s(t) = \beta \sum_{k=1}^{\infty} s_k(\beta) \varphi_k(t),$$

где

$$s_k(\beta) = (s_\beta, \varphi_k) = \int_0^T dt s_\beta(t) \varphi_k^*(t).$$

Таким образом, $W(t)$ можно записать как

$$W(t) = \lim_{M \rightarrow \infty} \sum_{k=1}^M W_k \varphi_k(t), \quad 0 \leq t \leq T, \quad (3.27)$$

где

$$W_k = b s_k(\beta) + \mathcal{N}_k, \quad \mathcal{N}_k = \int_0^T dt \mathcal{N}(t) \varphi_k^*(t). \quad (3.28)$$

Из соотношений (3.24) и (3.28) следует, что при наших исходных предположениях о процессе $Z_n(t)$ совместную плотность распределения коэффициентов W_k , фигурирующих в равенстве (3.27), можно представить в виде

$$\begin{aligned} f_{W_1 \dots W_M}(w_1, \dots, w_M) = \\ = \left[\prod_{k=1}^M (\pi \lambda_k)^{-1} \right] \exp \left[- \sum_{k=1}^M \frac{|w_k - b s_k(\beta)|^2}{\lambda_k} \right], \end{aligned} \quad (3.29)$$

где λ_k — собственные значения оператора R_n на отрезке $[0, T]$. Таким образом, мы найдем $\Lambda_{b, \beta}(W)$, вычислив сначала $\Lambda_{b, \beta}(W_1, \dots, W_M)$ и перейдя затем к пределу при $M \rightarrow \infty$.

При вычислении $\Lambda_{b, \beta}$ выберем в качестве опорной точки α_0 пару $b = 0$, β произвольно. Тогда, рассматривая отношение двух плотностей, получаем

$$\begin{aligned} \ln \Lambda_{b, \beta}(W_1, \dots, W_M) = \\ = \sum_{k=1}^M \{2 \operatorname{Re}[b s_k(\beta) W_k^*] - |b s_k(\beta)|^2\} / \lambda_k. \end{aligned} \quad (3.30)$$

Сходимость отношения правдоподобия

Нас будет интересовать предел правой части равенства (3.30) при $M \rightarrow \infty$. Выясним сначала, при каких условиях этот предел существует. Обозначим

$$U_k = \{2 \operatorname{Re}[b s_k(\beta) W_k^*] - |b s_k(\beta)|^2\} / \lambda_k.$$

Заметим, что случайные величины U_k независимы.

Задача 3.4. Показать, что

$$\begin{aligned} \mathbf{E}\{U_k\} &= |b|^2 |s_k(\beta)|^2 / \lambda_k, \\ \text{var}\{U_k\} &= 2 |b|^2 |s_k(\beta)|^2 / \lambda_k. \end{aligned}$$

Воспользуемся теперь следующей теоремой Колмогорова [5, стр. 102]: если U_1, U_2, \dots — последовательность независимых случайных величин и

$$\begin{aligned} V_M &= \sum_{k=1}^M U_k, \\ \sum_{k=1}^{\infty} \mathbf{E}\{U_k\} &< \infty, \quad \sum_{k=1}^{\infty} \text{var}\{U_k\} < \infty, \end{aligned}$$

то V_M сходится с вероятностью 1, а также в среднеквадратическом к некоторой случайной величине V , причём

$$\mathbf{E}\{V\} = \sum_{k=1}^{\infty} \mathbf{E}\{U_k\}, \quad \text{var}\{V\} = \sum_{k=1}^{\infty} \text{var}\{U_k\}.$$

В нашем случае величины U_k являются совместно гауссовскими, и, следовательно, случайная величина V также распределена по гауссовскому закону. Таким образом, если

$$\Sigma = \sum_{k=1}^{\infty} |s_k(\beta)|^2 / \lambda_k < \infty, \quad (3.31)$$

то

$$\ln \Lambda_{b, \beta}(W) = 2 \sum_{k=1}^{\infty} \text{Re}[b s_k(\beta) W_k^*] / \lambda_k - |b|^2 \Sigma \quad (3.32)$$

есть гауссовская случайная величина со средним $|b|^2 \Sigma$ и дисперсией $2|b|^2 \Sigma$.

Интересно поэтому понять, в каких случаях справедливо соотношение (3.31). Очевидно, для этого необходимо, чтобы слагаемые $|s_k(\beta)|^2 / \lambda_k$ достаточно быстро убывали — ряд из них должен сходиться. Физически это отвечает тому, что при больших частотах энергия сигнала

падает быстрее мощности шума. Келли, Рид и Рут [6, часть I] показали, что ряд (3.31) сходится, если

- (i) $s_\beta(t)$ — отклик фильтра H на сигнал конечной энергии;
- (ii) $\mathcal{N}(t)$ — отклик того же фильтра на белый шум (приблизительно постоянная спектральная плотность на всех частотах).

Учитывая, что в практической ситуации предположения (i) и (ii) оправданы, мы будем считать условие (3.31) выполненным. Таким образом, в дальнейшем при нахождении оценок максимального правдоподобия и построения границы Крамера — Рао оптимальным соотношением для нас будет равенство (3.32)¹⁾.

Заметим, однако, что представление в виде бесконечного ряда не слишком удобно при обращении с величиной $\ln \Lambda$. Поэтому предположим, что

$$\sum_{k=1}^{\infty} |s_k(\beta)|^2 / \lambda_k^2 < \infty \quad (3.33)$$

(это условие сильнее вытекающего из него неравенства (3.31)). Тогда можно определить семейство интегрируемых в квадрате функций

$$f_\beta(t) = \sum_{k=1}^{\infty} \frac{s_k(\beta)}{\lambda_k} \varphi_k(t), \quad 0 \leq t \leq T. \quad (3.34)$$

Непосредственной подстановкой убеждаемся, что $f_\beta(t)$ — решение интегрального уравнения

$$\int_0^T ds R_n(t-s) f_\beta(s) = s_\beta(t), \quad 0 \leq t \leq T, \quad (3.35)$$

¹⁾ Отметим, что интересующая нас величина $\Lambda_{b,\beta}(W)$ есть производная Радона — Никодима меры, связанной с процессом $W(t)$ и отвечающей значениям b, β , по мере, соответствующей $0, \beta$, а правая часть равенства (3.32) представляет собой предел логарифма производной Радона — Никодима, отвечающей мерам, порождаемым конечномерным представлением этого процесса. Однако, как показано Гренандером [7], этот предел равен логарифму производной мер, определяемых по полностью наблюдаемому процессу $W(t)$. Мы не будем здесь вдаваться подробнее в связанные с этим довольно тонкие рассуждения.

или, в операторных обозначениях,

$$R_n \hat{f}_\beta = s_\beta, \quad \hat{f}_\beta = R_n^{-1} s_\beta.$$

Решение интегрального уравнения (3.35) является единственным практическим путем для нахождения $\hat{f}_\beta(t)$. В том случае, когда R_n — ковариационная функция, отвечающая рациональной спектральной плотности, решение уравнения (3.35) можно свести к решению линейного дифференциального уравнения с постоянными коэффициентами [3, разд. A2-3].

Задача 3.5. Пусть

$$(\hat{f}_\beta, s_\beta) = \int_0^T dt \hat{f}_\beta(t) s_\beta^*(t), \quad (3.36)$$

$$(\hat{f}_\beta, W) = \int_0^T dt \hat{f}_\beta(t) W^*(t). \quad (3.37)$$

Показать, что число (\hat{f}_β, s_β) действительно и

$$\ln \Lambda_{b, \beta} = 2 \operatorname{Re} [b(\hat{f}_\beta, W)] - |b|^2 (\hat{f}_\beta, s_\beta), \quad (3.38)$$

$$E \{ \ln \Lambda_{b, \beta} \} = |b|^2 (\hat{f}_\beta, s_\beta), \quad (3.39)$$

$$\operatorname{var} \{ \ln \Lambda_{b, \beta} \} = 2 |b|^2 (\hat{f}_\beta, s_\beta). \quad (3.40)$$

Показать также, что $\ln \Lambda_{b, \beta}$ — гауссовская случайная величина.

Случай белого шума

Начиная с этого раздела, мы будем интересоваться тем, как реально находить оценки максимального правдоподобия параметров b и β и строить матрицу G , фигурирующую в неравенстве Крамера — Рао. В соответствии с этим сосредоточим внимание на ситуации, представляющей большой практический интерес и, к счастью, допускающей численное решение. Точнее, предположим, что $Z_n(t)$ и $s_\alpha(t)$ — отклики соответственно на шум и на сигнал фильтра, имеющего постоянную частотную характеристику на ограниченном интервале частот длиной $2W_0$, много большем, чем полоса частот сигнала $s_\alpha(t)$.

Таким образом, $s_\alpha(t)$ представляет собой неискаженный вариант слабого сигнала, принимаемого антенной, а сигнал $Z_n(t)$ мы считаем возникающим вследствие широкополосного шума (например, теплового) на входе

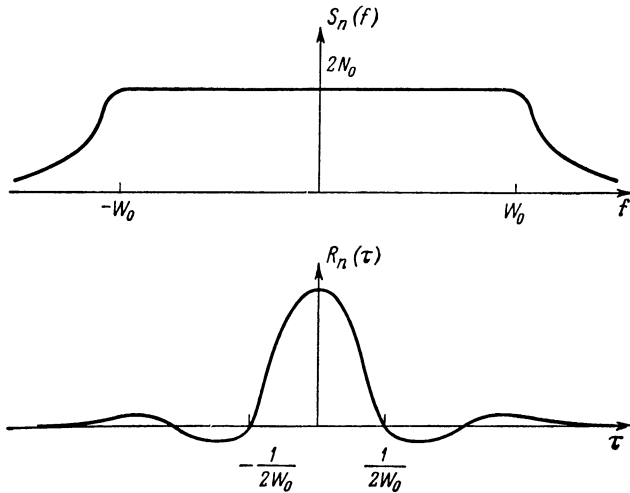


Рис. 3.2. Вид спектра и корреляционной функции «белого» шума.

фильтра, причем спектральная плотность шума постоянна на интервале, большем, чем полоса пропускания фильтра.

Поэтому

$$S_{z_n}(f) \approx \begin{cases} N_0/2 & \text{при } |f \pm f_0| \leq W_0, \\ \text{монотонно убывает} & \text{при } |f \pm f_0| > W_0, \end{cases}$$

и, следовательно,

$$S_n(f) \approx \begin{cases} 2N_0 & \text{при } |f| \leq W_0, \\ \text{монотонно убывает} & \text{при } |f| > W_0. \end{cases} \quad (3.41)$$

Соответствующая ковариационная функция $R_n(\tau)$ имеет вид, показанный на рис. 3.2, с шириной центрального пика, равной примерно $1/(2W_0)$. Однако если полоса частот $s_\beta(t)$ пренебрежимо мала по сравнению с W_0 , как мы предположили, то изменением $s_\beta(t)$ на интервале

длины $1/(2W_0)$ можно пренебречь. Таким образом, $R_n(\tau)$ по отношению к $s_\beta(t)$ приближенно есть импульс, ограничивающий площадь

$$\int_{-\infty}^{\infty} d\tau R_n(\tau) = S_n(0) = 2N_0,$$

и, следовательно,

$$\int_{-\infty}^{\infty} ds R_n(t-s) s_\beta(s) \approx 2N_0 s_\beta(t), \quad 0 \leq t \leq T.$$

Итак, $(1/2N_0)s_\beta(s)$ — приближенное решение интегрального уравнения (3.35), и мы полагаем

$$\hat{f}_\beta(t) \approx \frac{1}{2N_0} s_\beta(t), \quad 0 \leq t \leq T, \quad (3.42)$$

и ¹⁾

$$\ln \Lambda_{b, \beta}(W) \approx \frac{1}{2N_0} \{2 \operatorname{Re} [b(s_\beta, W)] - \|b\|^2 s_\beta\|^2\}. \quad (3.43)$$

Напомним, что $\ln \Lambda$ имеет гауссовское распределение со средним

$$E\{\ln \Lambda\} \approx \|b\|^2 s_\beta\|^2 / 2N_0 \quad (3.44)$$

и дисперсией

$$\operatorname{var}\{\ln \Lambda\} \approx \|b\|^2 s_\beta\|^2 / N_0. \quad (3.45)$$

3.3. РАДИОЛОКАЦИОННОЕ ИЗМЕРЕНИЕ ДАЛЬНОСТИ И ЗАДАЧА ОЦЕНИВАНИЯ

Опишем способы построения оценок максимального правдоподобия и границы точности, присущей задаче оценивания, в случае радиолокационного измерения дальности. Мы будем постепенно приспособлять наши вычисления к проблемам радиолокации, хотя значитель-

¹⁾ Математически строгий вывод этого выражения для отношения правдоподобия в случае, когда шум гауссовский белый, можно найти, например, в книге [15, стр. 90]. При этом (s_β, W) понимается как стохастический интеграл по наблюдаемому процессу. — *Прим. ред.*

ную часть наших результатов можно применить к более широкому классу прикладных задач.

При исследовании задач радиолокации необходимо выяснить, каким образом влияют интересующие нас параметры на фазу отраженного (принятого) сигнала. Такие важные параметры, как дальность цели или ее отражательная способность, влияют, конечно, на сдвиг фазы. С другой стороны, к изменению фазы приводят и такие факторы, как дрейф гетеродина приемника, не имеющие отношения к цели. Мы будем использовать обозначение $\varphi = \angle b$ для описания тех воздействий, которые не содержатся в β ; влияние же всех параметров, связанных с целью, которые мы хотим оценить, включим в β . Это приводит к необходимости различать следующие случаи:

1. Угол φ известен, влияние неизвестных параметров целиком содержится в β ; гетеродин приемника считается в этом случае вполне стабильным.

2. Параметр φ неизвестен и нуждается в оценке наряду с вектором β ; по-прежнему β и φ различаются своей связью с целью и соответственно с факторами, не относящимися к цели.

3. Угол φ является случайной величиной, значение которой неизвестно. Здесь мы рассмотрим лишь полностью некогерентный случай, когда случайная величина φ равномерно распределена на интервале $[0, 2\pi)$.

Случай 1 принято называть когерентным (случай известной фазы); случаи 2 и 3 называются некогерентными.

Первым нашим предположением при анализе этих возможностей, относящимся именно к задаче радиолокации, будет предположение о том, что

$$\|s_\beta\|^2 \equiv 1. \quad (3.46)$$

Заметим, что хотя на самом деле $\|s_\beta\|$ меняется в зависимости от дальности цели, этим изменением можно пренебречь в тех пределах дальности, которые обычно рассматриваются. Такая нормировка выбрана исключительно для удобства; при этом изменение амплитуды сказывается лишь на $A \equiv |b|$,

В первом из перечисленных выше случаев $\ln \Lambda$ можно представить в виде

$$\ln \Lambda_{b, \beta} = \frac{A}{N_0} \operatorname{Re}[(s_\beta, e^{-i\Phi} W)] - \frac{A^2}{2N_0}. \quad (3.47)$$

Таким образом, оценка максимального правдоподобия $\hat{\beta}_{\text{МП}}$ — это то значение β , которое максимизирует действительную часть скалярного произведения

$$(s_\beta, e^{-i\Phi} W),$$

и, следовательно, $\hat{\beta}_{\text{МП}}$ не зависит от A .

Задача 3.6. Показать, что если функция $\ln \Lambda_{b, \beta}$ определена соотношениями (3.43), (3.46), то значение b , максимизирующее $\ln \Lambda_{b, \beta}$, равно

$$\hat{b}_{\text{МП}} = (s_\beta, W)^*$$

и

$$\sup_b \ln \Lambda_{b, \beta} = \frac{1}{2N_0} |(s_\beta, W)|^2. \quad (3.48)$$

[Указание. Выделите полный квадрат.]

Во втором случае, когда параметр Φ неизвестен, оценка максимального правдоподобия вектора β — это значение β , максимизирующее правую часть равенства (3.48). Заметим, что эта оценка не зависит от A и Φ .

Рассмотрим теперь случай, когда Φ — равномерно распределенная случайная величина. Если мы вернемся к наблюдению только M координат W_1, \dots, W_M , то найдем, что отвечающая им плотность вероятности, входящая в отношение правдоподобия, равна

$$f_{A, \beta}(w_1, \dots, w_M) = \frac{1}{2\pi} \int_0^{2\pi} d\Phi_{A, \beta}(w_1, \dots, w_M).$$

Рассуждая так же, как при выводе формулы для $\Lambda_{b, \beta}$, получаем, что в этом случае интересующее нас отношение правдоподобия имеет вид

$$\Lambda_{A, \beta}(W) = \frac{1}{2\pi} \int_0^{2\pi} d\Phi \Lambda_{b, \beta}(W), \quad (3.49)$$

Задача 3.7. Показать, что

$$\Lambda_{A, \beta}(W) = \exp \left[-\frac{A^2 \|s_{\beta}\|}{2N_0} \right] I_0 \left(\frac{A | (s_{\beta}, W) |}{N_0} \right), \quad (3.50)$$

где $I_0(x)$ — функция Бесселя первого рода нулевого порядка, определяемая формулой

$$I_0(x) = \frac{1}{2\pi} \int_0^{2\pi} d\theta \exp(x \cos \theta).$$

Поскольку I_0 монотонно возрастает, при нашем предположении (3.46) (т. е. $\|s_{\beta}\| = 1$) оценка максимального правдоподобия вектора β по-прежнему совпадает с его значением, максимизирующим

$$| (s_{\beta}, W) | = \left| \int_0^T dt s_{\beta}(t) W^*(t) \right|. \quad (3.51)$$

Функция неопределенности

Введем функцию корреляции между двумя детерминированными сигналами $s_{\beta_1}(t)$ и $s_{\beta_2}(t)$ как

$$\int dt s_{\beta_1}(t) s_{\beta_2}^*(t).$$

В том случае, когда область интегрирования не указана, будем считать, что ее образуют те значения переменной интегрирования, для которых подинтегральное выражение отлично от нуля. В дальнейшем мы ограничимся специальным случаем, когда функция корреляции зависит лишь от разности $\beta_1 - \beta_2$, и обозначим эту функцию

$$\chi(\beta_1 - \beta_2) = (s_{\beta_1}, s_{\beta_2}) = \int dt s_{\beta_1}(t) s_{\beta_2}^*(t). \quad (3.52)$$

Такое допущение представляет интерес не только в задачах радиолокации, но и в некоторых случаях импульсной модуляции, таких, например, как время-импульсная модуляция. В силу условия (3.46) и неравенства Шварца имеем

$$| (s_{\beta_1}, s_{\beta_2}) | = | \chi(\beta_1 - \beta_2) | \leq \chi(0) = 1, \quad (3.53)$$

Согласно установившейся терминологии, функция $|\chi(\beta)|^2$ называется *функцией неопределенности*.

Если ϕ неизвестно, т. е. либо рассматривается как неизвестный параметр, либо предполагается равномерно распределенной случайной величиной, правило для построения оценки, как мы видели, состоит в выборе значения β , максимизирующего $|(s_\beta, W)|$. Это заставляет нас более тщательно изучить величину (s_β, W) . Заметим, что при любом фиксированном значении β скалярное произведение (s_β, W) является случайной величиной. Далее, для любого набора значений параметров β_k , $k = 1, \dots, M$, соответствующие случайные величины (s_{β_k}, W) имеют совместное гауссовское распределение. Таким образом, (s_β, W) представляет собой *гауссовский случайный процесс*, зависящий от векторного параметра β . (Читателю следует вспомнить, что в теории случайных процессов не требуется, чтобы параметр играл роль времени; этому случаю просто отвечает один важный и наиболее распространенный вид процессов.) Поскольку гауссовский случайный процесс полностью определяется своими средним и ковариационной функцией, интересно найти эти величины.

Задача 3.8. Обозначим снова через β' «истинное», или фактическое, значение параметра β . Запишем наше предположение о виде спектральной плотности шума как

$$R_n(\tau) \approx 2N_0\delta(\tau). \quad (3.54)$$

Показать, что среднее значение процесса (s_β, W) задается равенством

$$E\{(s_\beta, W)\} = b'^* \chi(\beta - \beta'), \quad (3.55)$$

а его ковариационная функция равна

$$\begin{aligned} E\{(s_{\beta_1}, W)(s_{\beta_2}, W)^*\} - E\{(s_{\beta_1}, W)\}E\{(s_{\beta_2}, W)^*\} = \\ = 2N_0\chi(\beta_1 - \beta_2). \end{aligned} \quad (3.56)$$

Исследуем теперь качественно, каким образом выбор функции неопределенности $|\chi(\cdot)|^2$ влияет на величину ошибок при оценке параметра β . Для того чтобы представить себе эту зависимость наглядно, проще всего

рассмотреть случай скалярного параметра β ; это позволит нам графически изобразить выборочные функции процесса, а сделанные нами замечания будут справедливы и для n -мерного вектора β .

Методически удобно представить (s_β, W) в виде среднего плюс аддитивный шум с нулевым математическим ожиданием:

$$(s_\beta, W) = b'^* \chi(\beta - \beta') + \mathcal{N}_\beta.$$

В силу (3.56) корреляционная функция процесса \mathcal{N}_β равна $2N_0\chi(\beta - \beta')$. Таким образом, вид выборочных функций процесса \mathcal{N}_β в частотном смысле должен напоминать функцию χ . Поучительно будет привести вид выборочных функций процесса (s_β, W) при

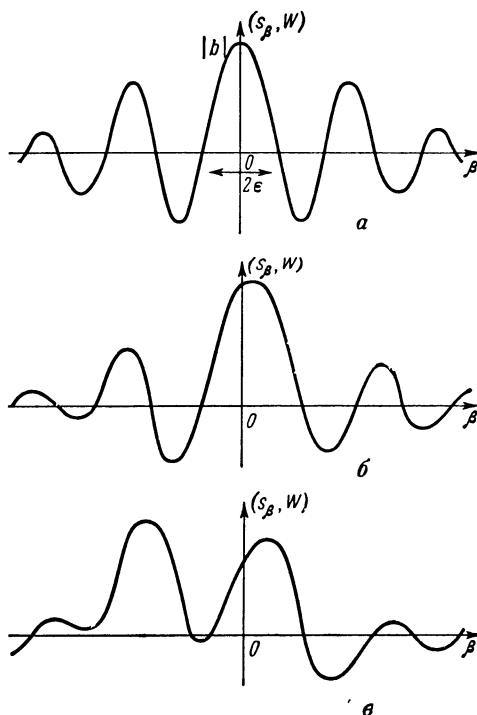
- а) $\mathcal{N}_\beta = 0$,
- б) $2N_0 \ll |b|^2$,
- в) $2N_0 \approx |b|^2$.

На рис. 3.3 показаны эти три случая. Центральная часть графика χ хорошо аппроксимируется квадратичной функцией от β в интервале длины 2ϵ , симметричном относительно нуля.

Напомним, что когда фаза неизвестна, правило построения оценки состоит в выборе значения β , максимизирующего $|(s_\beta, W)|$. В случае слабого шума, как видно из рис. 3.3, б, этот максимум лишь несколько сдвинут в пределах центрального интервала длины 2ϵ , и величина смещения прямо пропорциональна стандартному отклонению аддитивного шума; при этом она находится в обратной зависимости от крутизны χ в начале координат. В случае когда величина шума \mathcal{N}_β сравнима с математическим ожиданием, этот локальный максимум не только сдвинут относительно начала координат, но, что весьма вероятно, он уже может перестать быть глобальным максимумом, который при этом, как показано на рис. 3.3, в, может сместиться в точку какого-либо другого локального максимума. Величина ошибки теперь уже значительно превосходит величину смещения центрального локального максимума.

Такие большие ошибки называются *аномальными ошибками*. Как мы видели, они возникают тогда, когда

величина \mathcal{N}_β становится сравнимой с $|\chi(0)| - |\chi(\beta)|$ для β , лежащих вне интервала длины 2ϵ с центром в начале координат.



Р и с. 3.3. Вид выборочных функций процесса (s_β, W) .

Из этих замечаний вытекают следующие требования, предъявляемые к выбору сигнала s_β :

- (i) функция $|\chi(\beta)|^2$ должна иметь острый пик в начале координат,
- (ii) значения $|\chi(\beta)|$ вне интервала длины 2ϵ с центром в начале координат должны быть малы по сравнению с $|\chi(0)|$.

Цель первого требования — оптимизация качества приема при слабых шумах, а второе позволяет макси-

мально увеличить уровень N_0 шума, при котором аномальные ошибки происходят достаточно редко. Эта давно привлекавшая к себе внимание проблема выбора сигналов в настоящее время вызывает особый интерес ([8]—[11]).

Отметим, что задача нахождения вероятности того, что глобальный максимум выборочной функции $|s_B, W|$ окажется вне центрального интервала, эквивалентна нерешенной классической проблеме времени первого прохождения для гауссовского случайного процесса. Влияние аномальных ошибок, или так называемый пороговый эффект, рассматривается с иной (геометрической) точки зрения в разд. 4.3. Задачу построения оценок для величины среднеквадратической ошибки в пороговой области или хотя бы выяснение того, при каких значениях отношения сигнала к шуму A^2/N_0 возникает пороговый эффект, можно считать одной из важных открытых проблем современной теории связи.

Наиболее полезными в пороговой области оказались границы, установленные Зивом и Закаи [12]. По-видимому, эти границы вполне достаточны для задачи радиолокационного измерения дальности, а также для эквивалентной задачи время-импульсной модуляции. Однако все еще неясно, для сколь широкого класса задач эти границы точны или хотя бы достаточно просто вычислимы; исследования в этом направлении вряд ли можно считать завершенными.

В оставшейся части главы мы займемся вычислением границы Крамера — Рао. Эта граница не позволяет судить о качестве приема в пороговой области, где большие аномальные ошибки являются вероятными, однако, как следует из обсуждения задачи 2.7, она весьма точно описывает достижимое качество приема в случае, когда отношение сигнала к шуму A^2/N_0 велико.

Граница Крамера — Рао в задаче радиолокационного оценивания

Для того чтобы применить неравенство Крамера — Рао, обсуждавшееся в гл. 2, достаточно вычислить величины b_{ij} . Прежде чем приступить к этому, заметим, что

как в когерентном, так и в некогерентном случаях наше правило построения оценок параметра β не зависело от A . Чтобы избежать в дальнейшем обращения матриц высокого порядка, будем предполагать, что мощность A возвращенного сигнала известна; таким образом, неизвестные параметры — это

$$(\alpha_1, \dots, \alpha_{n+1}) = (\varphi, \beta_1, \dots, \beta_n).$$

При нашем допущении о том, что $\|s_\beta\| \stackrel{\pm}{=} 1$, из (2.25) и (3.47) следует, что

$$b_{ij} = -\frac{A'}{N_0} \mathbf{E} \left\{ \frac{\partial^2}{\partial \alpha_i \partial \alpha_j} \operatorname{Re} [(e^{i\varphi} s_\beta, W)] \right\} \Big|_{\alpha=\alpha'}.$$

Вынося дифференцирование за знак математического ожидания, получаем

$$\begin{aligned} b_{ij} &= -\frac{A'}{N_0} \left\{ \frac{\partial^2}{\partial \alpha_i \partial \alpha_j} \operatorname{Re} [(e^{i\varphi} s_\beta, A' e^{i\varphi'} s_{\beta'})] \right\} \Big|_{\alpha=\alpha'} = \\ &= -\frac{A'^2}{N_0} \left\{ \frac{\partial^2}{\partial \alpha_i \partial \alpha_j} \operatorname{Re} [e^{i(\varphi-\varphi')} \chi(\beta - \beta')] \right\} \Big|_{\alpha=\alpha'}. \end{aligned} \quad (3.57)$$

Это равенство подтверждает полученный нами ранее качественный вывод о зависимости величины ошибки в случае слабого шума от крутизны функции χ в начале координат.

Остановимся на классическом случае задачи радиолокационного измерения. Предположим, что передаваемый сигнал имеет вид

$$A' \operatorname{Re} \{s(t) e^{i\omega_0 t}\},$$

а возвращенный сигнал равен

$$A' \operatorname{Re} \{s(t - \tau') e^{i\varphi'} e^{i(\omega_0 + \omega')(t - \tau')}\},$$

где τ' — задержка, пропорциональная дальности цели, а ω' — доплеровское смещение, пропорциональное скорости цели. В наших предыдущих обозначениях

$$\begin{aligned} s_\beta(t) &= s(t - \tau') e^{-i\omega_0 \tau'} e^{i\omega' (t - \tau')}, \\ \beta_1 &= \tau, \quad \beta_2 = \omega. \end{aligned} \quad (3.58)$$

Производя несложные алгебраические преобразования, находим, что радиолокационная функция неопределенности имеет вид

$$\begin{aligned} |\chi(\tau - \tau', \omega - \omega')|^2 &= |s_\beta, s_{\beta'}|^2 = \\ &= \left| \int dt s(t) s^*(t + \tau - \tau') e^{it(\omega - \omega')} \right|^2. \end{aligned} \quad (3.59)$$

Отметим физическую интерпретацию этого выражения. Оно представляет собой квадрат величины отклика на сигнал $s(t)$ «рассогласованного» согласованного фильтра¹⁾, расстроенного на $\omega - \omega'$ и наблюдаемого с опозданием в $\tau - \tau'$ с.

Задача 3.9. Найти оценку максимального правдоподобия величин τ и ω в некогерентном случае с помощью операций над узкополосным сигналом $Z(t)$. Преобразовать выражение для величины, которую нужно максимизировать, к виду, удобному для практической реализации, и нарисовать блок-схему такой реализации.

Задача 3.10. Из условия $\|s_\beta\|^2 = 1$ следует, что $|\chi(0, 0)|^2 = 1$.

