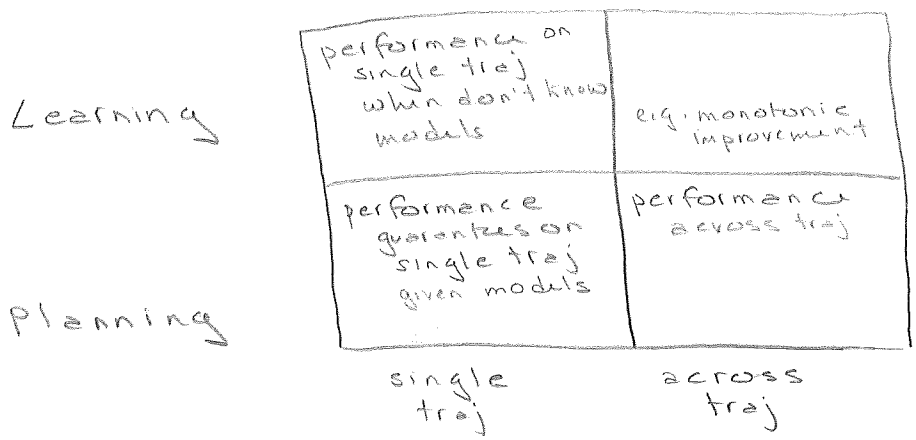
Risk, Safety & Beyond Expectation

may not just want to max expected total return as do RL

safety, risk



Phil Thomas's work: guaranteed monotonic improvement in expected total return of π deployed during learning

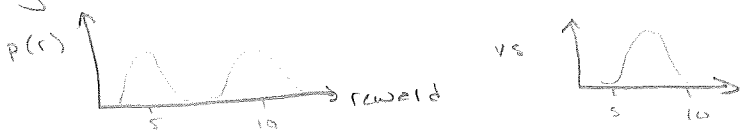
What type of guarantees would we like?

Beyond Expectation: Risk & Safety

so far only focussed on maximizing expected sum of rewards
- maybe bounds on that expectation, due to imprecise estimates of models or limited data to estimate expectation

but even if had ∞ data / perfect estimate of MDP parameters
- stochastic process
- still have uncertainty over reward obtain in a single trajectory

may care about properties of entire distrib of outcomes



what properties might we care about? examples from your / other application domains?

- worst case outcome
- median
- Markowitz mean variance (popular in finance lit)

$$= \max_{\pi} E^{\pi}[R] - \beta \, \mathrm{Var}^{\pi}[R]$$ where R is the total return from a traj

constrained var

$$= \max_{\pi} E^{\pi}[R] \text{ subj to } \mathrm{Var}^{\pi}[R] \leq \beta$$ gen NP hard to solve

sharpe-ratio obj

$$= \max_{\pi} \frac{E^{\pi}[R]}{\sqrt{\mathrm{Var}^{\pi}[R]}}$$

CVaR

Optimizing the CVaR via Sampling

CVaR = conditional value at risk

α-CVaR for random payoff R (whose distrib param by control param θ)
expected payoff over α% worst outcomes of $R(\theta)$

$$\Phi(\theta) = E^{\theta}[R \mid R \leq \underbrace{Va(\theta)}_{\alpha\text{-quantile of } R}]$$

CVaR optimization

$$\arg\max_{\theta} \Phi(\theta)$$

some domains $R = f_\theta(X)$
f_θ deterministic function
X random variable indep of θ
} can solve as stochastic program

examples?
portfolio optimization (for most individuals)

in RL, resource allocation, etc. θ impacts distrib of outcomes X
e.g. decisions/actions impact rewards received

contribution
derive gradient of CVaR when distrib of outcomes dep on θ
give a sampling based estimate for CVaR gradient
use to optimize CVaR by gradient descent

policy search method w/new objective: max CVaR instead of expected value

→ what choice of α makes optimizing CVaR equivalent to max expected value?
$\alpha = 100$

CVaR gradient for 1 dim var
random variable Z
prob density func of Z is $f_z(z,\theta)$ ← k-dim + cdf $F_z(z;\theta) = P(Z \leq z;\theta)$

(1) $Va_\alpha(Z;\theta) = \inf\{z : \int_0^z f_z(z,\theta) > \alpha\}$

(2) $\Phi_\alpha(Z;\theta) = E[Z \mid Z \leq Va(Z;\theta)]$

want $\partial/\partial\theta \; \Phi_\alpha(Z;\theta)$
analytic computation normally intractable
instead re-express as conditional expec & approxim by sampling

assumptions

1) Z continuous random var bounded in $[-b, b]$ $\forall \theta$
2) $\forall \theta,\ j = 1:k,$ $\partial V_\alpha(z;\theta)/\partial\theta_j$ and $\partial \bar{\Phi}(z;\theta)/\partial\theta_j$ exist & bounded
3) $\forall \theta,\ j = 1:k$ $\dfrac{\partial f_z(z;\theta)/\partial\theta_j}{f_z(z;\theta)} = \partial/\partial\theta_j \log f_z(z;\theta)$ exists & bounded

