# CO424 Reinforcement Learning Part 2: Coursework Part 1

Dr. Edward Johns

Wednesday 6th November 2019

This is Part 1 of the coursework for the second half of CO424 Reinforcement Learning. Part 2 of the coursework will be released next week. There are 5 questions, with each question requiring figures to be created, together with two short questions to be answered. You should submit a single PDF to CATE called "part_1_123456.pdf", where "123456" are the last 6 digits of your CID number. This PDF should contain the answers to the questions first, followed by the figures. The answers to the questions should take up no more than 2 pages of A4, font size 12. Top marks can be achieved with significantly less than 2 pages; these are intentionally short questions, with short answers expected. At the end of this document is an example of how the submitted document should be formatted.

**Submission deadline: 7pm on Thursday 21st November**

## 1 Online Learning and Mini-Batch Learning

*For this question, complete Sections 1 - 4 of the tutorial.*

**Figures 1 (a) and (b)** Plot two graphs of *Loss* ($y$-axis) vs *Steps* ($x$-axis), as in Section 4 of the Tutorial. The $y$-axes should have a logarithmic scale. For both graphs, the agent should be run for 500 steps. *Loss* is the Q-network's MSE loss for this step, and *Steps* is the total number of steps the agent has taken in the environment so far. Figure 1 (a) should be for online learning, and Figure 1 (b) should be when using the experience replay buffer. So for online learning, the loss is on a single sample, whereas with the experience replay buffer, it is the average loss over the mini-batch. Note that when using the replay buffer, the first data point on the plot should be at 50 steps (the size of the mini-batch). Remember to label the graphs.

**Question 1**

(i) Inspect the variance of the loss in each graph. State which of these two methods results in the most stable training, and explain the reason for this.

(ii) State which method is more efficient in terms of improving the Q-network's predictive accuracy in a given amount of wall clock time, and explain the reason for this.

# 2 Visualisation

*For this question, complete Setion 5 of the tutorial.*

**Figures 2 (a) and (b)** Show two images visualising what the agent has learned when a discount factor of 0 is used in the Bellman equation, as in Section 5 of the Tutorial. Figure 2 (a) should visualise the Q-values, and Figure 2 (b) should visualise the greedy policy. For both images, the agent should be run for 500 steps

**Question 2**

(i) Inspect the Q-value visualisations for the top-right and bottom-left of the environment. Is it likely that one of these two regions will have more accurate Q-value predictions than the other region? Explain your reasoning.

(ii) Inspect the visualisation of the policy. State whether the agent is able to reach the goal, and explain why this is.

# 3 Long-Term Prediction

*For this question, complete Section 6 of the tutorial.*

**Figures 3 (a) and (b)** Similarly to Section 1 of this coursework, plot two graphs of Loss ($y$-axis) vs *Steps* ($x$-axis) when the full Bellman equation is used, as in Section 6 of the Tutorial. The $y$-axes should have a logarithmic scale. Figure 3 (a) should be for when the Q-network is used in the Bellman equation, and Figure 3 (b) should be for when the target network is used in the Bellman equation. For both graphs, the agent should be run for 500 steps. Remember to label the graphs.

**Question 3**

(i) Inspect the shape of the loss curve in Figure 3 (a). Describe why the loss behaves in this way.

(ii) Inspect the shape of the loss curve in Figure 3 (b). Describe why the loss behaves in this way, and why it behaves differently to that in Figure 3 (a).

## 4 Exploration vs Exploitation

*For this question, there is no tutorial. It is up to you to decide how it should be implemented.*

**Figure 4** Starting from the code you have written by the end of the Tutorial, implement $\epsilon$-greedy exploration. Start $\epsilon$ at 1.0, and decrease $\epsilon$ by $\delta$ on every agent step. Stop decreasing $\epsilon$ if it reaches 0.0. Find the optimal value of $\delta$ which results in the best greedy policy after 500 steps in the environment. Plot a graph showing the agent's total sum of rewards per episode ($y$-axis) vs $\delta$ ($x$-axis), for five different values of $\delta$. The total sum of rewards should be for the greedy policy, after training the agent for 500 steps. The five values of $\delta$ should include $\delta = 0.0$, $\delta = 1.0$, the optimal value you found, plus two other values of your choosing which help to show the trend in the graph. Remember to label the graph.