Доказать, кроме того, что

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} d\tau \int_{-\infty}^{\infty} d\omega |\chi(\tau, \omega)|^2 = 1. \quad (3.60)$$

Равенство (3.60) показывает, что функция неопределенности не может одновременно иметь острый пик в начале координат и быть равномерно малой вне некоторого сужающегося центрального интервала, поскольку ограничиваемый ею объем постоянен. В работах [8]—[11] обсуждается проблема выбора сигнала для получения нужных свойств функции $|\chi|^2$ при ограничениях на пиковую и среднюю мощности, длительность и ширину полосы частот.

¹⁾ Фильтр называется *согласованным* с сигналом $s(t)$, $0 \leq t \leq T$, если его импульсная характеристика имеет вид $h(\tau) = s(T - \tau)$, $0 \leq \tau \leq T$. Этот фильтр применяется для оптимального обнаружения сигнала $s(t)$ или оценивания его амплитуды при наличии аддитивного белого гауссовского шума; см. [1, разд. 8.3 и 7.2].

Опишем некоторые свойства сигнала $s(t)$, облегчающие нахождение величин b_{ij} в равенстве (3.57). Хотя преобразование Фурье чаще определяется как функция обычной частоты f , нам будет удобнее ввести радианную частоту ω и положить

$$S(\omega) = \int_{-\infty}^{\infty} dt s(t) e^{i\omega t}.$$

Заметим, что в силу предположения о нормированности сигнала, т. е. $\|s_{\beta}\| = 1$, выполняется равенство

$$\int_{-\infty}^{\infty} dt |s(t)|^2 = \int_{-\infty}^{\infty} d\omega \frac{1}{2\pi} |S(\omega)|^2 = 1.$$

Зададим усредненные по времени и частотам величины

$$\begin{aligned} \bar{t}_s &= \int_{-\infty}^{\infty} dt t |s(t)|^2, & \bar{t}_s^2 &= \int_{-\infty}^{\infty} dt t^2 |s(t)|^2, \\ \text{var } t_s &= \bar{t}_s^2 - \bar{t}_s^2, \\ \bar{\omega}_s^2 &= \int_{-\infty}^{\infty} d\omega \omega^2 \frac{1}{2\pi} |S(\omega)|^2. \end{aligned} \quad (3.61)$$

Учитывая равенство (3.57) и свойства преобразования Фурье, находим

$$\begin{aligned} b_{\omega\omega} &= \frac{A^2}{N_0} \bar{t}_s^2, & b_{\tau\omega} &= \frac{A^2}{N_0} (\omega_0 - \omega') \bar{t}_s, \\ b_{\tau\tau} &= \frac{A^2}{N_0} [(\omega_0 - \omega')^2 + \bar{\omega}_s^2] \approx \frac{A^2}{N_0} (\omega_0 - \omega')^2, \\ b_{\varphi\tau} &= -\frac{A^2}{N_0} (\omega_0 - \omega'), & b_{\varphi\omega} &= \frac{A^2}{N_0} \bar{t}_s, \\ b_{\varphi\varphi} &= \frac{A^2}{N_0}. \end{aligned} \quad (3.62)$$

Приближенное равенство здесь относится к случаю узкополосного сигнала $\text{Re}[s(t)e^{i\omega_0 t}]$, ширина полосы частот которого мала по сравнению с его несущей частотой.

Задача 3.11. Показать, что в некогерентном случае, когда параметры τ , φ и ω неизвестны, симметричная (3×3) -матрица $G = B^{-1}$ равна

$$G = \frac{N_0}{A^2 \omega_s^2 \text{var } t_s} \begin{bmatrix} \tau & \omega & \varphi \\ \text{var } t_s & 0 & (\omega_0 - \omega') \text{var } t \\ \overline{\omega_s^2} & \overline{\omega_s^2 t_s} & \overline{t_s^2 \omega_s^2} + (\omega_0 - \omega')^2 \text{var } t_s \end{bmatrix}. \quad (3.63)$$

Таким образом,

$$\text{var}(\hat{\tau} - \tau') \geq \frac{N_0}{A^2 \omega_s^2} \quad (3.64)$$

и

$$\text{var}(\hat{\omega} - \omega') \geq \frac{N_0}{A^2 \text{var } t_s}. \quad (3.65)$$

Для сигнала, имеющего прямоугольный вид, т. е. при

$$s(t) = \begin{cases} \frac{1}{\sqrt{T_0}}, & 0 \leq t \leq T_0, \quad T_0 < T, \\ 0 & \text{в остальных случаях,} \end{cases}$$

$\overline{\omega_s^2} = \infty$ и, следовательно, правая часть неравенства (3.64) обращается в нуль. Этому на первый взгляд сомнительному результату наша физическая интуиция склонна искать объяснение в том, что никакой реальный импульс не может иметь строго прямоугольной формы. Однако это лишь уход от вопроса. В действительности трудность здесь связана с применением неравенства Крамера — Рао, поскольку для рассматриваемого импульса не всюду существует производная $\partial^2 \chi(\tau, \omega) / \partial \tau^2$. Здесь вместо границы Крамера — Рао надо воспользоваться некоторыми более общими границами. В рассматриваемом примере можно получить полезные результаты с помощью границы Баранкина [14]. Эти результаты можно еще более уточнить при использовании границы Зива — Закаи [12].

Задача 3.12. Показать, что в когерентном случае известной величины φ матрица G размера 2×2 , отвечающая параметрам τ и ω , имеет вид

$$G \approx \frac{N_0}{A^2 (\omega_0 - \omega')^2 \text{var } t_s} \begin{bmatrix} \overline{t_s^2} & (\omega_0 - \omega') \bar{t}_s \\ (\omega_0 - \omega') \bar{t}_s & (\omega_0 - \omega')^2 \end{bmatrix}. \quad (3.66)$$

Таким образом,

$$\text{var}(\hat{\tau} - \tau') \geq \frac{N_0 \overline{t_s^2}}{A^2 (\omega_0 - \omega')^2 \text{var } t_s} \quad (3.67)$$

и

$$\text{var}(\hat{\omega} - \omega') \geq \frac{N_0}{A^2 \text{var } t_s}. \quad (3.68)$$

Задача 3.13. Сравнить результаты (3.64) и (3.67). Какие трудности могут возникнуть при построении более эффективного метода оценивания в когерентном случае, существование которого подсказывает неравенство (3.67). Для большей конкретности предположите, что $s(t)$ — прямоугольный импульс длительности T_0 , и изобразите графически функцию χ при значениях $T_0 \omega_0 / 2\pi$ порядка нескольких единиц и при значениях $T_0 \omega_0 / 2\pi$, много больших единиц.

СПИСОК ЛИТЕРАТУРЫ

1. Сакрисон (Sakrison D. J.), Communication theory: Transmission of waveforms and digital information, New York, 1968.
2. Лоэв М., Теория вероятностей, ИЛ, М., 1962.
3. Давенпорт В. Б., Рут В. Л., Введение в теорию случайных сигналов и шумов, ИЛ, М., 1960.
4. Рисс Ф., Надь Б. С., Лекции по функциональному анализу, ИЛ, М., 1954.
5. Дуб Дж. Л., Вероятностные процессы, ИЛ, М., 1956.
6. Келли, Рид, Рут (Kelley E. J., Reed I. S., Root W. L.), The detection of radar echoes in noise, *J. Soc. Ind. Appl. Math.*, Part I: 8, 2 (1960), 309—341; Part II: 8, 2 (1960), 381—507.
7. Гренандер У., Случайные процессы и статистические выводы, ИЛ, М., 1961.
8. Титлбаум, Де Кларис (Tittlebaum E. L., DeClariss N.), Linear transformations of the ambiguity function, *IEEE Trans. Inform. Theory*, IT-12 (April 1966), 120—125.
9. Кей, Фоул, Хаггарти (Key E. L., Fowle E. N., Haggarty R. D.), A method of designing signals of large time-bandwidth product,

- The Inst. of Radio Eng. International Convention Record, Part 4, 1961, pp. 146—154.
10. Клудер (Kluder J. R.), The design of radar signals having both high range resolution and high velocity resolution, *Bell Syst. Tech. J.*, **39** (July 1960), 809—820.
 11. Фоул, Келли, Шихан (Fowle E. N., Kelly E. J., Sheehan J. A.), Radar system performance in a dense-target environment, The Inst. of Radio Eng. International Convention Record, Part 4, 1961, p. 136.
 12. Зив, Закаи (Ziv J., Zakai M.), Some lower bounds on signal parameter estimation, *IEEE Trans. Inform. Theory*, **IT-15** (May 1969), 386—391.
 13. Зейдман ((Seidman L.), Performance limitations and error calculations for parameter estimation, *IEEE Proceedings*, **58**, 5 (1970). (Русский перевод: *ТННЭР*, **58**, 5 (1970).)
 14. Сверлинг (Sverling P.), Parameter estimation for waveforms in additive Gaussian noise, *J. Soc. Ind. Appl. Math.*, **7** (June 1959), 152—166.
 - 15¹⁾. Гихман И. И., Скороход А. В., Стохастические дифференциальные уравнения, изд-во «Наукова думка», Киев, 1968.

¹⁾ Добавлено при переводе. — *Прим. ред.*

Глава 4

ИМПУЛЬСНАЯ МОДУЛЯЦИЯ

Рассмотрим ситуацию, в которой случайные величины из последовательности $\dots M_{-1}, M_0, M_1, \dots$ предъявляются по одной в каждые T с, причем значение любой из них, скажем M_k , передается с помощью сигнала продолжительности T с, занимающего интервал времени $kT \leq t < (k+1)T$. Пусть переданный сигнал подвергается в канале воздействию аддитивного гауссовского белого шума. Задача состоит в том, чтобы оценить величину M_k по принятому сигналу.

Последовательность случайных величин M_k может представлять собой отсчеты случайного процесса — сообщения, эффективная ширина полосы которого приближенно равна $1/2T$. Мы не будем обсуждать проблему восстановления этого случайного процесса из последовательности оценок выборочных значений \hat{M}_k . Читатель, интересующийся этим вопросом, может обратиться к разд. 9.1 книги Сакрисона [1], где содержится краткое обсуждение этой проблемы. Мы ограничимся единственной случайной величиной M и единственным переданным сигналом $s(t, m)$, $0 \leq t \leq T$, и будем искать среднеквадратическую ошибку, которую можно достичь при оценке M из сигнала

$$Z(t) = s(t, m) + \mathcal{N}(t), \quad 0 \leq t \leq T. \quad (4.1)$$

Здесь $\mathcal{N}(t)$ — гауссовский случайный процесс, спектр которого «белый» по сравнению с полосой частот сигнала $s(t, m)$.

У читателя может возникнуть вопрос: чем отличается эта задача от поставленной в предыдущей главе? Разница состоит в следующем. В гл. 3 предполагалось (как в случае радара), что функцию $s(t, \alpha)$ менять нельзя, и рассматривалась лишь задача оценивания. Здесь же

рассматривается задача модуляции, или кодирования: при заданных ограничениях выбрать семейство сигналов $s(t, m)$ (семейство потому, что необходимо выбрать свой сигнал для каждого m), минимизирующее достижимую ошибку оценивания. Наши ограничения на выбираемые семейства сигналов связаны с физическими ограничениями, продиктованными потребностями приложений. При введении этих ограничений и при определении эффективности таких импульсных передающих систем ситуация становится много более прозрачной при геометрическом подходе. Наше изложение основано на геометрических представлениях, впервые введенных Котельниковым [2, часть III] и более полно развитых в книге Возенкрафта и Джекобса [3, гл. 8]. Этот подход описан также в книге Сакрисона [1, гл. 9].

Сигналы продолжительности T с, имеющие конечную энергию, будут трактоваться как точки, или векторы, векторного пространства $L_2[0, T]$ функций, интегрируемых с квадратом на отрезке $[0, T]$. Для сигналов $Z(t)$ и $s(t, m)$, $0 \leq t \leq T$, будут употребляться обозначения Z и $s(m)$. Для скалярного произведения и нормы будут использоваться обычные обозначения:

$$(z, s(m)) = \int_0^T z(t) s(t, m) dt,$$

$$\|s(m)\|^2 = \int_0^T |s(t, m)|^2 dt.$$

Мы будем предполагать (исключительно ради упрощения изложения), что сигналы низкочастотные, так что их можно описать функциями, принимающими действительные значения. Результаты этой главы самым прямым образом применимы и в том случае, когда $s(t, m)$ — одна из квадратурно модулированных компонент узкополосного сигнала.

Рассмотрим ограничения на семейство $s(m)$. Пусть \mathcal{M} — множество возможных значений сообщения m ; предполагается, что \mathcal{M} — действительная прямая или ее подмножество. Тогда для любого $m \in \mathcal{M}$ естественно

требовать, чтобы сигнал $s(t, m)$ имел конечную энергию, т. е.

$$\|s(m)\|^2 \leq E \quad \text{для всех } m \in \mathcal{M}. \quad (4.2)$$

Заметим, что любая физическая система должна удовлетворять некоторой форме ограничения полосы частот. Любой сигнал ограниченной энергии можно разложить по полной ортонормальной системе функций φ_k :

$$s(t, m) = \sum_{k=1}^{\infty} s_k(m) \varphi_k(t), \quad 0 \leq t \leq T. \quad (4.3)$$

Если $\{\varphi_k\}$ — множество функций, выбранное так, что они эффективно заполняют полосу частот (например, преобразования Фурье от *зональных сферических функций* [3, приложение 5А; 4; 5]), то для больших N (больше 10) можно найти примерно $N=2WT$ ортонормальных функций из $\{\varphi_k\}$, которые заполняют частоты, не превосходящие W . Можно показать, что это наибольшее число ортонормальных функций, заполняющих частоты, не превосходящие W . Таким образом, в нашей геометрической структуре математически удобный путь введения ограничения полосы частот состоит в том, чтобы потребовать возможность представления

$$s(t, m) = \sum_{k=1}^N s_k(m) \varphi_k(t), \quad 0 \leq t \leq T. \quad (4.4)$$

Заметим, что если последовательность случайных величин образована отсчетами случайного процесса с эффективной шириной полосы частот $1/2T$, то W — ширина полосы нашей системы импульсной модуляции — в N раз больше, чем ширина полосы первоначального сигнала. Число N называется поэтому *множителем расширения полосы частот*.

Ограничение сигналами, представимыми в виде (4.4), удобно с чисто математической точки зрения. Однако оно не выполняется для таких схем, как *время-импульсная модуляция* (ВИМ)

$$s(t, m) = p(t - mT_0),$$

которая заполняет ограниченную полосу частот (полосу функции $p(t)$), но которую нельзя выразить в форме (4.4) для конечного N .

Введем, наконец, условие регулярности функции $s(m)$. Предположим, что функция $s(m)$ дифференцируема; точнее,

$$\left\| \frac{ds(m)}{dm} \right\|^2 = \int_0^T \left| \frac{\partial s(t, m)}{\partial m} \right|^2 dt < \infty \text{ для всех } m \in \mathcal{M}. \quad (4.5)$$

Конечно, из этого условия вытекает непрерывность $s(m)$, т. е.

$$\lim_{h \rightarrow 0} \|s(m+h) - s(m)\| = 0. \quad (4.6)$$

Это значит, что при изменении m в \mathcal{M} функция $s(m)$ вычерчивает непрерывную кривую в $L_2[0, T]$. Мы будем называть эту кривую *линией сигнала* или *кривой сигнала*.

Ограничиваясь аддитивным шумом $\mathcal{N}(t)$, предположим снова, что он представляет собой гауссовский процесс с нулевым средним и «белым» по сравнению с $s(t, m)$ спектром, так что

$$R_n(\tau) \approx \frac{N_0}{2} \delta(\tau). \quad (4.7)$$

Как и выше, мы говорим «белый шум», хотя $\mathcal{N}(t)$ имеет конечную мощность; однако предполагается, что $R_n(\tau)$ ведет себя как импульсная функция с весом $N_0/2$, когда она свертывается с $s(t, m)$. Из гл. 3 следует, что отношение правдоподобия имеет в этом случае вид (если произвести соответствующие упрощения для действительнозначного низкочастотного процесса $\mathcal{N}(t)$)

$$\ln \Lambda_m(Z) \approx \frac{2}{N_0} \{2(s(m), Z) - \|s(m)\|^2\}. \quad (4.8)$$

Приступим теперь к анализу потенциальной эффективности систем импульсной модуляции. Одна из главных целей последующего изложения — объяснить читателю, почему для *увеличения отношения сигнал/шум за*

счет расширения полосы частот необходима нелинейная модуляция и почему отсюда с необходимостью следует пороговый эффект.

4.1. ЭФФЕКТИВНОСТЬ ПРИ МАЛОМ ШУМЕ

Рассмотрим сначала вид оптимальной оценки. Так как наша цель состоит в минимизации среднеквадратической ошибки оценивания

$$\mathcal{E}_2 = \mathbf{E} \{ (M - \hat{M})^2 \},$$

то наилучшая оценка равна условному математическому ожиданию

$$\hat{M} = \mathbf{E} \{ M | Z(t), 0 \leq t \leq T \}.$$

В дальнейшем особенный интерес для нас представляет ситуация, в которой $s(m)$ — нелинейная функция. В этом случае задача точного вычисления условного математического ожидания \hat{M} безнадежна. Вместо этого обратимся к нашему старому другу — оценке максимального правдоподобия (МП). Правда, эта оценка игнорирует априорную информацию об M ; однако можно использовать вместо нее оценку максимальной апостериорной вероятности ¹⁾ (МАВ), т. е. оценку, максимизирующую величину $\ln f(m) + \ln \Lambda_m(z)$. В интересующем нас случае, когда $\mathbf{E} \{ (M - \hat{M})^2 \} \ll \text{var } M$, эффективность оценок МАВ и МП на самом деле совпадает.

Дадим геометрическую интерпретацию оценки МП. Так как $\|z\|$ не зависит от m , то, как видно из (4.8), оценка МП при данном z совпадает со значением m , минимизирующим величину

$$\|z - s(m)\|^2 = -[-\|z\|^2 + 2(s(m), z) - \|s(m)\|^2], \quad (4.9)$$

т. е. со значением m , соответствующим точке кривой сигнала, ближайшей к принятому вектору z . Это правило оценивания иллюстрируется на рис. 4.1, на котором показана двумерная проекция кривой бесконечномерного пространства $L_2 [0, T]$.

¹⁾ Точнее, плотности вероятности. — Прим. перев.

Вычислим \mathcal{E}_2 для слабых шумов. Обозначим через n вектор в $L_2[0, T]$, соответствующий выборочной функции шума, и пусть r_0 — наименьшее число, при котором справедливо неравенство

$$P\{\|\mathcal{N}\| > r_0\} \leq \delta \ll 1. \quad (4.10)$$

Зафиксируем произвольное значение m , скажем m_0 , и проанализируем величину \mathcal{E}_2 в предположении, что пе-

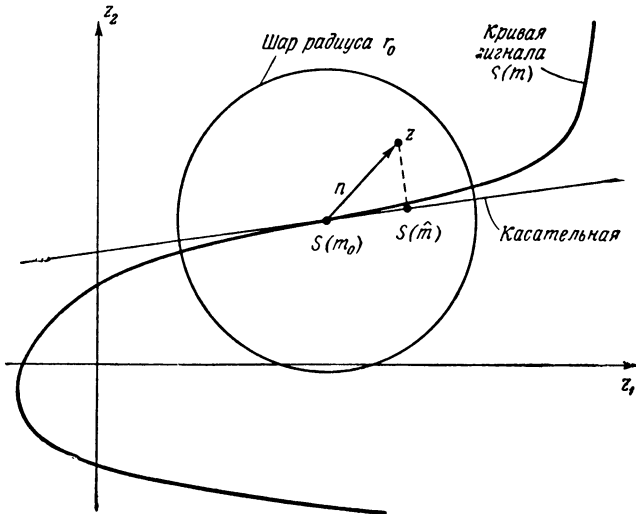


Рис. 4.1. Аппроксимация кривой сигнала касательной для определения ошибки при низком уровне шума.

реданное значение равно m_0 . Предположим, что шум достаточно мал (r_0 достаточно мало), так что

(i) в шаре радиуса r_0 с центром в точке $s(m_0)$ функцию $s(m)$ можно достаточно точно аппроксимировать касательной

$$s(m) = s(m_0) + (m - m_0) \left. \frac{ds(m)}{dm} \right|_{m=m_0}; \quad (4.11)$$

(ii) через этот шар не проходит никакой другой отрезок кривой сигнала, т. е. в шаре радиуса r_0 с центром в точке $s(m_0)$ находится только один связный сегмент

кривой, который к тому же можно аппроксимировать его касательной.

Для шара радиуса r_0 на рис. 4.1 эти предположения выполнены.

Найдем среднеквадратическую ошибку оценивания в случае, когда эти предположения выполнены. Нам удобно будет использовать обозначение

$$s'(m_0) = \left. \frac{ds(m)}{dm} \right|_{m=m_0}.$$

Тогда $s'(m_0)/\|s'(m_0)\|$ — вектор единичной длины в направлении касательной. Оценка МП \hat{m} соответствует точке, для которой расстояние от $s(m)$ до z кратчайшее. Из рис. 4.1 видно, что в предположении линейности эта точка $s(\hat{m})$ получается проектированием вектора n на касательную, так что

$$s'(m_0)(\hat{m} - m_0) \approx s(\hat{m}) - s(m_0) \approx \left(n, \frac{s'(m_0)}{\|s'(m_0)\|} \right) \frac{s'(m_0)}{\|s'(m_0)\|},$$

или, поскольку числовые коэффициенты слева и справа должны совпадать,

$$\hat{m} - m_0 \approx \frac{(n, s'(m_0))}{\|s'(m_0)\|^2}. \quad (4.12)$$

Это соотношение справедливо для любого значения, которое примет шум. Случайная ошибка оценки дается выражением

$$\hat{M} - m_0 \approx \frac{(\mathcal{N}, s'(m_0))}{\|s'(m_0)\|^2}.$$

При наших предположениях о процессе шума $\mathcal{N}(t)$ скалярное произведение $(\mathcal{N}, s'(m_0))$ — это гауссовская случайная величина с нулевым средним и дисперсией $\|s'(m_0)\|^2 N_0/2$. Таким образом ¹⁾.

$$\begin{aligned} \mathcal{E}_2(m_0) &= \mathbf{E} \{(\hat{M} - m_0)^2\} = \\ &= \frac{N_0/2}{\|s'(m_0)\|^2} = \frac{N_0/2}{\int_0^T \left[\left. \frac{\partial s(t, m)}{\partial m} \right|_{m=m_0} \right]^2 dt}. \end{aligned} \quad (4.13)$$

¹⁾ Точные условия, достаточные для справедливости соотношения (4.13) в некотором асимптотическом смысле, приведены в работе Ибрагимова И. А. и Хасьминского Р. З., Оценка параметра сигнала в гауссовском белом шуме, *Проблемы передачи информации*, 10, 1 (1974). — *Прим. перев.*

Задача 4.1. Найти границу Крамера — Рао для $\mathcal{E}_2(m_0)$.

Задача 4.2. Время-импульсная модуляция есть система передачи, описываемая равенством

$$s(t, m) = p(t - mT_0),$$

в котором $p(t)$ — импульс продолжительности T_1 , а T_0 и T_1 выбраны так, что функция $p(t - mT_0)$ отлична от нуля лишь на отрезке $0 \leq t \leq T$ для всех $m \in \mathcal{M}$. Выразить $\mathcal{E}_2(m_0)$ с помощью среднеквадратической ширины полосы функции $p(t)$, определенной формулой

$$\bar{f}_p^2 = \frac{\int_{-\infty}^{+\infty} f^2 |P(f)|^2 df}{\int_{-\infty}^{+\infty} |P(f)|^2 df},$$

где

$$P(f) = \mathcal{F}[p(t)] = \int_{-\infty}^{+\infty} e^{-i2\pi ft} p(t) dt.$$

4.2. ОПТИМАЛЬНОЕ ОТОБРАЖЕНИЕ НА КРИВУЮ СИГНАЛА

Формула (4.13) дает среднеквадратическую ошибку при условии, что передается величина m_0 . Найдем ошибку, усредненную также и по распределению случайной величины M . Прежде чем делать это, исследуем имеющиеся в нашем распоряжении возможности для оптимизации. ВИМ может в более общем виде мыслиться как семейство сигналов $p(t - \tau)$, где $p(t)$ — импульс продолжительности T_1 , а τ — параметр этого семейства, принимающий значения $0 \leq \tau \leq T_0 = T - T_1$. В задаче 4.2 предполагалось, что соотношение между величиной сообщения m и параметром задержки τ линейно. Однако это предположение не обязательно. В самом деле, если M — случайная величина с неограниченной областью возможных значений, как, например, гауссовская, то линейность

даже невозможна. В этом случае нужно использовать нелинейное отображение

$$\tau = \tau(m), \quad -\infty < m < \infty, \quad 0 < \tau \leq T_0.$$

Физически это отображение можно реализовать с помощью усилителя с существенно нелинейной характеристикой. Такое устройство называется *компандером*.

В данной ситуации желательно знать, как выбрать это нелинейное отображение, чтобы минимизировать \mathcal{E}_2 . В случае ВИМ для данной функции $p(t)$ функция $p(t - \tau)$ при изменении τ пробегает кривую в $L_2 [0, T]$. Однако мы не хотим ограничивать анализ системой ВИМ. Поэтому надо ввести параметризацию (τ в случае ВИМ), применимую для произвольной кривой. Удобным параметром будет просто длина кривой в L_2 , т. е. длина, которую измерил бы счетчик пройденного пути, укрепленный на велосипедном колесе, катящемся по кривой, или рулетка с гибкой лентой.

Обозначим через $2L$ полную длину кривой, а через l — длину (со знаком), измеренную от центра кривой, так что $-L \leq l \leq L$. Так как l есть по определению L_2 -длина кривой, то справедливо тождество

$$\left\| \frac{ds(l)}{dl} \right\| = 1, \quad -L \leq l \leq L, \quad (4.14)$$

в котором $s(l)$ означает точку на кривой как функцию от l . Тогда переменная сообщения m отображается на кривую посредством функционального соотношения

$$l = l(m), \quad s(m) = s[l(m)]. \quad (4.15)$$

Здесь мы некорректно употребляем обозначения, применяя символы $s(m)$ и $s(l)$ для двух различных функций с разными областями определения. О какой именно функции идет речь, ясно из аргумента.

Имеем

$$\left\| \frac{ds(m)}{dm} \right\| = \left\| \frac{ds(l)}{dl} \frac{dl}{dm} \right\| = \left\| \frac{ds(l)}{dl} \right\| \left| \frac{dl}{dm} \right| = \left| \frac{dl}{dm} \right|. \quad (4.16)$$

Подставляя (4.16) в (4.13), получаем

$$\mathcal{E}_2(m_0) = \frac{N_0/2}{\left| \frac{dl}{dm} \right|_{m=m_0}^2}.$$

Усредняя по распределению случайной величины M , находим

$$\mathcal{E}_2 = (N_0/2) \int_{-\infty}^{+\infty} \left| \frac{dl}{dm} \right|^{-2} f(m) dm. \quad (4.17)$$

Выберем $l(m)$ так, чтобы минимизировать правую часть равенства (4.17). При этом мы должны учесть ограничение, состоящее в том, что полная длина кривой равна $2L$:

$$\int_{-\infty}^{+\infty} \frac{dl}{dm} dm = 2L. \quad (4.18)$$

Кроме того, необходимо, чтобы между точками кривой и величинами m было взаимно однозначное соответствие. Поэтому мы потребуем, чтобы выполнялось неравенство

$$\frac{dl}{dm} \geq 0, \quad (4.19)$$

причем равенство возможно только для тех значений m , для которых $f(m) = 0$.

Найдем функцию dl/dm , минимизирующую \mathcal{E}_2 (формула (4.17)) при ограничении (4.18). Обозначим через $g_0(m)$ эту (пока не известную) оптимизирующую функцию, и пусть $\eta(m)$ — произвольная функция, удовлетворяющая условиям

$$\int_{-m_0}^{m_0} \eta(m) dm = 0, \quad (4.20)$$

$$\eta(m) = 0, \quad |m| > m_0. \quad (4.21)$$

Так как функция $g_0(m)$ удовлетворяет ограничению (4.18), то это же ограничение выполнено и для функции

$$\frac{dl}{dm} = g_0(m) + \varepsilon \eta(m) \quad (4.22)$$

при всех значениях ε . Для такого выбора dl/dm имеем

$$\mathcal{E}_2 = \mathcal{E}_2(\varepsilon, \eta) = \int_{-\infty}^{+\infty} f(m) [g_0(m) + \varepsilon \eta(m)]^{-2} dm. \quad (4.23)$$

Если $g_0(m)$ действительно является минимизирующей функцией при указанных выше ограничениях, то для любой функции $\eta(m)$, удовлетворяющей условиям (4.20) и (4.21), значение \mathcal{E}_2 должно быть минимально при $\varepsilon = 0$. Для этого необходимо, чтобы

$$\begin{aligned} \frac{\partial \mathcal{E}_2(\varepsilon, \eta)}{\partial \varepsilon} \Big|_{\varepsilon=0} &= -2 \int_{-\infty}^{+\infty} f(m) \eta(m) [g_0(m) + \varepsilon \eta(m)]^{-3} \Big|_{\varepsilon=0} dm = \\ &= -2 \int_{-\infty}^{+\infty} f(m) \eta(m) [g_0(m)]^{-3} dm = 0. \end{aligned} \quad (4.24)$$

Заметим, что

$$\frac{\partial^2 \mathcal{E}_2(\varepsilon, \eta)}{\partial \varepsilon^2} \Big|_{\varepsilon=0} = 6 \int_{-\infty}^{+\infty} f(m) \eta^2(m) [g_0(m)]^{-4} dm > 0,$$

так что выполнение равенства (4.24) — необходимое условие именно для минимума, а не для другой критической точки.