→ same as used in prior policy gradient lecture

proposition: Given assumptions 1-3,

$$\frac{\partial \bar{\Phi}_\alpha(z;\theta)}{\partial\theta_j} = E^\theta\left[\frac{\partial \log f_z(z;\theta)}{\partial\theta_j}\left(z - V_\alpha(z;\theta)\right) \,\middle|\, z \le V_\alpha(z;\theta)\right]$$

this is a conditional expectation

proof:

define level set $D_\theta = \{z \in [-b, b] : z \le V_\alpha(z;\theta)\}$
$= \{z \in [-b, V_\alpha(z;\theta)\}$

$\alpha = \int_{z \in D_\theta} f_z(z;\theta)\, dz$ by defn

$d/d\theta_j \alpha = d/d\theta_j \int_{-b}^{V_\alpha(z;\theta)} f_z(z;\theta)\, dz$ Leibniz rule

(3) $0 = \int_{-b}^{V_\alpha(z;\theta)} \partial f_z(z;\theta)/\partial\theta_j\, dz + \dfrac{\partial V_\alpha(z;\theta)}{\partial\theta_j} f_z(V_\alpha(z;\theta);\theta)$

recall $\bar{\Phi}_\alpha(z;\theta) = \int_{z \in D_\theta} f_z(z;\theta)\, z\, dz$
$= \int_{-b}^{V_\alpha(z;\theta)} f_z(z;\theta)\, z\, dz$

$d/d\theta_j \bar{\Phi}_z(z;\theta) = d/d\theta_j \int_{-b}^{V_\alpha(z;\theta)} f_z(z;\theta)\, z\, dz$ Leibniz rule

$= \int_{-b}^{V_\alpha(z;\theta)} d/d\theta_j f_z(z;\theta)\, z\, dz + \dfrac{\partial V_\alpha(z;\theta)}{\partial\theta_j} f_z(V_\alpha(z,\theta))\, V_\alpha(z;\theta)$

$= \int_{-b}^{V_\alpha(z;\theta)} d/d\theta_j f_z(z;\theta)\left[z - V_\alpha(z;\theta)\right] dz$ sub in eqn. 3

$= \int_{-b}^{V_\alpha(z;\theta)} \dfrac{f_z(z;\theta)}{f_z(z;\theta)} \dfrac{d}{d\theta_j} f_z(z;\theta)\left[z - V_\alpha(z;\theta)\right] dz$ multiply by 1

$= \int_{-b}^{V_\alpha(z;\theta)} f_z(z;\theta)\left[d/d\theta_j \log f_z(z;\theta)\left(z - V_\alpha(z;\theta)\right)\right] dz$ defn deriv log → same as used in policy grad

requires $d/d\theta_j \log f_Z(z;\theta)$: how do param impact performance distrib
in RL & some other applications, maybe in between variable
in RL θ policy param $\rightarrow$ traj of states/actions $\rightarrow$ rewards
easier to compute θ's impact on actions selected

- general CVaR gradient for RL
  - consider stochastic shortest path problems
  - finite state & action space MDP
  - parameterized policy π_θ
    - given history of states $s_0 \dots s_t = h_t$
    - yields distrib over actions $f_{a|h}(a|h;\theta)$
  - reward $\rho_t \sim f_{\rho|s,a}(\rho|s,a)$ (randomly sampled from distrib)
  - dynamics: $f_{s'|s,a}(s'|s,a)$
  - ξ_0 = initial state distrib
  - s^* = terminal state
  - $\forall\theta$ s^* reached w/prob 1
  - given π_θ let $s_0, a_0, \rho_0 \dots s_T$ be a traj from executing π_θ, ending $t=T$
  - decompose traj into discrete part & continuous
    - $Y \equiv s_0\, a_0\, s_1\, a_1 \dots s^*$
    - $X \equiv \rho_0\, \rho_1 \dots \rho_{T-1}$
  - $R \equiv \sum_{t=0}^{T} \rho_t$ ← random outcome of interest
  - goal: max CVaR of R by optimizing policy θ using gradient descent

interested in distrib of $R = r(X,Y)$
- $f_Y(y;\theta)$ = distrib of y (s, a traj) given θ
- $f_{X|Y}(x|y;\theta)$ = distrib of reward traj given a state-action traj
  - assume each element is bounded $[-b, b]$
  - X is a n-dim random var (1 per step)

assumptions very similar to 1-3 but now require gradients of
f_Y & $f_{X|Y}$ to exist & be bounded

Proposition 2. Under set of assumptions

$$\frac{\partial}{\partial\theta_j} \mathbb{E}_\alpha(R;\theta) = \mathbb{E}^\theta\left[\left(\underbrace{\frac{\partial \log f_Y(Y;\theta)}{\partial\theta_j}}_{\text{recall}} + \underbrace{\frac{\partial \log f_{X|Y}(X|Y;\theta)}{\partial\theta_j}}_{=0 \text{ in RL because } X \text{ indep of } \theta \text{ given } Y}\right)\Big(R - \mathrm{Va}(R;\theta)\Big)\,\Big|\, R \geq \mathrm{Va}(R;\theta)\right]$$

this holds for more general situations than RL - see paper

recall
$\log f_Y(Y;\theta) = \sum_t \log f_{a|h}(a_t|h_{t-1};\theta) + \log f_{s'|s,a}(s_{t+1}|s_t,a_t)$
second term indep of θ
so gradient of 1st term indep of dynamics params!

