Inteligentna odbijająca powierzchnia wspomagana komunikacją dla 6G

Tradycyjne metody realizacji systemów bezprzewodowych o dużej przepustowości, tj. wdrażanie gęstego i heterogenicznego sprzętu sieciowego, poprawa wydajności widmowej i pozyskiwanie większej przepustowości, wiążą się z wysokimi nakładami kapitałowymi i operacyjnymi, wysokim zużyciem energii i poważnymi wzajemnymi zakłóceniami. Aby sprostać surowym wymaganiom wydajnościowym systemu nowej generacji, wysoce pożądane jest znalezienie przełomowej i rewolucyjnej technologii, która pozwoli osiągnąć zrównoważony wzrost pojemności i wydajności przy przystępnych kosztach, niskiej złożoności i niskim zużyciu energii. Ostatnio inteligentna odbijająca powierzchnia (IRS), znana również jako rekonfigurowalna inteligentna powierzchnia (RIS), przyciągnęła wiele zainteresowań zarówno ze środowiska akademickiego, jak i przemysłu. Jest to obiecujący nowy paradygmat umożliwiający osiągnięcie inteligentnego i rekonfigurowalnego środowiska propagacji kanału bezprzewodowego. Mówiąc ogólnie, IRS to płaska powierzchnia składająca się z dużej liczby małych, pasywnych i niedrogich elementów odbijających, z których każdy jest w stanie niezależnie wywołać przesunięcie fazowe i tłumienie amplitudy padającej fali elektromagnetycznej. W przeciwieństwie do obecnych technik transmisji bezprzewodowej, które muszą pasywnie dostosowywać się do środowiska propagacji bezprzewodowej, IRS proaktywnie modyfikuje je poprzez inteligentnie kontrolowane odbicie, dzięki czemu wspólnie osiąga się drobnoziarniste pasywne lub odbijające kształtowanie wiązki. Poprzez rozsądne projektowanie współczynników odbicia, sygnały odbite przez IRS mogą być dodawane konstruktywnie z sygnałami za pośrednictwem innych ścieżek sygnałowych w celu zwiększenia pożądanej siły sygnału w odbiorniku lub destrukcyjnie w celu złagodzenia zakłóceń współkanałowych. W ten sposób zapewnia nowy stopień swobody w projektowaniu systemów bezprzewodowych za pomocą inteligentnego i programowalnego środowiska bezprzewodowego. Ten rozdział składa się z następujących głównych części:

● Podstawowa koncepcja IRS, techniczne zalety bezprzewodowej komunikacji wspomaganej przez IRS i jej potencjalne zastosowania.

● Podstawy pojedynczej transmisji wspomaganej przez IRS, w tym kaskadowy kanał IRS, model sygnału pasywnego kształtowania wiązki, optymalne współczynniki odbicia i specjalna cecha zwana stratą ścieżki odległości produktu.

● Podstawy wspomaganej przez IRS transmisji wieloantenowej, w tym wspólne aktywne i pasywne formowanie wiązki oraz wspólne wstępne kodowanie i optymalizacja odbicia.

● Wprowadzenie techniki dwuwiązkowej IRS, formowanie podwójnych wiązek nad hybrydowymi analogowo-cyfrowymi transceiverami i projektowanie optymalizacji.

● Podstawy wspomaganej przez IRS szerokopasmowej transmisji, w tym kaskadowy kanał zanikania selektywnego częstotliwościowo, model systemu wspomaganej przez IRS transmisji z ortogonalnym zwielokrotnieniem z podziałem częstotliwości (OFDM) i maksymalizacja szybkości.

● Wpływ starzenia się kanału na IRS, modelowanie nieaktualnych informacji o stanie kanału i analiza utraty wydajności. Ponadto wprowadzono zasadę predykcji kanału opartej na uczeniu maszynowym oraz podstawy rekurencyjnych sieci neuronowych, pamięci długoterminowej i krótkoterminowej oraz głębokiego uczenia się.

Podstawowa koncepcja

Tradycyjnie, rygorystyczne wymagania dotyczące wydajności komunikacji bezprzewodowej, na przykład bardzo wysoka szybkość transmisji danych, wysoka efektywność energetyczna, wszechobecny

zasięg, ogromna łączność, bardzo wysoka niezawodność i niskie opóźnienie, określone w 5G, zostały osiągnięte dzięki trzem głównym podejściom:

● Gęsta i heterogeniczna instalacja sieciowa: Wdrażanie coraz większej liczby urządzeń sieciowych, takich jak stacje bazowe, punkty dostępowe, zdalne głowice radiowe, przekaźniki i rozproszone anteny, może zwiększyć ponowne wykorzystanie zasobów widmowych na danym obszarze geograficznym i skrócić odległość propagacji między punktem obsługi a użytkownikiem. Chociaż takie podejście może znacznie rozszerzyć zasięg sieci i zwiększyć przepustowość systemu, wiąże się z wysokimi wydatkami kapitałowymi i operacyjnymi, wysokim zużyciem energii i poważnymi wzajemnymi zakłóceniami.

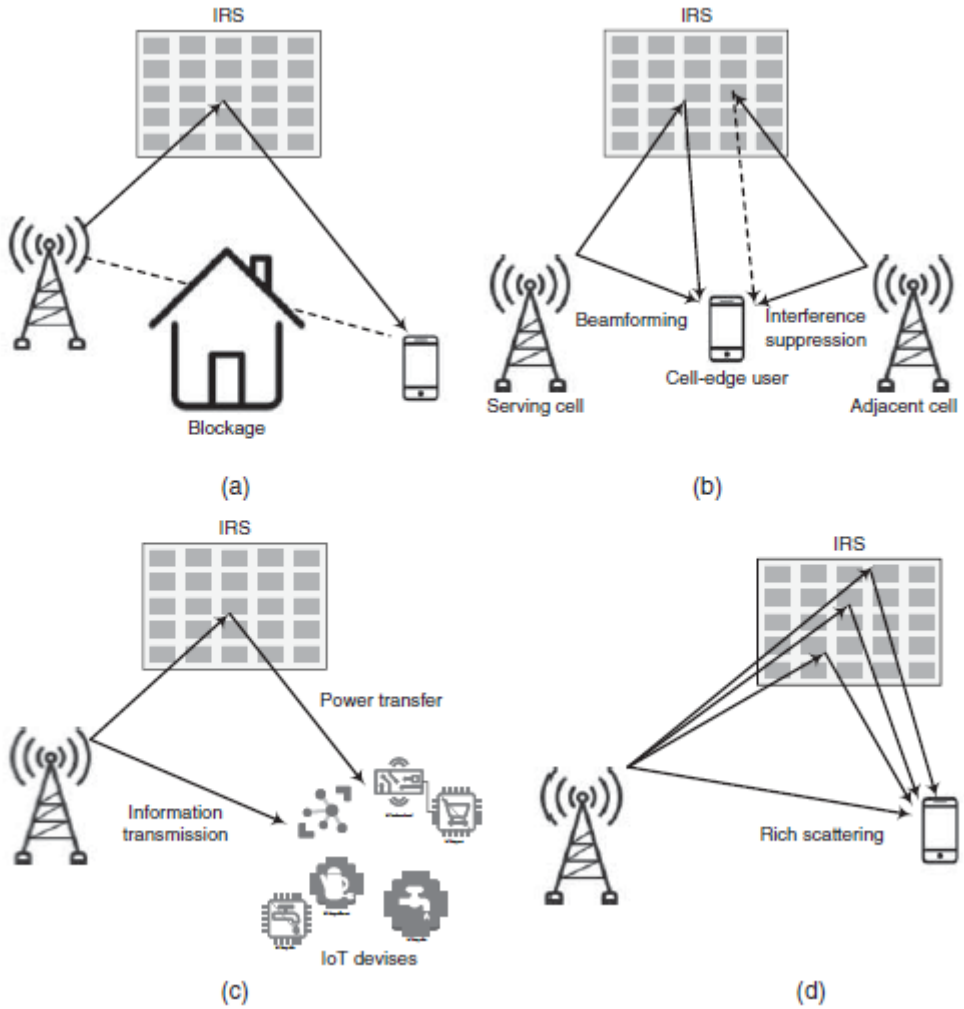
● Wysoka wydajność widmowa: Zintegrowanie ogromnej liczby anten na stacji bazowej wykorzystuje ogromny zysk multipleksowania przestrzennego za pośrednictwem technologii MIMO (Multi-Input Multi-Output). Takie podejście wymaga zaawansowanych technik przetwarzania sygnału, co wiąże się z wysokimi kosztami sprzętu i zużyciem energii. Ze względu na podstawowe ograniczenia środowiska propagacji, np. kanały niskiej rangi i wysoka korelacja między antenami, a także ograniczenia praktyczne, takie jak duży rozmiar tablicy, utrudniają, jeśli nie uniemożliwiają, dalszą znaczącą poprawę wydajności widmowej poprzez samo zwiększenie liczby anten.

● Większa przepustowość: Jednym z trendów technologicznych w transmisji bezprzewodowej jest to, że przepustowość sygnału staje się coraz szersza, od dziesiątek kHz w systemach pierwszej generacji do setek MHz w systemach piątej generacji, co ma na celu obsługę wyższej szybkości transmisji. Odpowiednio, wymagana jest duża liczba zasobów widmowych, co prowadzi do niedoboru widma. W następnej generacji migracja do pasm wyższych częstotliwości, takich jak fala milimetrowa (mmWave), teraherc (THz), a nawet fale świetlne, w celu wykorzystania ich obfitej przepustowości, staje się koniecznością w celu osiągnięcia rygorystycznej wydajności, na przykład ekstremalnej prędkości transmisji Terabitów na sekundę (Tbps). Poważne straty propagacyjne i podatność na blokowanie transmisji o wysokiej częstotliwości nieuchronnie wymagają gęstszego rozmieszczenia sieci i montażu większej liczby anten (tj. dużej macierzy antenowej dla uzyskania wysokiego wzmocnienia wiązki). Ten paradygmat dodatkowo wyolbrzymia problemy wysokich nakładów inwestycyjnych i operacyjnych, wysokiego zużycia energii i poważnych wzajemnych zakłóceń.

Oprócz dalszego rozwijania wyżej wymienionych technologii w kierunku zapotrzebowania systemów nowej generacji, wysoce pożądane jest znalezienie przełomowej i rewolucyjnej technologii, aby osiągnąć zrównoważony wzrost pojemności i wydajności przy przystępnych kosztach, niskiej złożoności i niskim zużyciu energii. Z drugiej strony, podstawowe wyzwanie, jakim jest poważne ograniczenie wydajności komunikacji bezprzewodowej, przypisuje się nieuchwytnym kanałom bezprzewodowym ze względu na ich znaczną utratę przepustowości, zacienianie, zmienność czasową, selektywność częstotliwości i propagację wielościeżkową. Tradycyjne podejścia do radzenia sobie z tym podstawowym ograniczeniem albo kompensują utratę kanału i losowość, wykorzystując różne solidne techniki modulacji, kodowania i różnorodności, albo dostosowują się do niego poprzez adaptacyjne sterowanie parametrami transmisji. Niemniej jednak techniki te nie tylko wymagają dużej ilości narzutu, ale także mają ograniczoną adaptowalność w w dużej mierze losowych kanałach bezprzewodowych, pozostawiając tym samym solidną barierę dla osiągnięcia wysoce niezawodnej komunikacji bezprzewodowej. W tym względzie, Intelligent Reflecting Surface (IRS), znana również jako Reconfigurable Intelligent Surface (RIS), Large Intelligent Surface , Large Intelligent Metasurface (LIM), Programmable Metasurface Reconfigurable Metasurface, Intelligent Walls i Reconfigurable Reflect-array, została zaproponowana jako obiecujący nowy paradygmat w celu osiągnięcia inteligentnego i rekonfigurowalnego środowiska propagacji kanału bezprzewodowego. Mówiąc ogólnie, IRS jest płaską powierzchnią składającą się z dużej liczby małych, pasywnych i niedrogich elementów odbijających, z których każdy jest w stanie niezależnie wywołać przesunięcie fazowe i/lub

tłumienie amplitudy (zwane zbiorczo współczynnikiem odbicia) padającej fali elektromagnetycznej. W przeciwieństwie do obecnych technik transmisji bezprzewodowej, które muszą pasywnie dostosowywać się do środowiska propagacji bezprzewodowej, IRS proaktywnie modyfikuje je poprzez inteligentnie kontrolowane odbicie, dzięki czemu wspólnie osiąga się drobnoziarniste pasywne lub odbijające kształtowanie wiązki. Poprzez rozsądne projektowanie współczynników odbicia, sygnały odbite przez IRS mogą być dodawane konstruktywnie z sygnałami za pośrednictwem innych ścieżek sygnałowych w celu zwiększenia pożądanej siły sygnału w odbiorniku lub destrukcyjnie w celu złagodzenia zakłóceń współkanałowych. W ten sposób zapewnia nowy stopień swobody w projektowaniu systemów bezprzewodowych za pomocą inteligentnego i programowalnego środowiska bezprzewodowego. Ponieważ jego elementy odbijające (np. niedrogie drukowane dipole) odbijają tylko pasywnie uderzającą falę elektromagnetyczną, łańcuchy częstotliwości radiowej (RF) do transmisji i odbioru sygnału stają się zbędne. W ten sposób można go wdrożyć przy rzędach wielkości niższych kosztach sprzętowych i zużyciu energii niż tradycyjne aktywne układy antenowe. Ponadto elementy odbijające są na ogół niskoprofilowe, lekkie i mają geometrię konforemną. Dlatego IRS można praktycznie wykonać tak, aby pasował do montażu na powierzchniach o dowolnym kształcie, aby sprostać szerokiej gamie scenariuszy wdrożeniowych i być zintegrowanym z istniejącymi sieciami bezprzewodowymi w sposób przejrzysty jako sprzęt pomocniczy, zapewniając w ten sposób dużą elastyczność i kompatybilność. Krótko mówiąc, IRS jest uznawany za przełomową technologię o wybitnych cechach niskiej złożoności, niskich kosztach, niskim zużyciu energii, wraz z potencjałem wysokiej wydajności. IRS wykazuje kilka szczególnych zalet w porównaniu z innymi pokrewnymi technologiami, tj. przekaźnikami bezprzewodowymi, komunikacją rozproszenia wstecznego i aktywną powierzchnią masywną MIMO . Przekaźniki bezprzewodowe zwykle działają w trybie półdupleksowym i w związku z tym cierpią na niską wydajność widmową w porównaniu z IRS działającym w trybie pełnego dupleksu. Chociaż przekaźnik pełnego dupleksu jest również możliwy do osiągnięcia, wymaga on wyrafinowanych technik samoniwelacji zakłóceń, których wdrożenie jest kosztowne. Ponadto IRS jest wolny od wzmocnienia szumów, ponieważ odbija jedynie uderzające fale elektromagnetyczne jako pasywny układ bez żadnego aktywnego modułu nadawczego (np. wzmacniacza mocy). W przeciwieństwie do tradycyjnego rozpraszania wstecznego, takiego jak znaczniki RFID (Radio Frequency IDentification), które komunikują się z czytnikiem RFID poprzez modulację odbitego sygnału emitowanego przez czytnik, IRS jest stosowany głównie w celu wspomagania istniejącego łącza komunikacyjnego bez własnych danych. Natomiast czytnik w komunikacji rozpraszania wstecznego musi wdrożyć anulowanie samozakłóceń w swoim odbiorniku, aby zdekodować wiadomość znacznika. Zarówno łącze bezpośrednie, jak i łącze odbite w transmisji wspomaganej przez IRS przenoszą identyczny sygnał, a zatem mogą być spójnie nałożone na odbiornik, aby zwiększyć siłę sygnału w celu lepszego wykrywania. Po trzecie, IRS różni się również od masywnego MIMO opartego na aktywnej powierzchni ze względu na różne architektury macierzy (pasywna kontra aktywna) i mechanizmy działania (odbijająca kontra nadawanie). Ze względu na wcześniej wymienione zalety, IRS nadaje się do masowego wdrażania w sieciach bezprzewodowych, aby znacznie zwiększyć ich wydajność widmową i efektywność energetyczną w sposób opłacalny. Przewiduje się, że IRS doprowadzi do fundamentalnej zmiany paradygmatu projektowania systemu bezprzewodowego, a mianowicie od skalowania kolejności systemów Massive MIMO pod względem liczby anten do MIMO o umiarkowanej skali wspomaganego przez IRS, a także od istniejącej heterogenicznej sieci bezprzewodowej do sieci hybrydowej wspomaganej przez IRS. W przeciwieństwie do Massive MIMO, które wykorzystuje dziesiątki, a nawet setki aktywnych anten do bezpośredniego tworzenia ostrych wiązek, system MIMO wspomagany przez IRS umożliwia wyposażenie stacji bazowej w znacznie mniej anten bez uszczerbku dla jakości doświadczeń użytkowników poprzez wykorzystanie dużej apertury IRS do tworzenia drobnoziarnistych wiązek odbitych poprzez inteligentne pasywne odbicie. Aby to zrobić, koszt sprzętu i zużycie energii systemu można znacznie obniżyć, szczególnie w przypadku systemów

bezprzewodowych migrujących do wyższych pasm częstotliwości. Z drugiej strony, chociaż istniejące sieci bezprzewodowe opierają się na heterogenicznej architekturze wielowarstwowej składającej się z makro, mikro i małych stacji bazowych, zdalnych głowic radiowych, przekaźników, rozproszonych anten itp., wszystkie są aktywnymi węzłami, które generują sygnały, co wymaga wyrafinowanej koordynacji i eliminacji zakłóceń. To tradycyjne podejście nieuchronnie pogarsza narzut działania sieci i dlatego może nie być w stanie utrzymać wzrostu przepustowości sieci bezprzewodowej w sposób opłacalny. Natomiast zintegrowanie IRS z siecią bezprzewodową przesunie istniejącą heterogeniczną sieć z aktywnymi komponentami tylko do nowej hybrydowej architektury obejmującej zarówno aktywne, jak i pasywne komponenty. Ponieważ IRS są znacznie tańsze w porównaniu ze swoimi aktywnymi odpowiednikami, można je gęściej rozmieszczać w sieci bezprzewodowej przy jeszcze niższych kosztach, ale bez wprowadzania zakłóceń dzięki ich pasywnemu odbiciu i wynikającemu z tego lokalnemu zasięgowi. Poprzez optymalne ustawienie współczynników między aktywnymi węzłami i pasywnymi IRS w sieci hybrydowej można osiągnąć zrównoważone, ekologiczne, tanie skalowanie przepustowości sieci. Rysunek ilustruje kilka obiecujących zastosowań bezprzewodowej transmisji wspomaganej przez IRS.



Pierwsze zastosowanie pokazuje martwy punkt, w którym bezpośrednie połączenie między użytkownikiem a obsługującą go stacją bazową jest poważnie blokowane przez przeszkodę, np. budynek z żelbetonu. W tym przypadku wdrożenie IRS, który ma silne połączenia zarówno ze stacją bazową, jak i użytkownikiem, może ominąć przeszkodę za pomocą inteligentnego odbicia sygnału, tworząc w ten sposób wirtualne łącze Line-Of-Sight (LOS). Jest to szczególnie pomocne w przypadku rozszerzenia zasięgu w komunikacji mmWave i THz, które są bardzo podatne na blokowanie. Drugie zastosowanie koncentruje się na użytkowniku na skraju komórki cierpiącym zarówno z powodu dużego tłumienia sygnału ze swojej obsługującej stacji bazowej, jak i znacznych zakłóceń współkanałowych ze stacji bazowej sąsiedniej. Wdrożenie IRS na skraju komórki może poprawić pożądaną siłę sygnału, jednocześnie tłumiąc zakłócenia międzykomórkowe poprzez odpowiednie zaprojektowanie pasywnego kształtowania wiązki, tworząc w ten sposób gorący punkt sygnału, a także strefę wolną od zakłóceń w jego pobliżu. Trzecia aplikacja rozważa użycie IRS w celu wspomagania implementacji Jednoczesnego Bezprzewodowego Przesyłania Informacji i Mocy (SWIPT). W scenariuszu wdrożenia masywnych urządzeń o niskim poborze mocy lub pasywnych w sieci Internetu Rzeczy (IoT) duża apertura IRS jest wykorzystywana do kompensacji znacznego tłumienia mocy na dużą odległość za pomocą odbicia wiązki, aby poprawić wydajność bezprzewodowego przesyłu mocy. Na koniec czwarta aplikacja dostarcza ogólny opis sztucznej manipulacji statystykami kanału poprzez dodawanie dodatkowych ścieżek sygnału w kierunku pożądanego kierunku, aby np. poprawić stan rangi kanału lub przekształcić kanał zanikający Rayleigha w kanał zanikający Rician.

Transmisja z pojedynczą anteną wspomagana przez IRS

W tej sekcji badamy podstawy podstawowej bezprzewodowej transmisji punkt-punkt wspomaganej przez IRS pod względem jej modeli sygnału i kanału. Rozważmy system trójwęzłowy składający się ze stacji bazowej, użytkownika i IRS z N pasywnymi elementami odbijającymi na płaskiej powierzchni, oznaczonymi odpowiednio jako $\mathbb{B}$, $\mathbb{U}$ i $\mathbb{I}$. Dla uproszczenia zakładamy najpierw, że zarówno stacja bazowa, jak i sprzęt użytkownika są wyposażone w pojedynczą antenę, podczas gdy szerokość pasma sygnału B_s jest wąska przy danej częstotliwości nośnej f_c, przy czym $B_s \ll f_c$. Jednak bardziej ogólne przypadki, tj. systemy szerokopasmowe z wieloma antenami, wieloma użytkownikami, wieloma komórkami i wieloma nośnymi, zostaną przedstawione w kolejnych sekcjach.

Model sygnału

Matematycznie równoważny sygnał transmisyjny o wartościach zespolonych pasma podstawowego jest oznaczany jako $s_b(t)$. Po konwersji w górę łańcucha RF antena transmisyjna podaje sygnał pasma przepustowego

$$s(t) = \Re\left[s_b(t)e^{j2\pi f_c t}\right] \quad (1)$$

do kanału bezprzewodowego, gdzie $\Re[\cdot]$ oznacza część rzeczywistą liczby zespolonej, a j jest jednostką urojoną $j^2 = -1$. Nie tracąc ogólności, najpierw skupiamy się na transmisji sygnału downlink ze stacji bazowej do użytkownika za pośrednictwem określonego elementu odbijającego IRS, oznaczonego przez n, gdzie $n \in \{1, 2,\ldots,N\}$. Odpowiedź impulsowa dla kanału zanikającego wielodrogowo między stacją bazową a n-tym elementem odbijającym może być modelowana przez

$$h_n(\tau) = \sum_{l=1}^{L} \alpha_{n,l}\delta(\tau - \tau_{n,l}), \quad (2)$$

gdzie L wyraża całkowitą liczbę możliwych do rozdzielenia ścieżek sygnału, $\alpha_{n,l}$ i $\tau_{n,l}$ oznaczają tłumienie i opóźnienie propagacji ścieżki l, zakładając, że odpowiedź kanału jest niezmienna w czasie podczas transmisji s(t). W rezultacie sygnał uderzający w n-ty element odbijający IRS jest wyrażony jako

$$\begin{aligned} r_n(t) &= h_n(\tau) * s(t) \\ &= \sum_{l=1}^{L} \alpha_{n,l} s(t-\tau_{n,l}) \\ &= \sum_{l=1}^{L} \alpha_{n,l} \Re\left[s_b(t-\tau_{n,l}) e^{j2\pi f_c(t-\tau_{n,l})}\right] \\ &= \Re\left[\sum_{l=1}^{L} \alpha_{n,l} s_b(t-\tau_{n,l}) e^{j2\pi f_c(t-\tau_{n,l})}\right] \\ &= \Re\left[\sum_{l=1}^{L} \alpha_{n,l} e^{-j2\pi f_c \tau_{n,l}} s_b(t-\tau_{n,l}) e^{j2\pi f_c t}\right] \\ &= \Re\left[(h_n^b(\tau) * s_b(t)) e^{j2\pi f_c t}\right], \end{aligned}$$

(3)

gdzie $*$ oznacza splot liniowy, a równoważną odpowiedź impulsową pasma podstawowego

$$h_n^b(\tau) = \sum_{l=1}^{L} \alpha_{n,l} e^{-j2\pi f_c \tau_{n,l}} \delta(\tau - \tau_{n,l}).$$

(4)

Piszemy $\beta_n \in [0, 1]$ i τn, aby oznaczyć tłumienie amplitudy i opóźnienie wywołane przez n-ty element odbijający. Ignorując uszkodzenia sprzętowe, np. szum fazowy i nieliniowość obwodu, sygnał odbity przez n-ty element odbijający można wyrazić jako

$$\begin{aligned} (t) &= \beta_n r_n(t-\tau_n) \\ &= \beta_n \Re\left[\left(h_n^b(\tau) * s_b(t-\tau_n)\right) e^{j2\pi f_c(t-\tau_n)}\right] \\ &\approx \Re\left[\left(h_n^b(\tau) * s_b(t)\right) \beta_n e^{-j2\pi f_c \tau_n} e^{j2\pi f_c t}\right] \\ &= \Re\left[\left(h_n^b(\tau) * s_b(t)\right) c_n e^{j2\pi f_c t}\right]. \end{aligned}$$

(5)

Warunek $s_b(t - \tau_n) \approx s_b(t)$ jest łatwy do spełnienia, ponieważ okres symbolu jest znacznie większy niż opóźnienie fizyczne wywołane przez element odbijający, tj. $T_s = 1/B_s \gg \tau_n$, w wąskopasmowym systemie bezprzewodowym. Zapisujemy $c_n = \beta n^{ej\theta n}$, aby oznaczyć współczynnik odbicia n-tego elementu odbijającego, gdzie $\theta_n = -2\pi f_c \tau_n \in [0, 2\pi)$ jest przesunięciem fazowym wywołanym przez element odbijający, a to przesunięcie fazowe jest okresowe względem 2π. Podobnie odpowiedź impulsowa dla kanału zanikania wielodrożnego między n-tym elementem odbijającym a użytkownikiem może być modelowana przez

$$g_n(\tau) = \sum_{\mathfrak{l}=1}^{\mathfrak{L}} \alpha_{n,\mathfrak{l}} \delta(\tau - \tau_{n,\mathfrak{l}}),$$

(6)

gdzie $\mathfrak{L}$ wyraża całkowitą liczbę możliwych do rozdzielenia ścieżek sygnału, $\alpha_{n,\mathfrak{l}}$ i $\tau_{n,\mathfrak{l}}$ oznaczają tłumienie i opóźnienie propagacji ścieżki $\mathfrak{l}$, zakładając, że odpowiedź kanału jest niezmienna w czasie podczas

transmisji $x_n(t)$. W konsekwencji odebrany sygnał u użytkownika ze względu na n-ty element odbijający jest podany przez

$$\begin{aligned} y_n(t) &= g_n(\tau) * x_n(t) \\ &= \sum_{I=1}^{\mathfrak{L}} \alpha_{n,I} x_n(t - \tau_{n,I}) \\ &= \sum_{I=1}^{\mathfrak{L}} \alpha_{n,I} \Re \left[\left(h_n^b(\tau) * s_b(t - \tau_{n,I}) \right) \beta_n e^{-j2\pi f_c \tau_n} e^{j2\pi f_c (t - \tau_{n,I})} \right] \\ &= \Re \left[\sum_{I=1}^{\mathfrak{L}} \alpha_{n,I} e^{-j2\pi f_c \tau_{n,I}} \left(h_n^b(\tau) * s_b(t - \tau_{n,I}) \right) c_n e^{j2\pi f_c t} \right]. \end{aligned} \quad (7)$$

Określenie równoważnej odpowiedzi impulsowej pasma podstawowego $g_n(\tau)$ jako

$$g_n^b(\tau) = \sum_{I=1}^{\mathfrak{L}} \alpha_{n,I} e^{-j2\pi f_c \tau_{n,I}} \delta(\tau - \tau_{n,I}). \quad (8)$$

