Lecture 12

LLMs, RLHF, DPO, Instruct GPT

Core Challenge with Language Models:

- **GPT models predict the next token based on historical data, lacking an innate ability to follow instructions.**

- GPT (Generative Pre-trained Transformer) models, at their core, predict the next word or token in a sequence based on the probabilities derived from pre-training on extensive text corpora.
- They don't inherently "understand" instructions or follow commands but generate what's statistically likely to come next, given their training.

The Challenge with Large Language Models (LMs)

- **Capability:** LMs can be prompted to perform a range of NLP tasks.
- **Issue:** They often exhibit unintended behaviors:
  - Making up facts
  - Generating biased or toxic text
  - Not adhering to user instructions
- **Reason:** The language modeling objective (predicting the next token) is misaligned with the goal of "following user's instructions helpfully and safely."

Alignment in Language Models

- **Objective**: Train LMs to act in accordance with the user's intention.
- **Explicit Intentions**: Following instructions.
- **Implicit Intentions**: Staying truthful, avoiding bias, toxicity, or harm.

Core Principles for Aligned LMs:

- **Helpful**: Assist users effectively.
- **Honest**: Offer accurate, non-misleading information.
- **Harmless**: Prioritize user safety and avoid harm.

The Goal of Alignment:

- **The aim is to align GPT's responses with specific user instructions and ethical standards, beyond just generating probable text.**

- The primary objective is to bridge the gap between these statistical predictions and meaningful adherence to instructions provided by users.
- This involves ensuring that the AI's responses are not just contextually appropriate or conversationally relevant but also aligned with the specific intentions, ethical expectations, and task-oriented goals of the user.

Why It Matters:

- **This alignment is crucial for enhancing GPT's reliability, ensuring it respects user intent and ethical norms.**

- Without this alignment, while a model like GPT might produce grammatically correct and contextually relevant content, it might diverge from user instructions or produce content that's inappropriate or misaligned with the user's ethical, cultural, or personal expectations.
- The goal is to enhance the model's reliability in following directives accurately, respecting ethical boundaries, and fulfilling the user's actual needs, thereby making the technology more trustworthy and effective in real-world applications.

Modifying Language Model Behavior

- Multiple strategies have been developed to refine the generation behavior of language models.

Filtering Pretraining Datasets

- **What**: Removing or excluding certain documents or content from the dataset used for initial training.
- **Why**: Prevents the model from learning harmful or undesirable patterns present in those documents.
- **Example**: Excluding documents that contain hate speech or misinformation to reduce the chances of the model generating such content.

Fine-Tuning on Value-Targeted Datasets

- **What**: Adjusting the model's parameters using a dataset that emphasizes specific values or behaviors.
- **Why**: Helps the model better adhere to desired values in its outputs.
- **Example**: If we want a model to generate environmentally-conscious responses, we might fine-tune it on a dataset full of eco-friendly content.

Key Areas of Focus in AI Alignment

1.Learning from Human Feedback:

- Tailoring AI through human interaction.

2. Training to Follow Instructions:

- Teaching AI specific task adherence.

3. Evaluating AI Harms:

- Identifying risks in AI outputs.

4. Modifying Behavior to Mitigate Harms:

- Adjusting AI to prevent negative impacts.

Training Strategies for GPT Models: RLHF vs. DPO

Supervised Fine-Tuning (SFT)

- Fine-tunes pre-trained models on curated datasets.

Reinforcement Learning from Human Feedback (RLHF)

- **Introduced (2017):** Involves a reward model and reinforcement learning to align behavior with human intent.

Direct Preference Optimization (DPO)

- **Introduced (2023):** A newer method that eliminates the need for a reward model, directly optimizing for human preferences with improved efficiency.

Addressing GPT Challenges – RLHF

- To address inherent challenges with GPT models, ChatGPT's training strategy mirrors the "Instruct GPT" approach, itself an amalgamation of strategies from preceding works.
- **The Three-Phased Training Approach:**

1.Supervised Fine-Tuning (SFT):

- "Refines a pre-trained GPT-3 model's responses for specific tasks or guidelines, enhancing its understanding and output relevance.“

2. Training a Reward Model (RM):

- "Develops a system that assesses the quality of text generated by the model, guiding it towards human-preferred responses."