Рассмотрим теперь пространство $L_2[-m_0, m_0]$. Из уравнений (4.20) и (4.24) вытекает, что

$$(f g_0^{-3}, \eta) = 0 \quad (4.24')$$

для любой функции $\eta \in L_2[-m_0, m_0]$ с носителем $[-m_0, m_0]$, удовлетворяющей соотношению

$$(1, \eta) = 0,$$

где 1 означает функцию, тождественно равную 1 на отрезке $[-m_0, m_0]$. Соотношение (4.24') должно выполняться для любой функции η , ортогональной к 1, так что функция $f g_0^{-3}$ должна быть «параллельна» функции 1.

Так как это рассуждение справедливо при любом m_0 , то окончательно получаем, что g_0 удовлетворяет соотношению

$$f(m) g_0^{-3}(m) \equiv c',$$

или

$$g_0(m) = c \{ + [f(m)]^{1/3} \}. \quad (4.25)$$

В последней формуле знак $+$ означает, что берется положительный кубический корень. Очевидно, это решение удовлетворяет ограничению (4.19). Чтобы удовлетворить ограничению (4.18), следует положить

$$c = \frac{2L}{\int_{-\infty}^{+\infty} f^{1/3}(m) dm}. \quad (4.26)$$

Обозначим

$$k = k[f(m)] = \left[\int_{-\infty}^{+\infty} f^{1/3}(m) dm \right]^3. \quad (4.27)$$

Подставляя (4.25) — (4.27) в выражение для \mathcal{E}_2 , получаем, что среднеквадратическая ошибка вычисляется по формуле

$$\mathcal{E}_2 = \frac{(N_0/2) k}{(2L)^2}. \quad (4.28)$$

Заметим, что эта минимальная величина для \mathcal{E}_2 представляет собой произведение трех сомножителей, первый из которых зависит лишь от уровня шумов, второй — только от длины кривой сигнала, третий — только от распределения сообщения.

Задача 4.3. Показать, что в случае, когда сообщение есть равномерно распределенная случайная величина с нулевым средним и дисперсией σ^2 ,

$$\mathcal{E}_2 = \frac{(N_0/2) 12\sigma^2}{(2L)^2}, \quad (4.29)$$

а для гауссовского сообщения с нулевым средним и дисперсией σ^2 справедлива формула ¹⁾

$$\mathcal{E}_2 = \frac{(N_0/2) 2\pi^{3/2}\sigma^2}{(2L)^2}. \quad (4.30)$$

¹⁾ Конечно, формулы (4.29) и (4.30) справедливы при любом математическом ожидании величины M . — *Прим. перев.*

4.3. ГРАНИЦЫ ЭФФЕКТИВНОСТИ. ПОРОГОВЫЙ ЭФФЕКТ

Из формулы (4.28) вытекает, что значение \mathcal{E}_2 можно уменьшить, если увеличить длину кривой. Интуитивно ясно, однако, что эту ошибку нельзя сделать произвольно малой за счет безграничного увеличения L . Выясним факторы, ограничивающие длину кривой. В соответствии с обсуждавшимся во введении к настоящей главе ограничением полосы частот предположим, что $s(m)$ принадлежит пространству, порожденному N ортонормальными функциями φ_k :

$$s(t, m) = \sum_{k=1}^N s_k(m) \varphi_k(t), \quad 0 \leq t \leq T. \quad (4.4)$$

Отметим, что для гауссовского случайного процесса с нулевым средним и корреляционной функцией

$$R_n(\tau) \approx \frac{N_0}{2} \delta(\tau) \quad (4.7)$$

проекция вектора $Z(t)$ на подпространство, порожденное функциями $\varphi_1, \dots, \varphi_N$ (т.е. N коэффициентами Фурье (Z, φ_k) , $k = 1, \dots, N$), есть достаточная статистика для оценки M .

Задача 4.4. Доказать это утверждение.

Обмен полосы частот и пороговый эффект

Учитывая сформулированное только что утверждение, будем рассматривать не векторы $s(m)$, Z и \mathcal{N} в $L_2[0, T]$, а лишь их проекции на N -мерное векторное пространство, порожденное функциями φ_k , $k = 1, \dots, N$. Тогда наши частотное и энергетическое ограничения приводят к тому, что кривая $s(m)$ лежит в N -мерном шаре радиуса \sqrt{E} .

Сделаем теперь следующие замечания. Во-первых, если потребовать, чтобы система модуляции была линейной, т.е. $s[l(m)] = l(m)\varphi$ для произвольной, но фиксированной функции φ в N -мерном подпространстве, то

(i) подпространство можно будет редуцировать к одномерному подпространству, порожденному функцией φ , а множитель расширения полосы частот будет равен 1;

(ii) кривая сигнала будет прямой линией с максимальной длиной $2L = 2\sqrt{E}$ и соответствующая минимальная среднеквадратическая ошибка будет равна

$$\mathcal{E}_2 = \frac{(N_0/2)k}{4E}.$$

Наше второе замечание состоит в том, что для увеличения длины кривой необходимо как можно больше «изгибать» ее в N -мерном шаре. Отсюда немедленно вытекают два следствия:

- (i) функция $s(l)$ должна быть *нелинейной*;
- (ii) кривая *должна* заполнять более чем одно измерение.

Таким образом, метод нелинейной модуляции и расширение полосы частот являются необходимыми предпосылками возрастания L и убывания среднеквадратической ошибки. Этот эффект называется *обменом ширины полосы на отношение сигнал/шум*. Использование этой фразеологии связано с тем, что при измерении эффективности чаще пользуются отношением сигнал/шум для сигнала на выходе

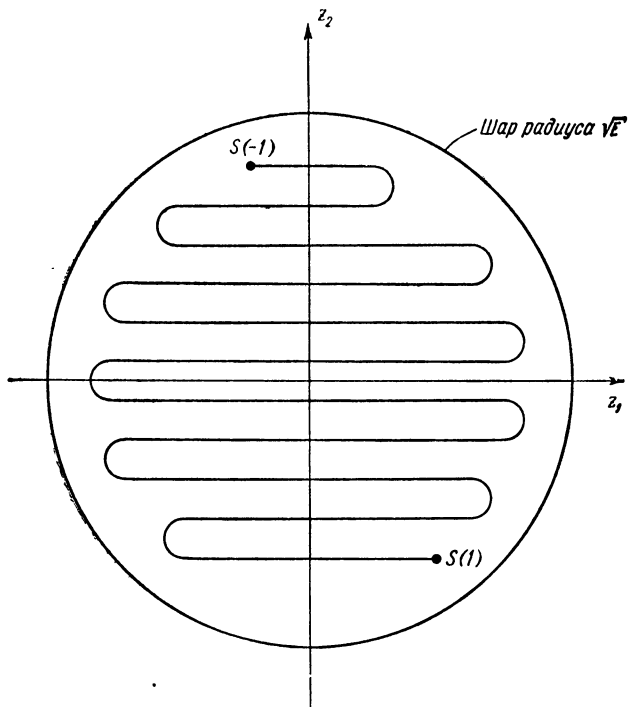
$$(S/\mathcal{N})_{\text{вых}} = \sigma_M^2 / \mathcal{E}_2, \quad (4.31)$$

чем самой величиной \mathcal{E}_2 .

Выясним теперь, насколько длинной можно сделать кривую сигнала и какой неблагоприятный эффект могло бы иметь слишком сильное ее изгибание. Заметим, что в N -мерный шар радиуса \sqrt{E} можно поместить сколь угодно длинную кривую просто путем многократного и очень частого изгибания этой кривой, как показано на рис. 4.2 (для двумерного сечения). Вспомним, однако, второе предположение (см. стр. 77), касающееся отсутствия других участков кривой в шаре радиуса r_0 . Это предположение существенно использовалось при выводе формулы для \mathcal{E}_2 , и оно совершенно необходимо для справедливости этой формулы.

Посмотрим (рис. 4.3), что произойдет, если различные ветви кривой сигнала будут слишком близки друг к другу. Предположим, что передается сигнал $s(m_0)$. Если уровень шума N_0 мал, то типичное значение приня-

того вектора есть z' , а соответствующий сигнал на выходе приемника m'_1 . При большем значении N_0 типичным значением принятого вектора будет z'' . В этом случае метод максимального правдоподобия дает на выходе



Р и с. 4.2. Сильно изогнутая кривая ограниченной энергии, но большой длины.

приемника сигнал m''_2 , в то время как наш линейный анализ привел бы к m'_2 . Фактическая ошибка была бы гораздо больше той, которая указывается соотношением (4.12), и поэтому формула (4.13) уже неверна. Когда N_0 становится настолько большим, что эти *аномальные ошибки* (большие ошибки, возникающие вследствие указания неправильной ветви кривой сигнала) вносят существенный вклад в среднеквадратическую ошибку, гово-

рят, что приемник оперирует в *пороговой области*. Это уменьшение эффективности приводит к кривой, изображенной на рис. 4.4 и отражающей зависимость от $1/N_0$ отношения сигнал/шум на выходе. Этот пороговый эффект неизбежно возникает в силу нелинейности кривой $s(l)$, и степень его проявления зависит (если N_0 становится достаточно большим) от того, насколько сильно изогнута кривая сигнала.

Граница плотной упаковки для эффективности

Если второе предположение (стр. 77), касающееся других ветвей кривой, выполнено, то тем самым ограничивается возможная длина кривой. Найдем количественную границу для этой длины. Наше изложение опирается на книгу Возенкрафта и Джекобса [3, § 8.4].

Задача 4.5. Рассмотрим проекцию шума на N -мерное подпространство, порожденное функциями $\varphi_1, \dots, \varphi_N$:

$$\mathcal{N} = (\mathcal{N}_1, \dots, \mathcal{N}_N), \quad \mathcal{N}_k = \int_0^T \mathcal{N}(t) \varphi_k(t) dt.$$

Найти совместное распределение величин \mathcal{N}_k и показать, что для случайной величины

$$\|\mathcal{N}\|^2 = \sum_{k=1}^N (\mathcal{N}_k)^2$$

справедливы соотношения

$$\begin{aligned} \mathbf{E} \{ \|\mathcal{N}\|^2 \} &= N(N_0/2), \\ \text{var} \{ \|\mathcal{N}\|^2 \} &= 2N(N_0/2)^2. \end{aligned} \quad (4.32)$$

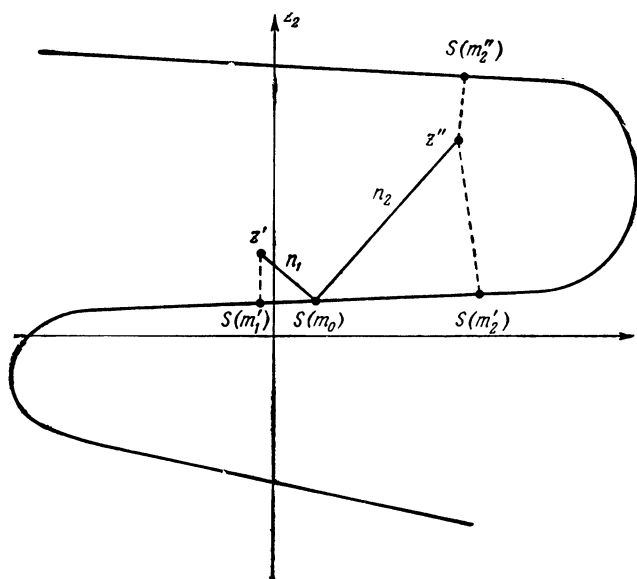
Из этих соотношений и из неравенства Чебышева получаем

$$\mathbf{P} \left\{ \left| \|\mathcal{N}\|^2 - \frac{NN_0}{2} \right| > \varepsilon' N \right\} \leq \frac{2(N_0/2)^2}{(\varepsilon')^2 N}. \quad (4.33)$$

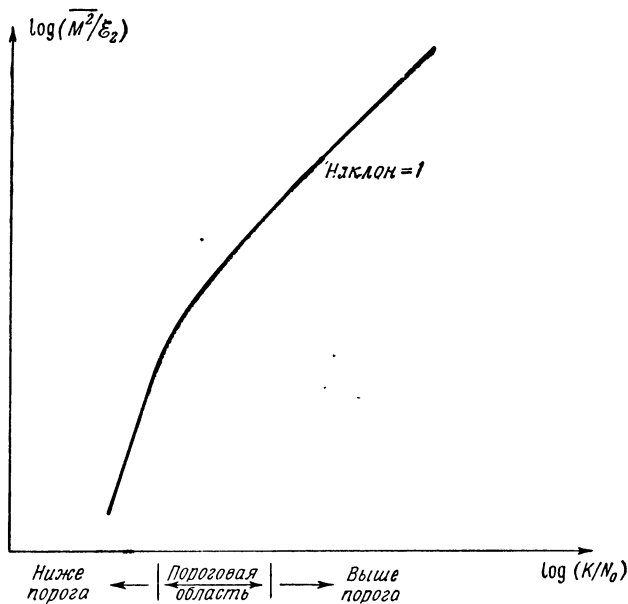
Таким образом, вероятность того, что для больших N случайная величина $\|\mathcal{N}\|$ лежит в интервале

$$\sqrt{N(N_0/2)}(1 - \varepsilon') \leq \|\mathcal{N}\| \leq \sqrt{N(N_0/2)}(1 + \varepsilon'), \quad (4.34)$$

стремится с ростом N к 1.



Р и с. 4.3. Возникновение малых и больших аномальных ошибок.



Р и с. 4.4. Пороговое поведение нелинейной системы импульсной модуляции.

Выясним, насколько должны отстоять для больших N разные ветви кривой при данном уровне шумов N_0 . Рассмотрим две соседние ветви кривой (рис. 4.5) и разложим вектор шума \mathcal{N} на перпендикулярную к кривой сигнала компоненту \mathcal{N}_\perp и параллельную ей компоненту \mathcal{N}_\parallel .

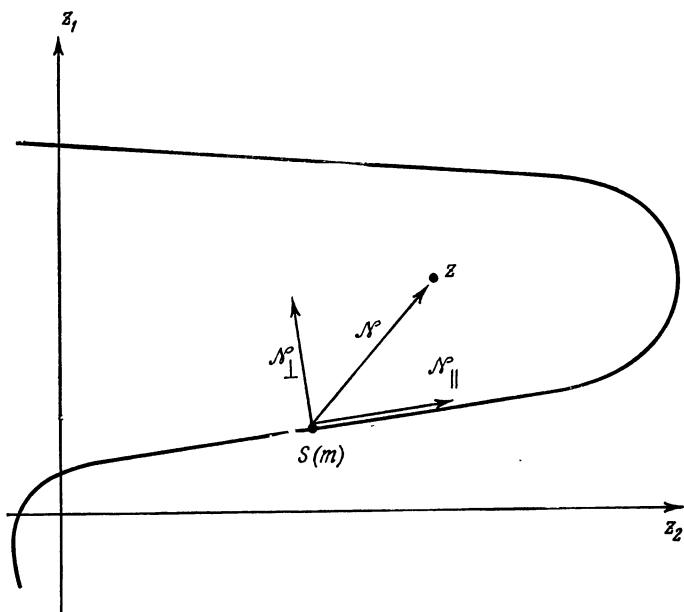


Рис. 4.5. Разложение вектора шума на две компоненты.

Вектор \mathcal{N}_\perp лежит в $(N-1)$ -мерном подпространстве. Следовательно, рассуждения использованные при выводе неравенств (4.33) и (4.34), применимы к \mathcal{N}_\perp с заменой N на $N-1$.

Из неравенства (4.34) видно, что для больших N аномальные ошибки встречаются с большой вероятностью, если расстояние между различными ветвями кривой меньше $2\sqrt{(N-1)N_0/2}$. Если же различные ветви удалены друг от друга на расстояние, немного превышающее $2\sqrt{(N-1)N_0/2}$, то вероятность аномальной ошибки мала. Это последнее условие выполнено,

если кривую можно заключить в «трубку» радиуса $\sqrt{(N-1)N_0/2}$ без самопересечений. Итак, мы получаем оценку

$$(2L) [\text{площадь сечения } ((N-1)\text{-мерный объем) диска радиуса } \leq \sqrt{(N-1)N_0/2}] \leq \leq N\text{-мерный объем, занятый принятыми сигналами.} \quad (4.35)$$

Это соотношение является неравенством, поскольку, вообще говоря, нельзя упаковать плотно (без «воздушных» зазоров) эту трубку в N -мерном объеме. Площадь сечения $(N-1)$ -мерного диска есть объем шара в $(N-1)$ -мерном пространстве радиуса $\sqrt{(N-1)N_0/2}$. Как хорошо известно [6, т. III, стр. 392 — 393], этот объем равен

$$V_{n-1}(\sqrt{(N-1)N_0/2}) = \frac{\pi^{(N-1)/2}}{\Gamma((N-1)/2 + 1)} [(N-1)N_0/2]^{(N-1)/2}. \quad (4.36)$$

Оценим объем, занятый принятыми сигналами. Так как $Z = s(m) + \mathcal{N}$, то $\|Z\|^2 = \|s(m)\|^2 + 2(\mathcal{N}, s(m)) + \|\mathcal{N}\|^2$. С помощью тех же соображений, которые привели к неравенству (4.34), можно показать, что с большой вероятностью

$$\|s(m)\|^2 + N(N_0/2)(1 - \varepsilon') \leq \|Z\|^2 \leq \leq \|s(m)\|^2 + N(N_0/2)(1 + \varepsilon'). \quad (4.37)$$

Так как $\|s(m)\|^2 \leq E$, то объем, занятый принятыми сигналами, не превосходит объема N -мерного шара радиуса $\sqrt{E + N(N_0/2)}$, равного

$$\frac{\pi^{N/2}}{\Gamma(N/2 + 1)} [E + N(N_0/2)]^{N/2}. \quad (4.38)$$

Из (4.35), (4.36) и (4.38) находим, что

$$2L \leq \frac{\pi^{1/2} \Gamma((N-1)/2 + 1) [NN_0/2]^{N/2}}{\Gamma(N/2 + 1) [(N-1)N_0/2]^{(N-1)/2}} \left[\frac{E + NN_0/2}{NN_0/2} \right]^{N/2}.$$

Используя формулу Стирлинга [6, т. II, стр. 792] для Г-функции при больших N и сокращая общие множители, получаем

$$2L \leq \sqrt{2\pi N_0/2} \left[1 + \frac{E}{N N_0/2} \right]^{N/2} \left(1 + \frac{1}{N-1} \right)^{(N-1)/2}.$$

Замечая, что последний множитель в правой части при больших N приближается к \sqrt{e} , окончательно имеем

$$2L \leq \sqrt{2\pi e (N_0/2)} \left[1 + \frac{E}{N (N_0/2)} \right]^{N/2}. \quad (4.39)$$

Подставляя (4.39) в (4.28), приходим к границе для достижимой среднеквадратической ошибки

$$\mathcal{E}_2 \geq \frac{k}{2\pi e} \left[1 + \frac{E}{N (N_0/2)} \right]^{-N}. \quad (4.40)$$

В терминах отношения сигнал/шум на выходе неравенство (4.40) можно переписать в виде

$$(S/\mathcal{N})_{\text{вых}} \leq \frac{\sigma^2}{K} 2\pi e \left[1 + \frac{E}{N (N_0/2)} \right]^N. \quad (4.41)$$

Эти границы нельзя понимать как точные границы эффективности достаточно «хорошей» системы импульсной модуляции по упомянутой ранее причине: трудно эффективно упаковать в данном объеме непрерывно дифференцируемую кривую (это даже более трудная задача, чем упаковка точек — задача, рассматриваемая в цифровом кодировании). На эти границы следует скорее смотреть как на указание того, как меняется эффективность для соответствующего семейства систем модуляции: отношение $(S/\mathcal{N})_{\text{вых}}$ должно меняться как N -я степень отношения сигнал/шум в канале, $(S/\mathcal{N})_{\text{кан}} = E/(N_0/2)$. Заметим, что этот вывод относится именно к *семейству* систем модуляции, в то время как для любой данной системы модуляции надпороговая эффективность растёт *линейно* с $E/(N_0/2)$.

Таким образом, при оперировании с широким диапазоном значений отношения сигнал/шум в канале эффективное использование канала требует использования

семейства систем модуляции, каждая из которых эффективна при данном значении отношения сигнал/шум. Такое семейство систем модуляции не обязано порождаться семейством различных физических схем. Его можно построить, изменяя один или более параметров в данной схеме; см., например, Мак-Аули и Сакрисон [7].

Очень трудно сравнить эффективность данной системы модуляции (например, ВИМ) с границей (4.40), так как эффективность данной системы трудно вычислить. Выше порога эффективность можно вычислить с помощью формулы (4.13) и последующего усреднения по m_0 :

$$\mathcal{E}_2 = \int_{-\infty}^{+\infty} \mathcal{E}_2(m_0) f(m_0) dm_0. \quad (4.42)$$

Однако эта формула верна *только* выше порога. Следовательно, без дополнительного анализа того, при каких значениях отношения $E/(N_0/2)$ справедлива формула (4.42), ее нельзя использовать для сравнения с границей.

Среднеквадратическую ошибку как *функцию* отношения $E/(N_0/2)$ (и линейное, и нелинейное поведение) можно иногда найти из границ. Возенкрафт и Джекобс [3, § 8.2] дали выражение для \mathcal{E}_2 как функции от $E/(N_0/2)$, но оно применимо только к сигналам с функцией корреляции вида

$$(s(m), s(m')) = c(m - m'),$$

где $c(m)$ имеет периодические нули и низкие боковые ветви.

Анализ ВИМ, использующий границу Баранкина, проведен Мак-Аули и Зейдманом [8]; граница Зива и Закаи значительно точнее и ее тоже можно применить к этой задаче. Единый подход к описанию этих трех границ в применении к ВИМ излагается в статье Зейдмана [9].

В области импульсной модуляции предстоит еще многое сделать. Главная из тем, требующих внимания, — нахождение сравнительно просто вычисляемых или же приближенных границ для нахождения эффективности данной системы модуляции как функции от $E/(N_0/2)$. Кроме того, большую область исследований составляет нахож-

дение новых систем модуляции, которые можно практически реализовать и эффективность которых ближе к границе (4.40), чем в существующих системах.

СПИСОК ЛИТЕРАТУРЫ

1. Сакрисон (Sakrison D. J.), *Communication theory: Transmission of waveforms and digital information*, New York, 1968.
2. Котельников В. А., Теория потенциальной помехоустойчивости, Изд-во ГЭИ, М., 1956.
3. Возенкрафт Дж., Джекобс И., Теоретические основы техники связи, изд-во «Мир», М., 1969.
4. Слепьян, Поллак, Ландау (Slepian D., Pollack H. O., Landau H. J.), Prolate spheroidal wave functions, Fourier analysis and uncertainty, Parts I and II, *Bell Syst. Tech. J.*, **40** (January 1961), 43—84.
5. Ландау, Поллак (Landau H. J., Pollack H. O.), Prolate spheroidal wave functions, Fourier analysis and uncertainty, Part III — The dimension of the space of essentially time and band limited signals, *Bell Syst. Tech. J.*, **41** (July 1962), 1295—1336.
6. Фихтенгольц Г. М., Курс дифференциального и интегрального исчисления, изд-во «Наука», М., 1966.
7. Мак-Аули, Сакрисон (McAuley R. J., Sakrison D. J.), A PPM-PM hybrid modulation system, *IEEE Trans. Commun. Tech.*, **COM-17** (August 1969), 458—468.
8. Мак-Аули, Зейдман (McAuley R. J., Seidman L. P.), A usefull form of the Barankin lower bound and its application to PPM threshold analysis, *IEEE Trans. Inform. Theory*, **IT-15** (March 1969), 273—278.
9. Зейдман (Seidman L. P.), Performance limitations and error calculations for parameter estimation, *IEEE Proc.*, **58**, 5 (1970). (Русский перевод: *ТННЭР*, **58**, 5 (1970).)

Глава 5

ОПТИМАЛЬНАЯ УГЛОВАЯ ДЕМОДУЛЯЦИЯ

В этой главе мы рассмотрим задачу оптимальной демодуляции сигналов систем модуляции, т. е. систем, в которых модулятор *непрерывно* во времени обрабатывает сигнал сообщения и непрерывно производит меняющийся модулированный сигнал $V(t)$. Мы ограничимся модуляторами вида, изображенного на рис. 5.1.

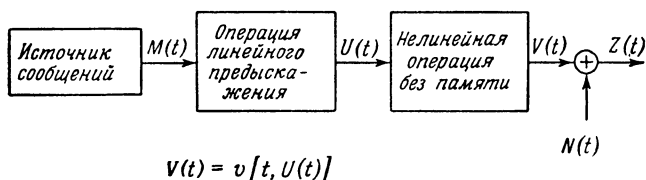


Рис. 5.1. Рассматриваемая система модуляции.

учитывая их практическую важность. В этих системах операция модуляции представляет собой каскад, состоящий из линейной операции и следующей за ней нелинейной операции без памяти. В конце нашего обсуждения мы сосредоточим внимание на одном классе систем модуляции, представляющем наибольший практический интерес, а именно на угловой модуляции. Изучение оптимального демодулятора даст повод кратко описать очень полезный субоптимальный демодулятор — *фазовую автоподстройку частоты*.

5.1. ДЕМОДУЛЯТОР ПО МАКСИМУМУ АПОСТЕРИОРНОЙ ВЕРОЯТНОСТИ (МАВ)

Определим оптимальный демодулятор для системы модуляции, показанной на рис. 5.1. Хотя обычно приходится рассматривать узкополосные сигналы, мы будем сразу описывать сигналы $V(t)$, $\mathcal{N}(t)$ и $Z(t)$ как действи-

тельнозначные узкополосные сигналы, а не с помощью комплексного низкочастотного представления. Мы будем считать шум стационарным гауссовским процессом с нулевым средним и белым спектром:

$$R_n(\tau) = \mathbf{E} \{ \mathcal{N}_{(t+\tau)} \mathcal{N}_{(t)} \} = \frac{N_0}{2} \delta(\tau). \quad (5.1)$$

Чтобы облегчить дальнейшее исследование, начнем с проблемы оценки процесса $U(t)$ по наблюдению $Z(t)$. В разд. 5.2, где будет обсуждаться демодулятор фазовой автоподстройки частоты, мы коротко исследуем вопрос о том, как выбор демодулятора должен учитывать эффект предискажения в сети связи.

Для некоторых важных нелинейных операций $v(\cdot, \cdot)$ вычисление условного математического ожидания процесса $U(t)$ по наблюдениям $Z(t)$ представляет собой совершенно безнадежную задачу. По этой причине мы будем рассматривать здесь оценку МАВ.

Предположим, что сообщение является стационарным гауссовским процессом с нулевым средним и известной корреляционной функцией $R_m(\tau)$. Тогда и $U(t)$ — тоже стационарный гауссовский процесс с нулевым средним.

Последующее изложение опирается на книгу Витерби [1, гл. 5]. Сначала сведем процесс $U(t)$ и наблюдаемый процесс $Z(t)$ к процессам с дискретным временем, считывая каждые $\Delta\tau$ с значения этих процессов в моменты времени

$$\tau_k = t_0 + k \Delta\tau, \quad k = 1, 2, \dots, K, \quad \tau_k = t. \quad (5.2)$$

Введем обозначения:

$$\begin{aligned} U_k &= U(\tau_k), & u_k &= u(\tau_k), \\ \mathcal{N}_k &= \mathcal{N}(\tau_k), & n_k &= n(\tau_k), \\ Z_k &= Z(\tau_k), & z_k &= z(\tau_k), \\ V_k &= v(\tau_k, U_k), & v_k &= v(\tau_k, u_k) \end{aligned} \quad (5.3)$$

и

$$\begin{aligned} \mathbf{u} &= (u_1, \dots, u_K), & \mathbf{n} &= (n_1, \dots, n_K), \\ \mathbf{v} &= (v_1, \dots, v_K), & \mathbf{z} &= (z_1, \dots, z_K). \end{aligned} \quad (5.4)$$

При наблюдении $\mathbf{Z} = \mathbf{z}$ задача состоит в нахождении значения \mathbf{u} , максимизирующего функцию

$$f(\mathbf{u} | \mathbf{z}) = \frac{f(\mathbf{z} | \mathbf{u}) f(\mathbf{u})}{f(\mathbf{z})},$$

или, что эквивалентно, максимизирующего ее натуральный логарифм.

Необходимое условие максимума — выполнение равенств

$$\frac{\partial \ln f(\mathbf{u})}{\partial u_k} + \frac{\partial \ln f(\mathbf{z} | \mathbf{u})}{\partial u_k} = 0, \quad k = 1, \dots, K. \quad (5.5)$$

При сделанных предположениях относительно гауссовости процесса $\mathcal{N}(t)$ имеем

$$\begin{aligned} \ln f(\mathbf{z} | \mathbf{u}) = \\ = -\frac{K}{2} \ln 2\pi - \frac{1}{2} \ln |R_n| - \frac{1}{2} (\mathbf{z} - \mathbf{v}, R_n^{-1} (\mathbf{z} - \mathbf{v})), \end{aligned} \quad (5.6)$$

где R_n — ковариационная матрица процесса $\mathcal{N}(t)$. Так как лишь v_k зависит от u_k , то

$$\frac{\partial}{\partial u_k} \left[-\frac{1}{2} (\mathbf{z} - \mathbf{v}, R_n^{-1} (\mathbf{z} - \mathbf{v})) \right] = \frac{\partial v_k}{\partial u_k} [R_n^{-1} (\mathbf{z} - \mathbf{v})]_k. \quad (5.7)$$

Зададим вектор \mathbf{g} равенством

$$\Delta \tau \mathbf{g} = R_n^{-1} (\mathbf{z} - \mathbf{v}). \quad (5.8)$$

Другими словами, $\mathbf{g} = \mathbf{g}(\mathbf{u})$ определяется из уравнения

$$\mathbf{z} - \mathbf{v}(\mathbf{u}) = R_n \mathbf{g}(\mathbf{u}) \Delta \tau. \quad (5.9)$$

Заметим, что хотя $v_k = v_k(u_k)$ есть функция только переменной u_k , но g_k зависит от всех компонент вектора \mathbf{u} .