Równanie (7) można zapisać w następujący sposób

$$y_n(t) = \Re \left[\left(g_n^b(\tau) * h_n^b(\tau) * s_b(t) \right) c_n e^{j2\pi f_c t} \right] \quad (9)$$

Oznaczając $y_{n,b}(t)$ przez równoważny sygnał odebrany pasma podstawowego, sygnał odebrany pasma przepustowego można również wyrazić jako

$$y_n(t) = \Re \left[y_{n,b}(t) e^{j2\pi f_c t} \right] \quad (10)$$

Porównując równanie (10) z równaniem (9), otrzymujemy

$$\begin{aligned} y_{n,b}(t) &= \left(g_n^b(\tau) * h_n^b(\tau) * s_b(t) \right) \beta_n e^{-j2\pi f_c \tau_n} \\ &= g_n^b(\tau) * c_n * h_n^b(\tau) * s_b(t). \end{aligned} \quad (11)$$

Do tej pory równoważną odpowiedź impulsową pasma podstawowego kanału kaskadowego od stacji bazowej do użytkownika za pośrednictwem n-tego elementu odbijającego można było modelować za pomocą

$$v_n(t) = g_n^b(\tau) * c_n * h_n^b(\tau). \quad (12)$$

Ponadto możemy wiedzieć, że model kanału równoważnego pasma podstawowego w czasie dyskretnym w systemie wąskopasmowym można podać wzorem

$$v_n = g_n h_n c_n, \quad (13)$$

który jest kanałem kaskadowym oznaczonym przez iloczyn trzech członów, tj. współczynnika kanału między stacją bazową a elementem odbijającym, współczynnika odbijającego i współczynnika kanału między elementem odbijającym a użytkownikiem. Model wąskopasmowy opiera się na fakcie, że pojedyncze dotknięcie jest wystarczające, aby wyrazić kanał o płaskiej częstotliwości, gdzie h_n i g_n oznaczają współczynniki kanału między stacją bazową a n-tym elementem odbijającym i n-tym elementem odbijającym a użytkownikiem. Ogólnie rzecz biorąc, h_n jest kołowo symetryczną zespoloną

zmienną losową Gaussa o zerowej średniej i wariancji σ^2_h, oznaczoną jako $h_n \sim \mathcal{CN}(0, \sigma_h^2)$, a także $g_n \sim \mathcal{CN}(0, \sigma_g^2)$. Warto wspomnieć, że odbite łącze ze stacji bazowej do użytkownika za pośrednictwem IRS jest również określane w literaturze jako dyadyczny kanał rozpraszania wstecznego lub kanał otworkowy , z całkiem odmiennymi zachowaniami od łącza bezpośredniego. Mówiąc konkretnie, każdy element odbijający na IRS zachowuje się jak otwór szpilkowy, który łączy wszystkie odebrane sygnały wielościeżkowe w jednym punkcie fizycznym i ponownie rozprasza połączony sygnał, jakby pochodził ze źródła punktowego. Załóżmy, że nie ma sprzężenia sygnału w odbiciu przez sąsiednie elementy IRS, tj. wszystkie elementy IRS niezależnie odbijają sygnały padające. Ze względu na znaczną utratę ścieżki, rozważamy tylko sygnały odbite przez IRS po raz pierwszy i ignorujemy te odbite wielokrotnie. W związku z tym odebrany sygnał ze wszystkich elementów odbijających można modelować jako superpozycję ich odpowiednich odbitych sygnałów. Na podstawie modelu odbijającego w równaniu (11) zatem dyskretny model sygnału pasma podstawowego uwzględniający wszystkie N elementów odbijających jest obliczany przez

$$y = \left(\sum_{n=1}^{N} g_n c_n h_n \right) \sqrt{P_t} s + z \tag{14}$$

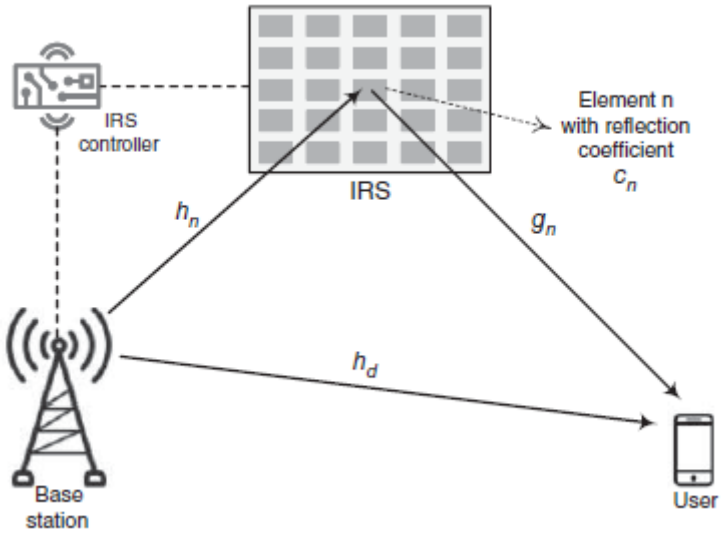
gdzie y oznacza odebrany symbol, s jest znormalizowanym przesłanym symbolem spełniającym $\mathbb{E}[|s|^2] = 1$, P_t wyraża przesłaną moc stacji bazowej, a n jest addytywnym białym szumem gaussowskim (AWGN) o zerowej średniej i wariancji σ^2_n, tj. $z \sim \mathcal{CN}(0, \sigma_z^2)$. Niech $h = [h_1, h_2, \ldots, h_N]^T$, $g = [g_1, g_2, \ldots, g_N]^T$ i $\boldsymbol{\Theta} = \text{diag}(c_1, c_2, \ldots, c_N)$, możemy otrzymać postać wektorową równania (14) jako

$$y = \mathbf{g}^T \boldsymbol{\Theta} \mathbf{h} \sqrt{P_t} s + z \tag{15}$$

Należy zauważyć, że zakłada się, że IRS wykonuje liniowe mapowanie sygnałów padających na sygnały odbite. Jeśli istnieje sprzężenie sygnału lub wspólne przetwarzanie nad elementami odbijającymi, macierz odbijająca $\boldsymbol{\Theta}$ nie byłaby już diagonalna. Ponadto, ponieważ odbiornik może podsłuchiwać sygnały zarówno z odbitego łącza oznaczonego jako $\mathbb{B} - \mathbb{I} - \mathbb{U}$, jak i łącza bezpośredniego oznaczonego jako $\mathbb{B} - \mathbb{U}$, obserwację u użytkownika można wyrazić jako

$$y = \left(\sum_{n=1}^{N} g_n c_n h_n + h_d \right) \sqrt{P_t} s + z = \left(\mathbf{g}^T \boldsymbol{\Theta} \mathbf{h} + h_d \right) \sqrt{P_t} s + z \tag{16}$$

gdzie $h_d \sim \mathcal{CN}(0, \sigma_d^2)$ jest współczynnikiem kanału bezpośredniego łącza. Dyskretny model równoważny pasma podstawowego bezprzewodowej transmisji sygnału wspomaganej przez IRS jest zilustrowany na rysunku .



Pasywne kształtowanie wiązki

Model sygnału IRS można stosować zarówno do bezprzewodowego przesyłu mocy, gdzie zebrana energia jest zazwyczaj modelowana jako wklęsła i rosnąca funkcja mocy odebranego sygnału, jak i do przesyłu informacji, gdzie osiągalna szybkość jest funkcją logarytmiczną współczynnika sygnału do szumu (SNR). Tak więc głównym zadaniem transmisji wspomaganej przez IRS jest wykonywanie pasywnego kształtowania wiązki lub odbijania wiązki poprzez rozsądne dostosowywanie współczynników odbicia, tak aby sygnały odbite przez IRS mogły być dodawane konstruktywnie z sygnałami za pośrednictwem innych ścieżek sygnałowych w celu zwiększenia pożądanej siły sygnału lub destrukcyjnie łagodząc zakłócenia współkanałowe. Zgodnie z tym odebrany współczynnik SNR u użytkownika jest podawany przez

$$\gamma = \frac{P_t|\sum_{n=1}^{N} g_n c_n h_n + h_d|^2}{\sigma_z^2} = \frac{P_t|\mathbf{g}^T\boldsymbol{\Theta}\mathbf{h} + h_d|^2}{\sigma_z^2}. \quad (17)$$

Pojemność kanału lub maksymalna osiągalna szybkość podstawowej transmisji punkt-punkt wspomaganej IRS jest obliczana przez R = log(1 + γ). Aby zmaksymalizować pojemność kanału, współczynniki odbijające c_n, ∀n = 1, 2,…,N są optymalizowane w odniesieniu do chwilowych warunków kanału. Zaniedbując stałe człony i stosując ciągłe amplitudy odbicia i przesunięcia fazowe, problem optymalizacji można sformułować jako

$$\begin{aligned} \max_{\boldsymbol{\beta},\,\boldsymbol{\theta}} \quad & \left|\sum_{n=1}^{N} g_n c_n h_n + h_d\right|^2 \\ \text{s.t.} \quad & \beta_n \in [0,1], \quad \forall n = 1,2,\ldots,N \\ & \theta_n \in [0,2\pi), \quad \forall n = 1,2,\ldots,N, \end{aligned} \quad (18)$$

gdzie $\boldsymbol{\beta} = [\beta_1, \beta_2,\ldots, \beta_N]^T$ i $\boldsymbol{\theta} = [\theta_1, \theta_2,\ldots, \theta_N]T$. Jak wiemy, siła sygnału jest maksymalizowana, gdy wszystkie przychodzące sygnały są spójnie łączone w odbiorniku poprzez wyrównanie ich faz. W konsekwencji optymalne przesunięcie fazowe dla odbijającego elementu n powinno wynosić

$$\theta_n^\star = \mathrm{mod}\,[\psi_d - (\phi_{h,n} + \phi_{g,n}), 2\pi], \quad (19)$$

gdzie $\phi_{h,n}$, $\phi_{g,n}$ i ψ_d oznaczają fazy h_n, g_n i h_d, odpowiednio, a mod [·] jest operacją modulo. To równanie oznacza, że współczynnik odbicia każdego elementu kompensuje obrót fazy wywołany przez połączenia $\mathbb{B}$-$\mathbb{I}$ i $\mathbb{I}$-$\mathbb{U}$, co skutkuje fazą resztkową zgodną z fazą połączenia bezpośredniego, tak aby osiągnąć spójne łączenie. Biorąc pod uwagę blokadę w połączeniu bezpośrednim, h_d zbliża się do zera, a zatem ψ_d w powyższym równaniu można zastąpić dowolną wartością fazy bez zmiany zmaksymalizowanego wyniku. To rozwiązanie optymalizacyjne opiera się na fakcie, że wartości β_n nie wpływają na optymalność w spójnym łączeniu, ponieważ różne sygnały są współfazowane. Stosując optymalne fazy w równaniu (19), problem optymalizacji w równaniu (7.18) jest uproszczony do

$$\begin{aligned} \max_{\boldsymbol{\beta}} \quad & \left|\sum_{n=1}^{N} |g_n||h_n|\beta_n + |h_d|\right|^2 \\ \text{s.t.} \quad & \beta_n \in [0,1], \quad \forall n = 1,2,\ldots,N, \end{aligned} \quad (20)$$

postępując zgodnie z pochodzeniem

$$\begin{aligned} \left|\sum_{n=1}^{N} g_n c_n h_n + h_d\right|^2 &= \left|\sum_{n=1}^{N} |g_n||h_n|\beta_n e^{j(\theta_n^\star + \phi_{h,n} + \phi_{g,n})} + h_d\right|^2 \\ &= \left|\sum_{n=1}^{N} |g_n||h_n|\beta_n e^{j\psi_d} + |h_d| e^{j\psi_d}\right|^2 \\ &= \left|\sum_{n=1}^{N} |g_n||h_n|\beta_n + |h_d|\right|^2 |e^{j\psi_d}|^2 \\ &= \left|\sum_{n=1}^{N} |g_n||h_n|\beta_n + |h_d|\right|^2. \end{aligned} \quad (21)$$

Łatwo jest stwierdzić, że optymalne amplitudy odbicia są podane przez $\beta^\star_n = 1$, $\forall n = 1, 2,\ldots,N$, ponieważ maksymalizacja siły sygnału w każdej odbitej ścieżce osiąga największą moc odbioru w warunkach spójnego łączenia. Ciekawą obserwacją jest to, że optymalny współczynnik odbicia dla każdego elementu IRS jest określany zgodnie ze znajomością odpowiadającego mu kaskadowego kanału jako całości, tj. $g_n h_n$, bez konieczności znajomości g_n i h_n indywidualnie. Tę właściwość można wykorzystać do znacznego obniżenia złożoności szacowania kanału. Następnie maksymalny odebrany SNR jest wyrażony jako

$$\gamma_{\max} = \frac{P_t \left|\sum_{n=1}^{N} |g_n||h_n| + |h_d|\right|^2}{\sigma_z^2}. \quad (22)$$

Podstawowym pytaniem dotyczącym osiągalnej wydajności transmisji sygnału wspomaganej przez IRS jest to, w jaki sposób odebrany SNR rośnie w odniesieniu do liczby odbijających elementów N. Przy założeniu niezależnych, identycznie rozłożonych (i.i.d.) kanałów Rayleigha, $\gamma_{\max}$ jest niecentralną zmienną losową chi-kwadrat o jednym stopniu swobody. Gdy N staje się wystarczająco duże, odbite łącza są bardziej dominujące, a bezpośrednie łącze można zaniedbać, tj. przyjmując wartość $|h_d| = 0$ w równaniu (22), aby uzyskać

$$\gamma_{\max} = \frac{P_t \left| \sum_{n=1}^{N} |g_n||h_n| \right|^2}{\sigma_z^2} \quad (23)$$

Zgodnie z centralnym twierdzeniem granicznym mamy

$$\gamma_{\max} \approx N^2 \frac{P_t \pi^2 \sigma_h^2 \sigma_g^2}{16\sigma_z^2}, \quad (24)$$

co oznacza, że zastosowanie IRS przynosi zysk SNR rzędu wielkości N2. Dzieje się tak, ponieważ IRS osiąga pasywny zysk kształtowania wiązki N w łączu $\mathbb{I}$-$\mathbb{U}$ i jednocześnie przechwytuje dodatkowy zysk apertury N w łączu $\mathbb{B}$-$\mathbb{I}$.

Utrata ścieżki iloczynu odległości

Na koniec warto wspomnieć, że kaskadowy kanał IRS podlega utracie ścieżki iloczynu odległości w przeciwieństwie do konwencjonalnej sumy utraty ścieżki odległości w typowym odbiciu lustrzanym. Jak wiemy, współczynniki kanału h_n i g_n są ustalane przez zanikanie na dużą skalę (tj. utratę ścieżki związaną z odległością i zacienianie) oraz zanikanie na małą skalę z powodu propagacji wielościeżkowej. W szczególności utrata ścieżki kanału odbitego przez IRS przechwytuje jego średnią moc i jest zatem niezbędna do analizy budżetu łącza i oceny wydajności. W środowisku propagacji dalekiego pola, w którym IRS znajduje się wystarczająco daleko zarówno od stacji bazowej, jak i użytkownika, odległości łączy $\mathbb{B}$-$\mathbb{I}$ i $\mathbb{I}$-$\mathbb{U}$ można oznaczyć odpowiednio jako d_h i d_g, niezależnie od niewielkiej różnicy między poszczególnymi elementami odbijającymi. Dla uproszczenia wzmocnienie mocy kanału $\mathbb{B}$-$\mathbb{I}$ można oznaczyć wzorem

$$\mathbb{E}\left[|h_n|^2\right] = \sigma_h^2 \propto \frac{1}{d_h^{\alpha_h}}, \quad (25)$$

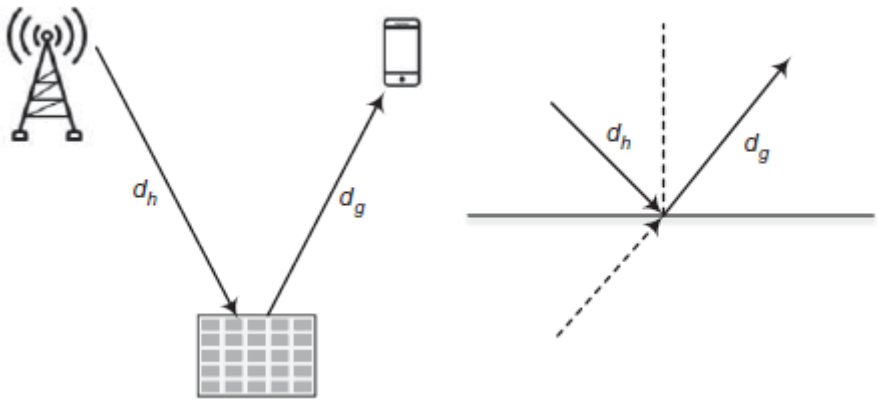
a kanał $\mathbb{I}$-$\mathbb{U}$ jest

$$\mathbb{E}\left[|g_n|^2\right] = \sigma_g^2 \propto \frac{1}{d_g^{\alpha_g}}, \quad (26)$$

ze składowymi strat ścieżki α_h i α_g. Następnie kanał kaskadowy odpowiada

$$\mathbb{E}\left[|h_n g_n|^2\right] = \sigma_h^2 \sigma_g^2 \propto \frac{1}{d_h^{\alpha_h} d_g^{\alpha_g}}. \quad (27)$$

wskazując, że kaskadowy kanał IRS podlega podwójnej utracie ścieżki, zwanej modelem utraty ścieżki odległości produktu. Stąd w praktyce wymagana jest duża liczba elementów odbijających IRS, aby zrekompensować tak poważną utratę mocy z powodu podwójnego tłumienia, poprzez wspólne projektowanie ich współczynników odbicia w celu uzyskania wysokiego pasywnego wzmocnienia kształtowania wiązki. Natomiast moc sygnału odebranego za pośrednictwem nieskończenie długiego, idealnego przewodnika elektrycznego jest odwrotnie proporcjonalna do sumy odległości dwuskokowych łączy, jak pokazano na rysunku tj.



$$\frac{1}{(d_h + d_g)^{\alpha}} \quad (28)$$

gdzie α jest składową stratności ścieżki. W takim modelu stratności ścieżki sumarycznej i odległościowej, odebrana moc odbitego sygnału jest równoważna sytuacji, gdy nadajnik znajduje się w punkcie obrazu oryginalnego nadajnika, przy tej samej odległości propagacji $d_h + d_g$.

Transmisja wieloantenowa wspomagana przez IRS

W tej sekcji dalej badana jest transmisja punkt-punkt wspomagana przez IRS w wąskopasmowym kanale o płaskiej częstotliwości, ale z wieloma antenami w stacji bazowej, gdzie transmisja w dół staje się systemem MISO (Multi-Input Single-Output), a transmisja w górę jest systemem SIMO (Single-Input Multi-Output). W związku z tym aktywne formowanie wiązki w stacji bazowej i pasywne formowanie wiązki w IRS muszą być wspólnie optymalizowane w celu zmaksymalizowania osiągalnej wydajności widmowej. Dla uproszczenia dyskusja koncentruje się tylko na transmisji MISO w dół, ale wyniki mają również zastosowanie do transmisji SIMO w górę. Ponadto staje się to systemem MIMO wspomaganym przez IRS, jeśli terminal użytkownika jest wyposażony w wiele anten. Po omówieniu wspólnego formowania wiązki MISO wspomaganego przez IRS nastąpi krótkie wprowadzenie do systemu MIMO wspomaganego przez IRS.

Wspólne aktywne i pasywne formowanie wiązki

Rozważmy system komunikacji MISO dla pojedynczego użytkownika, składający się ze stacji bazowej z anteną Nb, użytkownika z pojedynczą anteną i IRS z N pasywnymi elementami odbijającymi Aby scharakteryzować teoretyczny zysk wydajnościowy wnoszony przez IRS, zakładamy, że informacje o stanie kanału wszystkich zaangażowanych kanałów są doskonale znane i podążają za quasi-statycznym zanikaniem płaskiej częstotliwości. W stacji bazowej stosuje się liniowe formowanie wiązki oznaczone wektorem transmisji $\mathbf{w} \in \mathbb{C}^{Nb\times 1}$, przy czym $||\mathbf{w}||^2 \leqslant P_t$, gdzie $||\cdot||$ reprezentuje normę euklidesową wektora zespolonego. Następnie dyskretny sygnał równoważny pasma podstawowego odebrany przez użytkownika jest wyrażony jako

$$y = \left(\sum_{n=1}^{N} g_n c_n \mathbf{h}_n^T + \mathbf{h}_d^T\right)\mathbf{w}s + z = (\mathbf{g}^T \mathbf{\Theta} \mathbf{H} + \mathbf{h}_d^T)\mathbf{w}s + z, \quad (29)$$

gdzie $\mathbf{h}_d = [h_{d,1}, h_{d,2},\ldots, h_{d,Nb}]^T \in \mathbb{C}^{Nb\times 1}$ jest wektorem kanałowym od N_b anten stacji bazowych do użytkownika, $\mathbf{h}_n = [h_{n,1}, h_{n,2},\ldots, h_{n,Nb}]^T \in \mathbb{C}^{Nb\times 1}$ jest wektorem kanałowym od N_b anten stacji bazowych do n-tego elementu odbijającego, $\mathbf{g} = [g_1, g_2,\ldots, g_N]^T$, $\mathbf{\Theta} = \text{diag}(c_1, c_2,\ldots, c_N)$, a $\mathbf{H} \in \mathbb{C}^{N\times Nb}$ oznacza macierz

kanałową od stacji bazowej do IRS, gdzie n-ty wiersz tej macierzy jest równy hTn . Poprzez wspólne zaprojektowanie aktywnego kształtowania wiązki w i pasywnego współczynnika odbicia $\boldsymbol{\Theta}$, wydajność widmowa

$$R = \log\left(1 + \frac{|(\mathbf{g}^T\boldsymbol{\Theta}\mathbf{H} + \mathbf{h}_d^T)\mathbf{w}|^2}{\sigma_z^2}\right) \quad (30)$$

można zmaksymalizować, co prowadzi do następującego problemu optymalizacji

$$\begin{aligned} \max_{\boldsymbol{\Theta},\,\mathbf{w}} \quad & |(\mathbf{g}^T\boldsymbol{\Theta}\mathbf{H} + \mathbf{h}_d^T)\mathbf{w}|^2 \\ \text{s.t.} \quad & \|\mathbf{w}\|^2 \leqslant P_t \\ & \beta_n \in [0,1], \quad \forall n = 1,2,\ldots,N \\ & \theta_n \in [0,2\pi), \quad \forall n = 1,2,\ldots,N. \end{aligned} \quad (31)$$

Jak wykazano w poprzedniej sekcji, optymalna wartość tłumienia odbicia wynosi $\beta_n = 1$, $\forall n = 1, 2,\ldots,N$. Następnie wyżej wymieniony problem optymalizacji upraszcza się do

$$\begin{aligned} \max_{\boldsymbol{\theta},\,\mathbf{w}} \quad & |(\mathbf{g}^T\boldsymbol{\Theta}\mathbf{H} + \mathbf{h}_d^T)\mathbf{w}|^2 \\ \text{s.t.} \quad & \|\mathbf{w}\|^2 \leqslant P_t \\ & \theta_n \in [0,2\pi), \quad \forall n = 1,2,\ldots,N. \end{aligned} \quad (32)$$

z $\boldsymbol{\Theta}$ = diag ($e^{j\theta 1}$, $e^{j\theta 2}$,…, $e^{j\theta N}$) / . Niestety, nadal jest to problem optymalizacji niewypukłej, ponieważ jego funkcja celu nie jest wspólnie wklęsła względem $\boldsymbol{\theta}$ i w. W rzeczywistości sprzężenie między zmiennymi optymalizacji formowania wiązki aktywnej i pasywnej jest głównym wyzwaniem w ich wspólnej optymalizacji. Zaproponowano dwa rozwiązania tego problemu tj. relaksację półokreśloną (SDR), która jest stosowana w celu uzyskania wysokiej jakości przybliżenia, a także ograniczenia wydajności, oraz optymalizację naprzemienną, która proponuje naprzemienną optymalizację współczynników formowania wiązki nadawczej i odbijania w sposób iteracyjny.

Rozwiązanie SDR

Można zauważyć, że jeśli współczynniki odbicia $\boldsymbol{\theta}$ są stałe, problem optymalizacji staje się normalną metodą określania optymalnego wektora kształtowania wiązki w typowym wieloantenowym systemie transmisyjnym. Tak więc optymalnym kształtownikiem wiązki może być dopasowany filtr, znany również jako transmisja o maksymalnym współczynniku, który może maksymalizować siłę pożądanego sygnału, mamy

$$\mathbf{w}^\star = \sqrt{P_t}\frac{(\mathbf{g}^T\boldsymbol{\Theta}\mathbf{H} + \mathbf{h}_d^T)^H}{\|\mathbf{g}^T\boldsymbol{\Theta}\mathbf{H} + \mathbf{h}_d^T\|} \quad (33)$$

Podstawiając w⋆ do równania (32), sprowadza się ono do problemu optymalizacji tylko względem $\boldsymbol{\theta}$, tj.

