Lecture 11: Learning to act

Samuel Gershman

Harvard University

Roadmap

- Optimizing actions through error-driven learning algorithms, implemented in striatal circuits that receive dopaminergic error signals.

Roadmap

- Optimizing actions through error-driven learning algorithms, implemented in striatal circuits that receive dopaminergic error signals.
- Cognitive costs of action selection (quantified using information theory). A cost-sensitive learning algorithm can explain the origin of habits and a range of other apparently suboptimal behaviors. It is also compatible with data on the sensitivity of dopamine neuron activity to cognitive cost.

Roadmap

- Optimizing actions through error-driven learning algorithms, implemented in striatal circuits that receive dopaminergic error signals.
- Cognitive costs of action selection (quantified using information theory). A cost-sensitive learning algorithm can explain the origin of habits and a range of other apparently suboptimal behaviors. It is also compatible with data on the sensitivity of dopamine neuron activity to cognitive cost.
- Balancing exploration and exploitation.

Policy optimization

- Recall that a policy is a distribution over action a conditional on state s.

Policy optimization

- Recall that a policy is a distribution over action a conditional on state s.
- $V^{\pi}(s)$ denotes the value (expected discounted future return) of state s under policy π:

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, \pi\right] = \sum_a \pi(a|s) Q^{\pi}(s,a)$$

where r_t is the reward received at time t and γ is a discount factor.

Policy optimization

- Recall that a policy is a distribution over action a conditional on state s.
- $V^{\pi}(s)$ denotes the value (expected discounted future return) of state s under policy π:

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, \pi\right] = \sum_a \pi(a|s) Q^{\pi}(s,a)$$

  where r_t is the reward received at time t and γ is a discount factor.
- $Q^{\pi}(s,a)$ is the state-action value function, representing the value of taking action a in state s.

Policy optimization

- Key to efficient reinforcement learning (RL) algorithms: **Markovian dynamics**, so the value function can be expressed recursively (the Bellman equation).

Policy optimization

- Key to efficient reinforcement learning (RL) algorithms: **Markovian dynamics**, so the value function can be expressed recursively (the Bellman equation).
- Here we assume transition dynamics are Markovian conditional on both states and actions, with a transition distribution $T(s'|s, a)$.

Policy optimization

- Key to efficient reinforcement learning (RL) algorithms: **Markovian dynamics**, so the value function can be expressed recursively (the Bellman equation).
- Here we assume transition dynamics are Markovian conditional on both states and actions, with a transition distribution $T(s'|s, a)$.
- Equipped with this assumption, the environment is known as a **Markov decision process** (MDP).

Policy optimization

- The policy optimization problem is to maximize the value:

$$\pi^* = \underset{\pi}{\text{argmax}} \sum_s \mu^\pi(s) V^\pi(s)$$

where $\mu^\pi(s) = \lim_{t\to\infty} p(s_t = s|s_0, \pi)$ is the stationary distribution of the MDP.

Policy optimization

- The policy optimization problem is to maximize the value:

$$\pi^* = \operatorname*{argmax}_{\pi} \sum_{s} \mu^{\pi}(s) V^{\pi}(s)$$

  where $\mu^{\pi}(s) = \lim_{t \to \infty} p(s_t = s | s_0, \pi)$ is the stationary distribution of the MDP.
- Solving this problem efficiently generally requires gradient-based algorithms.

Policy gradient algorithms

- Assume policy $\pi_\theta(a|s)$ is a differentiable function of parameters θ.

Policy gradient algorithms

- Assume policy $\pi_\theta(a|s)$ is a differentiable function of parameters θ.
- Optimization problem:

$$\theta^* = \operatorname*{argmax}_\theta \mathcal{J}(\theta)$$

where $\mathcal{J}(\theta) = \sum_s \mu^\pi(s) V^\pi(s)$ is the expected state value under the stationary distribution.

Policy gradient algorithms

- Assume policy $\pi_\theta(a|s)$ is a differentiable function of parameters θ.
- Optimization problem:

$$\theta^* = \underset{\theta}{\text{argmax}}\, \mathcal{J}(\theta)$$

  where $\mathcal{J}(\theta) = \sum_s \mu^\pi(s) V^\pi(s)$ is the expected state value under the stationary distribution.
- Gradient descent:

$$\Delta\theta \propto \nabla_\theta \mathcal{J}(\theta)$$

Policy gradient algorithms

- Policy gradient can be approximated as:

$$\nabla_\theta \mathcal{J}(\theta) \approx \delta \nabla_\theta \log \pi_\theta(a|s)$$

where $\delta = r + \gamma \hat{V}^\pi(s') - \hat{V}^\pi(s)$ is the temporal difference (TD) error.

Policy gradient algorithms

- Policy gradient can be approximated as:

$$\nabla_\theta \mathcal{J}(\theta) \approx \delta \nabla_\theta \log \pi_\theta(a|s)$$

  where $\delta = r + \gamma \hat{V}^\pi(s') - \hat{V}^\pi(s)$ is the temporal difference (TD) error.
- This is known as an *actor-critic* algorithm because it involves interplay between an actor (the policy) and a critic (the value function and TD error).

Neural implementation of action selection

- Thalamic neurons controlling movement initiation (via connections to premotor cortex) are under tonic inhibition from the output nuclei of the basal ganglia, the globus pallidus internal segment (GPi) and substantia nigra pars reticulata (SNr).