Из (5.6) — (5.8) находим, что

$$\frac{\partial}{\partial u_k} \ln f(\mathbf{z} | \mathbf{u}) = \frac{\partial v_k}{\partial u_k} g_k(\mathbf{u}) \Delta \tau. \quad (5.10)$$

Подобным образом для плотности $f(\mathbf{u})$ получаем

$$\ln f(\mathbf{u}) = -\frac{1}{2} K \ln 2\pi - \frac{1}{2} \ln |R_u| - \frac{1}{2} (\mathbf{u}, R_u^{-1} \mathbf{u})$$

и

$$\frac{\partial}{\partial u_k} \ln f(\mathbf{u}) = -[R_u^{-1} \mathbf{u}]_k, \quad (5.11)$$

где R_u означает ковариационную матрицу процесса \mathbf{U} . Соотношения (5.5), (5.10) и (5.11) приводят к равенствам

$$[R_u^{-1} \mathbf{u}]_k = \frac{\partial v_k}{\partial u_k} g_k(\mathbf{u}) \Delta \tau, \quad k = 1, 2, \dots, K, \quad (5.12)$$

определяющим оценку МАВ. Положим

$$x_k = \frac{\partial v_k}{\partial u_k} g_k(\mathbf{u}) \Delta \tau. \quad (5.13)$$

Умножая обе части равенства (5.12) на R_u , получаем систему уравнений, которую в векторной форме можно записать в виде

$$\mathbf{u} = R_u \mathbf{x}. \quad (5.14)$$

Так как матрицы R_n и R_u имеют элементы

$$[R_n]_{ij} = R_n(\tau_i, \tau_j), \quad [R_u]_{ij} = R_u(\tau_i, \tau_j),$$

то, расписывая (5.9) и (5.14) по компонентам и учитывая (5.13), находим системы уравнений для оценки МАВ \hat{u}_k :

$$z(\tau_k) - v(\tau_k, \hat{u}_k) = \sum_{n=1}^K R_n(\tau_k, \tau_n) g(\tau_n, \hat{\mathbf{u}}) \Delta \tau, \quad (5.15)$$

$$\hat{u}(\tau_k) = \sum_{n=1}^K R_u(\tau_k, \tau_n) \frac{\partial v(\tau_n, \hat{u}_n)}{\partial u_n} g(\tau_n, \hat{\mathbf{u}}) \Delta \tau, \quad (5.16)$$

$$k = 1, \dots, K.$$

Эти две системы уравнений определяют оценку МАВ в случае дискретного времени. Однако на самом деле нас интересует случай непрерывного времени. Переходя к

пределу в (5.15) и (5.16) при $\Delta\tau \rightarrow 0$, формально приходим к системе двух интегральных уравнений

$$z(\tau) - v(\tau, \hat{u}(\tau)) = \int_{t_0}^{t_1} ds R_n(\tau, s) g[s; \hat{u}(\alpha), t_0 \leq \alpha \leq t_1], \quad (5.17)$$

$$\hat{u}(\tau) = \int_{t_0}^{t_1} ds R_u(\tau, s) \frac{\partial v[s, \hat{u}(s)]}{\partial \hat{u}(s)} g[s; \hat{u}(\alpha), t_0 \leq \alpha \leq t_1], \quad (5.18)$$

$$t_0 \leq \tau \leq t_1.$$

Уравнения (5.17) и (5.18) справедливы для произвольных гауссовских процессов $U(t)$ и $\mathcal{N}(t)$ с нулевым средним.

Сделаем теперь следующие дополнительные предположения, имеющие наибольший практический интерес:

- (i) процессы $U(t)$ и $\mathcal{N}(t)$ стационарны;
- (ii) процесс $\mathcal{N}(t)$ есть белый шум, т. е. $R_n(\tau) = (N_0/2) \delta(\tau)$;
- (iii) режим работы стационарен, $t_0 = -\infty$.

При выполнении этих предположений уравнение (5.17) принимает вид

$$z(\tau) - v(\tau, \hat{u}(\tau)) = \frac{N_0}{2} g[\tau; \hat{u}(\alpha), -\infty \leq \alpha \leq t_1], \quad -\infty \leq \tau \leq t_1. \quad (5.19)$$

Подставляя (5.19) в (5.18), видим, что оценку МАВ можно найти как решение одного интегрального уравнения¹⁾

$$\hat{u}(\tau) = \frac{2}{N_0} \int_{-\infty}^{t_1} ds R_u(\tau - s) \frac{\partial v[s, \hat{u}(s)]}{\partial \hat{u}(s)} \{z(s) - v(s, \hat{u}(s))\}. \quad (5.20)$$

Применим эти результаты к фазовой модуляции (ФМ). Заметим, что предсказывающий фильтр, предше-

¹⁾ Уравнению (5.20) в случае «чистого» гауссовского белого шума $\mathcal{N}(t)$ нелегко придать четкий математический смысл, так как в этом случае (5.20) фактически содержит стохастический интеграл с подынтегральной функцией, зависящей от будущего течения процесса. Этой трудности нет для уравнения (5.25), описывающего систему фазовой автоподстройки частоты. — Прим. ред.

ствующий нелинейному модулятору без памяти, позволит нам позже изучить частотную модуляцию (ЧМ), а также ФМ или ЧМ с предсказанием. Для фазовой модуляции

$$v(\tau, u(\tau)) = \sqrt{2P} \sin[\omega_0 \tau + u(\tau)] \quad (5.21)$$

и

$$\frac{\partial v}{\partial u} = \sqrt{2P} \cos[\omega_0 \tau + h(\tau)]. \quad (5.22)$$

Подставляя (5.21) и (5.22) в (5.20) и используя тригонометрическое тождество $2 \sin A \cos A = \sin 2A$, получаем

$$\begin{aligned} \hat{u}(\tau) &= \frac{2}{N_0} \int_{-\infty}^{\tau} ds R_u(\tau - s) \sqrt{2P} \cos(\omega_0 s + \hat{u}(s)) \times \\ &\quad \times \{z(s) - \sqrt{2P} \sin(\omega_0 s + \hat{u}(s))\} = \\ &= -\frac{\sqrt{2P}}{N_0} \int_{-\infty}^{\tau} ds R_u(\tau - s) \sin(2\omega_0 s + 2\hat{u}(s)) + \\ &\quad + \frac{2\sqrt{2P}}{N_0} \int_{-\infty}^{\tau} ds R_u(\tau - s) \cos(\omega_0 s + \hat{u}(s)) z(s) \approx \\ &\approx \frac{2\sqrt{2P}}{N_0} \int_{-\infty}^{\tau} ds R_u(\tau - s) \cos[\omega_0 s + \hat{u}(s)] z(s), \quad (5.23) \\ &\quad -\infty < \tau \leq t. \end{aligned}$$

Предположение, при котором справедливо последнее приближенное равенство, состоит в том, что $U(t)$ — низкочастотный процесс. Поэтому членом, содержащим двойную частоту, на выходе фильтра с импульсной переходной функцией $R_u(\cdot)$ можно пренебречь. Уравнение (5.23) определяет демодулятор МАВ для ФМ.

При попытке решить это уравнение, т.е. найти физическую реализацию оценки МАВ, обычно появляются некоторые трудности. Мы сможем найти лишь приближенное решение этого уравнения, которое приемлемо при некоторых дополнительных условиях. Для выяснения

физической интерпретации уравнения (5.23) полезно рассмотреть частный случай $\tau = t$. В этом случае

$$\hat{u}(t) = \frac{2\sqrt{2P}}{N_0} \int_{-\infty}^t ds R_u(t-s) \cos(\omega_0 s + \hat{u}(s)) z(s). \quad (5.24)$$

Заметим, что функция $\hat{u}(s)$ в подинтегральном выражении — это оценка МАВ для $U(s)$, основанная на наблюдении $Z(\tau)$ на *всем* интервале $-\infty \leq \tau \leq t$.

5.2. ФАЗОВАЯ АВТОПОДСТРОЙКА ЧАСТОТЫ — СУБОПТИМАЛЬНЫЙ ДЕМОДУЛЯТОР

Рассмотрим поведение цепи, изображенной на блок-схеме рис. 5.2. Сигнал на выходе этой цепи есть решение интегрального уравнения

$$u'(t) = \sqrt{2} \int_{-\infty}^t ds f(t-s) \cos(\omega_0 s + u'(s)) z(s). \quad (5.25)$$

Если положить

$$\hat{f}(t-s) = \frac{2\sqrt{2P}}{N_0} R_u(t-s),$$

то это интегральное уравнение примет вид

$$u'(t) = \frac{2\sqrt{2P}}{N_0} \int_{-\infty}^t ds R_u(t-s) \cos(\omega_0 s + u'(s)) z(s). \quad (5.26)$$

По внешнему виду уравнение (5.26) кажется точно таким же, как и уравнение (5.24), определяющее оценку МАВ. Однако имеется одна существенная разница. В (5.24) $\hat{u}(s)$ — это оценка МАВ для $U(s)$, основанная на наблюдении $Z(\tau)$, $-\infty \leq \tau \leq t$, в то время как в (5.26) $u'(s)$ есть отклик физически реализуемой системы и, следовательно, зависит лишь от $Z(\tau)$ при $-\infty \leq \tau \leq s < t$. Система фазовой автоподстройки частоты (5.25) поэтому не реализует оценку МАВ.

Чтобы показать, как приближенно реализовать оценку МАВ, вернемся к (5.23) и выпишем выражение для $z(s)$:

$$z(s) = \sqrt{2P} \sin(\omega_0 s + u(s)) = n(s).$$

Положим

$$n'(s) = \sqrt{2} \cos(\omega_0 s + \hat{u}(s)) n(s) \quad (5.27)$$

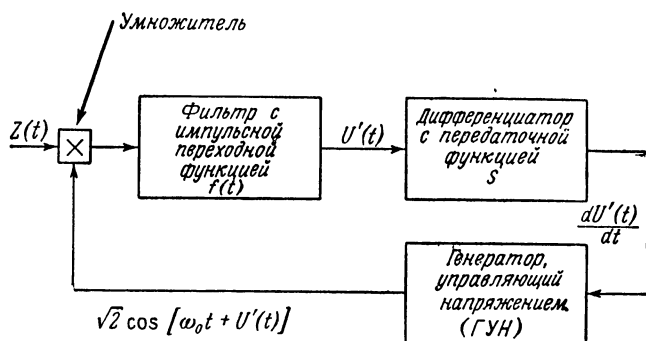


Рис. 5.2. Фазовая автоподстройка частоты.

и воспользуемся тригонометрическим тождеством

$$\cos A \sin B = \frac{1}{2} [\sin(A + B) + \sin(B - A)].$$

Тогда

$$\begin{aligned} \hat{u}(\tau) &= \frac{2\sqrt{2P}}{N_0} \int_{-\infty}^{\tau} ds R_u(\tau - s) \cos(\omega_0 s + \hat{u}(s)) \times \\ &\quad \times \{ \sqrt{2P} \sin(\omega_0 s + u(s)) + n(s) \} = \\ &= \frac{2\sqrt{2P}}{N_0} \int_{-\infty}^{\tau} ds R_u(\tau - s) \left\{ \frac{1}{\sqrt{2}} n'(s) + \right. \\ &\quad \left. + \sqrt{\frac{P}{2}} \sin(2\omega_0 s + \hat{u}(s) + u(s)) + \right. \\ &\quad \left. + \sqrt{\frac{P}{2}} \sin(u(s) - \hat{u}(s)) \right\} \approx \\ &\approx \frac{2P}{N_0} \int_{-\infty}^{\tau} ds R_u(\tau - s) \left\{ \sin(u(s) - \hat{u}(s)) + \frac{1}{\sqrt{P}} n'(s) \right\}, \quad (5.28) \\ &\quad -\infty < \tau \leq t, \end{aligned}$$

Приближенное равенство в (5.28) основано на том, что членом, содержащим двойную частоту, на выходе фильтра с переходной функцией $R_u(\cdot)$ можно пренебречь.

Предположим теперь, что ошибка оценки МАВ процесса $U(s)$ мала; в частности, с вероятностью, близкой к единице, величина $|U(s) - \hat{O}(s)|$ мала по сравнению с радианом, так что

$$\sin [u(s) - \hat{u}(s)] \approx u(s) - \hat{u}(s). \quad (5.29)$$

Учитывая это предположение и полагая

$$z'(s) = u(s) + \frac{n'(s)}{\sqrt{P}}, \quad (5.30)$$

приводим (5.28) к виду

$$u(\tau) \approx \frac{2P}{N_0} \int_{-\infty}^{\tau} ds R_u(\tau - s) [z'(s) - \hat{u}(s)]. \quad (5.31)$$

Прежде чем вычислять корреляционную функцию шумового процесса, определенного в (5.27), рассмотрим с иллюстративной целью сумму независимых одинаково распределенных случайных величин

$$\mathcal{N}_\Sigma = \frac{1}{M} \sum_{k=1}^M \mathcal{N}_k.$$

Легко видеть, что для больших M зависимость \mathcal{N}_Σ от любой одной величины \mathcal{N}_k слаба, а при $M \rightarrow \infty$ случайная величина \mathcal{N}_Σ становится статистически независимой от любой величины \mathcal{N}_k . Заметим, что $\hat{O}(s)$ зависит от $V(\tau)$ и $\mathcal{N}(\tau)$ при $-\infty < \tau \leq t$. Если можно приблизиться к $\hat{O}(s)$, производя действия над $V(\tau_k)$ и $\mathcal{N}(\tau_k)$ с достаточно частым дискретным множеством точек τ_k на временной оси, то, как и раньше, процесс $\hat{O}(s)$ зависит от большого числа независимых случайных величин $\mathcal{N}(\tau_k)$. Тогда с помощью приведенных выше эвристических соображений заключаем, что $\hat{O}(s)$ и $\mathcal{N}(\tau)$ независимы.

Так как процесс $\mathcal{N}(t)$ имеет нулевое среднее, то этим же свойством обладает и процесс $\mathcal{N}'(t)$. Чтобы изучить

спектральные свойства процесса $\mathcal{N}'(t)$, запишем шум $\mathcal{N}(t)$ в виде

$$\mathcal{N}(t) = \mathcal{N}_c(t) \cos \omega_0 t - \mathcal{N}_s(t) \sin \omega_0 t.$$

Заметим, что процесс $\mathcal{N}'(t)$ появляется лишь в свертке с R_u . Как и для других сигналов, мы интересуемся лишь низкочастотными членами в $\mathcal{N}'(t)$ и можем пренебречь членами с двойной частотой. Преобразуя $\mathcal{N}'(t) = \sqrt{2}\mathcal{N}(t) \cos(\omega_0 t + \hat{u}(t))$ с помощью тригонометрических тождеств и опуская члены с двойной частотой, получаем

$$\mathcal{N}'(t) = \frac{1}{\sqrt{2}} [\mathcal{N}_c(t) \cos \hat{u}(t) - \mathcal{N}_s(t) \sin \hat{u}(t)]. \quad (5.31')$$

Учитывая независимость $\mathcal{N}(t)$ и $\hat{u}(t)$, находим отсюда, что

$$\begin{aligned} R_{n'}(\tau) &= \mathbf{E} \{ \mathcal{N}'_{(t+\tau)} \mathcal{N}'(t) \} = \\ &= \frac{1}{2} [R_{n_c}(\tau) \mathbf{E} \{ \cos \hat{u}(t + \tau) \cos \hat{u}(t) \} + \\ &\quad + R_{n_s}(\tau) \mathbf{E} \{ \sin \hat{u}(t + \tau) \sin \hat{u}(t) \} - \\ &\quad - R_{n_c n_c}(\tau) \mathbf{E} \{ \cos \hat{u}(t + \tau) \sin \hat{u}(t) \} - \\ &\quad - R_{n_s n_c}(\tau) \mathbf{E} \{ \sin \hat{u}(t + \tau) \cos \hat{u}(t) \}]. \end{aligned}$$

Уточним теперь предположение о том, что $\mathcal{N}(t)$ — белый шум, а именно мы будем считать, что $S_n(f)$ имеет характер, указанный на рис. 5.3, где W_0 много больше ширины полосы сигнала $V(t)$. Тогда

$$\begin{aligned} R_{n'}(\tau) &\approx \frac{1}{2} N_0 \delta(\tau) [\mathbf{E} \{ \cos \hat{u}(t + \tau) \cos \hat{u}(t) \} + \\ &\quad + \mathbf{E} \{ \sin \hat{u}(t + \tau) \sin \hat{u}(t) \}] \approx \frac{1}{2} N_0 \delta(\tau). \end{aligned} \quad (5.32)$$

Отсюда находим выражение для корреляционной функции процесса $Z'(t)$:

$$R_{z'}(\tau) = R_u(\tau) + \frac{N_0}{2E} \delta(\tau). \quad (5.33)$$

Возвращаясь к (5.31) и подставляя выражение для $R_u(\tau)$ из (5.33), получаем

$$\hat{u}(\tau) = \frac{2P}{N_0} \left\{ -\frac{N_0}{2P} [z'(\tau) - \hat{u}(\tau)] + \int_{-\infty}^{\tau} ds R_{z'}(\tau - s) [z'(s) - \hat{u}(s)] \right\},$$

или

$$z'(\tau) = \frac{2P}{N_0} \int_{-\infty}^{\tau} ds R_{z'}(\tau - s) [z'(s) - \hat{u}(s)], \quad \tau \leq t. \quad (5.34)$$

Пусть теперь $h(\tau)$ — решение уравнения Винера — Хопфа (см. [2, § 11.2])

$$\int_0^{\infty} h(s) R_{z'}(\sigma - s) ds = R_u(\sigma), \quad \sigma \geq 0. \quad (5.35)$$

Тогда, свертывая обе части равенства (5.34) с $h(\cdot)$ и используя (5.35) и (5.31), имеем

$$\begin{aligned} \int_{-\infty}^t h(t - \tau) z'(\tau) d\tau &= \\ &= \frac{2P}{N_0} \int_{-\infty}^t h(t - \tau) d\tau \int_{-\infty}^{\tau} R_{z'}(\tau - s) [z'(s) - \hat{u}(s)] ds = \\ &= \frac{2P}{N_0} \int_{-\infty}^t [z'(s) - \hat{u}(s)] ds \int_{-\infty}^t h(t - \tau) R_{z'}(\tau - s) d\tau = \\ &= \frac{2P}{N_0} \int_{-\infty}^t [z'(s) - \hat{u}(s)] R_u(t - s) ds = \hat{u}(t). \end{aligned} \quad (5.36)$$

Итак, мы приходим к выводу: если ошибка $u(t)$ — $\hat{u}(t)$ мала по сравнению с одним радианом, то оценку $\hat{u}(t)$ можно получить на выходе линейного фильтра с импульсной переходной функцией, являющейся решением уравнения Винера — Хопфа (5.35).

Задача 5.1. Рассмотрим снова систему фазовой автоподстройки частоты (ФАЧ), показанную на рис. 5.2. Пусть значения $U(t) - U'(t)$ малы по сравнению с одним радианом, а частотная характеристика ФАЧ представляется в виде

$$\sqrt{P}F(s) = \frac{H(s)}{1 - H(s)}, \quad (5.37)$$

где $H(s)$ — частотная характеристика фильтра с импульсной переходной функцией $h(s)$ ($h(s)$ — решение уравне-

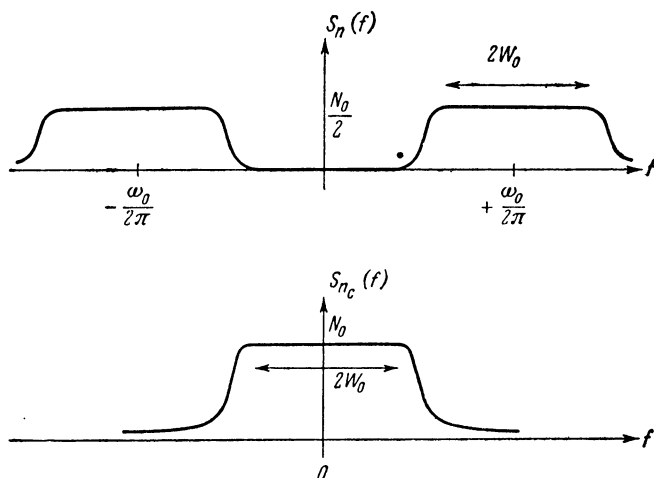


Рис. 5.3. Предполагаемый вид спектра шума.

ния (5.35)). Показать, что сигнал $u'(t)$ на выходе системы ФАЧ приближенно равен левой части равенства (5.36).

Результат предыдущей длинной выкладки можно сформулировать так: если ошибка наилучшей в среднеквадратическом оценки фазы мала по сравнению с одним радианом, то ФАЧ ведет себя как оптимальная (МАН) оценка. Полученный результат, однако, не может нас полностью удовлетворить, поскольку одна из главных причин широкого практического использования ФАЧ состоит в том, что ее эффективность при высоком

уровне шумов (пороговое поведение) превосходит более удобные приемники типа угловых демодуляторов, в то время как наш вывод применим лишь к изучению эффективности выше порога.

Остановимся теперь кратко на задаче исследования эффективности ФАЧ. На рис. 5.4, а изображена блок-схема всей системы фазовой модуляции. Из нашего анализа вытекает, что выше порога (среднеквадратическая фазовая ошибка мала по сравнению с одним радианом) всю систему можно заменить эквивалентной линейной системой, показанной на рис. 5.4, б. Для вычисления ошибки в оценке сообщения можно объединить два линейных фильтра $H(s)$ и $H_{pl}(s)$ и исследовать задачу фильтрации, показанную на рис. 5.4, в.

Для вычисления минимальной среднеквадратической ошибки фильтрации как функции от P/N_0 при оценке $M(t)$ в ситуации, отраженной на рис. 5.4, в, можно использовать винеровскую теорию оптимальной линейной фильтрации. Напомним, однако, что эти результаты справедливы лишь выше порога. Поэтому теорией оптимальной линейной фильтрации следует пользоваться для вычисления минимальной среднеквадратической ошибки фильтрации как функции от P/N_0 при оценке $U(t)$ в ситуации, отраженной на рис. 5.4, б. Вычисленная таким способом величина

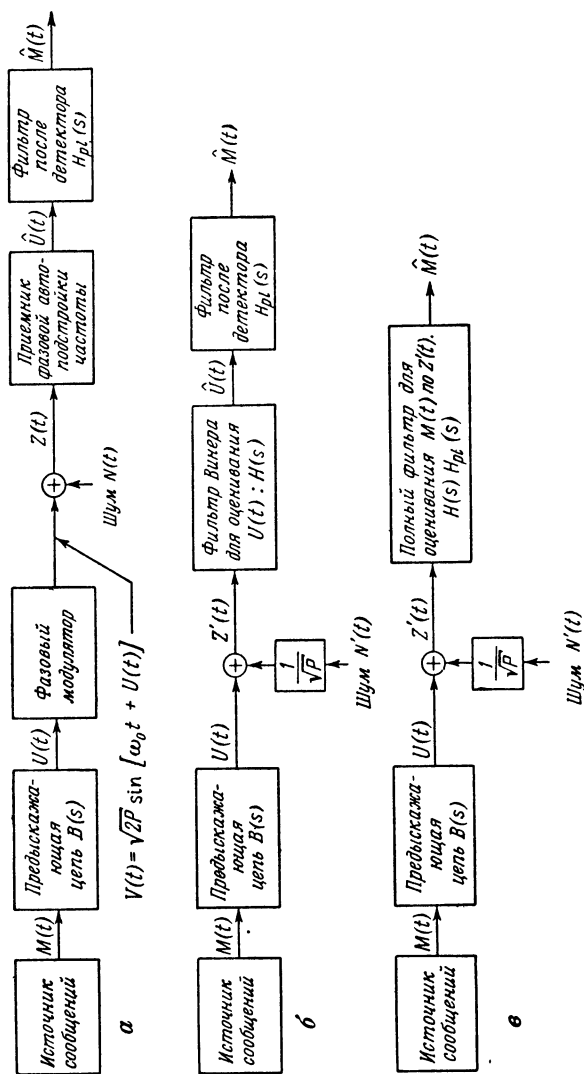
$$\mathcal{E}_{2, M} = \mathbf{E} \{ (M(t) - \hat{M}(t))^2 \}$$

совпадает с истинной лишь для тех значений P/N_0 , которые настолько велики, что

$$\mathcal{E}_{2, U} = \mathbf{E} \{ (U(t) - \hat{U}(t))^2 \} \leq \mathcal{E}_c,$$

где \mathcal{E}_c — число, заключенное между 0,16 и 0,33 (радиан)².

Этот подход позволяет вычислить не только надпороговую эффективность $(S/\mathcal{N})_{\text{вых}} = \sigma_M^2 / \mathcal{E}_{2, M}$ как функцию от P/N_0 , но также и значение P/N_0 , соответствующее порогу. Графики зависимости $(S/\mathcal{N})_{\text{вых}}$ от P/N_0 для частотной и фазовой модуляций, основанные на этом подходе,



Р и с. 5.4. Система угловой модуляции: а) вся система, б) эквивалентная линейная модель выше порога, в) упрощенный вид линейной модели.

приведены в книгах Витерби [1, гл. 6] и Сакрисона [3, гл. 10].

Бордман и Ван Трис [4] довели этот подход почти до конца, решив следующую оптимизационную задачу. При ограничениях

$$P/N_0 = \alpha_1,$$

$$\left[\frac{\int_0^\infty (f - f_0)^2 S_V(f) df}{\int_0^\infty S_V(f) df} \right]^{1/2} \leq \alpha_2$$

(левая часть этого неравенства представляет собой среднеквадратическую ширину полосы процесса $V(t)$) и

$$\mathcal{E}_{2,U} \leq \mathcal{E}_c$$

найти предсказание цепи $B(s)$, минимизирующее $\mathcal{E}_{2,m}$, и минимальное значение этой величины как функцию от α_1 для различных выборов α_2 .

Ван Трис [5] использовал эти результаты для сравнения эффективности без ограничения полосы ($\alpha_2 = \infty$) с эффективностью, которую можно достичь при передаче гауссовского источника с известным спектром по аддитивному гауссовскому каналу с белым шумом. Последнюю эффективность можно получить, приравняв скорость создания информации при заданном искажении источника и пропускную способность канала [6]. Согласно результату Ван Триса, в случае бесконечной полосы частот оптимальная угловая модуляция лишь на несколько децибелл хуже, чем полная эффективность, которую можно достичь методом модуляции, или кодирования. В противоположность этому следует заметить, однако, что для разумных значений множителя расширения полосы (т.е. отношения среднеквадратической ширины полосы V к ширине полосы M) и больших значений отношений сигнал/шум в канале угловая модуляция дает существенно худшую эффективность, чем теоретически достижимая.

СПИСОК ЛИТЕРАТУРЫ

1. Витерби Э. Д., Принципы когерентной связи, изд-во «Сов. радио», М., 1970.
2. Давенпорт В. Б., Рут В. Л., Введение в теорию случайных сигналов и шумов, ИЛ, М., 1960.
3. Сакрисон (Sakrison D. J.), Communication theory, Transmission of waveforms and digital information, New York, 1968.
4. Бордман, Ван Трис (Boardman C. J., Van Trees H. L., Jr.), Optimum angle modulation, *IEEE Trans. Commun. Tech.*, **COM-13**, 4 (December 1965), 452—469.
5. Ван Трис (Van Trees H. L., Jr.), A comparison of optimum angle modulation systems and rate distortion bounds, *IEEE Proceedings*, **53**, 12 (December 1965), 2123—2124. (Русский перевод: *ТННЭР*, **53**, 12 (1965), 2347—2348.)
6. Гоблик (Goblick T. J., Jr.), Theoretical limitations on the transmission of data from analog sources, *IEEE Trans. Inform. Theory*, **IT-11**, 4 (October 1965), 558—566.

Глава 6

КОДИРОВАНИЕ СЛУЧАЙНЫХ ИСТОЧНИКОВ СООБЩЕНИЙ

В этой главе мы подробно обсудим аспекты теории передачи аналоговых сообщений, связанные с вопросом точности, с которой можно окончательно восстановить эти сообщения в результате прохождения ими некоторой зашумленной или ограниченно передающей среды. Важную роль в нашем обсуждении будут играть некоторые концепции теории информации. Поэтому прежде чем сформулировать задачу, мы приведем интуитивные определения понятий скорости цифровых данных и пропускной способности канала.

Рассмотрим сначала цифровую последовательность, представленную, например, набором двоичных символов,

... 01101001 ...

Любая последовательность, состоящая из Q таких символов, может описать одно из $M = 2^Q$ возможных событий, так как существует ровно M двоичных последовательностей длины Q . Обратно, любое из M возможных событий можно описать с помощью Q двоичных символов, где Q — наименьшее целое число, не меньшее $\log_2 M$. Для описания последовательности таких событий, обладающих к тому же тем свойством, что каждое новое событие происходит один раз в T с, введем величину

$$R = \frac{\log_2 M}{T},$$

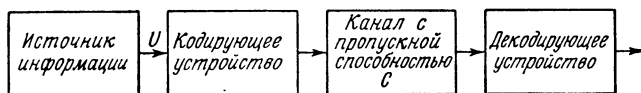
т.е. число двоичных символов в секунду, требуемое для описания реализации этих событий, которую назовем *цифровой скоростью* (в двоичных символах в секунду) последовательности дискретных событий. Ясно, что выбор логарифма по другому основанию, например по основанию 3 или 10, привел бы попросту к тому, что скорость выражалась бы в троичных или десятичных символах в секунду.