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & \|\mathbf{g}^T\boldsymbol{\Theta}\mathbf{H} + \mathbf{h}_d^T\|^2 \\ \text{s.t.} \quad & \theta_n \in [0,2\pi), \quad \forall n = 1,2,\ldots,N. \end{aligned} \quad (34)$$

Oznaczając $q = [q_1, q_2,\ldots, q_N]^H$, gdzie $q_n = e^{j\theta n}$, ograniczenia w równaniu (34) są równoważne ograniczeniom jednostkowego modułu $|q_n| = 1$, ∀n. Zgodnie z transformacją podaną przez Wu i Zhanga tj. $g^T\mathbf{\Theta}H = q^H\mathbf{\Phi}$, gdzie $\mathbf{\Phi} = \text{diag}(g^T)H \in \mathbb{C}_{N\times Nb}$, mamy

$$\|\mathbf{g}^T\mathbf{\Theta}\mathbf{H} + \mathbf{h}_d^T\|^2 = \|\mathbf{q}^H\mathbf{\Phi} + \mathbf{h}_d^T\|^2 \quad (35)$$

Wówczas równanie (34) jest równoważne

$$\begin{aligned} \max_{\boldsymbol{q}} \quad & \mathbf{q}^H\mathbf{\Phi}\mathbf{\Phi}^H\mathbf{q} + \mathbf{q}^H\mathbf{\Phi}\mathbf{h}_d^* + \mathbf{h}_d^T\mathbf{\Phi}^H\mathbf{q} + \|\mathbf{h}_d\|^2 \\ \text{s.t.} \quad & |q_n| = 1, \quad \forall n = 1, 2, \ldots, N. \end{aligned} \quad (36)$$

Uproszczony problem optymalizacji to nie wypukły kwadratowo ograniczony program kwadratowy, który jest ogólnie NP-trudny do rozwiązania. W szczególności ograniczenia modułu jednostkowego są nie wypukłe i trudne do opanowania, co narzuca kolejne wyzwanie w projektowaniu algorytmu optymalizacji. Dlatego w literaturze zaproponowano różne metody uzyskiwania wysokiej jakości suboptymalnych rozwiązań, w tym SDR z randomizacją Gaussa, AO, gdzie każde z przesunięć fazowych jest optymalizowane w formie zamkniętej, podczas gdy inne są ustalane iteracyjnie, aby zagwarantować lokalnie optymalne, oraz algorytm gałęzi i ograniczeń (BoB) w celu uzyskania globalnie optymalnego rozwiązania , aby wymienić tylko kilka. W poniższej części przedstawiono rozwiązanie SDR jako przykład, podczas gdy zainteresowani czytelnicy mogą zapoznać się z odpowiednią literaturą w celu uzyskania innych rozwiązań. Wprowadzając zmienną pomocniczą t, równanie (36) można zapisać jako

$$\begin{aligned} \max_{\boldsymbol{q}} \quad & \bar{\mathbf{q}}^H\mathbf{R}\bar{\mathbf{q}} + \|\mathbf{h}_d\|^2 \\ \text{s.t.} \quad & |q_n| = 1, \quad \forall n = 1, 2, \ldots, N+1, \end{aligned} \quad (37)$$

z

$$\mathbf{R} = \begin{pmatrix} \mathbf{\Phi}\mathbf{\Phi}^H & \mathbf{\Phi}\mathbf{h}_d^* \\ \mathbf{h}_d^T\mathbf{\Phi}^H & 0 \end{pmatrix} \quad (38)$$

i

$$\bar{\mathbf{q}} = \begin{bmatrix} \mathbf{q} \\ t \end{bmatrix} \quad (39)$$

Definiując $\mathbf{Q} = \bar{\mathbf{q}}\bar{\mathbf{q}}^H$, mamy $\bar{\mathbf{q}}^H\mathbf{R}\bar{\mathbf{q}} = \text{tr}(\mathbf{R}\bar{\mathbf{q}}\bar{\mathbf{q}}^H) = \text{tr}(\mathbf{RQ})$, co musi spełniać $\mathbf{Q} \succcurlyeq 0$ i rangę(Q) = 1. W rezultacie problem optymalizacji zostaje ostatecznie uproszczony do

$$\begin{aligned} \max_{\boldsymbol{Q}} \quad & \text{tr}(\mathbf{RQ}) \\ \text{s.t.} \quad & |q_n| = 1, \quad \forall n = 1, 2, \ldots, N+1, \\ & \mathbf{Q} \succcurlyeq 0, \end{aligned} \quad (40)$$

który staje się wypukłym półokreślonym programem. Rozłóż Q jako $Q = U\mathbf{\Sigma}U^H$, gdzie U jest macierzą unitarną, a $\mathbf{\Sigma}$ jest macierzą diagonalną, obie o rozmiarze (N + 1) × (N + 1). Według Wu i Zhanga [2018] suboptymalne rozwiązanie problemu optymalizacji w równaniu (37) jest podane przez

$$\bar{\mathbf{q}} = \mathbf{U}\boldsymbol{\Sigma}^{1/2}\mathbf{r}, \tag{41}$$

gdzie r jest wektorem losowym wygenerowanym zgodnie z $\mathbf{r} \in \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N+1})$. Ostatecznie rozwiązanie problemu optymalizacji w równaniu (36) można określić jako

$$\mathbf{q} = e^{j\arg\left(\left[\frac{\bar{\mathbf{q}}}{\bar{q}_{N+1}}\right]_{1:N}\right)}, \tag{42}$$

gdzie $[\cdot]_{1:N}$ oznacza podwektor wyodrębniający pierwsze N elementów. Chociaż wspólne aktywne i pasywne kształtowanie wiązki przy użyciu rozwiązania SDR zapewnia dobrą wydajność, wymaga globalnych informacji o stanie kanału, narzucając zaporowe operacje szacowania kanału i narzut wymiany sygnałów, szczególnie gdy liczba anten na stacji bazowej i liczba elementów odbijających w IRS są duże. Ponadto obliczenie optymalnego wektora kształtowania wiązki i współczynników odbijających wiąże się z dużym obciążeniem obliczeniowym, szczególnie w środowiskach o szybkim zanikaniu, w których kanały szybko się zmieniają.

Optymalizacja naprzemienna

Ta technika zapewnia wydajny algorytm obniżający złożoność poprzedniej metody SDR. Kluczowym pomysłem jest optymalizacja współczynników kształtowania wiązki nadawczej i współczynników odbijania naprzemiennie (zamiast łącznie) w sposób iteracyjny, aż do osiągnięcia zbieżności. Biorąc pod uwagę znany wektor kształtowania wiązki nadawczej w_0, problem optymalizacji w równaniu (32) jest uproszczony do

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & |(\mathbf{g}^T\boldsymbol{\Theta}\mathbf{H} + \mathbf{h}_d^T)\mathbf{w}_0|^2 \\ \text{s.t.} \quad & \theta_n \in [0, 2\pi), \quad \forall n = 1, 2, \ldots, N \end{aligned} \tag{43}$$

Równanie optymalizacji nadal nie jest wypukłe, ale może umożliwić rozwiązanie w formie zamkniętej, wykorzystując następującą nierówność

$$|(\mathbf{g}^T\boldsymbol{\Theta}\mathbf{H} + \mathbf{h}_d^T)\mathbf{w}_0| \leqslant |\mathbf{g}^T\boldsymbol{\Theta}\mathbf{H}\mathbf{w}_0| + |\mathbf{h}_d^T\mathbf{w}_0|. \tag{44}$$

Równość zachodzi wtedy i tylko wtedy, gdy $\arg(\mathbf{g}^T\boldsymbol{\Theta}\mathbf{H}\mathbf{w}_0) = \arg(\mathbf{h}_d^T\mathbf{w}_0) \triangleq \varphi_0$, gdzie $\arg(\cdot)$ oznacza fazę liczby zespolonej lub fazę składową wektora zespolonego. Oznaczając $\mathbf{q} = [q_1, q_2, \ldots, q_N]^H$ przy $q_n = e^{j\theta n}$ i $\boldsymbol{\chi} = \mathrm{diag}(\mathbf{g}^T)\mathbf{H}\mathbf{w}_0 \in \mathbb{C}^{N\times 1}$, mamy $\mathbf{g}^T\boldsymbol{\Theta}\mathbf{H}\mathbf{w}_0 = \mathbf{q}^H\boldsymbol{\chi} \in \mathbb{C}$. Następnie, ignorując stały człon $|\mathbf{h}_d^T\mathbf{w}_0|$, równanie (43) przekształcamy do

$$\begin{aligned} \max_{\mathbf{q}} \quad & |\mathbf{q}^H\boldsymbol{\chi}| \\ \text{s.t.} \quad & |q_n| = 1, \quad \forall n = 1, 2, \ldots, N \\ & \arg(\mathbf{q}^H\boldsymbol{\chi}) = \varphi_0. \end{aligned} \tag{45}$$

Łatwo jest uzyskać rozwiązanie optymalne maksymalizujące funkcję celu, tj.

$$\mathbf{q}^* = e^{j(\varphi_0 - \arg(\boldsymbol{\chi}))} = e^{j(\varphi_0 - \arg(\mathrm{diag}(\mathbf{g}^T)\mathbf{H}\mathbf{w}_0))}, \tag{46}$$

W związku z tym optymalne przesunięcie fazowe dla n-tego elementu odbijającego jest podane wzorem

$$\begin{aligned}\theta_n^\star &= \varphi_0 - \arg(g_n \mathbf{h}_n^T \mathbf{w}_0) \\ &= \varphi_0 - \arg(g_n) - \arg(\mathbf{h}_n^T \mathbf{w}_0),\end{aligned} \quad (47)$$

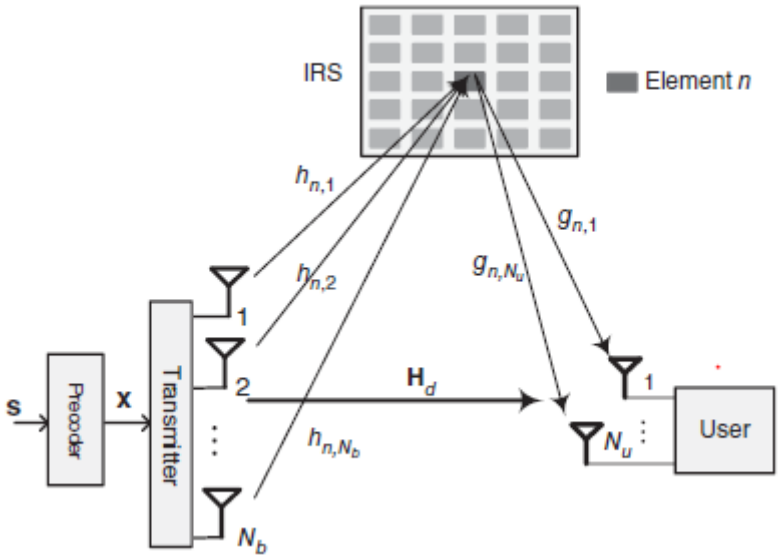
gdzie $h^T_n w_0 \in \mathbb{C}$ można uważać za efektywny kanał pojedynczego wejścia i pojedynczego wyjścia (SISO) odbierany przez n-ty element odbijający łączący efekty formowania wiązki nadawczej w0 i kanału od stacji bazowej do elementu odbijającego h_n, a gn oznacza współczynnik kanału od n-tego elementu odbijającego do użytkownika. W związku z tym równanie (47) oznacza, że optymalne przesunięcie fazowe powinno być dostrojone tak, aby faza sygnału przechodzącego przez łącza $\mathbb{B} - \mathbb{I}$ i $\mathbb{I} - \mathbb{U}$ była kompensowana, a faza resztkowa była wyrównana z fazą sygnału przez łącze bezpośrednie w celu uzyskania spójnego łączenia w odbiorniku. Gdy znane są współczynniki odbicia, rozwiązanie AO jest naprzemiennie modyfikowane w celu optymalizacji w przy danym $\theta^\star_n$ lub równoważnie przy znajomości całego kanału $g^T\mathbf{\Theta}H + h^T_d$. Optymalnym kształtownikiem wiązki może być dopasowany filtr, który może maksymalizować siłę pożądanego sygnału, tj.

$$\mathbf{w}^\star = \sqrt{P_t}\frac{(\mathbf{g}^T\mathbf{\Theta}\mathbf{H} + \mathbf{h}_d^T)^H}{\|\mathbf{g}^T\mathbf{\Theta}\mathbf{H} + \mathbf{h}_d^T\|} \quad (48)$$

Następnie rozwiązanie AO jest ponownie naprzemiennie optymalizowane w celu optymalizacji $\boldsymbol{\theta}$ przy danym $w^\star$. Ten proces jest powtarzany, aż do osiągnięcia zbieżności. To naprzemienne podejście optymalizacyjne jest praktycznie sensowne, ponieważ zarówno formowanie wiązki nadawczej, jak i przesunięcia fazowe są uzyskiwane w wyrażeniach o zamkniętej formie.

Wspólne kodowanie wstępne i odbijanie

Poprzednie dyskusje koncentrowały się głównie na systemach SISO lub MISO, w których sprzęt użytkownika ma tylko jedną antenę. W bardziej ogólnym przypadku MIMO wspomaganego przez IRS z wieloma antenami zarówno na stacji bazowej, jak i u użytkownika, wymagana jest wspólna optymalizacja współczynników odbijających IRS i macierzy kowariancji transmisji MIMO lub macierzy kodowania wstępnego. Jest to trudniejsze niż konwencjonalny system MIMO bez IRS. Model systemu i formuła optymalizacji komunikacji MIMO wspomaganej przez IRS zostaną przedstawione w następujący sposób. Rozważmy system komunikacji MIMO typu punkt-punkt składający się ze stacji bazowej z antenami Nb, pojedynczego użytkownika z antenami N_u i IRS z pasywnymi elementami odbijającymi N. Załóżmy quasi-statyczne blokowanie zanikania w kanałach o płaskiej częstotliwości i rozważmy jeden konkretny blok zanikania, w którym informacja o stanie kanału jest doskonale znana i pozostaje w przybliżeniu stała. Następnie, jak pokazano na rysunku ,



model systemu równoważnego pasma podstawowego w czasie dyskretnym można wyrazić jako

$$\mathbf{y} = \left(\sum_{n=1}^{N} \mathbf{g}_n e^{j\theta_n} \mathbf{h}_n^T + \mathbf{H}_d \right) \mathbf{x} + \mathbf{z}, \quad (49)$$

gdzie $y = [y_1, y_2,\ldots, y_{Nu}]^T \in \mathbb{C}^{Nu\times 1}$ oznacza odebrany wektor symboli u użytkownika, $H_d \in \mathbb{C}^{Nu\times Nb}$ jest macierzą kanałową bezpośredniego łącza od stacji bazowej do użytkownika, $h_n = [h_{n,1}, h_{n,2},\ldots, h_{n,Nb}]^T \in \mathbb{C}^{Nb\times 1}$ jest wektorem kanałowym od N_b anten stacji bazowej do n-tego elementu odbijającego, $g_n = [g_{n,1}, g_{n,2},\ldots, g_{n,Nu}]^T \in \mathbb{C}^{Nu\times 1}$ wyraża wektor kanałowy od n-tego elementu odbijającego do N_u anten użytkownika, θ_n jest indukowanym obrotem fazy na n-tym elemencie odbijającym (przyjmując maksymalną amplitudę $\beta_n = 1$), $z = [z_1, z_2,\ldots, z_{Nu}]^T \in \mathbb{C}^{Nu\times 1}$ oznacza wektor AWGN spełniający $\mathbf{z} \sim \mathcal{CN}\left(\mathbf{0}, \sigma_z^2 \mathbf{I}_{N_u}\right)$, a $x = [x_1, x_2,\ldots, xNb]^T \in \mathbb{C}^{Nb\times 1}$ oznacza wektor symboli przesyłanych, z ograniczeniem mocy przesyłania $\mathbb{E}[x^H x] \leqslant P_t$. Macierz kowariancji przesyłania jest zdefiniowana przez $Q = \mathbb{E}[xx^H] \in \mathbb{C}^{Nb\times Nb}$, a ograniczenie mocy można również wyrazić jako $tr(Q) \leqslant Pt$. Równoważnie, model systemu można również wyrazić w postaci macierzowej jako

$$\mathbf{y} = (\mathbf{G}\boldsymbol{\Theta}\mathbf{H} + \mathbf{H}_d)\mathbf{x} + \mathbf{z}, \quad (50)$$

gdzie macierz odbijająca diagonalnie jest zapisana jako $\boldsymbol{\Theta} = \text{diag}\{e^{j\theta 1}, e^{j\theta 2},\ldots, e^{j\theta N}\}$, a $H \in \mathbb{C}^{N\times Nb}$ oznacza macierz kanału od stacji bazowej do IRS, gdzie n-ty wiersz tej macierzy jest równy h^T_n, a $G \in \mathbb{C}^{Nu\times N}$ oznacza macierz kanału od IRS do użytkownika, wyrażoną przez $G = [g_1, g_2,\ldots, g_N]$. Traktując $\tilde{\mathbf{H}} = G\boldsymbol{\Theta}H + H_d$ jako efektywną macierz kanału w konwencjonalnym systemie MIMO, uzyskuje się pojemność kanału systemu MIMO wspomaganego przez IRS, która jest dana przez

$$\begin{aligned} C &= \log_2 \det\left(\mathbf{I}_{N_u} + \frac{\tilde{\mathbf{H}}\mathbf{Q}\tilde{\mathbf{H}}^H}{\sigma_z^2} \right) \\ &= \log_2 \det\left(\mathbf{I}_{N_u} + \frac{\left(\mathbf{G}\boldsymbol{\Theta}\mathbf{H} + \mathbf{H}_d\right) \mathbf{Q} \left(\mathbf{G}\boldsymbol{\Theta}\mathbf{H} + \mathbf{H}_d\right)^H}{\sigma_z^2} \right) \end{aligned} \quad (51)$$

W odróżnieniu od konwencjonalnego kanału MIMO bez IRS, dla którego pojemność jest określana wyłącznie przez macierz kanału Hd, pojemność kanału MIMO wspomaganego przez IRS zależy również od macierzy odbicia IRS $\boldsymbol{\Theta}$, ponieważ wpływa ona na efektywną macierz kanału $\tilde{\mathbf{H}} = \mathbf{G}\boldsymbol{\Theta}\mathbf{H} + \mathbf{H}_d$ a także na wynikową optymalną macierz kowariancji transmisji Q. Aby zmaksymalizować pojemność kanału MIMO wspomaganego przez IRS, musimy wspólnie zoptymalizować macierz odbicia IRS i macierz kowariancji transmisji, z zastrzeżeniem unimodularnych ograniczeń współczynników odbicia i ograniczenia mocy sumy w nadajniku. W rezultacie problem optymalizacji jest sformułowany jako

$$\begin{aligned} \max_{\boldsymbol{\Theta},\, \mathbf{Q}} \quad & \log_2 \det \left(\mathbf{I}_{N_u} + \frac{\left(\mathbf{G}\boldsymbol{\Theta}\mathbf{H} + \mathbf{H}_d\right) \mathbf{Q} \left(\mathbf{G}\boldsymbol{\Theta}\mathbf{H} + \mathbf{H}_d\right)^H}{\sigma_z^2} \right) \\ \text{s.t.} \quad & \theta_n \in [0, 2\pi), \quad \forall n = 1, 2, \ldots, N, \\ & \operatorname{tr}(\mathbf{Q}) \leqslant P_t. \end{aligned} \tag{52}$$

Jest to problem optymalizacji niewypukłej, ponieważ można wykazać, że funkcja celu jest niewklęsła nad macierzą odbicia, a ograniczenie unimodularne na każdym współczynniku odbicia jest również niewypukłe. Ponadto macierz kowariancji transmisji Q jest sprzężona z $\boldsymbol{\Theta}$ w funkcji celu, co utrudnia rozwiązanie tej optymalizacji. W Zhang i Zhang zaproponowano algorytm naprzemienny do rozwiązania tego problemu optymalizacji. Konkretnie, funkcja celu jest najpierw przekształcana do bardziej łatwej do rozwiązania formy pod względem zmiennych optymalizacji Q i θ_n, ∀n, na podstawie których rozwiązujemy następnie dwa podproblemy, aby zoptymalizować odpowiednio macierz kowariancji transmisji Q lub jeden współczynnik odbicia θ_n, ∀n przy wszystkich pozostałych zmiennych ustalonych. Optymalne rozwiązania obu podproblemów mają formę zamkniętą, co umożliwia wydajnemu algorytmowi optymalizacji naprzemiennej uzyskanie lokalnie optymalnego rozwiązania poprzez iteracyjne rozwiązywanie tych podproblemów. Oprócz optymalizacji macierzy kowariancji transmisji Q, istnieje alternatywne podejście polegające na maksymalizacji wydajności poprzez wspólne odbicie i wstępne kodowanie projektu. Niech x = Ps, gdzie $s \in \mathbb{C}^{Ns\times 1}$ oznacza wektor N_s symboli informacyjnych przesyłanych jednocześnie przez kanał, spełniający $\mathbb{E}[ss^H] = I_{Ns}$, a $P \in \mathbb{C}^{Nt\times Ns}$ oznacza macierz wstępnego kodowania używaną do kodowania Ns symboli informacyjnych w Nb przesyłanych symboli. Następnie model systemu można zapisać jako

$$\mathbf{y} = \left(\sum_{n=1}^{N} \mathbf{g}_n e^{j\theta_n} \mathbf{h}_n^T + \mathbf{H}_d \right) \mathbf{P}\mathbf{s} + \mathbf{z} = (\mathbf{G}\boldsymbol{\Theta}\mathbf{H} + \mathbf{H}_d)\mathbf{P}\mathbf{s} + \mathbf{z} \tag{53}$$

W tym przypadku problem optymalizacji staje się procesem znajdowania optymalnej macierzy prekodowania i macierzy współczynników odbicia. Dlatego problem optymalizacji można zapisać jako

$$\begin{aligned} \max_{\boldsymbol{\Theta},\, \mathbf{P}} \quad & \log_2 \det \left(\mathbf{I}_{N_u} + \frac{\left(\mathbf{G}\boldsymbol{\Theta}\mathbf{H} + \mathbf{H}_d\right) \mathbf{P}\mathbf{P}^H \left(\mathbf{G}\boldsymbol{\Theta}\mathbf{H} + \mathbf{H}_d\right)^H}{\sigma_z^2} \right) \\ \text{s.t.} \quad & \theta_n \in [0, 2\pi), \quad \forall n = 1, 2, \ldots, N, \\ & \operatorname{tr}(\mathbf{P}\mathbf{P}^H) \leqslant P_t. \end{aligned} \tag{54}$$

Oprócz maksymalizacji przepustowości kanału, macierz kowariancji transmisji lub macierz prekodowania można zoptymalizować w celu poprawy innych metryk wydajności. Na przykład Ye i inni proponują wspólną optymalizację prekodowania i odbijania w celu zminimalizowania współczynnika błędów symboli dla komunikacji MIMO wspomaganej przez IRS.

Podwójna wiązka inteligentnej powierzchni odbijającej

Korzystając z hybrydowego formowania wiązki, stacja bazowa (BS) może generować parę niezależnych wiązek w kierunku odpowiednio IRS i sprzętu użytkownika (UE). W rezultacie optymalne fazy odbijające są bezpośrednio obliczane na podstawie szacowanych informacji o stanie kanału (CSI), niezależnie od aktywnego formowania wiązki. W związku z tym optymalizacja pasywnego i aktywnego formowania wiązki jest rozdzielona, co skutkuje uproszczoną konstrukcją systemu. W przeciwieństwie do wspólnej optymalizacji pasywnego i aktywnego formowania wiązki, można uniknąć dużego obciążenia obliczeniowego i dużego opóźnienia spowodowanego iteracyjną optymalizacją. Ponadto zasięg jest optymalizowany pod kątem wydajności krawędzi komórki i środka komórki.

Podwójne wiązki w hybrydowym formowaniu wiązki

Istnieją trzy główne struktury formowania wiązki: cyfrowa, analogowa i hybrydowa cyfrowo-analogowa. Implementacja cyfrowego formowania wiązki w dużej macierzy, np. w transceiverze mmWave lub THz, wymaga wielu komponentów RF, np. wzmacniacze dużej mocy, co prowadzi do wysokich kosztów sprzętu i zużycia energii. To ograniczenie spowodowało zastosowanie analogowego kształtowania wiązki, wykorzystującego tylko jeden łańcuch RF. Analogowe kształtowanie wiązki jest implementowane jako podejście de facto dla wewnętrznych systemów mmWave. Jednak obsługuje ono tylko transmisję pojedynczego strumienia i cierpi z powodu upośledzenia sprzętowego analogowych przesuwników fazowych. W konsekwencji hybrydowe kształtowanie wiązki zostało zaproponowane jako wydajne podejście do obsługi transmisji wielostrumieniowej z wykorzystaniem tylko kilku łańcuchów RF i sieci przesuwników fazowych. W porównaniu z analogowym kształtowaniem wiązki, hybrydowe kształtowanie wiązki obsługuje multipleksowanie przestrzenne, różnorodność i dostęp wielokrotny z podziałem przestrzennym. Osiąga wydajność widmową porównywalną z cyfrowym kształtowaniem wiązki przy znacznie niższej złożoności sprzętowej i niższym koszcie. Hybrydowe kształtowanie wiązki można implementować za pomocą różnych form sieci obwodów, co skutkuje dwiema podstawowymi strukturami:

● W pełni połączone hybrydowe kształtowanie wiązki

Nadawane dane są najpierw wstępnie kodowane w paśmie podstawowym do strumieni Mdata, a każdy strumień jest przetwarzany przez niezależny łańcuch RF. Każdy łańcuch RF jest połączony ze wszystkimi N antenami za pośrednictwem sieci analogowej, gdzie $N \gg M$. W związku z tym wymagane są łącznie MN analogowych przesuwników fazowych.

● Częściowo połączone hybrydowe kształtowanie wiązki

Każdy łańcuch RF jest połączony tylko z podzbiorem wszystkich elementów anteny zwanym podmacierzą. Ta struktura jest preferowana z perspektywy praktyki, ponieważ znacznie obniża złożoność sprzętu (a także zużycie energii) poprzez drastyczne zmniejszenie liczby analogowych przesuwników fazowych z MN do N.

Bez utraty ogólności, w tej części skupiamy się na częściowo połączonym hybrydowym kształtowaniu wiązki w celu prostej analizy, ale jest ona również stosowalna do całkowicie połączonego (FC) hybrydowego kształtowania wiązki. Matematycznie wyjście prekodera pasma podstawowego jest oznaczone jako $s_m[t]$, $1 \le m \le M$. Po konwersji cyfrowo-analogowej i konwersji w górę m-ty łańcuch RF zasila

$$\Re\left[s_m[t]e^{j2\pi f_0 t}\right], \quad 1 \le m \le M, \tag{55}$$

do sieci analogowej, gdzie $\Re[\cdot]$ oznacza część rzeczywistą liczby zespolonej, a f_0 oznacza częstotliwość nośną. W częściowo połączonym formowaniu wiązki, tablica jest podzielona na kilka pod-tablic, a każda antena jest przydzielona tylko do jednego łańcucha RF. Każda pod-tablica zawiera $N_s = N/M$ elementów (załóżmy, że Ns jest liczbą całkowitą). Oznacz ważone przesunięcie fazowe na n-tej antenie m-tej pod-tablicy przez ψ_{nm}, gdzie $1 \le n \le N_s$ i $1 \le m \le M$ aby oznaczyć, a zatem sygnały transmisyjne m-tej pod-tablicy są wyrażone przez

$$\mathbf{s}_m(t) = \left[\Re\left[s_m[t]e^{j2\pi f_0 t}e^{j\psi_{1m}}\right], \ldots, \Re\left[s_m[t]e^{j2\pi f_0 t}e^{j\psi_{N_s m}}\right]\right]^T \quad (56)$$

Promieniejąc falę płaską do jednorodnego medium w kierunku wskazanym przez kąt wyjścia θ, różnica czasu między typowym elementem n m-tego podukładu a punktem odniesienia jest oznaczona jako $\tau_{nm}(\theta)$. W kanale bezprzewodowym o płaskim zanikaniu odebrany sygnał pasma przepustowego jest

$$\begin{aligned} y(t) &= \sum_{m=1}^{M}\sum_{n=1}^{N_s}\Re\left[h_m(t)s_m[t]e^{j2\pi f_0(t-\tau_{nm}(\theta))}e^{j\psi_{nm}}\right] + n(t) \\ &= \Re\left[\left(\sum_{m=1}^{M}h_m(t)s_m[t]\sum_{n=1}^{N_s}e^{-j2\pi f_0\tau_{nm}(\theta)}e^{j\psi_{nm}}\right)e^{j2\pi f_0 t}\right] + n(t), \end{aligned} \quad (57)$$

gdzie $h_m(t)$ reprezentuje odpowiedź kanału między m-tym podzestawem a odbiornikiem, a n(t) szum . Oznaczając wektor sterujący podzestawu m jako

$$\mathbf{a}_m(\theta) = \left[e^{-j2\pi f_0\tau_{1m}(\theta)}, e^{-j2\pi f_0\tau_{2m}(\theta)}, \ldots, e^{-j2\pi f_0\tau_{N_s m}(\theta)}\right]^T \quad (58)$$

i jego wektor ważenia (spowodowany przesunięciem fazy analogowej) jako

$$\mathbf{w}_m = \left[e^{j\psi_{1m}}, e^{j\psi_{2m}}, \ldots, e^{j\psi_{N_s m}}\right]^T, \quad (59)$$

Równanie (57) można zapisać w postaci

$$\begin{aligned} r(t) &= \Re\left[\left(\sum_{m=1}^{M}h_m(t)s_m[t]\mathbf{a}_m^T(\theta)\mathbf{w}_m\right)e^{j2\pi f_0 t}\right] + n(t), \\ &= \Re\left[\left(\sum_{m=1}^{M}h_m(t)s_m[t]B_m(\theta,t)\right)e^{j2\pi f_0 t}\right] + n(t), \end{aligned} \quad (60)$$

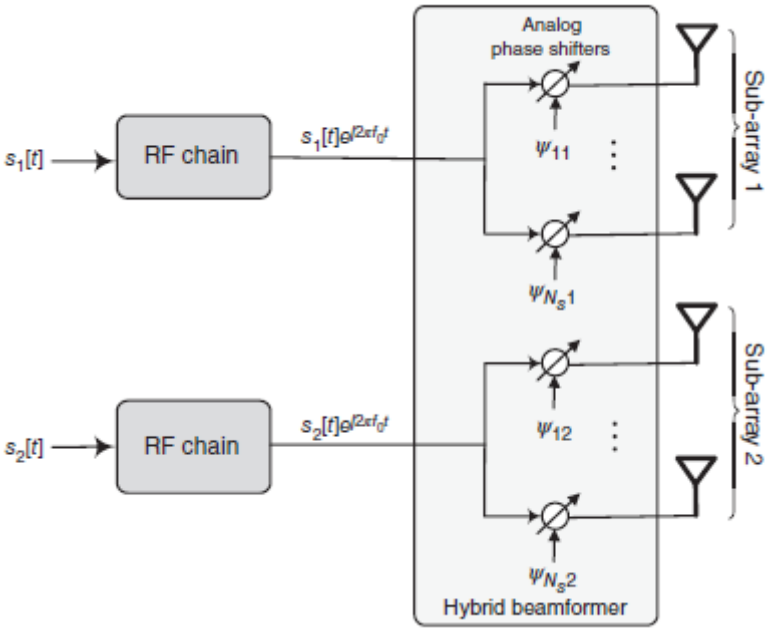
ze wzorem wiązki m-tego podzestawu:

$$B_m(\theta,t) = \mathbf{a}_m^T(\theta)\mathbf{w}_m = \sum_{n=1}^{N_s}e^{-j2\pi f_0\tau_{nm}(\theta)}e^{j\psi_{nm}} \quad (61)$$

Po konwersji w dół i próbkowaniu w odbiorniku, równoważny sygnał otrzymany w paśmie podstawowym można uprościć do postaci (pomijając indeks czasu dla uproszczenia):

$$r = \sum_{m=1}^{M}h_m B_m(\theta)s_m + n, \quad (62)$$

gdzie r jest odebranym symbolem, sm jest zmodulowanym symbolem dla m-tego łańcucha RF, a h_m oznacza współczynnik kanału między anteną odniesienia m-tego podzestawu a odbiornikiem. Nie tracąc ogólności, używamy częściowo połączonej struktury z dwoma odgałęzieniami w dalszej części, aby przedstawić i przeanalizować schemat IRS z podwójną wiązką. Jak pokazano na rysunku,



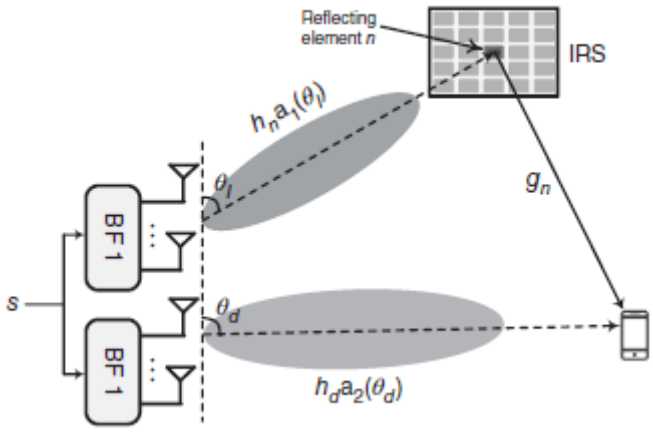
para wiązek o wzorach $B_1(\theta)$ i $B_2(\theta)$ nad hybrydowym kształtownikiem wiązki. Po konwersji w dół i próbkowaniu w odbiorniku, odebrany sygnał pasma podstawowego można uprościć do

$$r = h_1 B_1(\theta_1) s_1 + h_2 B_2(\theta_2) s_2 + n, \quad (63)$$

gdzie h_1 i h_1 oznaczają współczynniki kanału od anteny odniesienia pierwszego i drugiego podzestawu do UE, a θ_1 i θ_2 oznaczają kąty odejścia (DOA) sygnałów przesyłanych przez podzestawy 1 i 2, odpowiednio.