Neural implementation of action selection

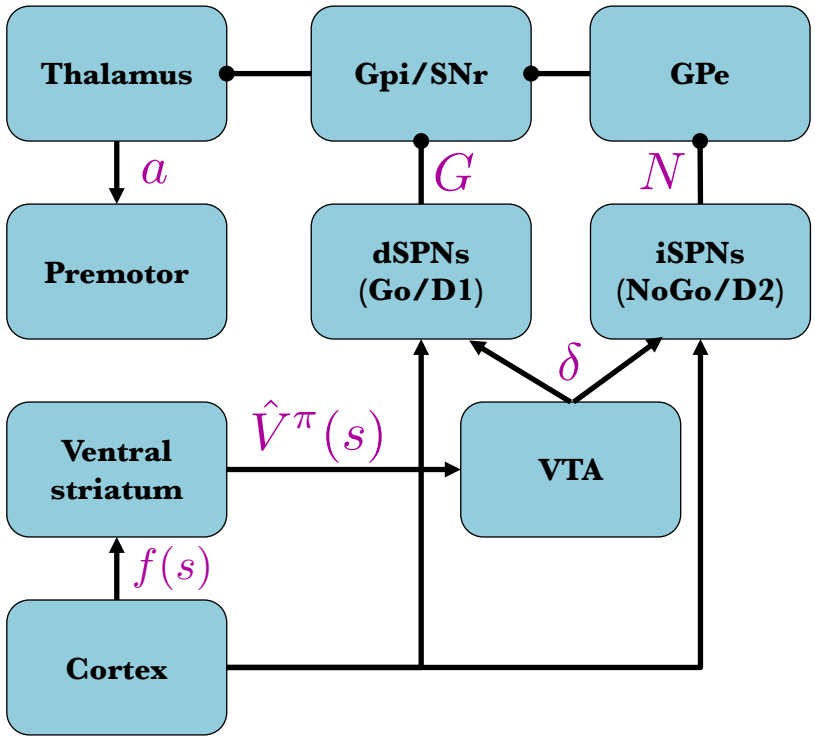
- Thalamic neurons controlling movement initiation (via connections to premotor cortex) are under tonic inhibition from the output nuclei of the basal ganglia, the globus pallidus internal segment (GPi) and substantia nigra pars reticulata (SNr).
- Movements are always "ready to go" upon disinhibition of the thalamus, when GPi and SNr are themselves inhibited by upstream structures in the basal ganglia: a "direct" pathway from the dorsal striatum (caudate and putamen), and an "indirect" pathway from the dorsal striatum through the globus pallidus external segment (GPe).

Neural implementation of action selection

- Thalamic neurons controlling movement initiation (via connections to premotor cortex) are under tonic inhibition from the output nuclei of the basal ganglia, the globus pallidus internal segment (GPi) and substantia nigra pars reticulata (SNr).
- Movements are always "ready to go" upon disinhibition of the thalamus, when GPi and SNr are themselves inhibited by upstream structures in the basal ganglia: a "direct" pathway from the dorsal striatum (caudate and putamen), and an "indirect" pathway from the dorsal striatum through the globus pallidus external segment (GPe).
- Direct ("Go") pathway promotes action production (disinhibition of the thalamus), whereas the indirect ("NoGo") pathway suppresses action production.

Action selection circuit

Excitatory connections are denoted by arrows; inhibitory connections are denoted by circles.



Neural implementation of action selection

- Within the dorsal striatum, separate populations of medium spiny neurons project to the direct and indirect pathways: the direct/indirect spiny projection neurons (dSPNs and iSPNs).

Neural implementation of action selection

- Within the dorsal striatum, separate populations of medium spiny neurons project to the direct and indirect pathways: the direct/indirect spiny projection neurons (dSPNs and iSPNs).
- Striatal policy parametrization:

$$\pi(a = j|s) \propto \exp\left[\alpha^G G_j - \alpha^N N_j\right]$$

where G_j is the input to the dSPNs ("Go" neurons) tuned to action j, and N_j is the input to the iSPNs ("NoGo" neurons).

Neural implementation of action selection

- Within the dorsal striatum, separate populations of medium spiny neurons project to the direct and indirect pathways: the direct/indirect spiny projection neurons (dSPNs and iSPNs).
- Striatal policy parametrization:

$$\pi(a = j|s) \propto \exp\left[\alpha^G G_j - \alpha^N N_j\right]$$

where G_j is the input to the dSPNs ("Go" neurons) tuned to action j, and N_j is the input to the iSPNs ("NoGo" neurons).

- Inputs are modeled as linear combinations of cortical state features, $x_d = f_d(s)$, where s is the state and $f_d(s)$ is the tuning function for cortical neuron d:

$$G_j = \sum_d \theta^G_{dj} x_d, \qquad N_j = \sum_d \theta^N_{dj} x_d$$

Policy parameters θ are corticostriatal synaptic strengths.

Neural implementation of policy learning

- The dSPNs express D1 dopamine receptors, whereas the iSPNs express D2 receptors. Dopamine has opposite effects on these neuron types, exciting dSPNs and inhibiting iSPNs.

