Lecture: Regret Policy Gradient

The Mathematical Bridge Between Actor-Critics and CFR

CSCE 631 — Intelligent Agents: Computational Game Solving

Alan Kuhnle

Today's Learning Goals

By the end of this lecture, you will understand:

1. Why standard policy gradients fail in adversarial games
2. The mathematical relationship: $q^{\sigma}(s, a) - v^{\sigma}(s) = \frac{r(s,a)}{B_{-i}(s)}$
3. How to derive RPG from this scaling relationship
4. Convergence guarantees in the tabular case
5. When and why RPG works in practice

Core insight: Actor-critic advantages are scaled counterfactual regrets [3]

Roadmap

Part 1: The Problem

- Why policy gradients cycle in games

Part 2: Mathematical Foundation

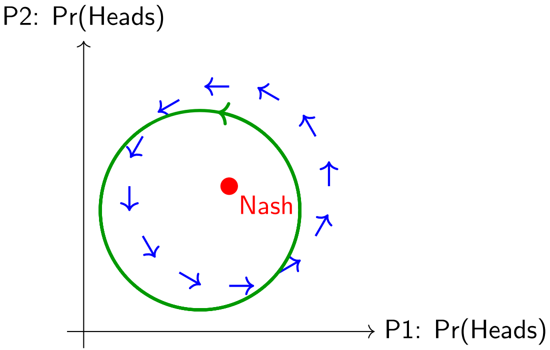
- Q-values vs. counterfactual values: full derivation
- The Bayes normalization constant $B_{-i}(s)$
- Advantages as scaled regrets

Part 3: RPG Algorithm

- Three variants: QPG, RPG, RMPG
- Convergence analysis (tabular case)
- Empirical results

Quick Reminder: Policy Gradients Cycle

Matching Pennies: Nash equilibrium is (0.5, 0.5) mixed strategy



Root cause: Non-stationarity—each player's environment changes as opponents learn [3]

What We Need: Regret Minimization

CFR converges to Nash by minimizing counterfactual regret

Question: Can we get Nash convergence with policy gradients?

Answer: Yes, if we connect actor-critic advantages to CFR regrets [3]

This lecture shows how.

Notation Review

Extensive-form game notation:

- $h \in H$: history (ground-truth state)
- s: information state (infoset) for player i
- σ: policy/strategy profile
- $\pi^\sigma(h) = \prod_{t<|h|} \sigma(s_t, a_t)$: reach probability under σ
- $\pi_i^\sigma(h)$: player i's contribution to reach probability
- $\pi_{-i}^\sigma(h)$: opponents' contribution (including chance)
- $\pi^\sigma(h) = \pi_i^\sigma(h) \cdot \pi_{-i}^\sigma(h)$

Key property (perfect recall):

$$\forall h, h' \in s, \quad \pi_i^\sigma(h) = \pi_i^\sigma(h') =: \pi_i^\sigma(s)$$

All histories in the same infoset have the same player i reach probability.

CFR Counterfactual Value (Review)

Definition:

$$v_i^c(\sigma, s, a) = \sum_{(h,z) \in Z(s,a)} \pi_{-i}^{\sigma}(h) \pi_i^{\sigma}(z) u_i(z)$$

where:

- $Z(s, a) = \{(h, z) \in H \times Z \mid h \in s,\ ha \sqsubseteq z\}$
- z: terminal history
- $u_i(z)$: utility at terminal

Infoset value:

$$v_i^c(\sigma, s) = \sum_a \sigma(s, a)\, v_i^c(\sigma, s, a)$$

Instantaneous regret:

$$r_i(\sigma, s, a) = v_i^c(\sigma, s, a) - v_i^c(\sigma, s)$$

RL Q-Value Definition

Standard Q-function:

$$q^{\sigma,i}(s,a) = \mathbb{E}_{\rho\sim\sigma}[G_t \mid S_t = s, A_t = a]$$

where $G_t = \sum_{t'=t}^{T} r_{t'}$ is the return.

Conditioning: We condition on *having reached* state s and taking action a.

