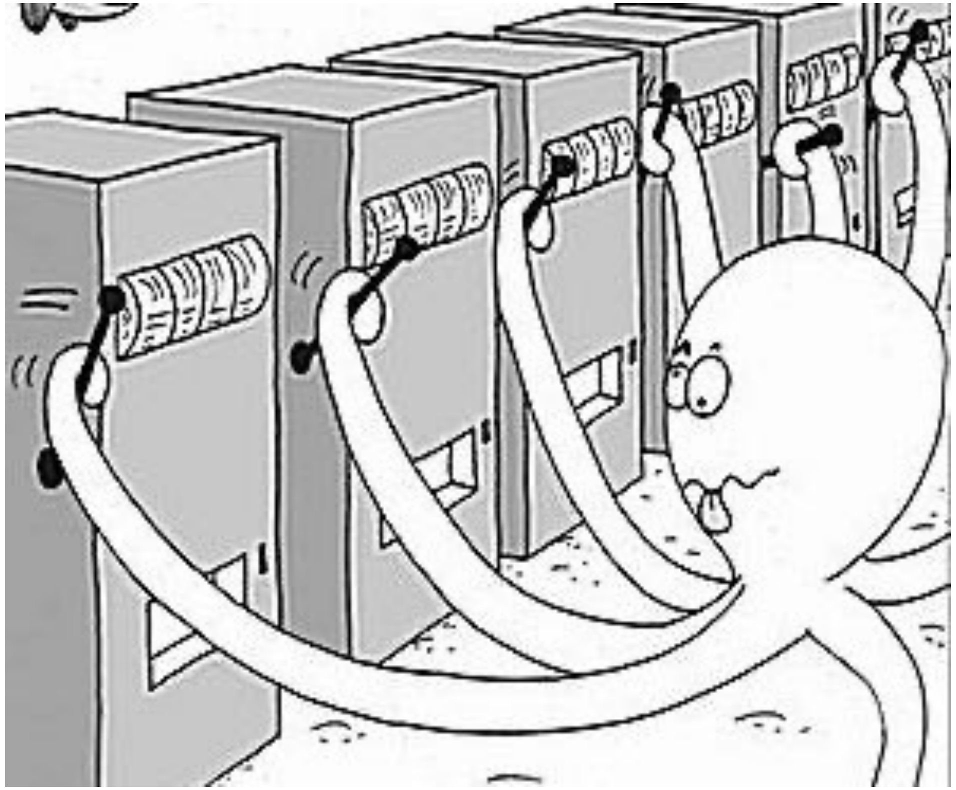
CSE 574 Planning and Learning Methods in AI