**Question 4**

(i) Write down your optimal value of $\delta$, and explain why this gives superior performance to a value of 0.0.

(ii) Explain why your optimal value of $\delta$ gives superior performance to a value of 1.0.

## 5 Reward Function

*For this question, there is no tutorial. It is up to you to decide how it should be implemented.*

**Figures 5 (a) and (b)** Starting from the code you have written by the end of Question 4, and using your chosen optimal value for $\delta$, experiment with different reward functions (this is a function in the Agent class). Score each reward function by how close the Agent is to the goal at the end of the episode, when testing with the agent's greedy policy. Note that this is not the same value as just the reward at the end of the episode. Similar to Section 5 in the Tutorial, this test should be run separately from the main training loop. Using this score, find a reward function which gives superior performance to the original reward function in the starter code. You do not need to exhaustively search all possible reward functions; just find one which shows a clear improvement. In Figure 5 (a), plot two curves, on the same graph, showing *Final Distance* ($x$-axis) vs *Steps* ($y$-axis), for both your chosen reward function, and the original reward function. *Final Distance* is the distance between the agent and the goal at the end of the episode, when the agent is tested with its greedy policy. *Steps* is the number of steps the agent has taken in the environment (not including the "test" steps). To generate these curves, train for 500 agent steps, and run a test with the greedy policy after every agent step. Remember to label the graph. In Figure 5 (b), visualise the final greedy policy (not the Q-values) for your chosen reward function using the method described in the Tutorial, after training for 500 agent steps.

**Question 5**

(i) State what your chosen reward function is, and explain why this gives superior performance to the original reward function.

(ii) Consider a sparse reward function, which gives zero reward everywhere, except for a reward of 1 if the agent is within a radius of 0.05 from the goal. For this particular task, explain whether this sparse reward function would give better or worse performance than your chosen reward function, when trained for 500 agent steps.

# Student Name

CID: 00123456

Reinforcement Learning Part 2: Coursework Part 1

*This is an example of how to format your coursework answers

**Question 1 (i)** Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.

**Question 1 (ii)** Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.

**Question 2 (i)** Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.

**Question 2 (ii)** Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.

**Question 3 (i)** Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.

**Question 3 (ii)** Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.

**Question 4 (i)** Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.

**Question 4 (ii)** Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.

**Question 5 (i)** Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.

**Question 5 (ii)** Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.

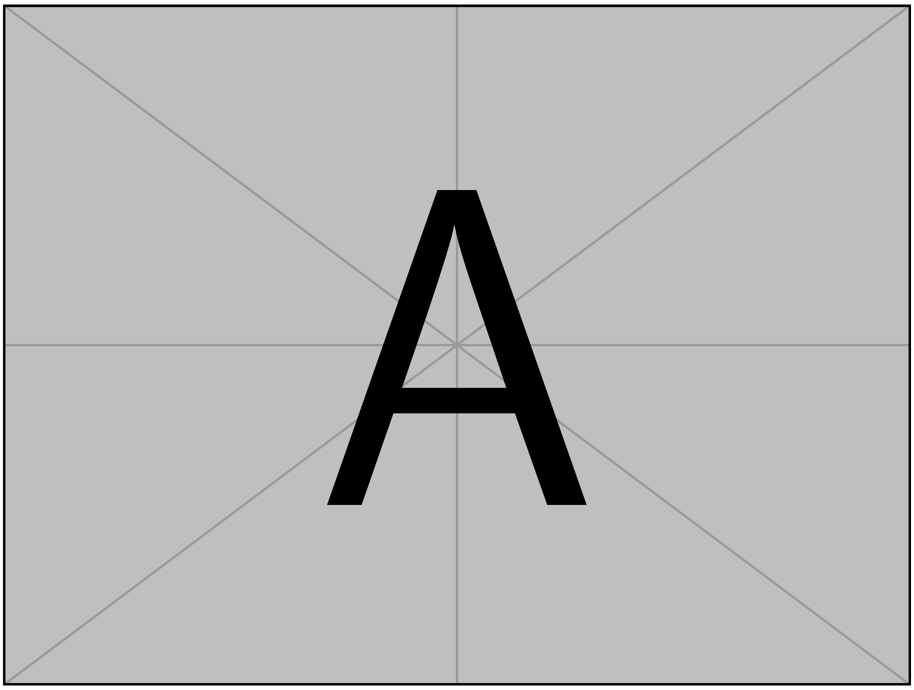

Figure 1 (a)

