



INSTITUTO LATINO-AMERICANO DE
TECNOLOGIA, INFRAESTRUTURA E
TERRITÓRIO (ILATIT)

ENGENHARIA DE ENERGIA

Vehicle-to-Grid (V2G) no mercado brasileiro de energia: Uma abordagem via *Reinforcement Learning*

Luan Carlos Covalski Cozer

Foz do Iguaçu
2026



INSTITUTO LATINO-AMERICANO DE
TECNOLOGIA, INFRAESTRUTURA E
TERRITÓRIO (ILATIT)

ENGENHARIA DE ENERGIA

Vehicle-to-Grid (V2G) no mercado brasileiro de energia: Uma abordagem via *Reinforcement Learning*

Luan Carlos Covalski Cozer

Monografia apresentada ao curso de Engenharia de Energia da Universidade Federal da Integração Latino-Americana (UNILA), como requisito parcial à obtenção do título de Bacharel em Engenharia de Energia.

Orientador: Prof. Dr. João Manoel Lenz Vianna da Silva

Foz do Iguaçu
2026

Luan Carlos Covalski Cozer

Vehicle-to-Grid (V2G) no mercado brasileiro de energia: Uma abordagem via Reinforcement Learning

Monografia apresentada ao curso de Engenharia de Energia da Universidade Federal da Integração Latino-Americana (UNILA), como requisito parcial à obtenção do título de Bacharel em Engenharia de Energia.

BANCA EXAMINADORA

Orientador: Prof. Dr. João Manoel Lenz
Vianna da Silva
UNILA

Prof. Dr. Jorge Javier Gimenez Ledesma
UNILA

Prof. Dr. Marcelo Nepomoceno Kapp
UNILA

Foz do Iguaçu, 08 de julho de 2026.

RESUMO

Este trabalho avalia o desempenho econômico da operação *Vehicle-to-Grid* (V2G) no mercado brasileiro de energia considerando a arbitragem de preços via Preço de Liquidação das Diferenças (PLD), utilizando um agente de Aprendizado por Reforço treinado pelo algoritmo *Proximal Policy Optimization* (PPO). Acoplado a um modelo digital eletroquímico construído no *framework* PyBaMM, o agente otimiza os ciclos de carga e descarga para maximizar o lucro de arbitragem do PLD horário do submercado Sudeste, assumindo a degradação física da bateria como custo interno da operação. O estudo compara duas químicas, a NMC-811, de alta densidade energética, e a LFP, robusta ao ciclamento, ao longo de três perfis de mobilidade residencial, confrontando o desempenho do agente com um carregamento passivo e uma heurística de referência. Os resultados demonstram que a arbitragem V2G com a LFP supera o carregamento passivo nos três perfis, gerando lucro absoluto nos cenários conservador e flexível e praticamente sem encurtar a vida útil da bateria. Por outro lado, a NMC-811 mostra-se viável apenas no perfil intenso, no qual o uso frequente já obriga o veículo a manter a carga elevada. Uma vez que estes ganhos pressupõem o acesso do consumidor ao preço horário do PLD, a adequação regulatória do V2G torna-se um pré-requisito para a sua adoção no Brasil, sendo que a futura inclusão de serviços ancilares poderá ampliar as receitas nos cenários atualmente marginais.

Palavras-chave: *Vehicle-to-Grid*; V2G; Aprendizado por Reforço; Degradação de baterias de íons de lítio; PPO; Mercado de Energia.

ABSTRACT

This work assesses the economic feasibility of Vehicle-to-Grid (V2G) operations in the Brazilian energy market using a Reinforcement Learning agent trained with the Proximal Policy Optimization (PPO) algorithm. Coupled with an electrochemical digital twin built in the PyBaMM framework, the agent optimizes charge and discharge cycles to maximize arbitrage profit from the hourly Settlement Price of Differences (PLD) in the Southeast submarket, internalizing physical battery degradation as an operating cost. The study compares two battery chemistries, the high-energy-density NMC-811 and the cycling-robust LFP, across three residential mobility profiles, benchmarking the agent against passive charging and a reference heuristic. Results indicate that V2G arbitrage with LFP outperforms passive charging across all profiles, yielding positive absolute profit in conservative and flexible scenarios without significantly shortening battery lifespan. Conversely, NMC-811 is only viable in the intensive profile, where heavy usage inherently requires maintaining a high state of charge. Since these financial gains require consumer access to hourly PLD rates, updating Brazilian regulations is a prerequisite for V2G adoption. Furthermore, integrating ancillary services could boost revenues in currently marginal scenarios.

Keywords: Vehicle-to-Grid; V2G; Reinforcement Learning; Lithium-ion battery degradation; PPO; Electricity market.

SUMÁRIO

1	Introdução	1
1.1	Objetivos	4
1.1.1	Objetivo Geral	4
1.1.2	Objetivos Específicos	4
2	Fundamentação Teórica	5
2.1	Veículos Elétricos e o V2G	5
2.1.1	Tipos de Veículos Elétricos	5
2.1.2	A Bateria de Íons de Lítio	5
2.1.2.1	Princípio de funcionamento	5
2.1.2.2	Estrutura hierárquica: célula, módulo e <i>pack</i>	6
2.1.2.3	Parâmetros operacionais essenciais	6
2.1.3	Mecanismos de Degradação	7
2.1.4	Químicas de Cátodo NMC e LFP	9
2.1.4.1	A família NMC	9
2.1.4.2	A química LFP (olivina)	9
2.1.5	Vehicle-to-Grid: Conceito e Justificativa	10
2.1.6	Arquitetura, Infraestrutura e Comunicação	11
2.1.6.1	Carregador bidirecional e fluxo de potência	11
2.1.6.2	Comunicação com a rede e protocolos	11
2.1.6.3	O papel do agregador	12
2.2	Sistema Elétrico Brasileiro e Mercado	12
2.2.1	Estrutura Institucional do Setor Elétrico Brasileiro	12
2.2.2	O Sistema Interligado Nacional e seus Submercados	13
2.2.3	Ambientes de Comercialização: ACR e ACL	13
2.2.4	O Preço de Liquidação das Diferenças (PLD)	14
2.2.5	Regulamentação Aplicável ao V2G: Estado Atual e Perspectivas	14
2.3	Aprendizado por Reforço	16
2.3.1	Inteligência Artificial e Aprendizado de Máquina: Panorama Conceitual	16
2.3.2	O Paradigma da Aprendizado por Reforço	17
2.3.3	Processo de Decisão de Markov	18
2.3.4	Políticas, Funções de Valor e a Equação de Bellman	19
2.3.5	Métodos de Solução - Modelo Conhecido ao Modelo Livre	20
2.3.6	Aprendizado por Reforço Profundo e Métodos <i>Policy-Gradient</i>	20
2.3.7	<i>Proximal Policy Optimization</i> (PPO)	21
2.4	Aprendizado por Reforço Aplicada ao V2G	22
3	Metodologia	25

3.1	Arquitetura Geral do Sistema de Simulação	25
3.1.1	As duas fases computacionais	25
3.1.2	Modelo Digital	26
3.1.3	<i>Framework</i> PyBaMM	27
3.2	Modelagem Eletroquímica da Bateria	27
3.2.1	Modelo DFN e aproximação térmica lumped	27
3.2.2	Mecanismos de degradação	28
3.2.3	Parametrização por química	29
3.3	Geração das <i>Lookup Tables</i> de Degradação	30
3.3.1	Estrutura geral da Fase Offline	30
3.3.2	Discretização do <i>Grid</i> 4D	30
3.3.3	Protocolo de Simulação por Ponto	31
3.3.4	Variáveis Extraídas	32
3.3.5	Monetização da Degradação	32
3.3.6	Consulta em Tempo Real	34
3.4	Formulação do MDP e da Função de Recompensa	35
3.4.1	Processo de Decisão de Markov	35
3.4.2	Modelagem Estocástica do Usuário	36
3.4.3	Espaço de Observações	39
3.4.4	Espaço de Ações e Dinâmica de Atualização	40
3.4.5	Função de Recompensa	42
3.5	Algoritmo de Otimização <i>Proximal Policy Optimization</i>	46
3.5.1	Escolha do Algoritmo	46
3.5.2	Aplicação ao Problema V2G	46
3.5.3	Hiperparâmetros	47
3.5.4	Matriz de Cenários	48
3.5.5	Divisão Temporal Treino / Validação	49
3.6	Dados de Entrada e Cenários de Validação	50
3.6.1	PLD Horário do Submercado Sudeste/Centro-Oeste	50
3.6.2	Procedimento de Simulação Final	51
3.6.3	Métricas Reportadas	52
3.6.4	Uso de ferramentas de inteligência artificial	53
4	Resultados e Discussão	54
4.1	Janela de oportunidade dos perfis	55
4.2	Convergência do treinamento e generalização	56
4.3	Desempenho econômico e comportamento aprendido	57
4.3.1	Arbitragem viável em todos os perfis com a LFP	60
4.3.2	A NMC-811, viável apenas quando a demanda exige carga alta	63
4.3.3	A frota de referência	66

4.3.4	A física por trás dos custos	69
4.3.5	Confronto com a heurística de referência	70
4.4	Impacto da química sobre a vida útil	71
4.5	Síntese dos resultados	73
4.6	Limitações e possibilidades de melhoria	74
5	Considerações Finais	76
	 Referências	 78
	 Apêndices	 83
	APÊNDICE A Parâmetros de Replicação	84

AGRADECIMENTOS

Primeiramente, gostaria de agradecer a Deus, pela vida e pelas oportunidades. Em seguida, agradeço à minha mãe, Helena, e ao meu falecido pai, Danilo, sem os quais nada desta trajetória seria possível, pela luta diária de tornar possível que seus filhos estudem. Aos meus irmãos e irmã, Lucas, Luciano e Fabiana, à minha namorada, Elisa, e à sua família, que me apoiaram muito sempre que precisei ao longo desses anos. Agradeço, também, aos meus amigos que fizeram parte dessa jornada, tornando os dias mais leves e “fáceis” de se viver dentro do ambiente acadêmico. Por fim, agradeço ao meu orientador, João, pelos conselhos e pela honestidade no desenvolvimento deste trabalho.

*“But you know that when the truth is told
That you can get what you want or you can just get old”
– Billy Joel*

LISTA DE ILUSTRAÇÕES

Figura 1 – Diagrama da célula de íons de lítio.	6
Figura 2 – Representação de célula, módulo e <i>pack</i>	7
Figura 3 – Arquitetura conceitual do V2G.	11
Figura 4 – Laço de interação agente-ambiente na Aprendizado por Reforço.	17
Figura 5 – Fluxograma da arquitetura de simulação V2G adotada.	26
Figura 6 – Preço médio horário do PLD no submercado Sudeste frente à probabilidade de conexão de cada perfil de mobilidade ao longo do dia.	55
Figura 7 – Recompensa média por episódio.	56
Figura 8 – Anatomia da decisão do agente em uma janela de 48 horas do cenário flexível LFP.	61
Figura 9 – Anatomia da decisão do agente em uma janela de 48 horas do cenário intenso LFP.	62
Figura 10 – Anatomia da decisão do agente em uma janela de 48 horas do cenário conservador LFP.	63
Figura 11 – Anatomia da decisão do agente em uma janela de 48 horas do cenário intenso NMC-811.	64
Figura 12 – Anatomia da decisão do agente em uma janela de 48 horas do cenário conservador NMC-811.	65
Figura 13 – Treino do agente da frota sob as duas formas de carga de emergência, recuperação a 0,5C e a 1C. (a) recompensa média por episódio, (b) desvio-padrão da política e (c) variância explicada do crítico.	67
Figura 14 – Comparação das três políticas em seis métricas por cenário.	73

LISTA DE TABELAS

Tabela 1 – Parametrização das duas químicas estudadas.	29
Tabela 2 – Discretização do espaço 4D adotada para a <i>lookup table</i>	31
Tabela 3 – Variáveis registradas em cada ponto da <i>lookup table</i>	32
Tabela 4 – Coeficientes da Eq. (3.1) por química.	33
Tabela 5 – Parâmetros estocásticos dos três perfis de mobilidade.	37
Tabela 6 – Hiperparâmetros do PPO adotados neste trabalho.	47
Tabela 7 – Horizonte efetivo do fator de desconto em função do passo $\Delta t = 5$ min.	47
Tabela 8 – Ferramentas de inteligência artificial utilizadas e suas atribuições.	53
Tabela 9 – Lucro líquido médio (R\$ por episódio de trinta dias) do agente avaliado sobre o PLD do ano de treino (2025) e sobre o PLD de um ano distinto (2026).	57
Tabela 10 – Lucro líquido trimestral (R\$) das três políticas e qualidade de serviço do PPO.	58
Tabela 11 – Operação do agente PPO por cenário no trimestre.	59
Tabela 12 – Operação do carregamento passivo por cenário no trimestre.	59
Tabela 13 – Economia marginal do V2G, como a diferença observada entre o agente e o passivo no ganho de mercado e na degradação.	60
Tabela 14 – Comparação do cenário baseline com recuperação 0,5C e 1C.	68
Tabela 15 – Custo de degradação trimestral (R\$) por mecanismo físico do agente PPO.	69
Tabela 16 – PPO x heurística de referência.	70
Tabela 17 – Vida útil estimada até o estado de saúde de 80% (anos).	72
Tabela 18 – Síntese do desempenho por cenário, combinando a economia marginal frente ao passivo com a qualidade de serviço ao usuário.	73
Tabela 19 – Constantes globais do ambiente de simulação.	84
Tabela 20 – Parâmetros das baterias por química.	84
Tabela 21 – Coeficientes da função de recompensa.	85
Tabela 22 – Submodelos do PyBaMM ativados por química.	85
Tabela 23 – Parâmetros do <i>patch</i> de degradação e do modelo térmico.	85
Tabela 24 – Discretização do <i>grid</i> 4D da <i>lookup table</i>	86
Tabela 25 – Parâmetros da distribuição Weibull de distância por viagem.	86

1 INTRODUÇÃO

O setor elétrico mundial atravessa uma transformação estrutural na década de 2020, marcada pela combinação da eletrificação progressiva dos transportes, da climatização de edificações e dos processos industriais a combustíveis fósseis com a expansão acelerada da infraestrutura computacional voltada à inteligência artificial (IA). Em 2024, a demanda global por eletricidade cresceu 4%, equivalente a um acréscimo aproximado de 1.080 TWh em apenas um ano, quase o dobro da média anual da década anterior (International Energy Agency, 2024). Em particular, o consumo elétrico atribuído a centros de processamento de dados (*data centers*), motor da expansão de modelos de IA, deverá saltar de aproximadamente 460 TWh em 2022 para mais de 1.000 TWh em 2030, podendo alcançar 1.300 TWh em 2035 no cenário-base da Agência Internacional de Energia (International Energy Agency, 2024).

O atendimento dessa demanda crescente impõe a expansão paralela do parque gerador, e essa expansão tem se materializado predominantemente por meio de fontes renováveis variáveis. No Brasil, o último Balanço Energético Nacional consolidou em 2024 uma matriz elétrica com 88,2% de participação renovável, um dos maiores índices do mundo (Empresa de Pesquisa Energética, 2025). A capacidade instalada de geração solar fotovoltaica atingiu 48.468 MW, um crescimento de 28,1% sobre o ano anterior, sendo a Micro e Minigeração Distribuída (MMGD) responsável por 35.892 MW desse total. A capacidade eólica chegou a 29.550 MW (Empresa de Pesquisa Energética, 2025). Juntas, essas duas fontes responderam por 23,7% da geração elétrica nacional, configurando uma hiperpenetração de fontes baseadas em conversores eletrônicos (*Inverter-Based Resources – IBR*). A baixa intensidade de carbono associada ao setor energético brasileiro, com apenas 2,0 t CO₂ equivalente por habitante em 2024 e razão de 0,13 kg CO₂ por dólar PPC base 2015, evidencia o elevado ganho ambiental embutido nessa estrutura (Empresa de Pesquisa Energética, 2025).

Apesar do ganho ambiental, a inserção massiva de fontes não-despacháveis altera profundamente a dinâmica operativa do Sistema Interligado Nacional (SIN). Diferente de geradores síncronos convencionais, parques eólicos e solares são acoplados à rede via conversores eletrônicos, o que reduz a inércia rotativa do sistema e compromete sua capacidade de absorver perturbações súbitas. Esses efeitos se manifestam de duas formas operacionais principais. Primeiro, em desperdício direto de energia primária, no qual ao longo de 2024 foram registradas aproximadamente 400 mil horas de restrição de geração (*curtailment* ou Geração Não Realizada Apurada) em usinas solares e eólicas, motivadas tanto por gargalos de transmissão quanto por sobreoferta sistêmica, chegando em prejuízos de R\$ 1,6 bilhão, com 14,6 TWh médios cortados (Portal Solar, 2025). Segundo, em fragilidades dinâmicas evidenciadas em eventos de grande perturbação, onde em 15 de agosto de 2023, uma falha na proteção da Linha de Transmissão de 500 kV Quixadá–Fortaleza II provocou a separação elétrica das regiões Norte e Nordeste do restante do SIN, resultando na interrupção instantânea de aproximada-

mente 23.368 MW de carga – cerca de 27% da carga operante naquele momento – e na falha de fornecimento em 25 estados e no Distrito Federal (Operador Nacional do Sistema Elétrico, 2023). Embora o evento tenha tido como gatilho inicial uma falha de equipamento, sua propagação foi favorecida pela menor reserva síncrona disponível e pela complexidade de coordenação entre subsistemas com alta penetração de fontes intermitentes.

Para mitigar esses efeitos, sistemas elétricos modernos vêm incorporando serviços ancilares de resposta rápida, com destaque para o controle primário e secundário de frequência, a reserva girante (*spinning reserve*) e a regulação de tensão. Kempton e Tomić (2005), em trabalho seminal, descreveram que tais serviços têm valor econômico desproporcional ao volume energético envolvido – respondem por aproximadamente 5 a 10% do custo elétrico total, mesmo movimentando uma fração pequena da energia comercializada – o que os torna particularmente atrativos para recursos de armazenamento bidirecional. Em mercados maduros, tais serviços são monetizados em produtos específicos que remuneram tanto a disponibilidade quanto a efetiva prestação do serviço pelo agente (KEMPTON; TOMIĆ, 2005). No contexto brasileiro, parte expressiva desses serviços – em especial o controle primário de frequência – historicamente foi tratada como obrigação técnica não remunerada para grandes geradores. Esse quadro começa a se reconfigurar com a Resolução Normativa nº 1.030/2022 da ANEEL, que abriu espaço para os chamados Sandboxes Regulatórios, e com a Consulta Pública nº 176/2024 do Ministério de Minas e Energia, que prepara a contratação de Sistemas de Armazenamento de Energia em Baterias (BESS – *Battery Energy Storage Systems*) em Leilões de Reserva de Capacidade (Agência Nacional de Energia Elétrica, 2022; Ministério de Minas e Energia, 2024).

Os BESS estacionários são, conceitualmente, a forma mais natural de prover esses serviços, pois respondem em frações de segundo a comandos de potência e possuem plena despachabilidade. Contudo, seu custo de capital por unidade de potência permanece elevado, o que tem limitado sua adoção em larga escala. Dessa forma, surge uma alternativa estruturalmente análoga porém economicamente mais favorável, de utilizar a frota crescente de veículos elétricos (VEs) – cujas baterias permanecem ociosas em estacionamentos em torno de 90% do tempo (DUBARRY *et al.*, 2018) – como recursos distribuídos de armazenamento bidirecional, o que é a essência da tecnologia denominada *Vehicle-to-Grid* (V2G), proposta originalmente por Kempton e Tomić (2005) e definida como a capacidade de uma frota de veículos elétricos trocar potência ativa com a rede em ambas as direções, sob coordenação de um agregador. Ou seja, o V2G converte cada veículo elétrico em um BESS móvel, com a vantagem estrutural de que o ativo principal – a bateria – tem seu custo de capital amortizado primariamente pela função veicular.

A pertinência do V2G no contexto brasileiro é reforçada pelos movimentos simultâneos do crescimento expressivo da frota nacional de veículos elétricos e a manutenção de uma das matrizes elétricas mais limpas do mundo, conforme já apresentado. Segundo a Associação Brasileira do Veículo Elétrico (2025), o ano de 2025 fechou com 223.912 emplacamentos de

veículos eletrificados leves no Brasil, com taxa de crescimento aproximadamente dez vezes superior à do mercado total de veículos automotores. Cada veículo eletrificado adicionado à frota brasileira é, em última análise, alimentado por uma matriz energeticamente limpa, o que confere ao Brasil uma vantagem ambiental rara no contexto internacional para a eletrificação dos transportes.

Entretanto, a transição não é livre de obstáculos, pois, embora o crescimento da infraestrutura pública de recarga brasileira, o Brasil ainda possui uma razão de 18 veículos elétricos por carregador disponível, o que demonstra um déficit da infraestrutura que pode acarretar na desaceleração dessa transição (Nexdigm, 2025). Esse cenário cria, de forma contraditória, uma janela de oportunidade, na qual a oferta de serviços bidirecionais V2G, com retorno econômico ao proprietário do veículo pode tornar-se um fator relevante para estimular a expansão da rede de recarga e justificar economicamente a aquisição do veículo.

No contexto brasileiro, a operação V2G ainda não está regulamentada como em outros países, mas este trabalho considera para arbitragem o Preço de Liquidação das Diferenças (PLD) horário – cuja volatilidade intradiária supera amplamente o *spread* tarifário convencional – uma fonte direta de retorno econômico. Em 2024, o PLD médio no Brasil situou-se em R\$ 127,95/MWh, com forte variação intradiária e inter-regional, refletindo as condições hidrológicas, a sazonalidade e os gargalos de intercâmbio entre submercados (Futura Energy, 2025). É essa volatilidade que viabiliza, em princípio, a estratégia de arbitragem por meio de armazenamento – carregar a bateria quando o preço está baixo e descarregar quando está alto.

No entanto, a formulação ótima da operação V2G, é computacionalmente desafiadora, porque o problema envolve múltiplas incertezas acopladas, como o preço futuro da energia, o padrão estocástico de uso do veículo pelo motorista, as restrições dinâmicas de Estado de Carga (*State of Charge* – SoC) impostas pela mobilidade, e os mecanismos de degradação eletroquímica da bateria, que se manifestam em escalas de tempo de milissegundos a anos. Métodos clássicos de otimização – como programação inteira mista ou Controle Preditivo Baseado em Modelo (MPC) – enfrentam dificuldades de escalabilidade frente a essa estocasticidade composta (ORFANOUDAKIS *et al.*, 2024). É nesse contexto que técnicas de Aprendizado por Reforço (*Reinforcement Learning* – RL), em particular algoritmos modernos como o *Proximal Policy Optimization* (PPO), apresentam-se como alternativa robusta. Tais algoritmos permitem que um agente aprenda, por interação com um ambiente simulado, políticas operacionais que maximizam uma recompensa de longo prazo – no presente caso, lucro líquido descontado pela degradação da bateria – sem necessitar de um modelo explícito da função de transição estocástica do ambiente (SUTTON; BARTO, 2018; SCHULMAN *et al.*, 2017). Niazi *et al.* (2026), em trabalho recente que serve de principal referência para esta monografia, demonstraram que tal abordagem supera estratégias heurísticas tanto em rentabilidade quanto em preservação do ativo. Dessa forma, este trabalho se propõe a implantar um agente PPO para realizar a otimização da carga e descarga de um veículo elétrico pela modalidade V2G, considerando o PLD do mercado de energia brasileiro como margem para

arbitragem, e considerando a degradação do veículo nesse processo de ciclagem como um custo adicional da tecnologia.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Avaliar o desempenho econômico da operação V2G no contexto do mercado de energia brasileiro, por meio do treinamento de um agente de Aprendizado por Reforço pelo algoritmo PPO acoplado a um modelo eletroquímico de bateria de íons de lítio, com foco no lucro de arbitragem do Preço de Liquidação das Diferenças (PLD) horário do submercado Sudeste, considerando a degradação da bateria como custo interno da operação.

1.1.2 Objetivos Específicos

- a) Aplicar um modelo eletroquímico e os mecanismos de degradação de baterias de íons de lítio;
- b) Construir um simulador estocástico de uso veicular residencial;
- c) Implementar e treinar um agente de Aprendizado por Reforço com o algoritmo PPO capaz de operar a bateria do veículo em modo bidirecional;
- d) Comparar o desempenho econômico do agente treinado contra uma linha de base de carregamento passivo (sem participação V2G);
- e) Avaliar o impacto da química da bateria sobre o lucro líquido acumulado e sobre a vida útil estimada do equipamento;
- f) Demonstrar a relevância e a aplicabilidade do Aprendizado por Reforço para problemas reais de energia, contribuindo para a formação no mundo atual.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 VEÍCULOS ELÉTRICOS E O V2G

2.1.1 Tipos de Veículos Elétricos

Sob a denominação genérica de Veículo Elétrico (VE) coexistem distintas arquiteturas de propulsão, classificadas conforme a relação entre motor elétrico, motor a combustão interna e bateria de tração (ACHARIGE *et al.*, 2023). Os veículos híbridos convencionais (*Hybrid Electric Vehicle*, HEV) combinam um motor a combustão com um motor elétrico e uma bateria de pequena capacidade, recarregada apenas pela frenagem regenerativa e pelo próprio motor a combustão, sem conexão com a rede, o que os mantém fora do escopo do V2G. Os veículos híbridos plug-in (*Plug-in Hybrid Electric Vehicle*, PHEV) incorporam uma bateria maior, recarregável diretamente pela rede, que garante autonomia puramente elétrica da ordem de 25 a 80 quilômetros antes do acionamento do motor a combustão, sendo candidatos válidos ao V2G ainda que a capacidade reduzida limite o serviço prestado (ACHARIGE *et al.*, 2023). Os veículos puramente elétricos a bateria (*Battery Electric Vehicle*, BEV) dispensam o motor a combustão e dependem exclusivamente da bateria de tração, cuja capacidade alcança dezenas de quilowatt-hora, a exemplo dos 62 kWh do Nissan Leaf e dos 68 kWh do Chevrolet Bolt (ACHARIGE *et al.*, 2023). Por reunirem maior capacidade, maior potência disponível e ausência de combustível fóssil no fluxo operacional, os BEV são o tipo mais adequado ao V2G e constituem o objeto deste trabalho.

2.1.2 A Bateria de Íons de Lítio

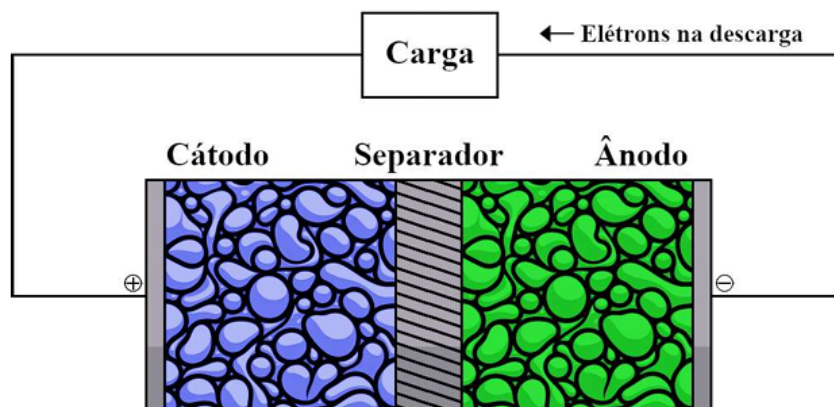
A bateria é o componente que define, simultaneamente, a autonomia do veículo, sua potência de aceleração, seu custo de capital e sua viabilidade para aplicações V2G. Nos veículos elétricos modernos, a tecnologia dominante é a bateria de íons de lítio (*Lithium-Ion Battery* – LIB), escolhida por sua combinação favorável de densidade energética, eficiência de ida e volta (*round-trip efficiency*) e ciclos de vida em condições nominais.

2.1.2.1 Princípio de funcionamento

Uma célula de íons de lítio é constituída por quatro elementos essenciais, sendo um eletrodo negativo (ânodo), tipicamente fabricado em grafite; um eletrodo positivo (cátodo), composto por um óxido metálico de lítio cuja química específica define grande parte das propriedades da célula; um eletrólito líquido, geralmente uma solução orgânica de sal de lítio; e um separador poroso, que impede o contato direto entre os eletrodos enquanto permite a passagem de íons (DOYLE *et al.*, 1993).

Durante a descarga, os íons de lítio migram do ânodo para o cátodo através do eletrólito, enquanto os elétrons percorrem o circuito externo gerando a corrente útil. Durante a carga, o processo se inverte, onde uma fonte externa força a migração dos íons de volta ao ânodo, armazenando energia química, conforme a Figura 1. Esse mecanismo de inserção e remoção reversível de íons em estruturas cristalinas hospedeiras é denominado *intercalação*, e é a base da reversibilidade que torna a bateria recarregável (REDDY, 2011).

Figura 1 – Diagrama da célula de íons de lítio.



Fonte: Oliveira (2025).

2.1.2.2 Estrutura hierárquica: célula, módulo e pack

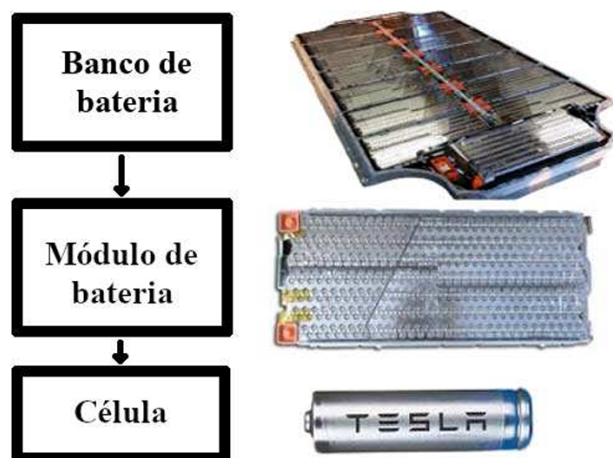
De acordo com Oliveira (2025), uma bateria de tração é composta por diversas estruturas hierárquicas. A unidade mínima é a célula, cujas tensões nominais situam-se na faixa de 2,5 a 4,2 V conforme a química. Múltiplas células são combinadas em módulos, agregando associações em série (para elevar a tensão) e em paralelo (para elevar a capacidade). Os módulos, por sua vez, são organizados em um *pack*, conforme Figura 2, que inclui também o sistema de gerenciamento (*Battery Management System* – BMS), os circuitos de proteção, o sistema de gerenciamento térmico e a estrutura mecânica. Em um BEV típico de 60 kWh, o *pack* pode conter de algumas centenas (em arquiteturas baseadas em células prismáticas ou pouch de grande capacidade) a alguns milhares (em arquiteturas baseadas em células cilíndricas de menor capacidade individual), a depender do formato adotado e da química.

2.1.2.3 Parâmetros operacionais essenciais

Para a operação V2G e para a modelagem desenvolvida nesta monografia, alguns parâmetros são especialmente relevantes (OLIVEIRA, 2025):

- **Capacidade nominal** (C_n , em Ah): quantidade de carga elétrica disponível na célula em condições padronizadas.

Figura 2 – Representação de célula, módulo e pack.



Fonte: Oliveira (2025).

- **Estado de Carga** (*State of Charge* – SoC): fração da capacidade atual em relação à capacidade nominal, expressa em percentual. O SoC governa diretamente a operação V2G – define quando a bateria pode descarregar e quando precisa carregar.
- **Estado de Saúde** (*State of Health* – SoH): fração da capacidade atual em relação à capacidade nominal de fábrica. Decresce ao longo da vida útil em razão da degradação irreversível. Em geral, considera-se que a bateria atingiu seu fim de vida automotiva quando o SoH cai abaixo de 80%.
- **Taxa C** (C_{rate}): razão entre a corrente de operação e a capacidade nominal. Uma operação a 1 C indica que a bateria descarrega completamente em uma hora; 0,5 C, em duas horas; 2 C, em meia hora. Taxas elevadas aceleram a degradação e geram aquecimento.
- **Eficiência de ida e volta** (*round-trip efficiency*): razão entre a energia recuperada na descarga e a energia injetada na carga, considerando perdas elétricas, térmicas e químicas. Como a operação V2G envolve uma conversão na carga e outra na descarga, cada uma com eficiência da ordem de 0,9 a 0,93 (KEMPTON; TOMIĆ, 2005), a eficiência de ida e volta entregue à rede fica tipicamente entre 80 e 90%.

2.1.3 Mecanismos de Degradação

Ao contrário de capacitores ou volantes de inércia, a bateria de íons de lítio sofre degradação irreversível com o uso e com o passar do tempo, e com isso, compreender os mecanismos físicos dessa degradação é essencial para o V2G, pois cada ciclo adicional de carga e descarga imposto pela operação bidirecional acelera, em alguma medida, o envelhecimento do equipamento.