3. Reinforcement Learning from Human Feedback (RLHF)

The Three-Phased Training Approach

1.Supervised Fine-Tuning (SFT):

- Refines a pre-trained GPT-3 model's responses for specific tasks or guidelines, enhancing its understanding and output relevance.

2.Training a Reward Model (RM):

- Develops a system that assesses the quality of text generated by the model, guiding it towards human-preferred responses.

3. Reinforcement Learning from Human Feedback (RLHF)

- Refining AI behavior through direct human feedback.

1. Supervised Fine-Tuning (SFT)

Data Collection Approach

- **Source**: OpenAI API with InstructGPT.
- **Human-in-the-loop**: Integral part of the process to ensure quality and diversity.
- **Goal**: Capture a broad spectrum of language patterns and ensure diversity in prompts.

Data Collection Approach

- **Writing Prompts & Responses**:

- Labelers generate diverse prompts and craft human-written responses.

Introduction to Labeler-Written Prompts

- **Objective**: Train the initial InstructGPT model with instruction-like prompts.
- **Source**: Prompts crafted by contractors (labelers).
- **Types of Prompts**: Plain, Few-shot, User-based.

Types of Labeler-Written Prompts

1. **Plain Prompts**:
    1. Goal: Generate arbitrary tasks ensuring diversity.
    2. Example: "Describe the process of evaporation."
2. **Few-shot Prompts**:
    1. Goal: Provide an instruction with multiple query/response pairs.
    2. Example:
        1. **Instruction**: "Give the sentiment for a tweet."
        2. **Queries**: Sample tweets.
        3. **Responses**: "Positive" or "Negative."

Types of Labeler-Written Prompts

3.User-based Prompts:

- Derived from real-world applications to the OpenAI API.
- Reflect specific needs or interests of potential users.
- Example:
  - Original Application: "I want to use the API for a bird identification app."
  - User-based Prompt: "Describe characteristics of a common bird."

User-based vs Plain Prompts

- Both prompts are crafted by labelers.
- User-based prompts are inspired by real-world applications, ensuring they align with practical needs.
- Plain prompts are more open-ended and not tied to any specific application.

Illustrative User Prompts

1. **Brainstorming**
   Example: "List five ideas for how to regain enthusiasm for my career."
2. **Classification**
   Example: "Given the following text, rate, on a scale from 1-10, how sarcastic the person is being."
3. **Extract**
   Example: "Extract all course titles from the table below:
   |Title|Lecturer|Room|"
4. **Generation**
   Example: "Write a short story where a brown bear goes to the beach, makes friends with a seal, and then returns home."

Illustrative User Prompts

5. Chat

Example: "This is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly. Human: 'Hello, who are you?' AI: 'I am an AI created by OpenAI. How can I help you today?’”

6. Rewrite

Example: "Translate this sentence to Spanish: <English sentence>"

7. Closed QA

Example: "Answer the following question: What shape is the earth?

A) A circle B) A sphere C) An ellipse D) A plane"

Illustrative User Prompts

8. Open QA

Example: "Who built the statue of liberty?"

9. Summarization

Example: "Summarize this for a second-grade student: {text}"

10. Other

Example: "Lookup 'cowboy' on Google and give me the results."

Supervised Fine-Tuning (SFT)

- **Objective**: Train the model to generate desired responses to given prompts.
- **Method**:
  - **Concatenate the prompt and the desired response.**
  - Use this concatenated text as input for the model.
  - Train the model to predict the next token in the sequence.

Supervised Fine-Tuning (SFT)

- **Example**:
  - **Prompt**: "What's the capital of France?"
  - **Desired Response**: "Paris."
  - **Concatenated Input**: "What's the capital of France? Paris."
- **Outcome**: The model learns to generate concise and accurate responses to a wide range of prompts.