IRS z podwójną wiązką

Podstawową zasadą DB-IRS jest skierowanie jednej wiązki w stronę użytkownika, podobnie jak w konwencjonalnym systemie formowania wiązki, podczas gdy druga wiązka jest używana do skupiania energii na IRS, jak pokazano na rysunku.



Ponieważ zarówno lokalizacja IRS, jak i BS jest celowo wybierana i ustalana, zwykle nie ma żadnej blokady pomiędzy nimi, a informacja DOA θ_I jest deterministyczna. Wiązka skierowana w stronę IRS jest ustalona, co znacznie upraszcza implementację systemu. Odstęp między antenami w hybrydowym formowaniu wiązki jest niewielki, zwykle wynosi połowę długości fali d = $\lambda/2$. W rezultacie anteny są silnie skorelowane. Bez utraty ogólności zakładamy, że pierwsza podtablica obsługuje IRS, a druga podtablica obsługuje UE. Wektor kanału z pierwszej podtablicy do n-tego elementu odbijającego można wyrazić jako $h_n = h_n a_1(\theta_I) \in \mathbb{C}^{Ns \times 1}$, gdzie h_n oznacza odpowiedź kanału z anteny odniesienia tej podtablicy na n-ty element odbijający. Podobnie wektor kanału z drugiej podtablicy do UE jest podany przez $h_d = h_d a_2(\theta_d)$, gdzie h_d oznacza odpowiedź kanału z anteny odniesienia drugiej podtablicy do UE, a θ_d jest DOA użytkownika. Piszemy w_1 i w_2, aby odpowiednio oznaczyć wektory ważenia dwóch podtablic. Tymczasem zakładamy, że dwie podtablice mają wspólny symbol transmisji s, przy czym $\mathbb{E}[|s|^2]= 1$. Obserwację u użytkownika w systemie DB-IRS można podać za pomocą

$$\begin{aligned} r &= \sqrt{P_t}\left(\sum_{n=1}^{N} g_n c_n \mathbf{h}_n^T \mathbf{w}_1 + \mathbf{h}_d^T \mathbf{w}_2\right) s + n \\ &= \sqrt{P_t}\left(\sum_{n=1}^{N} g_n c_n h_n \mathbf{a}_1^T(\theta_I)\mathbf{w}_1 + h_d \mathbf{a}_2^T(\theta_d)\mathbf{w}_2\right) s + n \\ &= \sqrt{P_t}\left(\sum_{n=1}^{N} g_n c_n h_n B_r(\theta_I) + h_d B_d(\theta_d)\right) s + n, \end{aligned} \tag{64}$$

ze wzorami wiązki odpowiadającymi IRS i UE $B_r(\theta) = a^T_1(\theta)w_1$ i $B_d(\theta) = a^T_2(\theta)w_2$, odpowiednio.

Projekt optymalizacji

W związku z tym otrzymany SNR u użytkownika jest podany przez

$$\gamma = \frac{P_t \left| \sum_{n=1}^{N} g_n c_n h_n B_r(\theta_I) + h_d B_d(\theta_d) \right|^2}{\sigma_n^2} \tag{65}$$

Celem projektu optymalizacyjnego jest maksymalizacja wydajności widmowej R = log (1 + γ) poprzez dobór optymalnych wektorów ważenia i współczynników odbicia, co skutkuje następującym problemem optymalizacyjnym

$$\begin{aligned}\max_{\boldsymbol{\Theta},\mathbf{w}_m} \quad & \left|\sum_{n=1}^{N} g_n c_n h_n B_r(\theta_I) + h_d B_d(\theta_d)\right|^2 \\ \text{s.t.} \quad & \|\mathbf{w}_m\|^2 \leqslant 1, \quad m = 1, 2, \\ & \phi_n \in [0, 2\pi), \quad \forall n = 1, 2, \ldots, N.\end{aligned} \tag{66}$$

Co najważniejsze, $\boldsymbol{\Theta}$ i w_m,m = 1, 2 w tym problemie optymalizacji jest odsprzęgnięte dzięki podejściu podwójnej wiązki. Zatem $\boldsymbol{\Theta}$ i wm, m = 1, 2 można optymalizować niezależnie, w przeciwieństwie do optymalizacji wspólnej w (32). Ponadto określenie w_1 i w_2 jest również niezależne. Aby uzyskać optymalną wiązkę w kierunku IRS, musimy rozwiązać

$$\begin{aligned}\max_{\mathbf{w}_1} \quad & |\mathbf{a}_1^T(\theta_I)\mathbf{w}_1|^2 \\ \text{s.t.} \quad & \|\mathbf{w}_1\|^2 \leqslant 1.\end{aligned} \tag{67}$$

Musimy znać informacje DOA, które można efektywnie oszacować za pomocą klasycznych algorytmów, takich jak klasyfikacja sygnałów wielokrotnych (MUSIC i ESPRINT). Biorąc pod uwagę szacowany θI , optymalny wektor ważenia to w1 =√1⁄N_sa*$_{1(}\theta_I$), co daje

$$B_r(\theta_I) = \sqrt{\frac{1}{N_s}}\mathbf{a}_1^T(\theta_I)\mathbf{a}_1^*(\theta_I) = \sqrt{N_s}. \tag{68}$$

Ponieważ BS i IRS są stałe, θ_d jest łatwiejsze do uzyskania i pozostaje stałe w długim okresie. Podobnie, bezpośrednie wzmocnienie wiązki $B_d(\theta_d)$ = √ N_s, jeśli wektor ważenia jest ustawiony na

$$\mathbf{w}_2 = \sqrt{\frac{1}{N_s}}\mathbf{a}_2^*(\theta_d). \tag{69}$$

Dla uproszczenia zakładamy, że płaty boczne dwóch wzorów wiązki w kierunku przeciwnym są pomijalne, tj. $|B_d(\theta_I)|$ = 0 i $|B_r(\theta_d)|$ = 0. Następnie problem optymalizacji w (66) zostaje zredukowany do

$$\begin{aligned}\max_{\boldsymbol{\Theta}} \quad & N_s\left|\sum_{n=1}^{N} g_n c_n h_n + h_d\right|^2 \\ \text{s.t.} \quad & \phi_n \in [0, 2\pi), \quad \forall n = 1, 2, \ldots, N,\end{aligned} \tag{70}$$

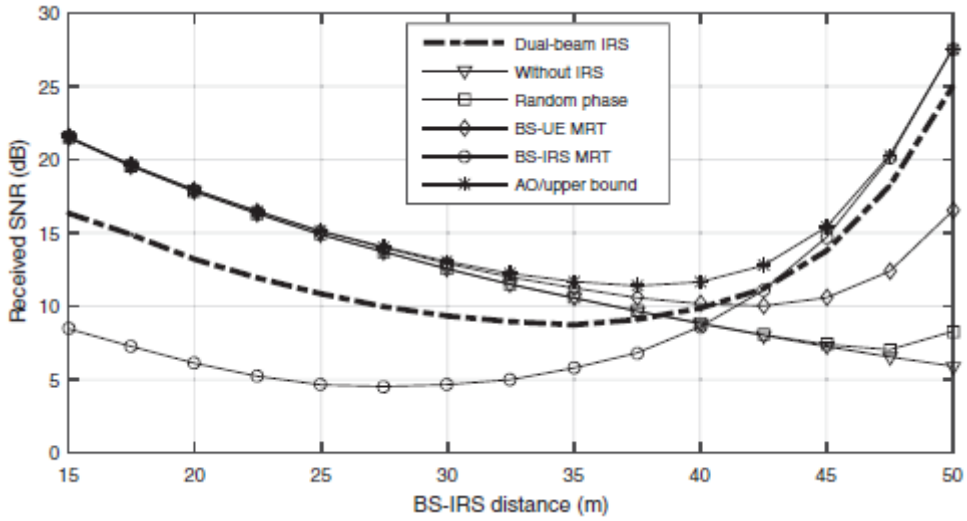
co jest równoważne systemowi IRS z pojedynczą anteną BS. Dlatego optymalne przesunięcie fazowe dla odbijającego elementu n jest równe

$$\phi_n^\star = \text{mod}\,[\phi_d - (\phi_{h,n} + \phi_{g,n}), 2\pi], \tag{71}$$

gdzie $\phi_{h,n}$, $\phi_{g,n}$ i ϕ_d oznaczają fazy h_n, g_n i h_d, odpowiednio, a mod [·] jest operacją modulo. Biorąc pod uwagę optymalne wiązki w (68) i (69) oraz optymalne przesunięcia fazowe, otrzymany SNR dwuwiązkowego systemu IRS, jak podano w (65), jest maksymalizowany, równy

$$\gamma_{\max} = \frac{N_s P_t\left|\sum_{n=1}^{N} |g_n||h_n| + |h_d|\right|^2}{\sigma_n^2} \tag{72}$$

implikując aktywne wzmocnienie formowania wiązki $N_s = N_b/2$ i pasywne wzmocnienie formowania wiązki N^2. W porównaniu ze wspólną optymalizacją formowania wiązki aktywnej i pasywnej, która wymaga iteracyjnego procesu optymalizacji, obliczenie parametrów IRS dla dwóch wiązek jest zasadniczo proste. W skrócie, system DB-IRS musi jedynie oszacować DOA użytkownika, a następnie skierować wiązkę w jego kierunku. Wiązka w kierunku IRS jest ustalona, co wymaga działania w perspektywie długoterminowej, ponieważ lokalizacje IRS i BS są ustalone i znane. Tymczasem optymalne przesunięcia fazowe dla IRS są uzyskiwane bezpośrednio w kategoriach uzyskanego CSI, jako (19). Aby zapewnić wgląd w tę technikę, Rysunek przedstawia porównanie wydajności odbioru SNR między kilkoma konfiguracjami transmisji: (1) Optymalizacja naprzemienna dla wspólnego aktywnego i pasywnego formowania wiązki.



Liczba iteracji jest ustawiona na trzy, co jest wystarczające do uzyskania zbieżności studni. Ponieważ AO osiąga optymalne wyniki tak samo jak schemat SDR lub górna granica wydajności, pozostałe dwa nie są pokazane na rysunku dla uproszczenia. (2) BS-UE transmisja o maksymalnym współczynniku (MRT), która ustawia $w^\star = h^*_d/||h_d||$ w celu osiągnięcia dopasowanego filtrowania w kategoriach bezpośredniego łącza $\mathbb{B} - \mathbb{U}$; (3) BS-IRS MRT, która ustawia $w^\star = h^*_n/||h_n||$ w celu osiągnięcia dopasowanego filtrowania w oparciu o łącze $\mathbb{B} - \mathbb{I}$, podczas gdy optymalne fazy odbijające są obliczane odpowiednio. Ponieważ to łącze jest LOS, lub kanałem pierwszego rzędu, h_n może być dowolnym wierszem H. (4) Losowe przesunięcia fazowe, gdzie θ_n, $\forall n$ jest wybierane losowo, a następnie przeprowadzane jest MRT w BS w oparciu o połączony kanał, tj. $w^\star = (g^T\mathbf{\Theta}H+h^T_d)^* / ||g^T\mathbf{\Theta}H+h^T_d||$. (5)Schemat porównawczy w systemie MISO bez pomocy IRS przez ustawienie $w^\star = h^*_d/||hd||$. W konwencjonalnym systemie bez IRS użytkownik na skraju komórki cierpi z powodu niskiego SNR z powodu poważnej utraty propagacji. Problem na skraju komórki jest łagodzony przez wdrożenie IRS, jak pokazano na rysunku powyżej, gdzie sieci bezprzewodowe wspomagane przez IRS wykazują porównywalną wydajność na skraju komórki jak w centrum komórki. Dzieje się tak, ponieważ użytkownik jest dalej od BS, a bliżej IRS, dzięki czemu może uzyskać silne odbite sygnały. Oznacza to, że zasięg systemu można skutecznie rozszerzyć, wdrażając pasywny IRS oprócz BS lub aktywnego przekaźnika. W szczególności schemat MRT BS-UE działa prawie optymalnie, gdy UE znajduje się w centrum komórki, podczas gdy cierpi na znaczną utratę SNR na krawędzi komórki. Dzieje się tak, ponieważ odebrany sygnał jest zdominowany przez bezpośrednie łącze w centrum komórki, podczas gdy odbite łącze dominuje na krawędzi komórki. Ponadto można zaobserwować, że BS-IRS MRT zachowuje się odwrotnie, gdy UE oddala się od BS w kierunku IRS. Ujawnia to również znaczenie dokładnych przesunięć fazowych, ponieważ iloczynowe przesunięcia fazowe przytłaczają potencjał IRS. Schemat IRS z podwójną wiązką może osiągnąć dobrą równowagę w centrum komórki i na krawędzi

komórki. Konkretnie, jego wydajność na krawędzi komórki zbliża się do optymalnego wyniku, ale złożoność implementacji jest zmniejszona.

Komunikacja szerokopasmowa wspomagana przez IRS

Poprzednie sekcje omawiały jedynie komunikację wspomaganą przez IRS w kanałach o zanikaniu częstotliwości (wąskopasmowych), które można po prostu modelować za pomocą pojedynczego odczepu kanału, np. $h \sim \mathcal{CN}(0, \sigma_h^2)$. Jednak jednym z trendów technologicznych w transmisji bezprzewodowej jest to, że szerokość pasma sygnału staje się coraz szersza, od dziesiątek kHz w systemach pierwszej generacji do setek MHz w systemach piątej generacji, co ma na celu obsługę wyższej szybkości transmisji. W rezultacie większość współczesnych systemów komunikacji bezprzewodowej jest szerokopasmowa z szerokością pasma sygnału znacznie szerszą niż szerokość pasma koherencji, co prowadzi do poważnej selektywności częstotliwości.

Kaskadowy kanał selektywny częstotliwościowo

Kanał zanikania wielodrogowego można opisać za pomocą odpowiedzi w czasie t na impuls w czasie t − τ, a mianowicie

$$h(\tau, t) = \sum_{l=1}^{L} a_l(t)\delta(\tau - \tau_l(t)) \tag{73}$$

gdzie $a_l(t)$ i $\tau_l(t)$ oznaczają tłumienie i opóźnienie propagacji l-tej ścieżki w czasie t, odpowiednio, a L jest całkowitą liczbą możliwych do rozwiązania ścieżek. W szczególnej sytuacji, gdy nadajnik, odbiornik i otaczające środowisko są nieruchome, opóźnienia tłumienia i propagacji nie zmieniają się w czasie. Stąd otrzymujemy liniowy, niezmienny w czasie model kanału z odpowiedzią impulsową

$$h(\tau) = \sum_{l=1}^{L} a_l \delta(\tau - \tau_l) \tag{74}$$

Praktyczna komunikacja bezprzewodowa to transmisja w paśmie przepustowym, która jest realizowana w paśmie o częstotliwości nośnej fc. Jednak większość przetwarzania sygnału w komunikacji bezprzewodowej, takiego jak kodowanie kanału, modulacja, wykrywanie i szacowanie kanału, jest zwykle implementowana w paśmie podstawowym. Stąd sensowne jest uzyskanie złożonego równoważnego modelu pasma podstawowego, który jest podany przez Tse i Viswanatha

$$h_b(\tau, t) = \sum_{l=1}^{L} a_l(t) e^{-2\pi j f_c \tau_l(t)} \delta(\tau - \tau_l(t)) \tag{75}$$

Stosując twierdzenie o próbkowaniu, możemy utworzyć bardziej użyteczny model kanału w czasie dyskretnym, ustalając ζ-ty odczep filtra kanałowego w (dyskretnym) czasie n, tj.

$$h_\zeta[n] = \sum_{l=1}^{L} a_l(nT_s) e^{-2\pi j f_c \tau_l(nT_s)} \operatorname{sinc}\left(\zeta - \frac{\tau_l(nT_s)}{T_s}\right), \quad \zeta = 0, 1, \ldots, Z-1, \tag{76}$$

gdzie $T_s = 1/B_s$ oznacza okres próbkowania z szerokością pasma sygnału B_s, a funkcja sinc jest zdefiniowana jako

$$\mathrm{sinc}(t) := \frac{\sin(\pi t)}{\pi t} \quad (77)$$

W szczególnym przypadku, gdy wzmocnienia i opóźnienia ścieżek są niezmienne w czasie, równanie (76) można uprościć do

$$h_\zeta = \sum_{l=1}^{L} a_l e^{-2\pi j f_c \tau_l} \mathrm{sinc}(\zeta - \frac{\tau_l}{T_s}). \quad (78)$$

W celu przedstawienia modelu systemowego transmisji OFDM wspomaganej przez IRS zakładamy, że zarówno stacja bazowa, jak i użytkownik są wyposażeni w jedną antenę.Równoważną odpowiedź impulsową pasma podstawowego kaskadowego kanału IRS od stacji bazowej do użytkownika za pośrednictwem n-tego elementu odbijającego można modelować za pomocą

$$v_n(t) = g_n^b(\tau) * c_n * h_n^b(\tau). \quad (79)$$

Innymi słowy, odpowiedź impulsowa kanału kaskadowego jest liniowym splotem odpowiedzi impulsowej stacji bazowej na kanał elementu n, współczynnika odbicia i odpowiedzi impulsowej elementu n na kanał użytkownika. Zgodnie z równaniem (78) kanał równoważny pasma podstawowego w czasie dyskretnym od stacji bazowej do n-tego odbijającego elementu w systemie szerokopasmowym można wyrazić wzorem

$$\mathbf{h}_n = [h_n^1, h_n^2, \ldots, h_n^\zeta, \ldots, h_n^{Z_n^{\mathbb{BI}}}]^T, \quad (80)$$

z liczbą opóźnionych odczepów dla elementu odbijającego n w łączu $\mathbb{B}$-$\mathbb{I}$ $Z^{\mathbb{BI}}{}_n$. Podobnie, kanał równoważny pasma podstawowego w czasie dyskretnym od n-tego elementu odbijającego do użytkownika jest oznaczony przez

$$\mathbf{g}_n = [g_n^1, g_n^2, \ldots, g_n^\zeta, \ldots, g_n^{Z_n^{\mathbb{IU}}}]^T, \quad (81)$$

gdzie $Z^{\mathbb{IU}}{}_n$ jest liczbą opóźnionych odczepów dla elementu odbijającego n w łączu $\mathbb{I}$-$\mathbb{U}$. Stąd, odnosząc się do równania (79), kanał kaskadowy od stacji bazowej do użytkownika za pośrednictwem n-tego elementu odbijającego można modelować jako

$$\mathbf{v}_n = \mathbf{g}_n * c_n * \mathbf{h}_n = c_n(\mathbf{g}_n * \mathbf{h}_n). \quad (82)$$

Możemy również zapisać kanał kaskadowy jako

$$\mathbf{v}_n = [v_n^1, v_n^2, \ldots, v_n^\zeta, \ldots, v_n^{Z_n^{\mathbb{BU}}}]^T, \quad (83)$$

z liczbą opóźnionych odczepów $Z^{\mathbb{BU}}{}_n = Z^{\mathbb{BI}}{}_n + Z^{\mathbb{IU}}{}_n - 1$. Na koniec, kanał równoważny pasmu bazowemu w czasie dyskretnym łącza bezpośredniego można zapisać jako

$$\mathbf{h}_d = [h_d^1, h_d^2, \ldots, h_d^\zeta, \ldots, h_d^{Z_d}]^T, \quad (84)$$

używając Z_d do oznaczenia liczby opóźnionych odczepów w łączu bezpośrednim

System OFDM wspomagany przez IRS

Ze względu na zdolność radzenia sobie z wielodrożnym zanikaniem częstotliwości bez konieczności skomplikowanej korekcji i prostej implementacji poprzez wykorzystanie cyfrowej transformaty Fouriera, OFDM lub multipleksowanie z ortogonalnym podziałem częstotliwości stało się dominującą techniką modulacji dla przewodowych i bezprzewodowych systemów komunikacyjnych w ciągu ostatnich dwóch dekad. Było szeroko stosowane w szerokiej gamie znanych standardów, np. Long-Term Evolution (LTE)-Advanced i 5GNR. Przewiduje się, że będzie ono stanowić kluczową technologię w nadchodzącym systemie 6G zarówno w konwencjonalnym paśmie sub-6 GHz, jak i pasmach wysokiej częstotliwości. Dlatego warto zwrócić szczególną uwagę na modelowanie i projektowanie transmisji OFDM wspomaganej przez IRS. Transmisja danych w systemie OFDM jest organizowana blokowo w kanale zanikania bloków, gdzie kanał pozostaje stały dla każdego symbolu OFDM. Piszemy

$$\tilde{\mathbf{x}}[t] = [\tilde{x}_0[t], \ldots, \tilde{x}_m[t], \ldots, \tilde{x}_{M-1}[t]]^T \quad (85)$$

oznacza blok transmisji w dziedzinie częstotliwości stacji bazowej na t-tym symbolu OFDM. (Tylda˜ oznacza zmienne w dziedzinie częstotliwości.) Przekształć $\tilde{\mathbf{x}}[t]$ w sekwencję w dziedzinie czasu

$$\mathbf{x}[t] = [x_0[t], \ldots, x_{m'}[t], \ldots, x_{M-1}[t]]^T \quad (86)$$

poprzez M-punktową odwrotną dyskretną transformację Fouriera (IDFT), tj.

$$x_{m'}[t] = \frac{1}{M}\sum_{m=0}^{M-1} \tilde{x}_m[t] e^{\frac{2\pi j m' m}{M}}. \quad (87)$$

∀m′ = 0, 1,…,M − 1. Definiowanie macierzy dyskretnej transformaty Fouriera (DFT)

$$\mathbf{F} = \begin{bmatrix} \omega_M^{0\cdot 0} & \cdots & \omega_M^{0\cdot(M-1)} \\ \vdots & \ddots & \vdots \\ \omega_M^{(M-1)\cdot 0} & \cdots & \omega_M^{(M-1)\cdot(M-1)} \end{bmatrix} \quad (88)$$

przy pierwotnym M-tym pierwiastku jedności $\omega^{m.m'}{}_M = e^{2\pi j m m'/M}$ modulację OFDM można również zapisać w postaci macierzowej jako

$$\mathbf{x}[t] = \mathbf{F}^{-1}\tilde{\mathbf{x}}[t] = \frac{1}{M}\mathbf{F}^*\tilde{\mathbf{x}}[t] \quad (89)$$

Aby uniknąć interferencji między symbolami (ISI) i zachować ortogonalność podnośnych, przedział ochronny znany jako prefiks cykliczny (CP) lub zwany rozszerzeniem cyklicznym jest pierwotnie dodawany pomiędzy dwoma kolejnymi blokami. Wstawianie CP oznacza, że ostatnia część symbolu OFDM jest kopiowana i wstawiana na początku tego symbolu OFDM. Zatem symbol OFDM z wstawieniem CP jest wyrażony przez

$$\mathbf{x}^{cp}[t] = [x_{M-N_{cp}}[t], \ldots, x_{M-1}[t], x_0[t], \ldots, x_{M-1}[t]]^T. \quad (90)$$

ISI można całkowicie wyeliminować, jeżeli długość CP nie jest mniejsza od długości żadnego filtra kanałowego, tj.

$$N_{cp} \geqslant \max\left(Z_1^{\mathrm{BU}}, \ldots, Z_N^{\mathrm{BU}}, Z_d\right) \quad (91)$$

Przesłany sygnał $x^{cp}[t]$ przechodzi przez kanał bezpośredni h_d, aby dotrzeć do użytkownika, co skutkuje otrzymaniem składowej sygnału $x^{cp}[t] * h_d$. Tymczasem kanał od stacji bazowej do użytkownika poprzez n-ty element odbijający odpowiada składowej sygnału $x^{cp}[t] * v_n = c_n x^{cp}[t] * (g_n * h_n)$. Zatem otrzymany wektor symboli u użytkownika jest obliczany przez

$$\begin{aligned}\mathbf{y}^{cp}[t] &= \sum_{n=1}^{N} \mathbf{v}_n * \mathbf{x}^{cp}[t] + \mathbf{h}_d * \mathbf{x}^{cp}[t] + \mathbf{z}[t] \\ &= \sum_{n=1}^{N} c_n(\mathbf{g}_n * \mathbf{h}_n) * \mathbf{x}^{cp}[t] + \mathbf{h}_d * \mathbf{x}^{cp}[t] + \mathbf{z}[t],\end{aligned} \tag{92}$$

z wektorem szumu addytywnego z[t]. Usuwając CP otrzymujemy

$$\begin{aligned}\mathbf{y}[t] &= \sum_{n=1}^{N} \ddot{\mathbf{v}}_n \otimes \mathbf{x}[t] + \ddot{\mathbf{h}}_d \otimes \mathbf{x}[t] + \mathbf{z}[t] \\ &= \left(\sum_{n=1}^{N} \ddot{\mathbf{v}}_n + \ddot{\mathbf{h}}_d \right) \otimes \mathbf{x}[t] + \mathbf{z}[t],\end{aligned} \tag{93}$$

gdzie $\otimes$ oznacza splot cykliczny, $\ddot{v}_n$ jest filtrem kanałowym M-punktowym utworzonym przez dodanie zer na końcu v_n, tj.