A literatura distingue dois regimes principais de envelhecimento (DUBARRY *et al.*, 2018):

- Envelhecimento calendário (*calendar aging*): degradação que ocorre mesmo quando a bateria está ociosa, em função do tempo, da temperatura ambiente e do SoC de armazenamento. É particularmente relevante para o V2G porque os veículos elétricos permanecem estacionados aproximadamente 90% do tempo (DUBARRY *et al.*, 2018).
- Envelhecimento por ciclos (*cycle aging*): degradação acumulada a cada ciclo de carga e descarga, função da profundidade do ciclo (*Depth of Discharge*), das taxas de corrente e da janela de SoC utilizada.

Karger *et al.* (2024) demonstraram que esses dois regimes não podem ser simplesmente somados, pois existem acoplamentos não-lineares entre os mecanismos físicos subjacentes que produzem efeitos sinérgicos.

Em nível microscópico, três mecanismos físicos dominam a perda de capacidade e o aumento de resistência interna das LIBs (O'KANE *et al.*, 2022):

1. Crescimento da camada de interfase eletrólito-sólido (*Solid Electrolyte Interphase* – SEI): na primeira carga da célula, parte do eletrólito decompõe-se quimicamente na superfície do ânodo de grafite, formando uma fina camada passivante. Essa camada cresce lentamente ao longo da vida útil, consumindo íons de lítio que ficam permanentemente indisponíveis (*Loss of Lithium Inventory* – LLI) e aumentando a resistência interna da célula.
2. Deposição metálica de lítio (*lithium plating*): em condições de baixa temperatura ou alta taxa de carga, em vez de se intercalar no grafite, parte dos íons de lítio é reduzida a lítio metálico na superfície do ânodo. Essa deposição é parcialmente reversível, mas a fração irreversível compromete a capacidade e pode formar dendritos potencialmente perigosos.
3. Fadiga mecânica e perda de material ativo (*Loss of Active Material* – LAM): a inserção e remoção repetida de íons nas partículas de material ativo provoca dilatação e contração volumétrica, gerando estresses mecânicos que levam à formação de fissuras nas partículas e ao desprendimento parcial de material ativo do coletor de corrente. Esse efeito é particularmente severo em químicas NMC com alto teor de níquel (KARGER *et al.*, 2024).

A temperatura é um agravante transversal, pois tanto temperaturas elevadas (que aceleram reações químicas indesejadas, como o crescimento da SEI) quanto baixas (que favorecem o *plating*) impactam negativamente a vida útil (ALIPOUR *et al.*, 2020).

2.1.4 Químicas de Cátodo NMC e LFP

A química do cátodo é o fator que mais diferencia uma bateria automotiva das demais, definindo densidade energética, estabilidade térmica, vida útil, custo e impacto socioambiental dos materiais. No mercado atual, as duas famílias que dominam são as químicas baseadas em óxido de níquel, manganês e cobalto (NMC) e as químicas baseadas em fosfato de ferro-lítio (*Lithium Iron Phosphate* – LFP) (DUBARRY *et al.*, 2018).

2.1.4.1 A família NMC

A nomenclatura “NMC-xyz” indica a proporção molar relativa de níquel, manganês e cobalto no cátodo. Em uma NMC-811, por exemplo, há 8 partes de níquel para 1 de manganês e 1 de cobalto. Quanto maior a proporção de níquel, maior a densidade energética da célula – pois o níquel é o elemento que mais contribui para a capacidade específica do cátodo – porém menor a estabilidade térmica e mecânica.

A bateria NMC-811 representa a fronteira atual em densidade energética automotiva, com aproximadamente 250 Wh/kg. A célula de referência Chen2020 (LG M50, formato 21700 cilíndrico) possui capacidade nominal de aproximadamente 5,0 Ah e é amplamente adotada como referência de modelagem (CHEN *et al.*, 2020). Veículos elétricos com NMC-811 oferecem maior autonomia para uma mesma massa de bateria, ao custo de maior suscetibilidade à fadiga mecânica (LAM) e à instabilidade térmica – com temperatura de início de *thermal runaway* próxima de 200 °C (FENG *et al.*, 2018).

A bateria NMC-532, de composição aproximada 50% Ni, 30% Mn e 20% Co, representa um compromisso entre densidade energética e estabilidade, com densidade energética inferior à da NMC-811 e estabilidade térmica superior, por conta do menor teor de níquel (FENG *et al.*, 2018). Soma-se a isso o menor teor de cobalto, elemento crítico tanto pelo custo quanto pelas controvérsias socioambientais associadas à sua mineração. A célula de referência associada a essa química é a *pouch graphite/NMC-532* caracterizada experimentalmente por Mohtat *et al.* (2020).

2.1.4.2 A química LFP (olivina)

A LFP difere fundamentalmente da família NMC por sua estrutura cristalina em fosfato com geometria de olivina, que confere excepcional estabilidade estrutural e térmica (FENG *et al.*, 2018). Apresenta densidade energética substancialmente menor que a das químicas NMC em base gravimétrica, mas oferece em compensação:

- maior vida útil em número de ciclos;
- maior tolerância a abuso térmico e a sobrecargas;
- ausência de cobalto e níquel, reduzindo custo de matéria-prima e impacto socioambiental;

- maior estabilidade do potencial de circuito aberto (*Open Circuit Voltage* – OCV) ao longo do SoC, o que dificulta a estimação precisa do SoC mas oferece tensão de saída mais constante.

A LFP tem ganhado participação de mercado expressiva, especialmente em modelos de entrada e em veículos comerciais leves, e tornou-se também referência em sistemas BESS estacionários de grande porte. A célula de referência Prada2013 (A123 Systems 26650, cilíndrica) possui capacidade nominal de aproximadamente 2,3 Ah e é tomada como base nas simulações desta monografia (PRADA *et al.*, 2013).

Além de LFP e NMC, o mercado emprega outras químicas de cátodo, como o óxido de cobalto-lítio (LCO), o óxido de manganês-lítio (LMO) e o óxido de níquel-cobalto-alumínio (NCA), e, dentro da própria família NMC, composições intermediárias como a NMC-622. Em estabilidade térmica, a literatura ordena esses materiais como $LFP > LMO > NMC > NCA > LCO$, com a LFP como o cátodo mais estável durante o *thermal runaway*; em densidade energética a ordem é aproximadamente inversa, pois os óxidos ricos em níquel oferecem maior energia específica ao custo de menor estabilidade (FENG *et al.*, 2018).

2.1.5 Vehicle-to-Grid: Conceito e Justificativa

O conceito de *Vehicle-to-Grid* (V2G) foi formalizado por Kempton e Tomić (2005) e refere-se à capacidade de uma frota de veículos elétricos trocar potência ativa com a rede elétrica em ambas as direções, sob coordenação de um agente intermediário denominado agregador. Em essência, o V2G transforma a bateria de tração do veículo em um recurso energético distribuído, capaz de prestar serviços à rede durante os períodos em que o veículo está conectado e ocioso.

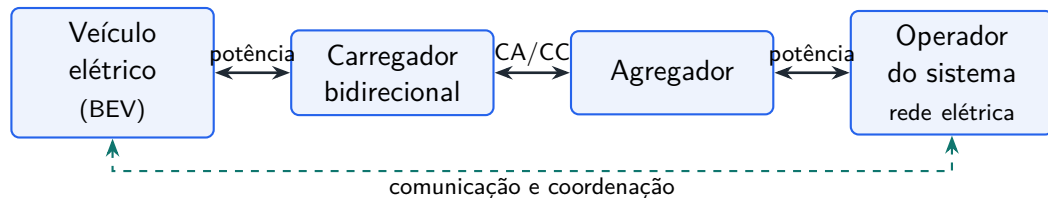
O argumento estrutural a favor do V2G repousa em três pilares (KEMPTON; TOMIĆ, 2005):

1. Ociosidade temporal: veículos privados permanecem estacionados em torno de 90 a 96% do tempo, o que significa que sua capacidade de armazenamento está, na prática, sub-aproveitada do ponto de vista da rede elétrica;
2. Despachabilidade rápida: a bateria de tração pode responder a comandos de potência em milissegundos, característica desejável para serviços ancilares de regulação de frequência e reserva girante, conforme já discutido na Introdução;
3. Baixo custo de capital marginal: como a bateria é amortizada primariamente pela função veicular, o custo de capital por unidade de potência para serviços de rede é significativamente inferior ao de um BESS estacionário dedicado.

É importante distinguir entre dois modos de operação. O V1G (também chamado *smart charging* ou recarga inteligente) refere-se à modulação unidirecional do horário e da

intensidade do carregamento – o veículo apenas consome energia, mas o agregador escolhe quando e quanto. Já o V2G propriamente dito refere-se à operação bidirecional plena, na qual o veículo pode também injetar potência na rede. O V2G é mais complexo do ponto de vista regulatório, técnico e econômico, mas oferece valor agregado proporcionalmente maior.

Figura 3 – Arquitetura conceitual do V2G.



Fonte: Adaptado de Kempton e Tomic (2005).

2.1.6 Arquitetura, Infraestrutura e Comunicação

A implementação prática do V2G requer três blocos físico-funcionais, o carregador bidirecional, o sistema de comunicação e o agregador.

2.1.6.1 Carregador bidirecional e fluxo de potência

A rede opera em corrente alternada e a bateria, em corrente contínua, de modo que toda transferência de potência entre as duas exige conversão eletrônica por um conversor bi-direcional, que opera no sentido CA-CC na carga e CC-CA na descarga (ACHARIGE *et al.*, 2023). A norma ABNT NBR IEC 61851 disciplina a recarga condutiva e padroniza quatro modos de carregamento, diferenciados pela potência, pela corrente e pela presença de comunicação e proteção dedicadas entre o veículo e a rede (ACHARIGE *et al.*, 2023). Esse conversor pode estar embarcado no veículo (*on-board*) ou alocado na estação (*off-board*), solução que transfere a complexidade da conversão para um equipamento estacionário, com restrições de massa e volume mais relaxadas, e é a topologia típica da recarga rápida em corrente contínua, e eficiência de conversão em equipamentos comerciais situa-se acima de 92% e chegando a 98% em conversores embarcados (ACHARIGE *et al.*, 2023); como um ciclo completo de V2G envolve a carga e a posterior descarga, somadas às perdas eletroquímicas e térmicas na célula, a eficiência de ida e volta entregue à rede fica tipicamente entre 80 e 90%, ou seja, cerca de 10 a 20% da energia não retorna como potência despachável.

2.1.6.2 Comunicação com a rede e protocolos

A coordenação entre veículo, ponto de recarga, agregador e operador de rede depende de protocolos padronizados de comunicação. Os principais padrões em uso ou em maturação, de acordo com Acharige *et al.* (2023) são:

- ISO 15118: padrão internacional para a comunicação entre veículo elétrico e infraestrutura de recarga, contemplando autenticação automática (*Plug & Charge*), suporte explícito a V2G e troca segura de informações de bilhetagem e potência.
- OCPP (*Open Charge Point Protocol*): padrão aberto para a comunicação entre ponto de recarga e plataforma de operação (agregador ou rede). Tornou-se padrão de fato no mercado de eletropostos públicos.
- CHAdeMO: padrão japonês que historicamente foi o primeiro a especificar a operação bidirecional comercialmente e que ainda equipa parte da frota.

No Brasil, a infraestrutura de comunicação para V2G ainda está em estágio inicial, e há regulamentação em desenvolvimento por meio da Consulta Pública nº 42/2025 da ANEEL (Agência Nacional de Energia Elétrica, 2025). A próxima seção aprofunda o quadro regulatório e de mercado brasileiro.

2.1.6.3 O papel do agregador

Segundo Kempton e Tomić (2005), o agregador é o agente que reúne dezenas, centenas ou milhares de veículos individuais sob uma única interface comercial e técnica com o operador da rede. Sua função é dupla. Em primeiro lugar, atinge a escala mínima de potência despachável exigida pelos mercados de serviços ancilares, que tipicamente operam em faixas de megawatts e não comportariam transações com um único veículo. Em segundo lugar, internaliza a estocasticidade da disponibilidade individual: enquanto um veículo isolado pode estar indisponível em uma janela arbitrária, uma carteira agregada de centenas de veículos apresenta disponibilidade aproximadamente determinística por compensação estatística.

2.2 SISTEMA ELÉTRICO BRASILEIRO E MERCADO

2.2.1 Estrutura Institucional do Setor Elétrico Brasileiro

O setor elétrico brasileiro opera sob um modelo institucional no qual as funções de planejamento, regulação, operação, comercialização e fiscalização são separadas entre entidades distintas, visando garantir transparência, neutralidade técnica e estabilidade regulatória. Para o desenvolvimento do V2G no país, é essencial compreender o papel dos principais agentes, pois a operação bidirecional de veículos transita por competências de mais de um órgão.

No nível federal, o Ministério de Minas e Energia (MME) formula as políticas energéticas e publica portarias e consultas públicas que orientam a expansão do setor, apoiado pela Empresa de Pesquisa Energética (EPE), autora do Balanço Energético Nacional e do Plano Decenal de Expansão de Energia (Empresa de Pesquisa Energética, 2025; Ministério de Minas e Energia; Empresa de Pesquisa Energética, 2025). A regulação técnica e econômica do setor cabe à Agência Nacional de Energia Elétrica (ANEEL), agência reguladora autárquica especial

responsável por disciplinar a conexão de agentes, a tarifação e os serviços ancilares – destacam-se para o V2G as Resoluções Normativas nº 1.000/2021 e nº 1.030/2022, a Lei nº 14.300/2022 (marco da Geração Distribuída) e a Consulta Pública nº 42/2025 (eletromobilidade).

A operação física do Sistema Interligado Nacional (SIN) é exercida pelo Operador Nacional do Sistema Elétrico (ONS), organização privada sem fins lucrativos que coordena em tempo real o despacho das usinas, o intercâmbio entre regiões e a manutenção dos parâmetros de qualidade da energia. Suas regras técnicas estão consolidadas nos Procedimentos de Rede, entre os quais o Submódulo 10.6 disciplina o controle da geração, o Controle Automático de Geração e a reserva de potência operativa (Operador Nacional do Sistema Elétrico, 2018). A liquidação financeira do mercado de curto prazo, por sua vez, é administrada pela Câmara de Comercialização de Energia Elétrica (CCEE), que calcula diariamente o Preço de Liquidação das Diferenças (PLD) – variável central do problema de arbitragem que motiva este trabalho (Câmara de Comercialização de Energia Elétrica, 2024a).

A esses agentes somam-se ainda o Conselho Nacional de Política Energética (CNPE), órgão consultivo da Presidência da República, e os agentes setoriais propriamente ditos, geradoras, transmissoras, distribuidoras e comercializadoras de energia, além dos consumidores cativos e livres.

2.2.2 O Sistema Interligado Nacional e seus Submercados

A operação do Sistema Interligado Nacional (SIN) otimiza o uso da matriz elétrica brasileira permitindo a transferência de energia entre regiões com diferentes regimes hídricos e de geração. Para a precificação e gestão física, o SIN divide-se em quatro submercados: Sudeste/Centro-Oeste, Sul, Nordeste e Norte. Essa divisão regional permite o descolamento de preços de curto prazo de acordo com as restrições físicas de intercâmbio e a oferta e demanda locais, resultando em valores que podem ser muito distintos no mesmo instante.

A definição desses preços decorre do despacho centralizado pelo ONS, que utiliza uma cadeia de modelos computacionais encadeados. Enquanto os modelos de longo e médio prazo (NEWAVE e DECOMP) planejam o uso dos reservatórios, o modelo de curtíssimo prazo (DESSEM) atua na discretização horária ou de meia-hora para calcular o Custo Marginal de Operação (CMO) – que é basicamente o custo adicional necessário para produzir uma unidade extra de energia elétrica no sistema – de cada submercado. É a partir do CMO calculado pelo DESSEM que a CCEE deriva o Preço de Liquidação das Diferenças (PLD) horário, conforme detalhado adiante.

2.2.3 Ambientes de Comercialização: ACR e ACL

A comercialização de energia elétrica no Brasil ocorre em dois ambientes distintos. O Ambiente de Contratação Regulada (ACR) congrega os consumidores cativos – tipicamente residências e pequenos comércios – que adquirem energia exclusivamente da distribuidora local, a um custo formado pela soma da Tarifa de Energia (TE) e da Tarifa de Uso do Sistema de

Distribuição (TUSD), submetidas à revisão periódica da ANEEL. Os contratos de fornecimento das distribuidoras junto a geradoras, por sua vez, são firmados por meio de leilões públicos coordenados pela CCEE e regulamentados pela ANEEL.

O Ambiente de Contratação Livre (ACL), popularmente conhecido como Mercado Livre de Energia, congrega os agentes que podem negociar livremente seus contratos bilaterais de compra e venda de energia – geradoras, comercializadoras e consumidores livres ou especiais. Historicamente, o acesso ao ACL era restrito a consumidores de alta tensão e elevada demanda, mas a abertura gradual desse ambiente tem reduzido continuamente os critérios mínimos de elegibilidade. Discussões em curso apontam para a possibilidade de extensão futura do ACL aos consumidores residenciais, o que abriria horizontes adicionais para modelos de negócio baseados em arbitragem e em recursos energéticos distribuídos.

2.2.4 O Preço de Liquidação das Diferenças (PLD)

O Preço de Liquidação das Diferenças (PLD) é a variável de preço do mercado de curto prazo brasileiro e, no escopo desta monografia, é a referência primária para a função de recompensa do agente de Aprendizado por Reforço. Em termos formais, o PLD é o preço utilizado pela CCEE para a valoração da energia comprada e vendida no Mercado de Curto Prazo, ou seja, o preço aplicado a toda quantidade de energia que não esteja coberta por contrato bilateral.

Em sua formação, o PLD parte do CMO, calculado pelo DESSEM para cada hora e cada submercado, sendo posteriormente truncado pelos limites máximo e mínimo regulatórios estabelecidos anualmente pela ANEEL. Assim, mesmo que o CMO calculado pelo modelo de otimização do ONS resulte em valores extremos – por exemplo, próximos a zero em momentos de sobreoferta hídrica, ou muito elevados em momentos de risco de déficit – o PLD efetivamente liquidado permanece na faixa regulamentada.

Uma transição metodológica relevante para o V2G ocorreu em 2021, quando a CCEE migrou da apuração do PLD em patamares semanais (com três faixas: leve, médio e pesado) para a apuração horária, alinhando o preço brasileiro às práticas dos mercados nodais maduros. Essa granularidade temporal é o que torna a arbitragem por armazenamento potencialmente viável, pois sem variação intradiária do preço, não haveria *spread* sobre o qual o V2G pudesse operar. A partir de 2021, o PLD passou a ser calculado diariamente para cada hora do dia seguinte e para cada um dos quatro submercados, totalizando 96 valores publicados por dia (Câmara de Comercialização de Energia Elétrica, 2024b). No ano de 2024, o PLD médio nacional ficou em torno de R\$ 127,95/MWh, com forte variação ao longo do dia e entre submercados (Futura Energy, 2025).

2.2.5 Regulamentação Aplicável ao V2G: Estado Atual e Perspectivas

Os serviços ancilares são as funções operacionais que mantêm a qualidade e a confiabilidade do fornecimento, como o controle de frequência e a reserva de potência, e seguem no

Brasil os Procedimentos de Rede do ONS sob regulação da ANEEL (Operador Nacional do Sistema Elétrico, 2018; Agência Nacional de Energia Elétrica, 2022). Historicamente boa parte deles foi tratada como obrigação não remunerada dos grandes geradores, premissa que deixou de se sustentar com a entrada de solar e eólica, levando a Resolução Normativa nº 1.030/2022 a abrir espaço para remunerar produtos de flexibilidade, inclusive em ambientes experimentais, o que interessa diretamente ao armazenamento em baterias e ao próprio V2G (Agência Nacional de Energia Elétrica, 2022). O marco recente, porém, mira o armazenamento estacionário de grande porte, como a Lei nº 15.269/2025 e o primeiro leilão de reserva de capacidade para baterias (Brasil, 2025), de modo que a remuneração de serviços ancilares por frotas de veículos ainda não tem regra definida.

O quadro regulatório aplicável ao V2G no Brasil resulta da combinação de duas frentes ainda não plenamente articuladas. Por um lado, as normas gerais de Geração Distribuída – atualmente moldadas pela Lei nº 14.300/2022 e pela REN nº 1.000/2021 da ANEEL (Brasil, 2022; Agência Nacional de Energia Elétrica, 2021) – que estabelecem o regime de compensação de energia para microgeradores e minigeradores, mas que não foram desenhadas considerando o caráter bidirecional, intermitente e estocástico de uma frota de veículos elétricos. Por outro lado, há a regulação específica de infraestrutura de recarga, em revisão pela ANEEL por meio da Consulta Pública nº 42/2025, aberta em dezembro de 2025 e com contribuições recebidas até março de 2026, que lista explicitamente o armazenamento, a recarga inteligente e o V2G entre seus temas e cujo desfecho deve introduzir disposições específicas para o acesso e a conexão de equipamentos de eletromobidade à rede de distribuição (Agência Nacional de Energia Elétrica, 2025; International Trade Administration, 2026).

Entre as medidas em discussão pela CP 42/2025 estão a previsão de contratos flexíveis de conexão para eletropostos, modelos mais claros de alocação de custos entre o agente e a distribuidora, e a obrigatoriedade de publicação, pelas distribuidoras, de Mapas de Disponibilidade de Conexão que indiquem, em nível de alimentador, em quais pontos da rede a instalação de carregadores de alta potência é viável sem a necessidade de reforço estrutural prévio. Essa medida enfrenta um gargalo antigo do setor, a demora nos pedidos de conexão e a incerteza sobre o custo de reforçar a rede.

Especificamente para o V2G, contudo, as questões pendentes são mais amplas. Ao contrário do armazenamento estacionário, que ganhou enquadramento próprio com a Lei nº 15.269/2025, o V2G permanece sem um regime que reconheça o veículo como ativo capaz de injetar energia, situado entre a compensação de GD, desenhada para geração e não para um fluxo bidirecional e intermitente, e o marco do armazenamento, voltado a agentes de grande porte. O caminho mais provável no curto prazo passa pela figura do agregador, que reuniria muitos veículos em um bloco de capacidade grande o suficiente para participar do mercado e ser visível ao ONS, modelo ainda não disciplinado no país. Não há, até a presente data, regulamentação específica que discipline a injeção bidirecional individual de pequena escala, nem mecanismos tarifários que distingam claramente a operação V2G da operação convencional de

GD. Há também questões em aberto sobre a remuneração de serviços ancilares prestados por frotas agregadas, sobre o tratamento fiscal da energia comercializada por motoristas individuais, e sobre a homologação técnica de equipamentos bidirecionais para o mercado nacional. As discussões em curso na ANEEL e o avanço da CP 42/2025 sinalizam, no entanto, um processo de amadurecimento institucional consistente com a maturação observada em mercados internacionais nos cinco a dez anos anteriores.

No horizonte de médio prazo, a combinação esperada de (i) expansão sustentada da frota de veículos elétricos, conforme já discutido na Introdução; (ii) maior pressão por flexibilidade operacional decorrente da hiperpenetração de fontes não-despacháveis; e (iii) abertura gradual do mercado livre de energia ao consumidor residencial cria condições estruturais favoráveis à consolidação do V2G como modalidade reconhecida no Brasil.

2.3 APRENDIZADO POR REFORÇO

2.3.1 Inteligência Artificial e Aprendizado de Máquina: Panorama Conceitual

O termo *Inteligência Artificial* (IA) designa o campo de pesquisa que se ocupa do desenvolvimento de sistemas computacionais capazes de executar tarefas que tradicionalmente exigiriam cognição humana – reconhecer padrões em imagens, raciocinar logicamente sobre relações, tomar decisões em ambientes incertos, processar linguagem natural. No interior desse campo amplo, o *Aprendizado de Máquina* (*Machine Learning* – ML) emergiu como abordagem dominante na prática contemporânea, onde em vez de programar explicitamente as regras de comportamento do sistema, instrumenta-se um modelo computacional para inferir regularidades a partir de dados ou de experiência, ajustando seus parâmetros internos por meio de algoritmos de otimização (SUTTON; BARTO, 2018).

A literatura de aprendizado de máquina organiza-se em três grandes paradigmas, distinguidos pela natureza do sinal de treinamento disponível ao algoritmo. No aprendizado supervisionado, o modelo é treinado a partir de pares (entrada, rótulo) fornecidos por um anotador humano ou por um processo externo confiável; o objetivo é aprender uma função que mapeie corretamente entradas a saídas em casos previamente não vistos. Aplicações típicas voltadas para energia incluem previsão de carga, classificação de defeitos em equipamentos elétricos e estimação do Estado de Saúde de baterias a partir de medições de corrente e tensão.

No aprendizado não-supervisionado, o algoritmo recebe apenas as entradas, sem rótulos, e busca descobrir estrutura latente nos dados – agrupamentos (*clustering*), redução de dimensionalidade ou anomalias. É utilizado em análise exploratória de dados, segmentação de consumidores e detecção de eventos anômalos em medições de redes elétricas.

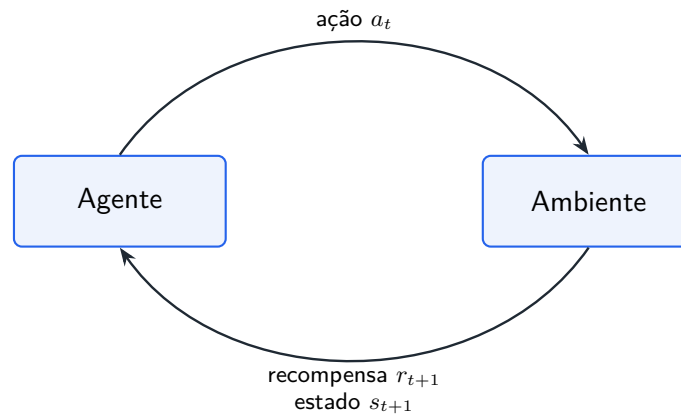
No aprendizado por reforço (*Reinforcement Learning* – RL), por fim, o algoritmo aprende por interação direta com um ambiente, recebendo apenas um sinal escalar de recompensa que avalia a qualidade de suas decisões ao longo do tempo. Diferentemente do paradigma supervisionado, não há uma resposta correta apontada a cada passo, e o agente

precisa descobrir, por experimentação repetida, quais sequências de ações conduzem a recompensas elevadas – frequentemente distantes no tempo da ação que as originou.

2.3.2 O Paradigma da Aprendizado por Reforço

A formulação da Aprendizado por Reforço descreve uma interação cíclica entre dois elementos, o agente, que toma decisões, e o ambiente, que responde a essas decisões produzindo consequências (SUTTON; BARTO, 2018). Em cada instante de tempo discreto t , o agente observa o estado s_t do ambiente, escolhe uma ação a_t com base em sua política, recebe uma recompensa imediata r_{t+1} e observa o próximo estado s_{t+1} conforme Figura 4. O processo se repete e o agente busca aprender uma política π – regra de mapeamento de estados a ações – que maximize a soma acumulada de recompensas ao longo do tempo.

Figura 4 – Laço de interação agente-ambiente na Aprendizado por Reforço.



Fonte: Adaptado de Sutton e Barto (2018).

Para tornar essa formulação mais palpável, recorre-se frequentemente a uma analogia simples. Imagine alguém ensinando um cachorro a sentar-se sob comando. O agente é o cachorro – quem decide o que fazer. O ambiente é tudo o que está ao redor: o dono, o cômodo, a hora do dia, eventuais distrações. O estado é a situação observada no momento (o dono falou “senta”, o cachorro está em pé, com fome). A ação é o comportamento que o cachorro escolhe – sentar-se, latir ou ignorar o comando. A recompensa positiva é o biscoito recebido quando obedece. A punição (ou recompensa negativa) é a repreensão verbal recebida quando ignora o comando ou faz algo indesejado. Ninguém diz ao cachorro explicitamente *qual* é a regra a ser aprendida e mesmo assim, ele descobre por tentativa e erro repetidos ao longo de muitas sessões, qual sequência de comportamentos maximiza biscoitos e minimiza repreensões.

Trazendo a analogia ao contexto V2G, o agente é um algoritmo de controle que decide, em cada intervalo de tempo, como operar a bateria de um veículo elétrico. O ambiente abrange a rede elétrica, o mercado de energia, o padrão estocástico de uso veicular do motorista e a física eletroquímica da própria bateria. O estado é o vetor de variáveis que o agente

observa em cada instante – tipicamente incluindo o Estado de Carga atual, sinais de preço da energia (corrente e previsto para o curto horizonte), indicadores temporais e informações sobre disponibilidade do veículo. A ação é a potência de carga ou descarga a ser aplicada à bateria, limitada pelo envelope físico do carregador. A recompensa positiva está associada a resultados desejáveis – receita pela venda de energia em momentos de preço elevado, economia ao carregar em momentos de preço baixo, ou eventual prestação de serviços ancilares à rede. A punição – recompensa negativa – é aplicada quando a operação acelera a degradação da bateria, compromete a disponibilidade do veículo para a função de mobilidade, ou viola algum limite operacional ou contratual. Ao longo de muitas iterações simuladas, o agente descobre – sem que ninguém lhe diga *qual* é a estratégia ótima – quais sequências de decisões maximizam as recompensas acumuladas e minimizam as penalidades.

Três características distinguem essa formulação dos paradigmas anteriores. Primeiro, o sinal de aprendizado é avaliativo, não instrutivo, pois a recompensa informa ao agente *quão boa* foi sua ação, mas não diz qual seria a melhor escolha. Segundo, o sinal de aprendizado é frequentemente atrasado, onde uma recompensa observada no instante $t + 10$ pode ser consequência de uma decisão tomada em t , e o agente precisa atribuir crédito retrospectivamente – um problema conhecido como atribuição de crédito temporal. Terceiro, as ações do agente influenciam os dados observados, no qual, ao alterar sua política, o agente também altera a distribuição de estados que visita, criando um problema acoplado de coleta de dados e otimização. Esse acoplamento contrasta diretamente com o aprendizado supervisionado, no qual o conjunto de dados é fixo e independente do modelo que se treina.

2.3.3 Processo de Decisão de Markov

A formalização matemática da interação agente-ambiente é o Processo de Decisão de Markov (*Markov Decision Process* – MDP). Um MDP é definido formalmente pela tupla:

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle \quad (2.1)$$

onde $\mathcal{S}$ é o conjunto de estados possíveis, $\mathcal{A}$ é o conjunto de ações disponíveis, $\mathcal{P}(s' | s, a)$ é a função de transição que descreve a probabilidade de o ambiente passar do estado s ao estado s' quando a ação a é executada, $\mathcal{R}(s, a, s')$ é a função de recompensa que retorna o valor escalar associado a essa transição, e $\gamma \in [0, 1]$ é o fator de desconto que pondera recompensas futuras em relação a recompensas imediatas.

A hipótese central da formulação é a propriedade de Markov, que postula que o futuro do sistema depende apenas do estado e da ação no instante presente, e não do histórico completo de estados e ações anteriores:

$$\Pr(s_{t+1} | s_t, a_t, s_{t-1}, a_{t-1}, \dots) = \Pr(s_{t+1} | s_t, a_t) \quad (2.2)$$

Embora simplificadora, essa hipótese pode ser respeitada na prática por meio do dese-

nho cuidadoso do espaço de estados, incluindo no vetor de estado todas as variáveis necessárias para que a dinâmica futura fique condicionada apenas a ele. Em problemas de controle V2G, por exemplo, é comum que o estado contenha não apenas variáveis instantâneas como Estado de Carga e hora atual, mas também previsões de curto horizonte do sinal de preço da energia e indicadores de uso recente do veículo, justamente para preservar a propriedade de Markov sob a presença de sinais exógenos correlacionados no tempo.

A grandeza que o agente busca maximizar é o retorno G_t , definido como a soma descontada das recompensas a partir do instante t :

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \quad (2.3)$$

O fator γ tem dupla função: matematicamente, garante a convergência da série em problemas com horizonte infinito; conceitualmente, codifica a preferência temporal do agente – valores próximos a 1 conferem peso quase igual a recompensas distantes; valores baixos privilegiam ganhos imediatos.

2.3.4 Políticas, Funções de Valor e a Equação de Bellman

A política π do agente é a regra pela qual ele escolhe ações a partir de estados. Em sua forma mais geral, é uma distribuição condicional $\pi(a | s)$ que define a probabilidade de executar a ação a quando o estado é s . Políticas determinísticas, em que $\pi(s) = a$ univocamente, são casos particulares dessa formulação.

Dada uma política, define-se a função de valor de estado $V^\pi(s)$ como o retorno esperado a partir do estado s seguindo π :

$$V^\pi(s) = \mathbb{E}_\pi[G_t | s_t = s] \quad (2.4)$$

e a função de valor de ação $Q^\pi(s, a)$ como o retorno esperado quando, no estado s , executa-se primeiro a ação a e depois segue-se π :

$$Q^\pi(s, a) = \mathbb{E}_\pi[G_t | s_t = s, a_t = a] \quad (2.5)$$

A função Q é particularmente útil porque encapsula em si toda a informação necessária para a tomada de decisão, onde, dado um estado, a escolha racional é a ação que maximiza $Q^\pi(s, a)$.