2. Training a Reward Model (RM)

Dataset Creation for RM

1. **Generating Responses**: For a given prompt, produce multiple responses using the SFT model.
2. **Pairing & Ranking**: Create pairs from these responses and have a human rank the better response in each pair.

- **Prompt**: "What is ice?"
- **Responses**:
  - A. "It's the solid form of water."
  - B. "Frozen water that's cold to touch."
  - C. "Water that has turned solid due to low temperatures."
  - D. "A crystalline substance formed when water freezes."

Dataset Creation for RM

Create Response Pairs: Pair the responses with each other.

- Total Pairs: $C(4,2) = 6$ pairs
- Example Pairs:
  1. (A, B)
     A: It's the solid form of water.
     B: Frozen water that's cold to touch.

  2. (A, C)
  3. (A, D)
  4. (B, C)
  5. (B, D)
  6. (C, D)

Human Ranking: Ask a human to rank responses within each pair.

- For pair (A, B), the human might prefer response A over B.

"It's the solid form of water."

"Frozen water that's cold to touch."

"It's the solid form of water."

"Frozen water that's cold to touch."

"Water that has turned solid due to low temperatures."

"A crystalline substance formed when water freezes."

"Water that has turned solid due to low temperatures."

"A crystalline substance formed when water freezes."

- **Total Prompts Used**: 33,000
- **Responses per Prompt**: Between 4 to 9
- **Sample Calculation**:
- For 4 responses per prompt: $33{,}000 \times C(4,2) = 198{,}000$ pairs
- For 9 responses per prompt: $33{,}000 \times C(9,2) = 1{,}188{,}000$ pairs

Reward Model Architecture

- **Starting Point**: Use the SFT model from Phase 1.

Modifications:

  - Remove the last linear layer (unembedding layer).
  - Add a randomly initialized linear layer.
  - The model now outputs a scalar value, effectively becoming a regressor.

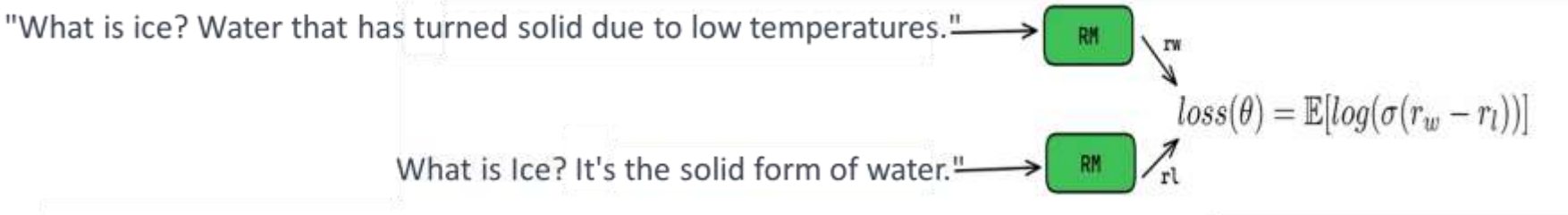
Training Regime for Reward Model

- **Pair Selection**:

- **Losing Pair**: "What is Ice? It's the solid form of water."
- **Winning Pair**: "What is ice? Water that has turned solid due to low temperatures."

Reward Calculation

1. **Winning Reward**: Pass the winning prompt-response to the reward model.
2. **Losing Reward**: Pass the losing prompt-response to the same model.
3. **Difference**: Calculate the difference between the two rewards.



Loss Function

$$\text{loss}(\theta) = \mathbb{E}\left[\log\left(\sigma(r_w - r_l)\right)\right]$$

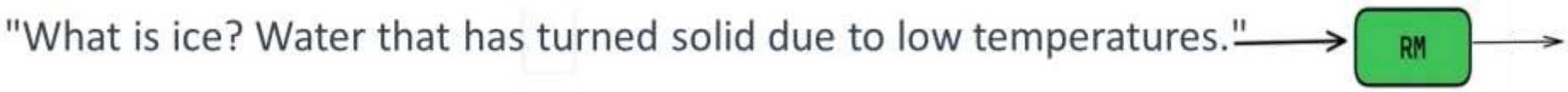
where

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

- The loss represents the difference in log-odds of a high reward (winning) and low reward (losing) response.
- Intuitively, the model aims to maximize the difference between the winning and losing responses.