- Estimating CVaR gradient using sampling (focus on RL)
  sample n traj : yields n pairs $x_1, y_1 \cdots x_n, y_n$
  ← given a fixed θ
  compute empirical α-quantile $\bar{v} = \inf_r \hat{F}(x,y) \geq \alpha$
  empirical cdf of R
  $\hat{F}(z) \equiv \frac{1}{n} \sum_{i=1}^{n} 1_{r(x_i,y_i) \leq z}$

  Monte-carlo estimate of gradient $\Delta_{j;n}$

$$\Delta_{j;n} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial \log f_y(y_i;\theta)}{\partial \theta_j} + \frac{\partial \log f_{x|y}(x_i|y_i;\theta)}{\partial \theta_j} \right) (r(x_i,y_i) - \bar{v}) \, 1_{r(x_i,y_i) \geq \bar{v}}$$

- Using for CVaR optimization
  initialize paremeters θ^0, $i = 0$
  sample n_i traj given θ^i
  compute estimated gradients $\Delta_{j;n_i}$ $\forall j = 1:k$
  update parameters

$$\theta_j^{i+1} = \Gamma(\theta_j^i + \epsilon_i \Delta_{j;n_i})$$

  Γ: projection to compact set w/ smooth boundary ← ensures params stay bounded
  ϵ_i: pos constant

→ only get this

- Properties of Estimates & Approach
  empirical estimate of α-quantile is biased
  → $\Delta_{j;n}$ biased estimator of CVaR gradient
  but $\Delta_{j;n}$ is still a consistent estimator
  under same assumptions as prop 2
  $\Delta_{j;n} \to \partial/\partial\theta_j \, \Phi_\alpha(R;\theta)$ with prob 1 as $n \to \infty$ (estimate from ∞ samples)

assumption 6 : $\forall \theta$ $f_R(\cdot;\theta)$ and $g(\beta;\theta) \equiv E^\theta\left[\left(\frac{\partial \log f_y(y;\theta)}{\partial \theta_j} + \frac{\partial \log f_{x|y}(x|y;\theta)}{\partial \theta_j} \right) (R - \text{Var}(R;\theta)) \,|\, R = \beta \right]$
  are continuous at $\text{Var}(R;\theta)$
  and $f_R(\text{Var}(R;\theta);\theta) > 0$

thm: under same assump as Prop 2 + assump 6 } bounds bias
$E[\Delta_{j;n}] - \partial/\partial\theta_j \, \Phi_\alpha(R;\theta)$ is $O(n^{-1/2})$

thm: under same assump as Prop 2 + assump 6 and
$\Phi_\alpha(R;\theta)$ cont. differentiable in θ
$\sum_{i=1}^{\infty} \epsilon_i = \infty$, $\sum_{i=1}^{\infty} \epsilon_i^2 < \infty$
(*) $\sum_{i=1}^{\infty} \epsilon_i \left| E[\Delta_{j;n_i}] - \partial/\partial\theta_j \, \Phi_\alpha(R,\theta^i) \right| < \infty$ w.p.1 $\forall j$
then CVaR optimization converges to a local optima almost surely

(*) requirement requires $\lim_{i\to\infty} n_i = \infty$
e.g. eventually estimating gradient w/∞ samples
but n_i can grow slowly
e.g. $\epsilon_i = 1/i$ $n_i = (\log i)^4$

- Implementing
  Does choice of α impact quality of estimate?
  yes. essentially using αn_i samples
  for small α can be very high variance
  can use importance sampling to reduce variance
  (θ from past iterations)

  are assumptions reasonable for RL? ← like in policy gradient approaches
  $\log f_{a|h}(a|h,\theta)$ needs to be smooth & gradient bounded
  softmax satisfies
  if reward is discrete, doesn't satisfy assumptions
  but empirically that was fine
  requires a simulator or access to real world
  n traj to make a step in policy gradient
  may not be as sample efficient as GP approaches
  note: for MDPs optimal policy for maximizing expected reward depends only on current state
  → is this true for CVaR criteria?
  not always
  in some cases s_t function of accumulated reward sufficient

used to play Tetris!
how do you expect policy to very relative to π that maximizes expected total reward?

follow up ideas