A relação recursiva entre o valor de um estado e os valores dos estados subsequentes é dada pela equação de Bellman, que ocupa posição central em toda a teoria da Aprendizagem por Reforço. Para a função de valor de estado, escreve-se:

$$V^\pi(s) = \sum_a \pi(a | s) \sum_{s'} \mathcal{P}(s' | s, a) [\mathcal{R}(s, a, s') + \gamma V^\pi(s')] \quad (2.6)$$

A interpretação é direta, o valor de um estado é a soma das recompensas imediatas obtidas a partir dele – ponderadas pelas probabilidades das ações da política e pelas probabilidades das transições do ambiente – adicionada ao valor descontado dos estados subsequentes. Uma equação análoga existe para a função Q^π , na qual a esperança sobre ações é substituída por seleção determinística da ação do par estado-ação considerado.

2.3.5 Métodos de Solução - Modelo Conhecido ao Modelo Livre

Três grandes famílias de métodos resolvem MDPs, distinguindo-se pela disponibilidade do modelo do ambiente e pela natureza da estimativa de valor (SUTTON; BARTO, 2018).

A Programação Dinâmica resolve a equação de Bellman por iteração, mas exige conhecer o modelo do ambiente (as funções $\mathcal{P}$ e $\mathcal{R}$), condição rara em problemas reais como o V2G. Os métodos de Monte Carlo dispensam esse modelo, porém só atualizam suas estimativas ao fim de cada episódio, o que os torna pouco eficientes em horizontes longos.

Os métodos de Diferença Temporal (*Temporal Difference* – TD) combinam as virtudes das duas abordagens anteriores: aprendem por interação direta, como Monte Carlo, mas atualizam suas estimativas a cada passo, como a Programação Dinâmica – utilizando a estimativa corrente do valor do próximo estado como sucedâneo para o retorno futuro real. A atualização básica do método TD(0) para a função V é:

$$V(s_t) \leftarrow V(s_t) + \alpha \underbrace{\left[r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \right]}_{\delta_t \text{ (erro TD)}} \quad (2.7)$$

O termo entre colchetes é o erro de Diferença Temporal (frequentemente denotado δ_t): a discrepância entre o valor estimado pela informação mais recente – recompensa observada acrescida do valor descontado do próximo estado – e o valor anteriormente estimado para o estado corrente. Esse sinal é a unidade fundamental sobre a qual praticamente todos os algoritmos modernos de Aprendizado por Reforço operam, incluindo o PPO utilizado neste trabalho. O fator $\alpha \in (0, 1]$ controla o tamanho do passo de atualização.

2.3.6 Aprendizado por Reforço Profundo e Métodos *Policy-Gradient*

A combinação da Aprendizado por Reforço com redes neurais profundas, o Aprendizado por Reforço Profundo (*Deep Reinforcement Learning* – DRL), tornou-se possível ao longo da década de 2010. Os primeiros marcos foram métodos baseados em valor, como o *Deep Q-Network* (DQN), que substitui a tabela de Q por uma rede neural. Esses métodos, porém, têm uma limitação em ações contínuas, porque o máximo sobre as ações não é diretamente computável quando $\mathcal{A}$ é um intervalo de números reais. Para problemas dessa natureza, como a potência de carga e descarga em V2G, recorre-se aos métodos *policy-gradient*.

Os métodos *policy-gradient* parametrizam diretamente a política $\pi_\theta(a | s)$ por meio de uma rede neural com parâmetros θ e otimizam esses parâmetros por subida do gradiente

sobre o retorno esperado. A formulação básica do gradiente da política é (SUTTON; BARTO, 2018):

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\pi_{\theta}} [\nabla_{\theta} \log \pi_{\theta}(a | s) \cdot A^{\pi_{\theta}}(s, a)] \quad (2.8)$$

onde $J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} [\sum_{t=0}^T \gamma^t r_t]$ é o objetivo a maximizar, ou seja, o retorno esperado descontado ao longo de uma trajetória τ gerada pela política π_{θ} , e

$$A^{\pi_{\theta}}(s, a) = Q^{\pi_{\theta}}(s, a) - V^{\pi_{\theta}}(s) \quad (2.9)$$

é a função de vantagem – a magnitude pela qual a ação a supera, no estado s , a média esperada da política. A vantagem é tipicamente estimada por uma segunda rede neural separada, configurando a arquitetura *Actor-Critic*, na qual o *ator* (*actor*) implementa a política e propõe as ações, enquanto o *crítico* (*critic*) implementa a função de valor que avalia a política e fornece o sinal de aprendizado.

Métodos *policy-gradient* são versáteis e adaptam-se naturalmente a ações contínuas, mas sofrem historicamente de instabilidade, uma vez que passos grandes na atualização da política podem destruir o progresso obtido em iterações anteriores, e o ajuste fino da taxa de aprendizado é delicado.

2.3.7 Proximal Policy Optimization (PPO)

O algoritmo *Proximal Policy Optimization* (PPO), proposto por Schulman *et al.* (2017) na OpenAI, endereça especificamente o problema de estabilidade dos métodos *policy-gradient*. Sua ideia central é restringir, a cada atualização, o quanto a política nova pode divergir da política anterior, garantindo que o aprendizado proceda em passos “proximais” – suficientes para progredir, mas não tão amplos a ponto de comprometer o desempenho já atingido.

A função objetivo característica do PPO é o objetivo *clipped*:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right] \quad (2.10)$$

onde

$$r_t(\theta) = \frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)} \quad (2.11)$$

é a razão entre a probabilidade da ação a_t sob a política nova e sob a política anterior, $\hat{A}_t$ é a estimativa da vantagem no instante t , e ϵ é o parâmetro de *clipping* (tipicamente entre 0,1 e 0,3). O operador *clip* recorta a razão $r_t(\theta)$ ao intervalo $[1 - \epsilon, 1 + \epsilon]$, e o operador *min* assegura que o objetivo permaneça uma cota inferior (pessimista) sobre a melhoria de política – desestimulando atualizações em direções com razão extrema, mesmo quando aparentemente vantajosas.

A estimativa de vantagem $\hat{A}_t$ que aparece no objetivo da Eq. (2.10) não é o valor exato da Eq. (2.9), que dependeria de Q e V verdadeiros, e sim uma estimativa obtida a partir das

trajetórias coletadas pelo método *Generalized Advantage Estimation* (GAE) (SCHULMAN *et al.*, 2015), definido por:

$$\hat{A}_t = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}, \quad \text{com} \quad \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \quad (2.12)$$

onde δ_t é o resíduo de diferença temporal (TD), que mede o quanto a recompensa observada r_t somada ao valor estimado do próximo estado supera o valor previsto pelo crítico para o estado atual, e $\lambda \in [0, 1]$ é o parâmetro do GAE, que regula o compromisso entre viés e variância da estimativa, com valores próximos de 1 reduzindo o viés ao custo de maior variância. O objetivo L^{CLIP} não altera diretamente os pesos da rede, apenas define a quantidade a ser maximizada, e os parâmetros são efetivamente ajustados por subida de gradiente sobre esse objetivo, com taxa de aprendizado α e gradiente obtido por retropropagação (SUTTON; BARTO, 2018; SCHULMAN *et al.*, 2017). O treinamento procede em ciclos (SCHULMAN *et al.*, 2017), em que a política corrente $\pi_{\theta_{\text{old}}}$ interage com o ambiente por um lote de passos coletando trajetórias, calcula-se a vantagem $\hat{A}_t$ de cada passo pela Eq. (2.12), e os pesos são ajustados por várias épocas de subida de gradiente sobre L^{CLIP} em *minibatches*, atualizando tanto a rede de política (ator) quanto a rede de valor (crítico, ajustada por descida de gradiente sobre o erro de previsão do retorno). Concluído o ciclo, a política atualizada torna-se a nova $\pi_{\theta_{\text{old}}}$ e o processo se repete. É o recorte de $r_t(\theta)$ que mantém cada ciclo dentro da região “proximal”, impedindo que a reutilização das mesmas trajetórias por várias épocas desestabilize a política.

Em termos práticos, três propriedades tornam o PPO particularmente adequado ao problema V2G abordado nesta monografia. Primeiro, ele opera nativamente em espaços de ação contínuos, condição necessária para representar a potência de carga e descarga como variável real. Segundo, possui estabilidade de treinamento superior aos métodos *policy-gradient* sem *clipping*, evitando colapsos de política em treinamentos longos – característica desejável quando o ambiente de simulação envolve estocasticidade composta (PLD, perfil veicular) e horizonte estendido. Terceiro, oferece eficiência amostral razoável para a classe de problemas em que se enquadra, sendo um dos algoritmos mais utilizados em aplicações de V2G (XIE *et al.*, 2025).

A implementação adotada neste trabalho utiliza o PPO em arquitetura *Actor-Critic*, com redes neurais separadas para política e função de valor, e estimação da vantagem por meio do GAE. Os detalhes específicos da arquitetura, dos hiperparâmetros e do procedimento de treinamento são apresentados no Capítulo 3.

2.4 APRENDIZADO POR REFORÇO APLICADA AO V2G

A aplicação de Aprendizado por Reforço ao problema V2G consolidou-se como campo de pesquisa expressivo ao longo da última década, conforme já contextualizado nas seções anteriores. A revisão sistemática de Xie *et al.* (2025) organiza essa literatura em oito grandes

áreas de aplicação, observadas sob três perspectivas distintas dos agentes envolvidos, a do operador de rede, a do agregador de pontos de recarga e a do usuário do veículo elétrico:

1. Controle colaborativo de frequência da rede – uso da potência bidirecional dos veículos para auxiliar o controle primário e secundário de frequência do sistema, em coordenação com a inércia rotativa convencional.
2. Despacho ótimo – programação dinâmica das operações de carga e descarga de frotas, sob restrições de SoC dos veículos individuais e padrões estocásticos de mobilidade.
3. Estratégia de precificação – definição, por parte de agregadores ou operadores, de sinais de preço (tarifa horária dinâmica, sinais de incentivo) que induzem comportamento de carga/descarga adequado às necessidades do sistema.
4. Estratégia de lances (*bidding*) – participação automatizada de agregadores em leilões de mercado de curto prazo e de serviços ancilares, otimizando lances sob incerteza de preço e de disponibilidade da frota.
5. Microrredes multi-energéticas – coordenação de frotas V2G em ambientes de microrrede que integram geração distribuída fotovoltaica, eólica, armazenamento estacionário e cargas residenciais ou industriais.
6. Programação de armazenamento de energia – gerenciamento conjunto da bateria veicular e de baterias estacionárias auxiliares, otimizando ciclos e arbitragem entre os ativos.
7. Otimização da rede de distribuição – coordenação de pontos de recarga ao longo de alimentadores para minimizar perdas, mitigar sobretensões e adiar reforços de infraestrutura.
8. Defesa contra ameaças cibernéticas – uso de RL para detecção e resposta a ataques que explorem a comunicação bidirecional entre veículos, pontos de recarga e plataformas agregadoras.

Cada uma dessas áreas demanda formulações particulares do espaço de estado, da função de recompensa e dos algoritmos adotados. Trabalhos *grid-side* tendem a priorizar métricas operacionais como desvio de frequência, relação pico-média (PAR) e perdas técnicas; trabalhos sob a ótica do agregador priorizam receita de mercado e cumprimento de contratos; trabalhos centrados no usuário priorizam o resultado financeiro individual descontado das penalidades de degradação e de conforto operacional (XIE *et al.*, 2025).

Entre os trabalhos da literatura recente, o de Niazi *et al.* (2026), publicado na *IEEE Access* sob o título *A Physics-Informed Reinforcement Learning Approach to Vehicle-to-Grid Control With Real-Time Battery Degradation*, é tomado como referência metodológica primária desta monografia. Os autores propõem um *framework* denominado *Physics-Informed Deep*

Reinforcement Learning (PI-DRL), no qual um Gêmeo Digital eletroquímico (assim denominado pelos autores) – baseado no modelo Doyle-Fuller-Newman implementado em PyBaMM – é integrado ao laço de decisão de um agente de Aprendizado por Reforço Profundo, fornecendo um sinal de custo de degradação fisicamente fundamentado em vez de aproximações empíricas lineares (NIAZI *et al.*, 2026). Esse trabalho é detalhado e adaptado ao contexto brasileiro no Capítulo 3 desta monografia.

Frente ao espectro de aplicações apresentado, o presente trabalho posiciona-se na intersecção das categorias despacho ótimo e estratégia de precificação (na ótica do usuário individual), com formulação adaptada à arbitragem do PLD horário brasileiro. A perspectiva adotada é a do usuário do veículo elétrico, escolha justificada pela incipiência do mercado V2G nacional e pela ausência, no curto prazo, de modelos de negócio de agregação plenamente constituídos no Brasil – conforme detalhado na Seção 2.2

3 METODOLOGIA

3.1 ARQUITETURA GERAL DO SISTEMA DE SIMULAÇÃO

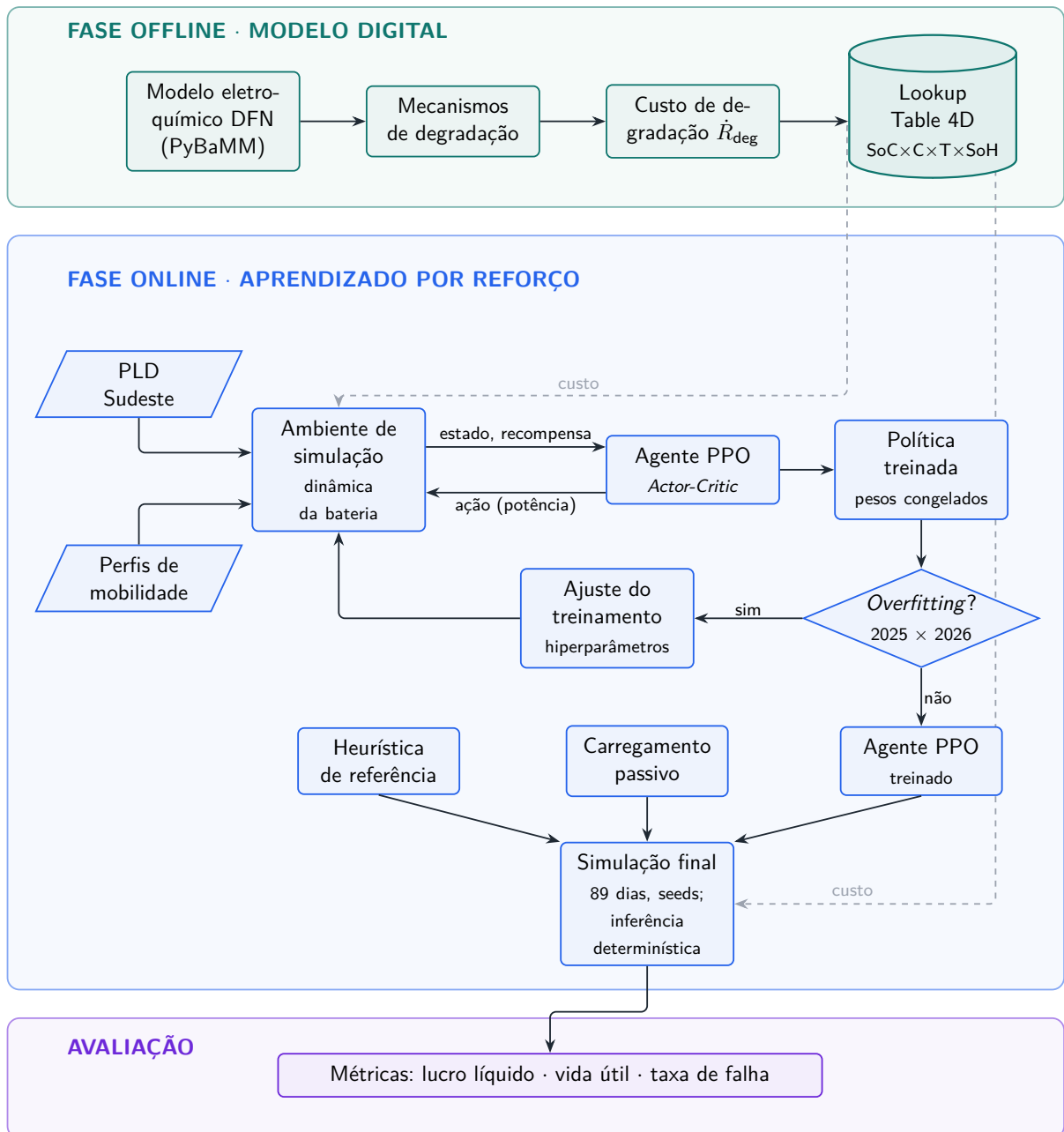
Este trabalho utilizou uma arquitetura de simulação computacional para a otimização de estratégias de V2G com agentes de Aprendizado por Reforço. A modelagem eletroquímica, a formulação do problema de otimização estocástica a estrutura da função de recompensa (*reward shaping*) e a dinâmica do mercado de energia baseiam-se em Niazi *et al.* (2026), entretanto, no contexto brasileiro para o mercado, considerou-se apenas o PLD para remuneração, o que será discutido mais a frente.

Para viabilizar a velocidade de processamento exigida pelo treinamento de RL, da ordem de 10^6 interações por cenário com $\Delta t = 5$ min cada, a arquitetura foi dividida em duas etapas. Na fase *offline*, o comportamento físico da bateria é simulado uma única vez no PyBaMM e tabulado em *lookup tables* 4D. Na fase *online*, o ambiente de RL consulta essas tabelas por interpolação a cada passo, recuperando o custo de degradação em microssegundos conforme a Figura 5, que sintetiza essa arquitetura.

3.1.1 As duas fases computacionais

A fase *offline* é responsável pela geração do modelo digital da bateria, onde modelos eletroquímicos espacialmente resolvidos simulam o comportamento dinâmico e o envelhecimento simultâneo da célula sob diferentes regimes elétricos, térmicos e de estado interno. Antes do treinamento, varre-se uma malha exaustiva de combinações de Estado de Carga (*State of Charge*, SoC), temperatura de operação, Estado de Saúde (*State of Health*, SoH) e taxa C (*C-rate*), e armazena-se o resultado em matrizes multidimensionais, as quais chamaremos de *lookup tables*. Cada entrada da tabela retorna as taxas instantâneas de degradação por mecanismo físico (crescimento da camada de interfase eletrólito-sólido, deposição de lítio metálico, perda de material ativo) e a variação térmica esperada naquela condição. Esta fase é descrita em detalhe nas seções 3.2 e 3.3.

A fase *online* é responsável pelo treinamento do agente, onde o ambiente estocástico de simulação V2G consome as tabelas por interpolação quadridimensional em tempo real. A cada passo de 5 minutos, o ambiente recebe a ação do agente (a potência normalizada aplicada à bateria), atualiza o estado físico do veículo, calcula a recompensa financeira pelo PLD horário e dela subtrai o custo de degradação obtido por consulta à *lookup table*, e os demais custos de orientação do agente. O algoritmo PPO otimiza a política ao longo de 10^6 passos por cenário. O ambiente é compatível com a biblioteca *Gymnasium*, escolha motivada pela sua adoção majoritária em aplicações de RL e pela facilidade de integração com bibliotecas de treinamento; trabalhos recentes de simulação V2G adotaram o mesmo padrão (ORFANOUDAKIS *et al.*, 2024). Detalhes nas Seções 3.4 e 3.5.

Figura 5 – Fluxograma da arquitetura de simulação V2G adotada.

Fonte: o autor, 2026.

3.1.2 Modelo Digital

O modelo digital da bateria é a representação virtual de uma célula, construída a partir de parâmetros físicos e das equações que governam seus mecanismos internos, reproduzindo seu comportamento eletroquímico e de envelhecimento com fidelidade estrutural. O modelo é construído e calibrado *offline*, a partir de parâmetros de literatura, e usado para gerar as taxas de degradação que alimentam o agente.

A motivação dessa escolha é que estratégias V2G submetem a bateria a regimes operacionais distintos do uso veicular convencional. O agente pode carregar em momentos de PLD baixo, descarregar em momentos de maior valor econômico e ainda preservar energia para as

viagens do usuário. Isso altera a profundidade e quantidade dos ciclos, a taxa de corrente e o perfil térmico, e exige que a degradação seja avaliada de forma sensível ao contexto operacional, e não como custo fixo por kWh transferido. O modelo digital, portanto, isola as variáveis geométricas das variáveis químicas, oferecendo uma base comparativa coerente para as duas composições estudadas, descritas na Seção 3.2.3.

3.1.3 Framework PyBaMM

Para resolver as equações que governam o modelo digital, adotou-se o *framework* aberto PyBaMM (*Python Battery Mathematical Modelling*) (SULZER *et al.*, 2021). Ele implementa de forma nativa o modelo Doyle-Fuller-Newman (DFN) (DOYLE *et al.*, 1993), que resolve a célula no espaço e é mais fiel que modelos reduzidos como o de Partícula Única (*Single Particle Model*, SPM), e permite acoplar a esse núcleo os submodelos de degradação usados neste trabalho, o crescimento da SEI, o *plating* de lítio e a mecânica de partícula, os mesmos mecanismos tratados por Niazi *et al.* (2026).

A combinação do DFN com esses submodelos é chamada de *DFN completo* no restante do capítulo, com a ressalva de que o LFP não inclui mecânica de partícula.

3.2 MODELAGEM ELETROQUÍMICA DA BATERIA

3.2.1 Modelo DFN e aproximação térmica lumped

A mecânica interna da célula foi modelada pela formulação Doyle-Fuller-Newman (DFN), também chamada de pseudo-bidimensional (P2D), proposta por Doyle *et al.* (1993) e usada por Niazi *et al.* (2026) como base da simulação eletroquímica. O DFN acompanha, ao longo da espessura do eletrodo, o transporte de íons e de elétrons nos eletrodos e no eletrólito, e resolve também a difusão do lítio no interior das partículas esféricas de material ativo. O nome pseudo-2D vem dessa combinação de duas escalas, uma macroscópica, no eixo que atravessa o eletrodo negativo, o separador e o eletrodo positivo, e uma microscópica, no raio de cada partícula, ligadas pela cinética de reação na superfície. Por enxergar os gradientes internos de concentração e de potencial que antecedem a degradação, é um modelo mais fiel que representações reduzidas, como o de partícula única.

Sobre esse núcleo, o PyBaMM acopla os submodelos de degradação tratados na próxima subseção. Os submodelos efetivamente ativados em cada química estão reunidos na Tabela 22 do Apêndice A.

A operação da bateria também gera calor, por perdas ôhmicas, pelos sobrepotenciais das reações e pela variação de entropia, e esse calor realimenta a cinética, já que as constantes de reação e os coeficientes de difusão dependem da temperatura. Está aqui a principal diferença em relação a Niazi *et al.* (2026), pois enquanto o artigo resolve a equação do calor ao longo da célula, este trabalho trata a temperatura como uniforme em todo o corpo da célula e acompanha apenas a sua evolução no tempo, a aproximação chamada de *lumped*. Ela é

adequada às células cilíndricas estudadas porque o número de Biot, que compara a resistência à condução de calor dentro da célula com a resistência à troca de calor com o ambiente, mantém-se bem abaixo da unidade nas condições simuladas. Sob convecção natural ($h \approx 10 \text{ W}/(\text{m}^2 \cdot \text{K})$) e taxas operacionais moderadas ($|C| \leq 1$), o gradiente interno de temperatura permanece de poucos graus, o que justifica tratar a temperatura como uniforme e dispensa resolver o campo espacial completo, que só aumentaria a complexidade computacional da geração das tabelas sem ganho relevante no resultado.

3.2.2 Mecanismos de degradação

As variáveis internas resolvidas pelo DFN alimentam os modelos de degradação. São tratados os três mecanismos que a operação V2G mais acelera, o crescimento da camada de interfase (SEI), a fadiga mecânica das partículas com perda de material ativo e a deposição de lítio metálico, cada um descrito pelo que representa fisicamente, como o PyBaMM o calcula e pelo ponto em que este trabalho difere de Niazi *et al.* (2026).

O Crescimento da SEI conforme já descrito na seção 2.1.3 se forma na primeira carga, quando parte do eletrólito se decompõe na superfície do ânodo de grafite, e continua crescendo devagar ao longo da vida. Esse crescimento consome lítio ciclável, o que reduz a capacidade, e engrossa o filme, o que aumenta a resistência interna, sendo a principal via de envelhecimento da célula em repouso. No PyBaMM, o crescimento é modelado na forma limitada por difusão de solvente, em que o passo lento é a chegada do solvente até a superfície do ânodo através da camada já formada, de modo que quanto mais espessa a SEI, mais devagar ela cresce. Essa auto-limitação surge do próprio transporte, sem depender de expoentes de ajuste, e é justamente o ponto em que o modelo difere de Niazi *et al.* (2026), que descreve a SEI por uma lei empírica de Arrhenius com dois expoentes cujos valores o artigo não publica. O PyBaMM ainda acompanha a SEI que se forma sobre as microfissuras, acoplando esse crescimento ao dano mecânico.

A Mecânica de partícula e perda de material ativo, referente a entrada e a saída de lítio que fazem a partícula de material ativo inchar e contrair, com o centro mudando de concentração mais devagar que a superfície, surgindo tensões internas que, repetidas a cada ciclo, abrem microfissuras e, no limite, desprendem material ativo do eletrodo, o qual se chama perda de material ativo (LAM). No PyBaMM, esse caminho é representado pelos submodelos de inchamento e fissuração e pela perda de material ativo dirigida por tensão, com as funções de propagação de fissura de Ai *et al.* (2020). O modelo resolve as tensões no interior da partícula e leva em conta apenas o esforço que estica o material, responsável por abrir as fissuras, já que a compressão tende a fechá-las. Em Niazi *et al.* (2026), a perda de material ativo é escrita a partir de uma tensão equivalente genérica, sem essa distinção.

Esse submodelo foi aplicado somente ao NMC, cuja estrutura em camadas, rica em níquel, sofre variação de volume anisotrópica durante a ciclagem e concentra tensões que nucleiam microfissuras, ao passo que a olivina do LFP varia pouco e de forma mais isotrópica,

amparada por ligações fósforo-oxigênio fortes, o que torna a perda mecânica de material praticamente irrelevante nessa química (WANG; SUN, 2015). A essa robustez física somam-se duas limitações práticas, a ausência, na literatura, de parâmetros de degradação mecânica confiáveis para o LFP e a falha de convergência do solver quando o submodelo é ativado sobre o conjunto de parâmetros dessa célula. A assimetria resultante recomenda cautela na comparação da magnitude absoluta de degradação entre as químicas, embora a comparação entre perfis dentro de uma mesma química permaneça válida.

A Deposição de lítio metálico ocorre em carga rápida ou em baixa temperatura, onde parte do lítio que deveria se intercalar no grafite se deposita como metal na superfície do ânodo. O fenômeno é parcialmente reversível, pois na descarga seguinte parte do metal volta a se dissolver, mas a fração que se isola eletricamente, o lítio morto, custa capacidade e representa risco de segurança. No PyBaMM, a deposição é modelada na forma reversível, e o sobrepotencial que a governa inclui a queda de tensão através do filme de SEI, de modo que, à medida que a SEI cresce, essa queda reduz o sobrepotencial efetivo de deposição e cria uma proteção parcial ao longo da vida. Esse acoplamento entre SEI e deposição, considerado nesta monografia, não aparece em Niazi *et al.* (2026), que adota um sobrepotencial termodinâmico simples.

3.2.3 Parametrização por química

Para avaliar o efeito da química do cátodo sobre o lucro V2G e a vida útil, foram simuladas duas configurações que representam os extremos do mercado atual, a NMC-811, de alta densidade energética, e a LFP, robusta ao ciclamento. A Tabela 1 reúne a capacidade, a topologia do *pack* e o conjunto de parâmetros do PyBaMM de cada uma. Duas químicas intermediárias da família NMC foram cogitadas e descartadas, a NMC-622, sem conjunto de parâmetros validado para ciclagem no PyBaMM, e a NMC-532, que disparava *cutoff voltage* em descargas moderadas durante a geração da tabela. As duas mantidas correspondem aos tipos A e C da frota de Niazi *et al.* (2026).

Tabela 1 – Parametrização das duas químicas estudadas.

Química	Capacidade	Pack assumido	Conjunto base	Célula de referência
NMC-811 (Type A)	80 kWh	111s × 40p (4440 células)	Chen2020	LG M50 (21700, 5,0 Ah)
LFP (Type C)	50 kWh	125s × 54p (6750 células)	Prada2013	A123 26650 (2,3 Ah)

Fonte: Chen *et al.* (2020) (NMC-811) e Prada *et al.* (2013) (LFP).

A topologia série-paralelo de cada *pack* assume tensão nominal de 400 V. O conjunto Chen2020 caracteriza a célula 21700 da LG Chem, amplamente usada como referência de modelagem DFN, e fornece também os parâmetros térmicos do NMC-811. Já o Prada2013, da célula A123 26650 (LFP), é mais antigo e não traz parâmetros térmicos nem cinéticas de degradação modernas, acrescentados a partir da literatura da própria célula.

Os parâmetros físicos de cada química, com os ajustes feitos neste trabalho, estão reunidos na Tabela 23 do Apêndice A, e vale destacar dois deles. A difusividade de solvente, que controla o crescimento da SEI, foi recalibrada por química, já que os valores nativos do PyBaMM são ilustrativos; os valores adotados ficam dentro do intervalo explorado por O’Kane *et al.* (2022), e o do LFP, ainda dentro desse intervalo, teve influência do envelhecimento em repouso medido por Naumann *et al.* (2018). A resistividade do filme de SEI, que determina quanto o crescimento da camada eleva a resistência interna, foi tomada de Tahmasbi *et al.* (2017) para reproduzir as taxas de crescimento de resistência observadas na literatura de V2G em operação intensa.

O segundo ajuste são os parâmetros mecânicos do NMC, também reunidos na Tabela 23 do apêndice. A tensão crítica de 150 MPa adotada para o cátodo situa-se dentro da faixa medida por Wheatcroft *et al.* (2024) para partículas de NMC-811, entre cerca de 86 e 180 MPa, uma escolha conservadora. O LFP não aparece porque a mecânica de partícula foi mantida desativada para essa química, conforme a Seção 3.2.2.

3.3 GERAÇÃO DAS *LOOKUP TABLES* DE DEGRADAÇÃO

3.3.1 Estrutura geral da Fase Offline

A Fase Offline executa, antes do treinamento do agente, uma varredura exaustiva da resposta físico-química da bateria sob condições representativas de operação V2G. Para cada ponto do espaço de estados, dado pela quádrupla (SoC, T, SoH, C_{rate}), o modelo digital simula um *pulso* de 5 minutos de duração e registra as taxas instantâneas de degradação por mecanismo físico e o custo monetário associado. O resultado consolidado é uma matriz tabular indexada pelas quatro variáveis de entrada e que retorna seis variáveis de saída (descritas no §3.3.4), serializada em arquivo CSV para consumo direto pelo ambiente Gymnasium.

O procedimento se divide em três etapas. A primeira é a discretização do espaço de estados (§3.3.2). A segunda é o protocolo de simulação ponto a ponto no PyBaMM, que trata de forma específica os pontos em repouso, com taxa C nula (§3.3.3). A terceira é a conversão das taxas físicas em custo financeiro por hora (§3.3.5), segundo a adaptação de Niazi *et al.* (2026). Na fase online (Figura 5), o ambiente consulta essa tabela a cada passo do treinamento.

3.3.2 Discretização do *Grid* 4D

Cada uma das duas químicas estudadas tem sua própria *lookup table* gerada sobre o mesmo *grid* de entrada, listado na Tabela 2.

O total é $21 \times 21 \times 11 \times 5 = 24\,255$ pontos por química, ou seja, 48 510 simulações eletroquímicas ao se considerar as duas químicas estudadas. A resolução de 0,1 na taxa C foi escolhida para dobrar a densidade do *grid* na faixa operacional efetiva do V2G residencial bra-

Tabela 2 – Discretização do espaço 4D adotada para a *lookup table*.

Variável	Intervalo	Passo	Nº de pontos	Comentário
Estado de carga, SoC	[0,0; 1,0]	0,05	21	Cobre o intervalo físico completo da célula
Taxa C, C_{rate}	[-1,0; +1,0]	0,1	21	$C > 0$ carga; $C < 0$ descarga V2G
Temperatura, T	[263,15; 313,15] K (-10 a 40 °C)	5 K	11	Faixa típica ambiente SE com sobras
Estado de saúde, SoH	[0,80; 1,00]	0,05	5	SoH = 0,80 é fim-de-vida automotivo

sileiro (entre 0,1C e $\sim 0,3C$, correspondente à potência de carregadores de 7-22 kW operando sobre baterias de 50-80 kWh).