Inference with Reward Model

- **Post-Training**: For any prompt-response combination, get a scalar reward.
- **Purpose**: Simulate human preference for a given prompt-response.



3. Reinforcement Learning from Human Feedback (RLHF)

Introduction to Reinforcement Learning Model (RL)

- **Objective**: Train a model that generates text aligning with human preferences.
- **Components**:
  - SFT Model: Generates responses but may not align with human preferences.
  - Reward Model: Provides a scalar reward but doesn't generate text.
- **Goal of RL**: Combine the capabilities of the above models to produce text that maximizes the reward.

Reinforcement Learning Definition

Reinforcement Learning (RL) is a type of machine learning in which an agent learns to make decisions by interacting with an environment, receiving rewards or penalties for its actions, and improving its behavior over time to maximize its total cumulative reward.

Reinforcements

- We use reinforcements to train animals
  - Food: (Positive reinforcements)
  - Hunger (Negative reinforcements)

Tic-tac-toe

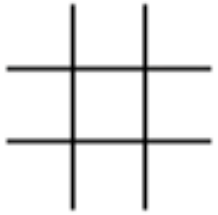


Tic-tac-toe

- Environment
- Action
- State
- Reward

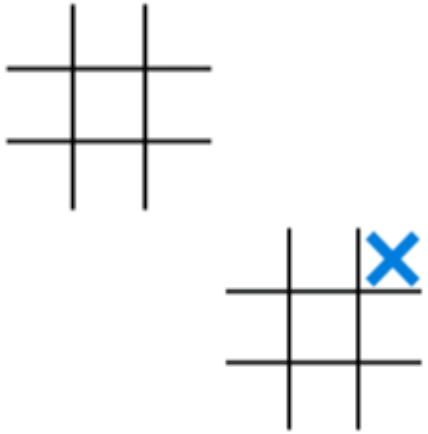
Tic-tac-toe

- Environment



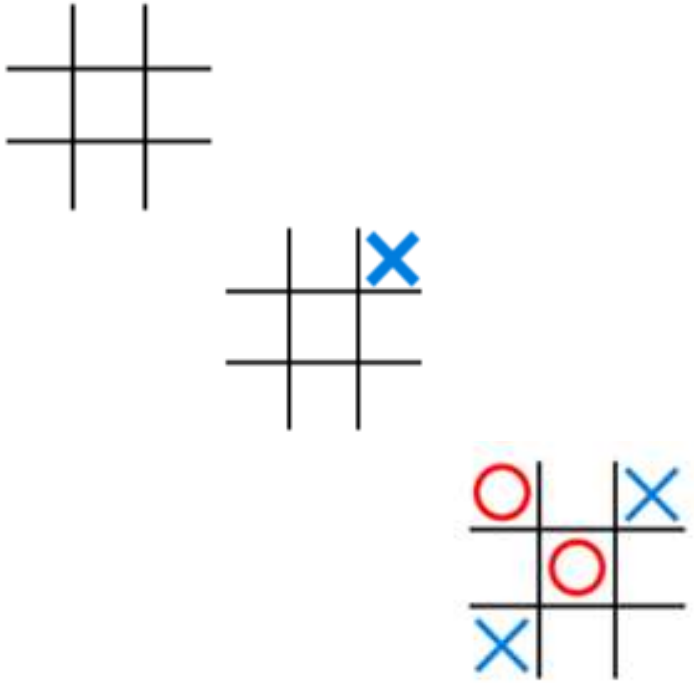
Tic-tac-toe

- Environment
- Action



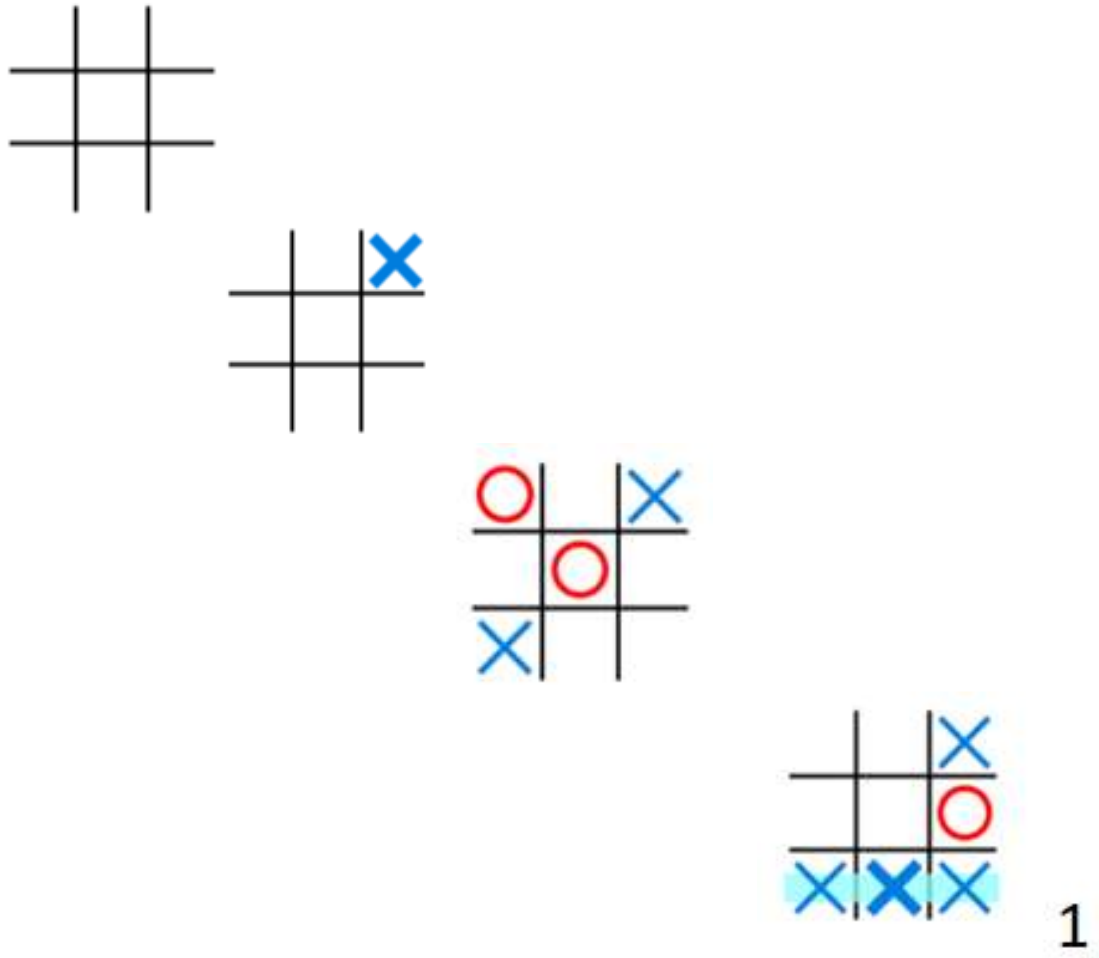
Tic-tac-toe

- Environment
- Action
- State



Tic-tac-toe

- Environment
- Action
- State
- Reward



Policy gradient

- Q-learning
  - Model-free value-based method

- Policy gradient
  - Model-free policy-based method

Deterministic policy vs Stochastic Policy

- Deterministic policy $a = \pi(s)$

- Stochastic policy $\pi_w(a|s) = \Pr(a|s)$

Supervised Learning

- We want to learn $\pi_w(a|s)$
- We have state-action pairs $\{(s_1, a_1^*), (s_2, a_2^*), \dots\}$