Key difference from CFV:

- CFV conditions on player i *playing to reach* s and taking a
- Q-value conditions on the *event* of reaching s (however it happened)

The Scaling Relationship: Derivation (1/4)

Goal: Relate $q^{\sigma,i}(s,a)$ to $v_i^c(\sigma,s,a)$

Start with Q-value definition:

$$q^{\sigma,i}(s,a) = \mathbb{E}_{\rho\sim\sigma}[G_{t,i} \mid S_t = s, A_t = a] = \sum_{h\in s} \sum_{z\in Z(s,a)} \Pr(h \mid s)\, \pi^{\sigma}(ha, z)\, u_i(z)$$

Apply Bayes' rule:

$$\Pr(h \mid s) = \frac{\Pr(h)}{\Pr(s)} = \frac{\Pr(h)}{\sum_{h'\in s} \Pr(h')}$$

The Scaling Relationship: Derivation (2/4)

Substitute Bayes' rule:

$$q^{\sigma,i}(s,a) = \sum_{h,z \in Z(s,a)} \frac{\Pr(h)}{\sum_{h' \in s} \Pr(h')} \pi^{\sigma}(ha,z)\, u_i(z)$$

Note: $\Pr(h) = \pi^{\sigma}(h)$ and $\pi^{\sigma}(ha,z) = \pi^{\sigma}(h)\sigma(s,a)\pi_i^{\sigma}(z)$ where z is the continuation from *ha*.

Simplify:

$$q^{\sigma,i}(s,a) = \sum_{h,z \in Z(s,a)} \frac{\pi^{\sigma}(h)}{\sum_{h' \in s} \pi^{\sigma}(h')} \pi^{\sigma}(h)\sigma(s,a)\pi_i^{\sigma}(z)\, u_i(z)$$

The Scaling Relationship: Derivation (3/4)

Factor out reach probabilities:

$$\pi^\sigma(h) = \pi_i^\sigma(h) \cdot \pi_{-i}^\sigma(h) = \pi_i^\sigma(s) \cdot \pi_{-i}^\sigma(h)$$

(using perfect recall: $\pi_i^\sigma(h) = \pi_i^\sigma(s)$ for all $h \in s$)

Substitute:

$$\begin{aligned} q^{\sigma,i}(s,a) &= \sum_{h,z \in Z(s,a)} \frac{\pi_i^\sigma(s)\pi_{-i}^\sigma(h)}{\sum_{h' \in s} \pi_i^\sigma(s)\pi_{-i}^\sigma(h')} \pi_i^\sigma(s)\pi_{-i}^\sigma(h)\sigma(s,a)\pi_i^\sigma(z)\, u_i(z) \\ &= \sum_{h,z \in Z(s,a)} \frac{\pi_{-i}^\sigma(h)}{\sum_{h' \in s} \pi_{-i}^\sigma(h')} \pi_i^\sigma(s)\pi_{-i}^\sigma(h)\sigma(s,a)\pi_i^\sigma(z)\, u_i(z) \end{aligned}$$

Cancel $\pi_i^\sigma(s)$ terms.

The Scaling Relationship: Derivation (4/4)

Define the Bayes normalizing constant:

$$B_{-i}(\sigma, s) := \sum_{h \in s} \pi^{\sigma}_{-i}(h)$$

This is the total opponent reach probability to infoset s.

Final result:

$$\begin{aligned} q^{\sigma,i}(s,a) &= \frac{1}{B_{-i}(\sigma,s)} \sum_{h,z \in Z(s,a)} \pi^{\sigma}_{-i}(h)\pi^{\sigma}_{i}(z)u_i(z) \\ &= \frac{v^c_i(\sigma,s,a)}{B_{-i}(\sigma,s)} \end{aligned}$$

Similarly: $v^{\sigma,i}(s) = \frac{v^c_i(\sigma,s)}{B_{-i}(\sigma,s)}$

The Key Result

Theorem (Scaling Relationship, from [3])

] For any policy σ and infoset s:

$$q^{\sigma,i}(s,a) = \frac{v_i^c(\sigma,s,a)}{B_{-i}(\sigma,s)}, \quad v^{\sigma,i}(s) = \frac{v_i^c(\sigma,s)}{B_{-i}(\sigma,s)}$$

where $B_{-i}(\sigma,s) = \sum_{h\in s} \pi_{-i}^{\sigma}(h)$ is the opponent reach probability.

Immediate corollary:

$$\begin{aligned} a^{\sigma,i}(s,a) &= q^{\sigma,i}(s,a) - v^{\sigma,i}(s) \\ &= \frac{v_i^c(\sigma,s,a) - v_i^c(\sigma,s)}{B_{-i}(\sigma,s)} \\ &= \frac{r_i(\sigma,s,a)}{B_{-i}(\sigma,s)} \end{aligned}$$

Actor-critic advantages are scaled counterfactual regrets!

Interpretation of the Scaling Factor

What is $B_{-i}(\sigma, s)$?

- Sum of opponent reach probabilities over all histories in s
- In single-agent settings: $B_{-i}(\sigma, s) = 1$ (deterministic environment)
- In games: depends on how opponents play

When are advantages regrets?