3.3.3 Protocolo de Simulação por Ponto

Cada ponto do *grid* é instanciado como um experimento independente no PyBaMM. A célula é inicializada nas condições (SoC_0, T_0, SoH_0), e um *step* de 5 minutos é aplicado com a taxa C correspondente. Esse intervalo é o mesmo Δt utilizado pelo ambiente de RL na Fase Online, garantindo consistência temporal entre a tabulação e a consulta. Os dois caminhos possíveis no protocolo são:

- Pontos de ciclagem ($C_{rate} \neq 0$): aplica-se um *Hold at C-rate for 5 minutes*. O solver IDAKLU integra o sistema completo de PDEs do DFN com todos os submodelos de degradação ativos (SEI, plating reversível e, no caso do NMC, *swelling/cracking*), e as variáveis de saída são lidas em $t = 5$ min.
- Pontos de calendar aging ($C_{rate} = 0$): aplica-se um *Rest for 5 minutes*. A célula permanece em circuito aberto, sem corrente líquida, mas a SEI continua crescendo lentamente pela difusão de solvente (mecanismo descrito na Seção 3.2.2) e a temperatura relaxa em direção a $T_{ambiente}$. As demais cinéticas (plating, mecânica de partícula) ficam inativas, pois não há *driving force* eletroquímica.

Em algumas condições extremas do *grid*, como temperatura muito baixa (abaixo de 0 °C) combinada com SoH reduzido, ou taxas próximas do limite com o SoC saturado, o solver não converge dentro do tempo de simulação. Nesses pontos, em vez de deixar a degradação indefinida, o procedimento recorre a um modelo empírico de envelhecimento, o de Schmalstieg *et al.* (2014) para o NMC-811 e o de Naumann *et al.* (2018) para o LFP, que estima a parcela de *calendar aging* a partir da temperatura e do SoC, ajustada pela taxa C nos pontos de ciclagem. Esse caminho afeta cerca de 2% do *grid* e evita pontos com degradação nula que o agente interpolaria como se fossem reais. A fração de pontos resolvidos por essa via é registrada nos metadados de cada tabela.

Como cada ponto da tabela é simulado a partir de uma célula nova, sem lítio metálico acumulado, o modelo de deposição reversível registra uma pequena corrente de deposição mesmo em regimes onde ela não deveria existir, como na descarga com SoC alto. Esse valor é um efeito da simulação ponto a ponto e levaria o agente a contabilizar um custo de degradação inexistente, então a deposição é zerada sempre que as condições do pulso indicam que ela não ocorre.

3.3.4 Variáveis Extraídas

Para cada ponto, o protocolo registra seis variáveis de saída, listadas na Tabela 3. As quatro primeiras são taxas instantâneas de degradação por mecanismo, expressas como fração por hora (h^{-1}), convenção dimensional uniforme que facilita a soma posterior na função custo. A quinta é a variação térmica da célula no intervalo de 5 minutos, em kelvin. A sexta é o custo monetário horário consolidado (detalhado em §3.3.5).

Tabela 3 – Variáveis registradas em cada ponto da *lookup table*.

Variável	Unidade	Significado físico	Origem
$\dot{Q}_{\text{SEI}}$	h^{-1}	Taxa de perda de capacidade por crescimento da SEI (inclui SEI on cracks)	PyBaMM (SEI)
$\dot{R}_{\text{SEI}}$	h^{-1}	Taxa de crescimento fracional da resistência interna ($\Delta R_{\text{cell}}/R_{\text{cell}}$)	PyBaMM (SEI)
$\dot{Q}_{\text{plat}}$	h^{-1}	Taxa de perda de capacidade por <i>plating</i> (líquida de <i>stripping</i> e <i>dead Li</i>)	PyBaMM (<i>plating</i>)
$\dot{f}_{\text{AML}}$	h^{-1}	Taxa de perda fracional de material ativo (ânodo + cátodo)	PyBaMM (mecânica, só NMC)
$\Delta T_{5\text{min}}$	K	Variação de temperatura da célula no pulso	PyBaMM (térmico, <i>lumped</i>)
$\dot{R}_{\text{deg}}$	R\$/h	Custo monetário horário consolidado da degradação	Eq. (3.1), apresentada a seguir

A normalização fracional da resistência ($\Delta R_{\text{cell}}/R_{\text{cell}}$) é invariante à topologia série/paralelo do *pack*, isto é, vale tanto em nível de célula quanto em nível de *pack*. Essa conversão é feita no próprio pipeline, dividindo ΔR_{SEI} (obtido em $\Omega \cdot \text{m}^2$ na saída do PyBaMM) pela área do eletrodo correspondente e pela resistência inicial da célula, $R_{\text{cell}} = R_{\text{pack}} \cdot (n_p/n_s)$, com a topologia (n_s, n_p) definida na seção §3.2.3.

3.3.5 Monetização da Degradação

O custo monetário horário da degradação é calculado pela adaptação da Eq. 16 de Niazi *et al.* (2026), aqui renumerada como Eq. (3.1):

$$\dot{R}_{\text{deg}} = \alpha_Q \cdot |\dot{Q}_{\text{SEI}} + \dot{Q}_{\text{plat}}| + \alpha_R \cdot |\dot{R}_{\text{SEI}}| + \alpha_{\text{AML}} \cdot |\dot{f}_{\text{AML}}| \quad (3.1)$$

Onde:

- $\dot{R}_{\text{deg}}$ é o custo monetário de degradação física por hora (R\$/h);
- α_Q é o custo de reposição total da bateria (R\$), associado à perda integral de capacidade, e corresponde, fisicamente, ao valor que o proprietário pagaria por um novo *pack* da mesma química quando o atual atingisse 0% de capacidade;
- α_R é o coeficiente de monetização do crescimento fracional de resistência (R\$ por unidade da fração $\Delta R/R$);
- α_{AML} é o coeficiente de perda de material ativo (R\$ por unidade da fração Δf_{AML});
- os módulos garantem que apenas a magnitude da degradação contribui, independentemente do sinal da taxa C .

A primeira contribuição agrupa as duas vias de perda de capacidade (SEI e plating) sob o mesmo coeficiente α_Q , pois ambos os mecanismos consomem lítio ciclável de forma equivalente do ponto de vista do proprietário. A segunda monetiza separadamente o crescimento de resistência, que reduz a eficiência energética e a potência disponível mesmo antes da capacidade chegar ao limite. A terceira monetiza a perda de material ativo, que se manifesta como degradação irreversível da arquitetura do eletrodo.

Sobre o efeito da temperatura, $\dot{R}_{\text{deg}}$ não inclui o termo de *reward shaping* térmico $\beta_T(T - T_{\text{opt}})^2$ presente nas funções de recompensa do agente. Esse termo entra explicitamente como P_t^{thermal} na Eq. (3.18) da §3.4.5 e é tratado como penalidade artificial separada, e não como custo físico, pois o efeito da temperatura sobre as taxas $\dot{Q}_{\text{SEI}}$, $\dot{Q}_{\text{plat}}$ e $\dot{f}_{\text{AML}}$ já é capturado pelo PyBaMM via Arrhenius nas constantes cinéticas. Somar $\beta_T(T - T_{\text{opt}})^2$ a $\dot{R}_{\text{deg}}$ implicaria dupla contagem do efeito térmico no custo total reportado.

Os coeficientes α adotados foram derivados do custo de reposição por kWh de cada química, segundo a Tabela 4. Os valores em USD/kWh de NMC-811 e LFP replicam os adotados por Niazi *et al.* (2026) (US\$ 145/kWh e US\$ 120/kWh, respectivamente), convertidos a R\$/kWh com câmbio (valor de referência em relação ao dia 30 de março de 2026) do dia 1 USD = 5,27 BRL.

Tabela 4 – Coeficientes da Eq. (3.1) por química.

Química	Custo (R\$/kWh)	α_Q (R\$)	$\alpha_R = \alpha_Q/4$ (R\$)	α_{AML} (R\$)
NMC-811 (80 kWh)	764,15	61.132	15.283	61.132
LFP (50 kWh)	632,40	31.620	7.905	31.620

Derivação do coeficiente $\alpha_R = \alpha_Q/4$: Niazi *et al.* (2026) sugerem na Eq. 16 o uso direto de α_Q para o crescimento de resistência, partindo da premissa de que uma resistência dobrada equivale economicamente à perda integral do *pack*, e nesta monografia alterou-se esse valor pelas seguintes motivações:

- Valor residual em *second-life*: estudos técnico-econômicos sobre o reaproveitamento de *packs* automotivos em aplicações estacionárias (armazenamento residencial, *Behind-The-Meter*, *peak-shaving*) indicam que um *pack* com resistência dobrada, critério clássico de fim de vida automotivo, ainda preserva aproximadamente 50% do valor econômico do *pack* novo (NEUBAUER; PESARAN, 2011; MARTINEZ-LASERNA *et al.*, 2018). Logo, o custo atribuível ao R-fade ao longo da vida automotiva é da ordem de $0,5 \cdot \alpha_Q$, e não α_Q integral.
- Causa física única, dois canais observáveis: o crescimento da SEI é um único processo físico que se manifesta simultaneamente como perda de capacidade ($\dot{Q}_{SEI}$) e como crescimento de resistência ($\dot{R}_{SEI}$). Como o canal de capacidade já precifica esse processo de forma integral por meio de α_Q , atribuir novamente α_Q ao canal de resistência contaria duas vezes o mesmo evento físico, e por isso apenas o canal de resistência recebe um fator multiplicativo adicional de 0,5.

A composição multiplicativa dos dois fatores ($0,5 \times 0,5 = 0,25$) resulta em $\alpha_R = \alpha_Q/4$. Para verificar a coerência, considera-se o estado de fim de vida automotivo caracterizado pelos dois critérios físicos clássicos, a perda de 20% de capacidade e a duplicação da resistência: o custo cumulativo seria $0,20 \cdot \alpha_Q$ no canal de capacidade somado a $0,25 \cdot \alpha_Q$ no canal de resistência (resistência dobrada multiplicada por $\alpha_R = \alpha_Q/4$), totalizando $\approx 0,45 \cdot \alpha_Q$, coerente com a meta de cerca de 50% do custo total. Esses dois critérios físicos não se confundem com o estado de saúde operacional da simulação, que combina capacidade e resistência de forma ponderada na Eq. (3.10).

3.3.6 Consulta em Tempo Real

Durante o treinamento, a cada passo de $\Delta t = 5$ min, o ambiente lê na *lookup table* as seis variáveis correspondentes ao estado atual da bateria (SoC_t, T_t, SoH_t, C_t), listadas na Tabela 3, entre elas o custo $\dot{R}_{deg,t}$. Como esse estado quase nunca coincide exatamente com um ponto tabelado, o valor é estimado por interpolação linear entre os pontos vizinhos do *grid*, nas quatro dimensões. Se o estado cair fora do *grid*, ele é limitado à borda mais próxima, situação rara porque o ambiente já mantém o SoC e a potência dentro das mesmas faixas usadas na tabela (clip de SoC em $[SoC_{min}, SoC_{max}]$ e de potência em $\pm 1C$).

O custo de degradação consumido em cada passo é então $\dot{R}_{deg,t} \cdot \Delta t$, em reais, e entra subtraído na função de recompensa do agente, formalizada no §3.4.

Troca de calor com o ambiente entre passos

A *lookup table* é gerada com a temperatura ambiente igual à temperatura inicial de cada ponto ($T_{amb} = T_0$), o que evita uma quinta dimensão na tabela e mantém o custo computacional viável. Essa escolha tem uma consequência. Quando a célula, na simulação

online, chega a temperaturas bem acima do ambiente real, por exemplo 50 °C com $T_{\text{amb,real}} = 25$ °C, a tabela não enxerga o gradiente entre a célula e o ambiente e subestima a dissipação de calor. Para recuperar essa troca de calor, o ambiente *online* soma ao $\Delta T_{5\text{min}}$ devolvido pela tabela um termo proporcional ao excesso de temperatura da célula sobre o ambiente real,

$$\Delta T_{\text{dis}}^{(t)} = -k_{\text{dis}} \cdot (T_t - T_{\text{amb,real}}) \quad (3.2)$$

onde $k_{\text{dis}} = 0,40$ é a fração do excesso dissipada por passo de 5 min, T_t é a temperatura da célula e $T_{\text{amb,real}}$ é a temperatura ambiente.

A cada passo, o ciclamento injeta calor por meio do $\Delta T_{5\text{min}}$ da tabela, e o termo acima devolve ao ambiente uma fração desse calor, tanto maior quanto mais quente está a célula. Esse mecanismo ativa a dissipação para o ambiente e reduz a velocidade com que a temperatura, de forma a reproduzir os mecanismos de segurança de um sistema BMS (*Battery Management System*). Os dois efeitos se equilibram quando a injeção iguala a dissipação, e a temperatura estaciona. Para o NMC-811 sob carga contínua de 1C, esse equilíbrio fica em torno de 39 °C, bem abaixo do teto de segurança de 60 °C. Nos períodos de repouso, sem geração de calor, o mesmo termo traz a temperatura de volta em direção ao ambiente. O coeficiente $k_{\text{dis}} = 0,40$ foi ajustado para reproduzir uma dissipação fisicamente plausível, com constante de tempo térmica de cerca de 10 min ($\tau = -\Delta t / \ln(1 - k_{\text{dis}}) \approx 10$ min), dentro da faixa esperada para um *pack* veicular com gerenciamento térmico ativo e coerente com os coeficientes de troca de calor reportados por Pesaran (2002) e Bandhauer *et al.* (2011), da ordem de dezenas de W/(m² · K) na refrigeração forçada a ar e de algumas centenas na refrigeração líquida.

No ponto de calibração de *calendar aging* ($T = 25$ °C, $C = 0$), a célula está em equilíbrio com o ambiente e o termo se anula, sem efeito sobre esse ponto. A temperatura ambiente é mantida fixa em 25 °C durante toda a simulação, simplificação que não cobre o ciclo dia-noite nem a sazonalidade.

3.4 FORMULAÇÃO DO MDP E DA FUNÇÃO DE RECOMPENSA

3.4.1 Processo de Decisão de Markov

A interação entre o agente e o sistema é formalizada como um Processo de Decisão de Markov (*Markov Decision Process*, MDP) discreto no tempo, representado pela tupla canônica:

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma \rangle \quad (3.3)$$

Onde:

- $\mathcal{S} \subseteq \mathbb{R}^{d_o}$ é o espaço de observações, com dimensão $d_o = 25$ no ambiente baseline e $d_o = 21$ no ambiente estocástico (detalhados no §3.4.3);

- $\mathcal{A} = [-1, +1]^{N_v}$ é o espaço de ações contínuo, com $N_v = 2$ veículos no baseline e $N_v = 1$ no estocástico (§3.4.4);
- $\mathcal{P}(s' | s, a)$ é a função de transição, determinística para a dinâmica de SoC e temperatura, e estocástica para a evolução do PLD horário e para a ocorrência de viagens no ambiente estocástico;
- $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ é a função de recompensa (§3.4.5);
- $\gamma \in [0, 1]$ é o fator de desconto, cujo valor adotado é justificado em §3.5.

Temporalidade: cada passo do ambiente corresponde a um intervalo fixo $\Delta t = 5$ min, ou seja, 288 passos por dia. Os episódios têm horizonte de 30 dias no treinamento e na validação Monte Carlo, e de 89 dias na simulação final (correspondente ao primeiro trimestre de 2026). Em cada reset do ambiente, o dia inicial do episódio é amostrado uniformemente do *dataset* de PLD durante o treinamento; na simulação final, esse dia é fixo, para garantir reprodutibilidade.

Esta monografia mantém duas instâncias do MDP executadas em paralelo:

- Ambiente baseline (`V2GBaselineEnv`): opera com dois veículos representando uma microfrota (um por química), com viagens determinísticas (chegada às 18:00, partida às 07:00).
- Ambiente estocástico (`V2GStochasticEnv`): opera com um veículo individual sob viagens amostradas a partir de distribuições truncadas (§3.4.2).

A manutenção de dois ambientes em paralelo permite comparar o aprendizado e as decisões do agente sob diferentes níveis de incerteza. O contraste entre o cenário determinístico e o estocástico serve para avaliar o impacto das duas penalidades comportamentais adicionadas (ansiedade e falha), verificando se tiveram papel relevante no treinamento quando comparadas à formulação original de Niazi *et al.* (2026).

3.4.2 Modelagem Estocástica do Usuário

Três perfis de mobilidade

O ambiente estocástico opera sobre três perfis de mobilidade representativos do uso veicular brasileiro, listados na Tabela 5, e foram calibrados na tentativa de representar 3 tipos de usuários brasileiros que poderiam vir a fazer o uso do V2G. Cada perfil é caracterizado por (i) um Estado de Carga mínimo de segurança e_{seg}^{SoC} , abaixo do qual o motorista experimenta ansiedade de autonomia (*range anxiety*; Neubauer e Wood (2014)); (ii) uma ou duas viagens diárias, cada uma com horário de partida, horário de retorno e distância percorrida amostrados de distribuições próprias descritas adiante.

A hierarquia de $e_{\text{seg}}^{\text{SoC}}$ entre os perfis reflete a dependência operacional do veículo. O perfil Intenso tem o limiar mais alto (0,40), pois o veículo é instrumento de trabalho e a falha de mobilidade compromete a renda (a exemplo de um motorista de aplicativo); o Conservador (a exemplo de um trabalhador presencial) e o Flexível (a exemplo de alguém que trabalha em *home-office*) compartilham o mesmo limiar (0,30), refletindo rotina regular com baixa criticidade operacional. O valor 0,30 no Flexível foi adotado deliberadamente acima do clamp $\text{SoC}^{\text{min}} = 0,20$ do NMC, de modo a manter a penalidade de ansiedade sensível mesmo quando o veículo opera próximo ao limite inferior do BMS.

Tabela 5 – Parâmetros estocásticos dos três perfis de mobilidade.

Parâmetro	Conservador	Flexível	Intenso
SoC de segurança $e_{\text{seg}}^{\text{SoC}}$	0,30	0,30	0,40
Viagem 1, partida ($\mu \pm \sigma$)	07:30 $\pm$ 15 min	14:00 $\pm$ 120 min	06:00 $\pm$ 20 min
Viagem 1, retorno ($\mu \pm \sigma$)	18:30 $\pm$ 45 min	17:30 $\pm$ 90 min	12:00 $\pm$ 30 min
Viagem 1, distância ($\mu \pm \sigma$)	15 $\pm$ 5 km	20 $\pm$ 15 km	90 $\pm$ 20 km
Probabilidade de ocorrência π_1	1,00	1,00	1,00
Viagem 2, partida	20:30 $\pm$ 30 min	21:00 $\pm$ 60 min	14:30 $\pm$ 30 min
Viagem 2, retorno	22:00 $\pm$ 30 min	02:00 (+1d) $\pm$ 90 min	21:30 $\pm$ 60 min
Viagem 2, distância	10 $\pm$ 5 km	25 $\pm$ 10 km	110 $\pm$ 30 km
Probabilidade π_2	0,15	0,60	0,95
Distância esperada/dia	$\sim$ 16,5 km	$\sim$ 35 km	$\sim$ 194 km
Janela média de conexão	$\sim$ 13 h	$\sim$ 18 h	$\sim$ 11 h

Amostragem das viagens

A cada reset do ambiente, e a cada novo dia do episódio, a presença das viagens 1 e 2 é amostrada de uma Bernoulli(π_i). Para cada viagem efetivamente ocorrida, três variáveis são amostradas:

- Horário de partida $t_{\text{dep},i}$ e horário de retorno $t_{\text{ret},i}$: amostrados de distribuições Normais Truncadas em $[\mu - 3\sigma, \mu + 3\sigma]$. A escolha pela Normal Truncada se justifica pelo fato de esses horários estarem ligados a relógio biológico e estrutura social de trabalho/escola, sem motivo empírico para distribuições assimétricas.
- Distância percorrida d_i : amostrada de uma distribuição Weibull de forma k_i e escala λ_i , calibrada para preservar a média μ_i e o desvio-padrão σ_i declarados na Tabela 5. A escolha pela Weibull, em vez da Normal Truncada, acompanha a assimetria positiva típica das distâncias diárias percorridas (PLÖTZ *et al.*, 2017), com forma e escala ajustadas à média e ao desvio-padrão de cada perfil.

Além das viagens, o estado de carga inicial de cada episódio também é amostrado a cada reset, a partir de uma distribuição Normal de média 0,30 e desvio-padrão 0,05, recortada aos limites operacionais da química ($[\text{SoC}_{\min}, \text{SoC}_{\max}]$).

Os parâmetros de forma e escala de cada viagem foram calibrados por perfil para reproduzir a média e o desvio-padrão da Tabela 5, e estão na Tabela 25 do Apêndice A. A amostra é restrita ao intervalo físico $[1; 500]$ km, e a energia consumida em cada viagem é $E_{\text{trip},i} = d_i \cdot \eta_{\text{EV}}$, com eficiência veicular fixada em $\eta_{\text{EV}} = 0,15$ kWh/km.

Probabilidade instantânea de partida

A penalidade de ansiedade (§3.4.5) requer uma estimativa da probabilidade de que uma partida ocorra no intervalo $[t, t + \Delta t]$, *dado que ainda não ocorreu até t* . Para cada viagem i ainda não realizada, define-se a probabilidade instantânea de partida no tempo discreto:

$$h_i(t) = \min\left(1, 0, \frac{F_i(t + \Delta t) - F_i(t)}{\max(1 - F_i(t), \varepsilon)}\right) \quad (3.4)$$

Onde F_i é a CDF da Normal Truncada associada ao horário de partida da viagem i , e $\varepsilon = 10^{-6}$ é um termo numérico que evita divisão por zero conforme $F_i(t) \rightarrow 1$. O operador $\min(1, \cdot)$ garante que h_i nunca exceda 1. Para múltiplas viagens, a probabilidade combinada de pelo menos uma partida no próximo Δt é obtida por independência:

$$p_{\text{dep}}(t) = 1 - \prod_{i \in I_t} [1 - \pi_i \cdot h_i(t)] \quad (3.5)$$

Onde π_i é a probabilidade Bernoulli de ocorrência da viagem i no dia em curso, e I_t é o conjunto de viagens ainda não realizadas no instante t .

A formulação condicional da Eq. (3.4) é preferível à probabilidade marginal $p_i \cdot (F_i(t + \Delta t) - F_i(t))$ por duas razões. A primeira é que o denominador $1 - F_i(t)$ redistribui a probabilidade que ainda resta, o que faz a probabilidade instantânea de partida crescer de forma contínua conforme t se aproxima de $\mu + 3\sigma$ e traduz a intuição de que, se a partida ainda não ocorreu, ela deve ocorrer logo. A segunda é que o corte $\min(1, \cdot)$ garante que h_i chegue a 1 no último passo antes do truncamento, levando a penalidade de ansiedade ao máximo nessa janela em vez de cair a zero. A variável $p_{\text{dep}}(t)$ é exposta no vetor de observação do ambiente estocástico (§3.4.3), o que permite ao agente antecipar o risco de partida sem esperar o evento, e a antecipação em horizontes mais longos, da ordem de 28 h, emerge naturalmente da função de valor do PPO ao longo do treinamento (§3.5.3), sem precisar de termos antecipativos explícitos na recompensa.

3.4.3 Espaço de Observações

Ambiente baseline (25 dimensões)

O vetor de observação do ambiente baseline segue a Eq. 17 de Niazi *et al.* (2026), adaptada para o contexto brasileiro:

$$\mathbf{o}_t^B = \left[\underbrace{g_t^{\text{time}}, \tilde{p}_t^{\text{PLD}}, \mathbf{l}_t^{\text{PLD}}}_{15 \text{ dim globais}}, \underbrace{\mathbf{v}_t^{(1)}, \mathbf{v}_t^{(2)}}_{5 \text{ dim} \times 2 \text{ veículos}} \right] \in \mathbb{R}^{25} \quad (3.6)$$

Onde:

- $g_t^{\text{time}} = (\sin(2\pi h_t), \cos(2\pi h_t))$ é a representação cíclica da hora do dia, com $h_t \in [0, 1]$ normalizado (2 dimensões);
- $\tilde{p}_t^{\text{PLD}}$ é o PLD horário corrente normalizado por média e desvio-padrão do *dataset* (1 dimensão);
- $\mathbf{l}_t^{\text{PLD}} \in \mathbb{R}^{12}$ é o *lookahead* das próximas 12 horas de PLD, também normalizado, com escolha justificada adiante;
- $\mathbf{v}_t^{(i)} = (\text{SoC}^{(i)}, \text{SoH}^{(i)}, \tilde{T}^{(i)}, \Delta t^{\text{dep},(i)}, c^{(i)})$ é o vetor por veículo (5 dimensões), em que $\tilde{T} = (T - 298,15)/20$ é a temperatura interna normalizada (em kelvin, centrada em 25 °C e escalada por 20 K), Δt^{dep} mede o tempo restante até a próxima partida e $c \in \{0, 1\}$ indica se o veículo está conectado.

Lookahead de 12 horas: Adotou-se um *lookahead* de 12 horas, equivalente a 12 valores na resolução horária do PLD, escolha que cobre a janela de decisão noturna típica (18:00 às 06:00 do dia seguinte). Para simplificação, assume-se que o PLD das próximas 24 horas está sempre disponível ao agente em qualquer instante t , independentemente da hora em que a consulta é feita. Abstrai-se, portanto, o ciclo fixo de publicação da CCEE (que anuncia o PLD do dia seguinte em um horário fixo do dia), de modo que o agente disponha permanentemente da janela $[t, t + 12 \text{ h}]$.

Ambiente estocástico (21 dimensões)

No ambiente individual, o vetor de observação acrescenta duas variáveis diretamente ligadas à ansiedade do motorista:

$$\mathbf{o}_t^S = \left[g_t^{\text{time}}, \tilde{p}_t^{\text{PLD}}, \mathbf{l}_t^{\text{PLD}}, \text{SoC}_t, \text{SoH}_t, \tilde{T}_t, p_{\text{dep}}(t), c_t, \tilde{E}_t^{\text{seg}} \right] \in \mathbb{R}^{21} \quad (3.7)$$

Onde, além das variáveis já definidas na Eq. (3.6):

- $p_{\text{dep}}(t)$ é a probabilidade instantânea de partida calculada pela Eq. (3.5);

- $\tilde{E}_t^{\text{seg}} = E^{\text{seg}}/C_{\text{bat}} = e_{\text{seg}}^{\text{SoC}}$ é o limiar de segurança do perfil ativo, normalizado pela capacidade.

A exposição explícita de $p_{\text{dep}}(t)$ permite ao agente aprender uma política antecipativa, descarregando agressivamente quando p_{dep} é baixo e recarregando antes que p_{dep} suba, sem aguardar o evento de partida. A exposição de $\tilde{E}_t^{\text{seg}}$ informa, em cada instante, o limiar abaixo do qual a penalidade de ansiedade está ativa.

O agente não faz previsão de preço nem de viagem. O PLD do horizonte de *lookahead* é fornecido como dado conhecido (§3.4.3) e a viagem entra como a probabilidade $p_{\text{dep}}(t)$ da Eq. (3.5), de modo que a política apenas antecipa decisões a partir desses sinais.

3.4.4 Espaço de Ações e Dinâmica de Atualização

Ação normalizada

A ação do agente é uma potência normalizada no intervalo $\mathcal{A} = [-1, +1]$ por veículo, mapeada para potência física em Watts pelo limite operacional da química:

$$P_t^{(i)} = a_t^{(i)} \cdot P_{\text{max}}^{(i)}, \quad P_{\text{max}}^{(i)} = C_{\text{nominal}} \cdot \text{capacity}^{(i)}, \quad a_t^{(i)} \in [-1, +1] \quad (3.8)$$

Com convenção de sinal: $P > 0$ indica carga (rede $\rightarrow$ veículo) e $P < 0$ indica descarga V2G (veículo $\rightarrow$ rede). O fator $C_{\text{nominal}} = 1C$ é a taxa máxima dessa simulação, e os limites por química são, portanto, 80 kW (NMC-811, 80 kWh) e 50 kW (LFP, 50 kWh), considera-se aqui que os carregadores possam atingir tais potências. Essa escolha, em vez de um limite global pelo cabo do carregador, alinha o espaço de ação acessível ao agente com o intervalo completo $[-1C, +1C]$ da *grid* da *lookup table* (Tabela 2), evitando saturação artificial em uma fração do espaço.

Dinâmica de SoC com convenção P_{CC} pura

A potência comandada pelo agente é interpretada como potência CC do lado da bateria, consistente com a definição de taxa C , $C_{\text{rate}} = P/C_{\text{bat}}$. A eficiência do conversor bidirecional CA-CC, $\eta = 0,95$, é aplicada apenas no lado da rede, no cálculo de receita e custo de mercado (§3.4.5). Essa convenção evita a dupla contagem da eficiência, pois aplicá-la também na transferência interna à bateria produziria uma eficiência *round-trip* simulada de $\eta^4 \approx 0,815$, em vez do $\eta^2 \approx 0,9$ esperado, o que distorceria a análise econômica do V2G em cerca de 10%. A bateria recebe ou descarrega, portanto, exatamente $P_t \cdot \Delta t$ em CC:

$$E_{t+1} = E_t + P_t \cdot \Delta t, \quad \text{SoC}_{t+1} = \text{clip}\left(\frac{E_{t+1}}{C_{\text{bat}}}; \text{SoC}^{\text{min}}, \text{SoC}^{\text{max}}\right) \quad (3.9)$$

Onde $\text{clip}(x; a, b) = \min(\max(x, a), b)$ implementa o limite operacional da química, com $[\text{SoC}^{\text{min}}, \text{SoC}^{\text{max}}] = [0,20; 0,90]$ para NMC e $[0,10; 0,95]$ para LFP.

Quando o SoC_{t+1} atinge os limites operacionais, parte da energia comandada pela ação não é absorvida ou entregue pela bateria. O ambiente registra, portanto, uma potência efetiva calculada a partir da variação real de SoC após o *clip*, $P_t^{\text{eff}} = (\text{SoC}_{t+1} - \text{SoC}_t) C_{\text{bat}} / \Delta t$, que coincide com P_t quando não há saturação e é menor em módulo quando o SoC satura, e utiliza P_t^{eff} em vez de P_t no cálculo da receita de mercado (§3.4.5). Essa mesma P_t^{eff} também alimenta a consulta à *lookup table* de degradação, evitando que estados em *clip* de SoC paguem custo por uma corrente que não foi efetivamente injetada na bateria. A correção é fundamental para a integridade da função de recompensa, pois sem ela o agente poderia comandar descarga máxima em SoC^{min} , receber receita do mercado e, com a bateria em *clip*, não perder energia, configurando um *exploit* estrutural do reward.

Dinâmica de SoH e atualização do estado de saúde

Ao fim de cada passo, o *Estado de Saúde* da bateria observado pelo agente é incrementalmente atualizado a partir das taxas físicas devolvidas pela *lookup table* (Tabela 3). Adota-se a Eq. 15 de Niazi *et al.* (2026), que decompõe o SoH como combinação convexa de dois canais, a perda de capacidade e o crescimento fracional de resistência (SEI). O canal de capacidade agrega as três vias que reduzem a capacidade utilizável da célula, isto é, o crescimento da SEI, o *plating* e a perda de material ativo, em coerência com a monetização da Eq. (3.1), na qual a perda de material ativo recebe o mesmo coeficiente da perda de capacidade ($\alpha_{\text{AML}} = \alpha_Q$), pois ambas retiram capacidade da célula:

$$\text{SoH}_{t+1} = \text{SoH}_t - \gamma_{\text{SoH}} \cdot (|\dot{Q}_{\text{SEI}}| + |\dot{Q}_{\text{plat}}| + |\dot{f}_{\text{AML}}|) \cdot \Delta t - (1 - \gamma_{\text{SoH}}) \cdot |\dot{R}_{\text{SEI}}| \cdot \Delta t \quad (3.10)$$

Onde:

- $\text{SoH}_t \in [\text{SoH}^{\text{min}}, 1]$ é o estado de saúde no início do passo, expresso como fração da capacidade nominal;
- $\gamma_{\text{SoH}} = 0,80$ é o peso atribuído à perda de capacidade e $1 - \gamma_{\text{SoH}} = 0,20$ o do canal de resistência; a forma composta segue Niazi *et al.* (2026), que define a Eq. 15 sem fixar o valor do peso, e adotou-se 0,80 porque o fim de vida automotivo é definido primariamente pela capacidade (o limiar de 80% é um critério de capacidade), de modo que a perda de capacidade é o critério primário de fim de vida e o aumento de resistência atua como secundário, sobretudo em *packs* de grande capacidade como os de veículos elétricos (ETXANDI-SANTOLAYA *et al.*, 2024) (§3.3.5);
- $\dot{Q}_{\text{SEI}}$, $\dot{Q}_{\text{plat}}$, $\dot{f}_{\text{AML}}$ e $\dot{R}_{\text{SEI}}$ são as taxas instantâneas (fração/h) interpoladas na *lookup table* para a quádrupla $(\text{SoC}_t, T_t, \text{SoH}_t, C_t)$;
- os módulos garantem que o SoH é monotonicamente decrescente, independentemente do sinal das taxas em saídas pontuais do interpolador;

- $\text{SoH}^{\min} = 0,50$ é o piso operacional no qual a Eq. (3.10) é clipada para evitar extrapolação fora do *grid* da *lookup table* (Tabela 2). Atingir esse piso configura fim de vida da célula e é registrado no *log* da simulação.