Supervised Learning

- We want to learn $\pi_w(a|s)$
- We have state-action pairs $\{(s_1, a_1^*), (s_2, a_2^*), \dots\}$
- Maximize log likelihood of the data

$$w^* = \underset{w}{\text{argmax}} \sum_n \log \pi_w(a_n^*|s_n)$$

$$w_{n+1} \leftarrow w_n + \alpha_n \nabla_w \log \pi_w(a_n^*|s_n)$$

Reinforcement Learning

- We want to learn $\pi_w(a|s)$

- We have state-action-reward triplets

  $\{(s_1, a_1, r_1), (s_2, a_2, r_2), \dots\}$

Reinforcement Learning

- We want to learn $\pi_w(a|s)$

- We have state-action-reward triplets

  $\{(s_1, a_1, r_1), (s_2, a_2, r_2), \dots\}$

- Maximize discounted sum of rewards

$$w^* = \underset{w}{\text{argmax}} \sum_n \gamma^n E_w[r_n|s_n, a_n]$$

Policy Gradient

- Maximize discounted sum of rewards

$$w^* = \underset{w}{\text{argmax}} \sum_n \gamma^n E_w[r_n | s_n, a_n]$$

Policy Gradient

- Maximize discounted sum of rewards

$$w^* = \underset{w}{\text{argmax}} \sum_n \gamma^n E_w[r_n | s_n, a_n]$$

$$w_{n+1} \leftarrow w_n + \alpha_n \nabla_w \log \pi_w(a_n^* | s_n)$$

Policy Gradient

- Maximize discounted sum of rewards

$$w^* = \underset{w}{\text{argmax}} \sum_n \gamma^n E_w[r_n|s_n, a_n]$$

$$w_{n+1} \leftarrow w_n + \alpha_n \nabla_w \log \pi_w(a_n^*|s_n)$$

$$w_{n+1} \leftarrow w_n + \alpha_n \gamma^n G_n \nabla_w \log \pi_w(a_n|s_n)$$

where $G_n = \sum_{t=0}^{\infty} \gamma^t r_{n+t}$

Policy Gradient

- **Objective**: Maximize the expected return of a stochastic, parameterized policy, π_w.
- **Expected Return**:

$$J(\pi_w) = \mathbb{E}_{\tau \sim \pi_w}[R(\tau)]$$

Where $R(\tau)$ is the total rewrad.

Gradient Ascent Optimization

- **Optimizing the Policy by Gradient Ascent**:

$$w_{k+1} = w_k + \alpha \nabla_w J(\pi_w)$$

- The gradient, $\nabla_w J(\pi_w)$, is the **policy gradient**(Vanilla Policy Gradien).

Useful Facts for Derivation

1. **Probability of a Trajectory**: Given a trajectory $\tau = (s_0, a_0, \ldots, s_{T+1})$ with actions from π_w:

$$P(\tau|w) = \rho_0(s_0) \prod_{t=0}^{T} P(s_{t+1}|s_t, a_t)\pi_w(a_t|s_t)$$

2. **The Log-Derivative Trick**: The derivative of $\log(u)$ is $\frac{\nabla u}{u}$. By rearrangement $\nabla u = u\nabla log(u)$:

$$\nabla_w P(\tau|w) = P(\tau|w)\nabla_w \log P(\tau|w)$$

Useful Facts for Derivation

3. **Log-Probability of a Trajectory**:

$$\log P(\tau|w) = \log \rho_0(s_0) + \sum_{t=0}^{T} (\log P(s_{t+1}|s_t, a_t) + \log \pi_w(a_t|s_t))$$

4. **Gradients of Environment Functions**: The environment has no dependence on w, so gradients of $\rho_0(s_0)$, $P(s_{t+1}|s_t, a_t)$, and $R(\tau)$ are zero.

Useful Facts for Derivation

- **Grad-Log-Prob of a Trajectory**:
  The gradient of the log-prob of a trajectory is:

$$\nabla_w \log P(\tau|w) = \cancelto{0}{\nabla_w \log \rho_0(s_0)} +$$

$$\sum_{t=0}^{T} \left(\cancelto{0}{\nabla_w \log P(s_{t+1}|s_t, a_t)} + \nabla_w \log \pi_w(a_t|s_t) \right)$$