$$\ddot{\mathbf{v}}_n = [v_n^1, v_n^2, \ldots, v_n^{\zeta}, \ldots, v_n^{Z_n^{\mathrm{IBU}}}, \underbrace{0, \ldots, 0}_{\text{Zero padding}}]^T, \tag{94}$$

i

$$\ddot{\mathbf{h}}_d = [h_d^1, h_d^2, \ldots, h_d^{\zeta}, \ldots, h_d^{Z_d}, \underbrace{0, \ldots, 0}_{\text{Zero padding}}]^T. \tag{95}$$

Oznaczając efektywny kanał obejmujący wszystkie ścieżki między stacją bazową a użytkownikiem przez $\ddot{\mathbf{h}} = \sum_{n=1}^{N} \ddot{\mathbf{v}}_n + \ddot{\mathbf{h}}_d$, model systemu w równaniu (93) upraszcza się do

$$\mathbf{y}[t] = \ddot{\mathbf{h}} \otimes \mathbf{x}[t] + \mathbf{z}[t]. \tag{96}$$

Następnie demodulator DFT wyprowadza odebrany sygnał w dziedzinie częstotliwości

$$\tilde{\mathbf{y}}[t] = \mathbf{F}\mathbf{y}[t] \tag{97}$$

Podstawiając (89) i (93) do (97) i stosując twierdzenie splotu dla DFT, otrzymujemy

$$\begin{aligned}\bar{\mathbf{y}}[t] &= \mathbf{F}\left(\dot{\mathbf{h}} \otimes \mathbf{x}[t]\right) + \mathbf{F}\mathbf{z}[t] \\ &= \sum_{n=1}^{N} \mathbf{F}(\ddot{\mathbf{v}}_n \otimes \mathbf{x}[t]) + \mathbf{F}(\dot{\mathbf{h}}_d \otimes \mathbf{x}[t]) + \mathbf{F}\mathbf{z}[t] \\ &= \sum_{n=1}^{N} \bar{\mathbf{v}}_n \odot \bar{\mathbf{x}}[t] + \bar{\mathbf{h}}_d \odot \bar{\mathbf{x}}[t] + \bar{\mathbf{z}}[t],\end{aligned} \tag{98}$$

gdzie $\odot$ oznacza iloczyn Hadamarda (tj. mnożenie elementów), a $\bar{\mathbf{v}}_n$, $\bar{\mathbf{h}}_d$, i $\bar{\mathbf{z}}[t]$ oznaczają odpowiednio odpowiedzi kanału w dziedzinie częstotliwości i szum, które są obliczane przez

$$\bar{\mathbf{v}}_n = \mathbf{F}\ddot{\mathbf{v}}_n, \quad \bar{\mathbf{h}}_d = \mathbf{F}\dot{\mathbf{h}}_d, \quad \bar{\mathbf{z}}[t] = \mathbf{F}\mathbf{z}[t]. \tag{99}$$

Alternatywnie możemy uzyskać ogólną odpowiedź kanału w dziedzinie częstotliwości jako

$$\bar{\mathbf{h}} = \mathbf{F}\dot{\mathbf{h}} = \mathbf{F}\left(\sum_{n=1}^{N} \ddot{\mathbf{v}}_n + \dot{\mathbf{h}}_d\right) = \sum_{n=1}^{N} \mathbf{F}\ddot{\mathbf{v}}_n + \mathbf{F}\dot{\mathbf{h}}_d = \sum_{n=1}^{N} \bar{\mathbf{v}}_n + \bar{\mathbf{h}}_d. \tag{100}$$

W tym zapisie charakterystykę częstotliwości kanału (CFR) przy typowej podnośnej OFDM m można wyrazić jako

$$\bar{h}_m = \mathbf{f}_m\left(\sum_{n=1}^{N} \ddot{\mathbf{v}}_n + \dot{\mathbf{h}}_d\right) = \sum_{n=1}^{N} \mathbf{f}_m\ddot{\mathbf{v}}_n + \mathbf{f}_m\dot{\mathbf{h}}_d = \sum_{n=1}^{N} \bar{v}_{m,n} + \bar{h}_{m,d}, \tag{101}$$

gdzie f_m jest m-tym wierszem macierzy DFT F, a $\bar{h}_m$ oznacza m-ty wpis $\bar{\mathbf{h}}$, a $\bar{v}_{m,n}$ i $\bar{h}_{m,d}$ są odpowiednio $\bar{\mathbf{v}}_n$ i $\bar{\mathbf{h}}_d$. Następnie model systemu w równaniu (98) można zapisać jako

$$\bar{\mathbf{y}}[t] = \bar{\mathbf{h}} \odot \bar{\mathbf{x}}[t] + \bar{\mathbf{z}}[t], \tag{102}$$

co jest równoważne z formą konwencjonalnego systemu OFDM. Na koniec każdy kanał selektywny częstotliwościowo w systemie OFDM wspomaganym przez IRS jest przekształcany w zestaw M niezależnych płaskich częstotliwościowo podnośnych. Transmisję sygnału na m-tej podnośnej można modelować za pomocą

$$\begin{aligned}\bar{y}_m[t] &= \bar{h}_m\bar{x}_m[t] + \bar{z}_m[t] \\ &= \left(\sum_{n=1}^{N} \bar{v}_{m,n} + \bar{h}_{m,d}\right)\bar{x}_m[t] + \bar{z}_m[t], \quad m = 0, 1, \ldots, M-1.\end{aligned} \tag{103}$$

Następnie możemy zaobserwować, że model sygnału wspomaganej przez IRS transmisji OFDM na każdej podnośnej jest równoważny modelowi wąskopasmowej transmisji wspomaganej przez IRS, jak podano w równaniu (16).

Maksymalizacja szybkości

Pomijając utratę przepustowości spowodowaną wstawieniem CP, osiągalną szybkość (tj. wydajność widmową w bps/Hz) systemu OFDM wspomaganego przez IRS można obliczyć za pomocą

$$R = \sum_{m=0}^{M-1} \log_2 \left(1 + \frac{P_m \left| \sum_{n=1}^{N} \tilde{v}_{m,n} + \tilde{h}_{m,d} \right|^2}{\sigma_z^2 / M} \right)$$
$$= \sum_{m=0}^{M-1} \log_2 \left(1 + \frac{P_m \left| \sum_{n=1}^{N} \mathbf{f}_m \ddot{\mathbf{v}}_n + \mathbf{f}_m \ddot{\mathbf{h}}_d \right|^2}{\sigma_z^2 / M} \right), \quad (104)$$

gdzie P_m oznacza moc nadawania przypisaną do m-tej podnośnej z ograniczeniem $\Sigma^{M-1}_{m=0} P_m \leqslant P_t$. Aby zmaksymalizować osiągalną szybkość, odbijające przesunięcia fazowe muszą uwzględniać kanały selektywne pod względem częstotliwości na różnych podnośnych lub, co za tym idzie, kanały w dziedzinie czasu przy różnych opóźnionych odczepach. Ponadto, $\boldsymbol{\Theta}$ musi zostać wspólnie zoptymalizowane z alokacjami mocy nadawania $\mathbf{p} = [P_0, P_1, \ldots, P_{M-1}]^T$ na M podnośnych, formułując następujący problem optymalizacji

$$\max_{\boldsymbol{\Theta},\, \mathbf{p}} \quad \sum_{m=0}^{M-1} \log_2 \left(1 + \frac{P_m \left| \sum_{n=1}^{N} \mathbf{f}_m \ddot{\mathbf{v}}_n + \mathbf{f}_m \ddot{\mathbf{h}}_d \right|^2}{\sigma_z^2 / M} \right)$$
$$\text{s.t.} \quad \theta_n \in [0, 2\pi), \quad \forall n = 1, 2, \ldots, N,$$
$$P_m \geqslant 0, \quad \forall m = 0, 1, \ldots, M-1,$$
$$\sum_{m=0}^{M-1} P_m \leqslant P_t, \quad (105)$$

co jest trudniejsze do rozwiązania w porównaniu z przypadkiem wąskopasmowym. Aby rozwiązać ten problem, zaproponowano wydajny algorytm oparty na sukcesywnym przybliżeniu wypukłym (SCA) poprzez aproksymację niewklęsłej funkcji szybkości w tej funkcji celu przy użyciu jej wklęsłej dolnej granicy w oparciu o rozwinięcie Taylora pierwszego rzędu. Algorytm oparty na SCA gwarantuje zbieżność do punktu stacjonarnego wspólnych współczynników odbicia IRS i problemu optymalizacji mocy transmisji, i wymaga jedynie złożoności wielomianowej w N i M. Aby jeszcze bardziej obniżyć złożoność, Zheng i Zhang [2020] zaproponowali uproszczony algorytm, w którym przesunięcia fazowe IRS są zaprojektowane tak, aby były zgodne tylko z kanałem domeny czasu o najsilniejszej sile ścieżki, określanym zatem jako maksymalizacja najsilniejszej odpowiedzi impulsowej kanału (CIR). Zakładając równy przydział mocy na podnośne OFDM, równanie (104) jest przepisywane jako

$$R = \sum_{m=0}^{M-1} \log_2 \left(1 + \frac{P_t \left| \sum_{n=1}^{N} \tilde{v}_{m,n} + \tilde{h}_{m,d} \right|^2}{\sigma_z^2} \right), \quad (106)$$

stosując $P_m = P_t/M$. Jest niewklęsły nad $\boldsymbol{\Theta}$ i dlatego trudno go optymalnie zmaksymalizować. Alternatywnie, rozważa się maksymalizację górnej granicy szybkości za pomocą nierówności Jensena, która jest dana przez

$$R \leqslant \log_2\left(1+\frac{1}{M}\sum_{m=0}^{M-1}\frac{P_t\left|\sum_{n=1}^{N}\bar{v}_{m,n}+\bar{h}_{m,d}\right|^2}{\sigma_z^2}\right). \quad (107)$$

Zaniedbując stałe wyrazy dla uproszczenia, formułujemy następujący problem optymalizacyjny

$$\max_{\boldsymbol{\Theta}} \quad \sum_{m=0}^{M-1}\left|\sum_{n=1}^{N}\bar{v}_{m,n}+\bar{h}_{m,d}\right|^2 \quad (108)$$

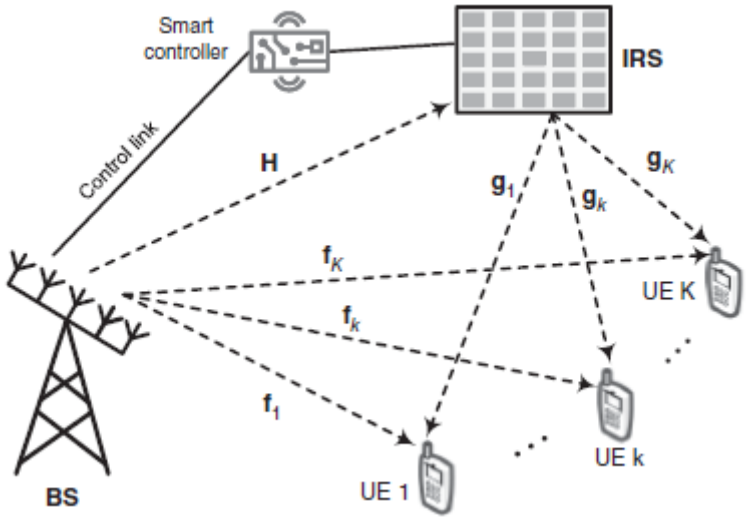
Wykorzystując właściwość dziedziny czasu, zastosowano metodę o niskiej złożoności, zwaną najsilniejszą maksymalizacją CIR, aby rozwiązać ten problem optymalizacji w sposób suboptymalny, co można odnieść do Zheng i Zhang [2020]. Na koniec warto zauważyć, że przesunięcia fazowe IRS θ_n, ∀n wpływają na odpowiedź kanału na każdej podnośnej OFDM identycznie. Innymi słowy, każdy element odbijający może wywołać tylko taką samą rotację fazy na wszystkich podnośnych OFDM w danym momencie i nie może zrealizować rotacji fazy specyficznej dla częstotliwości ze względu na ograniczenia implementacji sprzętowej. Wu i inni przedstawiają górną granicę osiągalnej szybkości, zakładając, że (w idealnym przypadku) różne współczynniki odbicia IRS można zaprojektować dla różnych podnośnych, dzięki czemu projekt odbicia jest specyficzny dla częstotliwości. Obserwuje się, że ta górna granica szybkości znacznie przewyższa rozwiązanie oparte na SCA z praktycznym odbiciem IRS o płaskiej częstotliwości, a luka szybkości zwiększa się wraz z liczbą podnośnych. Ujawnia to zatem, że fundamentalnym ograniczeniem systemów OFDM wspomaganych przez IRS jest brak odbicia IRS specyficznego dla częstotliwości ze względu na jego pasywne działanie.

Komunikacja IRS dla wielu użytkowników

Ze względu na swoją zdolność do zakłócania inteligentnej rekonfiguracji środowiska propagacji bezprzewodowej i ograniczenia sprzętowe, integracja IRS wprowadza pewne podstawowe cechy szczególne w koordynacji transmisji sygnału dla wielu użytkowników. Na przykład brak odbicia selektywnego częstotliwościowo prowadzi do utraty wydajności podejść z podziałem częstotliwości. Dlatego warto zwrócić szczególną uwagę na wpływ IRS na transmisję sygnału dla wielu użytkowników

Model wielokrotnego dostępu

Rysunek przedstawia wspomagany przez IRS wieloużytkownikowy system komunikacji MIMO downlink, w którym inteligentna powierzchnia z N elementami odbijającymi jest rozmieszczona w celu wspomagania transmisji z Nb-anteny BS do K jednoantenowych UE.



IRS jest pasywnym urządzeniem, w którym TDD jest zwykle przyjmowane w celu uproszczenia szacowania kanału. Użytkownicy wysyłają sygnały pilotażowe w treningu łącza w górę, tak aby BS mogła oszacować CSI łącza w górę, które jest wykorzystywane do optymalizacji transmisji danych łącza w dół ze względu na wzajemność kanałów. Aby scharakteryzować analizę teoretyczną, załóżmy, że CSI wszystkich zaangażowanych kanałów jest doskonale znany w BS. Ponadto kanały podążają za częstotliwościowo-płaskim blokiem zanikania. Ponieważ bezpośrednie ścieżki z BS lub IRS do UE mogą być zablokowane, odpowiadające im zanikanie na małą skalę podąża za rozkładem Rayleigha. W konsekwencji wzmocnienie kanału pomiędzy elementem anteny $n_b \in \{1, 2,\ldots,N_b\}$ a użytkownikiem $k \in \mathcal{K} \triangleq \{1, 2, \ldots, K\}$ jest kołowo symetryczną zespoloną zmienną losową Gaussa o zerowej średniej i wariancji $\sigma^2{}_f$, tj. $f_{kN_b} \sim \mathcal{CN}(0, \sigma_f^2)$. Zatem wektory kanałowe od BS i IRS do k-tego UE są oznaczone przez

$$\mathbf{f}_k = [f_{k1}, f_{k2}, \cdots, f_{kN_b}]^T, \quad (109)$$

i

$$\mathbf{g}_k = [g_{k1}, g_{k2}, \cdots, g_{kN}]^T, \quad (110)$$

odpowiednio, gdzie $g_{kn} \sim \mathcal{CN}(0, \sigma_g^2)$ jest wzmocnieniem kanału pomiędzy elementem IRS $n \in \mathcal{N} \triangleq \{1, 2, \ldots, N\}$ i użytkownikiem k. Piszemy

$$\mathbf{h}_n = [h_{n1}, h_{n2}, \ldots, h_{nN_b}]^T \quad (111)$$

aby oznaczyć wektor kanału pomiędzy BS a n-tym odbijającym elementem. Tak więc macierz kanału od BS do IRS jest wyrażona jako $H \in \mathbb{C}^{N \times Nb}$, gdzie n-ty wiersz H jest równy $h^T{}_n$. W przeciwieństwie do losowo rozłożonych i ruchomych UE, korzystna lokalizacja jest celowo wybierana dla IRS, aby wykorzystać ścieżkę LOS stałej BS bez żadnej blokady, co skutkuje zanikaniem Riciana, tj.

$$\mathbf{H} = \sqrt{\frac{\Gamma\sigma_h^2}{\Gamma+1}}\mathbf{H}_{\text{LOS}} + \sqrt{\frac{\sigma_h^2}{\Gamma+1}}\mathbf{H}_{NLOS}, \quad (112)$$

z czynnikiem Riciana Γ, składową LOS H_{LOS}, składową wielościeżkową H_{NLOS} składającą się z niezależnych wpisów następujących po $\mathcal{CN}(0,1),$ i stratą ścieżki BS-IRS σ^2_h. Powierzchnia odbijająca jest wyposażona w inteligentny kontroler, który może adaptacyjnie regulować przesunięcie fazowe każdego elementu IRS w kategoriach CSI uzyskanego poprzez estymację kanału okresowego. Piszemy $c_{nk} = \beta_{nk}e^{j\phi nk}$, aby oznaczyć współczynnik odbicia n-tego elementu IRS dla użytkownika k, z indukowanym przesunięciem fazowym $\phi_{nk} \in [0, 2\pi)$ i tłumieniem amplitudy $\beta_{nk} \in [0, 1]$. Jak ujawnili Wu i Zhang , optymalna wartość tłumienia odbicia wynosi $\beta_{nk} = 1$, ∀n, k, aby zmaksymalizować odebraną moc i uprościć implementację sprzętową. Ze względu na wysoką stratę ścieżki sygnały odbite przez IRS dwa lub więcej razy są pomijalne. Ignorując upośledzenia sprzętowe, takie jak ilościowe przesunięcia fazowe i szum fazowy, k-ty UE obserwuje odebrany sygnał

$$r_k = \sqrt{P_d}\left(\sum_{n=1}^{N} g_{kn}e^{j\phi_{kn}}\mathbf{h}_n^T + \mathbf{f}_k^T\right)\mathbf{s} + n_k, \quad (113)$$

gdzie s oznacza wektor $N_b \times 1$ sygnałów przesyłanych przez antenę BS, P_d wyraża ograniczenie mocy BS, n_k to AWGN z zerową średnią i wariancją σ^2_n, mianowicie $n_k \sim \mathcal{CN}(0, \sigma_n^2).$. Zdefiniuj$\mathbf{\Theta}$k = diag{$e^{j\phi 1k}$,…, $e^{j\phi Nk}$}, Równanie (113) można zapisać w postaci macierzowej jako

$$r_k = \sqrt{P_d}(\mathbf{g}_k^T\mathbf{\Theta}_k\mathbf{H} + \mathbf{f}_k^T)\mathbf{s} + n_k \quad (114)$$

Ortogonalny wielokrotny dostęp

W tej części analizowana jest osiągalna wydajność widmowa dwóch typowych schematów ortogonalnego wielokrotnego dostępu (OMA) w wielodostępnym systemie MIMO wspomaganym przez IRS, a także prezentowana jest naprzemienna metoda optymalizacji aktywnego kształtowania wiązki w stacji bazowej i pasywnego odbicia w IRS.

Wielodostęp z podziałem czasu

Ten schemat dzieli wymiary sygnalizacji wzdłuż osi czasu na ortogonalne części zwane przedziałami czasowymi. Każdy użytkownik transmituje przez całą szerokość pasma, ale cyklicznie uzyskuje dostęp do przypisanego mu przedziału czasowego. Oznacza to nieciągłą transmisję, co upraszcza projekt systemu, ponieważ niektóre przetwarzanie, takie jak szacowanie kanału, można wykonać w przedziałach czasowych innych użytkowników. Inną zaletą jest to, że TDMA jest w stanie przypisać wiele przedziałów czasowych dla pojedynczego użytkownika, zwiększając elastyczność systemu. Matematycznie ramka radiowa jest ortogonalnie dzielona na K przedziałów czasowych, gdzie CSI pozostaje stałe. W k-tym przedziale BS stosuje liniowe formowanie wiązki $w_k \in \mathbb{C}^{Nb\times 1}$, gdzie $||w_k||^2 \leqslant 1$, aby wysłać symbol niosący informacje s_k z zerową średnią i wariancją jednostkową, tj. $\mathbb{E}[\,|s_k|^2] = 1$, przeznaczony dla ogólnego użytkownika k. Podstawiając $s = w_k s_k$ do równania (114) otrzymujemy

$$r_k = \sqrt{P_d}(\mathbf{g}_k^T\mathbf{\Theta}_k\mathbf{H} + \mathbf{f}_k^T)\mathbf{w}_k s_k + n_k \quad (115)$$

Łączna optymalizacja aktywnego kształtowania wiązkiwk i odbicia $\mathbf{\Theta}_k$ pozwala na uzyskanie chwilowego SNR użytkownika k, tj.

$$\gamma_k = \frac{P_d |(\mathbf{g}_k^T \boldsymbol{\Theta}_k \mathbf{H} + \mathbf{f}_k^T)\mathbf{w}_k|^2}{\sigma_n^2} \quad (116)$$

można zmaksymalizować, formułując następującą optymalizację

$$\begin{aligned} \max_{\boldsymbol{\Theta}_k,\, \mathbf{w}_k} \quad & |(\mathbf{g}_k^T \boldsymbol{\Theta}_k \mathbf{H} + \mathbf{f}_k^T)\mathbf{w}_k|^2 \\ \text{s.t.} \quad & \|\mathbf{w}_k\|^2 \leqslant 1 \\ & \phi_{nk} \in [0, 2\pi), \quad \forall n = 1, \ldots, N, \ \forall k = 1, \ldots, K, \end{aligned} \quad (117)$$

który jest niewypukły, ponieważ funkcja celu nie jest wspólnie wklęsła względem $\boldsymbol{\Theta}_k$ i w_k. Aby rozwiązać ten problem, możemy zastosować optymalizację naprzemienną, która naprzemiennie optymalizuje $\boldsymbol{\Theta}_k$ i w_k w sposób iteracyjny. Biorąc pod uwagę zainicjowany wektor transmisji $w^{(0)}_k$, równanie (117) jest uproszczone do

$$\begin{aligned} \max_{\boldsymbol{\Theta}_k} \quad & |(\mathbf{g}_k^T \boldsymbol{\Theta}_k \mathbf{H} + \mathbf{f}_k^T)\mathbf{w}_k^{(0)}|^2 \\ \text{s.t.} \quad & \phi_{nk} \in [0, 2\pi), \quad \forall n = 1, \ldots, N, \ \forall k = 1, \ldots, K \end{aligned} \quad (118)$$

Funkcja celu nadal nie jest wypukła, ale umożliwia rozwiązanie w formie zamkniętej poprzez zastosowanie dobrze znanej nierówności trójkąta

$$|(\mathbf{g}_k^T \boldsymbol{\Theta}_k \mathbf{H} + \mathbf{f}_k^T)\mathbf{w}_k^{(0)}| \leqslant |\mathbf{g}_k^T \boldsymbol{\Theta}_k \mathbf{H} \mathbf{w}_k^{(0)}| + |\mathbf{f}_k^T \mathbf{w}_k^{(0)}| \quad (119)$$

Równość zachodzi wtedy i tylko wtedy, gdy

$$\arg\left(\mathbf{g}_k^T \boldsymbol{\Theta}_k \mathbf{H} \mathbf{w}_k^{(0)}\right) = \arg\left(\mathbf{f}_k^T \mathbf{w}_k^{(0)}\right) \triangleq \varphi_{0k}. \quad (120)$$

Zdefiniuj $q_k = q_{1k}, q_{2k}, \ldots, q_{Nk}]^H$ przy $q_{nk} = e^{j\phi nk}$ i $\boldsymbol{\chi}_k = \operatorname{diag}(\mathbf{g}_k^T)\mathbf{H}\mathbf{w}_k^{(0)} \in \mathbb{C}^{N\times 1}$, mamy $g^T_k \boldsymbol{\Theta}_k H w^{(0)}_k = q^H_k \boldsymbol{\chi}_k \in \mathbb{C}$. Zignoruj stały wyraz $|f^T k\, w^{(0)}_k|$, Równanie (118) jest przekształcane do

$$\begin{aligned} \max_{\mathbf{q}_k} \quad & |\mathbf{q}_k^H \boldsymbol{\chi}_k| \\ \text{s.t.} \quad & |q_{nk}| = 1, \quad \forall n = 1, \ldots, N, \ \forall k = 1, \ldots, K, \\ & \arg(\mathbf{q}_k^H \boldsymbol{\chi}_k) = \varphi_{0k}. \end{aligned} \quad (121)$$

Rozwiązanie równania (121) można wyprowadzić w następujący sposób

$$\mathbf{q}_k^{(1)} = e^{j(\varphi_{0k} - \arg(\chi_k))} = e^{j\left(\varphi_{0k} - \arg\left(\operatorname{diag}(\mathbf{g}_k^T)\mathbf{H}\mathbf{w}_k^{(0)}\right)\right)} \quad (122)$$

Odpowiednio,

$$\begin{aligned} \phi_{nk}^{(1)} &= \varphi_{0k} - \arg\left(g_{nk} \mathbf{h}_n^T \mathbf{w}_k^{(0)}\right) \\ &= \varphi_{0k} - \arg\left(g_{nk}\right) - \arg\left(\mathbf{h}_n^T \mathbf{w}_k^{(0)}\right) \end{aligned} \quad (123)$$

gdzie $h^T_n w^{(0)}_k \in \mathbb{C}$ można uznać za efektywny kanał SISO odbierany przez n-ty element odbijający, łączący efekty kształtowania wiązki nadawczej w(0) k i odpowiedzi kanału hn. W związku z tym równanie (123) oznacza, że reflektor IRS powinien być dostrojony tak, aby faza sygnału odbitego przez łącze kaskadowe była kompensowana, a faza resztkowa była wyrównana z fazą sygnału przez łącze bezpośrednie, aby uzyskać spójne łączenie w odbiorniku. Po ustaleniu faz odbijających w pierwszej iteracji, tj. $\boldsymbol{\Theta}^{(1)}_k = \text{diag}\{e^{j\phi(1)1k}, e^{j\phi(1)2k}, \ldots, e^{j\phi(1)Nk}\}$, optymalizacja jest naprzemiennie przeprowadzana w celu aktualizacji w_k. BS może zastosować dopasowane filtrowanie w celu zmaksymalizowania siły pożądanego sygnału, co skutkuje

$$\mathbf{w}_k^{(1)} = \frac{(\mathbf{g}_k^T \boldsymbol{\Theta}_k^{(1)} \mathbf{H} + \mathbf{f}_k^T)^H}{\|\mathbf{g}_k^T \boldsymbol{\Theta}_k^{(1)} \mathbf{H} + \mathbf{f}_k^T\|}. \quad (124)$$