Рассмотрим теперь канал связи с шумом (например, аддитивный гауссовский канал), по которому в течение T с передается сигнал $s(t)$, $0 \leq t \leq T$, в результате чего на выходе канала наблюдается выборочная функция случайного процесса $Z(t)$ той же длительности. Оказывается, что в такой схеме любопытным образом возникает следующий пороговый эффект. Существует такое число C , что если R — произвольное число, меньшее C , то при достаточно больших T можно найти $M = 2^{RT}$ различных сигналов $s_m(t)$, $0 \leq t \leq T$, $m = 1, \dots, M$, так, что по наблюдению $Z(t)$ с некоторой малой вероятностью ошибки $P(\mathcal{E})$ можно определить, какой из этих сигналов был передан, причем вероятность ошибки можно произвольно уменьшить, если T достаточно велико, а M растет, как 2^{RT} . Если же R больше C , то такая надежная передача оказывается невозможной. Величина C называется *пропускной способностью* канала и измеряется в битах (двоичных единицах информации) за секунду. Другой выбор основания (вместо основания 2, использованного выше) привел бы попросту к соответствующему изменению размерности пропускной способности.

Таким образом, число C является критической границей для скорости R , при которой возможна *надежная* передача по каналу с шумом. Это замечательное открытие принадлежит Клоду Шеннону. Читателю, не знакомому с его основными идеями, мы настоятельно рекомендуем прочитать отправную статью Шеннона [1], в которой приведено доказательство указанной теоремы кодирования для аддитивного канала с белым гауссовским шумом. Для произвольного случайного канала число C можно определить на основе теоретико-информационных понятий. В настоящей главе те же понятия будут использованы двойственным образом для определения информационной скорости аналогового источника. С теоретико-информационной точкой зрения на понятие пропускной способности канала и на теорему кодирования читатель может познакомиться по работе Шеннона [2, части I—IV] и книге Галлагера [3, гл. 2—5].

Исследуем ситуацию, схематически изображенную на рис. 6.1. Здесь выход источника U может представлять случайный процесс или случайное поле (например,

двумерное случайное поле, возникающее при передаче изображений). Каналом может быть линия телеметрической цифровой связи, не содержащая по существу ошибок, либо цифровое запоминающее устройство, либо канал связи с шумами, имеющий пропускную способность, равную C . Задача состоит в построении кодирующего устройства, позволяющего добиться эффективной передачи в том смысле, что для достижения удовлетворительного уровня воспроизведения исходного сообщения U на



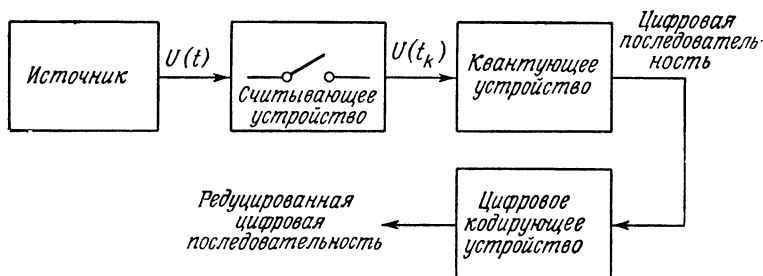
Р и с. 6.1. Проблема кодирования источника.

выходе источника выходным сообщением \hat{U} потребовалась бы минимальная пропускная способность передающего канала.

Если выход источника представляет случайный процесс $U(t)$, а каналом является канал связи, содержащий шумы, эта задача эквивалентна построению модуляционной системы, приводящей к максимизации на приемном конце отношения сигнал/шум при заданных ограничениях на передаваемую мощность, ширину полосы частот и уровень шума в канале. Именно по этой причине сформулированная выше задача отвечает тематике данной книги, а в той формулировке, в какой она рассматривается ниже, эта задача применима к значительно более широкому классу реальных физических проблем, таких, как эффективное хранение аналоговой информации с помощью цифровых запоминающих устройств. В настоящей главе мы не будем пытаться ответить на вопрос, как добиться эффективного кодирования, а лишь определим, какова минимальная пропускная способность канала, требуемая для достижения заданного уровня искажения сообщения, и выясним структуру близкого к оптимальному метода кодирования.

Для пояснения методов, которыми мы будем в дальнейшем пользоваться, опишем один частный подход к нашей задаче и выясним некоторые его недостатки. Рас-

смотрим источник, выходом которого является случайный процесс, и кодирующее устройство (рис. 6.2), состоящее из считывающего устройства, квантователя (преобразующего аналоговое сообщение в цифровое) и цифрового кодирующего устройства. Частота считывания и уровень квантования определяют так, чтобы представить сигнал $U(t)$ с заданной степенью точности. Избыточность



Исходная цифровая последовательность \longrightarrow

1100, 1100, 1100, 1101, 1101, 1101, 1110, 1110, ...

11100, 0, 0, 11101, 0, 0, 11110, 0, ...

Редуцированная цифровая последовательность \longrightarrow

Рис. 6.2. Типичный частотный метод редукции данных.

в полученных таким образом отсчетах используется затем цифровым кодирующим устройством. Простой способ такого использования избыточности состоит в том, что повторившееся значение отсчета заменяется передачей нуля; о появлении же нового отсчета приемник информируется с помощью передачи единицы, вслед за чем передается двоичное представление этого отсчета. Пример такой редукции показан на рис. 6.2, где для представления отсчета используются четыре бита. Этот метод называется интерполяцией нулевого порядка. Он эффективен, если $U(t)$ изменяется не слишком быстро, в результате чего кодируемая последовательность содержит устойчивые интервалы, приводящие к передаче длинных серий нулей. Соответствующий коэффициент сжатия

определяется как A/B , где A — число двоичных символов в секунду в исходной цифровой последовательности, B — число двоичных символов в секунду в закодированной цифровой последовательности.

Хотя указанный метод редукции в некоторых случаях может оказаться весьма эффективным, все же этот частный подход во многом несовершенен, если его рассматривать с точки зрения общей конструкции для изучения задачи. Во-первых, в самой основе этого метода предполагается, что считывание значений функции с определенной частотой и их квантование должны составлять первый шаг при редукции данных к цифровому виду. На самом деле это не так, и существует много различных способов первичного сведения данных к цифровому виду; некоторые из них могут оказаться гораздо более эффективными, чем дискретизация по времени с последующим квантованием. Во-вторых, использование в качестве меры эффективности коэффициента сжатия само обладает рядом серьезных недостатков. Прежде всего эта величина позволяет производить лишь относительное сравнение различных возможных схем, но не дает *абсолютной* меры, с помощью которой можно было бы выделить оптимальное решение задачи. Кроме того, она может оказаться обманчивым показателем успеха, если, например, исходная скорость считывания слишком велика.

При введении теоретической конструкции, применяемой при изучении проблемы кодирования источников, важным первым шагом является определение *меры искажения*, позволяющей количественно измерить расхождение между сигналом на выходе декодирующего устройства \hat{U} и сигналом на выходе источника U . Это позволит сравнить скорости различных методов кодирования произвольного вида при одном и том же уровне искажения. Введенная мера позволит также определить шенноновскую функцию скорости $R(d)$ при заданном искажении.

Важность этой функции подчеркивает следующая теорема Шеннона о кодировании источников:

При передаче сообщений с уровнем искажения d необходим канал связи с пропускной способностью, не меньшей $R(d)$ (независимо от наличия в канале шумов).

Обратно, при соответствующем кодировании сообщения источника могут передаваться со скоростью, незначительно превышающей $R(d)$ при уровне искажения d .

Приступим теперь к изложению шенноновской теории, устанавливающей связь между уровнем искажения и скоростью передачи. В разд. 6.1 мы начнем с ряда рассмотрений, которые приведут нас к геометрическому доказательству теоремы кодирования для источников, порождающих гауссовские случайные величины. Несмотря на то что этот результат носит частный характер, он дает отчетливое представление о методе, с помощью которого достигается эффективное кодирование источников. В разд. 6.2 мы введем необходимые для нас концепции и определения теории информации, в разд. 6.3 докажем теорему кодирования для произвольного вида источников и, наконец, в разд. 6.4 найдем для гауссовского случайного процесса скорость создания информации при заданном искажении, используя критерий среднеквадратической ошибки.

6.1. ГАУССОВСКИЙ ИСТОЧНИК СО СРЕДНЕКВАДРАТИЧЕСКОЙ МЕРОЙ ИСКАЖЕНИЯ

В этом разделе мы дадим геометрический вывод функции скорости при заданном искажении для гауссовских случайных величин в случае, когда мерой искажения служит среднеквадратическая ошибка. Этот результат вряд ли нов — по сути дела он вытекает из ранней работы Шеннона [1]. Однако наглядность геометрического доказательства имеет определенное преимущество для уяснения метода, с помощью которого производится кодирование источника. Став на геометрическую точку зрения, мы лучше уясним как приемы кодирования, так и свойства, которыми должен обладать близкий к оптимальному метод кодирования. Это достигается, однако, ценой того, что геометрические рассуждения ограничиваются случаем гауссовского источника со среднеквадратической ошибкой в качестве меры искажения.

Итак, рассмотрим ситуацию, когда по каналу, допускающему безошибочную передачу со средней скоростью

R бит на случайную величину, необходимо передать последовательность независимых гауссовских случайных величин с нулевым средним и дисперсией σ^2 . Наша цель состоит в том, чтобы с помощью геометрических соображений найти минимальную среднеквадратическую ошибку, с которой можно передать по имеющемуся каналу связи указанные случайные величины.

Начнем с того, что сгруппируем L данных случайных величин в отдельные блоки. Каждый такой блок будет описываться словом в LR двоичных единиц информации, которую можно безошибочно передать получателю. Мы будем рассматривать такой блок, состоящий из случайных величин U_1, U_2, \dots, U_L как L -мерный случайный вектор \mathbf{U} :

$$\mathbf{U} = (U_1, U_2, \dots, U_L), \quad (6.1)$$

предполагая, что L велико. Нашу задачу можно теперь переформулировать следующим образом. Поскольку передачу цифровых символов можно вести со скоростью R двоичных символов на случайную величину, возможно различие $M = 2^{RL}$ значений вектора \mathbf{U} , и, следовательно, в пространстве значений \mathbf{U} можно выбрать M соответствующих *кодowych*, или *изображающих*, векторов $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_M$. Разобьем L -мерное пространство на M областей так, чтобы j -я область состояла из значений \mathbf{u} , находящихся ближе (в евклидовой метрике) к \mathbf{u}_j , чем к любому другому изображающему вектору (точки, лежащие на равном расстоянии, произвольным образом приписываются к одной из областей).

Если случайный вектор \mathbf{U} принял значение \mathbf{u} , то мы будем передавать (не совершая ошибок) номер j ближайшего к \mathbf{u} изображающего вектора \mathbf{u}_j , т.е. номер области, содержащей \mathbf{u} . После этого приемник использует \mathbf{u}_j в качестве аппроксимации значения \mathbf{u} , принятого вектором \mathbf{U} . Если в качестве скалярного произведения в L -мерном евклидовом пространстве, содержащем наши наблюдения, принять

$$(\mathbf{u}, \tilde{\mathbf{u}}) = \frac{1}{L} \sum_{k=1}^L u_k \tilde{u}_k, \quad (6.2)$$

то расстояние

$$\|\mathbf{u} - \mathbf{u}_j\|^2 = \frac{1}{L} \sum_{k=1}^L (u_k^j - u_k)^2 \quad (6.3)$$

будет представлять собой среднюю ошибку на одну компоненту, когда вектор \mathbf{U} принял значение $\mathbf{u} = (u_1, u_2, \dots, u_L)$, а в качестве аппроксимации вектора \mathbf{u} использован вектор $\mathbf{u}_j = (u_1^j, u_2^j, \dots, u_L^j)$. Наша цель — найти минимальную возможную среднюю ошибку аппроксимации, когда размерность L велика, а $M = 2^{RL}$ изображающих векторов выбираются оптимальным образом.

Сделаем ряд замечаний, касающихся наиболее вероятных значений вектора \mathbf{U} . Случайная величина $L\|\mathbf{U}\|^2$ имеет χ^2 -распределение с L степенями свободы. Используя известный вид плотности этого распределения [4], нетрудно показать, что

$$\mathbf{E}\{\|\mathbf{U}\|\} = \sigma \sqrt{\frac{2}{L} \frac{\Gamma(L/2 + 1/2)}{\Gamma(L/2)}}, \quad (6.4)$$

$$\mathbf{E}\{\|\mathbf{U}\|^2\} = \sigma^2 \quad (6.5)$$

и, следовательно,

$$\text{var}\{\|\mathbf{U}\|\} = \sigma^2 \left[1 - \frac{2}{L} \frac{\Gamma^2(L/2 + 1/2)}{\Gamma^2(L/2)} \right]. \quad (6.6)$$

С помощью разложения для $\ln \Gamma$ с остаточным членом [5] нетрудно получить оценку

$$\text{var}\{\|\mathbf{U}\|\} \leq \frac{\sigma^2}{1,2L}. \quad (6.7)$$

Таким образом, при больших значениях L величина $\|\mathbf{U}\|$ близка в среднеквадратическом к своему математическому ожиданию. Зададим поэтому случайный вектор \mathbf{U}_p равенством

$$\mathbf{U}_p = \mathbf{U} \frac{\mathbf{E}\{\|\mathbf{U}\|\}}{\|\mathbf{U}\|}, \quad (6.8)$$

так что длина \mathbf{U}_p есть

$$r_0 = \|\mathbf{U}_p\| = \mathbf{E}\{\|\mathbf{U}\|\}.$$

Так как

$$\mathbf{E}\{\|\mathbf{U} - \mathbf{U}_p\|^2\} = \text{var}\{\|\mathbf{U}\|\} \leq \frac{\sigma^2}{1,2L}, \quad (6.9)$$

то при больших L среднеквадратическая ошибка аппроксимации \mathbf{U} вектором \mathbf{U}_p мала. Далее, квантуя значение \mathbf{U}_p ближайшей точкой из набора \mathbf{U}_j , мы получим, что общая среднеквадратическая ошибка приемника равна сумме ошибки квантования и величины $\text{var} \{ \|\mathbf{U}\| \}$:

$$\begin{aligned} d &= \mathbf{E} \{ \|\mathbf{U} - \mathbf{u}_j\|^2 \} = \mathbf{E} \{ \|\mathbf{U} - \mathbf{U}_p + \mathbf{U}_p - \mathbf{u}_j\|^2 \} = \\ &= \mathbf{E} \{ \|\mathbf{U}_p - \mathbf{u}_j\|^2 \} + \text{var} \{ \|\mathbf{U}\| \}. \end{aligned} \quad (6.10)$$

Задача 6.1. Показать, что перекрестный член

$$\mathbf{E} \{ (\mathbf{U} - \mathbf{U}_p, \mathbf{U}_p - \mathbf{u}_j) \}$$

равен нулю.

Поскольку величину $\text{var} \{ \|\mathbf{U}\| \}$ можно сделать сколь угодно малой за счет выбора достаточно большого L , в процессе кодирования, или «квантования», мы ничего не теряем, проектируя сначала вектор \mathbf{U} на сферу радиуса $\mathbf{E} \{ \|\mathbf{U}\| \}$. Преимущество такого проектирования состоит в том, что вектор \mathbf{U}_p имеет равномерное распределение на L -мерной сфере радиуса $\mathbf{E} \{ \|\mathbf{U}\| \} = r_0$.

В применении к передаче последовательности независимых гауссовских случайных величин с дисперсией σ^2 шенноновская теория скорости при заданном искажении [2, часть V; 6] приводит к следующему результату:

Теорема 6.1 (теорема кодирования). Каков бы ни был метод кодирования при передаче по каналу с пропускной способностью R бит на случайную величину, нельзя добиться передачи указанных выше величин со среднеквадратической ошибкой, меньшей чем

$$d(R) = \sigma^2 2^{-2R}. \quad (6.11)$$

Однако, кодируя случайные величины достаточно длинными блоками (т.е. выбирая L достаточно большим), можно добиться передачи со среднеквадратической ошибкой, сколь угодно близкой к $d(R)$.

Учитывая сделанные выше замечания относительно распределения \mathbf{U}_p и величины $\text{var} \{ \|\mathbf{U}\| \}$, дадим геометрическое доказательство этого результата. В наших рассуждениях мы опустим некоторые, необходимые, конечно, для достижения полной строгости, выкладки (их не-

трудно будет воспроизвести); это позволит сделать основные этапы доказательства более доступными.

Докажем сначала негативную часть теоремы кодирования, т.е. утверждение о том, что величина $d(R)$ в (6.11) определяет минимальную ошибку, совместимую с передачей со скоростью R . Это утверждение будет доказано в три этапа.

Во-первых, заметим, что поскольку $\text{var} \{\|U\|\}$ стремится к нулю при $L \rightarrow \infty$, асимптотически мы ничего не проигрываем, проектируя U на сферу радиуса r_0 .

Во-вторых, выясним, какова наиболее эффективная форма области, отвечающей данному кодовому вектору. Воспользуемся тем, что поскольку вектор U_p равномерно распределен на сфере сообщений радиуса r_0 , оптимальной формой области будет форма, при которой для данного значения A минимизируется величина \mathcal{E}^2 , где

A — площадь пересечения сферы сообщений с рассматриваемой областью,

\mathcal{E}^2 — момент инерции этой поверхности относительно кодового вектора (точки u_j), равный среднеквадратической ошибке при квантовании проектируемых на эту поверхность векторов данным кодовым вектором.

Нетрудно видеть, каким должен быть оптимальный вид области, представляемой кодовым вектором. Пусть направление кодового вектора u_j выбрано произвольно, и пусть u_p — точка пересечения сферы сообщений с прямой, проходящей через центр сферы в направлении u_j . Эта ситуация показана на рис. 6.3. Очевидно, что точка u_p ближе к u_j , чем любая другая точка на сфере. Кроме того, беря u_p в качестве полюса нашей сферы, мы видим, что все точки, имеющие фиксированный полярный угол ϕ , находятся на одинаковом расстоянии от u_j , причем это расстояние монотонно возрастает с ростом ϕ .

Таким образом, желая выделить на сфере сообщений область с заданной площадью, мы можем минимизировать ее момент инерции относительно u_j , располагая ее как можно ближе к полюсу. Тогда оптимальным видом области будет часть сферы, имеющая форму шапки, симметричной относительно полярной оси. Если L велико,

то почти вся площадь поверхности $((L - 1)$ -мерного объема) такой области сосредоточена у самого края шапки, так что для оптимальной области значение \mathcal{E}^2 пропорционально квадрату расстояния от \mathbf{u}_j до края шапки.

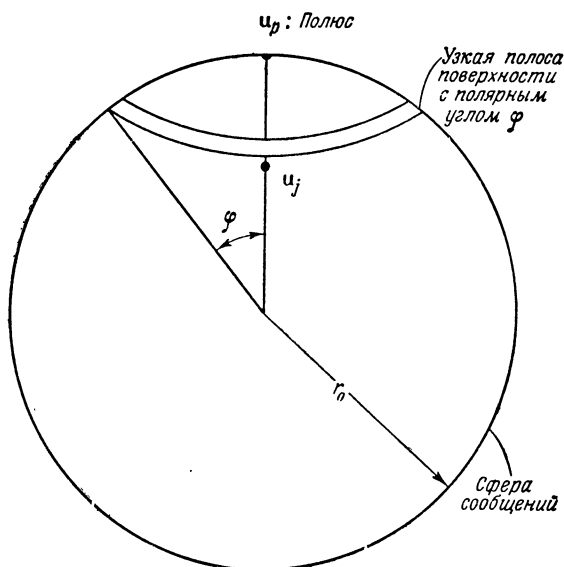


Рис. 6.3. Геометрическое изображение вида оптимальной области представляемой кодовым вектором.

Ясно, что эта величина минимальна, когда точка \mathbf{u}_j отстоит от центра сферы на расстоянии $\sqrt{r_0^2 - \mathcal{E}^2}$ (напомним, что оптимальный вид области получен независимо от расположения \mathbf{u}_j на полярном радиусе). Итак, оптимальной формой изображаемой области является шар радиуса \mathcal{E} с центром в изображающей точке, находящейся на расстоянии $\sqrt{r_0^2 - \mathcal{E}^2}$ от начала координат.

Последний этап нашего доказательства — нахождение с учетом оптимального вида изображаемых областей минимального их числа, необходимого в процессе кодирования. Поскольку \mathbf{U}_p имеет равномерное распределение на сфере сообщений, любая часть поверхности, за исключе-

нием, быть может, пренебрежимо малой доли, должна быть покрыта изображаемыми областями. Следовательно, число M требуемых областей ограничено снизу:

$$\begin{aligned} \text{площадь сферы сообщений} &= A_L r_0^{L-1} \leq \\ &\leq M [\text{площадь полярной шапки с диаметром} \\ &\quad \text{основания } 2\mathcal{E}]; \quad (6.12) \end{aligned}$$

здесь A_L — площадь L -мерной сферы единичного радиуса, а число областей M связано со скоростью передачи R равенством $M = 2^{2L}$. Поскольку площадь поверхности полярной шапки, очевидно, не превосходит площади сферы радиуса \mathcal{E} , неравенство (6.12) можно заменить более грубым неравенством

$$A_L r_0^{L-1} \leq M A_L \mathcal{E}^{L-1} = 2^{RL} A_L \mathcal{E}^{L-1}.$$

Так как r_0 асимптотически совпадает с σ , после простых преобразований получаем

$$\frac{\mathcal{E}}{\sigma} \geq 2^{-RL/(L-1)},$$

откуда при достаточно больших L следует, что

$$\frac{\mathcal{E}}{\sigma} \geq 2^{-R}. \quad (6.13)$$

Это неравенство доказывает негативную часть теоремы кодирования.

При доказательстве позитивного утверждения теоремы воспользуемся методом случайного кодирования. Суть метода состоит в том, что кодовые векторы выбираются случайным образом, причем так, что для любой фиксированной точки \mathbf{U} вероятность не находиться внутри шара радиуса \mathcal{E} с центром в одной из $M = 2^{RL}$ кодовых точек пренебрежимо мала, если $\mathcal{E} > \sqrt{d(R)}$.

Учитывая сделанные выше замечания относительно вида оптимальных областей и оптимального расположения кодовых векторов, выберем $M = 2^{RL}$ точек $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_M$ независимо друг от друга в соответствии с равномерным распределением на сфере радиуса $\sqrt{r_0^2 - \mathcal{E}^2}$.

Рассмотрим произвольную точку u_p , полученную проектированием сообщения на сферу радиуса $\sqrt{r_0^2 - \mathcal{E}^2}$. Оценим вероятность того, что внутри шара радиуса \mathcal{E} с центром u_p не найдется ни одного кодового вектора.

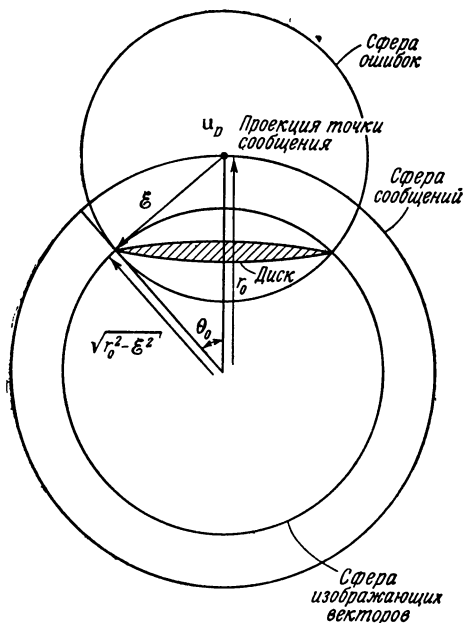


Рис. 6.4. Иллюстрация построений, используемых для вычисления P_g .

Обозначим эту вероятность через P_g . Для удобства введем обозначения

V_L — объем L -мерного шара единичного радиуса,

A_L — площадь $((L-1)$ -мерный объем) L -мерный сфера единичного радиуса.

На рис 6.4 показаны произвольное спроектированное сообщение u_p и сфера радиуса \mathcal{E} с центром в точке u_p .

Из этого рисунка и из вида распределения, с которым выбираются кодовые векторы, ясно, что

$$\mathbf{P}_{\mathcal{E}} = \left[1 - \frac{\text{площадь сферической шапки радиуса } \sqrt{r_0^2 - \mathcal{E}^2} \text{ с полярным углом } \theta_0}{A_L (r_0^2 - \mathcal{E}^2)^{(L-1)/2}} \right]^M, \quad (6.14)$$

где

$$\theta_0 = \arctg \frac{\mathcal{E}}{\sqrt{r_0^2 - \mathcal{E}^2}}, \quad M = 2^{RL}. \quad (6.15)$$

Далее, как видно из рис. 6.4, площадь сферической шапки радиуса $\sqrt{r_0^2 - \mathcal{E}^2}$ с полярным углом θ_0 ограничена снизу площадью $(L-1)$ -мерного диска радиуса $\sin \theta_0 \sqrt{r_0^2 - \mathcal{E}^2}$, которая равна

$$V_{L-1} (\sin \theta_0 \sqrt{r_0^2 - \mathcal{E}^2})^{L-1} = V_{L-1} \left(\frac{\mathcal{E}}{r_0} \sqrt{r_0^2 - \mathcal{E}^2} \right)^{L-1}. \quad (6.16)$$

Заметим, что

$$V_n = \int_0^1 A_n r^{n-1} dr = \frac{1}{n} A_n. \quad (6.17)$$

Из (6.14), (6.16) и (6.17) имеем

$$\mathbf{P}_{\mathcal{E}} \leq \left[1 - \frac{A_{L-1} (\mathcal{E}/r_0)^{L-1}}{(L-1) A_L} \right]^M. \quad (6.18)$$

Используя (6.17) и хорошо известное выражение для V_L (см. [7]), находим

$$\frac{A_{L-1}}{(L-1) A_L} = \frac{1}{\sqrt{\pi}} \frac{\Gamma(L/2 + 1)}{L \Gamma((L-1)/2 + 1)}.$$

Применяя формулу Стирлинга, получаем при больших L

$$\frac{A_{L-1}}{(L-1) A_L} \approx \frac{1}{\sqrt{2\pi L}}. \quad (6.19)$$

Неравенство (6.18) вместе с (6.15) и (6.19) дает

$$\mathbf{P}_{\mathcal{E}} \leq \left[1 - \frac{1}{\sqrt{2\pi L}} \left(\frac{\mathcal{E}}{r_0} \right)^{L-1} \right]^{2^{LR}}. \quad (6.20)$$

Пусть

$$\mathcal{E} = (1 + \varepsilon) \frac{r_0}{\sigma} \sqrt{d(R)}.$$

Тогда, согласно (6.11),

$$\mathcal{E} = (1 + \varepsilon) r_0 2^{-R}, \quad \varepsilon > 0. \quad (6.21)$$

Так как r_0 асимптотически совпадает с σ , то при больших L значение \mathcal{E}^2 близко к значению d , определяемому (6.11). Подставляя (6.21) в (6.20), находим

$$P_{\mathcal{E}} \approx \left[1 - \frac{1}{\sqrt{2\pi L}} 2^{-R(L-1)} (1 + \varepsilon)^{L-1} \right]^{2RL}, \quad (6.22)$$

или (при больших L)

$$P_{\mathcal{E}} \approx \exp \left[- \frac{(1 + \varepsilon)^L 2^R}{\sqrt{2\pi L} (1 + \varepsilon)} \right]. \quad (6.23)$$

Итак, вероятность данной точке, лежащей на сфере сообщений, не попасть внутрь шара радиуса \mathcal{E} с центром в одной из кодовых точек стремится к нулю, если \mathcal{E} определяется соотношением (6.21) с $\varepsilon > 0$.

Посмотрим теперь, в какой мере мы использовали предположение о гауссовости распределения случайного вектора \mathbf{U} . При доказательстве негативного утверждения теоремы мы воспользовались сферической симметрией распределения \mathbf{U} , и, следовательно, предположение о гауссовости здесь оказалось существенным. При доказательстве позитивной части теоремы мы опирались лишь на тот факт, что можно пренебречь ошибкой, связанной с проектированием вектора сообщений на сферу радиуса r_0 . Таким образом, здесь нам понадобилось лишь выполнение неравенства (6.9).

Если распределение случайной величины U не является гауссовским, то соотношение (6.10) можно заменить оценкой

$$d \leq E \{ \|\mathbf{U}_p - \mathbf{u}_j\|^2 \} + V + 2r_0 V^{1/2},$$

где

$$V = E \{ (\|\mathbf{U}\| - r_0)^2 \},$$

а r_0 — среднее значение $\|\mathbf{U}\|$, отвечающее по-прежнему гауссовскому распределению. Следовательно, среднеква-

дратической ошибкой, вызванной проектированием вектора \mathbf{U} на сферу радиуса r_0 , можно пренебречь, если V стремится к нулю при $L \rightarrow \infty$.

Задача 6.2. Показать, что $\lim_{L \rightarrow \infty} V = 0$ для любой случайной величины U с конечной дисперсией.

Теперь мы можем сформулировать следующее утверждение для случая передачи сообщений, имеющих конечную дисперсию.