Simplifying, we get:

$$\nabla_w \log P(\tau|w) = \sum_{t=0}^{T} \nabla_w \log \pi_w(a_t|s_t)$$

Basic Policy Gradient

$$
\begin{aligned}
\nabla_w J(\pi_w) &= \nabla_w \mathop{\mathrm{E}}_{\tau \sim \pi_w} [R(\tau)] \\
&= \nabla_w \int_\tau P(\tau \mid w) R(\tau) \\
&= \int_\tau \nabla_w P(\tau \mid w) R(\tau) \\
&= \int_\tau P(\tau \mid w) \nabla_w \log P(\tau \mid w) R(\tau) \\
&= \mathop{\mathbf{E}}_{\tau \sim \pi_w} [\nabla_w \log P(\tau \mid w) R(\tau)] \\
\therefore \nabla_w J(\pi_w) &= \mathop{\mathbf{E}}_{\tau \sim \pi_w} \left[\sum_{t=0}^{T} \nabla_w \log \pi_w (a_t \mid s_t) R(\tau) \right]
\end{aligned}
$$

Basic Policy Gradient

- The policy gradient is an expectation, which can be estimated via sample mean.
- Using trajectories $\mathcal{D} = \{\tau_i\}_{i=1,\dots,N}$ from the policy π_w, we get:

$$\hat{g} = \frac{1}{|\mathcal{D}|} \sum_{\tau \in \mathcal{D}} \sum_{t=0}^{T} \nabla_w \log \pi_w(a_t|s_t) R(\tau)$$

- $|\mathcal{D}|$ represents the number of trajectories (N in this case).
- This expression is our desired computable form.
- With a policy that allows $\nabla_w \log \pi_w(a|s)$ calculations and by collecting trajectory datasets, we can compute the gradient and update.

Aligning Generation with Human Preference

- The RL model's aim is to maximize the reward.
- This can be formulated mathematically as an objective maximization problem.
- Objective:

$$J(w) = \mathbb{E}_{x \sim p_w}[r(x)]$$

where:

- $J(w)$ is the objective function.
- x is the generated response.
- $r(x)$ is the reward for response x.

Training the RL Model

- Use gradient ascent to maximize the objective function.
- Gradient Ascent Update:

$$w \leftarrow w + \alpha \nabla_w J(w)$$

where:

- α is the learning rate.
- $\nabla_w J(w)$ is the gradient of the objective function.

With some mathematical manipulations, the gradient expression becomes:

$$\nabla_w J(w) = \mathbb{E}_{x \sim p_w} \left[r(x) \nabla_w \log p_w(x) \right]$$

$$= \frac{1}{D} \sum_{D} \sum_{t=0}^{T} r(x) \nabla_w \log p_w(x_t | x_{t-1}, x_{t-2}, \ldots)$$

where:

- D refers to all the prompt-response pairs.

Generating Text with the Model

1. The model generates a token x_t based on the previous tokens $x_0, x_1, \ldots, x_{t-1}$.
2. It continues to generate tokens until the sequence is complete.
3. After generating the entire sequence, the Reward Model (RM) evaluates it and assigns a reward $r(x)$.
4. This reward informs the gradient for the policy gradient update.

Note: During sequence generation, the model is unaware of the sequence's reward. It learns the reward only post-generation, using it to refine its parameters for better future sequences.

Model, Dataset, and Training Regime

- **Model**: The model structure remains similar to the SFT model but is trained with the new objective.
- **Dataset**: Use the same dataset but now with the goal of maximizing the reward.
- **Training Regime**: Iteratively generate responses, compute rewards, and update the model using gradient ascent.

Exploring the Reinforcement Learning Phase

Model:

- Copy the SFT model trained in phase 1.
- No modifications needed.

Dataset:

- A collection of prompts.
- No responses attached.

Training Regime

1. Feed a prompt to the model.
2. The model generates a corresponding response.
3. Concatenate the prompt and response.
4. Pass this concatenated pair to the reward model to obtain a reward score.
5. Use the reward and probabilities in the gradient equation.
6. Update the model parameters using this gradient.

Challenge with Direct Model Updates

- Using the initial training method can make the model unpredictable.
- These updates can push the model to produce text that doesn't make sense or is off-topic.