Neural implementation of policy learning

- The dSPNs express D1 dopamine receptors, whereas the iSPNs express D2 receptors. Dopamine has opposite effects on these neuron types, exciting dSPNs and inhibiting iSPNs.
- Because D2 receptors have a higher affinity for dopamine, inhibitory effects of dopamine predominate at low concentrations.

Neural implementation of policy learning

- The dSPNs express D1 dopamine receptors, whereas the iSPNs express D2 receptors. Dopamine has opposite effects on these neuron types, exciting dSPNs and inhibiting iSPNs.
- Because D2 receptors have a higher affinity for dopamine, inhibitory effects of dopamine predominate at low concentrations.
- Because D2 receptors saturate at relatively low concentrations compared to D1 receptors, excitatory effects of dopamine predominate at high concentrations.

Neural implementation of policy learning

- The dSPNs express D1 dopamine receptors, whereas the iSPNs express D2 receptors. Dopamine has opposite effects on these neuron types, exciting dSPNs and inhibiting iSPNs.
- Because D2 receptors have a higher affinity for dopamine, inhibitory effects of dopamine predominate at low concentrations.
- Because D2 receptors saturate at relatively low concentrations compared to D1 receptors, excitatory effects of dopamine predominate at high concentrations.
- Dopamine signaling induces synaptic plasticity with opposite signs, promoting potentiation in dSPNs and depression in iSPNs, following a three-factor Hebbian rule (coincidence of presynaptic and postsynaptic firing with dopamine).

Neural implementation of policy learning

- Policy gradient for striatal parametrization:

$$\Delta\theta_{dj}^{G} \propto \alpha^{G}\delta x_d y_j, \qquad \Delta\theta_{dj}^{N} \propto -\alpha^{N}\delta x_d y_j$$

where $y_j = \mathbb{I}[a = j] - \pi_\theta(a = j|s)$ is an *action prediction error*; y_j is positive when action j occurs unexpectedly, and is negative when action j is expected but fails to occur.

Neural implementation of policy learning

- Policy gradient for striatal parametrization:

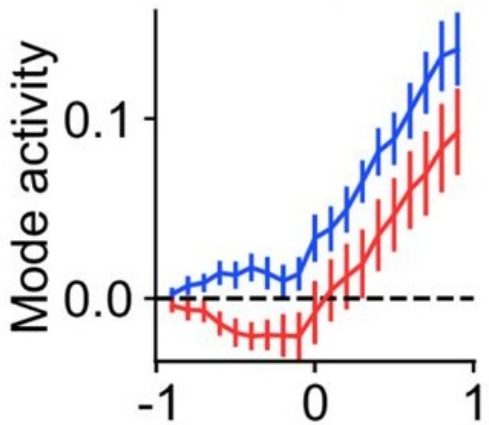
$$\Delta\theta^G_{dj} \propto \alpha^G \delta x_d y_j, \qquad \Delta\theta^N_{dj} \propto -\alpha^N \delta x_d y_j$$

  where $y_j = \mathbb{I}[a = j] - \pi_\theta(a = j|s)$ is an *action prediction error*; y_j is positive when action j occurs unexpectedly, and is negative when action j is expected but fails to occur.

- To interpret the updates as 3-factor Hebbian rules (presynaptic $\times$ postsynaptic $\times$ TD error), y_j must correspond to the postsynaptic (striatal) activity, and this must be the same for both dSPNs and iSPNs. Thus, dSPNs and iSPNs should be negatively correlated prior to a decision (since they push action selection in opposite directions), but should be positive correlated after action selection

SPN activity before and after action selection

On the X-axis, 0 is the time of action selection. Blue line shows activation of dSPNs associated with the selected action; red line shows activation of iSPNs.



[Lindsey et al 2025]

Opponency

- Tonic (slowly changing) levels of dopamine control the balance between direct and indirect pathways.

Opponency

- Tonic (slowly changing) levels of dopamine control the balance between direct and indirect pathways.
- Too much dopamine (e.g., Huntington's disease) $\rightarrow$ hyperkinesia (too much / too fast movement).

Opponency

- Tonic (slowly changing) levels of dopamine control the balance between direct and indirect pathways.
- Too much dopamine (e.g., Huntington's disease) $\rightarrow$ hyperkinesia (too much / too fast movement).
- Too little dopamine (e,g., Parkinson's disease) $\rightarrow$ hypokinesia (too little / too slow movement).

Chorea (hyperkinesia)

People with chorea (derived from the Greek word for "dance") experience rapid, intrusive movements

[La Médecine Illustrée 1880]

Sensitivity and tonic dopamine

- Tonic dopamine effects can be modeled via the sensitivity parameters.

Sensitivity and tonic dopamine

- Tonic dopamine effects can be modeled via the sensitivity parameters.
- Function relating ρ (tonic dopamine level) to sensitivity is based on the dose-occupancy functions for D1 and D2 receptors, combined with their postsynaptic effects:

$$\alpha^G = 1 + \tanh(\rho), \qquad \alpha^N = 1 - \tanh(\rho),$$

where $\tanh(\cdot)$ is the hyperbolic tangent function.