Ransalu Senanayake

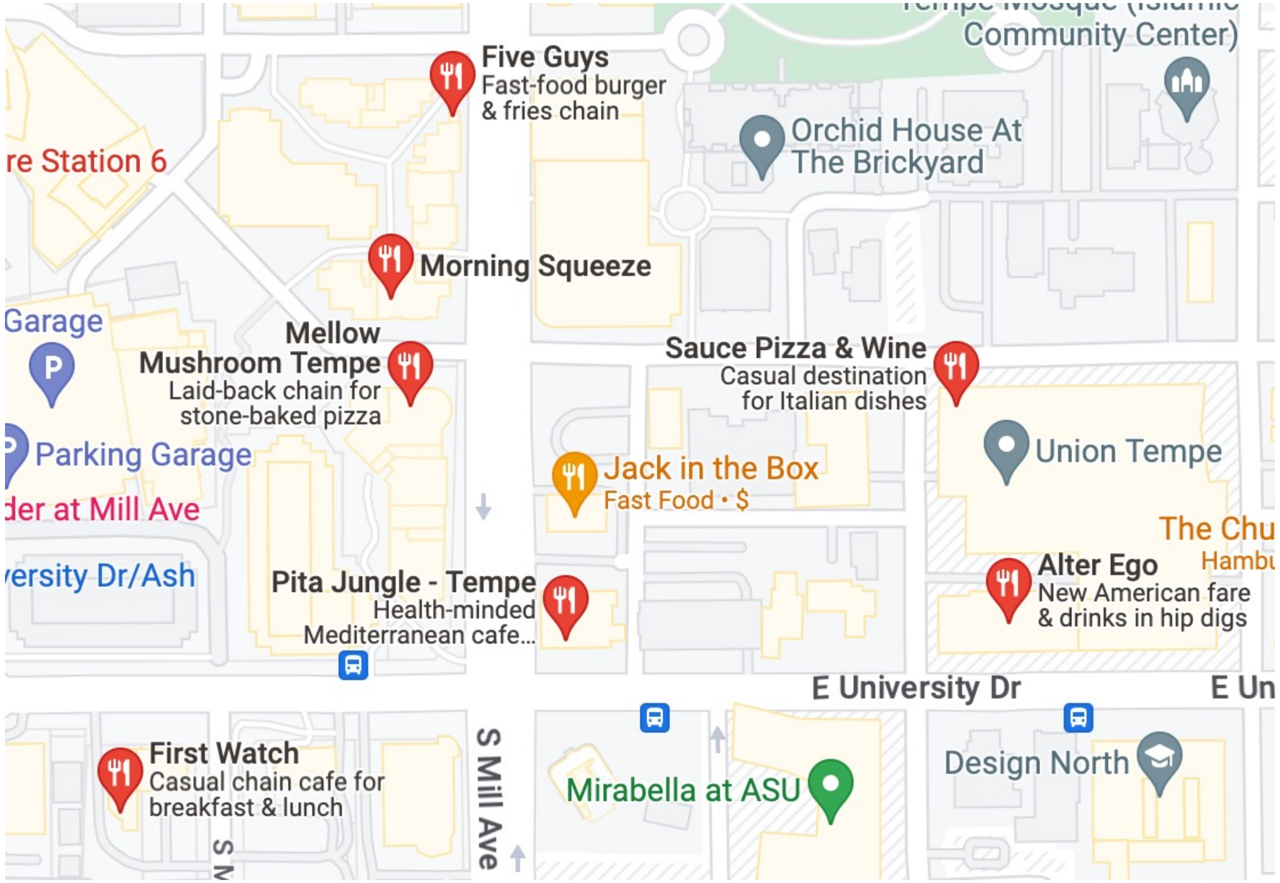
Week 3

Multi-Arm Bandits



Multi-Arm Bandits Solutions

- Exploitation only
- Exploration only (greedy)
- ϵ-first (exploration-first)
- ϵ-greedy
- UCB



ϵ-greedy

$$\text{arm}_t = \begin{cases} \text{arm that maximizes reward, with probability } 1-\epsilon \\ \text{random arm, with probability } \epsilon \end{cases}$$

- What is the effect of ϵ?
- Fixed ϵ (e.g., ϵ=10%), decreasing ϵ, adaptive ϵ, etc.
- Can we utilize more information than the average?

UCB1 Algorithm

```
Randomly pull arms k={1,…,K} several times (n) to get an initial
estimate of expected rewards r̄_k

For iteration t=1,…,T
   Play machine k_{t+1} = argmax( r̄_k + α sqrt(2 logN_t / n_k) )
end
```

$$k_{t+1} = argmax\left(\bar{r}_k + \alpha\sqrt{\frac{2\,\log N_t}{n_k}}\right)$$

Expected regret

Initial phase (figuring out the reward from each arm): $\mathcal{O}(\sqrt{K\mathrm{T}\log T})$

Later phase (when we get to know about arms/δr_k): $\mathcal{O}\left(\sum_k \frac{1}{\delta r_k}\log T\right)$