O símbolo γ_{SoH} é distinto do fator de desconto γ do PPO (Eq. (3.3)); ambos são tradicionais na literatura, e a notação preserva a correspondência direta com o artigo-base. O SoH atualizado por (3.10) entra no vetor de observação do passo seguinte (§3.4.3) e em todas as consultas futuras à *lookup table*, fechando o laço de *feedback* entre degradação acumulada e decisão do agente.

Consumo de energia nas viagens (estocástico)

No instante de partida $t = t^{\text{dep}}$, a bateria sofre um decremento abrupto:

$$\text{SoC}_{t^{\text{dep}}+} = \max\left(\text{SoC}_{t^{\text{dep}}-} - \frac{E_{\text{trip}}}{C_{\text{bat}}}, \text{SoC}^{\min}\right) \quad (3.11)$$

Durante a viagem, o indicador c_t vai a zero e nenhuma ação do agente é aplicada, dado que o veículo está desconectado da rede. O *clip* em $\text{SoC}^{\min}$ (em vez de 0) representa uma recarga emergencial implícita: se o agente comandou V2G demais e a bateria não tem energia para cumprir a viagem, simula-se que o usuário recorreu a um carregador externo, e o custo dessa falha entra na penalidade P^{fail} (Eq. (3.17)).

3.4.5 Função de Recompensa

A função de recompensa é o núcleo da metodologia, e é justamente onde os dois ambientes divergem. O baseline usa a Eq. 19 de Niazi de forma direta, enquanto o estocástico estende essa formulação com duas penalidades adicionais (Eqs. (3.16) e (3.17)) que modelam o comportamento de um motorista individual sob incerteza.

Cenário Baseline, Eq. 19 de Niazi

No ambiente baseline, a recompensa é a soma sobre os dois veículos da Eq. 19 de Niazi *et al.* (2026):

$$R_t^B = w_{\text{profit}} \cdot \sum_{i=1}^2 R_t^{\text{profit},(i)} - w_{\text{deg}} \cdot \sum_{i=1}^2 \dot{R}_t^{\text{deg},(i)} \cdot \Delta t - \sum_{i=1}^2 P_t^{\text{thermal},(i)} \quad (3.12)$$

Onde $w_{\text{profit}} = w_{\text{deg}} = 1,0$ são os pesos da recompensa, e $\dot{R}_t^{\text{deg},(i)}$ é o custo horário de degradação obtido por interpolação na *lookup table* (Eq. (3.1), §3.3.5). O lucro de mercado por veículo segue a Eq. 20 de Niazi adaptada ao contexto brasileiro:

$$R_t^{\text{profit},(i)} = \eta_{\text{dis}} \cdot |P_t^{\text{eff},(i)}| \cdot \Delta t \cdot p_t^{\text{PLD}} \cdot \mathbb{1}[P^{\text{eff}} < 0] - \frac{P_t^{\text{eff},(i)} \cdot \Delta t}{\eta_{\text{ch}}} \cdot p_t^{\text{PLD}} \cdot \mathbb{1}[P^{\text{eff}} > 0] \quad (3.13)$$

Onde:

- $\eta_{\text{dis}} = \eta_{\text{ch}} = \eta = 0,95$ é a eficiência do conversor bidirecional, aplicada apenas no lado da rede (consistente com a convenção P_CC pura);
- p_t^{PLD} é o PLD horário corrente no submercado Sudeste/Centro-Oeste (R\$/kWh);
- $\mathbb{I}[\cdot]$ é a função indicadora;
- o primeiro termo é a receita de venda em descarga ($P^{\text{eff}} < 0$), em que a bateria entrega energia CC e a rede recebe η dessa energia, monetizada pelo PLD;
- o segundo termo é o custo de compra em carga ($P^{\text{eff}} > 0$), em que a bateria absorve energia CC, mas a rede precisa fornecer $1/\eta$ desse valor, custo igualmente monetizado pelo PLD.

A Eq. 20 original de Niazi inclui ainda termos de serviços ancilares de regulação de frequência (*regulation capacity* P^{cap} e *regulation mileage* P^{mil} do mercado PJM). Como o submercado brasileiro do PLD não remunera esses serviços sob a regulação vigente (§2.2), esses termos foram zerados ($P^{\text{cap}} = P^{\text{mil}} = 0$) na implementação. A estrutura é mantida no código, preparada para ativação futura caso o arcabouço regulatório brasileiro venha a contemplar mercados de capacidade e *mileage*.

Restrição rígida de partida: a Eq. 24 de Niazi *et al.* (2026) exige que o SoC no instante de partida seja maior ou igual a um valor mínimo $\text{SoC}_{\text{req}}^{\text{dep}}$:

$$\text{SoC}^{(i)}(t^{\text{dep}}) \geq \text{SoC}_{\text{req}}^{\text{dep}} \quad (3.14)$$

Esta monografia adota $\text{SoC}_{\text{req}}^{\text{dep}} = 0,60$ no ambiente baseline. O estado de carga de retorno de cada veículo é resultado do estado de carga com que o veículo partiu menos o consumo da viagem, cuja distância segue a Weibull da viagem 1 do perfil intenso e é convertida em energia pela eficiência de 0,15 kWh/km, exatamente como no ambiente estocástico. Assim o excedente carregado acima da exigência de partida é conservado através da viagem em vez de descartado, tornando o baseline consistente com o estocástico nesse aspecto e deixando a janela noturna livre para arbitragem do PLD.

A restrição da Eq. (3.14) é implementada como uma projeção de viabilidade. Em cada passo a ação do agente é aplicada normalmente, e o ambiente apenas verifica antes se ela mantém a partida viável, isto é, se o tempo de conexão restante, carregando a 0,5 C, ainda é suficiente para alcançar $\text{SoC}_{\text{req}}^{\text{dep}}$ no horário de saída. Enquanto essa folga existe a ação passa intacta, e o agente permanece livre para carregar ou vender energia. Apenas quando a ação escolhida deixaria a partida inalcançável o ambiente a sobrescreve, impondo a taxa C mínima que, distribuída pelo tempo que resta até a saída, recompõe o SoC exigido, taxa essa que nunca ultrapassa 0,5 C. Na prática ela fica abaixo desse teto, porque a barreira passa a agir assim que a folga a 0,5 C se esgota, ainda restando cerca de meia hora até a partida, de modo

que a carga necessária se espalha por essa janela. A política passiva adotada como referência carrega a 0,5 C até $\text{SoC}_{\text{req}}^{\text{dep}}$ e então permanece ociosa.

Cenário Estocástico, extensão da Eq. 19

No ambiente individual, o instante de partida é aleatório e desconhecido a priori pelo agente, e a Eq. (3.14) não pode ser implementada como projeção determinística. A solução adotada, que constitui contribuição original desta monografia, é estender a Eq. 19 de Niazi com duas penalidades *soft* que modelam o comportamento subjetivo do motorista:

$$R_t^S = \underbrace{w_{\text{profit}} \cdot R_t^{\text{profit}} - w_{\text{deg}} \cdot \dot{R}_t^{\text{deg}} \cdot \Delta t}_{\text{Eq. 19 de Niazi, um veículo}} - P_t^{\text{anx}} - P_t^{\text{fail}} - P_t^{\text{thermal}} \quad (3.15)$$

A primeira penalidade modela a ansiedade de autonomia de forma contínua, ativando-se sempre que o SoC está abaixo do limiar de segurança do perfil:

$$P_t^{\text{anx}} = \lambda_{\text{anx}} \cdot p_{\text{dep}}(t) \cdot \max(0, E^{\text{seg}} - E_t) \quad (3.16)$$

Onde:

- $\lambda_{\text{anx}} = 3,0$ é o coeficiente de ansiedade (em R\$/kWh equivalente), calibrado empiricamente após sucessivas rodadas e auditorias de política. O valor anterior de 0,5 tornava a penalidade pequena diante do custo marginal de ciclar, e o agente PPO racionalizava aceitar a penalidade em vez de preservar a margem de autonomia;
- $p_{\text{dep}}(t)$ vem da Eq. (3.5) e modula a intensidade da penalidade, isto é, quanto mais iminente a partida, maior a punição por ter SoC baixo;
- $E^{\text{seg}} = e_{\text{seg}}^{\text{SoC}} \cdot C_{\text{bat}}$ é a energia-alvo de segurança;
- o operador $\max(0, \cdot)$ desliga a penalidade quando o SoC está acima do limiar.

A segunda penalidade modela a falha crítica (a impossibilidade de cumprir a viagem por falta de energia) e é aplicada apenas no instante de partida:

$$P_t^{\text{fail}} = \lambda_{\text{fail}} \cdot \max(0, E_{\text{trip}} - E_{t^{\text{dep}}}^{\text{disp}}) \cdot \mathbb{I}[t = t^{\text{dep}}] \quad (3.17)$$

Onde:

- $\lambda_{\text{fail}} = 50$ é o coeficiente de falha (R\$/kWh de déficit), que pune o agente em proporção à energia que falta para cumprir a viagem no instante da partida;
- E_{trip} é a energia necessária para completar a viagem amostrada (distância $\times$ eficiência do veículo);

- $E_{t_{dep-}}^{disp} = \max(0, E_{t_{dep-}} - \text{SoC}^{\min} \cdot C_{bat})$ é a energia efetivamente disponível para a viagem imediatamente antes da partida, descontada a reserva inacessível pelo BMS abaixo de $\text{SoC}^{\min}$;
- o indicador garante que a penalidade só é aplicada uma única vez por viagem, no instante da partida.

Descontar a reserva $\text{SoC}^{\min} \cdot C_{bat}$ é necessário porque, sem esse desconto, uma partida com o SoC no limite mínimo produziria um déficit nulo no cálculo direto $E_{trip} - E_{t_{dep-}}$, ainda que o veículo não tenha energia útil para a viagem. Em rodadas anteriores, o agente PPO aproveitou essa falha de contagem e passou a manter o veículo parado em $\text{SoC} = \text{SoC}^{\min}$, e o desconto corrige isso ao reconhecer que a energia abaixo do limite mínimo fica indisponível para a viagem.

A razão $\lambda_{fail}/\lambda_{anx} \approx 17$ reflete que uma falha real é cerca de dez vezes mais grave que o desconforto de ansiedade. Os valores vêm de calibração empírica e podem ser refinados em trabalhos futuros. As duas penalidades têm papéis distintos e propositais, pois P^{anx} , proporcional à margem $E^{seg} - E_t$, molda a política de forma contínua para preservar a margem psicológica do motorista durante toda a janela de risco, enquanto P^{fail} , proporcional ao déficit $E_{trip} - E_{t_{dep-}}^{disp}$, age como um sinal forte no instante da partida, quando a viagem não pode ser cumprida por falta de energia útil. A margem de segurança E^{seg} entra apenas pela P^{anx} , sem repetir o critério de viabilidade física que já está na P^{fail} .

Penalidade térmica de reward shaping

A penalidade térmica P_t^{thermal} presente em ambas as recompensas (Eqs. (3.12) e (3.15)) replica diretamente o termo $\beta_T(T - T_{opt})^2$ da Eq. 19 de Niazi *et al.* (2026):

$$P_t^{\text{thermal}} = \beta_T \cdot (T_t - T_{opt})^2 \cdot \Delta t \quad (3.18)$$

Onde:

- $\beta_T = 0,01 \text{ R\$}/(\text{h} \cdot \text{K}^2)$ é o coeficiente da penalidade quadrática;
- $T_{opt} = 298,15 \text{ K}$ (25 °C) é a temperatura ótima de operação;
- T_t é a temperatura da célula no início do passo (em Kelvin).

O termo é *reward shaping* puro, isto é, uma penalidade artificial imposta ao agente para guiá-lo a evitar regimes térmicos adversos durante o treinamento. Diferentemente de $\dot{R}_t^{\text{deg}}$ (Eq. (3.1)), P_t^{thermal} não representa um custo monetário físico real, pois o efeito da temperatura sobre o envelhecimento da bateria já é capturado dentro do PyBaMM via dependência de Arrhenius das constantes cinéticas, e por isso $\dot{R}_t^{\text{deg}}$ contém implicitamente o impacto térmico físico. Somar $\beta_T(T - T_{opt})^2$ a $\dot{R}_t^{\text{deg}}$ implicaria dupla contagem do efeito térmico no custo total

reportado, motivo pelo qual a contabilidade econômica desta monografia (lucros, margens marginais V2G, decomposição por mecanismo) exclui o termo P^{thermal} acumulado, conforme detalhado em §4.3 do Cap. 4. A função aqui é exclusivamente pedagógica para o treinamento, orientando a política a operar próximo de T_{opt} além do que o impacto físico da temperatura sobre a degradação por si só conseguiria forçar.

3.5 ALGORITMO DE OTIMIZAÇÃO *PROXIMAL POLICY OPTIMIZATION*

3.5.1 Escolha do Algoritmo

A formulação do problema apresentada no §3.4 combina três características que limitam a escolha do algoritmo de otimização: (i) ações contínuas em $\mathcal{A} = [-1, +1]^{N_v}$, o que descarta algoritmos baseados puramente em Q -valores tabulares ou em DQN; (ii) recompensa multi-objetivo com componentes de magnitudes muito distintas (receita financeira em reais, custo de degradação em reais, penalidades em reais por kWh equivalente), o que exige estabilidade de atualização da política; (iii) estocasticidade composta sobre preço PLD, viagens e estado interno da bateria, que recomenda métodos *on-policy* para evitar viés de dados desatualizados.

O *Proximal Policy Optimization* (PPO), em arquitetura *Actor-Critic*, atende a esses três requisitos. É o mesmo algoritmo adotado por Niazi *et al.* (2026) e é o padrão na maior parte da literatura recente de RL aplicado a V2G (XIE *et al.*, 2025). A implementação utilizada é a do *stable-baselines3*, biblioteca Python amplamente validada e mantida sob revisão por pares.

3.5.2 Aplicação ao Problema V2G

A formulação matemática do PPO em arquitetura *Actor-Critic* foi apresentada no §2.3.7 do Cap. 2, com a função objetivo *clipped* dado pela Eq. (2.10) e a razão de política $r_t(\theta)$ Eq. (2.11). Esta seção registra como essa formulação genérica é instanciada no problema V2G desta monografia.

O retorno a ser maximizado é a soma descontada das recompensas r_t ao longo da trajetória, com r_t dado pela Eq. (3.12) no ambiente baseline e pela Eq. (3.15) no ambiente estocástico. A estimativa de vantagem $\hat{A}_t$ que entra no objetivo *clipped* é calculada pelo *Generalized Advantage Estimation* (GAE; Schulman *et al.* (2015)), formulado na Eq. (2.12) do Cap. 2, que combina estimativas em horizontes distintos por meio de um parâmetro λ_{GAE} próprio, distinto dos coeficientes λ_{anx} e λ_{fail} do reward. A rede crítica é treinada conjuntamente com a política, cada uma como uma rede MLP separada com a mesma arquitetura detalhada na Tabela 6.

A implementação adotada é a do *stable-baselines3*, escolha que se justifica pela validação extensiva da biblioteca em problemas de controle contínuo e pela manutenção ativa sob revisão por pares. Os hiperparâmetros do PPO são apresentados na Tabela 6 a seguir, em

paralelo com os valores adotados por Niazi *et al.* (2026), e a única divergência intencional, o fator de desconto γ , é justificada na sequência.

3.5.3 Hiperparâmetros

Os hiperparâmetros adotados estão na Tabela 6, ao lado dos correspondentes valores do artigo-base.

Tabela 6 – Hiperparâmetros do PPO adotados neste trabalho.

Hiperparâmetro	Valor adotado	Niazi (2026)	Observação
Taxa de aprendizado	3×10^{-4}	3×10^{-4}	Idêntico
Fator de desconto γ	0,997	0,99	Ajustado, ver discussão abaixo
Parâmetro GAE λ	0,95	0,95	Idêntico
Clip range ϵ	0,2	0,2	Idêntico
Coeficiente de entropia	0,01	0,01	Idêntico
Coeficiente da <i>value function</i>	0,5	0,5	Idêntico
Norma máxima do gradiente	0,5	0,5	Idêntico
N_{steps} (rollout)	2.048	2.048	Idêntico
Batch size	64	64	Idêntico
N_{epochs} (update)	10	10	Idêntico
Arquitetura ator/crítico	MLP [256, 256], <i>tanh</i>	MLP [256, 256]	Idêntico
Total timesteps	1×10^6 por cenário	n/a	Não explícito em Niazi
Seed	42	n/a	Fixo para reprodutibilidade

Com exceção de γ , os valores da Tabela 6 são os padrões do PPO usados por Niazi *et al.* (2026), que seguem a formulação de Schulman *et al.* (2017) e os valores de omissão do *stable-baselines3*, e não um ajuste específico deste trabalho. O único parâmetro alterado é o fator de desconto γ , discutido a seguir.

O valor adotado $\gamma = 0,997$ difere do $\gamma = 0,99$ utilizado por Niazi por uma razão estrutural relacionada ao passo de simulação. Como $\Delta t = 5$ min neste trabalho, o horizonte efetivo do desconto exponencial, ou seja, o tempo, em passos ou em horas, em que o peso γ^h cai a $1/e$, é aproximadamente $h_{\text{eff}} \approx 1/(1 - \gamma)$, conforme a Tabela 7.

Tabela 7 – Horizonte efetivo do fator de desconto em função do passo $\Delta t = 5$ min.

γ	h_{eff} (passos)	h_{eff} (horas)	h_{eff} (dias)
0,99	100	8,3	0,35
0,997	333	27,8	1,16
0,999	1.000	83	3,47

A arbitragem do PLD horário opera em ciclo diário (24 h), pois é fundamentalmente um padrão dia-noite, com preços baixos na madrugada e picos no fim da tarde. Para que o

agente reconheça esse padrão e o explore (carregar de madrugada, descarregar à tarde), o horizonte efetivo precisa cobrir pelo menos um ciclo completo. Com $\gamma = 0,99$, o horizonte efetivo é de apenas 8,3 h, e o agente “esquece” o pico vespertino ao planejar a carga matutina. Por outro lado, $\gamma = 0,999$ (~ 83 h efetivos) eleva a variância do estimador de valor a ponto de comprometer a estabilidade do treinamento em janelas de 10^6 passos.

Critério de parada e modelo final

O treinamento de cada cenário usa um orçamento fixo de 10^6 passos, sem parada antecipada por convergência. O modelo avaliado na validação e na simulação final é o obtido ao fim desses 10^6 passos, com os pesos congelados.

3.5.4 Matriz de Cenários

São treinados sete cenários independentes com os mesmos hiperparâmetros da Tabela 6:

1. **Um cenário baseline:** ambiente com microfrota de dois veículos, um por química (NMC-811 e LFP).
2. **Seis cenários estocásticos:** ambiente V2GStochasticEnv cobrindo o produto 3 perfis $\times$ 2 químicas = 6 combinações (perfil Conservador, Flexível e Intenso, cada um com NMC-811 ou LFP).

A política utilizada na validação e na simulação final é determinística, ou seja, a ação escolhida é a média da distribuição gerada pela rede do ator, em vez de uma amostra. Essa escolha garante reprodutibilidade dos resultados financeiros e remove ruído de amostragem que confundiria a interpretação dos números reportados no Cap. 4.

Para que o ganho marginal do agente PPO sobre uma operação sem inteligência possa ser quantificado, cada cenário é avaliado em paralelo a uma política passiva de referência π^{pas} , que modela o comportamento típico de um motorista doméstico em 2026, isto é, chegar em casa, plugar o carregador e iniciar a carga imediatamente até atingir o SoC alvo, sem qualquer consideração sobre o preço da energia ou o tempo restante até a próxima viagem. No ambiente estocástico,

$$\pi^{\text{pas}}(s_t) = \begin{cases} +0,5 & \text{se } c_t = 1 \text{ e } \text{SoC}_t < \text{SoC}^{\text{alvo}}, \\ 0 & \text{caso contrário,} \end{cases} \quad (3.19)$$

onde $c_t \in \{0, 1\}$ é o indicador de conexão do veículo, $\text{SoC}^{\text{alvo}} = 0,80$ é o nível alvo de carga e a ação $+0,5$ corresponde a 0,5 C nominal, valor que mantém a célula fora do regime de *plating* ativo da *lookup* ($C \geq 0,7$ em SoC alto). O passivo nunca descarrega para a rede e não participa do mercado V2G. No ambiente *baseline* a política passiva é ainda mais simples, com $\pi^{\text{pas}} = 0$ em todo passo, e o cumprimento da Eq. (3.14) fica a cargo da projeção determinística

embutida no ambiente, conforme Niazi *et al.* (2026). O intuito dessa política é de reproduzir o cenário que o usuário não tem coordenação com os sinais de preço (PLD), e isolar o efeito da inteligência de operação na comparação econômica, pois qualquer diferença entre PPO e passivo é, por construção, atribuível à decisão de *quando* carregar e *quando* descarregar, e não a outras escolhas de arquitetura.

Além da política passiva, cada cenário é também confrontado com uma política heurística de referência, baseada em regras fixas de preço e sem qualquer aprendizado. Se o passivo representa o piso de comportamento, um motorista que apenas carrega, a heurística representa uma operação de arbitragem simples e existe para responder a uma pergunta direta de validação: o agente PPO de fato aprende uma estratégia melhor do que uma regra ingênua de carregar quando o PLD está barato e descarregar quando está caro? Caso a heurística superasse o PPO, isso seria indício de que o agente havia ficado preso em um ótimo local. A heurística carrega a 0,2 C quando o veículo está conectado (a fim de minimizar a degradação), o SoC está abaixo de um alvo e o PLD está abaixo de um limiar inferior; descarrega a 0,5 C quando o SoC está acima de um piso de segurança e o PLD está acima de um limiar superior; e permanece ociosa no restante do tempo:

$$\pi^{\text{heu}}(s_t) = \begin{cases} +0,2 & \text{se } c_t = 1, \text{ SoC}_t < \text{SoC}^{\text{alvo}} \text{ e } p_t^{\text{PLD}} < p_{\text{low}}, \\ -0,5 & \text{se } c_t = 1, \text{ SoC}_t > \text{SoC}_{\text{min}}^{\text{dis}} \text{ e } p_t^{\text{PLD}} > p_{\text{high}}, \\ 0 & \text{caso contrário,} \end{cases} \quad (3.20)$$

com alvo de carga $\text{SoC}^{\text{alvo}} = 0,85$, piso de descarga $\text{SoC}_{\text{min}}^{\text{dis}} = 0,50$ e limiares de preço $p_{\text{low}} = 0,25$ e $p_{\text{high}} = 0,35$ R\$/kWh. Esses limiares e faixas de SoC não foram otimizados; foram definidos por inspeção da distribuição do PLD do período e do comportamento observado nas simulações, apenas o suficiente para separar de forma limpa as horas baratas das caras. A intenção não é que a heurística seja ótima, e sim que seja uma referência honesta e difícil de superar por acaso, contra a qual o ganho real do agente PPO possa ser medido.

3.5.5 Divisão Temporal Treino / Validação

Para garantir avaliação genuinamente *out-of-sample*, o *dataset* de PLD horário é particionado em duas janelas disjuntas no tempo:

- **Treino:** PLD horário do submercado Sudeste/Centro-Oeste de 01/01/2025 a 31/12/2025 (365 dias). O dia inicial do episódio é amostrado uniformemente em cada reset.
- **Validação Monte Carlo:** PLD do mesmo submercado de 01/01/2026 a 30/03/2026 (89 dias, primeiro trimestre de 2026), em 30 episódios independentes com dia inicial aleatório em cada um.
- **Simulação final:** mesma janela 2026-Q1 (89 dias), em um único episódio determinístico começando em 01/01/2026, repetido para múltiplas *seeds* de inicialização interna do

ambiente. As políticas Passiva e Heurística são aplicadas nas mesmas *seeds*, e neste momento, em paralelo à simulação do agente PPO treinado.

Nenhum passo do período de validação ou da simulação final é exposto durante o treinamento. Isso garante que as métricas reportadas no Cap. 4 refletem capacidade de generalização do agente, e não memorização da série temporal de treino.

3.6 DADOS DE ENTRADA E CENÁRIOS DE VALIDAÇÃO

3.6.1 PLD Horário do Submercado Sudeste/Centro-Oeste

A série temporal exógena consumida pelo simulador é o Preço de Liquidação das Diferenças (PLD) horário do submercado Sudeste/Centro-Oeste (SE/CO), publicado pela Câmara de Comercialização de Energia Elétrica (Câmara de Comercialização de Energia Elétrica, 2024b). A escolha pelo SE/CO segue três motivações práticas:

- é o submercado de maior carga e maior frota de veículos elétricos do Brasil, com concentração expressiva da malha de eletropostos (§2.2);
- apresenta maior volatilidade intradiária em razão do mix de geração hidrotérmica e da forte penetração de geração distribuída fotovoltaica, o que cria *spread* economicamente significativo entre vales e picos;
- é o submercado com maior representatividade na literatura nacional recente sobre arbitragem de PLD, o que facilita comparações futuras.

O *dataset* utilizado é o arquivo `pld_brasileiro_horario.csv`, com resolução horária (uma linha por hora, 8.760 linhas por ano), cobrindo o período de 1º de janeiro de 2025 a 30 de março de 2026 (15 meses de dados consecutivos). A primeira janela de 12 meses (todo o ano de 2025) é utilizada para treinamento; a janela subsequente de 89 dias (primeiro trimestre de 2026) é utilizada para validação e simulação final, conforme detalhado no §3.5.5.

Pré-processamento

O preço bruto do PLD vem em R\$/MWh; o ambiente converte para R\$/kWh dividindo por 1.000 a cada consulta. Para alimentar a rede neural do agente (Eq. (3.6)), o PLD é padronizado por *z-score*:

$$\tilde{p}_t^{\text{PLD}} = \frac{p_t^{\text{PLD}} - \mu_{\text{PLD}}}{\sigma_{\text{PLD}} + 10^{-8}} \quad (3.21)$$

onde μ_{PLD} e σ_{PLD} são a média e o desvio-padrão do PLD calculados sobre a janela de preços efetivamente carregada em cada fase, isto é, os 365 dias de 2025 no treino e a janela de 2026 na validação e na simulação final. A normalização é, portanto, local a cada janela,

e não uma estatística fixa transportada do treino; como as distribuições de PLD de 2025 e de 2026 são próximas, a diferença entre as duas é pequena. A avaliação permanece fora da amostra no sentido essencial de que nenhum passo do período de 2026 é exposto ao agente durante o treino. O termo 10^{-8} é apenas estabilização numérica contra divisão por zero.

Resolução temporal $\times$ passo do MDP

O PLD é publicado em resolução horária, ao passo que o passo do MDP é de 5 minutos. Para fechar essa diferença de granularidade, o preço a ser usado em um passo t é interpolado por valor constante dentro da hora corrente, de modo que todos os 12 passos de simulação de uma mesma hora compartilham o mesmo p_t^{PLD} . Essa é a representação fiel do que ocorre no mercado físico, dado que o PLD é uma função em degraus por hora, e não uma função contínua, o que dispensa qualquer interpolação suave.

3.6.2 Procedimento de Simulação Final

Os resultados reportados no Cap. 4 são obtidos por meio de uma simulação final determinística de 89 dias consecutivos, executada após o treinamento dos sete cenários listados no §3.5.4. As condições da simulação final são:

- **Janela temporal fixa:** 01/01/2026 a 30/03/2026 (89 dias), começando sempre no dia $d_0 = 01/01/2026$, sem amostragem aleatória do dia inicial;
- **Política determinística:** ação igual à média da distribuição da rede do ator, sem ruído de amostragem;
- **Cenários:** os mesmos sete cenários treinados (1 baseline + 6 estocásticos), executados independentemente;
- **Seeds:** a simulação final de cada cenário é repetida com dez *seeds* distintas na inicialização do ambiente, de modo que os perfis estocásticos amostram viagens diferentes a cada repetição e a dispersão dos resultados reportados possa ser quantificada.

Uma *seed* é o valor que inicializa o gerador de números aleatórios do ambiente, e variá-la faz cada perfil estocástico sortear viagens diferentes. Adotaram-se dez *seeds* por cenário, número suficiente para observar a variabilidade dos resultados sem tornar a simulação lenta demais em computador pessoal.

A escolha de uma janela determinística contínua de 89 dias, em vez de muitos episódios curtos de 30 dias com início aleatório, é importante para a interpretabilidade dos resultados financeiros. A métrica primária a ser reportada no Cap. 4 é o lucro líquido cumulativo em reais ao longo de um trimestre representativo do regime hidrológico de início de ano (estação úmida no SE/CO), com PLDs tipicamente baixos no início do período e crescentes ao longo

do trimestre. Episódios curtos com janela aleatória diluem essa estrutura temporal e dificultam a comparação entre cenários.

Para cada cenário, o simulador registra, em resolução de 5 minutos:

- a ação a_t , a potência efetiva P_t^{eff} , o SoC_t , o SoH_t e a temperatura T_t ;
- a recompensa r_t , decomposta em receita de mercado, custo de degradação e penalidades P^{anx} e P^{fail} ;
- as taxas instantâneas $\dot{Q}_{\text{SEI}}$, $\dot{R}_{\text{SEI}}$, $\dot{Q}_{\text{plat}}$, $\dot{f}_{\text{AML}}$ obtidas da *lookup table*;
- o registro de cada evento de viagem (instante, distância, energia consumida, indicador de falha).

3.6.3 Métricas Reportadas

As métricas principais reportadas no Cap. 4, para cada cenário, são:

- Lucro líquido cumulativo no trimestre (R\$/Q1), composto pela receita de venda de energia ao mercado, descontados o custo de compra e o custo de degradação. As penalidades P^{anx} e P^{fail} são termos de *reward shaping* usados apenas no treino e não entram no lucro líquido reportado, sendo a qualidade de serviço ao usuário avaliada à parte pela taxa de falha;
- Receita V2G antes da degradação, isto é, a parcela que vem da arbitragem de PLD pura, isolada do custo físico da bateria, conforme a Eq. (3.13);
- Custo de degradação acumulado (R\$/Q1), reportado tanto em valor total quanto decomposto por mecanismo físico (SEI-Q, SEI-R, plating, LAM e térmica) para identificar qual mecanismo domina em cada química;
- SoH final ao término dos 89 dias e vida útil projetada em anos até atingir o limiar de fim de vida automotivo ($\text{SoH} = 0,80$), extrapolada linearmente a partir do ΔSoH observado no período;
- Energia transacionada com a rede (kWh), separada entre energia comprada e energia vendida, que serve de base para o cálculo de *spread* e ciclos efetivos;
- Taxa de falha (apenas no estocástico), definida como a fração de viagens em que a energia útil disponível na partida ficou abaixo da energia requerida pelo trajeto.

O lucro líquido de cada cenário é a receita de venda de energia à rede, descontados o custo de compra e o custo de degradação da bateria:

$$L_{\text{liq}} = R_{\text{venda}} - C_{\text{compra}} - C_{\text{deg}} \quad (3.22)$$

Quando o lucro líquido é positivo, há lucro, e quando é negativo, há prejuízo.

3.6.4 Uso de ferramentas de inteligência artificial

A elaboração desta monografia contou com o apoio de modelos de linguagem de grande porte em tarefas auxiliares, sempre sob revisão e validação do autor. O uso concentrou-se na escrita e depuração do código Python do simulador, na geração dos *scripts* que produzem as figuras e tabelas, na formatação de trechos em LaTeX, no apoio à pesquisa bibliográfica e no estudo dos fundamentos de aprendizado por reforço e do funcionamento do treinamento do PPO.

Em parte do trabalho adotou-se também a revisão cruzada entre modelos, em que a análise de um modelo foi submetida à crítica de um segundo e, em caso de divergência persistente, a um terceiro, reduzindo o risco de erro associado a um único modelo. A Tabela 8 resume as ferramentas e suas atribuições.

Tabela 8 – Ferramentas de inteligência artificial utilizadas e suas atribuições.