Sensitivity and tonic dopamine

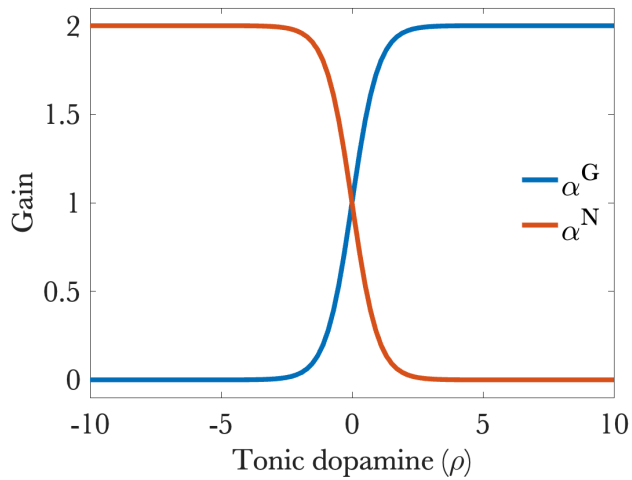
- Tonic dopamine effects can be modeled via the sensitivity parameters.
- Function relating ρ (tonic dopamine level) to sensitivity is based on the dose-occupancy functions for D1 and D2 receptors, combined with their postsynaptic effects:

$$\alpha^G = 1 + \tanh(\rho), \qquad \alpha^N = 1 - \tanh(\rho),$$

  where $\tanh(\cdot)$ is the hyperbolic tangent function.
- High levels of tonic dopamine amplify dSPNs and suppress iSPNs, whereas low levels suppress dSPNs and amplify iSPNs.

Sensitivity and tonic dopamine

Sensitivity for Go and NoGo components as a function of tonic dopamine.



Sensitivity and tonic dopamine

- Stimulating dSPNs produces hyperkinetic symptoms, whereas stimulating iSPNs produces hypokinetic symptoms [Kravitz et al 2010].

Sensitivity and tonic dopamine

- Stimulating dSPNs produces hyperkinetic symptoms, whereas stimulating iSPNs produces hypokinetic symptoms [Kravitz et al 2010].
- Similar effects observed in mice with genetic knockouts that selectively impair one of the pathways [Bateup et al 2010].

The role of the critic

- In the last lecture, we argued that the ventral striatum (nucleus accumbens) is responsible for encoding an approximation of the state value function, $\hat{V}$.

The role of the critic

- In the last lecture, we argued that the ventral striatum (nucleus accumbens) is responsible for encoding an approximation of the state value function, $\hat{V}$.
- Parameters of this function approximator are the synapses linking cortical inputs to neurons in the ventral striatum, updated by TD learning using the error signal δ conveyed by dopamine.

The role of the critic

- In the last lecture, we argued that the ventral striatum (nucleus accumbens) is responsible for encoding an approximation of the state value function, $\hat{V}$.
- Parameters of this function approximator are the synapses linking cortical inputs to neurons in the ventral striatum, updated by TD learning using the error signal δ conveyed by dopamine.
- Same TD error is hypothesized to update the policy parameters (synapses linking cortical inputs to the dorsal striatum).

Separate functions of ventral and dorsal striatum

- Division of labor between ventral (value) and dorsal (policy) striatum, consistent with studies showing that action preference signals are prevalent in dorsal striatum but are typically weak or absent in ventral striatum [e.g. Samejima et al 2005, Kim et al 2009].

Separate functions of ventral and dorsal striatum

- Division of labor between ventral (value) and dorsal (policy) striatum, consistent with studies showing that action preference signals are prevalent in dorsal striatum but are typically weak or absent in ventral striatum [e.g. Samejima et al 2005, Kim et al 2009].
- Ventral striatum, but not the dorsal striatum, should be active during classical conditioning, when value learning (but not policy updating) is engaged, whereas both regions should be active during instrumental conditioning (when both value and policy updating are engaged), consistent with evidence from human brain imaging [O'Doherty et al 2004].

Separate functions of ventral and dorsal striatum

- Division of labor between ventral (value) and dorsal (policy) striatum, consistent with studies showing that action preference signals are prevalent in dorsal striatum but are typically weak or absent in ventral striatum [e.g. Samejima et al 2005, Kim et al 2009].
- Ventral striatum, but not the dorsal striatum, should be active during classical conditioning, when value learning (but not policy updating) is engaged, whereas both regions should be active during instrumental conditioning (when both value and policy updating are engaged), consistent with evidence from human brain imaging [O'Doherty et al 2004].
- Error signal is uniform across ventral and dorsal striatum. This is consistent with recordings of dopamine neuron axons projecting to different parts of the striatum [Tsutsui et al 2020].

Tonic dopamine and motivation

- Motivational effects of tonic dopamine can be captured by assuming that it encodes average reward [Niv et al 2007].

Tonic dopamine and motivation

- Motivational effects of tonic dopamine can be captured by assuming that it encodes average reward [Niv et al 2007].
- High tonic dopamine $\rightarrow$ greater willingness to exert effort (Go $>$ NoGo; benefits of effort weighed more than costs).

Tonic dopamine and motivation

- Motivational effects of tonic dopamine can be captured by assuming that it encodes average reward [Niv et al 2007].
- High tonic dopamine $\rightarrow$ greater willingness to exert effort (Go $>$ NoGo; benefits of effort weighed more than costs).
- Dopamine-depleted rodents are disinclined to climb over a barrier to obtain a more desirable food reward [Salamone et al 1994].