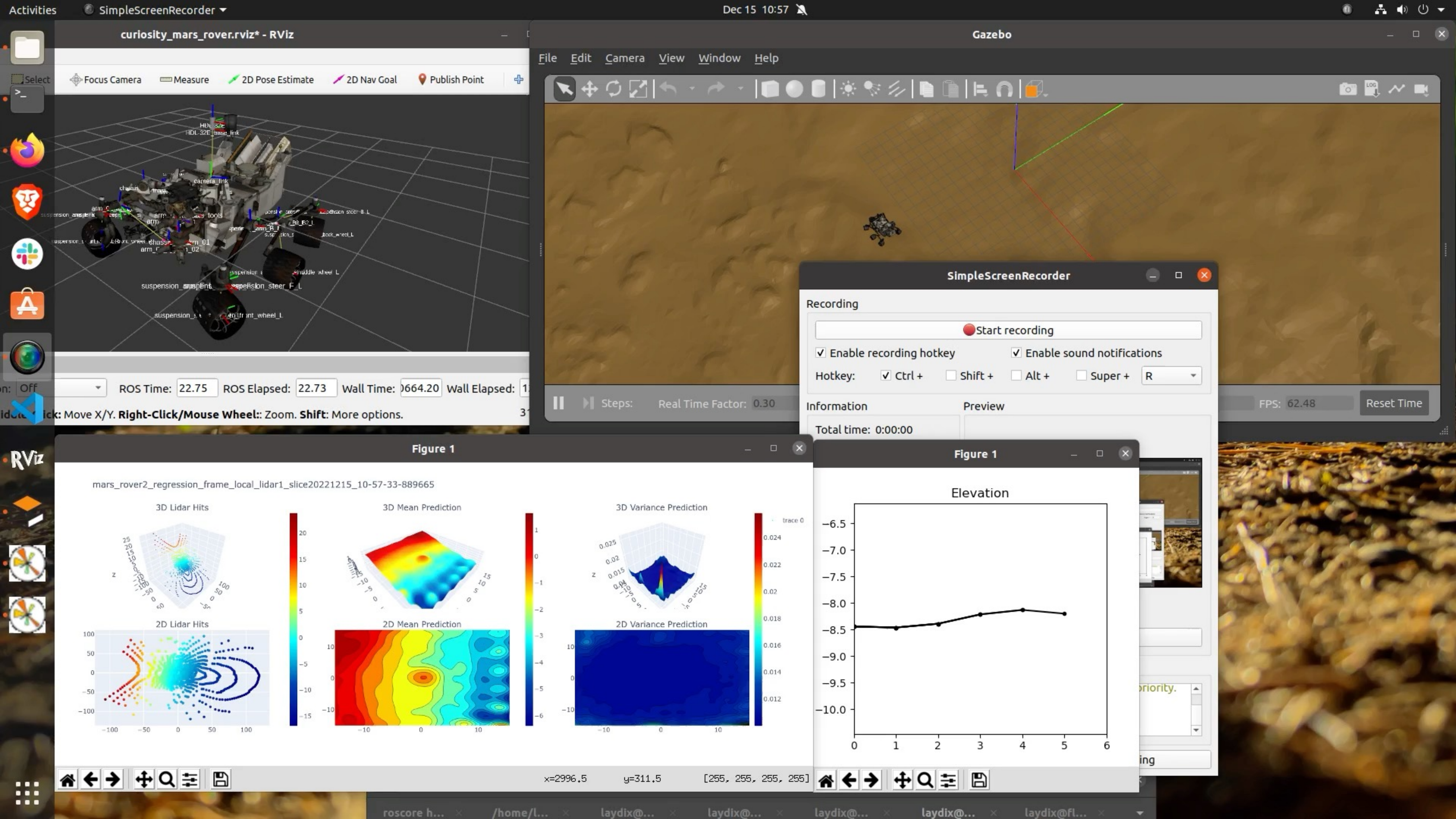
δr_k is the reward gap of the kth arm compared to the arm with the best reward

Bayesian Bandits and Thompson Sampling

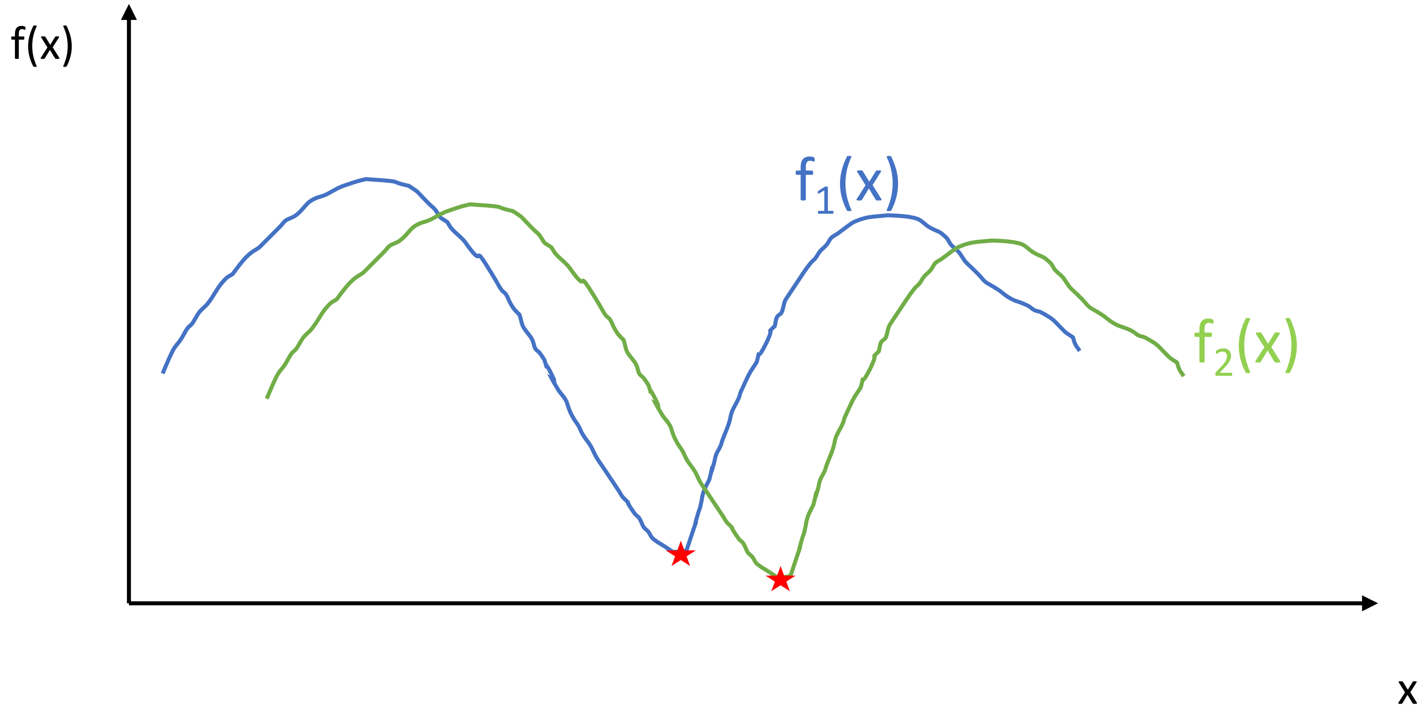
```
Assume parameterized distributions for the prior and likelihood

For T iterations
  Compute the posterior p(θ|D) ∝ p(D|θ)p(θ)
  Sample parameters from each arm
  Compute the reward for each sample
  Pick the arm that maximizes the reward
  Append the dataset with D={(arm,reward)}
end
```