1. $B_{-i}(\sigma, s) \approx 1$: opponent reach is near uniform
2. Single-agent: exactly equal
3. Deterministic transitions: exactly equal
4. Frequently visited states: B_{-i} is stable

Implication: Actor-critics implicitly do regret minimization, scaled by opponent behavior [3]

Example: Matching Pennies

At Nash equilibrium: Both players play (0.5, 0.5)

For Player 1:

- Both actions have same Q-value: $q^\sigma(s, \mathrm{H}) = q^\sigma(s, \mathrm{T}) = 0$
- Value: $v^\sigma(s) = 0$
- Advantage: $a^\sigma(s, \mathrm{H}) = a^\sigma(s, \mathrm{T}) = 0$

Counterfactual side:

- $v_i^c(\sigma, s, \mathrm{H}) = v_i^c(\sigma, s, \mathrm{T}) = 0$ (zero-sum, symmetric)
- $r(\mathrm{H}) = r(\mathrm{T}) = 0$

If P2 plays 70% Heads:

- P1 should play more Heads: $a^\sigma(s, \mathrm{H}) > 0$, $r(\mathrm{H}) > 0$
- Both frameworks detect the same signal!

Policy Gradient Theorem (Standard Form)

For maximizing $J(\sigma_\theta) = v^{\sigma_\theta}(s_0)$:

$$\nabla_\theta J(\sigma_\theta) \propto \sum_s \mu(s) \sum_a \nabla_\theta \sigma_\theta(s,a)\, q^{\sigma_\theta}(s,a)$$

where $\mu(s)$ is the on-policy state distribution.

Baseline-reduced form (actor-critic):

$$\nabla_\theta J(\sigma_\theta) \propto \sum_s \mu(s) \sum_a \nabla_\theta \sigma_\theta(s,a)\, (q^{\sigma_\theta}(s,a) - v^{\sigma_\theta}(s))$$

But wait: We just showed $q - v =$ scaled regret!

Q-Based Policy Gradient (QPG)

Rewrite using Q-based critic:

$$\nabla_\theta^{\text{QPG}}(s) = \sum_a \nabla_\theta \sigma(s, a; \theta) \left[q(s, a; w) - \sum_b \sigma(s, b; \theta) q(s, b; w) \right]$$

Interpretation:

- The term in brackets is the advantage: $a^\sigma(s, a)$
- From our derivation: $a^\sigma(s, a) = \frac{r(s,a)}{B_{-i}(s)}$
- So we're doing gradient ascent on scaled regret

This is standard actor-critic with all-action enumeration [3]

Regret Policy Gradient (RPG)

Motivation: CFR uses *thresholded* cumulative regret:

$$\sigma^{t+1}(a|s) \propto \max\left(0, \sum_{\tau=1}^{t} r_\tau(s,a)\right)$$

RPG gradient [3]:

$$\nabla_\theta^{\text{RPG}}(s) = -\sum_a \nabla_\theta \left[q(s,a;w) - \sum_b \sigma(s,b;\theta) q(s,b;w)\right]^+$$

where $(x)^+ = \max(0, x)$.

Key differences from QPG:

1. Negative sign: gradient *descent* on regret (instead of ascent on value)
2. Thresholding: only positive advantages contribute
3. Minimizes upper bound on cumulative regret

Regret Matching Policy Gradient (RMPG)

Alternative inspired by regret-matching weighting:

$$\nabla_\theta^{\text{RMPG}}(s) = \sum_a \nabla_\theta \sigma(s, a; \theta) \left[q(s, a; w) - \sum_b \sigma(s, b; \theta) q(s, b; w) \right]^+$$

Interpretation:

- Weight policy gradient by thresholded advantage
- Actions with positive regret get positive weight
- Actions with negative regret get zero weight

Trade-off: Most direct connection to regret matching, but empirically can plateau [3]

Relationship Between Variants

Mathematical connection (Appendix F of [3]): At equilibrium (when advantages are balanced):

$$\nabla_\theta^{\text{RPG}}(s) \propto \nabla_\theta^{\text{QPG}}(s)$$

Variant	**Update Rule**	**Connection to CFR**
QPG	Ascent on advantage	Scaled regret (no threshold)
RPG	Descent on thresholded advantage	Minimizes regret upper bound
RMPG	Weighted by thresholded advantage	Direct regret-matching analog

Training Setup

Architecture:

- Actor: $\sigma_\theta(a|s)$ (softmax output)
- Critic: $q_w(s,a)$ (outputs Q-value for each action)

Training loop:

1. Generate trajectory via self-play using σ_θ
2. For N_q steps: update critic via TD or Monte Carlo

$$w \leftarrow w - \alpha_c \nabla_w \left(q_w(s,a) - \hat{G}_t \right)^2$$

3. Update actor using chosen gradient (QPG/RPG/RMPG)

$$\theta \leftarrow \theta + \alpha_a \nabla_\theta$$

Key hyperparameters [3]: $N_q = 100$-1000; learning rates annealed; entropy regularization

Convergence: Tabular Case

Theorem (Theorem 1 from [3)

, simplified] In two-player zero-sum games with tabular policies, if:

- *Learning rate:* $\alpha_{s,k} = k^{-1/2}\pi_i^{\sigma}(s)B_{-i}(\sigma, s)$ *at iteration k*
- *Policy parameters projected to simplex*
- *All policies have positive support:* $\sigma_\theta(a|s) > 0$

Then projected actor-critic policy iteration has regret:

$$R_i^K \leq \frac{\sqrt{|S_i|}}{\pi_i^{\min}}\sqrt{K} + O(\sqrt{K})$$

This is $O(1/\sqrt{K})$ convergence—same rate as CFR!

Understanding the Learning Rate

Required learning rate: $\alpha_{s,k} = k^{-1/2}\pi_i^\sigma(s)B_{-i}(\sigma, s)$

Two components:

1. $k^{-1/2}$: standard decreasing rate for stochastic optimization
2. $\pi_i^\sigma(s)B_{-i}(\sigma, s)$: frequency-dependent scaling

Why this form?

- $\pi_i^\sigma(s)$: how often player i reaches s
- $B_{-i}(\sigma, s)$: scaling factor from our derivation
- Product: effective sampling frequency of (s, a) pairs

In practice: Use global annealed rate; on-policy sampling provides implicit weighting [3]

State-Local Gradients (Stronger Result)

Theorem (Theorem 2 from [3)

] Using state-local objectives:

$$\frac{\partial}{\partial \theta_{s,a}} J^{PG}(\sigma_\theta, s) = \frac{\partial v^{\sigma_\theta, i}(s)}{\partial \theta_{s,a}}$$

with learning rate $\alpha_k = k^{-1/2}$*, regret bound improves to:*

$$R_i^K \leq \sqrt{|S_i|}\sqrt{K} + O(\sqrt{K})$$

(no dependence on $\pi_i^{\min}$*)*

Intuition: Update each state's parameters based only on local value, not global objective

Trade-off: Stronger guarantee but requires tabular parameterization [3]

Function Approximation: The Gap

Challenge: Convergence theorems assume tabular policies.

With neural networks:

- No theoretical Nash guarantee
- Q-function approximation introduces bias
- Generalization can help or hurt
- Rare states may never be visited

Empirical observation [3]:

- RPG/QPG converge to low exploitability in practice
- Performance comparable to or better than NFSP
- Current policy often beats NFSP's average policy

Open question: Can we derive probabilistic bounds for the sampled, function-approximation case?

Proof Sketch: Why It Works

Key steps in convergence proof [3]:

1. **Regret decomposition:** Show actor-critic updates minimize a regret-like quantity

$$\text{advantage} = \frac{\text{regret}}{B_{-i}(s)}$$

2. **Projection analysis:** Projecting to simplex maintains regret bounds
3. **Variance control:** Learning rate schedule balances bias-variance
4. **Martingale argument:** Stochastic updates converge in expectation
5. **No-regret property:** Average policy converges to Nash

Critical assumption: Tabular + exact Q-values (or consistent estimates) [3]

Domains: Kuhn and Leduc Poker

Kuhn Poker:

- 3-card deck (J, Q, K); 2 or 3 players
- One betting round: Check/Bet, Fold/Call
- Simple but requires mixed strategies

Leduc Poker:

- 2-suit deck (6 cards for 2-player)
- Two betting rounds; public card after first
- Bet limits: 2 chips (round 1), 4 chips (round 2)
- Standard benchmark for multiagent RL [3]

Evaluation metric:

$$\text{NASH CONV}(\sigma) = \sum_i \left(\max_{\sigma_i'} \mathbb{E}_{\sigma_i', \sigma_{-i}}[G_{0,i}] - \mathbb{E}_{\sigma}[G_{0,i}] \right)$$

Measures exploitability (distance from Nash)

2-Player Leduc: Convergence

Results from [3]:

- **Short-term:** NFSP converges faster initially
- **Long-term:** RPG and QPG reach similar or lower NashConv
- **RMPG:** Tends to plateau at higher exploitability
- **A2C baseline:** Much slower (lacks regret structure)