Tonic dopamine and motivation

- Motivational effects of tonic dopamine can be captured by assuming that it encodes average reward [Niv et al 2007].
- High tonic dopamine $\rightarrow$ greater willingness to exert effort (Go $>$ NoGo; benefits of effort weighed more than costs).
- Dopamine-depleted rodents are disinclined to climb over a barrier to obtain a more desirable food reward [Salamone et al 1994].
- Dopamine fluctuations on the timescale of minutes covary with both average reward and response vigor [Hamid et al 2016].

Tonic dopamine and motivation

- Motivational effects of tonic dopamine can be captured by assuming that it encodes average reward [Niv et al 2007].
- High tonic dopamine $\rightarrow$ greater willingness to exert effort (Go $>$ NoGo; benefits of effort weighed more than costs).
- Dopamine-depleted rodents are disinclined to climb over a barrier to obtain a more desirable food reward [Salamone et al 1994].
- Dopamine fluctuations on the timescale of minutes covary with both average reward and response vigor [Hamid et al 2016].
- Under some assumptions, the time integral of TD errors (phasic dopamine) equals average reward (tonic dopamine).

Risk sensitivity

- Consider a choice between a risky option that delivers reward R with probability P (otherwise 0) and a safe option that always delivers reward $S < R$.

Risk sensitivity

- Consider a choice between a risky option that delivers reward R with probability P (otherwise 0) and a safe option that always delivers reward $S < R$.
- **Certainty equivalent**: value of S that would be required to make an agent indifferent between the risky and safe options.

Risk sensitivity

- Consider a choice between a risky option that delivers reward R with probability P (otherwise 0) and a safe option that always delivers reward $S < R$.
- **Certainty equivalent**: value of S that would be required to make an agent indifferent between the risky and safe options.
- **Risk premium**: difference between the expected payoff for the risky option (RP in this case) and the certainty equivalent. Quantifies how much an agent is willing to pay to avoid the risk—i.e., *risk aversion*.

Risk sensitivity

- Consider a choice between a risky option that delivers reward R with probability P (otherwise 0) and a safe option that always delivers reward $S < R$.
- **Certainty equivalent**: value of S that would be required to make an agent indifferent between the risky and safe options.
- **Risk premium**: difference between the expected payoff for the risky option (RP in this case) and the certainty equivalent. Quantifies how much an agent is willing to pay to avoid the risk—i.e., *risk aversion*.
- While humans are typically risk averse for positive outcomes, they tend to be risk seeking for negative outcomes, preferring risky over safe options with the same expected value.

Risk sensitivity in the brain

- Let the net drive for option j (Direct - Indirect) be denoted by $D_j = \alpha^G G_j - \alpha^N N_j$.

Risk sensitivity in the brain

- Let the net drive for option j (Direct - Indirect) be denoted by $D_j = \alpha^G G_j - \alpha^N N_j$.
- Asymptotically:

$$D_j \propto \mu_j + \beta\sigma_j, \qquad \beta = \frac{\alpha^G - \alpha^N}{\alpha^G + \alpha^N}$$

where $\mu_j \propto G_j - N_j$ is the expected reward for option j, and $\sigma_j \propto G_j + N_j$ is the reward standard deviation.

Risk sensitivity in the brain

- Let the net drive for option j (Direct - Indirect) be denoted by $D_j = \alpha^G G_j - \alpha^N N_j$.
- Asymptotically:

$$D_j \propto \mu_j + \beta\sigma_j, \qquad \beta = \frac{\alpha^G - \alpha^N}{\alpha^G + \alpha^N}$$

  where $\mu_j \propto G_j - N_j$ is the expected reward for option j, and $\sigma_j \propto G_j + N_j$ is the reward standard deviation.
- $\alpha^G > \alpha^N \rightarrow$ risk-seeking. $\alpha^G < \alpha^N \rightarrow$ risk-averse.

Risk sensitivity in the brain

- Recall that tonic dopamine is hypothesized to increase α^G and decrease α^N.

Risk sensitivity in the brain

- Recall that tonic dopamine is hypothesized to increase α^G and decrease α^N.
- Thus, we should see risk aversion at low levels of tonic dopamine and risk seeking at high levels.

Risk sensitivity in the brain

- Recall that tonic dopamine is hypothesized to increase α^G and decrease α^N.
- Thus, we should see risk aversion at low levels of tonic dopamine and risk seeking at high levels.
- Unmedicated Parkinson's patients (low tonic dopamine) are relatively more risk-averse than healthy controls, and this difference is eliminated by dopaminergic medication [Cherkasova et al 2019].

Risk sensitivity in the brain

- Recall that tonic dopamine is hypothesized to increase α^G and decrease α^N.
- Thus, we should see risk aversion at low levels of tonic dopamine and risk seeking at high levels.
- Unmedicated Parkinson's patients (low tonic dopamine) are relatively more risk-averse than healthy controls, and this difference is eliminated by dopaminergic medication [Cherkasova et al 2019].
- Medication can even cause pathological gambling, which is reduced after cessation of medication [Dodd et al 2005].

Risk sensitivity and average reward

- Under the hypothesis that tonic dopamine tracks average reward, risk seeking should increase with average reward.