Po zakończeniu pierwszej iteracji BS otrzymuje $\boldsymbol{\Theta}^{(1)}_k$ i $w^{(1)}_k$, które służą jako początkowe dane wejściowe dla drugiej iteracji w celu wyprowadzenia $\boldsymbol{\Theta}^{(2)}_k$ i $w^{(2)}_k$. Proces ten powtarza się, aż do osiągnięcia zbieżności z optymalnym kształtownikiem wiązki $w^{\star}_k$ i optymalnym odbiciem $\boldsymbol{\Theta}^{\star}_k$. Podstawiając $w^{\star}k$ i $\boldsymbol{\Theta}^{\star}_k$ do równania (116), możemy wyprowadzić możliwą do osiągnięcia wydajność widmową użytkownika k jako

$$R_k = \frac{1}{K} \log \left(1 + \frac{P_d \left| (\mathbf{g}_k^T \boldsymbol{\Theta}_k^{\star} \mathbf{H} + \mathbf{f}_k^T) \mathbf{w}_k^{\star} \right|^2}{\sigma_n^2} \right). \quad (125)$$

W ten sposób można obliczyć stawkę sumaryczną systemu TDMA IRS

$$R_{\text{TDMA}} = \sum_{k=1}^{K} R_k. \quad (126)$$

Dostęp wielokrotny z podziałem częstotliwości

W FDMA pasmo systemu jest dzielone wzdłuż osi częstotliwości na K ortogonalnych podkanałów. Każdy użytkownik zajmuje dedykowany podkanał przez cały czas. BS wykorzystuje liniowe formowanie wiązki w_k do transmisji s_k przez k-ty podkanał z równomiernie przydzieloną mocą transmisji P_d/K. Zatem osiągalna wydajność widmowa użytkownika k wynosi

$$R_k = \frac{1}{K} \log \left(1 + \frac{P_d/K \left| (\mathbf{g}_k^T \boldsymbol{\Theta}_k \mathbf{H} + \mathbf{f}_k^T) \mathbf{w}_k \right|^2}{\sigma_n^2/K} \right). \quad (127)$$

W przeciwieństwie do TDMA, gdzie przesunięcia fazowe IRS mogą być dynamicznie dostosowywane w różnych slotach, powierzchnia może być optymalizowana tylko dla konkretnego użytkownika, podczas gdy inni użytkownicy cierpią na odbicie niewspółosiowe fazowo. Dzieje się tak z powodu ograniczeń sprzętowych pasywnych elementów IRS, które mogą być wytwarzane w sposób selektywny czasowo, a nie selektywny częstotliwościowo. Bez utraty ogólności zakładamy, że system FDMA optymalizuje IRS, aby pomóc w transmisji sygnału użytkownika k, optymalne parametry $\boldsymbol{\Theta}^{\star}_k$ i $w^{\star}_k$ można wyprowadzić, stosując tę samą optymalizację naprzemienną, co w przypadku TDMA. Gdy przesunięcia

fazowe powierzchni zostaną całkowicie dostosowane do k, to co pozostali użytkownicy K − 1, oznaczeni jako {i|i = 1, 2,…, K, i ≠ k}, mogą zrobić, to zrealizować częściową optymalizację (zamiast wspólnej optymalizacji) poprzez aktualizację ich odpowiednich aktywnych formowania wiązki na podstawie połączonego wzmocnienia kanału $g^T_i \ \mathbf{\Theta}^\star_k H + f^T_i$. Dla użytkownika i, formowanie wiązki można zoptymalizować jako

$$\mathbf{w}_i^\star = \frac{(\mathbf{g}_i^T \mathbf{\Theta}_k^\star \mathbf{H} + \mathbf{f}_i^T)^H}{\|\mathbf{g}_i^T \mathbf{\Theta}_k^\star \mathbf{H} + \mathbf{f}_i^T\|}. \quad (128)$$

Następnie można obliczyć stawkę sumaryczną systemu FDMA IRS,

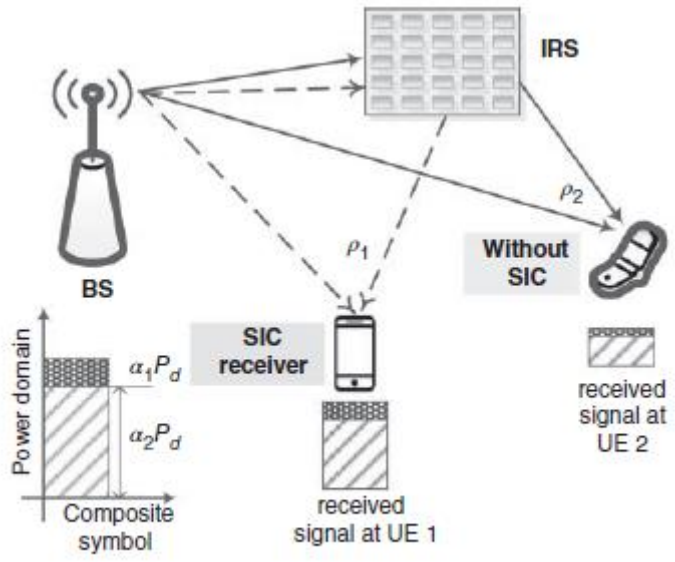
$$R_{FDMA} = \frac{1}{K}\log\left(1 + \frac{P_d|(\mathbf{g}_k^T \mathbf{\Theta}_k^\star \mathbf{H} + \mathbf{f}_k^T)\mathbf{w}_k^\star|^2}{\sigma_n^2}\right) + \sum_i \frac{1}{K}\log\left(1 + \frac{P_d|(\mathbf{g}_i^T \mathbf{\Theta}_k^\star \mathbf{H} + \mathbf{f}_i^T)\mathbf{w}_i^\star|^2}{\sigma_n^2}\right) \quad (129)$$

Nieortogonalny wielokrotny dostęp

Chociaż interferencja między użytkownikami wśród użytkowników multipleksowanych ortogonalnie jest łagodzona w celu ułatwienia wykrywania wielu użytkowników o niskiej złożoności w odbiorniku, powszechnie wiadomo, że OMA nie może osiągnąć łącznej przepustowości systemu bezprzewodowego dla wielu użytkowników. Kodowanie superpozycji i sukcesywne usuwanie zakłóceń (SIC) umożliwiają ponowne wykorzystanie każdej jednostki zasobów ortogonalnych przez więcej niż jednego użytkownika. W nadajniku wszystkie pojedyncze symbole informacji są nakładane na jeden przebieg, podczas gdy SIC w odbiorniku dekoduje sygnały iteracyjnie, aż otrzyma pożądany sygnał. Matematycznie BS nakłada K symboli niosących informacje na złożony przebieg

$$\mathbf{s} = \sum_{k=1}^{K} \sqrt{\alpha_k}\mathbf{w}_k s_k, \quad (130)$$

gdzie α_k reprezentuje współczynnik przydziału mocy podlegający $\Sigma^K_{k=1} \alpha_k \leqslant 1$. Wyzwaniem jest podjęcie decyzji, jak przydzielić moc użytkownikom, co jest krytyczne dla usuwania zakłóceń w odbiorniku. Dlatego NOMA jest uważana za rodzaj wielokrotnego dostępu w domenie mocy. Ogólnie rzecz biorąc, więcej mocy jest przydzielane użytkownikom o mniejszym wzmocnieniu kanału, np. znajdującym się dalej od BS, w celu poprawy odebranego SNR, tak aby można było zagwarantować wysoką niezawodność wykrywania. Pomimo mniejszej mocy przydzielonej użytkownikowi o silniejszym wzmocnieniu kanału, np. blisko BS, jest on w stanie prawidłowo wykryć swój sygnał przy rozsądnym SNR. Jako przykład, ilustracja systemu NOMA z wykorzystaniem IRS w łączu wstecznym, składającego się z BS, powierzchni i K = 2 użytkowników, jest podana na rysunku .



Zastąp równanie (130) równaniem (114) aby uzyskać obserwację użytkownika k jako

$$
\begin{aligned}
r_k &= \sqrt{P_d}(\mathbf{g}_k^T \boldsymbol{\Theta}_k \mathbf{H} + \mathbf{f}_k^T) \sum_{k'=1}^{K} \sqrt{\alpha_{k'}} \mathbf{w}_{k'} s_{k'} + n_k \\
&= \underbrace{\sqrt{\alpha_k P_d}(\mathbf{g}_k^T \boldsymbol{\Theta}_k \mathbf{H} + \mathbf{f}_k^T) \mathbf{w}_k s_k}_{\text{Desired signal}} \\
&\quad + \underbrace{\sqrt{P_d}(\mathbf{g}_k^T \boldsymbol{\Theta}_k \mathbf{H} + \mathbf{f}_k^T) \sum_{k'=1, k' \neq k}^{K} \sqrt{\alpha_{k'}} \mathbf{w}_{k'} s_{k'}}_{\text{Multi-user interference}} + n_k
\end{aligned}
\tag{131}
$$

Ze względu na ograniczenia sprzętowe, IRS może pomóc tylko jednemu użytkownikowi, podczas gdy inni użytkownicy muszą dzielić wspólne przesunięcia fazowe, które nie są dla nich korzystne. Jako FDMA, zakładamy, że system optymalizuje IRS, aby pomóc w transmisji sygnału użytkownika $\hat{k}$. Optymalne parametry $\boldsymbol{\Theta}^\star_{\hat{k}}$ i $\mathbf{w}^\star_{\hat{k}}$ można wyprowadzić, stosując tę samą optymalizację naprzemienną, co TDMA. Gdy przesunięcia fazowe powierzchni zostaną całkowicie dostosowane do $\hat{k}$, użytkownik $k \neq \hat{k}$ może częściowo zoptymalizować swoją transmisję, wyprowadzając swoje aktywne formowanie wiązki, biorąc pod uwagę łączne wzmocnienie kanału $\mathbf{g}^T_k \boldsymbol{\Theta}^\star_{\hat{k}} \mathbf{H} + \mathbf{f}^T_k$. Podobnie jak w równaniu (128), formowanie wiązki dla użytkownika k jest obliczane jako

$$
\mathbf{w}_k^\star = \frac{(\mathbf{g}_k^T \boldsymbol{\Theta}_{\hat{k}}^\star \mathbf{H} + \mathbf{f}_k^T)^H}{\|\mathbf{g}_k^T \boldsymbol{\Theta}_{\hat{k}}^\star \mathbf{H} + \mathbf{f}_k^T\|}. \tag{132}
$$

Ten sam sygnał x, który zawiera wszystkie symbole informacyjne, jest dostarczany do wszystkich użytkowników. Optymalna kolejność usuwania zakłóceń to wykrywanie użytkownika z największym przydziałem mocy (najsłabszym wzmocnieniem kanału) do użytkownika z najmniejszym przydziałem mocy (najsilniejszym wzmocnieniem kanału). Piszemy $\rho_k = (\mathbf{g}^T_k \boldsymbol{\Theta}^\star_k \mathbf{H} + \mathbf{f}^T_k)\mathbf{w}^\star_k$, $\forall k$ aby oznaczyć efektywne wzmocnienie połączonego kanału dla użytkownika k. Bez utraty ogólności załóżmy, że użytkownik 1 ma największe łączne wzmocnienie kanału, a użytkownik K jest najsłabszy, tj.

$$\|\rho_1\|^2 \geqslant \|\rho_2\|^2 \geqslant \cdots \geqslant \|\rho_K\|^2. \quad (133)$$

W tym zamówieniu każdy użytkownik NOMA najpierw dekoduje sK, a następnie odejmuje jego wynikowy składnik od otrzymanego sygnału. W rezultacie typowy użytkownik k po pierwszej iteracji SIC otrzymuje

$$\bar{r}_k = r_k - \rho_k \sqrt{\alpha_K P_d} s_K = \rho_k \sum_{k=1}^{K-1} \sqrt{\alpha_k P_d} s_k + n_k. \quad (134)$$

zakładając bezbłędne wykrywanie i doskonałą znajomość kanału. W drugiej iteracji użytkownik dekoduje sK−1 przy użyciu pozostałego sygnału $\bar{r}_k$. Anulowanie powtarza się, aż każdy użytkownik otrzyma symbol przeznaczony dla niego. W szczególności najsłabszy użytkownik dekoduje swój własny sygnał bezpośrednio bez SIC, ponieważ przydzielono mu największą moc. Traktując zakłócenia wielu użytkowników jako szum, SNR dla użytkownika K można zapisać jako

$$\gamma_K = \frac{\|\rho_K\|^2 \alpha_K P_d}{\|\rho_K\|^2 \sum_{k=1}^{K-1} \alpha_k P_d + \sigma_n^2}. \quad (135)$$

Ogólnie rzecz biorąc, użytkownik k pomyślnie anuluje sygnały od użytkownika k + 1 do K, ale cierpi z powodu zakłóceń od użytkownika 1 do k − 1. W konsekwencji otrzymany SNR dla użytkownika k wynosi

$$\gamma_k = \frac{\|\rho_k\|^2 \alpha_k P_d}{\|\rho_k\|^2 \sum_{k'=1}^{k-1} \alpha_{k'} P_d + \sigma_n^2}, \quad (136)$$

co daje osiągalną szybkość $R_k = \log(1 + \gamma_k)$. Suma szybkości transmisji NOMA w łączu wstecznym wspomaganej przez IRS jest obliczana przez

$$R_{\text{NOMA}} = \sum_{k=1}^{K} \log\left(1 + \frac{\|\rho_k\|^2 \alpha_k P_d}{\|\rho_k\|^2 \sum_{k'=1}^{k-1} \alpha_{k'} P_d + \sigma_n^2}\right) \quad (137)$$

Starzenie się kanału i prognozowanie

Aby w pełni wykorzystać potencjał IRS w zwiększaniu mocy i wydajności widmowej, stacja bazowa musi znać natychmiastowe informacje o stanie kanału łącza bezpośredniego i łącza kaskadowego. W systemie FDD (Frequency-Division Duplex) CSI jest szacowany u użytkownika i przekazywany z powrotem do stacji bazowej za pośrednictwem ograniczonego kanału sprzężenia zwrotnego. Ze względu na opóźnienie sprzężenia zwrotnego dostępny CSI w stacji bazowej może ulec starzeniu przed jego faktycznym wykorzystaniem. Chociaż opóźnienia sprzężenia zwrotnego można uniknąć w systemie TDD wykorzystującym wzajemność kanałów, nadal możliwe jest, że CSI ulegnie starzeniu z powodu opóźnienia przetwarzania. Oznacza to, że powinniśmy wziąć pod uwagę czas potrzebny na obliczenie optymalnych faz odbijających i konfigurację elementów IRS, szczególnie w środowiskach o wysokiej mobilności lub wysokiej częstotliwości. W praktyce wydajność systemu jest podatna na takie opóźnienia sprzężenia zwrotnego i przetwarzania, ponieważ wiedza o CSI szybko się dezaktualizuje. Zjawisko to nazywane jest starzeniem się kanału, podlegającym zanikaniu kanału i uszkodzeniom sprzętu. Oprócz systemów wspomaganych przez IRS powszechnie wiadomo, że starzenie się kanałów poważnie pogarsza wydajność szerokiej gamy adaptacyjnych systemów komunikacji bezprzewodowej, w tym prekodowania MIMO , MIMO dla wielu użytkowników , masywnego MIMO, masywnego MIMO

bezkomórkowego, kształtowania wiązki, oportunistycznego wyboru przekaźnika , wyrównania interferencyjnego, różnorodności transmisji w pętli zamkniętej, wyboru anteny nadawczej , częstotliwości ortogonalnej Starzenie się kanału i prognozowanie. Aby w pełni wykorzystać potencjał IRS w zwiększaniu mocy i wydajności widmowej, stacja bazowa musi znać natychmiastowe informacje o stanie kanału łącza bezpośredniego i łącza kaskadowego. W systemie FDD (Frequency-Division Duplex) CSI jest szacowany u użytkownika i przekazywany z powrotem do stacji bazowej za pośrednictwem ograniczonego kanału sprzężenia zwrotnego. Ze względu na opóźnienie sprzężenia zwrotnego dostępny CSI w stacji bazowej może ulec starzeniu przed jego faktycznym wykorzystaniem. Chociaż opóźnienia sprzężenia zwrotnego można uniknąć w systemie TDD wykorzystującym wzajemność kanałów, nadal możliwe jest, że CSI ulegnie starzeniu z powodu opóźnienia przetwarzania. Oznacza to, że powinniśmy wziąć pod uwagę czas potrzebny na obliczenie optymalnych faz odbijających i konfigurację elementów IRS, szczególnie w środowiskach o wysokiej mobilności lub wysokiej częstotliwości. W praktyce wydajność systemu jest podatna na takie opóźnienia sprzężenia zwrotnego i przetwarzania, ponieważ wiedza o CSI szybko się dezaktualizuje. Zjawisko to nazywane jest starzeniem się kanału, podlegającym zanikaniu kanału i uszkodzeniom sprzętu. Oprócz systemów wspomaganych przez IRS, powszechnie wiadomo, że starzenie się kanału poważnie pogarsza wydajność szerokiej gamy adaptacyjnych systemów komunikacji bezprzewodowej, w tym MIMO precoding, MIMO dla wielu użytkowników , massive MIMO, cell-free massive MIMO , beamforming, oportunistyczny wybór przekaźnika , wyrównanie zakłóceń, różnorodność transmisji w pętli zamkniętej, wybór anteny nadawczej [Yu i in., 2017], dostęp z ortogonalnym podziałem częstotliwości, transmisja Coordinated Multi-Point (CoMP), bezpieczeństwo warstwy fizycznej, zarządzanie mobilnością, aby wymienić tylko kilka. Aby rozwiązać problem starzenia się kanału, w literaturze zaproponowano wiele algorytmów i protokołów łagodzących. Te techniki albo pasywnie kompensują utratę wydajności za cenę ograniczonych zasobów radiowych, albo mają na celu osiągnięcie pełnego potencjału wydajności tylko częściowo, projektując system przy założeniu przestarzałego CSI. Z kolei alternatywna metoda znana jako predykcja kanału otwiera nowy sposób na wydajne i skuteczne podejście do bezpośredniej poprawy dokładności CSI bez marnowania zasobów radiowych, przyciągając wiele uwagi. Dwie oparte na modelach techniki predykcyjne, a mianowicie model autoregresyjny (AR) i model parametryczny , zostały zastosowane poprzez statystyczne modelowanie kanałów bezprzewodowych. Metoda predykcji AR modeluje kanał bezprzewodowy jako proces autoregresyjny i wyprowadza stan kanału w następnym czasie, wykorzystując ważoną kombinację liniową przeszłych i bieżących stanów kanału. Chociaż model AR jest prosty, jest podatny na szum i propagację błędów, co czyni go nieatrakcyjnym w predykcji dalekiego zasięgu. Model parametryczny zakłada, że zanikający kanał jest superpozycją kilku złożonych sinusoid, a jego parametry, np. tłumienie, fazy, kąty przestrzenne, przesunięcia Dopplera i liczba rozproszeń, zmieniają się znacznie wolniej w stosunku do szybkości zanikania kanałów i można je dokładnie ocenić. Oprócz żmudnego procesu szacowania, szacowane parametry wkrótce wygasną w kanale zmieniającym się w czasie i dlatego muszą być ponownie szacowane iteracyjnie, co prowadzi do wysokiej złożoności obliczeniowej. W marcu 2016 r., gdy AlphaGo, program komputerowy opracowany przez Google Deep-Mind, odniósł miażdżące zwycięstwo nad ludzkim mistrzem w grze Go, pasja do eksploracji technologii sztucznej inteligencji (AI) rozbudziła się niemal we wszystkich dziedzinach nauki i inżynierii. Właściwie społeczność badawcza zajmująca się sieciami bezprzewodowymi zaczęła stosować techniki AI do rozwiązywania problemów komunikacyjnych dawno temu. Wykorzystując możliwość predykcji szeregów czasowych , klasyczna technika AI zwana rekurencyjną siecią neuronową (RNN) została zastosowana do tworzenia predyktorów kanałów dla kanałów zanikających o płaskiej częstotliwości i pojedynczej antenie , a następnie rozszerzona na kanały zanikające selektywnie pod względem częstotliwości MIMO . Ostatnio zbadano również wykonalność i skuteczność zastosowania głębokiej sieci neuronowej opartej na pamięci krótkoterminowej (LSTM) i bramkowanej jednostce rekurencyjnej

(GRU) do przewidywania zanikających kanałów . Ta sekcja zapewni czytelnikom kompleksowy pogląd na starzenie się kanałów, predykcję kanałów i podstawy wykorzystania głębokiego uczenia się do przewidywania zanikających kanałów.

Nieaktualne informacje o stanie kanału

Z powodu opóźnień sprzężenia zwrotnego i przetwarzania istnieje luka między momentem, w którym sygnały pilotażowe badają kanały łącza w górę, a momentem, w którym następuje transmisja danych łącza w dół, z przesunięciami fazowymi elementów IRS dostrojonymi pod względem zmierzonego CSI. Zmierzone CSI może być nieaktualne w wyniku wahań kanałów wywołanych głównie przesunięciami Dopplera (z powodu mobilności użytkownika i wysokiej częstotliwości), a także szumem fazowym w transceiverze. Przypomnijmy, że wzmocnienie odczepu kanału

$$h_l[t] = \sum_i a_i(tT_s)e^{-j2\pi f_c \tau_i(tT_s)}\text{sinc}\left[l - \frac{\tau_i(tT_s)}{T_s}\right] \quad (138)$$

dla l = 0,…, L − 1, przy częstotliwości nośnej fc, próbkowanym tłumieniu $a_i(tT_s)$ i opóźnieniu $\tau_i(tT_s)$ i-tej ścieżki sygnału, okresie próbkowania T_s i sinc(x) ≜ sin(x) x dla x ≠ 0. Aby modelować, jak szybko odczepy $h_l[t]$ ewoluują w czasie t, zdefiniowano wielkość statystyczną znaną jako funkcja autokorelacji wzmocnienia odczepu jako

$$R_l[\tau] = \mathbb{E}[h_l^*[t]h_l[t+\tau]]. \quad (139)$$

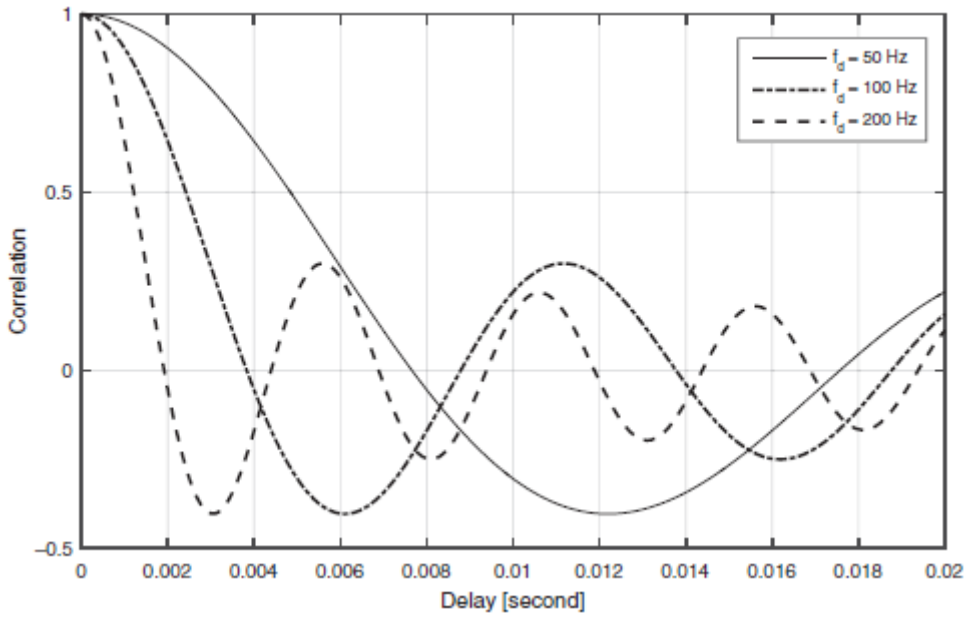
W przypadku klasycznego widma Dopplera modelu Jakesa funkcja autokorelacji przyjmuje wartość

$$R_l[\tau] = J_0(2\pi f_d \tau), \quad (140)$$

gdzie f_d oznacza maksymalne przesunięcie Dopplera, τ oznacza opóźnienie między przestarzałym a rzeczywistym CSI, a $J_0(\cdot)$ oznacza funkcję Bessela zerowego rzędu pierwszego rodzaju. W szczególności maksymalną częstotliwość Dopplera można obliczyć za pomocą

$$f_d = \frac{f_c v}{c} = \frac{v}{\lambda}, \quad (141)$$

gdzie v oznacza prędkość poruszającego się obiektu, c jest prędkością światła w wolnej przestrzeni, a λ reprezentuje długość fali częstotliwości nośnej. Aby zapewnić konkretny obraz, wartości autokorelacji kanałów zanikających z przesunięciami Dopplera o 50, 100 i 200 Hz są pokazane na rysunku.



Przesunięcie Dopplera

Dla uproszczenia ignorujemy indeksy czasu i oznaczamy rzeczywistą realizację kanału MIMO flat-fading przez H = [h_{nrnt}]$_{Nr×Nt}$ i jego przestarzałą wersję H' = [h'$_{nrnt}$]$_{Nr×Nt}$. W kontekście IRS ten kanał MIMO może modelować połączenia między wieloantenową stacją bazową a wieloantenowym UE, między wieloantenową stacją bazową a IRS lub między IRS a wieloantenowym UE. Współczynnik korelacji jest używany do ilościowego określenia niedokładności przestarzałego CSI w niezależnych i identycznie rozłożonych (i.i.d.) kanałach, tj.

$$\rho = \frac{|cov(h_{n_r n_t}, h'_{n_r n_t})|}{\mu_h \mu_{h'}}, \quad (142)$$

gdzie co$v(\cdot)$ oznacza kowariancję dwóch zmiennych losowych, μ oznacza odchylenie standardowe, a h_{nrnt} oznacza wzmocnienie kanału między anteną nadawczą n_t a anteną odbiorczą n_r lub n-tym elementem odbijającym. Ze względu na elementy i.i.d. w H i H', ρ jest niezależne od n_r i n_t. Zatem równanie (142) można uprościć do

$$\rho = \frac{|cov(h, h')|}{\mu_h \mu_{h'}}. \quad (143)$$

Ponieważ elementy HandH' są, według Ramyi i Bhashyama [2009], obydwa mają rozkład Gaussa symetryczny kołowo-średnicowy, ich związek można podać wzorem

$$\mathbf{H}' = \rho \mathbf{H} + \sqrt{1-\rho^2}\mathbf{E} \quad (144)$$

gdzie E = [ε_{nrnt}]$_{Nr×Nt}$ jest macierzą składającą się ze znormalizowanych zmiennych losowych Gaussa, tj. $\varepsilon_{n_r n_t} \sim \mathcal{CN}(0,1)$. W przypadku płaskiego zanikania częstotliwości w układzie wąskopasmowym pojedyncze dotknięcie i jego przestarzała wersja są oznaczane odpowiednio przez h i h'. Zgodnie z tym mamy

$$h' = \sigma_{h'} \left(\frac{\rho}{\sigma_h} h + \varepsilon \sqrt{1 - \rho^2} \right) \quad (145)$$

gdzie ε jest zmienną losową o rozkładzie standardowym normalnym $\varepsilon \sim \mathcal{CN}(0,1)$, a $\sigma^2_{h'}$ jest wariancją h'. Przyjmując model Jakesa, H i H' podążają za wspólnym złożonym rozkładem Gaussa, gdzie współczynnik korelacji przyjmuje wartość $\rho = J_0(2\pi f_d \tau)$. Zatem H warunkowane przez H' podąża za rozkładem Gaussa, który jest modelowany przez

$$\mathbf{H}|\mathbf{H}' \sim \mathcal{CN}(\rho \mathbf{H}', 1 - \rho^2).$$

Ze względu na założenie znormalizowanego wzmocnienia kanału $\mathbb{E}[|h|^2] = 1$, średni SNR $\bar{\gamma} = \mathbb{E}\left[\frac{|h|^2 P_t}{\sigma_n^2}\right]$ zostaje uproszczony d $\bar{\gamma} = P_t / \sigma_n^2$. Zatem natychmiastowy SNR $\gamma = ||H||^2 P_t / \sigma^2_n$ można zapisać jako $\gamma = ||\mathbf{H}||^2 \bar{\gamma}$. Uwarunkowany przestarzałą wersją $\gamma' = ||\bar{\mathbf{H}'}||^2 \bar{\gamma}$, γ podąża za niecentralnym rozkładem chi-kwadrat z dwoma stopniami swobody.

$$f_{\gamma|\gamma'}(\gamma|\gamma') = \frac{1}{\bar{\gamma}(1-\rho^2)} e^{-\frac{\gamma + \rho^2 \gamma'}{\bar{\gamma}(1-\rho^2)}} J_0\left(\frac{2\sqrt{\rho^2 \gamma \gamma'}}{\bar{\gamma}(1-\rho^2)} \right) \quad (146)$$

Szum fazowy

Z powodu niedoskonałych oscylatorów w nadajniku, przesyłane sygnały cierpią na szum fazowy podczas przetwarzania konwersji w górę z sygnałów pasma podstawowego do pasma przepustowego i odwrotnie w odbiorniku. Taki szum fazowy jest nie tylko losowy, ale także zmienny w czasie, co prowadzi do przestarzałego CSI, który jest równoważny efektowi przesunięcia Dopplera. Wykorzystując dobrze ugruntowany proces Wienera, szum fazowy BS i UE w dyskretnym momencie czasu t można modelować jako

$$\begin{cases} \Delta\phi_t = \phi_t - \phi_{t-1}, & \Delta\phi_t \sim \mathcal{CN}(0, \sigma^2_\phi) \\ \Delta\varphi_t = \varphi_t - \varphi_{t-1}, & \Delta\varphi_t \sim \mathcal{CN}(0, \sigma^2_\varphi) \end{cases} \quad (147)$$

gdzie wariancje przyrostu są podane przez $\sigma^2_i = 4\pi^2 f_c c_i T_s$, $\forall i = \phi, \varphi$ z okresem symbolu T_s i stałą zależną od oscylatora c_i. Ponieważ pasywne elementy odbijające w IRS nie mają żadnych komponentów RF, nie ma szumu fazowego na powierzchni odbijającej.