Modelo	Desenvolvedor	Atribuição principal
Claude Opus 4	Anthropic	Código, figuras e tabelas, LaTeX, apoio bibliográfico e ao estudo de aprendizado por reforço
GPT-5.5 (Codex)	OpenAI	Revisão crítica de código
Gemini 3 Pro	Google	Desempate na revisão cruzada

4 RESULTADOS E DISCUSSÃO

Este capítulo percorrerá os resultados e discussões dos sete cenários treinados, começando pela caracterização da oportunidade econômica que cada perfil de mobilidade oferece, seguindo pela verificação de que o agente convergiu e generaliza para dados não vistos no treino, e culminando na discussão do desempenho econômico do V2G frente ao carregamento passivo, na qual se explica, química a química e perfil a perfil, o comportamento que o agente aprendeu e a razão de tê-lo aprendido. Não teremos uma seção de resultados dedicada a *lookup table*, em vez disso, serão compartilhados aqui de forma introdutória os principais resultados dela, que valem a pena serem destacados, pois é o ponto de partida para a fase online conforme visto na figura 5.

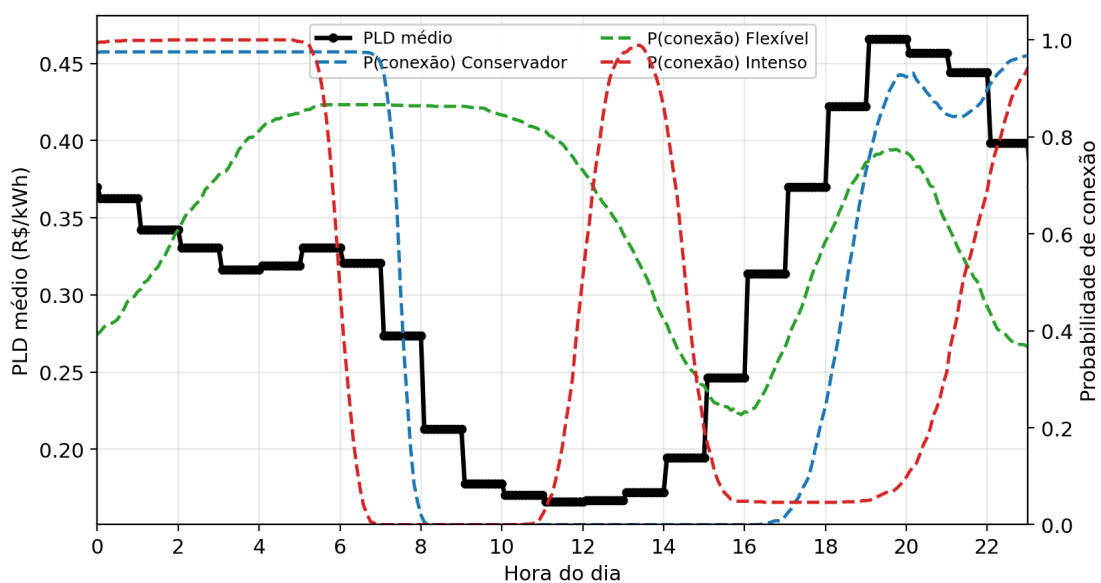
A *lookup* construída na fase offline revela uma separação química marcante entre as duas baterias. No NMC-811, a mecânica de partícula está ativa em 90,7% dos pontos do *grid*, com taxa máxima de perda de material ativo de $2,5 \times 10^{-4}$ frac/h, e exibe assimetria expressiva entre carga e descarga, uma vez que em $c\text{-rate} = +1$ a taxa é vinte e três vezes maior do que em $c\text{-rate} = -1$, reflexo direto da contração do cátodo durante a saída de lítio. O LFP, em contraste, apresenta $\dot{f}_{\text{LAM}}$ nulo em toda a grade, comportamento que reproduz a estabilidade estrutural da olivina e justifica a omissão deliberada da mecânica de partícula para esta química, conforme discutido na seção §3.2. O *plating* de lítio metálico, por fim, se distingue entre as duas químicas quanto à janela operacional segura, no NMC-811 ele se ativa já em temperatura ambiente, a partir de taxas de carga em torno de 0,7C a 25 °C, e seu limiar desce conforme a célula esfria, alcançando 0,5C a 15 °C e 0,3C a 0 °C. No LFP, em contraste, não há *plating* algum a 25 °C em toda a faixa operacional, e o fenômeno só aparece em frio, restrito a 1C a 15 °C e a partir de 0,5C a 0 °C. Essa diferença demarca para o NMC-811 uma zona de risco bem mais ampla, que a política online precisará aprender a evitar limitando a corrente de carga, sobretudo em dias frios, ao passo que no LFP a carga rápida permanece praticamente livre de *plating* nas temperaturas usuais do Sudeste brasileiro. A sensibilidade da SEI ao *cycling*, por sua vez, separa as químicas de forma mais sutil mas economicamente relevante, pois no NMC-811 o aumento de $c\text{-rate}$ de zero para ± 1 adiciona aproximadamente 15% à taxa $\dot{Q}_{\text{SEI}}$, enquanto no LFP o *cycling* pouco diferencia do regime calendário, com $\dot{Q}_{\text{SEI}}$ essencialmente constante na faixa $[-1, +1]$. Esse último resultado antecipa a intuição operacional de que um agente bem treinado tende a explorar mais ciclos no LFP, uma vez que a degradação incremental por operação quase não se separa do envelhecimento de calendário, diferentemente do NMC-811, no qual cada ciclo cobra um custo identificável.

4.1 JANELA DE OPORTUNIDADE DOS PERFIS

A viabilidade do V2G depende de dois fatores que o motorista não controla, sendo o primeiro a amplitude da variação do preço da energia ao longo do dia, que define o ganho possível por arbitragem, e o segundo o tempo em que o veículo permanece disponível para operar (desconsiderando o cenário *baseline*), que define a frequência com que esse ganho pode ser capturado. No submercado Sudeste durante o trimestre analisado, o PLD horário teve mediana de R\$ 0,314/kWh, com vale típico em torno de R\$ 0,18 (percentil 25) e pico em torno de R\$ 0,40 (percentil 75), o que estabelece um intervalo de cerca de R\$ 0,20/kWh como a margem bruta máxima que um único ciclo de arbitragem pode capturar antes de descontar as perdas de conversão e o custo de degradação. Como esse intervalo é modesto frente ao que mercados com serviços ancilares oferecem, já se antecipa que a viabilidade dependerá fortemente de quão barato cada química torna o ato de ciclar.

A Figura 6 sobrepõe esse preço médio horário à probabilidade de conexão de cada perfil, e revela uma tensão temporal importante, porque o preço atinge seu vale ao redor do meio-dia e tem momentos também mais baratos na madrugada, ainda que acima desse vale, ao passo que os picos se concentram no início da noite, exatamente o intervalo em que o perfil intenso costuma estar em trânsito e, portanto, indisponível para vender. O perfil flexível, por trabalhar em casa, é o que mais tempo permanece conectado, ficando disponível cerca de 63% do tempo, contra 52% do conservador e 47% do intenso. Interessante notar que no conservador a probabilidade de conexão cresce exatamente junto com o PLD, o que abre margem para chegar e vender energia.

Figura 6 – Preço médio horário do PLD no submercado Sudeste frente à probabilidade de conexão de cada perfil de mobilidade ao longo do dia.



Fonte: o autor, 2026.

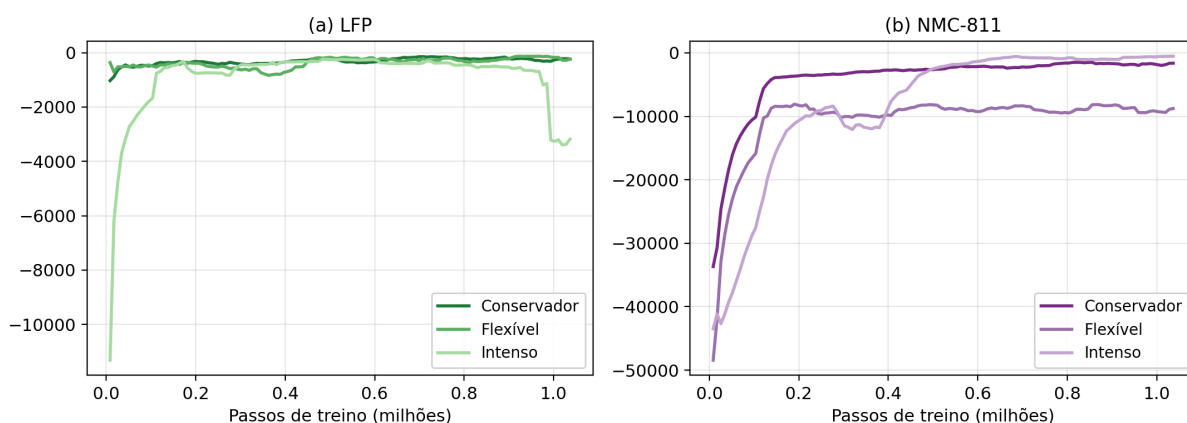
Há, contudo, uma diferença qualitativa entre os perfis que se mostrará decisiva na

análise econômica, e que não aparece no tempo total de conexão, mas sim na previsibilidade do horário de partida. O conservador e o intenso saem em horários regulares, com desvio-padrão de quinze a trinta minutos, enquanto o flexível parte em horários muito incertos, com desvio-padrão de duas horas na viagem principal, conforme a parametrização da Tabela 5. Como veremos, é essa imprevisibilidade da demanda, e não a falta de tempo conectado, que mais compromete a capacidade do agente de garantir energia para o usuário no perfil flexível.

4.2 CONVERGÊNCIA DO TREINAMENTO E GENERALIZAÇÃO

Cada um dos sete cenários foi treinado por um milhão de passos de interação, o equivalente a aproximadamente 116 episódios completos de trinta dias, e a Figura 7 acompanha a evolução da recompensa média por episódio ao longo desse treino, separando os cenários por química para tornar visível a diferença de dinâmica entre elas. As curvas da LFP partem de patamares já próximos do valor final e estabilizam cedo, porque o baixo custo de ciclar essa química faz com que erros cometidos durante a exploração custem pouco, ao passo que as curvas da NMC-811 começam muito abaixo e sobem de forma acentuada, com o cenário intenso evoluindo de uma recompensa inicial da ordem de -39.000 para cerca de -500 , uma melhora de duas ordens de grandeza que reflete o quanto o agente precisou aprender a evitar a degradação cara dessa química. Em todos os casos a curva alcança um platô antes do final do treino, indicando que o agente esgotou o ganho de aprendizado disponível sob a função de recompensa adotada.

Figura 7 – Recompensa média por episódio.



Fonte: o autor, 2026.

A oscilação visível ao final da curva do perfil intenso com LFP não representa perda de desempenho da política, mas sim a variância natural de um perfil que percorre quase 200 km por dia, no qual alguns poucos episódios sorteados sobre janelas de preço desfavoráveis bastam para puxar a média móvel para baixo, como confirma o aumento do desvio-padrão entre episódios nesse trecho. Como o agente é avaliado pelo modelo final, com os pesos congelados, e não

pela média móvel do treino, essa variância não compromete o resultado, conforme atestam o lucro e a ausência de falhas desse cenário na avaliação.

Convergência durante o treino, no entanto, não garante que o agente generalize, e como o treinamento foi conduzido sobre a série de PLD de 2025, é necessário descartar a hipótese de que a política tenha apenas memorizado aquela sequência específica de preços. Para isso, a mesma política treinada foi avaliada em duas séries de PLD distintas, a do próprio ano de treino (2025) e a do ano de teste (2026), em ambos os casos por trinta episódios de início aleatório, de modo que um desempenho muito superior dentro do ano de treino denunciaria sobreajuste. A Tabela 9 mostra que isso não ocorre, já que os lucros líquidos permanecem na mesma ordem de grandeza nas duas séries, e em vários cenários o desempenho sobre o ano de teste é até superior, o que indica que o agente aprendeu a estrutura recorrente do PLD, ou seja, a alternância diária entre vales e picos, e não os valores particulares do ano de treino.

Tabela 9 – Lucro líquido médio (R\$ por episódio de trinta dias) do agente avaliado sobre o PLD do ano de treino (2025) e sobre o PLD de um ano distinto (2026).

Cenário	PLD 2025 (treino)	PLD 2026 (distinto)	Diferença relativa
Baseline (frota)	−380,4	−480,3	−26,3%
NMC-811	−272,1	−316,4	−16,3%
LFP	−108,4	−163,9	−51,2%
Conservador NMC-811	−170,9	−151,2	+11,5%
Flexível NMC-811	−253,9	−281,8	−11,0%
Intenso NMC-811	−387,6	−337,2	+13,0%
Conservador LFP	−15,7	+37,4	comparável (valores pequenos)
Flexível LFP	+11,1	+125,2	comparável (valores pequenos)
Intenso LFP	−222,4	−308,0	−38,5%

4.3 DESEMPENHO ECONÔMICO E COMPORTAMENTO APRENDIDO

A pergunta central do objetivo de comparação com a linha de base, enunciado na Seção 1.1, é se a operação bidirecional inteligente compensa frente ao carregamento passivo, e a resposta depende fortemente da combinação entre química e perfil, encontrando explicação no balanço entre o ganho de arbitragem e o custo de degradação que cada decisão acarreta. Para responder a essa pergunta, cada cenário foi avaliado sob três políticas, sendo a primeira o carregamento passivo, que representa o usuário comum que apenas carrega ao chegar em casa e nunca devolve energia à rede, a segunda o agente PPO treinado, e a terceira uma política heurística de regras fixas, sem aprendizado, que carrega lentamente quando o preço está baixo e descarrega quando está alto, e que serve de referência para distinguir aquilo que o agente realmente aprendeu de mais sofisticado daquilo que regras simples já entregariam, confronto retomado em detalhe na Seção 4.3.5. A Tabela 10 reúne o resultado econômico das três políticas, acompanhado da frequência de falhas de autonomia do agente.

Tabela 10 – Lucro líquido trimestral (R\$) das três políticas e qualidade de serviço do PPO.

Cenário	Passivo	PPO	Heurística	Falhas/dia	% partidas
Baseline (frota)	−1.760	−1.420	−1.452	~ 0	0%
NMC-811	−1.074	−928	−909	~ 0	0%
LFP	−686	−492	−546	~ 0	0%
Conservador NMC-811	−565	−447	−441	0,34	30%
Flexível NMC-811	−729	−769	−370	1,38	100%
Intenso NMC-811	−1.458	−1.011	−517	0,01	1%
Conservador LFP	−246	+120	−153	0,11	10%
Flexível LFP	−331	+315	−75	0,12	9%
Intenso LFP	−971	−901	−263	0,00	0%

O lucro líquido reportado corresponde à receita de venda descontada do custo de compra de energia e do custo de degradação, conforme a Eq. (3.22), e um valor negativo representa prejuízo. Esse resultado não inclui o custo de o motorista ficar sem energia, razão pela qual a tabela traz, ao lado do lucro, a frequência de falhas de autonomia, que contabiliza quantas vezes o veículo partiu sem energia suficiente acima da reserva do sistema de gerenciamento para completar a viagem planejada, obrigando o usuário a recorrer a recarga externa. Essa distinção é necessária porque um cenário pode exibir lucro competitivo e, ainda assim, estar deixando o usuário a pé com frequência inaceitável, de modo que só consideramos plenamente viável o cenário que combina lucro superior ao passivo com baixa taxa de falha.

A dispersão dos resultados entre as dez *seeds* é pequena na maioria dos cenários, com desvio-padrão do lucro líquido em torno de R\$ 30, mas cresce de forma marcante nos dois cenários flexíveis, alcançando R\$ 174 no flexível com LFP e R\$ 214 no flexível com NMC, o que não é um ruído numérico, mas a própria imprevisibilidade da partida desse perfil propagando-se até o resultado financeiro, de modo que a incerteza do horário de viagem, discutida na Seção 4.1, reaparece aqui como incerteza econômica. A operação física por trás desses lucros aparece nas Tabelas 11 e 12, que reportam, para o agente e para o passivo em todos os sete cenários, a energia comprada e vendida, o custo de compra e a receita de venda correspondentes, a intensidade de ciclagem e o custo de degradação.

A leitura conjunta das duas tabelas revela o que distingue cada modo, porque o passivo limita-se a comprar a energia necessária para a carga e cicla muito pouco, de 0,02 a 0,31 vez por dia, ao passo que o agente movimenta volumes muito maiores e devolve boa parte deles à rede. Para isolar o quanto a operação bidirecional acrescenta sobre o passivo, a Tabela 13 decompõe a diferença observada entre os dois modos em duas parcelas, o ganho de mercado, definido como a melhora no resultado de mercado (receita de venda menos custo de compra) frente ao passivo, e a degradação adicional que o agente impõe à bateria, cuja subtração fornece a economia marginal do V2G. Convém tratar essas parcelas como diferenças observadas entre os modos, e não como relação de causa isolada, já que o agente altera compra, venda, estado

Tabela 11 – Operação do agente PPO por cenário no trimestre.

Cenário	Compra (kWh)	Venda (kWh)	Custo compra (R\$)	Receita (R\$)	Ciclos/ dia	Degrad. (R\$)
Baseline (frota)	5.297	2.570	1.714	1.086	0,196	791
NMC-811	2.974	1.591	961	674	0,196	642
LFP	2.323	980	753	411	0,196	150
Conservador NMC-811	780	546	242	259	0,093	464
Flexível NMC-811	121	114	74	101	0,017	796
Intenso NMC-811	5.952	2.864	1.638	1.284	0,619	658
Conservador LFP	3.300	2.768	986	1.251	0,682	145
Flexível LFP	3.753	3.028	752	1.243	0,762	176
Intenso LFP	3.394	566	1.079	309	0,445	131

Tabela 12 – Operação do carregamento passivo por cenário no trimestre.

Cenário	Compra (kWh)	Venda (kWh)	Custo compra (R\$)	Receita (R\$)	Ciclos/ dia	Degrad. (R\$)
Baseline (frota)	2.512	0	1.059	0	0,063	701
NMC-811	1.247	0	523	0	0,063	552
LFP	1.265	0	536	0	0,063	150
Conservador NMC-811	274	0	105	0	0,019	459
Flexível NMC-811	468	0	159	0	0,033	570
Intenso NMC-811	2.778	0	846	0	0,195	612
Conservador LFP	258	0	102	0	0,029	144
Flexível LFP	451	0	155	0	0,051	176
Intenso LFP	2.760	0	840	0	0,310	131

de carga e degradação ao mesmo tempo.

$$\text{Ganho de mercado} = (R_{\text{venda}}^{\text{PPO}} - C_{\text{compra}}^{\text{PPO}}) - (R_{\text{venda}}^{\text{pas}} - C_{\text{compra}}^{\text{pas}}), \quad (4.1)$$

Nota-se que nos três cenários com LFP o agente cicla muito mais que o passivo, chegando a ciclar quinze a vinte e quatro vezes mais nos perfis conservador e flexível, nos quais sobra folga de energia para arbitrar, e ainda assim a degradação permanece praticamente idêntica à do passivo, variando no máximo um real, o que indica um custo marginal de ciclagem muito baixo, ainda que não nulo, de modo que a arbitragem se converte quase integralmente em ganho. Essa baixa sensibilidade da LFP à ciclagem é observada no início do capítulo, quando a variação de potência não variou notavelmente a degradação. A segunda leitura, de sinal oposto, aparece no flexível com NMC, no qual o agente compra menos energia que o próprio passivo, 121 kWh contra 468, e mesmo assim degrada mais, R\$ 796 contra R\$ 570, comportamento aparentemente contraditório que a discussão da Seção 4.3.2 explica e que já

Tabela 13 – Economia marginal do V2G, como a diferença observada entre o agente e o passivo no ganho de mercado e na degradação.

Cenário	Ganho de mercado (R\$/trim.)	Degrad. adicional (R\$/trim.)	Econ. marginal (R\$/trim.)	Anualizada (R\$)
Baseline (frota)	+430	+90	+340	+1.359
NMC-811	+236	+90	+146	+584
LFP	+194	+0	+194	+775
Conservador NMC-811	+122	+5	+118	+473
Flexível NMC-811	+186	+226	-40	-158
Intenso NMC-811	+492	+46	+447	+1.789
Conservador LFP	+367	+1	+366	+1.465
Flexível LFP	+646	+0	+646	+2.585
Intenso LFP	+70	+0	+70	+280

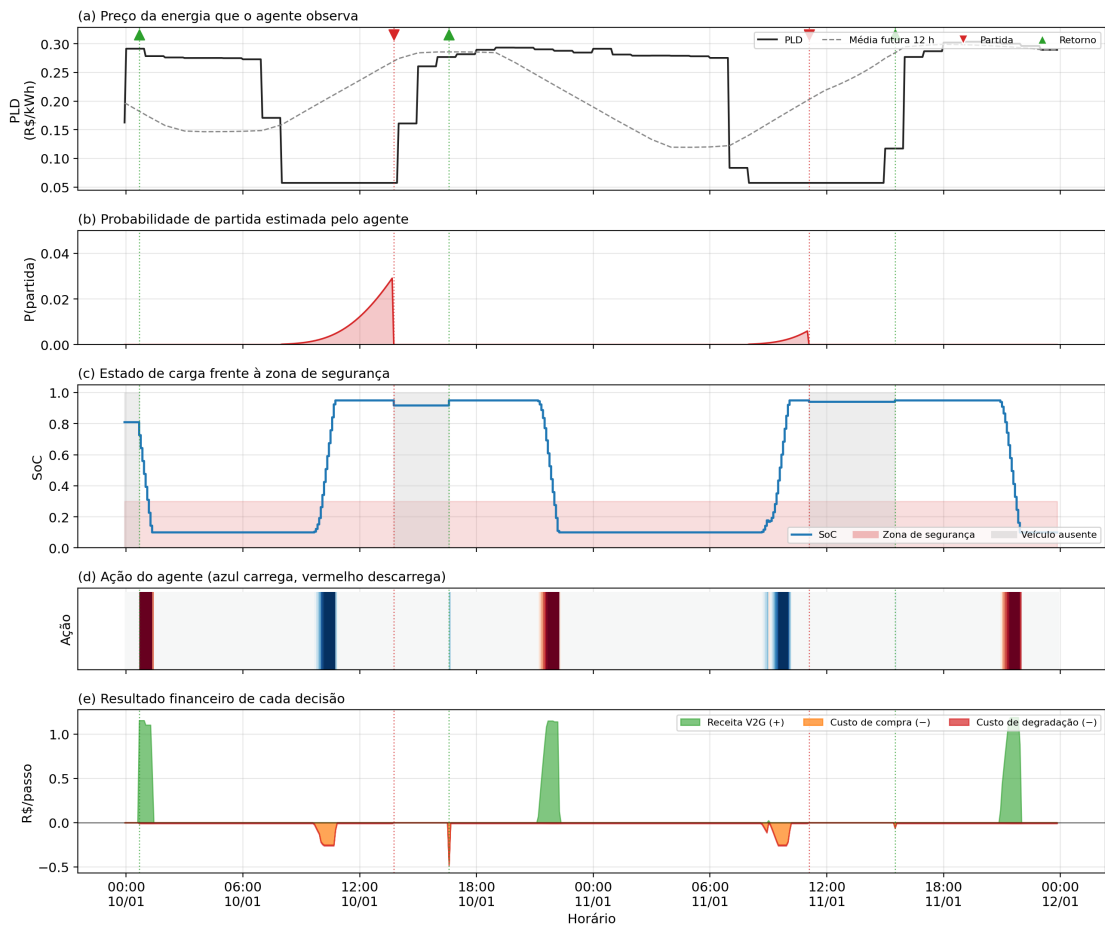
antecipa o caráter problemático desse cenário.

4.3.1 Arbitragem viável em todos os perfis com a LFP

Nos três cenários com LFP o agente compra entre 3.300 e 3.753 kWh ao longo do trimestre e devolve a maior parte dessa energia à rede, ciclando a bateria de 0,45 a 0,76 vez por dia, e paga por isso um custo de degradação que varia de R\$ 131 a R\$ 176, cerca de quatro vezes menor que o observado na NMC e composto quase inteiramente pelo lento crescimento da SEI. É justamente esse baixo custo de ciclar que liberta o agente para arbitrar com folga, comprando nos vales e vendendo nos picos sem que o desgaste consuma o ganho obtido, e o resultado aparece de forma direta no balanço econômico, já que os perfis conservador e flexível com LFP são os dois únicos cenários do estudo com lucro absoluto positivo, de R\$ 120 e R\$ 315 respectivamente, contra prejuízos de R\$ 246 e R\$ 331 do carregamento passivo correspondente.

O melhor desempenho, alcançado pelo perfil flexível com LFP, ilustra com clareza o que o agente aprendeu, e a figura 8 decompõe uma janela de 48 horas dessa operação para tornar visível a lógica da decisão. O painel superior mostra o preço da energia oscilando entre o vale da madrugada e o pico do início da noite, o segundo painel mostra a probabilidade de partida que o agente estima a partir do horário e do histórico do perfil, o terceiro acompanha o estado de carga com a faixa de segurança sombreada em vermelho, e o painel inferior registra a ação tomada a cada instante, em azul para carga e em vermelho para descarga.

O padrão que emerge é economicamente coerente, porque o agente carrega quando o preço recua em direção ao vale, eleva o estado de carga até próximo do máximo, e descarrega nos picos de preço do início da noite, sempre cuidando de recompor a carga conforme a probabilidade de partida sobe, de modo a chegar ao horário provável de viagem com energia suficiente. Vale notar que o agente permite que o estado de carga repouse na faixa de

Figura 8 – Anatomia da decisão do agente em uma janela de 48 horas do cenário flexível LFP.

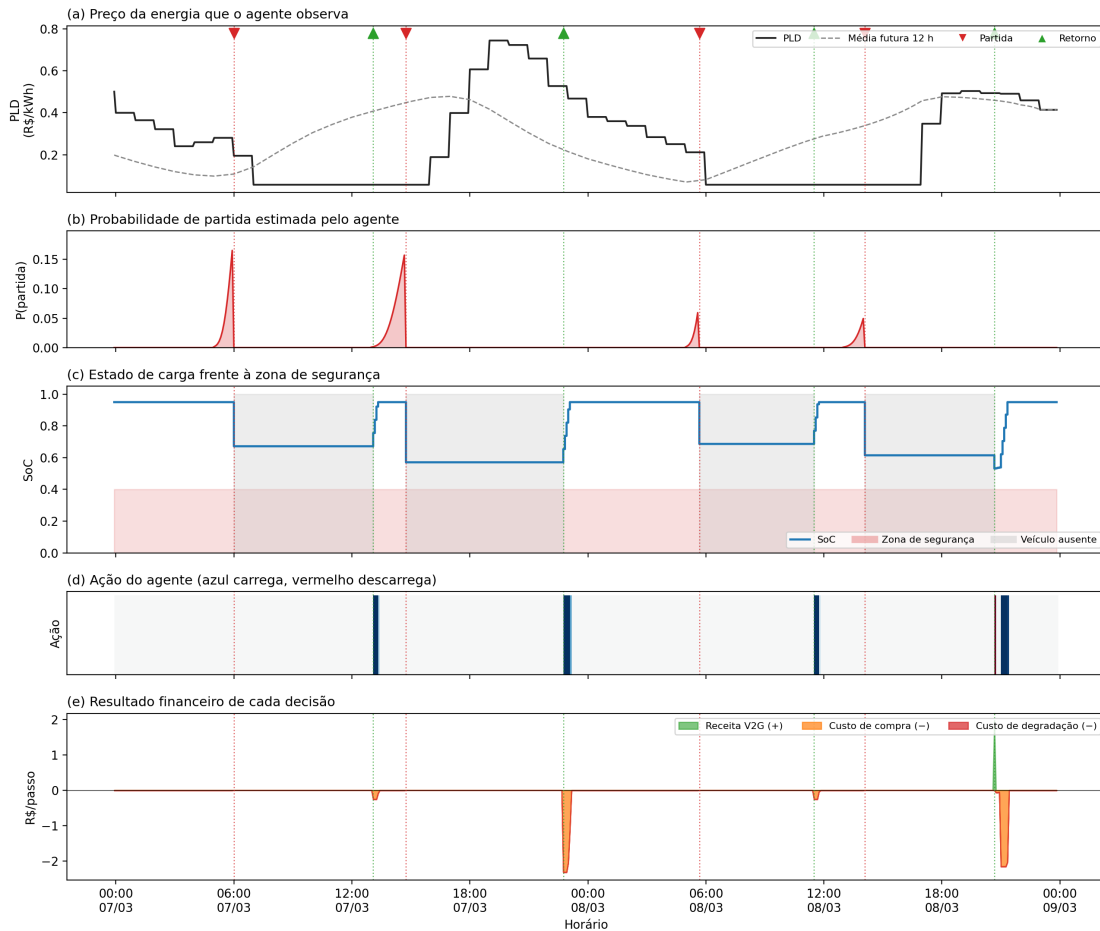
Fonte: o autor, 2026.

segurança durante os longos períodos de madrugada em que nenhuma viagem está próxima, comportamento que não configura risco porque a penalidade de ansiedade só se torna relevante quando a probabilidade de partida cresce, e nesses instantes a bateria já se encontra recarregada. É essa coordenação fina entre o preço futuro conhecido e a probabilidade de partida estimada que distingue o agente da heurística de regras fixas, e a diferença econômica confirma o ponto, já que no flexível com LFP o agente lucra R\$ 315 enquanto a heurística perde R\$ 75, sinal de que o aprendizado capturou um sincronismo entre vale, pico e janela de viagem que regras simples não reproduzem.

O perfil intenso com LFP constitui uma exceção instrutiva, porque a bateria de 50 kWh é pequena para uma demanda diária próxima de 30 kWh, o que leva o agente a um modo de operação conservador no qual compra 3.394 kWh mas vende apenas 566, reservando quase toda a energia para as próprias viagens, conforme podemos visualizar na figura 9 a seguir. É interessante de notar que o agente aprende a manter o SoC sempre alto, pois as janelas em casa são mais curtas, e assim que a probabilidade de partida começa a subir, o veículo já possui a carga necessária para realizar a viagem e manter a segurança. O lucro supera o passivo por margem estreita, de R\$ -901 contra R\$ -971, e o valor do resultado não está na

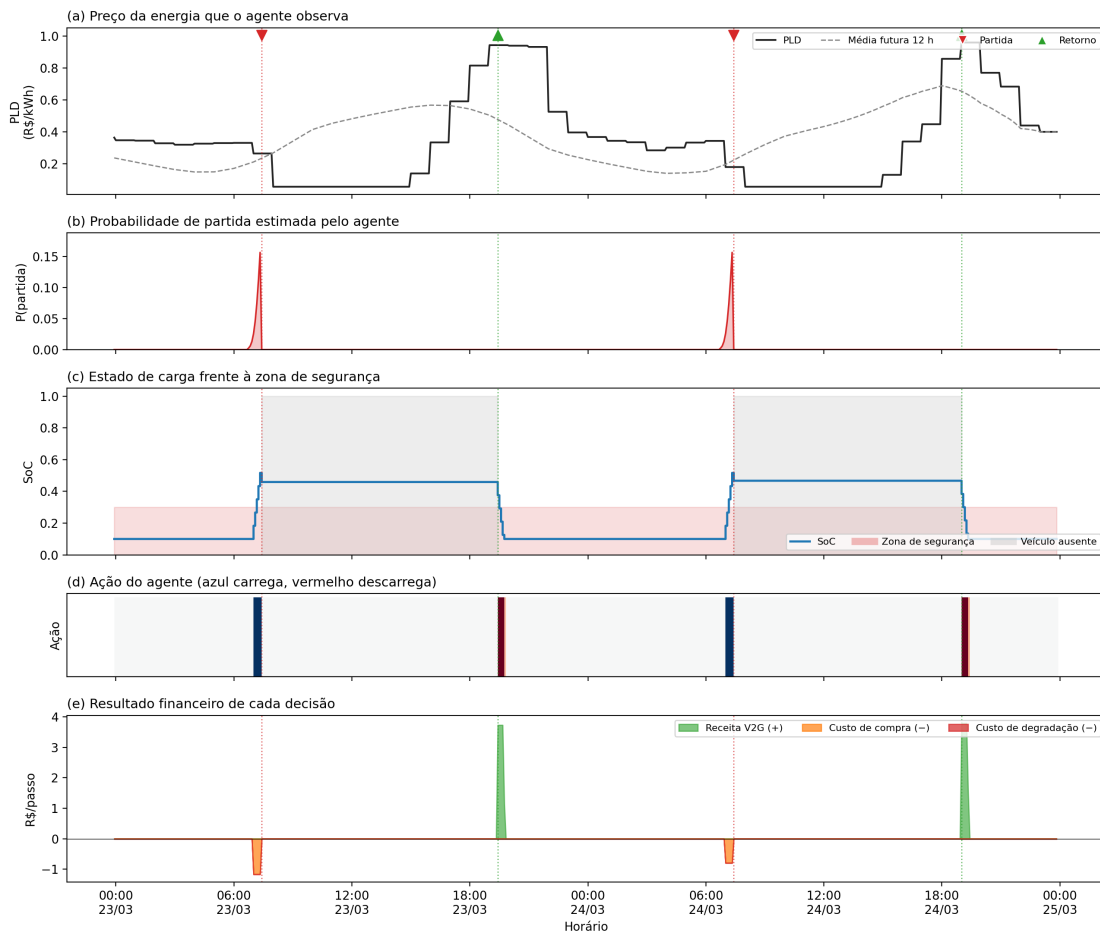
arbitragem, que é mínima, mas na constatação de que o agente reconhece quando ciclar não compensa e se abstém de fazê-lo constantemente, sem registrar nenhuma falha de autonomia ao longo do trimestre.