Risk sensitivity and average reward

- Under the hypothesis that tonic dopamine tracks average reward, risk seeking should increase with average reward.
- Risk-seeking in wild chimpanzees (engaging in risky hunting rather than safe foraging) increases during periods of higher diet quality [Gilby & Wrangham 2007].

Risk sensitivity and average reward

- Under the hypothesis that tonic dopamine tracks average reward, risk seeking should increase with average reward.
- Risk-seeking in wild chimpanzees (engaging in risky hunting rather than safe foraging) increases during periods of higher diet quality [Gilby & Wrangham 2007].
- People shift from risk aversion toward risk seeking immediately following a meal [Symmonds et al 2010], a shift from low to high reward context [Rigoli et al 2016], and after a single prior gain [Thaler & Johnson 1990] or incidental positive outcome such as a win by the local sports team [Otto et al 2016].

Exploration

- **Exploration-exploitation dilemma**: to maximize long-term reward, an agent needs to balance exploring (gathering information about possibly low-reward actions) and exploiting (choosing actions that are believed to be best based on current estimates, which might be wrong).

Exploration

- **Exploration-exploitation dilemma**: to maximize long-term reward, an agent needs to balance exploring (gathering information about possibly low-reward actions) and exploiting (choosing actions that are believed to be best based on current estimates, which might be wrong).
- Optimal solution is intractable, but some heuristic solutions have theoretical guarantees.

Exploration

- **Exploration-exploitation dilemma**: to maximize long-term reward, an agent needs to balance exploring (gathering information about possibly low-reward actions) and exploiting (choosing actions that are believed to be best based on current estimates, which might be wrong).
- Optimal solution is intractable, but some heuristic solutions have theoretical guarantees.
- **Uncertainty bonus heuristic**: choose based on the sum of a mean reward estimate $\hat{\mu}_j$ and an uncertainty bonus $\beta\sigma_j$.

Exploration

- **Exploration-exploitation dilemma**: to maximize long-term reward, an agent needs to balance exploring (gathering information about possibly low-reward actions) and exploiting (choosing actions that are believed to be best based on current estimates, which might be wrong).
- Optimal solution is intractable, but some heuristic solutions have theoretical guarantees.
- **Uncertainty bonus heuristic**: choose based on the sum of a mean reward estimate $\hat{\mu}_j$ and an uncertainty bonus $\beta\sigma_j$.
- This looks a lot like the risk-reward equation we derived from the striatal parametrization, and it is!

Exploration

- **Exploration-exploitation dilemma**: to maximize long-term reward, an agent needs to balance exploring (gathering information about possibly low-reward actions) and exploiting (choosing actions that are believed to be best based on current estimates, which might be wrong).
- Optimal solution is intractable, but some heuristic solutions have theoretical guarantees.
- **Uncertainty bonus heuristic**: choose based on the sum of a mean reward estimate $\hat{\mu}_j$ and an uncertainty bonus $\beta\sigma_j$.
- This looks a lot like the risk-reward equation we derived from the striatal parametrization, and it is!
- However, the uncertainty bonus heuristic requires that $\beta \geq 0$ (risk seeking), whereas our earlier derivation allows $\beta < 0$ (risk aversion) if $\alpha^N > \alpha^G$. This will tend to happen early during learning, when average reward (and hence tonic dopamine) is low.

Exploration and novelty

- To solve this puzzle, we need some way of boosting tonic dopamine early during learning, so that the uncertainty bonus is positive during the period when exploration is needed.

Exploration and novelty

- To solve this puzzle, we need some way of boosting tonic dopamine early during learning, so that the uncertainty bonus is positive during the period when exploration is needed.
- **Novelty bonuses** offer a solution: boost tonic dopamine in response to novel stimuli, thereby promoting risk seeking ($\alpha^G > \alpha^N$). This boost will diminish with repeated exposure, eventually shifting toward risk aversion.

Exploration and novelty

- To solve this puzzle, we need some way of boosting tonic dopamine early during learning, so that the uncertainty bonus is positive during the period when exploration is needed.
- **Novelty bonuses** offer a solution: boost tonic dopamine in response to novel stimuli, thereby promoting risk seeking ($\alpha^G > \alpha^N$). This boost will diminish with repeated exposure, eventually shifting toward risk aversion.
- Dopamine is in fact boosted in response to novelty [Horvitz 200; Kutlu et al 2021].

Exploration and novelty

- Pharmacologically elevating dopamine increases novelty seeking [Costa et al 2014] and antagonizing D1 receptors reduces novelty seeking [Peters et al 2007].

Exploration and novelty

- Pharmacologically elevating dopamine increases novelty seeking [Costa et al 2014] and antagonizing D1 receptors reduces novelty seeking [Peters et al 2007].
- Consistent with the tonic dopamine average reward story, novelty seeking increases with average reward [Gershman & Niv 2015].

Exploration and novelty

- Pharmacologically elevating dopamine increases novelty seeking [Costa et al 2014] and antagonizing D1 receptors reduces novelty seeking [Peters et al 2007].
- Consistent with the tonic dopamine average reward story, novelty seeking increases with average reward [Gershman & Niv 2015].
- This is also true in the real world: novel restaurant choice increases with average restaurant quality in a local area, and this is amplified high variance, consistent with an uncertainty bonus [Schulz et al 2019].