Wpływ starzenia się kanału na IRS

Z praktycznego punktu widzenia szacowany CSI używany do obliczania optymalnych faz odbijających może znacząco różnić się od rzeczywistego CSI w chwili użycia wybranych faz do odbicia sygnałów. Wykorzystanie przestarzałej wersji CSI zamiast rzeczywistego CSI może poważnie pogorszyć wydajność systemu z ulepszonym IRS. Biorąc pod uwagę wpływ przesunięć Dopplera i szumu fazowego, możemy napisać

$$\mathfrak{h}_d = h_d e^{j(\phi_u + \varphi_u)} \quad (148)$$

oznaczać całkowite wzmocnienie kanału bezpośredniego łącza między BS i UE podczas transmisji sygnału oznaczonego przez czas u, gdzie ϕ_u i φ_u oznaczają odpowiedni szum fazowy BS i UE w chwili u. Jego przestarzała wersja uzyskana podczas procesu szacowania kanału (czas p) jest podana przez

$$\mathfrak{h}'_d = h'_d e^{j(\phi_p+\varphi_p)} = \left(\rho h_d + \varepsilon\sqrt{1-\rho^2}\right) e^{j(\phi_u-\Delta\phi_u+\varphi_u-\Delta\psi_u)} \quad (149)$$

gdzie ϕ_p i φ_p oznaczają odpowiadające szumy fazowe BS i UE w chwili p. Podobnie możemy uzyskać rzeczywisty CSI między BS a n-tym elementem IRS podczas transmisji sygnału

$$\mathfrak{h}_n = h_n e^{j\phi_u} \quad (150)$$

ponieważ elementy IRS są pasywne. Odpowiedni przestarzały CSI w chwili p jest podany przez

$$\mathfrak{h}'_n = h'_n e^{j\phi_p} = \left(\rho h_n + \varepsilon\sqrt{1-\rho^2}\right) e^{j(\phi_u-\Delta\phi_u)}. \quad (151)$$

Rzeczywisty i nieaktualny CSI pomiędzy n-tym elementem IRS a UE jest wyrażony jako

$$\mathfrak{g}_n = g_n e^{j\psi_u}. \quad (152)$$

i

$$\mathfrak{g}'_n = g'_n e^{j\psi_p} = \left(\rho g_n + \varepsilon\sqrt{1-\rho^2}\right) e^{j(\psi_u-\Delta\psi_u)} \quad (153)$$

odpowiednio. W dobrych warunkach, gdy kanały wykazują powolne zanikanie, jeśli przesunięcia Dopplera są niewielkie, a jakość oscylatorów jest wysoka, efekt starzenia się kanału nie jest wyraźny, a utrata wydajności może być pomijalna. W przeciwnym razie wpływ powinien być poważny albo w środowiskach o szybkim zanikaniu, albo przy wykorzystaniu taniego sprzętu. Otrzymany sygnał w UE w systemie z ulepszeniem IRS można wyrazić jako

$$y = \left(\sum_{n=1}^{N} \mathfrak{g}_n c_n \mathfrak{h}_n + \mathfrak{h}_d\right)\sqrt{P_t}s + z. \quad (154)$$

Jednakże BS ma tylko przestarzałe informacje o kanale $\mathfrak{g}'_n$, $\mathfrak{h}'_n$ i $\mathfrak{h}'_d$. Optymalne przesunięcie fazowe dla odbijającego elementu n jest ustawione na

$$\theta_n^{\star} = \mathrm{mod}\,[\psi_d - (\phi_{h,n} + \phi_{g,n}), 2\pi], \quad (155)$$

gdzie $\phi_{h,n}$, $\phi_{g,n}$ i ψ_d oznaczają fazy $\mathfrak{h}'_n$, $\mathfrak{g}'_n$ i $\mathfrak{h}'_d$, odpowiednio, a mod [·] jest operacją modulo. Następnie maksymalny odebrany SNR jest wyrażony jako

$$\gamma_{\max} = \frac{P_t\left|\sum_{n=1}^{N}|g_n||h_n|e^{j\phi_e} + |h_d|e^{j\psi_e}\right|^2}{\sigma_z^2} \quad (156)$$

gdzie używamy $e^{j\phi e}$ i $ej\psi e$ do oznaczenia faz resztkowych (błędów) wynikających ze starzenia się kanału, które niszczą pożądane spójne łączenie i zdecydowanie pogarszają wydajność systemu.

Klasyczna prognoza kanału

Na podstawie typowych metod statystycznych można budować modele predykcyjne, aby przybliżać dynamikę zanikającego kanału przy użyciu zestawu parametrów propagacji. Dzięki znajomości bieżących i przeszłych informacji o kanale, parametry te są wyprowadzane, a następnie CSI w następnym kroku może być ekstrapolowane. Istniejące prognozy kanału oparte na modelu są głównie różnicowane na dwie główne kategorie, tj. modele autoregresyjne i parametryczne. Podstawowe zasady, metody szacowania parametrów i ograniczenia tych dwóch modeli przedstawiono poniżej.

Model autoregresyjny

Wykorzystując autokorelację kanału zmieniającego się w czasie w dziedzinie czasu, ta technika modeluje odpowiedź impulsową kanału jako proces AR i szacuje jego współczynniki przy użyciu filtru Kalmana (KF). Następnie tworzony jest liniowy predyktor w celu ekstrapolacji przyszłego współczynnika kanału poprzez połączenie ważonych bieżących i przeszłych współczynników kanału. Piszemy AR(p), aby oznaczyć złożony proces AR rzędu p, który można oznaczyć za pomocą rekurencji w dziedzinie czasu, tj.

$$x[n] = \sum_{k=1}^{p} a_k x[n-k] + w[n] \quad (157)$$

gdzie w[n] jest zespolonym szumem gaussowskim o zerowej średniej i wariancji σ^2_p, a {a_1, a_2,…, a_p} oznaczają współczynniki AR. Odpowiednia gęstość widmowa mocy (PSD) AR(p) ma postać wymierną

$$S_{xx}(f) = \frac{\sigma_p^2}{\left|1 + \sum_{k=1}^{p} a_k e^{-2\pi jfk}\right|^2}. \quad (158)$$

W kanale zanikającym Rayleigha teoretyczna gęstość widmowa sygnału związana z częściami sygnału zanikającego w fazie lub w kwadraturze ma dobrze znaną postać o kształcie litery U ograniczoną pasmem, tj.

$$S(f) = \begin{cases} \frac{1}{\pi f_d \sqrt{1-\left(\frac{f}{f_d}\right)^2}}, & |f| \leqslant f_d \\ 0, & f > f_d \end{cases} \quad (159)$$

gdzie f_d jest maksymalnym przesunięciem Dopplera w hercach. Odpowiednia funkcja autokorelacji dyskretnego czasu jest podana przez

$$R[n] = J_0(2\pi f_m |n|), \quad (160)$$

ze znormalizowanym maksymalnym przesunięciem Dopplera $f_m = f_d T_s$. Dowolne widmo może być blisko przybliżone przez model AR o wystarczająco dużym rzędzie. Podstawowy związek pomiędzy pożądaną funkcją autokorelacji R[n] a parametrami modelu AR(p) można podać w postaci macierzowej przez

$$\mathbf{v} = \mathbf{Ra}, \quad (161)$$

gdzie

$$\mathbf{R} = \begin{bmatrix} R[0] & R[-1] & \cdots & R[-p+1] \\ R[1] & R[0] & \cdots & R[-p+2] \\ \vdots & \vdots & \ddots & \vdots \\ R[p-1] & R[p-2] & \cdots & R[0] \end{bmatrix}, \quad (162)$$

$$\mathbf{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_p \end{bmatrix}^T, \quad (163)$$

$$\mathbf{v} = \begin{bmatrix} R[1] & R[2] & \cdots & R[p] \end{bmatrix}^T, \quad (164)$$

i

$$\sigma_p^2 = R[0] + \Sigma_{k=1}^{p} a_k R[k]. \quad (165)$$

Podstawiając równania (162)-(164) do równania (161), otrzymujemy współczynniki AR. Następnie możemy zbudować predyktor KF dla kanału zanikającego z pojedynczą anteną o płaskiej częstotliwości, jako

$$\hat{h}[t+1] = \sum_{k=1}^{p} a_k h[t-k+1]. \quad (166)$$

Traktując kanał MIMO jako zbiór niezależnych podkanałów, można również podać predyktor KF dla systemu wieloantenowego za pomocą

$$\hat{\mathbf{H}}[t+1] = \sum_{k=1}^{p} a_k \mathbf{H}[t-k+1]. \quad (167)$$

Ten schemat nie jest optymalny, ponieważ wykorzystuje jedynie korelację czasową poszczególnych podkanałów, ignorując jednocześnie korelacje przestrzenne i częstotliwościowe między wieloma antenami w kanale MIMO. Ponadto ten schemat jest podatny na szum i cierpi na propagację błędów, co ogranicza jego znaczenie w praktyce.

Model parametryczny

Ten schemat modeluje kanał zanikający jako superpozycję skończonej liczby złożonych sinusoid, z których każda ma swoją odpowiednią amplitudę, przesunięcie Dopplera i fazę [Adeogun i in., 2014]. Uzasadnienie opiera się na obserwacji, że parametry wielościeżkowe zmieniają się powoli w porównaniu ze współczynnikiem zaniku kanałów, a przyszły CSI w pewnym zakresie można ekstrapolować, jeśli te parametry są znane. Zgodnie z powszechnie stosowanym modelem sumy sinusoid, kanał MIMO jest wyrażony jako superpozycja źródeł rozpraszania

$$\mathbf{H}(t) = \sum_{p=1}^{P} \alpha_p \mathbf{a}_r(\theta_p) \mathbf{a}_t^T(\phi_p) e^{j\omega_p t}, \quad (168)$$

gdzie α_p jest amplitudą p-tego źródła rozpraszania, ωp oznacza jego przesunięcie Dopplera, θ_p i ϕ_p oznaczają odpowiednio kąty przybycia i odejścia, a_r oznacza wektor odpowiedzi układu anten odbiorczych, podczas gdy at oznacza układ anten nadawczych. Używając jako przykładu jednorodnego układu liniowego z M równomiernie rozmieszczonymi elementami,

jego wektor sterujący można sformułować jako

$$\mathbf{a}(\psi) = \left[1, e^{-j\frac{2\pi}{\lambda}d\sin(\psi)}, \ldots, e^{-j\frac{2\pi}{\lambda}(M-1)d\sin(\psi)}\right]^T \quad (169)$$

gdzie ψ oznacza kąt przybycia lub odejścia, d jest odstępem między antenami, a λ oznacza długość fali częstotliwości nośnej. Przewidywanie kanału MIMO w kategoriach tego modelu jest zasadniczo problemem szacowania parametrów, w którym należy oszacować liczbę źródeł rozpraszania, amplitudę i przesunięcie Dopplera dla każdej ścieżki, a także jej kąty przybycia i odejścia. Innymi słowy, główną pracą przy budowaniu modelu parametrycznego jest ustalenie $\hat{P}$ i $\{\hat{\alpha}_p, \hat{\theta}_p, \hat{\phi}_p, \hat{\omega}_p\}_{p=1}^{\hat{P}}$ przy znajomości liczby dyskretnych próbek wzmocnienia kanału {H[k]|k = 1,…, K}. Procedura szacowania parametrów dla modelu parametrycznego jest podzielona na następujące etapy:

i. Używając K dostępnych macierzy kanałów do utworzenia wystarczająco dużej macierzy wykazującej wymaganą strukturę niezmienniczości translacyjnej we wszystkich wymiarach. Dlatego tworząc macierz blokową $N_rQ \times N_tL$-Hankel, którą można zapisać jako

$$\hat{\mathbf{D}} = \begin{bmatrix} \mathbf{H}[1] & \mathbf{H}[2] & \cdots & \mathbf{H}[S] \\ \mathbf{H}[2] & \mathbf{H}[3] & \cdots & \mathbf{H}[S+1] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}[Q] & \mathbf{H}[Q+1] & \cdots & \mathbf{H}[K] \end{bmatrix} \quad (170)$$

gdzie Q jest rozmiarem macierzy Hankela, a S = K − Q + 1.

ii. Na podstawie przekształconych danych obliczana jest macierz kowariancji zawierająca korelację czasową i przestrzenną. Macierz kowariancji czasowo-przestrzennej $\hat{C}$ jest następnie wyprowadzana jako $\hat{C} = \hat{D}\,\hat{D}^H/(N_tS)$, gdzie $(\cdot)^H$ oznacza transpozycję sprzężenia hermitowskiego.

iii. Następnie liczbę dominujących źródeł rozpraszania można oszacować, stosując kryterium minimalnej długości opisu (MDL), jako

$$\hat{P} = \arg \min_{p=1,\ldots,N_rQ-1} \left[S\log(\lambda_p) + \frac{1}{2}(p^2+p)\log S\right] \quad (171)$$

gdzie λp jest p-tą wartością własną $\hat{C}$.

iv. Struktura niezmienniczości w $\hat{C}$ jest wykorzystywana do wspólnego oszacowania parametrów strukturalnych. Wykorzystując w pełni klasyczne algorytmy estymacji, takie jak MUSIC i Estimation of Signal Parameters by Rotational Invariance Techniques (ESPRIT) , można obliczyć kąty przybycia i odejścia, a także przesunięcia Dopplera, tj. $\{\hat{\theta}_p, \hat{\phi}_p, \hat{\omega}_p\}_{p=1}^{\hat{P}}$.

v. Uzyskano oszacowane parametry strukturalne $\{\hat{\theta}_p, \hat{\phi}_p, \hat{\omega}_p\}_{p=1}^{\hat{P}}$, a następnie razem z with $\hat{P}$, można obliczyć amplitudy zespolone $\{\hat{\alpha}_p\}_{p=1}^{\hat{P}}$.

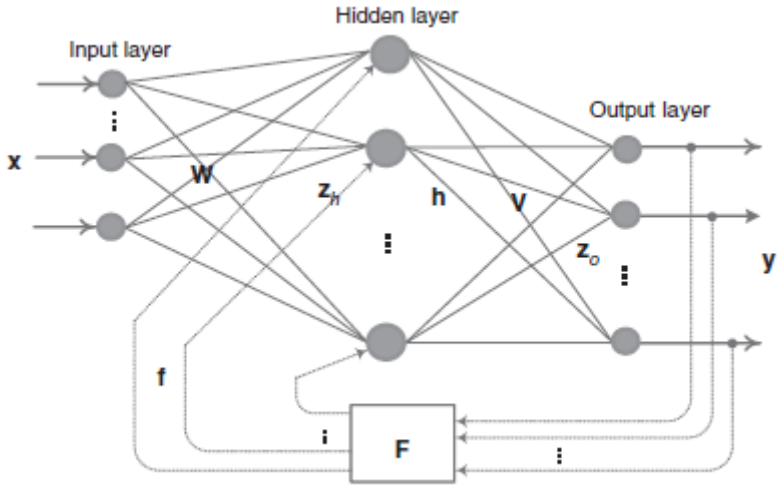
vi. Po ustaleniu wszystkich parametrów, prognozowanie kanału przeprowadza się w następujący sposób

$$\hat{\mathbf{H}}(\tau) = \sum_{p=1}^{\hat{P}} \hat{\alpha}_p \mathbf{a}_r(\hat{\theta}_p)\mathbf{a}_t^T(\hat{\phi}_p)e^{j\hat{\omega}_p\tau} \quad (172)$$

gdzie τ oznacza zakres czasu, dla którego CSI ma być przewidywane. Jak widać, proces szacowania parametrów jest żmudny, co prowadzi do wysokiej złożoności obliczeniowej. Co ważniejsze, szacowane parametry szybko stają się nieważne wraz ze zmianą środowisk propagacji mobilnej, szczególnie w szybko zanikającym kanale. Oznacza to, że te parametry muszą być okresowo szacowane, co jest nieatrakcyjne z praktycznego punktu widzenia. Aby przezwyciężyć słabość konwencjonalnej predykcji kanału, niektóre techniki uczenia maszynowego wykazują duży potencjał. W poniższej części zostaną przedstawione zasady predykcji kanału opartej na uczeniu maszynowym, w tym podstawy rekurencyjnej sieci neuronowej, pamięci długoterminowej i krótkoterminowej oraz bramkowanej jednostki rekurencyjnej w sieciach płytkich i głębokich.

Rekurencyjna sieć neuronowa

Rekurencyjna sieć neuronowa to klasa uczenia maszynowego, która wykazała duży potencjał w dziedzinie przewidywania szeregów czasowych. W przeciwieństwie do sieci typu feed-forward, która uczy się tylko na podstawie danych treningowych, rekurencyjna sieć neuronowa może również wykorzystywać swoją pamięć stanów przeszłych do przetwarzania sekwencji danych wejściowych. RNN ma kilka wariantów, wśród których sieć Jordana jest obecnie używana do budowy predyktora kanału. Zasadniczo prosta sieć składa się z trzech warstw: warstwy wejściowej z neuronami N_i, warstwy ukrytej z neuronami N_h i warstwy mającej N_o wyjść, jak pokazano na rysunku .



Każdemu połączeniu między aktywacją neuronu w warstwie poprzedniej a wejściem neuronu w warstwie następczej przypisuje się wagę. Niech w_{ln} oznacza wagę łączącą n-ty neuron wejściowy i l-ty neuron ukryty, zaś v_{ol} jest wagą neuronu ukrytego l i wyjścia o, gdzie $1 \le n \le N_i$, $1 \le l \le N_h$ i $1 \le o \le N_o$. Konstruowanie macierzy wag $N_h \times N_i$ W jako

$$\mathbf{W} = \begin{bmatrix} w_{11} & \cdots & w_{1N_i} \\ \vdots & \ddots & \vdots \\ w_{N_h1} & \cdots & w_{N_hN_i} \end{bmatrix} \quad (173)$$

i oznaczając wektor aktywacji warstwy wejściowej i składową rekurencyjną (sprzężenie zwrotne) w kroku czasowym t jako

$$\mathbf{x}(t) = \left[x_1(t), x_2(t), \ldots, x_{N_i}(t)\right]^T \quad (174)$$

i

$$\mathbf{f}(t) = \left[f_1(t), f_2(t), \ldots, f_{N_h}(t)\right]^T, \quad (175)$$

odpowiednio, dane wejściowe dla warstwy ukrytej są wyrażone w formie macierzowej przez

$$\mathbf{z}_h(t) = \mathbf{W}\mathbf{x}(t) + \mathbf{f}(t) + \mathbf{b}_h \quad (176)$$

gdzie $b_h = [\ b^h_1, \ldots, b^h_{Nh}\ ]^T$ oznacza wektor odchyleń w warstwie ukrytej. Używając macierzy F do przedstawienia odwzorowania z wyjścia w poprzednim kroku czasowym, tj. $y(t-1) = [\ y_1(t-1), \ldots, y_{No}(t-1)\ ]^T$, na składnik rekurencyjny, mamy

$$\mathbf{f}(t) = \mathbf{F}\mathbf{y}(t-1). \quad (177)$$

Zachowanie sieci neuronowej zależy od funkcji aktywacji, które zazwyczaj mieszczą się w następujących kategoriach: liniowe, prostowane liniowe, progowe, sigmoidalne i styczne. Ogólnie rzecz biorąc, funkcja sigmoidalna jest stosowana do radzenia sobie z nieliniowością, która jest definiowana jako

$$S(x) = \frac{1}{1+e^{-x}}. \quad (178)$$

Podstawiając równanie (177) do równania (178), wektor aktywacji warstwy ukrytej ma postać

$$\mathbf{h}(t) = S(\mathbf{z}_h(t)) = S(\mathbf{W}\mathbf{x}(t) + \mathbf{F}\mathbf{y}(t-1) + \mathbf{b}_h), \quad (179)$$

gdzie S(zh) dla uproszczenia oznacza operację elementarną, tj.

$$S(\mathbf{z}_h) = \left[S(z_1), S(z_2), \ldots, S(z_{N_h})\right]^T. \quad (180)$$

Analogicznie do równania (173) wprowadzono inną macierz wagową V o wymiarze $N_o \times N_h$ z wpisami $\{v_{oi}\}$. Następnie dane wejściowe dla warstwy wyjściowej to $z_o(t) = Vh(t) + b_y$, gdzie b_y jest wektorem odchyleń w warstwie wyjściowej, co daje wektor wyjściowy:

$$\mathbf{y}(t) = S\left(\mathbf{z}_o(t)\right) = S\left(\mathbf{V}\mathbf{h}(t) + \mathbf{b}_y\right) \quad (181)$$

Podobnie jak inne techniki sztucznej inteligencji oparte na danych, działanie RNN jest podzielone na dwie fazy: trenowanie i przewidywanie. Trening sieci neuronowej jest zazwyczaj oparty na szybkim algorytmie znanym jako Back-Propagation (BP). Po otrzymaniu zestawu danych treningowych sieć przekazuje dane wejściowe i porównuje wynikowe dane wyjściowe y z wartością oczekiwaną y0. Błędy predykcji są propagowane wstecz przez sieć, co powoduje iteracyjne aktualizowanie wag i odchyleń, aż do osiągnięcia pewnego warunku zbieżności. Aby zapewnić wstępne wrażenie tego procesu, algorytm BP w połączeniu z nauką gradientu zstępującego dla sieci z przekazem do przodu jest krótko przedstawiony:

Rozpocznij od początkowego stanu sieci, w którym $\{W, V, b_h, b_y\}$ są losowo ustawione.

1. Wprowadź przykład treningowy (x, y_0).

2. Feed-Forward: Dla warstwy ukrytej jej wejście zh i aktywację h można obliczyć odpowiednio za pomocą równań (176) i (179). Dla uproszczenia zilustrowano algorytm BP stosowany w sieci feed forward bez składnika rekurencyjnego. Stąd dokładne równanie do obliczenia wejścia to $zh(t) = W_x(t) + b_h$. Ponadto z_o i y w równaniu (181) dla warstwy wyjściowej są uzyskiwane.

3. Oblicz błąd wyjściowy $e^y = [e^y_1, e^y_2, \ldots, e^y_{No}]^T$ w kategoriach

$$\mathbf{e}^y = \nabla_y C \odot S'(\mathbf{z}_o), \tag{182}$$

gdzie ∇ oznacza wektor, którego elementy są pochodnymi cząstkowymi, mianowicie ∇y C = $\left[\frac{\partial C}{\partial y_1}, \ldots, \frac{\partial C}{\partial y_{N_o}}\right]^T$. Ponadto $S'(z_o)$ oznacza pochodną funkcji aktywacji względem odpowiadającego jej wejścia zo. Biorąc pod uwagę funkcję sigmoidalną w równaniu (178), mamy na przykład

$$S'(\mathbf{z}_o) = \frac{\partial S(\mathbf{z}_o)}{\partial \mathbf{z}_o} = S(\mathbf{z}_o)(1 - S(\mathbf{z}_o)) \tag{183}$$

4. BP: e^y jest propagowane z powrotem do warstwy ukrytej w celu wyprowadzenia tam wektora błędu, tj.

$$\mathbf{e}^h = \mathbf{V}^T \mathbf{e}^y \odot S'(\mathbf{z}_h). \tag{184}$$

5. Metoda gradientu zstępującego: dzięki propagacji wstecznej błędów wagi i odchylenia można aktualizować zgodnie z następującymi zasadami:

$$\begin{cases} \mathbf{W} = \mathbf{W} - \eta \mathbf{e}^h \mathbf{x}^T \\ \mathbf{V} = \mathbf{V} - \eta \mathbf{e}^y \mathbf{h}^T \\ \mathbf{b}_h = \mathbf{b}_h - \eta \mathbf{e}^h \\ \mathbf{b}_y = \mathbf{b}_y - \eta \mathbf{e}^y \end{cases} \tag{185}$$

gdzie η oznacza szybkość uczenia.

Wagi i odchylenia są iteracyjnie aktualizowane, dopóki funkcja kosztu nie spadnie poniżej wstępnie zdefiniowanego progu lub liczba epok nie osiągnie maksymalnej wartości. Po zakończeniu procesu szkolenia wyszkolona sieć może zostać użyta do przetworzenia nadchodzących próbek. Szkolenie RNN zazwyczaj wykorzystuje wariant algorytmu BP, znany jako Back-Propagation Through Time (BPTT). Wymaga on rozwinięcia rekurencyjnej sieci neuronowej w krokach czasowych w celu utworzenia sieci pseudo-feed-forward, w której można zastosować algorytm BP. Na tej podstawie zaprojektowano inne bardziej zaawansowane lub wydajne podejścia, takie jak rekurencyjne uczenie się w czasie rzeczywistym i rozszerzone filtrowanie Kalmana.

Przewidywanie kanału oparte na RNN

Obserwując model kanału MIMO i strukturę sieci neuronowej, można znaleźć duże podobieństwo, które oba mają wiele wejść i wyjść z w pełni ważonymi połączeniami. Sieć neuronowa dobrze nadaje się do przetwarzania kanałów MIMO poprzez dostosowywanie liczby neuronów wejściowych i

wyjściowych w odniesieniu do liczby anten nadawczych i odbiorczych. Predyktor RNN można dość elastycznie skonfigurować do prognozowania odpowiedzi kanału lub obwiedni na żądanie w kanałach o płaskim lub selektywnym zanikaniu częstotliwości. Poniższa dyskusja rozpoczyna się od najprostszego przypadku, w którym zastosowano RNN do przewidywania płaskiego zanikania kanału w systemie pojedynczej anteny, a następnie rozszerza się krok po kroku aż do predyktora domeny częstotliwości dla selektywnych częstotliwościowo kanałów MIMO.