Figura 9 – Anatomia da decisão do agente em uma janela de 48 horas do cenário intenso LFP.



Fonte: o autor, 2026.

O perfil conservador LFP se assemelha na arbitragem com o perfil flexível de mesma química, no qual é ciclado majoritariamente para operar V2G na venda de energia, o qual percebe-se pelo contraste de ciclos entre o modo passivo e PPO das tabelas 11 e 12. Assim, o perfil conseguiu um lucro absoluto de R\$ 120 contra os custos de R\$ -246 e R\$ -153 do passivo e da heurística respectivamente, tendo uma economia marginal de R\$ 366 no trimestre e R\$ 1465 anualizados. A figura 10 nos traz o comportamento típico do agente, em que ele aprende a deixar a bateria mais próximo ao limite da zona de segurança e carregando no momento em que ele sente a probabilidade de partida atuar. É justamente esse comportamento aprendido que faz com que ele falhe 10% das partidas, onde ele tenta carregar assim que a probabilidade começa a atuar, ele começa a carregar, mas falha por uma margem mínima.

Figura 10 – Anatomia da decisão do agente em uma janela de 48 horas do cenário conservador LFP.

Fonte: o autor, 2026.

4.3.2 A NMC-811, viável apenas quando a demanda exige carga alta

Na NMC-811 o resultado se bifurca conforme o perfil de uso obrigue ou não o agente a manter o estado de carga elevado, e o caso bem-sucedido é o do perfil intenso, no qual os dois turnos diários consomem cerca de 30 kWh e forçam o agente a manter a bateria cheia, necessidade que ele converte em oportunidade ao carregar nos horários de menor preço, sobretudo no vale do meio-dia entre os dois turnos, e vender parte da energia no pico noturno antes de recompor a carga para o turno seguinte. Esse cenário cicla 0,62 vez por dia, a maior intensidade entre os perfis com NMC, supera o passivo com folga, de R\$ -1.011 contra R\$ -1.458, e quase não falha, deixando o usuário a pé em apenas 1% das partidas. A figura 11 mostra claramente a arbitragem, que de fato o agente aprendeu a reconhecer os momentos mais oportunos para recarregar a bateria e a necessidade de mantê-la com SoC elevado mesmo antes da probabilidade de partida atuar, pois ele entendeu que a degradação da NMC é mais acentuada, o que é nítido ao visualizar a inclinação da curva do SoC nos momentos de carga e descarga, se comparada em relação à química LFP. Ele é oportuno ao carregar assim que chega em casa e reconhece que é o vale, como também ao chegar em casa e reconhecer o momento alto do PLD e realizar a venda da energia.

Figura 11 – Anatomia da decisão do agente em uma janela de 48 horas do cenário intenso NMC-811.

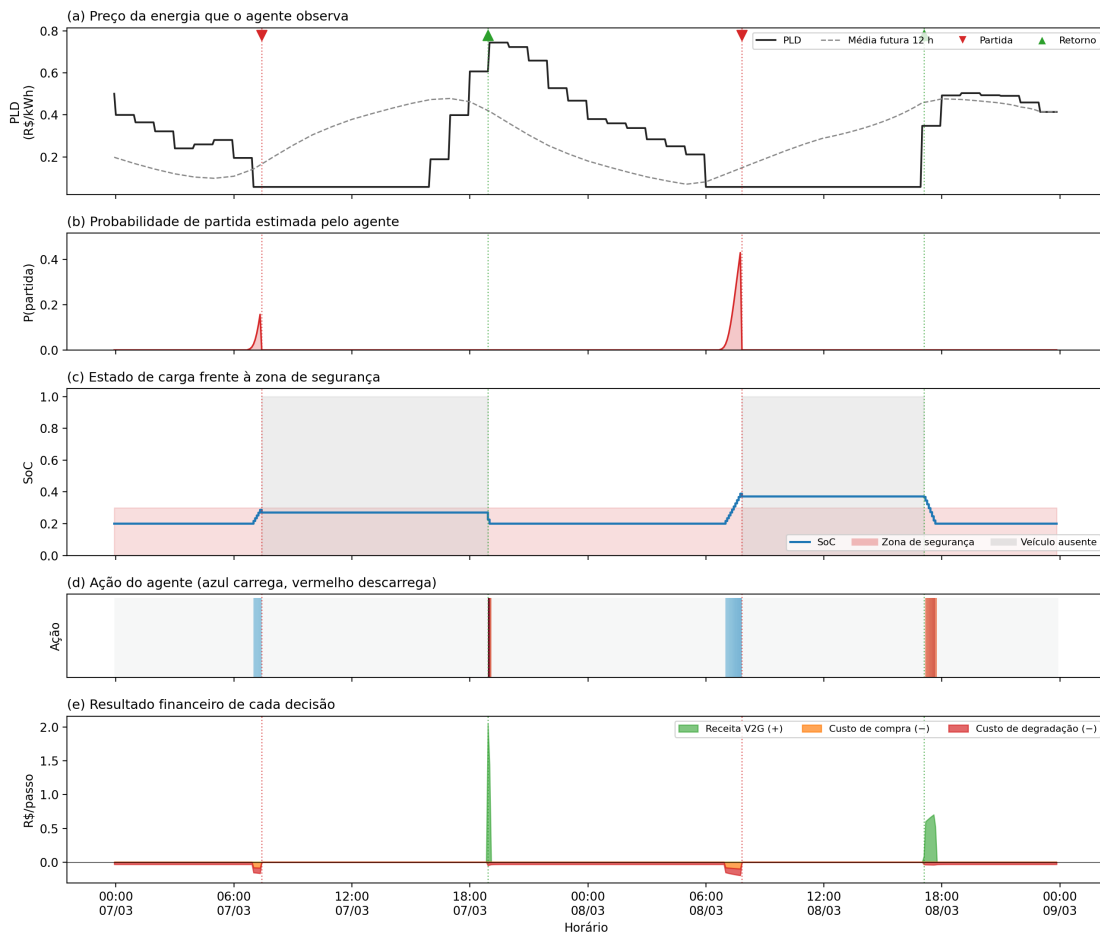
Fonte: o autor, 2026.

Um aspecto desse cenário merece atenção, porque a degradação do intenso com NMC, de R\$ 658 no trimestre, é dominada não pela deposição de lítio metálico, que é nula, mas pela perda de material ativo, responsável por R\$ 243, fato coerente com a forma de operar do agente, que carrega devagar e evita o regime de *plating*, mas cicla profundamente e com isso desgasta mecanicamente o cátodo. Trata-se, portanto, de degradação acumulada pelo uso intenso, e não pelo abuso de carga rápida, distinção que a decomposição da Tabela 15 torna explícita.

Os perfis conservador e flexível com NMC contam a história oposta e expõem o limite do agente. No conservador, as viagens curtas permitem que ele corte o estado de carga rente à reserva mínima, mantendo a bateria quase vazia e descarregando o excedente nos picos, estratégia que lhe rende lucro superior ao passivo, de R\$ -447 contra R\$ -565, mas ao custo de falhar em 30% das partidas, o que significa que, em média a cada três dias, o motorista encontra o veículo sem energia para a viagem, conforme a figura 12. A vantagem econômica desse cenário é, portanto, parcialmente construída sobre um serviço que deixa de ser prestado. Na figura 12 é demonstrado um dia típico, porém o agente não chega a falhar, mas mostra que de fato ele fica muito próximo da falha, o que faz com que ocasionalmente (30% das vezes)

ela ocorra. O agente permite o veículo falhar, pois devido a viagens curtas que o perfil realiza, a punição no sinal do reward é fraco, pois a diferença da energia requerida para a viagem é muito próxima da energia que o veículo parte, o que faz com que ele aprenda a permitir falhas de vez em quando para poder arbitrar.

Figura 12 – Anatomia da decisão do agente em uma janela de 48 horas do cenário conservador NMC-811.



Fonte: o autor, 2026.

O flexível com NMC é o caso de fracasso, e também o mais instrutivo do trabalho, porque nele o agente permanece praticamente inativo, ciclando apenas 0,02 vez por dia, menos até que o próprio passivo, mantendo a carga colada no mínimo e falhando em todas as partidas, mais de uma vez por dia. O lucro de R\$ -769 aparenta proximidade com o passivo, mas essa aparência engana, porque o agente mal toca a rede doméstica, comprando apenas 121 kWh no trimestre, de modo que as viagens são servidas quase inteiramente por recarga externa que o lucro não contabiliza, configurando não uma economia, mas um abandono do serviço. A comparação com o passivo nas Tabelas 11 e 12 torna esse abandono ainda mais nítido, porque o agente compra quatro vezes menos energia que o passivo e, ainda assim, degrada mais, R\$ 796 contra R\$ 570, uma vez que as poucas cargas que realiza acontecem em pânico, próximas à partida e em taxa elevada o suficiente para ativar a deposição de lítio metálico, de modo que a política consegue o pior dos dois mundos ao nem servir o usuário nem preservar

a bateria.

A causa desse colapso não é falta de tempo para carregar, dado que o flexível é o perfil mais conectado, com 63% de disponibilidade, mas sim a imprevisibilidade da partida, pois como a viagem principal parte em horário muito incerto, o agente nunca recebe um sinal antecipado e confiável de que precisa carregar, e quando a partida se torna iminente já é tarde para uma carga lenta, restando a opção de não carregar. Diante de uma química cara de ciclar e de uma demanda que não consegue prever, a política aprendida converge para a inércia, e o balanço de recompensa desse cenário explicita o quão custosa é essa inércia, já que a recompensa média por episódio, da ordem de -23.700 , é dominada quase inteiramente pela penalidade de falha, próxima de -18.400 , somada à penalidade de ansiedade, em torno de -4.500 , enquanto o lucro de arbitragem mal alcança $+30$. O ponto que torna esse cenário instrutivo é que o agente teria como melhorar a própria recompensa em quase 20.000 apenas carregando de forma preventiva, o que mostra que ele não está fazendo uma escolha econômica deliberada de falhar, mas sim deixando de aprender, sob esta formulação de estado, recompensa e incerteza de partida, a associação entre carregar agora e evitar a falha que ocorreria horas depois em um instante que não consegue prever.

Essa interpretação se confirma na comparação com a heurística, que, dispondo exatamente da mesma informação que o agente, carrega de forma preventiva nos vales de preço e com isso falha em apenas 4% das partidas, além de perder bem menos, R\$ -370 contra os R\$ -769 do agente, o que demonstra que uma regra simples de bom senso resolve o problema que o aprendizado não resolveu. O contraste entre o conservador e o flexível com NMC reforça a interpretação, porque ambos compartilham a mesma química e viagens curtas, diferindo apenas na previsibilidade da partida, e essa única diferença separa um cenário que falha em 30% das vezes de outro que falha em todas, o que mostra que a previsibilidade da demanda pesa tanto quanto o preço ou a química na viabilidade do V2G para o agente PPO nesses cenários.

4.3.3 A frota de referência

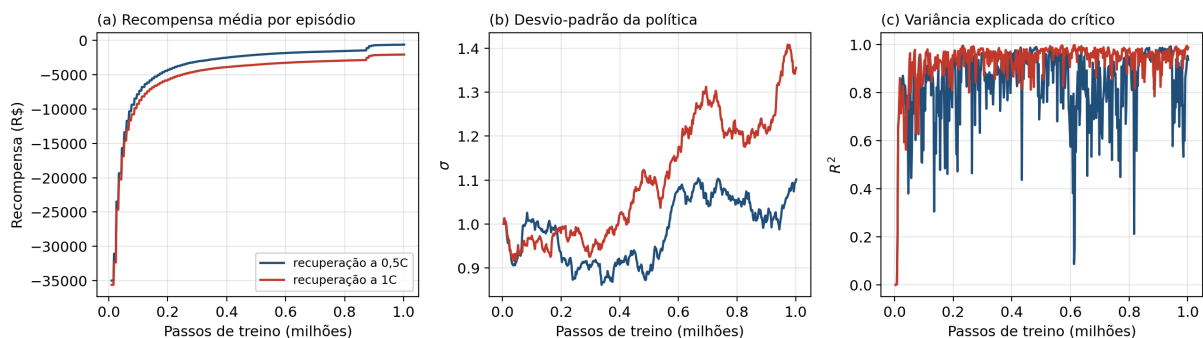
O cenário *baseline* combina um veículo NMC-811 e um veículo LFP sob horários de viagem fixos, e serve para situar os resultados estocásticos dos perfis construídos frente a um modelo pré-determinado. Nele o agente opera os dois veículos ao mesmo tempo, comprando 5.297 kWh e vendendo 2.570 kWh ao longo do trimestre, e o lucro líquido da frota, de R\$ -1.420 , supera o do carregamento passivo, de R\$ -1.760 , o que indica que a operação bidirecional compensou nessa configuração, com economia marginal de R\$ 340 no trimestre e cerca de R\$ 1.359 anualizados. A desagregação por veículo mostra que o ganho vem das duas químicas, já que o NMC melhora de R\$ -1.074 no passivo para R\$ -928 com o agente e o LFP de R\$ -686 para R\$ -492 , embora ambos permaneçam em terreno negativo porque a recarga necessária para garantir a partida, feita numa janela noturna de PLD elevado, pesa sobre o resultado absoluto das duas políticas. A heurística de regras fixas chega a um valor muito

próximo do agente, R\$ -1.452 contra R\$ -1.420 , de modo que as três políticas convergem para a mesma ordem de grandeza e o agente, nessa configuração de viagens fixas, iguala o que uma regra simples já entrega.

No *baseline* o horário de partida é fixo, como já mencionado, então o ambiente precisa garantir que cada veículo saia com a carga mínima exigida. Quando o agente deixa a bateria baixa e a hora de partir se aproxima, o ambiente assume o controle e carrega o que falta. A forma como ele faz essa carga de emergência teve um efeito grande sobre o que o agente conseguiu aprender, e vale abrir a discussão da frota por aqui, porque traz reflexões valiosas de como o aprendizado por reforço funciona.

Treinamos o mesmo agente sob duas versões dessa carga de emergência, uma que recupera a meia taxa (0,5C) e outra que recupera na taxa máxima (1C), e a Figura 13 acompanha os dois treinos. A diferença é nítida. Com a recuperação a 0,5C a recompensa sobe e estabiliza em torno de R\$ -600 por episódio, e o desvio-padrão da política fica estável perto de 1,1, sinal de que o agente assentou numa forma de operar. Com a recuperação a 1C a recompensa empaca num patamar bem pior, perto de R\$ -2.000 , e o desvio-padrão sobe ao longo do treino até cerca de 1,4, ou seja, em vez de afinar a política o agente segue tentando sem convergir para uma saída boa. Nos dois casos o crítico estima bem o retorno, o que indica que a dificuldade a 1C está na política em si, não na leitura que o agente faz do estado.

Figura 13 – Treino do agente da frota sob as duas formas de carga de emergência, recuperação a 0,5C e a 1C. (a) recompensa média por episódio, (b) desvio-padrão da política e (c) variância explicada do crítico.



Fonte: o autor, 2026.

A razão por trás disso está em como o PPO aprende. O agente é ajustado com base na ação que ele mesmo propôs, medindo o quanto ela rendeu acima ou abaixo do esperado. No instante da carga de emergência, porém, a ação aplicada é a que o ambiente impôs, e o custo pesado dessa carga a 1C, que vem da deposição de lítio metálico no NMC, acaba sendo creditado a uma ação que não foi escolhida pelo agente. Esse custo ainda chega tarde, no momento da partida, longe das decisões de horas antes que deixaram a bateria baixa. O agente recebe um sinal confuso e atrasado, e tem grande dificuldade de ligar o prejuízo do fim da noite à escolha de não ter carregado mais cedo. É por isso que, mesmo com mais de um milhão de passos de treino, ele não se antecipa e continua caindo na carga forçada.

A Tabela 14 põe as quatro políticas lado a lado, com o lucro líquido da frota e o custo de degradação separado por química. O passivo, a heurística e o agente treinado a 0,5C ficam todos na mesma faixa de prejuízo, entre R\$ −1.400 e R\$ −1.800 no trimestre, com degradação parecida. O agente treinado a 1C cai para R\$ −5.467, e a tabela mostra de onde vem a diferença. O custo de degradação do LFP é de R\$ 150 em todas as políticas, inclusive a 1C, porque a olivina não sofre deposição de lítio na temperatura de operação, então essa química fica indiferente à mudança. Toda a piora se concentra no NMC, cujo custo de degradação salta de R\$ 642 a 0,5C para R\$ 4.707 a 1C. Desse total a 1C, R\$ 3.897 vêm da deposição de lítio provocada pela carga forçada em alta taxa, R\$ 472 do crescimento da SEI e R\$ 338 da perda de material ativo. Já o ganho de mercado dos dois agentes, ou seja, o que cada um tira da arbitragem de preço, é praticamente igual e positivo, R\$ 430 a 0,5C e R\$ 448 a 1C, então o agente aprende a arbitrar bem nas duas versões e o afundamento do de 1C vem inteiramente do desgaste do NMC.

Tabela 14 – Comparação do cenário baseline com recuperação 0,5C e 1C.

Política (frota)	Lucro líquido (R\$/trim.)	Degrad. NMC (R\$)	Degrad. LFP (R\$)
Passivo	−1.760	552	150
Heurística	−1.452	534	150
PPO, recuperação a 0,5C	−1.420	642	150
PPO, recuperação a 1C	−5.467	4.707	150

O efeito na vida útil acompanha esse desgaste. O NMC operado a 0,5C projeta 5,8 anos de vida, contra 6,8 do passivo, uma perda pequena, enquanto a 1C a projeção cai para menos de um ano. O LFP fica praticamente intacto nas duas versões, em torno de 12,8 anos.

Esse comportamento do baseline ilumina, por contraste, o que faz os cenários estocásticos funcionarem, já que lá o agente conta com dois sinais que o baseline não tem, a penalidade de ansiedade, que cresce a cada passo conforme a carga cai e a partida se aproxima, e a penalidade de falha, cobrada quando ele parte sem energia suficiente. São sinais fortes, frequentes e imediatos, e é justamente isso que ensina o agente a se preparar com antecedência e a chegar ao horário de partida já com carga. A comparação confirma que uma regra rígida imposta pelo ambiente cria dificuldade de aprendizado quando o custo da recompensa não é atribuído à ação que o agente de fato escolheu, que é justamente o que acontece com a sobrescrição.

Cabe registrar que Niazi *et al.* (2026) trata a exigência de carga na partida como uma restrição rígida, mas não detalha como ela é imposta na prática. A forma adotada inicialmente neste trabalho foi a sobrescrição da ação pelo ambiente, e o treino a 1C mostra que essa escolha, quando a carga forçada é agressiva, atrapalha o aprendizado. Uma alternativa que provavelmente instruiria melhor o agente seria trocar a sobrescrição por uma penalidade na própria recompensa, cobrada quando o agente deixa a carga chegar ao ponto em que o ambiente precisaria intervir, de modo que o custo recaísse sobre a decisão do próprio agente e lhe desse um sinal mais claro para se antecipar. Essa direção fica como sugestão de melhoria

para trabalhos futuros. Diante disso, os resultados do baseline reportados neste capítulo usam a recuperação suave a 0,5C.

4.3.4 A física por trás dos custos

A Tabela 15 decompõe o custo de degradação de cada cenário estocástico nos seus mecanismos físicos, e permite enxergar a razão das diferenças discutidas acima. O contraste mais evidente é entre as químicas, porque os cenários LFP acumulam custo exclusivamente por crescimento da SEI, sem qualquer parcela de *plating* ou perda de material ativo, ao passo que os cenários NMC distribuem o custo entre os três mecanismos. Dentro da NMC, o perfil flexível é o único a registrar *plating* significativo, de R\$ 248, que provém das raras cargas de emergência em taxa elevada disparadas quando a partida se torna iminente, enquanto o perfil intenso concentra seu custo na perda de material ativo, de R\$ 243, decorrente da ciclagem profunda e frequente. Essa decomposição sustenta, em termos físicos, a afirmação de que a degradação do intenso vem do uso e a do flexível vem do mau sincronismo da carga.

Tabela 15 – Custo de degradação trimestral (R\$) por mecanismo físico do agente PPO.

Cenário	SEI (<i>Q</i>)	SEI (<i>R</i>)	<i>Plating</i>	LAM	Total
Baseline (frota)	267	351	0	174	791
NMC-811	205	263	0	174	642
LFP	62	88	0	0	150
Conservador NMC-811	193	248	0	23	464
Flexível NMC-811	234	301	248	13	796
Intenso NMC-811	181	233	0	243	658
Conservador LFP	60	85	0	0	145
Flexível LFP	73	103	0	0	176
Intenso LFP	54	77	0	0	131

A ausência completa de *plating* e de perda de material ativo na LFP sustenta o baixo custo de ciclagem observado na Tabela 13, e tem origem em duas propriedades da química. A perda de material ativo é nula porque o submodelo de mecânica de partícula, que descreve o inchaço e a fratura do material durante a ciclagem, não foi ativado para a LFP, de modo que a fratura de partícula e a consequente perda de material ativo são desprezíveis. Já o *plating*, embora ativo no modelo, não se manifesta porque a operação ocorre em torno de 25 °C, temperatura na qual, conforme adiantado na abertura do capítulo, a LFP permanece fora do regime de deposição de lítio, que nessa química só se abre abaixo de 15 °C. Resta, portanto, apenas a SEI, cujo crescimento é dominado pelo envelhecimento de calendário e responde de forma fraca à ciclagem, razão pela qual ciclar dez ou mais vezes mais que o passivo acrescenta apenas cerca de um real de degradação por trimestre, um custo baixo, mas não nulo. Cabe uma ressalva para a química da LFP, pois como mencionado, o submodelo de perda de material ativo foi desativado para o LFP, e nos perfis de ciclagem mais intensa, como

o flexível, que é justamente onde essa perda começaria a aparecer, essa desativação pode estar subestimando a degradação da química, pois ainda que a expansão da olivina seja isotrópica (ao contrário da anisotrópica da NMC) ela pode ser diferente de zero, o que acaba por ser uma das limitações deste trabalho, e uma sugestão de melhoria futura.

4.3.5 Confronto com a heurística de referência

A política heurística serve de régua para separar aquilo que o agente aprendeu de genuinamente inteligente do que regras simples já entregariam, e por isso convém confrontá-la com o PPO cenário a cenário. A heurística decide apenas pelo preço, pois carrega de forma lenta sempre que o PLD está baixo e descarrega quando está alto, sem nunca olhar a probabilidade de partida, o que a torna um bom contraponto, já que qualquer vantagem do PPO sobre ela mede o valor de antecipar a viagem, e qualquer desvantagem aponta para uma falha de aprendizado. A Tabela 16 reúne o lucro e a taxa de falha das duas políticas, e a leitura conjunta dessas duas colunas, e não do lucro isolado, é o que revela o quadro real.

Tabela 16 – PPO x heurística de referência.

Cenário	Lucro PPO (R\$)	Lucro heur. (R\$)	Falhas PPO (%)	Falhas heur. (%)
Baseline (frota)	−1.420	−1.452	0	0
NMC-811	−928	−909	0	0
LFP	−492	−546	0	0
Conservador NMC-811	−447	−441	30	24
Flexível NMC-811	−769	−370	100	4
Intenso NMC-811	−1.011	−517	1	41
Conservador LFP	+120	−153	10	33
Flexível LFP	+315	−75	9	7
Intenso LFP	−901	−263	0	56

Três padrões emergem do confronto. O primeiro aparece nos perfis conservador e flexível com LFP, nos quais o PPO supera a heurística nas duas dimensões ao mesmo tempo, lucrando mais e falhando de forma comparável ou menor, com destaque para o flexível, em que o agente obtém R\$ +315 contra os R\$ −75 da heurística com taxa de falha semelhante. Esse é o território em que o aprendizado de fato agregou valor, porque o agente descobriu um sincronismo entre vale, pico e janela de viagem que as regras fixas, presas a limiares de preço, não conseguem reproduzir.

O segundo padrão é o oposto e aparece de forma isolada no flexível com NMC, único cenário em que a heurística domina o agente nas duas dimensões, perdendo bem menos, R\$ −370 contra R\$ −769, e falhando em apenas 4% das partidas contra os 100% do PPO. Como a heurística dispõe exatamente da mesma informação e ainda assim resolve o problema

que o agente não resolveu, esse confronto é a evidência mais direta de que o PPO ficou preso em uma política ruim nesse cenário, e não de que o problema seja intratável.

O terceiro padrão é o mais sutil e cobre os perfis intensos, com as duas químicas, além em parte do conservador com NMC, nos quais a heurística exhibe lucro melhor que o do agente, mas apenas porque falha muito mais. No intenso com NMC ela perde R\$ –517 contra os R\$ –1.011 do PPO, porém abandona o usuário em 41% das partidas contra apenas 1% do agente, e no intenso com LFP a diferença é ainda mais gritante, com 56% de falhas contra zero. A razão é que a heurística, decidindo só pelo preço, não dá prioridade a chegar à viagem com energia, de modo que nos perfis de alta demanda ela deixa o carro sem carga com frequência, ao passo que o agente, penalizado pela falha, aprende a proteger o usuário e paga por isso em economia. Nesses casos o lucro superior da heurística é ilusório, porque se constrói sobre um serviço que ela deixa de prestar, e a comparação correta mostra o PPO fazendo a escolha certa.

O confronto, portanto, cumpre três funções de uma vez, ao confirmar que o agente aprendeu arbitragem real onde a química permite, ao expor o único cenário em que ele ficou aquém de uma regra trivial, e ao evidenciar que, nos perfis intensos, o que parece desvantagem econômica do agente é na verdade o preço justo de manter o usuário atendido, prioridade que a probabilidade de partida e a penalidade de falha souberam incutir no PPO, mas que a heurística ignora por construção.

4.4 IMPACTO DA QUÍMICA SOBRE A VIDA ÚTIL

O objetivo de avaliar o efeito da química sobre o ativo se traduz na vida útil estimada até o veículo atingir 80% do estado de saúde, valor usual de fim de vida automotivo, obtida por extrapolação da degradação observada no trimestre. A Tabela 17 reúne essa estimativa para o agente e para o passivo, e o contraste entre as químicas reproduz, em anos, a diferença de robustez já evidenciada nas seções anteriores, com a LFP projetando de doze a quase quinze anos de vida útil contra cerca de cinco a oito anos da NMC-811. A projeção é linear e supõe que a taxa de degradação observada no trimestre se mantém até o estado de saúde atingir 0,80:

$$\text{vida útil} = \frac{0,20}{1 - \text{SoH}_{\text{final}}} \cdot \frac{N_{\text{dias}}}{365,25} \quad [\text{anos}] \quad (4.2)$$

onde $\text{SoH}_{\text{final}}$ é o estado de saúde ao fim dos $N_{\text{dias}} = 89$ dias de simulação e 0,20 é a perda de capacidade acumulada até o fim de vida automotivo ($\text{SoH} = 0,80$).

O que mais chama a atenção nessa tabela é que, na maioria dos cenários, a operação bidirecional praticamente não encurta a vida útil frente ao passivo, resultado que confirma que o custo de degradação embutido na recompensa foi suficiente para conter a ciclagem abusiva, levando o agente a arbitrar apenas quando o ganho supera o desgaste marginal. As exceções estão na NMC. O flexível com NMC perde cerca de um ano e meio de vida útil em relação ao

Tabela 17 – Vida útil estimada até o estado de saúde de 80% (anos).

Cenário	PPO	Passivo	Heurística	Perda atribuível ao V2G
Baseline – NMC-811	5,8	6,8	7,0	um ano
Baseline – LFP	12,9	12,9	12,9	desprezível
Conservador NMC-811	8,1	8,2	8,3	desprezível
Flexível NMC-811	5,7	7,3	7,3	aprox. um ano e meio
Intenso NMC-811	5,7	6,1	8,1	desprezível
Conservador LFP	13,4	13,5	13,5	desprezível
Flexível LFP	12,3	12,3	12,3	nula
Intenso LFP	14,7	14,7	14,7	nula

passivo, encurtamento que provém das cargas de emergência em taxa elevada já discutidas, as quais ativam o *plating* e fazem desse cenário, mesmo quase inativo, o de maior custo de degradação por *plating* do estudo. O veículo NMC da frota de referência encurta cerca de um ano, de 6,8 para 5,8 anos, agora por conta da ciclagem normal da arbitragem, que combina crescimento de SEI e alguma perda de material ativo, sem nenhuma deposição de lítio, de modo que o custo do V2G sobre a vida útil dessa química existe mas é modesto e não decorre de qualquer mecanismo anômalo de carga.

Um segundo aspecto da tabela separa as duas químicas. Na LFP, o perfil intenso, que percorre quase 200 km por dia, projeta vida útil *maior* que o conservador, 14,7 contra 13,4 anos, o que parece contraintuitivo à primeira vista. A explicação está em uma característica do modelo de simulação, na qual a degradação só é contabilizada enquanto o veículo está conectado ao carregador, permanecendo congelada durante as viagens, somada ao fato de a degradação da LFP ser inteiramente de calendário. Como o perfil intenso passa mais tempo em trânsito, ficando fora da tomada cerca de 12,7 horas por dia contra 11,6 do conservador, ele acumula menos exposição ao envelhecimento de calendário, e essa redução supera o desgaste da ciclagem, desprezível nessa química. A correlação é exata nesses cenários, e a ordem da perda anual de saúde, de 1,78% no flexível, 1,46% no conservador e 1,33% no intenso, reproduz exatamente a ordem do tempo de conexão, de 15,1, 12,4 e 11,3 horas por dia, respectivamente. Convém registrar que esse comportamento é, em parte, um artefato do recorte do modelo, já que na operação real o envelhecimento de calendário prossegue mesmo com o veículo estacionado, de modo que as vidas úteis da LFP nos perfis de alta quilometragem estão provavelmente superestimadas em termos absolutos.

Na NMC, ao contrário, a ordem é a intuitiva. A ciclagem intensa do perfil intenso aciona fortemente a perda de material ativo, mecanismo que domina sua degradação e o leva à menor vida útil entre os perfis dessa química, 5,7 anos, abaixo dos 8,1 anos do conservador, cuja operação leve envelhece sobretudo por calendário. Aqui o desgaste mecânico do uso pesado supera o efeito do maior tempo em trânsito, e a vida útil projetada reflete diretamente a intensidade de uso.

4.5 SÍNTESE DOS RESULTADOS

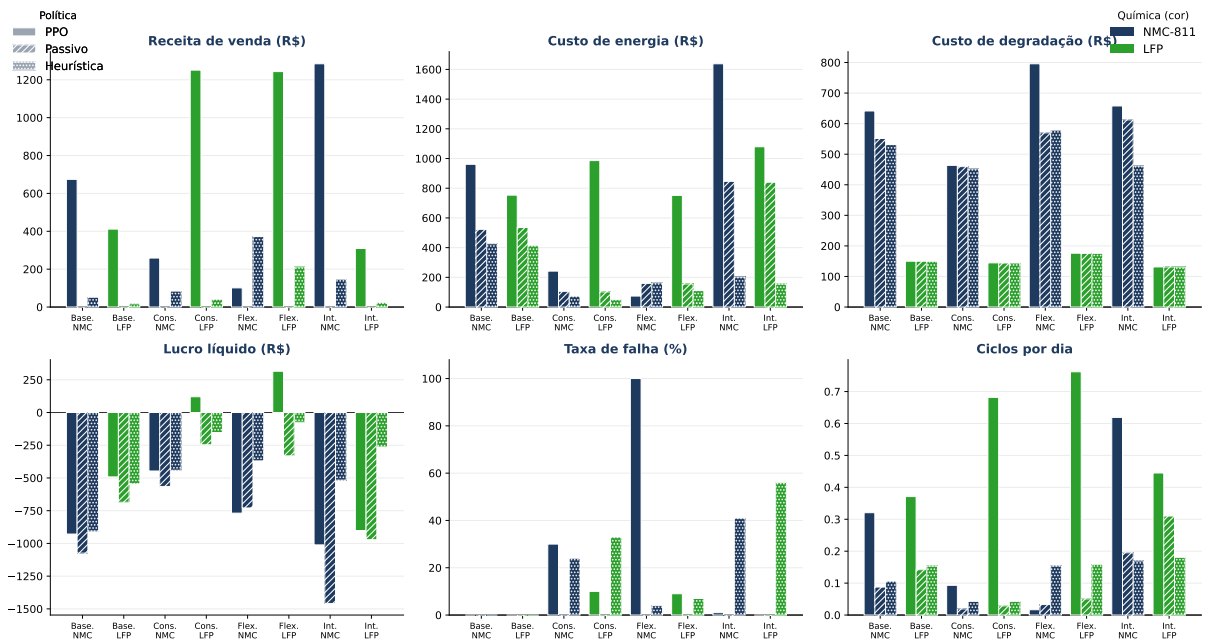
Reunindo o veredito econômico e a qualidade de serviço de cada cenário, a Tabela 18 fecha a análise em duas perguntas que precisam ser respondidas em conjunto, a primeira sobre se a operação bidirecional gera economia frente ao passivo, e a segunda sobre se o agente mantém o veículo utilizável para o motorista, de modo que um cenário só recebe veredito favorável quando responde sim às duas.