Study question

Novelty responses in dopamine neurons may function as "uncertainty bonuses" for exploration. How does this mechanism help resolve the exploration-exploitation dilemma, and how might it fail under pathological conditions?

Policy compression

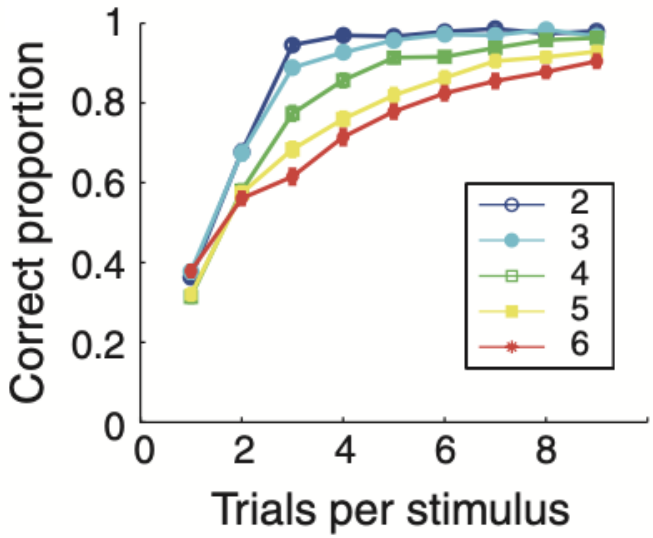
- Human performance declines with the number of states. Why?

Policy compression

- Human performance declines with the number of states. Why?
- Hypothesis: capacity constraint on policies.

Performance declines with set size

On each trial, subjects had to choose the correct action for a specific stimulus (indicating the state). The number of distinct stimuli in a block is the set size. Each curve shows the proportion of correct actions as a function of trial in a learning block for a given set size.



[Collins & Frank 2012]

Policy compression

- Capacity constraint may reflect a limit on the state encoding function $f(s)$; if this produces representations (x) that overlap across states, then there can potentially be confusion between states.

Policy compression

- Capacity constraint may reflect a limit on the state encoding function $f(s)$; if this produces representations (x) that overlap across states, then there can potentially be confusion between states.
- Alternatively, the mapping from representations to action probabilities may be constrained (e.g., the policy weights θ can't get too large).

Policy compression

- Capacity constraint may reflect a limit on the state encoding function $f(s)$; if this produces representations (x) that overlap across states, then there can potentially be confusion between states.
- Alternatively, the mapping from representations to action probabilities may be constrained (e.g., the policy weights θ can't get too large).
- More generally, we can quantify how state-dependent the policy is using the mutual information between states and actions, $\mathcal{I}[s; a]$. We will refer to this quantity as the **policy complexity** to capture the intuition that policies are more complex when they are more sensitive to variations in the state.

Policy compression

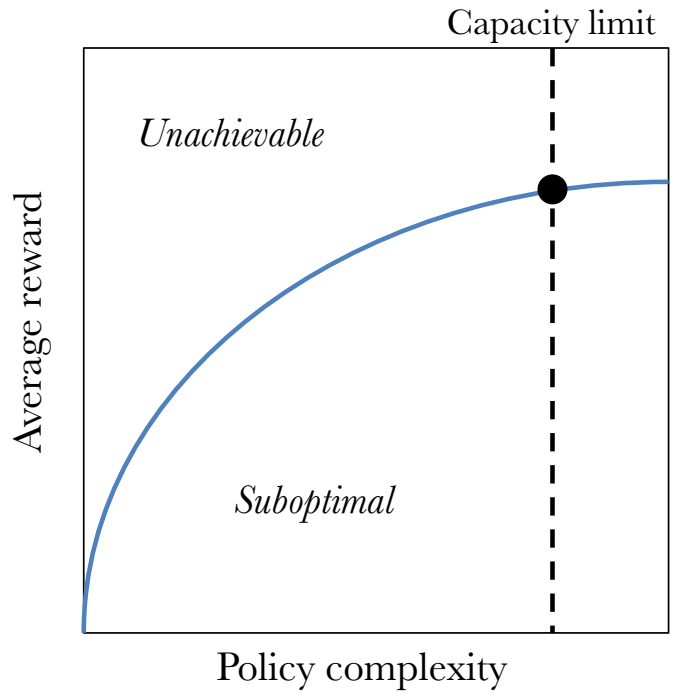
- Set size effects can be explained by imposing an upper bound (capacity limit) $\mathcal{C}$ on policy complexity.

Policy compression

- Set size effects can be explained by imposing an upper bound (capacity limit) $\mathcal{C}$ on policy complexity.
- We can then study the achievable average reward for a given capacity limit—the *reward-complexity frontier*.

Reward-complexity frontier

The curve show the reward-complexity frontier, separating unachievable from suboptimal policies. The circle shows the optimal achievable average reward for a given capacity limit (upper bound on policy complexity).



Policy compression

- Capacity-limited reward optimization problem as a Lagrangian:

$$\pi^* = \underset{\pi}{\text{argmax}}\, \rho(\pi) - \lambda c(\pi), \qquad \lambda = \frac{\partial \rho(\pi)}{\partial c(\pi)}$$

where we have expressed the average reward $\rho(\pi)$ as a function of the policy, and $c(\pi)$ is the complexity of π.