Теорема 6.2. Пусть некоторая случайная величина имеет дисперсию σ^2 . Рассмотрим блоковую передачу независимых случайных величин с тем же распределением, т. е. передачу вектора

$$\mathbf{U} = (U_1, U_2, \dots, U_L).$$

Тогда если L достаточно велико, то при скорости передачи R бит на случайную величину возможна блоковая передача со среднеквадратической ошибкой, сколь угодно близкой к $\sigma^2 2^{-2R}$.

Таким образом, среди всех распределений с конечной дисперсией σ^2 гауссовское распределение будет «наихудшим» в том смысле, что скорость передачи при заданном искажении для гауссовских случайных величин служит верхней границей скорости при заданном искажении, отвечающей произвольному распределению с конечной дисперсией. Это свойство, связанное со сферической симметрией гауссовского распределения, впервые было отмечено Шенноном [2] в 1948 г.

6.2. СВОЙСТВА ВЗАИМНОЙ ИНФОРМАЦИИ

Для нахождения функции скорости при заданном искажении в случае произвольного источника сообщений нам понадобятся некоторые определения теории информации. Часть необходимых сведений мы постараемся вкратце изложить в этом разделе; читателю, интересующемуся более полным обсуждением, мы предлагаем ознакомиться с книгами Галлагера [3, гл. 2] и Пинскера [8, гл. 2, 3].

Всюду в этом разделе буквами U , X , Y и V будут обозначаться векторные случайные величины счетной размерности. Таким образом, X может быть как скалярной случайной величиной или конечномерным случайным вектором, так и сепарабельным случайным процессом (полем). Через \mathcal{X} мы будем обозначать отвечающее случайной величине X пространство ее выборочных значений x .

Пусть \mathcal{F}_x обозначает σ -поле в пространстве \mathcal{X} , т. е. множество событий (подмножеств из \mathcal{X}), удовлетворяющих следующим условиям:

1) \mathcal{F}_x не пусто, т. е. состоит хотя бы из одного множества;

2) дополнение любого множества из \mathcal{F}_x , а также объединение или пересечение любого счетного набора таких множеств принадлежат \mathcal{F}_x .

Случайная величина X является функцией выборочной точки ω , принадлежащей основному выборочному пространству Ω . При этом исходная вероятностная мера, заданная на подмножествах из Ω , порождает распределение вероятностей величины X , т. е. вероятностную меру — обозначим ее $\mathbf{P}(\cdot)$, — заданную на подмножествах, или событиях, из \mathcal{F}_x . Эта функция множеств удовлетворяет равенствам $\mathbf{P}(\Omega) = 1$ и

$$\mathbf{P}\left(\bigcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} \mathbf{P}(E_k),$$

где $\{E_k\}$ — произвольная последовательность непересекающихся подмножеств, принадлежащих \mathcal{F}_x . Функция множеств, удовлетворяющая последнему соотношению, называется *счетно аддитивной*.

В частном случае, когда X — скалярная случайная величина, соответствующее распределение вероятностей определяется функцией распределения

$$F_X(x) = \mathbf{P}\{X(\omega) \leq x\};$$

в этом случае вероятность любого события E , связанного с X , равна

$$\mathbf{P}(E) = \int_E d\mathbf{P}(\omega) = \int_E dF_X(x).$$

Если функция $F_X(x)$ дифференцируема по x (существует производная Радона — Никодима меры $\mathbf{P}_X(x)$ относительно меры Лебега на действительной прямой), то распределение вероятностей описывается функцией плотности

$$f_X(x) = \frac{d}{dx} F_X(x),$$

$$\mathbf{P}(E) = \int_E f_X(x) dx.$$

Предположим, что возможными распределениями величины X являются две вероятностные меры \mathbf{P}_1 и \mathbf{P}_2 . Для любого события $E \in \mathcal{F}_X$, для которого $\mathbf{P}_2(E) \neq 0$, рассмотрим отношение

$$\mathbf{P}_1(E)/\mathbf{P}_2(E).$$

В дальнейшем мы ограничимся случаем, когда $\mathbf{P}_2(E) \neq 0$ для всех событий $E \in \mathcal{F}_X$, для которых $\mathbf{P}_1(E) \neq 0$. Тогда мы сможем рассмотреть «исчезающее» множество E , т. е. последовательность все меньших и меньших событий, каждое из которых содержит выборочную точку x . Естественно возникает вопрос, будет ли при этом отношение

$$\mathbf{P}_1(E)/\mathbf{P}_2(E)$$

сходиться к какому-нибудь пределу, который в этом случае представлял бы собой функцию от x ? В действительности это так и будет, соответствующий предел обозначается

$$\Lambda(x) = \frac{d\mathbf{P}_1(x)}{d\mathbf{P}_2(x)}$$

и называется производной Радона — Никодима меры \mathbf{P}_1 относительно \mathbf{P}_2 . Можно также показать, что

$$\mathbf{P}_1(E) = \int_E \Lambda(x) d\mathbf{P}_2(x).$$

Отметим, что если X — скалярная случайная величина, а \mathbf{P}_1 и \mathbf{P}_2 определяются своими функциями плотности $f_1(x)$ и $f_2(x)$, то

$$\Lambda(x) = \frac{d\mathbf{P}_1(x)}{d\mathbf{P}_2(x)} = \frac{f_1(x)}{f_2(x)},$$

$$\mathbf{P}_1(E) = \int_E \Lambda(x) f_2(x) dx = \int_E f_1(x) dx.$$

В гл. 3 предел отношения правдоподобия $\Lambda_{b, \beta}(W)$ также представлял собой производную Радона — Никодима меры, отвечающей $W(t)$ при значениях параметров b, β , относительно той же меры при $b = 0$. Эти два частных случая более привычны, и читателю полезно убедиться в том, что определение Λ в этих случаях совпадает с общим определением Λ , данным выше.

Хотя мы не требуем от читателя свободного владения понятием производной Радона — Никодима, нам придется, к сожалению, пользоваться им в этом и в следующем разделах. Тем, для кого это понятие ново, можно посоветовать при чтении этих разделов понимать под величиной Λ отношение двух вероятностных плотностей. С подробным обсуждением относящихся сюда результатов читатель может ознакомиться по любой книге, посвященной теории меры, или же по книге Лозва [9, часть I].

Рассмотрим теперь две векторные случайные величины X и Y с совместным распределением \mathbf{P}_{XY} , определенным в пространстве соответствующих им значений. Пусть E и F — произвольные события соответственно в \mathcal{F}_x и \mathcal{F}_y . Произведение мер, отвечающих X и Y , определяется соотношением

$$\mathbf{P}_{X \times Y}(EF) = \mathbf{P}_X(E) \mathbf{P}_Y(F), \quad (6.24)$$

где \mathbf{P}_X и \mathbf{P}_Y — маргинальные распределения, соответствующие X и Y . В том случае, когда распределение случайных величин X и Y характеризуется совместной плотностью вероятностей $f_{XY}(x, y)$, введенное нами произведение мер определяется плотностью $f_X(x)f_Y(y)$, где $f_X(x)$ и $f_Y(y)$ — маргинальные плотности, получающиеся интегрированием $f_{XY}(x, y)$ соответственно по y или по x .

Определим разбиение пространства \mathcal{X} значений вектора X как произвольный конечный набор непересекающихся множеств $E_1, \dots, E_N \in \mathcal{F}_x$, объединение которых совпадает с \mathcal{X} :

$$\bigcup_{n=1}^N E_n = \mathcal{X}. \quad (6.25)$$

Аналогично определяется разбиение пространства \mathcal{Y} .

Зададим среднюю взаимную информацию между X и Y соотношением

$$I(X; Y) = \sup \sum_{m, n} \mathbf{P}_{XY}(E_m F_n) \ln \left[\frac{\mathbf{P}_{XY}(E_m F_n)}{\mathbf{P}_{X \times Y}(E_m F_n)} \right], \quad (6.26)$$

где верхняя грань берется по всевозможным разбиениям пространств \mathcal{X} и \mathcal{Y} .

Отметим важную характерную особенность определения взаимной информации, данного с помощью разбиений пространств \mathcal{X} и \mathcal{Y} . Если величина $I(X; Y)$ конечна, то она произвольно мало отличается от средней взаимной информации двух дискретных случайных величин (аппроксимирующих X и Y), и, следовательно, ей присущи многие важные свойства, которыми обладает величина взаимной информации в дискретном случае. Трудность, связанная с определением (6.26), состоит в том, что оно не дает алгоритма для вычисления $I(X; Y)$. Однако, как можно ожидать из внешнего вида формулы (6.26), величину под знаком логарифма можно приближенно заменить производной Радона — Никодима, а всю сумму можно приблизить интегралом. Этот факт действительно можно обосновать, и мы приведем здесь следующий результат, принадлежащий Гельфанду и Яглому¹⁾.

Теорема 6.3. Пусть $\mathbf{P}_{X \times Y} \neq 0$ для всех событий $E \times F$, принадлежащих $\mathcal{F}_{x \times y}$, для которых $\mathbf{P}_{XY} \neq 0$. Тогда

$$I(X; Y) = \int_{\mathcal{X} \times \mathcal{Y}} \ln \Lambda_{XY}(x, y) d\mathbf{P}_{XY}(x, y), \quad (6.27)$$

¹⁾ Доказательство см. в разд. 2.4 статьи Добрушина Р. Л., Общая формулировка основной теоремы Шеннона в теории информации, *УМН*, 14, вып. 6 (1959), 3—104. — *Прим. перев.*

где

$$\Lambda_{XY}(x, y) = \frac{d\mathbf{P}_{XY}(x, y)}{d\mathbf{P}_{X \times Y}(x, y)}. \quad (6.28)$$

В противном случае $I(X; Y) = \infty$.

Заметим, что если распределение случайных величин X и Y задается совместной плотностью $f_{XY}(x, y)$, то несколько отпугивающая формула (6.27) принимает более простой вид

$$I(X; Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} dx dy f_{XY}(x, y) \ln \left[\frac{f_{XY}(x, y)}{f_X(x) f_Y(y)} \right]. \quad (6.29)$$

Перечислим некоторые свойства средней взаимной информации.

$$1) I(X; Y) \geq 0. \quad (6.30)$$

Это неравенство следует из определения и соответствующего свойства количества информации между двумя дискретными случайными величинами (см. Галлагер [3, теорема 2.3.2]).

$$2) I(X; Y) = I(Y; X). \quad (6.31)$$

Это равенство непосредственно вытекает из определения.

3) Пусть пара X_1, X_2 не зависит от пары Y_1, Y_2 . Тогда

$$I(X_1, Y_1; X_2, Y_2) = I(X_1; X_2) + I(Y_1; Y_2). \quad (6.32)$$

Это равенство также следует из определения и из соотношения

$$\mathbf{P}_{X_1 X_2 Y_1 Y_2}(E_1 E_2 F_1 F_2) = \mathbf{P}_{X_1 X_2}(E_1 E_2) \mathbf{P}_{Y_1 Y_2}(F_1 F_2).$$

4) Пусть Z — измеримая функция от Y , т.е. вероятность, заданная на множествах из \mathcal{F}_Y , определяет вероятностную меру на \mathcal{F}_Z . Тогда

$$I(X; Y, Z) = I(X; Y) \quad (6.33)$$

(заметим, что всегда $I(X; Y, Z) \geq I(X; Y)$) и

$$I(X; Y) \geq I(X; Z). \quad (6.34)$$

Равенство (6.33) верно в силу того, что любое разбиение, определяемое величинами Y и Z , можно определить и

с помощью одной лишь величины Y . (Аналогично доказывается в общем случае соответствующее неравенство.) Неравенство (6.34) справедливо, поскольку любое разбиение, определяемое Z , можно определить с помощью Y ; равенство здесь выполняется в том случае, когда соответствие между Y и Z взаимно однозначно.

5) Обозначим через I_{XY} случайную величину, получаемую при подстановке X и Y вместо x и y в функцию $\ln \Lambda_{XY}(x, y)$. Тогда

$$\mathbf{E}\{|I_{XY}|\} \leq I(X; Y) + \frac{2}{e} = \mathbf{E}\{I_{XY}\} + \frac{2}{e}. \quad (6.35)$$

Задача 6.3. Доказать неравенство (6.35).

Опишем теперь специальный способ задания совместного распределения двух случайных величин (процессов) U и V , важный с точки зрения практики. Пусть величина U имеет распределение $\mathbf{P}_U(u)$ ¹⁾, порождаемое на выходе источника, и для любого значения u задано условное распределение $\mathbf{P}(v|u)$ величины V (определяемое, например, переходной функцией канала). Снабдим распределение источника индексом α , а переходное распределение — индексом γ . Тогда совместное распределение величин U и V можно задать равенством

$$\mathbf{P}_{UV}(u, v) = \mathbf{P}_\gamma(v|u) \mathbf{P}_\alpha(u). \quad (6.36)$$

Заметим, что маргинальное распределение величины V зависит как от α , так и от γ .

Исследуем свойства совместного распределения как функции от γ при фиксированном α . Пусть, в частности, \mathbf{P}_{γ_1} и \mathbf{P}_{γ_2} — два переходных распределения. Определим для любого θ , $0 \leq \theta \leq 1$, распределение $\mathbf{P}_{\bar{\gamma}}$ формулой

$$\mathbf{P}_{\bar{\gamma}}(v|u) = \theta \mathbf{P}_{\gamma_1}(v|u) + (1 - \theta) \mathbf{P}_{\gamma_2}(v|u). \quad (6.37)$$

Тогда

$$I_{\alpha\bar{\gamma}}(U; V) \leq \theta I_{\alpha\gamma_1}(U; V) + (1 - \theta) I_{\alpha\gamma_2}(U; V), \quad (6.38)$$

¹⁾ Обозначение $\mathbf{P}_U(u)$ не вполне корректно, поскольку аргументом функции $\mathbf{P}_U(\cdot)$ является событие, т. е. подмножество из \mathcal{F}_u . Однако оно удобно, так как указывает, какой случайной величиной порождается σ -поле, на котором рассматривается функция множеств $\mathbf{P}(\cdot)$.

где индексы α, γ при I указывают, при каком совместном распределении величин U и V вычисляется взаимная информация. Другими словами, средняя взаимная информация выпукла (вниз) как функция от γ .

Задача 6.4. Доказать неравенство (6.38). [Указания. 1) Выберите надлежащее разбиение. 2) При $N = 2$ воспользуйтесь неравенством

$$\sum_{n=1}^N r_n \ln \frac{r_n}{u_n} \geq (r_1 + \dots + r_N) \ln \frac{r_1 + \dots + r_N}{u_1 + \dots + u_N}, \quad (6.39)$$

справедливым для любых наборов неотрицательных чисел r и u .]

Рассмотрим три случайные величины (процесса) X , Y и V и введем величину средней взаимной информации между X и Y при фиксированном значении v , принятом случайной величиной V . Считая $\mathbf{P}_{XY|V}(x, y|v)$ для произвольного v совместным распределением величин X и Y , можно, как и выше, определить произведение соответствующих маргинальных распределений, а также производную Радона — Никодима

$$\Lambda_{XY|V}(x, y|v) = \frac{d\mathbf{P}_{XY|V}}{d\mathbf{P}_{X \times Y|V}}.$$

Тогда для каждого значения v мы получаем условную взаимную информацию

$$I(X; Y|v) = \int_{\mathcal{X} \times \mathcal{Y}} \ln \Lambda_{XY|V}(x, y|v) d\mathbf{P}_{XY|V}(x, y|v).$$

Усредненная по всевозможным значениям v , эта величина называется *средней условной взаимной информацией*:

$$I(X; Y|V) = \int_{\mathcal{X} \times \mathcal{Y} \times \mathcal{V}} \ln \Lambda_{XY|V}(x, y|v) d\mathbf{P}_{XYV}(x, y, v). \quad (6.40)$$

Предположим на время, что X , Y и V — случайные величины, распределение которых задается плотностью $f_{XYV}(x, y, v)$. Тогда

$$\begin{aligned} \frac{d\mathbf{P}_{XYV}}{d\mathbf{P}_{X \times YV}} &= \frac{f_{XYV}(x, y, v)}{f_X(x) f_{YV}(y, v)} = \frac{f_{XY|V}(x, y|v) f_V(v) f_{X|V}(x|v)}{f_{Y|V}(y|v) f_V(v) f_X(x) f_{X|V}(x, v)} = \\ &= \frac{f_{XY|V}(x, y|v)}{f_{X|V}(x|v) f_{Y|V}(y|v)} \cdot \frac{f_{XV}(x, v)}{f_X(x) f_V(v)} = \frac{d\mathbf{P}_{XY|V}}{d\mathbf{P}_{X \times Y|V}} \cdot \frac{d\mathbf{P}_{XV}}{d\mathbf{P}_{X \times V}}. \end{aligned} \quad (6.41)$$

Равенство между крайними членами, полученное в предположении существования совместной плотности, выполняется и в общем случае, когда X , Y и V — произвольные векторные случайные величины счетной размерности.

Беря логарифм и усредняя по мере \mathbf{P}_{XYV} , получаем формулу Колмогорова

$$I(X; Y, V) = I(X; V) + I(X; Y|V). \quad (6.42)$$

Рассмотрим теперь случай, когда случайные величины (процессы) X , Y и V образуют цепь Маркова в том смысле, что при фиксированном значении средней крайние величины независимы, т. е.

$$\mathbf{P}_{XV|Y}(E, F|y) = \mathbf{P}_{X|Y}(E|y) \mathbf{P}_{V|Y}(F|y), \quad E \in \mathcal{F}_x, \quad F \in \mathcal{F}_v. \quad (6.43)$$

Задача 6.5. Показать, что если величины X , Y и V образуют цепь Маркова, то

$$I(X; V|Y) = 0. \quad (6.44)$$

Задача 6.6. Пусть величины U , X , Y и V образуют цепь Маркова в том смысле, что равенство (6.43) справедливо как для U , X и Y , так и для X , Y и V . Показать, что

$$I(U; V) \leq I(X; Y). \quad (6.45)$$

Этот результат известен под названием *теоремы обработки данных*.

6.3. СКОРОСТЬ ПРИ ЗАДАННОМ ИСКАЖЕНИИ И ТЕОРЕМА КОДИРОВАНИЯ ИСТОЧНИКОВ

Функция скорости при заданном искажении

Рассмотрим источник сообщений, выходом которого является случайная величина U с фиксированным распределением.

Пусть \tilde{U} — величина на выходе некоторого декодирующего устройства, образующего каскадную цепь вместе с кодирующим устройством и каналом связи. Обозначим через γ переходное распределение величины \tilde{U} при заданном значении $U = u$. Можно считать, что распределение величины \tilde{U} индуцируется некоторой пробной или гипотетической каскадной цепью.

Мы будем предполагать, что искажение, возникающее при передаче по каналу связи, измеряется расстоянием $d(u, \tilde{u})$ между величинами выходных сигналов источника u и декодирующего устройства \tilde{u} . В связи с этим введем количественную меру искажения, возникающего при передаче сообщений, как среднее значение случайной величины $D = d(U, \tilde{U})$:

$$d = \mathbf{E}_{\alpha\gamma}\{D\} = \int \int d(u, \tilde{u}) d\mathbf{P}_{\alpha\gamma}(u, \tilde{u}). \quad (6.46)$$

Индексы α и γ указывают на то, что эта величина зависит как от распределения источника α , так и от переходного распределения γ .

Обозначим через $\Gamma_\alpha(d)$ множество переходных распределений γ , для которых средняя величина искажения не превосходит заданного уровня d (мы считаем, что функция $d(\cdot, \cdot)$ неотрицательна):

$$\Gamma_\alpha(d) = \{\gamma: \mathbf{E}_{\alpha\gamma}\{d(U, \tilde{U})\} \leq d\}. \quad (6.47)$$

Заметим, что средняя взаимная информация $I_{\alpha\gamma}(U; \tilde{U})$ между U и \tilde{U} также зависит от γ . Принадлежащее Шеннону понятие функции скорости при заданном искажении определяется формулой

$$R_\alpha(d) = \inf_{\gamma \in \Gamma_\alpha(d)} I_{\alpha\gamma}(U; \tilde{U}). \quad (6.48)$$

Отметим, что $R_\alpha(d)$ зависит от

распределения источника α ,
заданного допустимого уровня искажения d ,
расстояния $d(\cdot, \cdot)$, определяющего меру искажения.

Можно показать, что $R_\alpha(d)$ как функция от d выпукла (вниз) и монотонно не возрастает.

Задача 6.7. Доказать это утверждение.

Обратная теорема кодирования

Важность функции $R_\alpha(d)$, определенной несколько формально, объясняется принадлежащей Шеннону теоремой о кодировании источника. Эта теорема утверждает, что $R_\alpha(d)$ определяет минимальную пропускную способность канала, требуемую для передачи сообщений с уровнем искажения d . Мы посвятим доказательству этой теоремы оставшуюся часть раздела.

Ситуация, рассматриваемая нами, вместе с используемыми обозначениями представлена на рис. 6.5. Примем в качестве величины искажения при блоковом кодировании L выходных символов источника, одновременно предъявляемых для кодирования (см. буфер на рис. 6.5), среднюю по блоку величину искажения d между выходными сообщениями источника и декодирующего устройства:

$$\bar{D} = \frac{1}{L} \sum_{l=1}^L D_l = \frac{1}{L} \sum_{l=1}^L d(U_l, \tilde{U}_l), \quad (6.49)$$

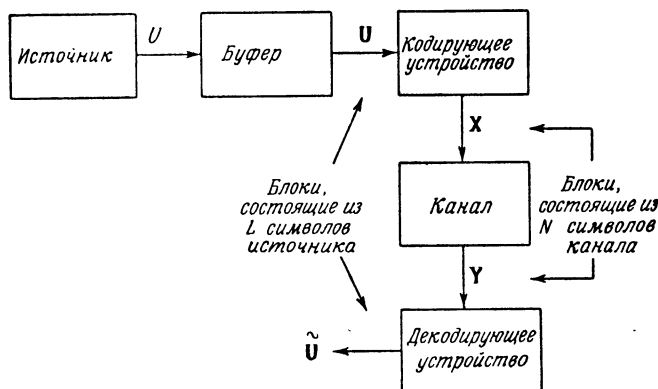
$$d = \mathbf{E}_\alpha\{\bar{D}\} = \frac{1}{L} \sum_{l=1}^L \mathbf{E}_\alpha\{d(U_l, \tilde{U}_l)\}. \quad (6.50)$$

Если U — случайная величина, то \mathbf{U} будет блоком из L экземпляров таких случайных величин; если же U — выборочная функция случайного процесса продолжительностью T с, то \mathbf{U} — выборка случайного процесса длительностью LT с.

Предположим, что последовательные символы U_1, U_2, \dots , поступающие на выход источника сообщений,

независимы и одинаково распределены. Сформулируем и докажем обратную теорему кодирования источника.

Теорема 6.4 (Шеннон). *Рассмотрим блочное кодирование L выходных символов источника с целью передачи по дискретному (во времени) каналу с пропускной способностью C нат¹) на одну передачу. Пусть N — число передач, имеющих в нашем распоряжении за время,*



Р и с. 6.5. Схематическое изображение кодирования источника

в течение которого источник вырабатывает L векторных сообщений. Тогда не существует каскадной цепи, состоящей из кодирующего и декодирующего устройств и канала связи, для которой среднее искажение на сообщение источника

$$E_{\alpha}\{\bar{D}\} = \frac{1}{L} \sum_{i=1}^L E_{\alpha}\{D_i\}$$

¹) Нат — единица информационной пропускной способности, получающаяся при использовании натуральных логарифмов для определения минимальной скорости, совместимой с надежной передачей информации; при использовании двоичных логарифмов ей соответствует один бит.

не превосходит d_0 , где d_0 — наибольшее решение уравнения

$$\frac{N}{L} C = R_\alpha(d_0), \quad (6.51)$$

а C — пропускная способность канала в каскаде.

Заметим, что это негативное утверждение теоремы кодирования выполняется независимо от

- 1) каких-либо предположений о виде кодирующего и декодирующего устройств (другими словами, блоки этих устройств, изображенные на рис. 6.5, могут совершать произвольные операции над величинами, поступающими на их вход);
- 2) длины L блока кодирования;
- 3) вида канала, который может представлять собой как передающее устройство с шумами, так и бесшумное цифровое запоминающее устройство; необходимо лишь, чтобы это устройство характеризовалось своей пропускной способностью.

Доказательство (Галлагер [3, гл. 9]). Для некоторой пары измеримых функций f и g

$$\mathbf{X} = f(\mathbf{U}), \quad \tilde{\mathbf{U}} = g(\mathbf{Y}).$$

Таким образом, в силу свойства 4) из предыдущего раздела

$$I(\mathbf{U}, \mathbf{X}; \mathbf{Y}) = I(\mathbf{U}; \mathbf{Y}) \geq I(\mathbf{U}; \tilde{\mathbf{U}}). \quad (6.52)$$

Так как величины \mathbf{U} , \mathbf{X} и \mathbf{Y} образуют цепь Маркова, то, учитывая формулу Колмогорова и равенство (6.44), имеем

$$I(\mathbf{U}, \mathbf{X}; \mathbf{Y}) = I(\mathbf{X}; \mathbf{Y}) + I(\mathbf{Y}; \mathbf{U} | \mathbf{X}) = I(\mathbf{X}; \mathbf{Y}).$$

Отсюда и из неравенства (6.52) получаем теорему обработки данных для величин \mathbf{U} , \mathbf{X} , \mathbf{Y} , $\tilde{\mathbf{U}}$:

$$I(\mathbf{X}; \mathbf{Y}) \geq I(\mathbf{U}; \tilde{\mathbf{U}}). \quad (6.53)$$

При N независимых использованиях канала имеем

$$NC \geq I(\mathbf{X}; \mathbf{Y}). \quad (6.54)$$

Неравенства (6.53) и (6.54) вместе дают

$$NC \geq I(U, \tilde{U}). \quad (6.55)$$

Вновь используя формулу Колмогорова, находим

$$I(U; \tilde{U}) = \sum_{i=1}^L I(U_i; \tilde{U} | U_1, \dots, U_{i-1}), \quad (6.56)$$

где в силу той же формулы Колмогорова и независимости величин U_i

$$\begin{aligned} I(U_i; \tilde{U} | U_1, \dots, U_{i-1}) &= \\ &= I(U_i; \tilde{U}, U_1, \dots, U_{i-1}) - I(U_i; U_1, \dots, U_{i-1}) = \\ &= I(U_i; \tilde{U}, U_1, \dots, U_{i-1}) \geq I(U_i; \tilde{U}_i). \end{aligned} \quad (6.57)$$

Из соотношений (6.55) — (6.57) следует, что

$$NC \geq \sum_{i=1}^L I(U_i; \tilde{U}_i). \quad (6.58)$$

Полученное неравенство справедливо при любом совместном распределении, в том числе при совместном распределении, задаваемом распределением источника α и произвольным переходным распределением γ . Рассмотрим теперь произвольную каскадную цепь, состоящую из кодирующего и декодирующего устройств и канала связи, для которой средняя величина искажения не превосходит d . Пусть γ_i — переходное распределение величины \tilde{U}_i при заданном значении U_i , индуцируемое рассматриваемым каскадом. Так как среднее искажение не превосходит d , то

$$\begin{aligned} E\{\bar{D}\} &= E\left\{\frac{1}{L} \sum_{i=1}^L d(U_i, \tilde{U}_i)\right\} = \\ &= \frac{1}{L} \sum_{i=1}^L E_{\alpha\gamma_i}\{d(U_i, \tilde{U}_i)\} \leq d. \end{aligned} \quad (6.59)$$

Обозначим через $P_{\tilde{\gamma}}(u | \tilde{u})$ переходное распределение

$$P_{\tilde{\gamma}}(u | \tilde{u}) = \frac{1}{L} \sum_{i=1}^L P_{\gamma_i}(u | \tilde{u}). \quad (6.60)$$

Тогда

$$\begin{aligned} E_{\alpha\tilde{\gamma}}\{D\} &= \frac{1}{L} \sum_{l=1}^L \int_{\mathcal{U} \times \mathcal{U}} d(u, \tilde{u}) dP_{\gamma_l}(u | \tilde{u}) dP_{\alpha}(u) = \\ &= \frac{1}{L} \sum_{l=1}^L E_{\alpha\gamma_l}\{d(U_l, \tilde{U}_l)\} \leq d. \end{aligned} \quad (6.61)$$

Таким образом, если некоторая каскадная цепь, состоящая из кодирующего и декодирующего устройств и канала связи, приводит к среднему искажению, не превосходящему d , то распределение $\tilde{\gamma}$ также должно принадлежать $\Gamma_{\alpha}(d)$. В силу выпуклости (вниз) функции $I_{\alpha\gamma}(U; \tilde{U})$ (задача 6.4)

$$\sum_{l=1}^L I_{\alpha\gamma_l}(U_l; \tilde{U}_l) \geq L I_{\alpha\tilde{\gamma}}(U; \tilde{U}). \quad (6.62)$$

Объединяя неравенства (6.58) и (6.62) и учитывая, что $\tilde{\gamma} \in \Gamma_{\alpha}(d)$, получаем

$$NC \geq L I_{\alpha\tilde{\gamma}}(U; \tilde{U}) \geq L \inf_{\gamma \in \Gamma_{\alpha}(d)} I_{\alpha\gamma}(U; \tilde{U}) = L R_{\alpha}(d). \quad (6.63)$$

Позитивное утверждение теоремы кодирования

Для доказательства позитивного утверждения теоремы кодирования необходимо сделать некоторые предположения о «хвостах» распределения величины U . Достаточно удобным и содержательным предположением такого рода является существование некоторой фиксированной точки в области значений U , обозначим ее 0, для которой

$$E_{\alpha}\{d(U, 0)\} < \infty. \quad (6.64)$$

Негативная часть теоремы кодирования утверждает, что *независимо* от применяемых методов кодирования и декодирования, функция $R_{\alpha}(d)$ определяет минимальную пропускную способность канала, при которой возможен уровень искажения d . При доказательстве позитивной части теоремы мы покажем, что с помощью системы импульсно-кодовой модуляции (ИКМ) можно

добиться передачи, при которой скорость сколь угодно близка к величине, определяемой функцией $R_\alpha(d)$.