Tabela 18 – Síntese do desempenho por cenário, combinando a economia marginal frente ao passivo com a qualidade de serviço ao usuário.

Cenário	Compensa?	Serve o usuário?	Veredito
Baseline (frota)	sim	sim	viável
NMC-811	sim	sim	viável
LFP	sim	sim	viável
Conservador NMC-811	sim (marginal)	não (30% falha)	inviável pela qualidade de serviço
Flexível NMC-811	não	não (100% falha)	fracasso, abandono do usuário
Intenso NMC-811	sim	sim	viável
Conservador LFP	sim	aceitável (10%)	viável
Flexível LFP	sim (melhor)	aceitável (9%)	viável, melhor caso
Intenso LFP	sim (marginal)	sim	viável

A Figura 14 reúne as seis métricas de todos os cenários e políticas em um único painel.

Figura 14 – Comparação das três políticas em seis métricas por cenário.



Fonte: o autor, 2026.

O quadro consolida a ideia de que a viabilidade do V2G no PLD Sudeste se decide pela interação entre química e perfil, e não por um fator isolado, porque a LFP entrega veredito

favorável nos três perfis, ao passo que a NMC só é viável quando o perfil intenso a obriga a manter carga alta, falhando nos perfis de baixa demanda, seja por sacrificar o serviço, como no conservador, seja por colapsar inteiramente, como no flexível. O baseline, ao combinar as duas químicas sob viagens determinísticas, confirma em escala de frota que é a imoprevisibilidade dos perfis alinhada a alta degradação da NMC que dificulta a viabilidade do V2G.

4.6 LIMITAÇÕES E POSSIBILIDADES DE MELHORIA

A operação V2G compensa quando a química tolera o ciclamento, como a LFP em todos os perfis, ou quando o perfil obriga o agente a manter carga elevada, como o motorista intenso com NMC, e falha quando uma química cara de ciclar encontra um perfil de baixa demanda ou de partida imprevisível. Essas falhas vêm da calibração/modelagem da função de recompensa, não de um defeito do simulador, e apontam melhorias para trabalhos futuros

Antes de discutir essas melhorias, é honesto reconhecer duas limitações do próprio modelo que condicionam a leitura dos resultados. A primeira é que a degradação só é contabilizada enquanto o veículo está conectado ao carregador, ficando congelada durante as viagens, o que subestima o envelhecimento de calendário dos perfis de alta quilometragem e, como discutido na Seção 4.4, faz com que o perfil intenso projete vida útil maior que o conservador na LFP, um resultado que deve ser lido como artefato do recorte do modelo e não como propriedade física. A segunda é que as vidas úteis da Tabela 17 resultam da extrapolação da degradação de um trimestre para vários anos, sob a hipótese de estacionariedade do preço, do uso, da temperatura e da política, hipótese que não se sustenta literalmente no longo prazo, de modo que esses valores valem como indicador comparativo entre cenários e não como previsão absoluta de durabilidade.

A mais direta diz respeito à penalidade de falha de autonomia, cujo coeficiente foi fixado em $\lambda_{\text{fail}} = 50$ durante o treino, valor que torna a falha relativamente barata para o agente diante de uma química cara de carregar, de modo que elevá-lo para algo como $\lambda_{\text{fail}} = 500$ aumentaria em dez vezes o custo de não servir o usuário e é a correção mais provável para o abandono observado no flexível com NMC, sem recorrer a restrições rígidas que eliminariam a liberdade de operação do agente. Uma segunda melhoria atua sobre o horizonte de decisão, porque o fator de desconto $\gamma = 0,997$ corresponde a um horizonte efetivo de cerca de 28 horas, curto demais para que o agente associe uma carga feita hoje a uma viagem que ocorrerá amanhã, e elevá-lo para 0,999 ampliaria esse horizonte para vários dias, favorecendo o carregamento antecipado justamente nos perfis de partida imprevisível.

Um terceiro caminho seria de enriquecer a percepção do agente, já que a política atual decide com base apenas no estado instantâneo, sem memória do que ocorreu antes no episódio, de maneira que acrescentar à observação variáveis como o tempo decorrido desde a conexão e a energia já comprada no dia, ou ainda substituir a rede do agente por uma arquitetura recorrente capaz de reter memória ao longo do tempo, permitiria inferir melhor a aproximação de uma viagem e atacaria diretamente a dificuldade de antecipação que caracteriza o perfil

flexível. Por fim, no plano estrutural, o intervalo de preço do PLD como única fonte de receita mostrou-se estreito para viabilizar a NMC nos perfis de baixa demanda, e a inclusão de serviços ancilares, como a regulação de frequência, ampliaria a receita por ciclo e poderia inverter o sinal econômico nos cenários hoje marginais, em linha com o que Niazi *et al.* (2026) apontam como condição para o V2G em químicas de alta densidade energética.

5 CONSIDERAÇÕES FINAIS

O agente de aprendizado por reforço, treinado sobre o PLD horário do submercado Sudeste e penalizado pela degradação real da bateria, obteve lucro com arbitragem nos cenários em que a célula suporta o ciclamento sem se desgastar rápido, e falhou onde a química de ciclagem mais custosa se combinou a um uso de baixa demanda, tornando assim, a viabilidade do V2G dependente da química da bayeria e do perfil de uso do veículo

A LFP sustentou esse ganho nos três perfis testados, com o agente comprando energia barata nos vales de preço e vendendo nos picos do início da noite, e o custo de degradação que essa operação acrescentou sobre o carregamento passivo ficou inexpressivo por trimestre. Essa robustez decorre de duas propriedades da química, porque a 25°C a LFP permanece fora do regime de deposição de lítio metálico e a estabilidade mecânica da estrutura olivina mantém o desgaste restrito ao lento crescimento da SEI, o que explica seu custo total de degradação quatro vezes menor que o da NMC. A arbitragem compensou frente ao passivo nos três perfis, e o lucro absoluto positivo apareceu no conservador e no flexível, o melhor caso do estudo.

A NMC-811 foi viável apenas no perfil intenso, pois nele, a alta quilometragem obriga o veículo a manter carga elevada, e o agente aproveita essa necessidade para arbitrar com folga. Nos perfis de baixa demanda a química não compensou, sendo que no flexível com partida imprevisível, o agente não conseguiu garantir energia ao usuário, e no conservador ele chegou a falhar 30% das partidas, comrpometendo a utilização da tecnologia no serviço ao usuário. Para químicas de alta densidade energética, o V2G só compensa quando o próprio uso do veículo já exige carga alta.

Esses resultados só valem enquanto se aceita a premissa que organiza todo o trabalho, segundo a qual o usuário pode comprar e vender energia ao preço horário do PLD, e essa premissa ainda não se realiza no Brasil, onde o consumidor residencial não tem acesso ao mercado de curto prazo nem remuneração definida para a energia que devolveria à rede. Enquanto essa lacuna regulatória não for resolvida, o V2G permanece inviável por uma razão que antecede a econômica, de modo que a abertura do mercado ao usuário é o passo anterior a qualquer discussão de ganho. Mesmo depois dessa abertura, a arbitragem pelo PLD deve ser entendida como a receita mínima do modelo, já que o intervalo diário de preço no Sudeste é estreito, da ordem de R\$ 0,20 por kWh entre o vale e o pico, o que limita o ganho por ciclo. Os serviços ancilares, como a regulação de frequência, atuam justamente sobre essa limitação, pois remuneram a disponibilidade e a resposta rápida da bateria e ampliariam a margem o suficiente para viabilizar os casos que a arbitragem pura deixou em prejuízo.

A continuação mais direta deste trabalho passa por incorporar os serviços ancilares ao próprio ambiente de simulação e, com eles, adotar premissas de mercado distintas da brasileira, reproduzindo arranjos regulatórios já consolidados em outros países, onde a remuneração da regulação de frequência e da reserva de potência torna o V2G uma operação madura. Esse

caminho permitiria estimar quanta receita ancilar bastaria para inverter os cenários que a arbitragem pura deixou em prejuízo, e aproximaria a simulação das condições em que o V2G já opera no exterior.

No plano do modelo, há espaço para substituir o modelo eletroquímico de degradação por formulações alternativas, capazes de capturar mecanismos ou químicas não contemplados aqui, e para refinar o que ficou simplificado, como a contabilização do envelhecimento de calendário durante as viagens, hoje congelado, que corrigiria a superestimação da vida útil nos perfis de maior quilometragem. No plano do agente, os refinamentos de recompensa e de percepção discutidos no capítulo de resultados atacariam diretamente as falhas observadas nos perfis de partida imprevisível. Na mesma linha, o cenário de referência mostrou que a forma como o ambiente garante uma restrição rígida pesa sobre o que o agente consegue aprender, porque uma intervenção que sobrescreve a ação do motorista entrega a ele um custo que não veio de uma escolha sua, o que dificulta associar a degradação à decisão que de fato a causou. Substituir essa sobrescrição por um sinal de recompensa que recaia sobre a própria decisão do agente, cobrado quando ele se aproxima da condição que obrigaria a intervenção, é uma direção promissora tanto para essa restrição rígida quanto para os perfis de partida imprevisível. Vale ainda comparar o PPO com algoritmos *off-policy* como SAC ou TD3, já que o PPO foi o único algoritmo investigado neste trabalho. A extensão a outros submercados e a outras janelas sazonais completa esse conjunto de direções.

Cabe reconhecer, por fim, que todos os resultados derivam de simulação computacional. A modelagem eletroquímica emprega parâmetros experimentais de células comerciais e o PLD utiliza dados históricos reais da CCEE, mas a integração entre eletroquímica, mercado e comportamento do motorista é inteiramente simulada, de modo que as conclusões sobre o lucro V2G são estimativas baseadas de modelagem física, não medições. A validação cruzada experimental, possivelmente no contexto de um *sandbox* regulatório no Brasil, é a continuidade direta desta linha de pesquisa.

REFERÊNCIAS

ACHARIGE, S. S. G.; HAQUE, M. E.; ARIF, M. T.; HOSSEINZADEH, N. Review of Electric Vehicle Charging Technologies, Configurations, and Architectures. *IEEE Access*, v. 11, p. 41218–41255, 2023. Citado 2 vezes nas páginas 5 e 11.

Agência Nacional de Energia Elétrica. *Resolução Normativa n.º 1.000, de 7 de dezembro de 2021*. Brasília, 2021. Disponível em: <<https://www2.aneel.gov.br/cedoc/ren20211000.pdf>>. Citado na página 15.

Agência Nacional de Energia Elétrica. *Resolução Normativa n.º 1.030, de 26 de julho de 2022. Aprimora as regras para prestação e remuneração de serviços ancilares no SIN*. Brasília, 2022. Disponível em: <<https://www2.aneel.gov.br/cedoc/ren20221030.pdf>>. Citado 2 vezes nas páginas 2 e 15.

Agência Nacional de Energia Elétrica. *Consulta Pública n.º 42/2025 – Eletromobilidade e conexão à rede*. Brasília, 2025. Disponível em: <<https://www.gov.br/aneel/pt-br/assuntos/noticias/2025/aneel-abre-consulta-publica-para-aprimorar-regras-de-conexao-de-eletromobilidade-a-rede-eletrica>>. Citado 2 vezes nas páginas 12 e 15.

AI, W.; KRAFT, L.; STURM, J.; JOSSEN, A.; WU, B. Electrochemical Thermal-Mechanical Modelling of Stress Inhomogeneity in Lithium-Ion Pouch Cells. *Journal of The Electrochemical Society*, v. 167, n. 1, p. 013512, 2020. Citado 2 vezes nas páginas 28 e 84.

ALIPOUR, M.; ZIEBERT, C.; CONTE, F. V.; KIZILEL, R. A Review on Temperature-Dependent Electrochemical Properties, Aging, and Performance of Lithium-Ion Cells. *Batteries*, MDPI, v. 6, n. 3, p. 35, 2020. Citado na página 8.

Associação Brasileira do Veículo Elétrico. *ABVE Data: Números do mercado de eletromobilidade*. São Paulo, 2025. Disponível em: <<https://abve.org.br/abve-data/>>. Citado na página 2.

BANDHAUER, T. M.; GARIMELLA, S.; FULLER, T. F. A critical review of thermal issues in lithium-ion batteries. *Journal of The Electrochemical Society*, v. 158, n. 3, p. R1–R25, 2011. Citado na página 35.

Brasil. *Lei n.º 14.300, de 6 de janeiro de 2022. Institui o marco legal da microgeração e minigeração distribuída*. 2022. Diário Oficial da União. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2022/lei/L14300.htm>. Citado na página 15.

Brasil. *Lei n.º 15.269, de 24 de novembro de 2025. Reconhece o armazenamento de energia como atividade do setor elétrico e institui o agente de armazenamento*. 2025. Diário Oficial da União. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/lei/l15269.htm>. Citado na página 15.

Câmara de Comercialização de Energia Elétrica. *Conceitos de Preços*. São Paulo, 2024. Disponível em: <<https://www.ccee.org.br/precos/conceitos-precos>>. Citado na página 13.

Câmara de Comercialização de Energia Elétrica. *Info PLD 2024 – Boletins semanais*. São Paulo, 2024. Disponível em: <<https://www.ccee.org.br/precos/conceitos-precos>>. Citado 2 vezes nas páginas 14 e 50.

CHEN, C.-H.; PLANELLA, F. B.; O'REGAN, K.; GASTOL, D. E.; WIDANAGE, W. D.; KENDRICK, E. Development of Experimental Techniques for Parameterization of Multi-scale Lithium-ion Battery Models. *Journal of The Electrochemical Society*, v. 167, n. 8, p. 080534, 2020. Citado 3 vezes nas páginas 9, 29 e 84.

DOYLE, M.; FULLER, T. F.; NEWMAN, J. Modeling of Galvanostatic Charge and Discharge of the Lithium/Polymer/Insertion Cell. *Journal of the Electrochemical Society*, v. 140, n. 6, p. 1526–1533, 1993. Citado 2 vezes nas páginas 5 e 27.

DUBARRY, M.; QIN, N.; BROOKER, P. Calendar aging of commercial Li-ion cells of different chemistries – A review. *Current Opinion in Electrochemistry*, v. 9, p. 106–113, 2018. Citado 3 vezes nas páginas 2, 8 e 9.

Empresa de Pesquisa Energética. *Balanco Energético Nacional 2025: ano base 2024 – Relatório Síntese*. Rio de Janeiro, 2025. Disponível em: <<https://www.epe.gov.br/pt/publicacoes-dados-abertos/publicacoes/balanco-energetico-nacional-ben>>. Citado 2 vezes nas páginas 1 e 12.

ETXANDI-SANTOLAYA, M.; CASALS, L. C.; CORCHERO, C. Extending the electric vehicle battery first life: Performance beyond the current end of life threshold. *Heliyon*, v. 10, n. 4, p. e26066, 2024. Citado na página 41.

FENG, X.; OUYANG, M.; LIU, X.; LU, L.; XIA, Y.; HE, X. Thermal runaway mechanism of lithium ion battery for electric vehicles: A review. *Energy Storage Materials*, v. 10, p. 246–267, 2018. Citado 2 vezes nas páginas 9 e 10.

Futura Energy. *CCEE Projeta PLD Médio de R\$ 59/MWh até Maio*. 2025. Futura Energy. Fonte secundária; análise de projeção de PLD para 2025. Acesso em: 26 jun. 2026. Citado 2 vezes nas páginas 3 e 14.

International Energy Agency. *Electricity 2024*. Paris, 2024. Disponível em: <<https://www.iea.org/reports/electricity-2024>>. Citado na página 1.

International Trade Administration. *Brazil Electric Vehicle Grid*. 2026. U.S. Department of Commerce. Fonte secundária; análise da CP 42/2025 e do mercado brasileiro de eletromobilidade. Acesso em: 28 jun. 2026. Citado na página 15.

KARGER, A.; O'KANE, S. E. J.; ROGGE, M.; KIRST, C.; SINGER, J. P.; MARINESCU, M.; OFFER, G. J.; JOSSEN, A. Modeling Particle Versus SEI Cracking in Lithium-Ion Battery Degradation: Why Calendar and Cycle Aging Cannot Simply be Added. *Journal of The Electrochemical Society*, v. 171, p. 090512, 2024. Citado na página 8.

KEMPTON, W.; TOMIĆ, J. Vehicle-to-grid power fundamentals: Calculating capacity and net revenue. *Journal of Power Sources*, v. 144, n. 1, p. 268–279, 2005. Citado 5 vezes nas páginas 2, 7, 10, 11 e 12.

MARTINEZ-LASERNA, E.; GANDIAGA, I.; SARASKETA-ZABALA, E.; BADEDA, J.; STROE, D.-I.; SWIERCZYNSKI, M.; GOIKOETXEA, A. Battery second life: Hype, hope or reality? A critical review of the state of the art. *Renewable and Sustainable Energy Reviews*, v. 93, p. 701–718, 2018. Citado na página 34.

Ministério de Minas e Energia. *Consulta Pública n.º 176/2024 – BESS em Leilão de Reserva de Capacidade*. Brasília, 2024. Disponível em: <<https://www.gov.br/participamaisbrasil/cp-176>>. Citado na página 2.

Ministério de Minas e Energia; Empresa de Pesquisa Energética. *Plano Decenal de Expansão de Energia 2034*. Brasília, 2025. Versão final aprovada pela Portaria MME 831/2025 (DOU 09/04/2025). Disponível em: <https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-804/topico-758/PDE2034_Aprovado.pdf>. Citado na página 12.

MOHTAT, P.; LEE, S.; SULZER, V.; SIEGEL, J. B.; STEFANOPOULOU, A. G. Differential Expansion and Voltage Model for Li-ion Batteries at Practical Charging Rates. *Journal of The Electrochemical Society*, v. 167, n. 11, p. 110561, 2020. Citado na página 9.

NAUMANN, M.; SCHIMPE, M.; KEIL, P.; HESSE, H. C.; JOSSEN, A. Analysis and modeling of calendar aging of a commercial LiFePO₄/graphite cell. *Journal of Energy Storage*, v. 17, p. 153–169, 2018. Citado 2 vezes nas páginas 30 e 31.

NEUBAUER, J.; PESARAN, A. The ability of battery second use strategies to impact plug-in electric vehicle prices and serve utility energy storage applications. *Journal of Power Sources*, v. 196, n. 23, p. 10351–10358, 2011. Citado na página 34.

NEUBAUER, J.; WOOD, E. The impact of range anxiety and home, workplace, and public charging infrastructure on simulated battery electric vehicle lifetime utility. *Journal of Power Sources*, v. 257, p. 12–20, 2014. Citado na página 36.

Nexdigm. *Brazil's EV Charging Infrastructure Industry*. 2025. Relatório setorial. Disponível em: <<https://www.nexdigm.com/market-research/insights/blog/brazil-ev-charging-infrastructure-industry/>>. Citado na página 3.

NIAZI, M. A.; KUMAR, V.; SAJID, Q.; KOONDHAR, M. A.; KIM, Y.-S.; ZAINURI, M. A. A. M.; TOUTI, E. A Physics-Informed Reinforcement Learning Approach to Vehicle-to-Grid Control With Real-Time Battery Degradation. *IEEE Access*, v. 14, jan. 2026. Citado 22 vezes nas páginas 3, 23, 24, 25, 27, 28, 29, 30, 32, 33, 36, 39, 41, 42, 43, 45, 46, 47, 49, 68, 75 e 84.

O'KANE, S. E. J.; AI, W.; MADABATTULA, G.; ALONSO-ALVAREZ, D.; TIMMS, R.; SULZER, V.; EDGE, J. S.; WU, B.; OFFER, G. J.; MARINESCU, M. Lithium-ion battery degradation: how to model it. *Physical Chemistry Chemical Physics*, v. 24, n. 13, p. 7909–7922, 2022. Citado 3 vezes nas páginas 8, 30 e 84.

OLIVEIRA, S. A. d. *Gêmeos Digitais para Baterias de Veículos Elétricos*. Dissertação (Mestrado) — Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Campina Grande, Campina Grande, mar. 2025. Citado 2 vezes nas páginas 6 e 7.

Operador Nacional do Sistema Elétrico. *Submódulo 10.6 – Controle da geração*. Rio de Janeiro, 2018. Disponível em: <<https://www.ons.org.br/ProcedimentosDeRede/MÃşdulo%2010/SubmÃşdulo%2010.6/SubmÃşdulo%2010.6%202018.08.pdf>>. Citado 2 vezes nas páginas 13 e 15.

Operador Nacional do Sistema Elétrico. *Relatório de Análise de Perturbação – Ocorrência no SIN em 15 de agosto de 2023*. Rio de Janeiro, 2023. Disponível em:

<<https://www.ons.org.br/AcervoDigitalDocumentosEPublicacoes/RAP%202023.08.15%2008h030min%20vers%C3%A3o%20final.pdf>>. Citado na página 2.

ORFANOUDAKIS, S.; DIAZ-LONDONO, C.; YILMAZ, Y. E.; PALENSKY, P.; VERGARA, P. P. *EV2Gym: A Flexible V2G Simulator for EV Smart Charging Research and Benchmarking*. 2024. ArXiv preprint arXiv:2404.01849. Disponível em: <<https://arxiv.org/abs/2404.01849>>. Citado 2 vezes nas páginas 3 e 25.

PESARAN, A. A. *Battery Thermal Management in EVs and HEVs: Issues and Solutions*. [S.l.], 2002. Disponível em: <<https://www.nrel.gov/docs/fy02osti/31901.pdf>>. Citado na página 35.

PLÖTZ, P.; JAKOBSSON, N.; SPREI, F. On the distribution of individual daily driving distances. *Transportation Research Part B: Methodological*, v. 101, p. 213–227, 2017. Citado na página 37.

Portal Solar. *Cortes na geração solar e eólica somaram 400 mil horas em 2024*. 2025. Portal de notícias do setor solar. Fonte secundária; reporta dado consolidado da Volt Robotics. Disponível em: <<https://www.portalsolar.com.br/noticias/operacao-e-expansao/oem/cortes-na-geracao-solar-e-eolica-somaram-400-mil-horas-em-2024>>. Citado na página 1.

PRADA, E.; DOMENICO, D. D.; CREFF, Y.; BERNARD, J.; SAUVANT-MOYNOT, V.; HUET, F. A Simplified Electrochemical and Thermal Aging Model of LiFePO₄-Graphite Li-ion Batteries: Power and Capacity Fade Simulations. *Journal of The Electrochemical Society*, v. 160, n. 4, p. A616–A628, 2013. Citado 3 vezes nas páginas 10, 29 e 84.

REDDY, T. B. (Ed.). *Linden's Handbook of Batteries*. 4. ed. New York: McGraw-Hill, 2011. Citado na página 6.

SCHMALSTIEG, J.; KÄBITZ, S.; ECKER, M.; SAUER, D. U. A holistic aging model for Li(NiMnCo)O₂ based 18650 lithium-ion batteries. *Journal of Power Sources*, v. 257, p. 325–334, 2014. Citado na página 31.

SCHULMAN, J.; MORITZ, P.; LEVINE, S.; JORDAN, M. I.; ABBEEL, P. *High-Dimensional Continuous Control Using Generalized Advantage Estimation*. 2015. ArXiv preprint arXiv:1506.02438. Disponível em: <<https://arxiv.org/abs/1506.02438>>. Citado 2 vezes nas páginas 22 e 46.

SCHULMAN, J.; WOLSKI, F.; DHARIWAL, P.; RADFORD, A.; KLIMOV, O. *Proximal Policy Optimization Algorithms*. 2017. ArXiv preprint arXiv:1707.06347. Disponível em: <<https://arxiv.org/abs/1707.06347>>. Citado 4 vezes nas páginas 3, 21, 22 e 47.

SULZER, V.; MARQUIS, S. G.; TIMMS, R.; ROBINSON, M.; CHAPMAN, S. J. Python Battery Mathematical Modelling (PyBaMM). *Journal of Open Research Software*, v. 9, n. 1, p. 14, 2021. Citado na página 27.

SUTTON, R. S.; BARTO, A. G. *Reinforcement Learning: An Introduction*. 2. ed. Cambridge, MA: MIT Press, 2018. Citado 6 vezes nas páginas 3, 16, 17, 20, 21 e 22.

TAHMASBI, A. A.; KADYK, T.; EIKERLING, M. H. Statistical Physics-Based Model of Solid Electrolyte Interphase Growth in Lithium Ion Batteries. *Journal of The Electrochemical Society*, v. 164, n. 6, p. A1307–A1313, 2017. Citado 2 vezes nas páginas 30 e 85.

WANG, J.; SUN, X. Olivine LiFePO₄: the remaining challenges for future energy storage. *Energy & Environmental Science*, v. 8, p. 1110–1138, 2015. Citado na página 29.

WHEATCROFT, L.; BIRD, A.; GOLLAPALLY, N.; BOOTH, S. G.; CUSSEN, S. A.; INKSON, B. J. In Situ Fracture Behavior of Single Crystal LiNi_{0.8}Mn_{0.1}Co_{0.1}O₂ (NMC811). *Batteries & Supercaps*, v. 7, n. 6, p. e202400077, 2024. Citado na página 30.

XIE, H.; SONG, G.; SHI, Z.; ZHANG, J.; LIN, Z.; YU, Q.; FU, H.; SONG, X.; ZHANG, H. Reinforcement learning for vehicle-to-grid: A review. *Advances in Applied Energy*, v. 17, p. 100214, 2025. Citado 3 vezes nas páginas 22, 23 e 46.

Apêndices

APÊNDICE A – PARÂMETROS DE REPLICAÇÃO

Este apêndice reúne, os valores numéricos e as opções de configuração necessários para reproduzir os resultados desta monografia. Os parâmetros de cada química (Tabela 1), os coeficientes de monetização (Tabela 4), os perfis de mobilidade (Tabela 5) e os hiperparâmetros do PPO (Tabela 6) estão no Capítulo 3 e não são repetidos aqui; as tabelas a seguir consolidam o restante.

Quanto às fontes, as opções de submodelo (Tabela 22) e os parâmetros de base do *patch* de degradação (Tabela 23) vêm de O’Kane *et al.* (2022) e Ai *et al.* (2020), sobre os *parameter sets* Chen2020 (CHEN *et al.*, 2020) (NMC-811) e Prada2013 (PRADA *et al.*, 2013) (LFP). Os custos de reposição que originam os coeficientes α seguem Niazi *et al.* (2026), e os coeficientes λ_{anx} e λ_{fail} da recompensa são calibração própria desta monografia (§3.4.5).

Tabela 19 – Constantes globais do ambiente de simulação.

Constante	Valor
Passo de tempo Δt	5 min (288 passos/dia)
Eficiência do conversor bidirecional η	0,95 (aplicada no lado da rede)
Eficiência veicular η_{EV}	0,15 kWh/km
Coeficiente de dissipação entre passos k_{dis}	0,40 por passo ($\tau \approx 10$ min)
Horizonte do episódio (treino)	30 dias
Janela da simulação final	89 dias (01/01 a 30/03 de 2026)
Submercado do PLD	Sudeste/Centro-Oeste
<i>Seed</i> do treino	42 (idêntica em todos os cenários)
<i>Seeds</i> da simulação final	1000 a 1009 (10 <i>seeds</i> de inicialização)
Validação <i>out-of-sample</i> (<i>overfitting</i>)	30 episódios, <i>seed</i> 123
<i>Total timesteps</i> por cenário	1×10^6

Tabela 20 – Parâmetros das baterias por química.

Parâmetro	NMC-811	LFP
Capacidade nominal	80 kWh	50 kWh
$\text{SoC}^{\text{min}} / \text{SoC}^{\text{max}}$	0,20 / 0,90	0,10 / 0,95
Taxa C máxima	1C	1C
Custo de reposição	R\$ 764,15/kWh	R\$ 632,40/kWh
α_Q	R\$ 61.132	R\$ 31.620
$\alpha_R = \alpha_Q/4$	R\$ 15.283	R\$ 7.905
$\alpha_{\text{AML}} = \alpha_Q$	R\$ 61.132	R\$ 31.620
Resistência interna inicial	0,08 Ω	0,025 Ω
<i>Parameter set</i> PyBaMM	Chen2020	Prada2013

Tabela 21 – Coeficientes da função de recompensa.

Coeficiente	Valor
Peso do lucro w_{profit}	1,0
Peso da degradação w_{deg}	1,0
Coeficiente de ansiedade λ_{anx}	3,0 R\$/kWh
Coeficiente de falha λ_{fail}	50 R\$/kWh
Coeficiente térmico β_T	0,01 R\$/(K ² ·h)
Temperatura ótima T_{opt}	25 °C

Tabela 22 – Submodelos do PyBaMM ativados por química.

Submodelo (opção PyBaMM)	NMC-811	LFP
SEI	<i>solvent-diffusion limited</i>	<i>solvent-diffusion limited</i>
SEI film resistance	<i>distributed</i>	<i>distributed</i>
lithium plating	<i>reversible</i>	<i>reversible</i>
particle mechanics	<i>swelling and cracking</i>	não ativado
loss of active material	<i>stress-driven</i>	não ativado
thermal	<i>lumped</i>	<i>lumped</i>

A integração numérica usa o *solver* IDAKLU do PyBaMM, com tolerâncias absoluta e relativa de 10^{-8} e discretização radial de 30 nós nos domínios de partícula.

Tabela 23 – Parâmetros do *patch* de degradação e do modelo térmico.

Parâmetro	Valor
SEI solvent diffusivity D_{sol} (NMC-811)	$8,3 \times 10^{-23} \text{ m}^2/\text{s}$
SEI solvent diffusivity D_{sol} (LFP)	$3,6 \times 10^{-23} \text{ m}^2/\text{s}$
Bulk solvent concentration c_{sol}	2.636 mol/m^3
SEI partial molar volume $\bar{V}_{\text{SEI}}$	$9,585 \times 10^{-5} \text{ m}^3/\text{mol}$
SEI resistivity ρ_{SEI}	$5,714 \times 10^3 \Omega \cdot \text{m}$ (TAHMASBI <i>et al.</i> , 2017)
Lithium plating kinetic rate constant	$1 \times 10^{-10} \text{ m/s}$
Lithium plating transfer coefficient	0,65
Dead lithium decay constant	$1 \times 10^{-6} \text{ s}^{-1}$
Negative electrode critical stress (grafite)	60 MPa
Positive electrode critical stress (NMC-811)	150 MPa
Young's modulus E (cátodo NMC-811 / grafite)	375 / 15 GPa
Poisson's ratio ν (cátodo / grafite)	0,20 / 0,30
LAM rate β_{LAM}	$2,78 \times 10^{-7} \text{ s}^{-1}$
Convection coefficient h (modelo <i>lumped</i>)	$10 \text{ W}/(\text{m}^2 \cdot \text{K})$

Tabela 24 – Discretização do *grid* 4D da *lookup table*.

Eixo	Mínimo	Máximo	Passo	Pontos
Estado de carga (SoC)	0,00	1,00	0,05	21
Taxa C	-1,0	+1,0	0,1	21
Temperatura	263,15 K	313,15 K	5 K	11
Estado de saúde (SoH)	0,80	1,00	0,05	5

O produto dos quatro eixos resulta em $21 \times 21 \times 11 \times 5 = 24.255$ pontos por química, cada um simulado de forma independente no PyBaMM conforme descrito no Capítulo 3.

A Tabela 25 reúne os parâmetros da distribuição Weibull de distância por viagem usados no ambiente estocástico (§3.4.2). Os pares de forma e escala (k_i, λ_i) de cada viagem são obtidos do coeficiente de variação σ_i/μ_i por busca de raiz (método de Brent).

Tabela 25 – Parâmetros da distribuição Weibull de distância por viagem.

Perfil	Viagem	$\mu \pm \sigma$ (km)	k (forma)	λ (escala)
Conservador	1	15 ± 5	3,30	16,7
Conservador	2	10 ± 5	2,10	11,3
Flexível	1	20 ± 15	1,35	21,8
Flexível	2	25 ± 10	2,70	28,1
Intenso	1	90 ± 20	5,17	97,8
Intenso	2	110 ± 30	4,13	121,1