Policy compression

- Capacity-limited reward optimization problem as a Lagrangian:

$$\pi^* = \underset{\pi}{\text{argmax}}\, \rho(\pi) - \lambda c(\pi), \qquad \lambda = \frac{\partial \rho(\pi)}{\partial c(\pi)}$$

  where we have expressed the average reward $\rho(\pi)$ as a function of the policy, and $c(\pi)$ is the complexity of π.
- The parameter $\lambda \geq 0$ is a Lagrange multiplier that monotonically decreases with the capacity limit $\mathcal{C}$.

Policy compression

- Capacity-limited reward optimization problem as a Lagrangian:

$$\pi^* = \operatorname*{argmax}_{\pi} \rho(\pi) - \lambda c(\pi), \qquad \lambda = \frac{\partial \rho(\pi)}{\partial c(\pi)}$$

  where we have expressed the average reward $\rho(\pi)$ as a function of the policy, and $c(\pi)$ is the complexity of π.
- The parameter $\lambda \geq 0$ is a Lagrange multiplier that monotonically decreases with the capacity limit $\mathcal{C}$.
- In the limit $\mathcal{C} \to \infty$ (no bound on policy complexity), $\lambda \to 0$ and we recover average reward optimality.

Policy compression

- Optimal solution can be written explicitly:

$$\pi^*(a|s) \propto \exp\left[Q^\pi(s,a) + \lambda \log p^*(a)\right], \qquad p^*(a) = \sum_s \pi^*(a|s)p(s)$$

Policy compression

- Optimal solution can be written explicitly:

$$\pi^*(a|s) \propto \exp\left[Q^\pi(s,a) + \lambda \log p^*(a)\right], \qquad p^*(a) = \sum_s \pi^*(a|s)p(s)$$

- Standard softmax policy from first principles.

Policy compression

- Optimal solution can be written explicitly:

$$\pi^*(a|s) \propto \exp\left[Q^\pi(s,a) + \lambda \log p^*(a)\right], \qquad p^*(a) = \sum_s \pi^*(a|s)p(s)$$

- Standard softmax policy from first principles.
- Produces stochasticity even asymptotically and under perfect knowledge of rewards; reflects cognitive resource constraints rather than exploration (although it can induce useful exploration as a side effect).

Policy compression

- Response bias $p^*(a)$ in the softmax reflects frequently chosen actions across all states.

Policy compression

- Response bias $p^*(a)$ in the softmax reflects frequently chosen actions across all states.
- This bias only appears when capacity is limited ($\lambda > 0$). Human behavioral studies have shown that such a bias exists (a form of perseveration when the bias is updated online), and that the bias increases with set size [Lai & Gershman 2024], consistent with the idea that the bias arises from a limited resource that is shared across states.

Policy compression

- Response bias $p^*(a)$ in the softmax reflects frequently chosen actions across all states.
- This bias only appears when capacity is limited ($\lambda > 0$). Human behavioral studies have shown that such a bias exists (a form of perseveration when the bias is updated online), and that the bias increases with set size [Lai & Gershman 2024], consistent with the idea that the bias arises from a limited resource that is shared across states.
- For a similar reason, choice stochasticity increases with set size.

Learning compressed policies

- We can derive a learning algorithm for the capacity-limited optimization problem by noting that it is equivalent to:

$$\pi^* = \underset{\pi}{\text{argmax}}\, \mathbb{E}\left[r - \lambda \log \frac{\pi(a|s)}{p^*(a)}\right].$$

Learning compressed policies

- We can derive a learning algorithm for the capacity-limited optimization problem by noting that it is equivalent to:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}\left[r - \lambda \log \frac{\pi(a|s)}{p^*(a)} \right].$$

- Standard policy gradient algorithm can be used to find the optimal policy, simply by adding a complexity penalty (how much the state-dependent policy deviates from the action bias) to the rewards.

Learning compressed policies

- We can derive a learning algorithm for the capacity-limited optimization problem by noting that it is equivalent to:

$$\pi^* = \underset{\pi}{\text{argmax}}\, \mathbb{E}\left[r - \lambda \log \frac{\pi(a|s)}{p^*(a)}\right].$$

- Standard policy gradient algorithm can be used to find the optimal policy, simply by adding a complexity penalty (how much the state-dependent policy deviates from the action bias) to the rewards.
- Predicts that phasic dopamine signals encoding TD errors should be suppressed by policy complexity, as observed empirically [Gershman & Lak 2025].

Summary

- A biologically plausible policy parametrization, based on opponency in the direct and indirect pathways of the basal ganglia, can be used to learn optimal actions.

Summary

- A biologically plausible policy parametrization, based on opponency in the direct and indirect pathways of the basal ganglia, can be used to learn optimal actions.
- Tonic dopamine plays an important role in this architecture, governing both exploration during learning and asymptotic risk sensitivity.

Summary

- A biologically plausible policy parametrization, based on opponency in the direct and indirect pathways of the basal ganglia, can be used to learn optimal actions.
- Tonic dopamine plays an important role in this architecture, governing both exploration during learning and asymptotic risk sensitivity.
- Capacity-limited policy optimization by augmenting the reward function with a complexity penalty, which naturally explains behavioral stochasticity and perseveration, as well as the sensitivity of dopamine neuron activity to policy complexity.