С помощью системы ИКМ проблема кодирования решается в два этапа. Если канал является случайным или содержит шумы, то сначала строится код, позволяющий с пренебрежимо малой вероятностью ошибки передавать по каналу цифровую информацию со скоростью, сколь угодно близкой к C . Мы не будем здесь касаться этой проблемы, отсылая читателя к гл. 5—8 книги Галлагера [3]. Таким образом, мы предполагаем, что в нашем распоряжении имеется канал связи, по которому можно надежно передавать цифровую информацию со скоростью, сколь угодно близкой к C .

Очевидно, что вторая задача, возникающая при использовании метода ИКМ, заключается в нахождении способа эффективного представления величины U на выходе источника в цифровом или дискретном виде. Этой задачей мы сейчас и займемся.

Метод ее решения состоит в представлении источника (L, M) -кодом, т. е. множеством из M L -мерных векторов u_1, u_2, \dots, u_M . Кодирование источника осуществляется выбором M кодовых слов (L -мерных векторов) и сопоставлением с помощью некоторого устройства вектора u на выходе источника с кодовым словом u_m , для которого величина

$$d(u, u_m) = \frac{1}{L} \sum_{l=1}^L d(u_l, u_l^m), \quad m = 1, 2, \dots, M,$$

минимальна. Отметим во избежание недоразумений, что мы здесь воспользовались одной и той же буквой для обозначения кодового слова, или L -вектора, (u) и отдельного выходного символа (u) . Метод декодирования состоит в сопоставлении индексу m величины u_m , которая и принимается за представление выхода источника.

Однако при доказательстве позитивной части теоремы кодирования мы не будем для достижения эффективного представления строить фиксированный метод кодирования, соответствующий определенному выбору множества из M L -векторов. Вместо этого мы рассмотрим некоторую совокупность кодов и покажем, что

усредненное по этой совокупности качество кодирования можно сделать сколь угодно близким к величине, определяемой функцией $R_\alpha(d)$. Отсюда будет следовать, что найдется по крайней мере один код, качество которого также будет оптимальным.

Для определения совокупности кодов введем распределение в пространстве L -векторов. Положим

$$P_\alpha(u) = \prod_{l=1}^L P_\alpha(u_l), \quad (6.65)$$

$$P_\gamma(\tilde{u} | u) = \prod_{l=1}^L P_\gamma(\tilde{u}_l | u_l), \quad (6.66)$$

$$P_{\alpha\gamma}(u, \tilde{u}) = P_\gamma(\tilde{u} | u) P_\alpha(u), \quad (6.67)$$

$$P_{\alpha\gamma}(\tilde{u}) = \int_{\mathcal{U}} P_\gamma(\tilde{u} | u) dP_\alpha(u), \quad (6.68)$$

$$P_{\alpha\gamma}(\tilde{u}) = \int_{\mathcal{U} \times \dots \times \mathcal{U}} P_\gamma(\tilde{u} | u) dP_\alpha(u) = \prod_{l=1}^L P_{\alpha\gamma}(\tilde{u}_l). \quad (6.69)$$

Рассмотрим множество кодов, т. е. наборов, состоящих из M кодовых слов, или L -векторов, порождаемое независимым выбором каждого кодового вектора согласно распределению $P_{\alpha\gamma}(\tilde{u})$. Тогда распределение случайной величины

$$\bar{D} = d(U, u_m) = \frac{1}{L} \sum_{l=1}^L d(U_l, u_l^m), \quad (6.70)$$

где $m \in \{1, \dots, M\}$ — индекс, минимизирующий $d(U, u_h)$, зависит как от распределения источника α , так и от распределения $P_{\alpha\gamma}(\tilde{u})$, порождающего рассматриваемое множество кодов. Заметим, что распределение именно этой величины (обозначим его P_c) определяет величину ошибки кодирования.

Найдем теперь связь между распределением вероятностей P_c и вероятностной мерой $P_{\alpha\gamma}$, согласно которой определяется наше множество кодов,

Лемма 6.1 (Галлагер [3, лемма 9.3.1]). Пусть R^* и d^* — произвольные положительные числа, а α_γ — произвольное распределение. Зададим множество

$$A = \{\mathbf{u}, \tilde{\mathbf{u}}: I_{\alpha_\gamma}(\mathbf{u}, \tilde{\mathbf{u}}) > LR^* \text{ или } d(\mathbf{u}, \tilde{\mathbf{u}}) > d^*\}. \quad (6.71)$$

Тогда

$$\mathbf{P}_c\{\bar{D} > d^*\} \leq \mathbf{P}_{\alpha_\gamma}(A) + \exp\{-Me^{-LR^*}\}. \quad (6.72)$$

Доказательство. Мы следуем доказательству этой леммы, принадлежащему Галлагеру, обобщая лишь его с дискретного случая на произвольное измеримое пространство. Для произвольного вектора \mathbf{u} определим $A_{\mathbf{u}}$ как множество векторов $\tilde{\mathbf{u}}$, для которых $\mathbf{u}, \tilde{\mathbf{u}} \in A$:

$$A_{\mathbf{u}} = \{\tilde{\mathbf{u}}: I_{\alpha_\gamma}(\mathbf{u}, \tilde{\mathbf{u}}) > LR^* \text{ или } d(\mathbf{u}, \tilde{\mathbf{u}}) > d^*\} \quad (6.73)$$

(заметим, что множество $A_{\mathbf{u}}$ измеримо). Рассмотрим условную вероятность¹⁾

$$\begin{aligned} \mathbf{P}_c\{\bar{D} > d^* | \mathbf{u}\} &= \mathbf{P}\{d(\mathbf{u}, \mathbf{u}_m) > d^*, m = 1, 2, \dots, M\} = \\ &= \prod_{m=1}^M \mathbf{P}\{d(\mathbf{u}, \mathbf{u}_m) > d^*\} = \left[1 - \int_{A_{\mathbf{u}}^c} d\mathbf{P}_{\alpha_\gamma}(\tilde{\mathbf{u}})\right]^M. \end{aligned} \quad (6.74)$$

Очевидно, что для любых $\mathbf{u}, \tilde{\mathbf{u}} \in A_{\mathbf{u}}^c$

$$\Lambda_{\alpha_\gamma}(\mathbf{u}, \tilde{\mathbf{u}}) \leq e^{LR^*}, \quad (6.75)$$

так что из неравенства (6.74) следует, что

$$\mathbf{P}_c\{\bar{D} > d^* | \mathbf{u}\} \leq \left[1 - e^{-LR^*} \int_{A_{\mathbf{u}}^c} \Lambda_{\alpha_\gamma}(\mathbf{u}, \tilde{\mathbf{u}}) d\mathbf{P}_{\alpha_\gamma}(\tilde{\mathbf{u}})\right]^M. \quad (6.76)$$

Воспользуемся неравенством

$$(1 - \beta x)^M \leq 1 - x + e^{-M\beta}$$

¹⁾ Здесь и ниже через A^c обозначено дополнение к множеству A . — Прим. перев.

(см. Галлагер [3, (9.3.22), (9.3.23)]). Обозначив e^{-LR^*} через β , а через x — интеграл в (6.76), из последнего неравенства получим

$$P_c\{\bar{D} > d^* | u\} \leq \int_{A_u} \Lambda_{\alpha\gamma}(u, \tilde{u}) dP_{\alpha\gamma}(\tilde{u}) + \exp\{-Me^{-LR^*}\}.$$

Интегрируя обе части этого неравенства по мере $P_\alpha(u)$ на множестве \mathcal{U} , находим

$$\begin{aligned} P_c\{\bar{D} > d^*\} &\leq \int_A dP_{\alpha\gamma}(u, \tilde{u}) + \exp\{-Me^{-LR^*}\} = \\ &= P_{\alpha\gamma}\{A\} + \exp\{-Me^{-LR^*}\}. \end{aligned} \quad (6.72)$$

Покажем теперь, что величину U на выходе источника можно представить с точностью, произвольно мало отличающейся от d , так, что скорость создания цифровой информации будет сколь угодно близка к $R(d)$. Приводимая ниже теорема, а также ее доказательство воспроизводятся из книги Галлагера [3].

Теорема 6.5 (Галлагер [3, теорема 9.6.2]). Пусть $R_\alpha(d)$ — скорость при заданном искажении, определяемая распределением источника α . Тогда для любых $d > 0$ и $\delta > 0$ существуют такое (достаточно большое) число L и такой $(M+1, L)$ -код, что

$$M \leq \exp\{L[R_\alpha(d) + \delta]\} \quad (6.77)$$

и

$$E_\alpha\{\bar{D}\} \leq d + \delta. \quad (6.78)$$

Доказательство (Галлагер). Пусть γ_0 — распределение, принадлежащее $\Gamma_\alpha(d)$ и такое, что

$$I_{\alpha\gamma_0}(U; \tilde{U}) \leq R_\alpha(d) + \frac{\delta}{4}. \quad (6.79)$$

Рассмотрим множество (L, M) -кодов, порождаемое распределением $\mathbf{P}_{\alpha\gamma_0}(\tilde{u})$, и применим доказанную лемму, полагая

$$\begin{aligned}d^* &= d + \frac{\delta}{2}, \\R^* &= R_\alpha(d) + \frac{\delta}{2}, \\M &= \exp\left\{L\left[R_\alpha(d) + \frac{3}{4}\delta\right]\right\}.\end{aligned}$$

Тогда, согласно лемме,

$$\mathbf{P}_c\left\{\bar{D} > d + \frac{\delta}{2}\right\} \leq \mathbf{P}_{\alpha\gamma_0}(A) + \exp\{-e^{L\delta/4}\}, \quad (6.80)$$

где

$$\begin{aligned}\mathbf{P}_{\alpha\gamma_0}(A) &= \mathbf{P}_{\alpha\gamma_0}\left\{I_{\alpha\gamma_0}(\mathbf{u}; \tilde{\mathbf{u}}) > L\left[R_\alpha(d) + \frac{\delta}{2}\right]\right. \\&\quad \text{или } \bar{D} > \left(d + \frac{\delta}{2}\right)\bigg\} \leq \\&\leq \mathbf{P}_{\alpha\gamma_0}\left\{\frac{1}{L}I_{\alpha\gamma_0}(\mathbf{u}; \tilde{\mathbf{u}}) > \left[I_{\alpha\gamma_0}(\mathbf{U}, \tilde{\mathbf{U}}) + \frac{\delta}{4}\right]\right\} + \\&\quad + \mathbf{P}_{\alpha\gamma_0}\left\{\bar{D} > \left(d + \frac{\delta}{2}\right)\right\}. \quad (6.81)\end{aligned}$$

Используя независимость пар (U_l, \tilde{U}_l) , $l = 1, \dots, L$, при совместном распределении, описываемом соотношениями (6.65) — (6.69), получаем

$$\frac{1}{L}I_{\alpha\gamma_0}(\mathbf{u}; \tilde{\mathbf{u}}) = \frac{1}{L}\sum_{l=1}^L I_{\alpha\gamma_0}(u_l, \tilde{u}_l) \quad (6.82)$$

и

$$\bar{D} = \frac{1}{L}\sum_{l=1}^L D_l = \frac{1}{L}\sum_{l=1}^L d(U_l, \tilde{U}_l), \quad (6.83)$$

причем правые части этих равенств являются суммами независимых одинаково распределенных случайных величин. Кроме того, в силу задачи 6.2

$$\mathbf{E}_{\alpha\gamma_0}\{I_{\alpha\gamma_0}\} \leq \frac{2}{e} + I_{\alpha\gamma_0}(U; \tilde{U}) < \infty,$$

$$\mathbf{E}_{\alpha\gamma_0}\{D_l\} = \mathbf{E}_{\alpha\gamma_0}\{D_l\} \leq d < \infty.$$

Таким образом, применяя слабый закон больших чисел [10], убеждаемся, что величину $\mathbf{P}_{\alpha\gamma_0}\{A\}$ можно сделать сколь угодно малой за счет выбора достаточно большого L . Учитывая неравенства (6.80) и (6.81), находим

$$\mathbf{P}_c\left\{\bar{D} > d + \frac{\delta}{2}\right\} \leq \beta(L), \quad (6.84)$$

где $\beta(L)$ стремится к нулю при $L \rightarrow \infty$.

Замечание. Мы уже на полпути к завершению доказательства теоремы. Действительно, пусть B — событие, состоящее в том, что $\bar{D} > d + \delta/2$. Тогда, как мы показали, вероятность $\mathbf{P}(B)$ можно сделать сколь угодно малой. Однако это не означает еще, что $\mathbf{E}\{\bar{D}\} < d + \delta$, поскольку влияние «хвостов» распределения может оказаться существенным. Для завершения доказательства осталось учесть влияние этих «хвостов».

Из неравенства (6.84) вытекает, что

$$\mathbf{E}_{\alpha\gamma_0}\{\bar{D}\} \leq \left(d + \frac{\delta}{2}\right) + \mathbf{P}_{\alpha\gamma_0}(B) \mathbf{E}_{\alpha\gamma_0}\{\bar{D} | B\}. \quad (6.85)$$

Эта оценка справедлива для кода, состоящего из M случайно выбранных L -векторов. Пополним множество слов этого кода, добавив к нему нулевой вектор. Если теперь при осуществлении события B величине на выходе источника поставить в соответствие добавленный нулевой вектор, то получим

$$\mathbf{P}_{\alpha\gamma_0}(B) \mathbf{E}_{\alpha\gamma_0}\{\bar{D} | B\} = \frac{1}{L} \sum_{l=1}^L \mathbf{P}_{\alpha\gamma_0}(B) \mathbf{E}_{\alpha\gamma_0}\{d(U_l, 0) | B\}. \quad (6.86)$$

Рассмотрим произвольную положительную случайную величину V и произвольное множество B ; выберем d_0 так, чтобы

$$\{V > d_0\} \supset \{B\}.$$

Тогда

$$\mathbf{P}(B) \mathbf{E}\{V | B\} = \int_B v d\mathbf{P}(v) \leq \int_{v \geq d_0} v d\mathbf{P}(v). \quad (6.87)$$

Объединяя неравенства (6.86) и (6.87), находим, что при $V = d(U, 0)$

$$\mathbf{P}_{\alpha\gamma_0}(B) \mathbf{E}_{\alpha\gamma_0}\{\bar{D} | B\} \leq \int_{v \geq d_0} v d\mathbf{P}(v). \quad (6.88)$$

Так как $\mathbf{E}\{V\} \leq K_0 < \infty$, то

$$\lim_{d_0 \rightarrow \infty} \int_{v \geq d_0} v d\mathbf{P}(v) = 0. \quad (6.89)$$

Следовательно, правую часть неравенства (6.88) можно сделать меньше $\delta/2$, если d_0 взять достаточно большим. Выполнение этого условия можно обеспечить, положив L достаточно большим, поскольку $\beta(L) \rightarrow 0$ при $L \rightarrow \infty$. Это замечание вместе с неравенством (6.85) завершает доказательство теоремы.

Задача 6.8. Предположим, что мы определили взаимную информацию по-иному, нежели в (6.26) или (6.27), используя вместо $\ln(\Lambda)$ другую функцию $g(\Lambda)$. Если бы при этом новом определении оба утверждения (и позитивное, и негативное) теоремы кодирования выполнялись с иной функцией $R_\alpha(d)$, то мы получили бы две противоречащие друг другу теоремы. Проследить доказательство обоих утверждений теоремы кодирования и указать, где в этом доказательстве оказался существенным однозначный выбор функции $g(\cdot) = \ln(\cdot)$.

6.4. СКОРОСТЬ ПРИ ЗАДАННОМ ИСКАЖЕНИИ ДЛЯ ГАУССОВСКОГО СЛУЧАЙНОГО ПРОЦЕССА СО ВЗВЕШЕННОЙ КВАДРАТИЧЕСКОЙ МЕРОЙ ИСКАЖЕНИЯ

Вид скорости при заданном искажении известен лишь для немногих источников сообщений и ограниченного числа функций $d(\cdot, \cdot)$. В этом разделе мы рассмотрим источник, представляющий наибольший интерес в приложениях теории связи, а именно источник, выходом которого является гауссовский случайный процесс.

Другими словами, мы предполагаем, что U — выборочная функция гауссовского случайного процесса $U(t)$ продолжительностью T с, $0 \leq t \leq T$. Будем также считать, что процесс $U(t)$ стационарен, имеет нулевое ма-

тематическое ожидание и известную ковариационную функцию $R_u(\tau) = \mathbf{E}\{U_{t+\tau}U_t\}$.

Определим меру искажения, которую мы будем в дальнейшем рассматривать. Обозначим через A однородную линейную систему, действующую в $L_2[0, T]$, с импульсной переходной функцией $a(t)$ и частотной характеристикой $A(f)$. Предположим, что $a(t) \neq 0$ лишь на некотором конечном отрезке $[0, T_a]$. Определим взвешенную ошибку между выборочной функцией $U(t)$ на выходе источника сообщений и ее аппроксимацией $\tilde{U}(t)$:

$$\mathbf{E}(t) = \int_0^T a(t-s)[U(s) - \tilde{U}(s)]ds, \quad 0 < t \leq T + T_a. \quad (6.90)$$

Поскольку на отрезке $T_a \leq t \leq T$ невозможны ложные переходные эффекты, определим среднее расхождение как

$$d = \mathbf{E} \left\{ \frac{1}{T'} \int_{T_a}^T |\mathbf{E}(t)|^2 dt \right\}, \quad (6.91)$$

где $T' = T - T_a$.

Весовая функция $A(f)$ была введена для того, чтобы приспособить меру искажения к учету субъективных оценок ошибки сигнала $\tilde{U}(t)$. Например, если $u(t)$ — сигнал, воспроизводящий музыкальное звучание, то естественно потребовать, чтобы величина $|A(f)|^2$ была, грубо говоря, обратно пропорциональна $S_u(f)$, т. е. чтобы высокие звуки, имеющие обычно слабую мощность, звучали более отчетливо. С другой стороны, вне диапазона слышимости $|A(f)|^2$ может быстро падать до нуля.

Если же $u(t)$ — развертка телевизионного изображения, то, вероятно, следует предположить, что выбор $a(t)$ должен способствовать аппроксимации как самого импульса, так и его производной, чтобы, во-первых, улавливать изменение интенсивности и, во-вторых, сохранить четкость изображения предметов. Функция $|A(f)|^2$ может, как и раньше, быстро падать до нуля вне области разрешения применяемой оптической системы.

Опишем подход к нахождению $R(d)$ в рассматриваемом случае. Он состоит в таком представлении процесса

$U(t)$, $0 \leq t \leq T$, счетным набором независимых гауссовских случайных величин, чтобы среднюю взвешенную ошибку можно было выразить через эти величины в «диагональном» виде.

Для этого определим линейные интегральные операторы A и A^* равенствами

$$[Au](t) = \begin{cases} \int_0^T a(t-s)u(s)ds, & T_a \leq t \leq T, \\ 0 & \text{в остальных случаях,} \end{cases} \quad (6.92)$$

$$[A^*u](s) = \begin{cases} \int_{T_a}^T a(t-s)u(t)dt, & 0 \leq s \leq T, \\ 0 & \text{в остальных случаях.} \end{cases} \quad (6.93)$$

Зададим случайный процесс $V(t)$, $T_a \leq t \leq T$, положив

$$V(t) = [AU](t). \quad (6.94)$$

В введенных операторных обозначениях расхождение (6.91) принимает вид

$$d = \mathbf{E} \left\{ \frac{1}{T'} \|A(U - \tilde{U})\|_{[T_a, T]}^2 \right\}. \quad (6.91)$$

Обозначим через R_v оператор $A \circ R_u \circ A^*$, т. е. интегральный оператор в $L_2[T_a, T]$ с ядром

$$\begin{aligned} R_v(t_1 - t_2) &= \int_0^T ds_1 a(t_1 - s_1) \int_0^T ds_2 a(t_2 - s_2) R_u(s_1 - s_2) = \\ &= \mathbf{E} \{V_{t_1} V_{t_2}\}, \quad T_a \leq t_1, t_2 \leq T. \end{aligned} \quad (6.95)$$

Поскольку процесс $U(t)$ стационарен в квадрате $T_a \leq t_1, t_2 \leq T$, функция R_v зависит лишь от разности $t_1 - t_2$.

Рассмотрим базис разложения Карунена — Лоэва для $V(t)$ на $[T_a, T]$, т. е. введем систему собственных функций и собственных значений уравнения

$$\lambda_k \Phi_k(t_1) = \int_{T_a}^T dt_2 R_v(t_1 - t_2) \Phi_k(t_2), \quad T_a \leq t_1 \leq T. \quad (6.96)$$

С помощью $\{\varphi_k\}$ построим также систему функций для разложения $U(t)$. Рассмотрим оператор $A^* \circ A$, действующий из $L_2[0, T]$ в $L_2[0, T]$. Этот оператор симметричен, и на множестве его значений можно определить оператор, обратный к нему. Имеем

$$A^* \varphi_k = \frac{1}{\lambda_k} A^* R_v \varphi_k = \frac{1}{\lambda_k} A^* A R_u A^* \varphi_k.$$

Таким образом, функция $A^* \varphi_k$ принадлежит множеству значений оператора $A^* \circ A$, и, следовательно, функции

$$\theta_k(s) = [A^* \circ A]^{-1} A^* \varphi_k = \frac{1}{\lambda_k} R_u A^* \varphi_k, \quad (6.97)$$

$$k = 1, 2, \dots, \quad s \in [0, T],$$

принадлежат $L_2[0, T]$. Заметим, что

$$[A \theta_k](t) = \frac{1}{\lambda_k} [A R_u A^* \varphi_k](t) = \varphi_k(t), \quad t \in [T_0, T]. \quad (6.98)$$

Возьмем теперь произвольную функцию u в области определения оператора A и положим

$$v(t) = [A u](t), \quad T_a \leq t \leq T,$$

$$v_N(t) = \sum_{k=1}^N v_k \varphi_k(t),$$

$$v_k = \{v, \varphi_k\},$$

$$u_N(s) = \sum_{k=1}^N v_k \theta_k(s), \quad 0 \leq s \leq T.$$

В силу (6.98)

$$[A u_N](t) = v_N(t),$$

а в силу полноты системы $\{\varphi_k\}$ в $L_2[T_a, T]$

$$0 = \lim_{N \rightarrow \infty} \|v - v_N\|_{[T_a, T]}^2 = \lim_{N \rightarrow \infty} \|A(u - u_N)\|_{[T_a, T]}^2. \quad (6.99)$$

Итак, при взвешенном квадратическом критерии ошибок любую функцию в области определения оператора A можно представить с помощью набора $\{\theta_k\}$. Более того, если оператор A ограничен снизу (по норме) числом k , то в силу (6.99) u_N сходится к u в смысле среднеквадра-

тической интегральной ошибки, и в этом случае функции θ_k порождают область определения оператора A .

Рассуждая подобным образом, можно показать, что процесс на выходе источника представим в виде

$$\lim_{N \rightarrow \infty} U_N(s) = \lim_{N \rightarrow \infty} \sum_{k=1}^N V_k \theta_k(s), \quad 0 \leq s \leq T, \quad (6.100)$$

где коэффициенты V_k определяются равенством

$$V_k = (V, \varphi_k) = \int_{T_a}^T dt V(t) \varphi_k(t).$$

Полагая

$$V_N(t) = \sum_{k=1}^N V_k \varphi_k(t), \quad T_a \leq t \leq T,$$

в силу разложения Карунена — Лозва получаем

$$V(t) = \text{l.i.m. } V_N(t)^{1)}$$

равномерно для всех $t \in [T_a, T]$. Таким образом,

$$\begin{aligned} 0 &= \lim_{N \rightarrow \infty} \mathbf{E} \{ \| V - V_N \|^2_{[T_a, T]} \} = \\ &= \lim_{N \rightarrow \infty} \mathbf{E} \{ \| A(U - U_N) \|^2_{[T_a, T]} \}. \end{aligned} \quad (6.101)$$

Если оператор A ограничен снизу, то из (6.101) следует, что $U_N(s)$ сходится к $U(s)$ и в смысле среднеквадратической интегральной ошибки.

Выразим величину искажения через координаты V_k . Для этого представим процесс источника и закодированный процесс в виде

$$U(s) = \sum_{k=1}^{\infty} V_k \theta_k(s), \quad \tilde{U}(s) = \sum_{k=1}^{\infty} \tilde{V}_k \theta_k(s), \quad 0 \leq s \leq T.$$

¹⁾ Общепринятое обозначение для сходимости в среднеквадратическом. — *Прим. перев.*

Согласно равенству Парсеваля,

$$\begin{aligned} d &= \frac{1}{T'} \mathbf{E} \{ \| A(U - \tilde{U}) \|_{[T_a, T]}^2 \} = \\ &= \frac{1}{T'} \mathbf{E} \{ \| V - \tilde{V} \|_{[T_a, T]}^2 \} = \\ &= \frac{1}{T'} \sum_{k=1}^{\infty} \mathbf{E} \{ (V_k - \tilde{V}_k)^2 \}. \end{aligned} \quad (6.102)$$

Заметим, что величины V_k независимы, так как они являются коэффициентами разложения Карунена — Лозва процесса $V(t)$. Кроме того, они являются главными осями для функции искажения, как это следует из последнего равенства, поскольку в (6.102) отсутствуют перекрестные члены.

В разд. 6.2 мы видели, что скорость при заданном искажении для гауссовской случайной величины U со средним 0 и дисперсией λ определяется равенством

$$R(d) = \frac{1}{2} \log_2 \frac{\lambda}{d} \quad (\text{бит на случайную величину}). \quad (6.103)$$

Задача 6.9. Вывести (6.103) из определения $R(d)$, данного в разд. 6.3, и найти распределение γ , приводящее к этой функции $R(d)$, а также соответствующее распределение, индуцируемое на \tilde{U} . [Указание. Рассмотрите плотность $f(u)$ и условную плотность $f(u|\tilde{u})$ распределения U и величины

$$\begin{aligned} H(U) &= - \int_{-\infty}^{\infty} du f(u) \log f(u), \\ H(U|\tilde{U}) &= - \int_{-\infty}^{\infty} d\tilde{u} f(\tilde{u}) \int_{-\infty}^{\infty} du f(u|\tilde{u}) \log f(u|\tilde{u}). \end{aligned}$$

Имеем

$$I(U; \tilde{U}) = H(U) - H(U|\tilde{U}),$$

откуда

$$\inf_{\gamma \in \Gamma_{\alpha}(d)} I_{\alpha\gamma}(U; \tilde{U}) = H(U) - \sup_{\gamma \in \Gamma_{\alpha}^*(d)} H(U|\tilde{U}).$$

Далее,

$$H(U | \tilde{U}) = H(U - \tilde{U} | \tilde{U}) \leq H(U - \tilde{U}),$$

причем равенство возможно лишь в том случае, когда величины $U - \tilde{U}$ и \tilde{U} независимы.]

Найдем теперь функцию скорости при заданном искажении для $U(t)$, $0 \leq t \leq T$. Поскольку величины V_k независимы, то, применяя (6.56) и (6.57) к последовательности $\{V_k\}$, получаем

$$I(\mathbf{V}; \tilde{\mathbf{V}}) \geq \sum_{k=1}^{\infty} I(V_k; \tilde{V}_k), \quad (6.104)$$

причем равенство выполняется, если величины \tilde{V}_k независимы. Таким образом, используя (6.102) и (6.103), находим

$$I(\mathbf{V}; \tilde{\mathbf{V}}) \geq \frac{1}{2} \sum_{k=1}^{\infty} \log \frac{\lambda_k}{d_k}, \quad (6.105)$$

причем равенство выполняется, если $V_k - \tilde{V}_k$ являются независимыми гауссовскими случайными величинами с дисперсиями d_k . Заметим, что $d_k \leq \lambda_k$, поскольку V_k можно воспроизвести со среднеквадратической ошибкой λ_k , полагая $\tilde{V}_k \equiv 0$ при нулевой скорости передачи. Итак, $R_\alpha(d)$ — точная нижняя грань правой части неравенства (6.105) при ограничениях

$$0 \leq d_k \leq \lambda_k, \quad d = \frac{1}{T'} \sum_{k=1}^{\infty} d_k. \quad (6.106)$$

Задача 6.10. Показать, что $R_\alpha(d)$ определяется параметрически уравнениями

$$dT'(\mu) = \frac{1}{T'} \left[\mu N(\mu) + \sum_{k: \lambda_k \leq \mu} \lambda_k \right], \quad (6.107)$$

$$R_\alpha, T'(\mu) = \frac{1}{2T'} \sum_{k: \lambda_k > \mu} \log_2 \frac{\lambda_k}{\mu} \quad \text{бит/с}, \quad (6.108)$$

где $N(\mu)$ — число тех k , для которых $\lambda_k > \mu$.

Мы ввели здесь индекс T' для указания того, что полученная скорость относится к выборочной функции $U(t)$ длительностью T с. [Указание. Воспользуйтесь неравенством

$$\log \frac{1}{n} \left(\sum_{k=1}^n a_k \right) \geq \frac{1}{n} \sum_{k=1}^n \log(a_k),$$

причем равенство выполняется тогда и только тогда, когда все a_k равны между собой.]

Уравнения (6.107) и (6.108) определяют в параметрическом виде скорость при заданном искажении для выборки $U(t)$ длительностью T с. Посмотрим, как меняется $R_{a,T'}(d)$ в зависимости от $T = T' + T_a$.