Predykcja płaskiego zanikania kanału

Na początek rozważmy model równoważny pasma podstawowego w czasie dyskretnym dla płaskiego zanikania kanału SISO:

$$r[t] = h[t]s[t] + n[t], \quad (186)$$

Celem predyktora RNN jest uzyskanie przewidywanej wartości $\hat{h}[t+\tau]$, która jest jak najbliższa jej rzeczywistej wartości h[t + τ]. Aby poradzić sobie ze wzmocnieniami kanału o wartościach zespolonych, potrzebna jest sieć z wagami o wartościach zespolonych, zwana dalej RNN o wartościach zespolonych. W czasie t, h[t] jest uzyskiwane poprzez oszacowanie kanału, podczas gdy seria d przeszłych wartości h[t − 1], h[t − 2],…, h[t − d] może być zapamiętana po prostu poprzez odczepioną linię opóźniającą. Te wzmocnienia kanału d + 1 są wprowadzane do RNN jako dane wejściowe, tj.

$$\mathbf{x}[t] = \left[h[t], h[t-1], \ldots, h[t-d]\right]^T. \quad (187)$$

Wraz z opóźnionym sprzężeniem zwrotnym uzyskuje się przewidywanie przyszłego wzmocnienia kanału $\mathbf{y}[t] = \left[\hat{h}[t+1]\right]^T$. Rozszerzenie tego predyktora na kanały MIMO o płaskim zanikaniu jest proste. Aby dostosować się do warstwy wejściowej a RNN, wymagane jest zwektoryzowanie macierzy kanałów do wektora $N_r N_t \times 1$ w następujący sposób:

$$\mathbf{h}[t] = \vec{\mathbf{H}}[t] = \left[h_{11}[t], h_{12}[t], \ldots, h_{N_r N_t}[t]\right]. \quad (188)$$

Razem z liczbą d przeszłych wartości H[t − 1], …,H[t − d], wejście RNN w tym przypadku wynosi

$$\mathbf{x}[t] = [\mathbf{h}[t], \mathbf{h}[t-1], \ldots, \mathbf{h}[t-d]]^T, \quad (189)$$

co daje wartość predykcyjną y[t] = $\hat{h}^T$[t + 1], którą można przekształcić w macierz przewidywaną $\hat{H}$ [t + 1]. W porównaniu z RNN o wartościach zespolonych rekurencyjna sieć neuronowa z wagami o wartościach rzeczywistych, zwana RNN o wartościach rzeczywistych, ma niższą złożoność i wyższą dokładność przewidywania, podczas gdy może ona obsługiwać tylko dane o wartościach rzeczywistych. Na szczęście wzmocnienie kanału o wartościach zespolonych można rozłożyć na dwie wartości rzeczywiste, mianowicie h = hr + jhi. W związku z tym w Jiang i Schotten [2018b] zaproponowano RNN o wartościach rzeczywistych w celu zbudowania prostszego predyktora o wyższej dokładności poprzez rozdzielenie części rzeczywistej i urojonej. Bez konieczności używania dwóch RNN, sztuki rzeczywiste i urojone można przetwarzać łącznie w jednym predyktorze. W tym przypadku dane wejściowe sieci to

$$\mathbf{x}[t] = \left[h^r[t], h^i[t], \ldots, h^r[t-d], h^i[t-d]\right]^T, \quad (190)$$

generowanie wyjścia $\mathbf{y}[t] = \left[\hat{h}^r[t+1], \hat{h}^i[t+1]\right]^T$ który syntetyzuje do przewidywanego wzmocnienia kanału $\hat{h}[t+1] = \hat{h}^r[t+1] + j\hat{h}^i[t+1]$. Podobnie H[t] rozkłada się na

$$\mathbf{H}[t] = \mathbf{H}_R[t] + j\mathbf{H}_I[t], \quad (191)$$

gdzie $H_R = \Re(H) = [h^r_{nrnt}]N_r \times N_t$ oznacza macierz złożoną z części rzeczywistych wzmocnień kanału, a $H_I = \Im(H) = [h^i_{nrnt}]N_r \times N_t$ jest jej urojonym odpowiednikiem. Podobnie jak równanie (188), macierze te są wektoryzowane jako

$$\mathbf{h}_r[t] = \vec{\mathbf{H}}_R[t] = \left[h^r_{11}[t], h^r_{12}[t], \ldots, h^r_{N_r N_t}[t]\right] \quad (192)$$

Transmisja

$$\mathbf{x}[t] = [\mathbf{h}_r[t], \mathbf{h}_i[t], \ldots, \mathbf{h}_r[t-d], \mathbf{h}_i[t-d]]^T \quad (193)$$

do sieci, wynikowy wynik jest zapisywany jako $\mathbf{y}[t] = \left[\hat{\mathbf{h}}_r[t+1], \hat{\mathbf{h}}_i[t+1]\right]^T$ które można przekształcić na $\hat{H}_R[t + D]$ i $\hat{H}_I[t + D]$. Następnie przewidywana macierz jest po prostu zbierana przez $\hat{H}[t + 1] = \hat{H}_R[t + 1] + j\,\hat{H}_I[t + 1]$. Wiele adaptacyjnych systemów transmisyjnych potrzebuje jedynie znać obwiednię odpowiedzi kanału $|h|$, a nie wzmocnienie h o wartościach zespolonych. Dlatego też można bezpośrednio zastosować RNN o wartościach rzeczywistych, co z kolei może obniżyć złożoność, przyspieszyć proces szkolenia i poprawić dokładność przewidywania w porównaniu z przewidywaniem wzmocnień kanału. Obwiednia kanału w czasie t oznaczona jako $|h[t]|$ jest znana, z liczbą d minionych wartości $|h[t-1]|$, $|h[t-2]|$,…, $|h[t-d]|$, wejście w tym przypadku jest zapisane jako

$$\mathbf{x}[t] = \left[|h[t]|, |h[t-1]|, \ldots, |h[t-d]|\right]^T, \quad (194)$$

które generują $|\hat{h}[t+1]|$ przez sieć. Ponadto niech $Q[t] = [|h_{nrnt}[t]|]_{Nr \times Nt}$ oznacza macierz, w której (n_r, n_t)-ty wpis jest obwiednią $h_{nrnt}[t]$ w H[t]. Podobnie Q[t] jest wektoryzowane jako

$$\mathbf{q}[t] = \vec{\mathbf{Q}}[t] = \left[|h_{11}[t]|, |h_{12}[t]|, \ldots, |h_{N_r N_t}[t]|\right] \quad (195)$$

Przy danych wejściowych $x[t] = [q[t], q[t-1],\ldots, q[t-d]]^T$ otrzymujemy przewidywanie $y[t] = \hat{q}[t + 1]$ i dalej przekształcamy je na $\hat{Q}[t + 1]$.

Predykcja kanału zaniku częstotliwości selektywnej

Na początek rozważmy model dyskretnego czasu dla układu SISO częstotliwości selektywnej:

$$r[t] = \sum_{l=0}^{L-1} h_l[t]s[t-l] + n[t], \quad (196)$$

gdzie s i r oznaczają odpowiednio symbol przesłany i odebrany, $h_l[t]$ oznacza l-ty odczep dla filtru kanałowego o zmiennym czasie w czasie t, a n to szum addytywny. Dla uproszczenia pominięto indeks czasu, kanał selektywny pod względem częstotliwości jest modelowany jako liniowy filtr L-odczep

$$\mathbf{h} = [h_0, h_1, \ldots, h_{L-1}]^T, \quad (197)$$

Można go przekształcić w zestaw N ortogonalnych kanałów wąskopasmowych, znanych jako podnośne, poprzez modulację OFDM, która jest reprezentowana przez

$$\tilde{r}_n[t] = \tilde{h}_n[t]\tilde{s}_n[t] + \tilde{n}_n[t], \quad n = 0, 1, \ldots, N-1, \quad (198)$$

gdzie $\tilde{s}_n[t], \tilde{r}_n[t]$ i $\tilde{n}_n[t]$ oznaczają odpowiednio sygnał przesłany, sygnał odebrany i szum na podnośnej n. Zgodnie z efektem płotu pikietowego w dyskretnej transformacie Fouriera, charakterystyka częstotliwościowa filtru kanałowego oznaczona jako $\tilde{\mathbf{h}} = [\tilde{h}_0, \tilde{h}_1, \ldots, \tilde{h}_{N-1}]^T$ jest DFT

$$\mathbf{h}' = [h_0, h_1, \ldots, h_{L-1}, 0, \ldots, 0]^T \quad (199)$$

który uzupełnia h w równaniu (197) zerami N − L na końcu. Rozszerzenie równania (198) na system wieloantenowy jest proste, chociaż system MIMO-OFDM, który jest modelowany jako

$$\tilde{\mathbf{r}}_n[t] = \tilde{\mathbf{H}}_n[t]\tilde{\mathbf{s}}_n[t] + \tilde{\mathbf{n}}_n[t], \quad n = 0, 1, \ldots, N-1, \quad (200)$$

gdzie $\tilde{\mathbf{s}}_n[t]$ reprezentuje wektor symboli transmisji $N_t \times 1$ na podnośnej n w czasie t, $\tilde{\mathbf{r}}_n[t]$ jest wektorem symboli odebranych $N_r \times 1$, a $\tilde{\mathbf{n}}[t]$ jest wektorem szumu addytywnego. Podkanał pomiędzy anteną nadawczą n_t i anteną odbiorczą n_r jest równoważny kanałowi SISO selektywnemu pod względem częstotliwości, oznaczonemu jako filtr kanałowy

$$\mathbf{h}^{n_r n_t} = [h_0^{n_r n_t}, h_1^{n_r n_t}, \ldots, h_{L-1}^{n_r n_t}]^T, \quad (201)$$

Podobnie, charakterystykę częstotliwościową tego filtru można uzyskać poprzez przeprowadzenie DFT, tj.

$$\tilde{\mathbf{h}}^{n_r n_t} = [\tilde{h}_0^{n_r n_t}, \tilde{h}_1^{n_r n_t}, \ldots, \tilde{h}_{N-1}^{n_r n_t}]^T, \quad (202)$$

Następnie macierz kanału na podnośnej n można zapisać jako

$$\tilde{\mathbf{H}}_n[t] = \left[\tilde{h}_n^{n_r n_t}[t]\right]_{N_r \times N_t}. \quad (203)$$

Głównym pomysłem predykcji kanału w dziedzinie częstotliwości jest konwersja kanału selektywnego częstotliwościowo na zestaw ortogonalnych płaskich zanikających podnośnych, a następnie wykorzystanie predyktora w dziedzinie częstotliwości do prognozowania odpowiedzi częstotliwościowej na każdej podnośnej. W czasie t nad podnośną n, $\tilde{H}_n[t]$, a także jego opóźnienia d-krokowe $\tilde{H}_n[t-1]$,…, $\tilde{H}_n[t-d]$, są wprowadzane do RNN. Pomijając indeks czasu dla uproszczenia, te macierze są wektoryzowane jako

$$\tilde{\mathbf{h}}_n = \mathrm{vec}\left(\tilde{\mathbf{H}}_n\right) = [\tilde{h}_n^{11}, \tilde{h}_n^{12}, \ldots, \tilde{h}_n^{N_r N_t}]. \quad (204)$$

RNN wyprowadza prognozę D-step, tj. $\hat{\mathbf{h}}_n[t+D] = [\hat{h}_n^{11}[t+D], \ldots, \hat{h}_n^{N_r N_t}[t+D]]^T$ przekształcając się w przewidywaną macierz $\hat{H}_n[t + D]$ za pośrednictwem modułu wektorowo-tomatrix. Chociaż prognoza jest przeprowadzana na poziomie podnośnych, nie musimy zajmować się wszystkimi N podnośnymi, biorąc pod uwagę korelację częstotliwości kanału. Zintegrowany z systemem wspomaganym pilotem, wystarczy przewidywać CSI tylko na podnośnych niosących symbole pilota. Załóżmy, że jeden pilot jest wstawiany równomiernie co NP podnośnych, co daje łącznie $P = \lfloor N/N_P \rfloor$podnośnych pilota, gdzie $\lfloor \cdot \rfloor$ oznacza funkcję sufitową. Biorąc pod uwagę przewidywane wartości $\hat{H}_p[t + D]$, p = 1,…, P, można zastosować interpolację w dziedzinie częstotliwości, aby uzyskać przewidywane wartości dla wszystkich podnośnych $\hat{H}_n[t + D]$, n = 0,…,N − 1.

Pamięć długo-krótkoterminowa

RNN jest dobra w przetwarzaniu sekwencji danych poprzez przechowywanie nieokreślonych informacji historycznych w swoim stanie wewnętrznym, wykazując duży potencjał w przewidywaniu szeregów czasowych. Niemniej jednak cierpi na problemy z eksplozją i zanikaniem gradientu w technice szkoleniowej BPTT opartej na gradiencie, gdzie wstecznie propagowany sygnał błędu ma tendencję do bycia bardzo dużym, co prowadzi do oscylujących wag lub zmierza do zera, co oznacza niewspółmiernie długi czas szkolenia lub szkolenie w ogóle nie działa. W tym celu Hochreiter i Schmidhuber zaprojektowali elegancką strukturę RNN - pamięć długo-krótkoterminową - w 1997 r. w swojej pionierskiej pracy [Hochreiter i Schmidhuber, 1997]. Kluczową innowacją LSTM w radzeniu sobie z długoterminową zależnością jest wprowadzenie specjalnych jednostek zwanych komórkami pamięci w rekurencyjnej ukrytej warstwie i bramach mnożnikowych, które regulują przepływ informacji. W oryginalnej strukturze LSTM każdy blok pamięci zawiera dwie bramki: bramkę wejściową chroniącą zawartość pamięci przechowywaną w komórce przed zakłóceniami spowodowanymi przez nieistotne zakłócenia oraz bramkę wyjściową kontrolującą zakres, w jakim informacje pamięci są stosowane do generowania aktywacji wyjściowej. Aby rozwiązać słabość LSTM, a mianowicie stan wewnętrzny rośnie w nieskończoność i ostatecznie powoduje awarię sieci podczas przetwarzania ciągłych strumieni wejściowych, które nie są segmentowane na podsekwencje, dodano bramkę zapominania [Gers i in., 2000]. Skaluje ona stan wewnętrzny komórki pamięci przed powrotem przez połączenia samorekurencyjne. Chociaż jej historia nie jest długa, LSTM została pomyślnie zastosowana do zadań przewidywania sekwencji i etykietowania. Osiągnęła już najnowocześniejsze wyniki technologiczne w wielu dziedzinach, takich jak tłumaczenie maszynowe, rozpoznawanie mowy i rozpoznawanie pisma ręcznego, a także osiągnęła wielki sukces komercyjny, uzasadniony wieloma bezprecedensowymi inteligentnymi usługami, takimi jak Google Tłumacz i Apple Siri. Podobnie jak głęboka sieć RNN składająca się z wielu rekurencyjnych warstw ukrytych, głęboka sieć LSTM jest budowana przez układanie w stosy wielu warstw LSTM. Bez utraty ogólności, Rysunek 7.12 pokazuje przykład głębokiej sieci LSTM, która składa się z warstwy wejściowej, trzech warstw ukrytych i warstwy wyjściowej. W dowolnym kroku czasowym wektor danych x przechodzi przez wejściową warstwę sprzężenia zwrotnego, aby uzyskać $d^{(1)}$, która jest aktywacją komórek pamięci w pierwszej ukrytej warstwie. Wraz z jednostką rekurencyjną sprzężenia zwrotnego z poprzedniego kroku czasowego, $d^{(2)}$ jest generowane, a następnie przekazywane do drugiej ukrytej warstwy. Ten rekurencyjny proces trwa, aż warstwa wyjściowa uzyska y zgodnie z $d^{(4)}$. Rozwijając sieć w czasie, blok pamięci w l-tej ukrytej warstwie ma dwa stany wewnętrzne w kroku czasowym t − 1, tj. stan krótkotrwały $s^{(l)}{}_{t-1}$ i stan długotrwały $c^{(l)}{}_{t-1}$. Przechodząc przez komórki pamięci od lewej do prawej, c(l) t−1 najpierw wyrzuca stare wspomnienia w bramce zapomnienia, integruje nowe informacje wybrane przez bramkę wejściową, a następnie wysyła jako bieżący stan długoterminowy $c^{(l)}{}_t$. Wektor wejściowy $d^{(l)}t$ i poprzednia pamięć krótkotrwała $s^{(l)}{}_{t-1}$ są wprowadzane do czterech różnych warstw w pełni połączonych (FC), generując wektory aktywacji bramek jako

$$\mathbf{f}_t^{(l)} = \sigma_g\left(\mathbf{W}_f^{(l)}\mathbf{d}_t^{(l)} + \mathbf{U}_f^{(l)}\mathbf{s}_{t-1}^{(l)} + \mathbf{b}_f^{(l)}\right) \quad (205)$$

$$\mathbf{i}_t^{(l)} = \sigma_g\left(\mathbf{W}_i^{(l)}\mathbf{d}_t^{(l)} + \mathbf{U}_i^{(l)}\mathbf{s}_{t-1}^{(l)} + \mathbf{b}_i^{(l)}\right) \quad (206)$$

$$\mathbf{o}_t^{(l)} = \sigma_g\left(\mathbf{W}_o^{(l)}\mathbf{d}_t^{(l)} + \mathbf{U}_o^{(l)}\mathbf{s}_{t-1}^{(l)} + \mathbf{b}_o^{(l)}\right) \quad (207)$$

gdzie W i U są macierzami wagowymi dla warstw FC, b oznacza wektory polaryzacji, indeksy dolne f, i i o są powiązane odpowiednio z bramką zapomnienia, wejściową i wyjściową, a σg oznacza funkcję aktywacji sigmoidalnej

$$\sigma_g(x) = \frac{1}{1+e^{-x}}. \quad (208)$$

Usunięcie kilku starych wspomnień przy bramce zapomnienia i dodanie kilku nowych informacji wybranych z bieżącego wejścia pamięci, które jest zdefiniowane jako

$$\mathbf{g}_t^{(l)} = \sigma_h\left(\mathbf{W}_g^{(l)}\mathbf{d}_t^{(l)} + \mathbf{U}_g^{(l)}\mathbf{s}_{t-1}^{(l)} + \mathbf{b}_g^{(l)}\right) \quad (209)$$

poprzednia pamięć długotrwała $c_{(l)t-1}$ jest w ten sposób przekształcana w

$$\mathbf{c}_t^{(l)} = \mathbf{f}_t^{(l)} \otimes \mathbf{c}_{t-1}^{(l)} + \mathbf{i}_t^{(l)} \otimes \mathbf{g}_t^{(l)} \quad (210)$$

gdzie $\otimes$ oznacza iloczyn Hadamarda (mnożenie elementów) dla macierzy, a σh jest funkcją tangensa hiperbolicznego oznaczaną jako tanh, definiowaną przez

$$\sigma_h(x) = \frac{e^{2x}-1}{e^{2x}+1} \quad (211)$$

Oprócz funkcji sigmoidalnych i tanh istnieją inne powszechnie stosowane funkcje aktywacji, np. Rectified Linear Unit (ReLU), którą można zapisać jako

$$\sigma_r(x) = \max(0, x) \quad (212)$$

która zwraca 0, jeśli otrzyma jakiekolwiek ujemne wejście, ale dla każdej dodatniej wartości x zwraca tę wartość. Ponadto c_t przechodzi przez funkcję tanh, a następnie jest filtrowany przez bramkę wyjściową, aby wytworzyć bieżącą pamięć krótkotrwałą, a także wyjście dla tego bloku pamięci, tj.

$$\mathbf{s}_t^{(l)} = \mathbf{d}_t^{(l+1)} = \mathbf{o}_t^{(l)} \otimes \sigma_h\left(\mathbf{c}_t^{(l)}\right). \quad (213)$$

Od czasu pojawienia się LSTM jego oryginalna struktura nadal ewoluuje. Cho i inni przedstawiają uproszczoną wersję z mniejszą liczbą parametrów w 2014 r., znaną jako bramkowana jednostka rekurencyjna lub GRU, która wykazuje jeszcze lepszą wydajność niż LSTM w przypadku niektórych mniejszych zestawów danych. W bloku pamięci GRU stany krótkoterminowy i długoterminowy są

scalane w jeden, a pojedyncza bramka $z^{(l)}_t$ jest używana do zastąpienia bramek forget i input, mianowicie:

$$\mathbf{z}_t^{(l)} = \sigma_g\left(\mathbf{W}_z^{(l)}\mathbf{d}_t^{(l)} + \mathbf{U}_z^{(l)}\mathbf{s}_{t-1}^{(l)} + \mathbf{b}_z^{(l)}\right) \quad (214)$$

Bramka wyjściowa została usunięta, ale wprowadzono nowy stan pośredni $r^{(l)}_t$, tj.

$$\mathbf{r}_t^{(l)} = \sigma_g\left(\mathbf{W}_r^{(l)}\mathbf{d}_t^{(l)} + \mathbf{U}_r^{(l)}\mathbf{s}_{t-1}^{(l)} + \mathbf{b}_r^{(l)}\right) \quad (215)$$

Podobnie ukryty stan w poprzednim kroku czasowym przechodzi przez komórki pamięci, usuwa część starej pamięci i ładuje pewne aktualne informacje, co skutkuje bieżącym stanem:

$$\mathbf{s}_t^{(l)} = (1 - \mathbf{z}_t^{(l)}) \otimes \mathbf{s}_{t-1}^{(l)} + \mathbf{z}_t^{(l)} \otimes \sigma_h\left(\mathbf{W}_s^{(l)}\mathbf{d}_t^{(l)} + \mathbf{U}_s^{(l)}(\mathbf{r}_t^{(l)} \otimes \mathbf{s}_{t-1}^{(l)}) + \mathbf{b}_s^{(l)}\right). \quad (216)$$

Głębokie uczenie się oparte na prognozowaniu kanału

Aby rzucić światło na działanie predyktora łącza w dół (DL), badany jest system MIMO typu punkt-punkt z płaskim zanikaniem z antenami nadawczymi Nt i odbiorczymi Nr. Jego transmisja jest modelowana jako

$$\mathbf{r}[t] = \mathbf{H}[t]\mathbf{s}[t] + \mathbf{n}[t], \quad (217)$$

gdzie r[t] i s[t] oznaczają odpowiednio wektory sygnału odebranego i przesłanego w kroku czasowym t, n[t] jest wektorem szumu addytywnego, a H[t] reprezentuje macierz kanału $N_r \times N_t$, której wejście (n_r, n_t) $h_{nrn}t$ jest zespolonym wzmocnieniem kanału między anteną nadawczą nt a anteną odbiorczą nr. Nadajnik wymaga sprzężenia zwrotnego CSI, aby dostosować swoje parametry transmisji do kanału zmieniającego się w czasie. Ze względu na opóźnienie sprzężenia zwrotnego τ, gdy nadajnik używa H[t] do wyboru parametrów, natychmiastowe wzmocnienie kanału zmienia się już na H[t + τ]. Prawdopodobnie H[t] ≠ H[t + τ], szczególnie w środowisku o wysokiej mobilności. Przestarzałe CSI powoduje poważne straty wydajności w przypadku szerokiej gamy adaptacyjnych technik bezprzewodowych. Dlatego warto przeprowadzić predykcję kanału w odbiorniku, aby uzyskać przewidywany CSI $\hat{H}$ [t + D], gdzie D $\geqslant \tau$, aby przeciwdziałać efektowi opóźnienia sprzężenia zwrotnego lub, co jest równoważne, wykonać predykcję w nadajniku. Natychmiastowa macierz kanału H[t] jest szacowana w odbiorniku, który jest wprowadzany do predyktora, a nie jest wprowadzana bezpośrednio do nadajnika, jak to ma miejsce w przypadku typowego adaptacyjnego systemu MIMO. Na podstawie obserwacji, że wielkość wzmocnienia kanału, tj. $|h_{nrnt}|$, jest już wystarczająca dla większości zadań adaptacyjnych, predykcja jest stosowana do takich rzeczywistych danych kanałowych. Aby dostosować warstwę wejściową sieci neuronowej, dodajemy warstwę wstępnego przetwarzania danych w predyktorze, gdzie macierz kanału jest przenoszona do wektora wielkości kanału, takiego jak:

$$\mathbf{H}[t] \rightarrow \left[|h_{11}[t]|, |h_{12}[t]|, \ldots, |h_{N_r N_t}[t]|\right]^T \quad (218)$$

Zastępując xt na powyższym rysunku tym wektorem kanału i przechodząc przez szereg ukrytych warstw, warstwa wyjściowa generuje $\hat{H}$ [t + D], co jest prognozą D-kroku. Oprócz wielkości, predyktor może być również stosowany do przetwarzania wzmocnień kanału o wartościach zespolonych. W tym

celu predyktor musi być zazwyczaj zbudowany na głębokiej sieci neuronowej o wartościach zespolonych, która nie jest dobrze zaimplementowana w obecnych narzędziach oprogramowania AI. Zamiast tego sieć o wartościach rzeczywistych jest stosowana do przewidywania części rzeczywistych i urojonych wzmocnień kanału, gdzie warstwa wstępnego przetwarzania danych przekształca H[t] w

$$[\Re(h_{11}[t]), \ldots, \Re(h_{N_r N_t}[t]), \Im(h_{11}[t]), \ldots, \Im(h_{N_r N_t}[t])]^T \quad (219)$$

gdzie$\Re(\cdot)$ i$\Im(\cdot)$ oznaczają odpowiednio jednostki rzeczywiste i urojone. Wprowadzenie tego wektora do predyktora w czasie t,$\hat{H}$ [t + D] można uzyskać po prostu łącząc przewidywane jednostki rzeczywiste i urojone. Oprócz kanałów o płaskim zanikaniu, predyktor można również zastosować do kanałów selektywnych częstotliwościowo, po prostu konwertując go na zestaw N wąskopasmowych podnośnych, na przykład poprzez modulację OFDM. Przy n-tej podnośnej transmisja sygnału jest reprezentowana przez

$$\tilde{\mathbf{r}}_n[t] = \tilde{\mathbf{H}}_n[t]\tilde{\mathbf{s}}_n[t] + \tilde{\mathbf{n}}_n[t], \quad n = 0, 1, \ldots, N-1 \quad (220)$$

gdzie $\tilde{\mathbf{r}}_n[t]$ reprezentuje N_r odebranych symboli dla podnośnej n w czasie t, $\tilde{\mathbf{s}}_n[t]$ odpowiada N_t symbolom transmisji, $\tilde{\mathbf{n}}[t]$ jest wektorem szumu addytywnego. Piszemy $\tilde{H}_n[t] = [\tilde{h}_n^{n_r n_t}[t]]_{N_r \times N_t}$ aby oznaczyć macierz kanału w dziedzinie częstotliwości, gdzie $1 \leqslant n_r \leqslant N_r$, $1 \leqslant n_t \leqslant N_t$, a $\tilde{h}_n^{n_r n_t} \in \mathbb{C}^{1\times 1}$ oznacza współczynnik CFR na podnośnej n między anteną nadawczą n_t a anteną odbiorczą n_r. Proces predykcji na każdej podnośnej jest taki sam jak w przypadku kanału o płaskim zanikaniu.

Podsumowanie

Dzięki dużej liczbie małych, pasywnych i niedrogich elementów odbijających, IRS może proaktywnie tworzyć inteligentne i programowalne środowisko bezprzewodowe. W ten sposób zapewnia nowy stopień swobody w projektowaniu systemów bezprzewodowych 6G, umożliwiając zrównoważony wzrost pojemności i wydajności przy przystępnych kosztach, niskiej złożoności i niskim zużyciu energii. W tym rozdziale przedstawiono model systemu i transmisję sygnału systemów wspomaganych przez IRS z pojedynczą lub wieloma antenowymi stacjami bazowymi zarówno w kanałach o płaskiej częstotliwości, jak i kanałach o selektywnym zanikaniu częstotliwości. Przedstawiono również wspólną optymalizację aktywnego i pasywnego formowania wiązki oraz transmisji IRS z podwójną wiązką. Ponadto przedstawiono wpływ starzenia się kanału na IRS i predykcję kanału opartą na uczeniu maszynowym. Ta część może służyć jako samouczek dla czytelników jako punkt wyjścia do dalszych prac badawczych nad tym obiecującym tematem.