- Non-informative/uniform/flat/broad prior. Conjugate prior.

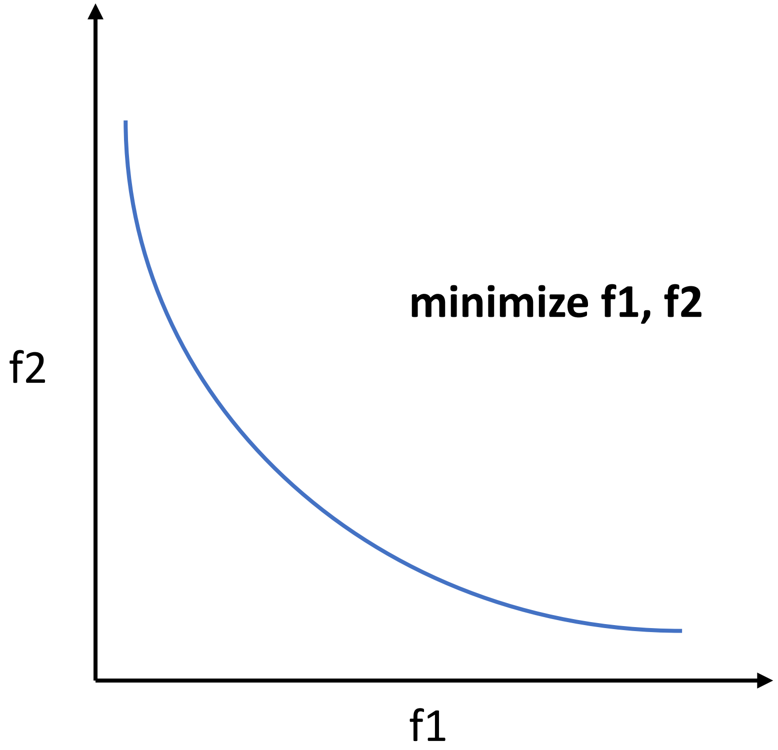

Activities
SimpleScreenRecorder
Dec 15 10:57
curiosity_mars_rover.rviz* - RViz
Select
Focus Camera
Measure
2D Pose Estimate
2D Nav Goal
Publish Point
ROS Time: 22.75
ROS Elapsed: 22.73
Wall Time:
Wall Elapsed:
Move X/Y. Right-Click/Mouse Wheel:: Zoom. Shift: More options.
Gazebo
File
Edit
Camera
View
Window
Help
Steps:
Real Time Factor: 0.30
FPS: 62.48
Reset Time
SimpleScreenRecorder
Recording
Start recording
Enable recording hotkey
Enable sound notifications
Hotkey:
Ctrl +
Shift +
Alt +
Super +
R
Information
Preview
Total time: 0:00:00
Figure 1
mars_rover2_regression_frame_local_lidar1_slice20221215_10-57-33-889665
3D Lidar Hits
3D Mean Prediction
3D Variance Prediction
2D Lidar Hits
2D Mean Prediction
2D Variance Prediction
trace 0
x=2996.5
y=311.5
[255, 255, 255, 255]
Figure 1
Elevation
roscore h...
/home/l...
laydix@...
laydix@...
laydix@...
laydix@...
laydix@fl...

Multi-Objective Optimization



Multi-Objective Optimization

- A choice is *Pareto optimal* if it is impossible to improve in one objective without worsening at least one other objective



- $wf_1(x) + (1 - w)f_2(x)$
- Possible solutions
- What if we maximize
- Hypervolume

Multi-Objective Bayesian Optimization (MOBO)

Differentiable Expected Hypervolume Improvement for Parallel Multi-Objective Bayesian Optimization

Samuel Daulton
Facebook
sdaulton@fb.com

Maximilian Balandat
Facebook
balandat@fb.com

Eytan Bakshy
Facebook
ebakshy@fb.com