Typical final NashConv after 20M steps:

- NFSP: $\sim$0.5
- QPG: $\sim$0.4
- RPG: $\sim$0.4
- A2C: $\sim$1.5

Conclusion: RPG/QPG competitive with NFSP using simpler architecture [3]

Performance vs. Fixed Opponents

Test: Evaluate learned policies vs. CFR500 (CFR with 500 iterations)

Results from [3]:

- **RPG:** Positive expected reward; beats CFR500 consistently
- **QPG:** Similar to RPG
- **NFSP:** Lower reward; average policy more conservative
- **A2C:** Negative reward; fails to learn robust strategies

Interpretation:

- RPG's *current* policy is stronger than NFSP's *average* policy against fixed bots
- On-policy learning produces more exploitative (but still robust) strategies

Caveat: Current policy may be more exploitable by adaptive opponents

3-Player Results

Challenge: No Nash guarantee for $n > 2$ player games

Findings [3]:

- RPG/QPG still converge to low exploitability
- NFSP also works but with higher variance
- No formal guarantees, but empirically effective

Open question: Can regret-based framework extend to n-player general-sum with guarantees?

Implementation Details

From [3]:

- **Architecture:** 2 FC layers, 128 units, ReLU
- **Optimizers:** Adam for both networks
- **Learning rates:**
  - Critic: fixed 10^{-3}
  - Actor: annealed from 10^{-3} to 0
- **Critic updates per policy update:** $N_q = 100\text{-}1000$
- **Reward normalization:** Z-score (streaming)
- **Temperature annealing:** $\tau : 1 \to 0$ over 1M steps
- **Entropy regularization:** $\beta = 10^{-3}$–10^{-2}

Why RPG Works in Practice

Despite lack of guarantees with function approximation:

1. **On-policy sampling:** Implicitly weights states by $\pi_i^\sigma(s)B_{-i}(s)$
2. **Q-network generalization:** Captures advantage patterns across similar states
3. **Regret structure:** Thresholding prevents runaway updates
4. **Entropy regularization:** Maintains exploration
5. **Multiple critic updates:** Reduces Q-function bias

Key insight: Don't need perfect regret estimates—just need to capture important strategic patterns [3]

Comparison: RPG vs. Deep CFR vs. NFSP

Property	**RPG**	**Deep CFR**	**NFSP**
On/off-policy	On-policy	Off-policy	Off-policy
Replay buffer	No	Yes (2)	Yes (2)
Theoretical guarantee	Tabular only	None	None
Regret connection	Explicit	Explicit	Implicit
Best response	Implicit	None	Explicit (DQN)
Implementation	Simple	Complex	Moderate
Convergence	$O(1/\sqrt{K})$ (tabular)	Empirical	Empirical

RPG sweet spot: Theoretical grounding + practical simplicity [3]

The Mathematical Journey

1. **Problem:** Policy gradients cycle in games (non-stationarity)
2. **Key insight:** Derive scaling relationship

$$q^{\sigma}(s, a) - v^{\sigma}(s) = \frac{r(s, a)}{B_{-i}(s)}$$

3. **Implication:** Actor-critics minimize scaled regret
4. **Algorithm:** Design RPG variants inspired by CFR's regret matching
5. **Theory:** Prove $O(1/\sqrt{K})$ convergence (tabular case)
6. **Practice:** Empirically effective with function approximation

Key Takeaways

1. **Mathematical connection:** Advantages = scaled regrets (exact relationship) [3]
2. **Algorithm design:** RPG inherits CFR's convergence properties (in tabular case)
3. **Practical advantage:** On-policy, model-free, simpler than alternatives
4. **Empirical success:** Competitive with NFSP in benchmark domains
5. **Open problem:** Extend guarantees to function approximation setting

Big picture: RPG demonstrates that regret minimization and policy gradients are fundamentally connected through the Bayes normalization constant [3]

References

[3] Srinivasan, S., Lanctot, M., et al. (2018). *Actor-Critic Policy Optimization in Partially Observable Multiagent Environments*. NeurIPS 2018.

- Zinkevich, M., et al. (2007). *Regret Minimization in Games with Incomplete Information*. NIPS (CFR).
- Heinrich, J., & Silver, D. (2016). *Deep Reinforcement Learning from Self-Play in Imperfect-Information Games* (NFSP).
- Brown, N., et al. (2019). *Deep Counterfactual Regret Minimization*. ICML.
- Sutton, R., & Barto, A. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press.

Code:

- OpenSpiel: `https://github.com/deepmind/open_spiel`

Thank you!

Questions?