Задача 6.11. Показать, что для любого $T' > 0$ и целого N

$$R_{a,NT'}(d) \leq R_{a,T'}(d). \quad (6.109)$$

[Указание. Рассмотрите кодирование выборочных функций процесса $U(t)$ длительностью $NT' + T_a$ и выборки продолжительностью $T' + T_a = T$.]

Неравенство (6.109) показывает, что $R_{a,T'}$ монотонно не возрастает по T' , когда T' , возрастая, пробегает множество точек, кратных любому фиксированному значению. Естественнo предположить, что

$$\lim_{T' \rightarrow \infty} R_{a,T'}(d) = R_a(d) \quad (6.110)$$

есть $\inf_T R_{a,T'}(d)$. Это действительно так; интересующийся читатель может найти доказательство этого утверждения в книге Галлагера [3, дополнение 4А, лемма 2].

Итак, предел (6.110) представляет собой скорость, достижимую при кодировании длинных сообщений $U(t)$. Для ее оценки нам понадобится следующая теорема Каца, Мурдока и Сегё.

Теорема 6.6. Пусть $\lambda_{k,T'}$ — собственные значения интегрального оператора с ядром $R_b(t_1 - t_2)$ на $[0, T']$.

Предположим, что функция $R_v(\cdot)$ ограничена, абсолютно интегрируема на $(-\infty, \infty)$ и допускает действительное преобразование Фурье

$$S_v(f) = \int_{-\infty}^{\infty} d\tau e^{-i2\pi f\tau} R_v(\tau).$$

Тогда для любого целого $p \geq 1$

$$\lim_{T' \rightarrow \infty} \frac{1}{T'} \sum_{k=1}^{\infty} (\lambda_{k,T'})^p = \int_{-\infty}^{\infty} [S_v(f)]^p df. \quad (6.111)$$

Доказательство этой теоремы приведено в книге Гренандера и Серге [11, разд. 8.6]. Мы воспользуемся следующим вытекающим из нее результатом.

Следствие 6.1. Пусть $[a, b]$ — произвольный отрезок, не содержащий нуля, $g(\lambda)$ — непрерывная функция на $[a, b]$. Предположим, что мера Лебега множества $\{f: S_v(f) = a \text{ или } S_v(f) = b\}$ равна 0. Тогда

$$\lim_{T' \rightarrow \infty} \frac{1}{T'} \sum_{a \leq \lambda_k < b} g(\lambda_{k,T'}) = \int_{a \leq S_v(f) < b} g[S_v(f)] df. \quad (6.112)$$

Наметим достаточно простое доказательство этого утверждения.

Пусть $\chi_{[a, b]}(\lambda)$ — характеристическая функция множества $[a, b]$, т. е.

$$\chi_{[a, b]}(\lambda) = \begin{cases} 1, & a \leq \lambda \leq b, \\ 0 & \text{в остальных случаях.} \end{cases}$$

Тогда функция $\frac{g(\lambda)}{\lambda} \chi_{[a, b]}(\lambda)$ непрерывна всюду, кроме точек a и b . Далее, при наших предположениях о R_v

$$-\infty < \lambda_{\min} \leq S_v(f) < \lambda_{\max} < \infty.$$

Построим на отрезке $[\lambda_{\min}, \lambda_{\max}]$ два полинома P_1 и P_2 , ограничивающих функцию $g(\lambda) \chi_{[a, b]}(\lambda)$ сверху и снизу и произвольно близких друг к другу всюду, кроме окрест-

ностей точек a и b . Применяя (6.111) к каждому слагаемому построенных полиномов, находим

$$\begin{aligned} \int_{-\infty}^{\infty} P_1[S_v(f)] df &\leq \lim_{T' \rightarrow \infty} \frac{1}{T'} \sum_k g(\lambda_{k,T'}) \chi_{[a, b]}(\lambda_{k,T'}) \leq \\ &\leq \int_{-\infty}^{\infty} P_2[S_v(f)] df. \end{aligned}$$

Поскольку по условию $S_v(f) = a$ и $S_v(f) = b$ лишь на множестве меры 0, эти интегралы можно сделать сколь угодно близкими друг к другу. Отсюда следует (6.112).

Применяя (6.112) к параметрическим уравнениям для нахождения $R_{\alpha, T'}(d)$, мы получим пару параметрических уравнений, определяющих предельное значение скорости при заданном искажении:

$$d(\mu) = \mu \int_{S_v(f) > \mu} df + \int_{S_v(f) \leq \mu} S_v(f) df, \quad (6.113)$$

$$R_{\alpha}(\mu) = \frac{1}{2} \int_{S_v(f) > \mu} \log_2 \frac{S_v(f)}{\mu} df \text{ (бит/с)}. \quad (6.114)$$

Здесь

$$S_v(f) = S_u(f) |A(f)|^2. \quad (6.115)$$

Заметим, что второе слагаемое в (6.113) возникает при применении следствия 6.1 к сумме

$$\sum_{\lambda_k \leq \mu} \lambda_{k,T'} = R_v(0) - \sum_{\lambda_k > \mu} \lambda_{k,T'}$$

с целью исключения собственных значений, находящихся в окрестности нуля.

Напомним, что справедливость негативного утверждения теоремы кодирования не зависела от длины блока и, следовательно, оно непосредственно применимо к величине $R_{\alpha}(d)$, являющейся пределом для $R_{\alpha, T'}(d)$. Напротив, позитивное утверждение теоремы, связанное с указанием метода кодирования, применимо к ситуации, когда для кодирования предъявляются независимые одинаково распределенные случайные величины. К счастью,

равенство (6.112) показывает, что этого можно добиться. Рассмотрим коэффициенты V_k , порождаемые процессом на отрезке $[0, T]$, когда T растет. Сгруппируем коэффициенты V_k в блоки, относя в n -й блок те из них, для которых

$$\mu + (n-1)\varepsilon \leq \lambda_k < \mu + n\varepsilon,$$

где $\varepsilon \ll \mu$ и $n = 1, 2, \dots, (\lambda_{\max} - \mu)/\varepsilon$. Полагая в (6.112) $g(u) = 1$, $a = \mu + (n-1)\varepsilon$ и $\beta = \mu + n\varepsilon$, получаем, что число коэффициентов в каждом блоке растет линейно по T' . Таким образом можно образовать длинные блоки из независимых гауссовских случайных величин, имеющих примерно одинаковую дисперсию, и эффективно их закодировать¹⁾.

Метод построения разложения Карунена — Лозва для процесса $U(t)$ большой длительности с последующим квантованием коэффициентов с помощью L -векторов весьма сложен, и, по-видимому, его реализация крайне трудна. Гоблик и Холзингер [12] провели интересное сравнение функции $R_\alpha(d)$ со скоростью, достижимой практически более целесообразными субоптимальными методами кодирования гауссовских случайных величин.

Уравнения (6.113) и (6.114), определяющие скорость при заданном искажении для гауссовского случайного процесса, впервые были выведены Колмогоровым [13] для случая $A(f) = 1$ (интегрально-квадратическая ошибка). Для случая взвешенной ошибки эти уравнения впервые вывел Пинскер [14]. Результаты Пинскера были обобщены Добрушиным и Цыбаковым в работе [15]. Аналогичные результаты были независимо получены в работах [16, 17].

Столь же большой интерес, как и гауссовскому процессу, уделялся гауссовскому случайному полю $U(x, y)$, представляющему собой некоторое плоское изображение. Если поле U стационарно со спектральной функцией $S_u(f_x, f_y)$ и мы рассматриваем весовую функцию с дву-

¹⁾ В действительности нет необходимости прибегать к довольно сложной процедуре объединения в блоки случайных величин с примерно равными дисперсиями. Как показано Галлагером [3, разд. 9.7], можно эффективно кодировать блоки и таких *независимых* гауссовских случайных величин, дисперсии которых *различны*.

мерным преобразованием $A(f_x, f_y)$, то уравнения (6.113), (6.114) принимают вид

$$d(\mu) = \mu \int_{S_v(f_x, f_y) > \mu} df_x df_y + \int_{S_v(f_x, f_y) \leq \mu} S_v(f_x, f_y) df_x df_y, \quad (6.116)$$

$$R_a(\mu) = \frac{1}{2} \int_{S_v(f_x, f_y) > \mu} \log_2 \frac{S_v(f_x, f_y)}{\mu} df_x df_y \quad (\text{бит на единицу площади}), \quad (6.117)$$

где

$$S_v(f_x, f_y) = |A(f_x, f_y)|^2 S_u(f_x, f_y). \quad (6.118)$$

Вывод этих соотношений аналогичен выводу в случае временного процесса с учетом разложения Карунена — Лозва для двумерного случайного поля

$$V(x, y) = \sum_{k=1}^{\infty} V_k \Phi_k(x, y),$$

где $E\{V_k^2\} = \lambda_k$ — собственные значения, отвечающие ядру $R_v(x, y)$ в прямоугольнике. Теорема Каца, Мурдока и Сегё применима также и в N -мерном случае [18]; отсюда и вытекают уравнения (6.116), (6.117). В изотропном случае, когда $R_v(x, y)$ зависит лишь от $\sqrt{x^2 + y^2}$, $S_v(f_x, f_y)$ можно представить как функцию лишь от $\sqrt{f_x^2 + f_y^2}$, и соотношения (6.116), (6.117) преобразуются к виду, содержащему обычные интегралы. Подробности см. в работе [19].

Задача 6.12. Рассмотрим два источника сообщений с распределениями α_1 и α_2 . Пусть соответствующими функциями скорости при заданном искажении будут $R_{\alpha_1}(d)$ и $R_{\alpha_2}(d)$. Рассмотрим *общее* кодирующее устройство, с помощью которого сообщения *каждого* источника могут передаваться при уровне искажения, не превосходящем d . Найти нижнюю и верхнюю границы скорости, необходимой для такого кодирующего устройства. Выводы должны быть четкими и обоснованными на каждом шаге; одни лишь правдоподобные рассуждения недостаточны.

6.5. ГРАНИЦЫ ПЕРЕДАЧИ ПО АДДИТИВНОМУ ГАУССОВСКОМУ КАНАЛУ

Вернемся к задаче, рассматривавшейся в гл. 3—5, т. е. к задаче передачи сообщений по аддитивному гауссовскому каналу.

Предположим сначала, что в течение каждых T с мы можем передавать сигнал, имеющий полосу частот W , и произведение полосы на время равно $N = 2WT$. Как и в гл. 4, допустим, что энергия передаваемого сигнала не должна превышать E , а аддитивный шум в канале белый гауссовский, со спектральной плотностью $N_0/2$. Можно показать [1, часть IV], что пропускная способность такого канала равна

$$C = \frac{N}{2} \log_2 \left[1 + \frac{E}{N(N_0/2)} \right] \text{ (бит на сигнал длительностью } T \text{ с)}. \quad (6.119)$$

Если по рассматриваемому каналу передается последовательность независимых гауссовских случайных величин с дисперсией σ^2 при скорости одна случайная величина в T с, то минимальное среднеквадратическое искажение d_0 можно найти, приравняв C в (6.119) к функции $R_\alpha(d)$, определяемой формулой (6.103) (с заменой λ на σ^2):

$$\frac{1}{2} \log_2 \frac{\sigma^2}{d_0} = \frac{N}{2} \log_2 \left[1 + \frac{E}{N(N_0/2)} \right].$$

Таким образом,

$$d_0 = \sigma^2 \left[1 + \frac{E}{N(N_0/2)} \right]^{-N}. \quad (6.120)$$

Если этот результат выразить в терминах отношения выходного сигнала к шуму, то мы получим границу

$$\left(\frac{S}{\mathcal{N}} \right)_{\text{вых}} \leq \frac{\sigma^2}{d_0} = \left[1 + \left(\frac{S}{\mathcal{N}} \right)_{\text{кан}} \frac{1}{N} \right]^N, \quad (6.121)$$

где

$$\left(\frac{S}{\mathcal{N}} \right)_{\text{кан}} = \frac{E}{(N_0/2)} \quad (6.122)$$

есть отношение сигнал/шум на входе канала.

Если бы вместо гауссовских случайных величин мы рассмотрели передачу последовательности независимых случайных величин, равномерно распределенных на $[-u_0/2, u_0/2]$, то для $R_\alpha(d_0)$ мы имели бы границу [1, часть VI]

$$R_\alpha(d) \geq \frac{1}{2} \log_2 \left[\frac{u_0^2}{2\pi e d} \right]. \quad (6.123)$$

Приравнявая $R_\alpha(d_0)$ к пропускной способности в (6.119) и производя преобразования, находим

$$\left(\frac{S}{\mathcal{N}} \right)_{\text{вых}} \leq \frac{\sigma^2}{d_0} = \frac{(u_0^2/12)}{d_0} \leq \frac{\pi e}{6} \left[1 + \left(\frac{S}{\mathcal{N}} \right)_{\text{кан}} \frac{1}{N} \right]^N. \quad (6.124)$$

Заметим, что при значениях $(S/\mathcal{N})_{\text{кан}}$, много больших N , отношение $(S/\mathcal{N})_{\text{вых}}$ в обоих неравенствах (6.12) и (6.124) растет как N -я степень $(S/\mathcal{N})_{\text{кан}}$ в отличие от *линейного* роста $(S/\mathcal{N})_{\text{вых}}$ для любого фиксированного метода модуляции в надпороговой области. Заметим также, что границы (6.121) и (6.124) имеют тот же вид, что и граница плотной упаковки, обсуждавшаяся в разд. 4.3, отличающаяся от нее лишь множителями при $\left[1 + (S/\mathcal{N})_{\text{кан}} \frac{1}{N} \right]^N$. На самом деле если (4.29) и (4.30) объединить с (4.39), то можно показать, что все эти множители весьма близки друг к другу.

Важный результат теории скорости при заданном искажении состоит в том, что она позволяет взглянуть на структуру, а также оценить сложность близких к оптимальным методов передачи. Однако ее основное значение заключается в описании точных границ качества передачи (6.121) и (6.124), позволяющих производить абсолютное сравнение существующих и предлагаемых систем с оптимальными системами, имеющими, быть может, высокую сложность. Такое сравнение было проведено Зейдманом [20] для системы время-импульсной модуляции (ВИМ), использующей сигнал $(\sin \omega t)/\tau$ с произведением полосы на время, равным $N = 40$. Полученные результаты применимы также и к задаче радиолокационного измерения дальности при том же виде

сигналов. Эти результаты представлены на рис. 6.6, где приводятся

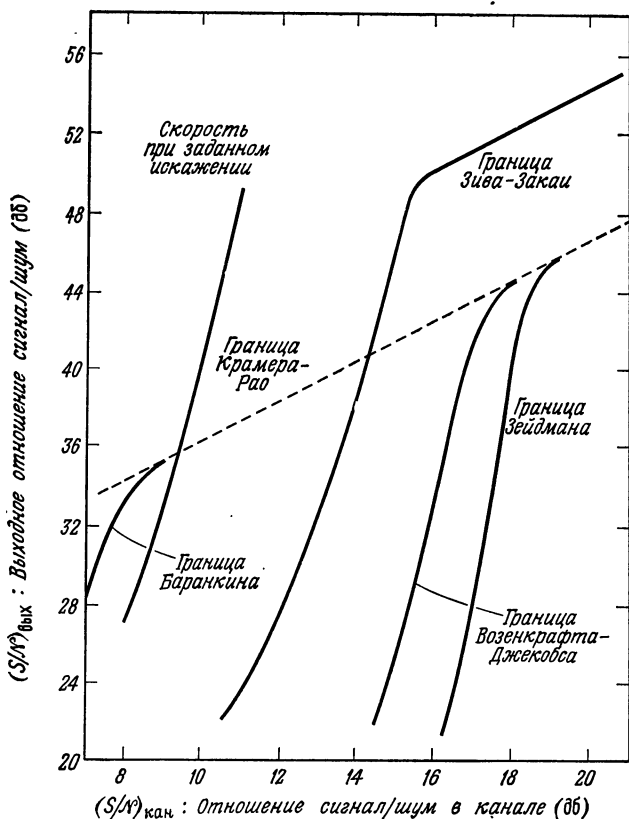


Рис. 6.6. Сравнение кривой скорости при заданном искажении (пропускной способности канала) с качеством время-импульсной модуляции.

- 1) граница скорости при заданном искажении (граница пропускной способности канала);
- 2) верхние границы Баранкина, Крамера — Рао и Зива — Закаи;
- 3) нижняя граница Зейдмана [21];

- 4) приближенная граница Возенкрафта — Джекобса, упомянутая в гл. 4.

Кривая, характеризующая качество произвольной системы с $N = 40$, должна проходить ниже кривой скорости при заданном искажении. Любая же кривая, относящаяся к время-импульсной модуляции, проходит выше границы Зейдмана и ниже всех трех верхних границ.

На рисунке видно, что ниже порога наиболее точной является граница Зива — Закаи, а выше порога — граница Крамера — Рао. Ниже порога граница Баранкина хуже границы Зива — Закаи и даже границы скорости при заданном искажении. По-видимому, есть основание опираться и на приближенный анализ Возенкрафта — Джекобса, который, согласно их сообщению, подтверждается проведенными измерениями.

Сравнение кривой Возенкрафта — Джекобса с кривой скорости при заданном искажении показывает, что ниже порога для системы ВИМ требуется примерно на 7 дБ больше мощности передачи, чем ее затрачивает оптимальная система с тем же отношением сигнал/шум. Однако выше порога это различие быстро возрастает. Таким образом, выше порога система ВИМ, по-видимому, для эффективной связи непригодна.

Границу скорости при заданном искажении (пропускной способности) можно также использовать для определения точности, с которой можно передать гауссовский случайный процесс по аддитивному гауссовскому каналу. Пропускная способность гауссовского канала с шириной полосы частот W_c , мощностным ограничением P_c и спектральной плотностью аддитивного шума $N_0/2$ равна [1, часть IV]

$$C = W_c \log_2 \left[1 + \frac{P_c}{2W_c N_0} \right] \text{ бит/с.} \quad (6.125)$$

Задача 6.13. Показать, что для гауссовского процесса, спектральная плотность которого имеет вид

$$S_u(f) = \begin{cases} P_u/2W_u, & |f| \leq W_u, \\ 0 & |f| > W_u, \end{cases} \quad (6.126)$$

скорость передачи при искажении, определяемом интегрально-квадратической ошибкой, равна

$$R_a(d) = W_u \log_2(P_u/d) \text{ бит/с.} \quad (6.127)$$

Для нахождения минимальной интегрально-квадратической погрешности, с которой такой процесс можно передать по гауссовскому каналу, приравняем (6.122) и (6.127). Тогда

$$\begin{aligned} \left(\frac{S}{\mathcal{N}}\right)_{\text{вых}} &\leq \frac{P_u}{d_0} = \left[1 + \frac{P_c}{2W_c N_0}\right]^{W_c/W_u} = \\ &= \left[1 + \left(\frac{S}{\mathcal{N}}\right)_{\text{кан}}\right]^{W_c/W_u}. \end{aligned} \quad (6.128)$$

Неравенство (6.128) представляет интерес лишь в связи с видом его зависимости от $(S/\mathcal{N})_{\text{кан}}$ и W_c/W_u . Это связано с тем, что процесс с постоянной ограниченной полосой частот плохо описывает большинство реальных физических источников.

Для семейства гауссовских процессов со спектром Буттерворта численная оценка P_u/d_0 скорости при искажении, определяемом интегрально-квадратической ошибкой, была получена Гобликом. По поводу кривых, описанных Гобликом, отсылаем читателя к его статье [22].

Как было указано в гл. 5, результаты Гоблика были использованы Ван Трисом [23] для сравнительного изучения качества систем импульсной модуляции с оптимальным предсказанием для каналов с бесконечной полосой пропускания.

СПИСОК ЛИТЕРАТУРЫ

1. Шеннон К. Э., Математическая теория связи, сб. «Работы по теории информации и кибернетике», ИЛ, М., 1963, стр. 243—332.
2. Шеннон К. Э., Связь при наличии шума, сб. «Работы по теории информации и кибернетике», ИЛ, М., 1963, стр. 433—460.
3. Галлагер Р. Ж., Теория информации и надежная связь, изд-во «Сов. радио», М., 1974.
4. Крамер Г., Математические методы статистики, ИЛ, М., 1948.
5. Уиттекер Э. Т., Ватсон Г. Н., Курс современного анализа, изд. 2, ч. I, Физматгиз, М., 1962.
6. Шеннон К. Э., Теоремы кодирования для дискретного источника при заданном критерии точности, сб. «Работы по теории информации и кибернетике», ИЛ, М., 1963, стр. 587—621.

7. Фихтенгольц Г. М., Курс дифференциального и интегрального исчисления, изд-во «Наука», М., 1966.
8. Пинскер М. С., Информация и информационная устойчивость случайных величин и процессов, изд-во АН СССР, М., 1960.
9. Лозв М., Теория вероятностей, ИЛ, М., 1962.
10. Феллер В., Введение в теорию вероятностей и ее приложения, изд-во «Мир», М., 1967, стр. 291—292.
11. Гренандер У., Серё Г., Тёплицевы формы и их приложения, ИЛ, М., 1961.
12. Гоблик, Холзингер (Goblick T. J., Holsinger J. L.), Analog source digitization: a comparison of theory and practice, *IEEE Trans. Inform. Theory*, IT-13 (July 1967), 323—326.
13. Колмогоров А. Н., Теория передачи информации, кн. «Сессия Академии наук СССР по научным проблемам автоматизации производства 15—20 октября 1956 г.», М., 1957, стр. 66—99.
14. Пинскер М. С., Вычисление скорости создания сообщений стационарным случайным процессом и пропускной способности стационарного канала, *ДАН СССР*, 111 (1956), 753—756.
15. Добрушин Р. Л., Цыбаков Б. С., Передача информации с дополнительным шумом, сб. «Проблемы передачи информации», вып. 14, М., 1963, стр. 21—42.
16. Сакрисон (Sakrison D. J.), The rate distortion function of a Gaussian process with a weighted-square error criterion, *IEEE Trans. Inform. Theory*, IT-14 (May 1968), 506—508.
17. Сакрисон (Sakrison D. J.), Добавление к [14], *IEEE Trans. Inform. Theory*, IT-15 (September 1969), 610—611.
18. Сакрисон (Sakrison D. J.), An extension of the theorem of Каč, Murdock and Szegö to N dimensions, *IEEE Trans. Inform. Theory*, IT-15 (September 1969), 608—610.
19. Сакрисон, Алгази (Sakrison D. J., Algazi V. R.), A comparizon of line-by-line and two-dimensional encoding of random images, *IEEE Trans. Inform. Theory*, IT-17 (June 1971), 386—397.
20. Зейдман (Seidman L. P.), Performance limitations and error calculations for parameter estimation, *IEEE Proceedings*, 58, 5 (1970). (Русский перевод: *ТИИЭР*, 58, 5 (1970).)
21. Зейдман (Seidman L. P.), An upper bound on average estimation error in nonlinear systems, *IEEE Trans. Inform. Theory*, IT-14 (March 1968), 243—250.
22. Гоблик (Goblick T. J., Jr.), Theoretical limitations on the transmission of data from analog sources, *IEEE Trans. Inform. Theory*, IT-11 (October 1965), 558—566.
23. Ван Трис (Van Trees H. L., Jr.), A comparison of optimum angle modulation systems and rate distortion bounds. *IEEE Proceedings*, 53 (December 1965), 2123—2124. (Русский перевод: *ТИИЭР*, 53, 12 (1965), 2347—2348.)

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

- аддитивный гауссовский канал 72
- — — граница качества передачи 76, 84, 158
- — — — — Баранкина 36, 92, 160
- — — — — Возенкрфта — Джекобса 92, 160
- — — — — Зейдмана 160
- — — — — Зива — Закаи 76, 84, 160
- — — — — Крамера — Рао 79, 160
- — — — — скорости при заданном искажении 160
- аномальные ошибки 63, 86
- априорная плотность 14

- байесовская оценка 16
- байесовский риск 16
- бит 111

- вероятностная мера 126
- взаимная информация 129
- — свойства 130, 131
- — условная 132
- время-импульсная модуляция (ВИМ) 74
- выборочная точка 126
- выборочное пространство 126
- — разбиение 129
- выпуклая функция 15

- границы ошибок; см. аддитивный гауссовский канал, нелинейная модуляция, радиолокационное оценивание

- достаточная статистика 24

- зональные сферические функции 74

- изображающий вектор 116
- импульсная модуляция 72
- интегральное уравнение 50
- — собственные функции 50
- интегральный оператор 49
- — положительно определенный 49
- — симметричный 49

- кодový вектор 116
- компа́ндер 80
- комплексная огибающая 45
- коэффициент сжатия 113
- кривая сигнала 75

- мера искажения 114
- — взвешенная квадратичная 147
- метод оценивания 15
- множество кодов 141
- множитель расширения полосы частот 74
- модуляция угловая 94; см. также нелинейная модуляция

- нелинейная модуляция, граница эффективности 84
- — пороговый эффект 84
- — эффективность при малом шуме 76
- неравенство Крамера — Рао 30
- — — в радиолокационном оценивании 65
- — — выполнение равенства 33
- — — для случайных параметров 39

- обмен отношения сигнал/шум на полосу частот 85

- отношение правдоподобия 26
 — — для гауссовского случайного процесса 52
 — — сходимость 53
 оценка 15
 — асимптотически эффективная 52
 — максимального правдоподобия 22, 37
 — — — свойства 37
 — максимума апостериорной вероятности (МАВ) 18, 95
 — несмещенная 23
 — состоятельная 23
 — эффективная 32
- пороговая область 87
 пороговый эффект 76
 преобразование Гильберта 44
 произведение мер 128
 производная Радона — Никодима 128
 пропускная способность 111
- радиолокационное измерение дальности и оценивание 42, 58
 — — — — — границы эффективности 65
 разбиение 129
 разложение Карунена — Лозва 49, 51, 148
 регулярность отношения правдоподобия 27
- скорость при заданном искажении 134
 — — — — гауссовского случайного поля 157
 — — — — — процесса 146, 155
 — — — — гауссовской случайной величины 118, 151
 случайное кодирование 121
- согласованный фильтр 67
 средний риск 16
 средняя взаимная информация 129
 — — — свойства 130, 131
 — — — условная 132
- теорема Каца, Мурдока и Сергё 153
 — обработки данных 133, 137
 — Шеннона о кодировании источников 114, 118
 — — — — — геометрическое доказательство 115
 — — — — — обратная 135
 — — — — — позитивное утверждение 139
- угловая модуляция 94
 узкополосный процесс 44
 уравнение Винера — Хопфа 104
 условная медиана 17
 условное математическое ожидание 17
 условный риск 16
- фазовая автоподстройка частоты 94
 формула Колмогорова 133
 функция множеств 126
 — — счетно аддитивная 126
 — — неопределенности 62
 — — свойство фиксированного объема 67
 — плотности 127
 — потеря 15
 — правдоподобия 23
 — распределения 126
- цифровая скорость 110
 σ -поле 126

ОГЛАВЛЕНИЕ

Предисловие редактора перевода	5
Предисловие к русскому изданию	7
От автора	9
Глава 1. Введение	11
Глава 2. Оценка параметров	14
2.1. Оценка случайных параметров	15
2.2. Оценивание неслучайных параметров	21
Список литературы	41
Глава 3. Оценка параметров сигнала в аддитивном гауссовском шуме	42
3.1. Представление узкополосных случайных процессов	44
3.2. Отношение правдоподобия	52
3.3. Радиолокационное измерение дальности и задача оце- нивания	58
Список литературы	70
Глава 4. Импульсная модуляция	72
4.1. Эффективность при малом шуме	76
4.2. Оптимальное отображение на кривую сигнала	79
4.3. Границы эффективности. Пороговый эффект	84
Список литературы	93
Глава 5. Оптимальная угловая демодуляция	94
5.1. Демодулятор по максимуму апостериорной вероятности (MAB)	94
5.2. Фазовая автоподстройка частоты — субоптимальный де- модулятор	100
Список литературы	109
Глава 6. Кодирование случайных источников сообщений	110
6.1. Гауссовский источник со среднеквадратической мерой искажения	115
6.2. Свойства взаимной информации	125
6.3. Скорость при заданном искажении и теорема кодиро- вания источников	134
6.4. Скорость при заданном искажении для гауссовского случайного процесса со взвешенной квадратической ме- рой искажения	146
6.5. Границы передачи по аддитивному гауссовскому каналу	158
Список литературы	162
Предметный указатель	164

УВАЖАЕМЫЙ ЧИТАТЕЛЬ!

Ваши замечания о содержании книги, ее оформлении, качестве перевода и другие просим присылать по адресу: 129820, Москва, И-110, ГСП, 1-й Рижский пер., д. 2, издательство «Мир».

Д. Сакрисон
ЛЕКЦИИ ОБ АНАЛОГОВОЙ СВЯЗИ

Редактор Л. Б. Штейнпресс
Художник В. М. Новоселова
Художественный редактор В. И. Шаповалов
Технический редактор Н. И. Манохина
Корректоры К. Л. Водяницкая, И. С. Соколова

Сдано в набор 17/VIII 1973 г.

Подписано к печати 26/II 1974 г.

Бумага тип. № 1. для глуб. печ. $84 \times 108^{1/32} = 2,63$
бум. л., печ. л., 8,82. Уч.-изд. л. 7,72. Изд. № 1/7164

Цена 56 к. Зак. 767.

ИЗДАТЕЛЬСТВО «МИР»
Москва, 1-й Рижский пер., 2

Ордена Трудового Красного Знамени
Ленинградская типография № 2
имени Евгении Соколовой Союзполиграфпрома
при Государственном комитете Совета Министров
СССР по делам издательств, полиграфии
и книжной торговли,
198052, Ленинград, Л-52, Измайловский проспект, 29

56 к.