Solution: KL Divergence

- Ensure updates don't deviate too much from the SFT trained in phase 1.
- Use KL divergence to measure the "distance" between the SFT and the RL model.
- KL divergence compares two distributions.

Training Process

1. For a given prompt, the RL model generates a response.
2. At each generation step, a probability distribution over the vocabulary is produced.
3. Feed the same prompt-response to the SFT model to get its probability distribution over the vocabulary.
4. Calculate the KL divergence between the distributions from the RL model and the SFT model.
5. Subtract this divergence value from the reward for the prompt-response pair.

Subtract this divergence value from the reward

$$J(p_w) = \mathop{\mathbb{E}}_{x \sim p_w} [r(x) - \beta \mathrm{KL}(p_w(x), p_{SFT}(x))]$$

- $p_w(x)$ is the probability distribution over the vocabulary produced by the RL model for the sequence x.
- $p_{SFT}(x)$ is the probability distribution over the vocabulary produced by the original SFT model for the sequence x.
- β is a hyperparamete.

Issue with the Cost Function

Problem:

- The model's performance on some datasets was inferior to the original pretrained model.

Solution:

- Introduce an additional term in the cost function to keep the RL model close to the original pretrained base.

Updated Cost Function

$$J(p_w) = \mathop{\mathbb{E}}_{x \sim p_w} [r(x) - \beta \mathrm{KL}\left(p_w(x), p_{SFT}(x)\right)] + \gamma \mathop{\mathbb{E}}_{x \sim \text{pretrain}} [\log\left(p_w(x)\right)]$$

Where:

- p_w is the RL model with parameters w.
- γ is a hyperparameter, ensuring the RL model aligns with the original pretrained base.

In Conclusion

The model is optimized to:

- Maximize reward obtained using the reward model.
- Minimize divergence from SFT model.
- Maximize likelihood of observing sequences seen during pretraining.

After updating the RL model for some steps:

- Initialize a new reward model from it.
- Ask humans to generate another dataset following the procedure described in Phase 2.
- Conduct another training session for the reward model.

Addressing GPT Challenges – DPO

- **DPO**: Direct Preference Optimization
- **Two-Stage Training Approach**:
  1. **Supervised Fine-Tuning (SFT)**:
     - Refines a pre-trained model's responses for specific tasks or guidelines, enhancing its understanding and output relevance.
     - Similar to RLHF's SFT phase.
  2. **Direct Preference Optimization (DPO)**:
     - Fine-tunes the model to directly optimize preferred behavior using a pairwise comparison loss.

Core Concept of DPO

- Directly aligns the model's behavior with human preferences without using Reinforcement Learning.
- Uses **pairwise comparisons** between preferred and less-preferred responses for optimization.
- Avoids the need for a separate reward model.

How DPO Works

1. **Initialize**:
   - Start with a pre-trained model.
   - Keep a frozen reference model π_{ref} for relative comparison.
2. **Data**:
   - Collect human-labeled data as pairs (y_w, y_l) based on preference.
3. **Optimization**:
   - Minimize L_{DPO}, which directly optimizes preferred responses without introducing complex reward models.

DPO Loss Function

The DPO loss function is:

$$L_{DPO} = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right]$$

- x **(prompt)**: e.g., *What is Ice?*
- y_w **(preferred response)**: e.g., *It's the solid form of water.*
- y_l **(less preferred response)**: e.g., *Water that has turned solid due to low temperatures.*
- π_θ: Current model's output probabilities.
- π_{ref}: Reference model's frozen probabilities.
- β: Temperature scaling factor.

$$L_{DPO} = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right]$$

$$L_{DPO} = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right]$$

$$L_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log\sigma\left(\beta\log\frac{\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}}{\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}}\right)\right]$$

Advantages of DPO

- **Efficiency**: Avoids reinforcement learning, simplifying the fine-tuning process.
- **Stability**: Retains the model's prior knowledge while aligning with preferences.
- **Direct Optimization**: Fine-tunes preferences directly using supervised methods.
- **No Reward Model**: Reduces the complexity of training pipelines.