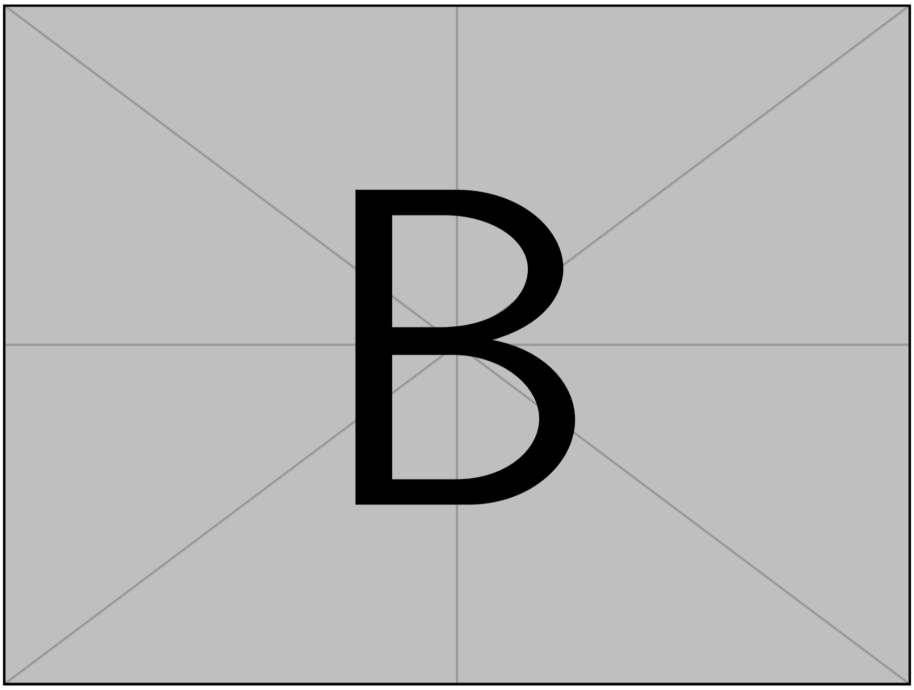

Figure 1 (b)

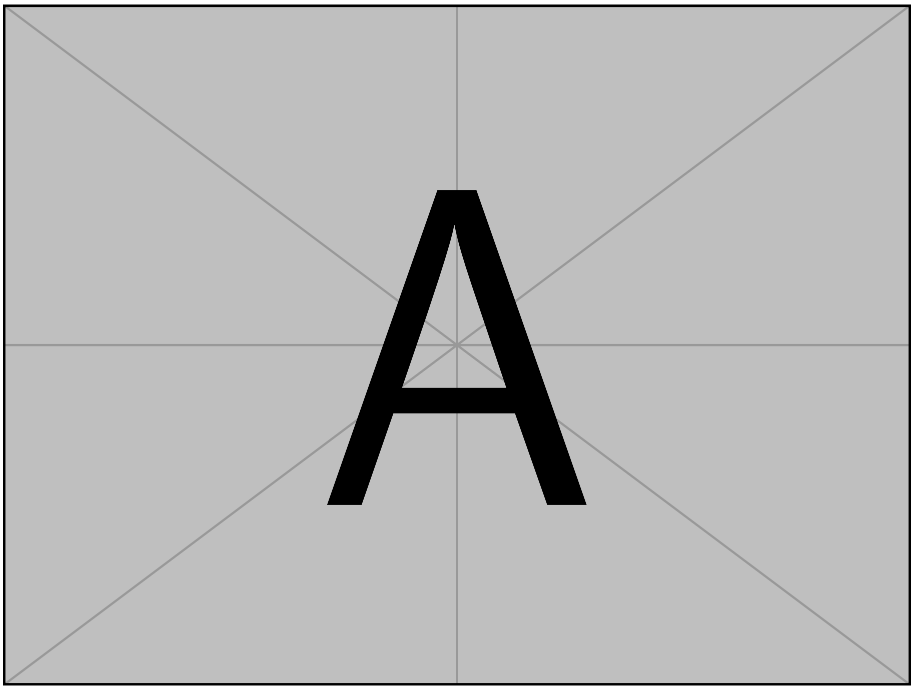

Figure 2 (a)

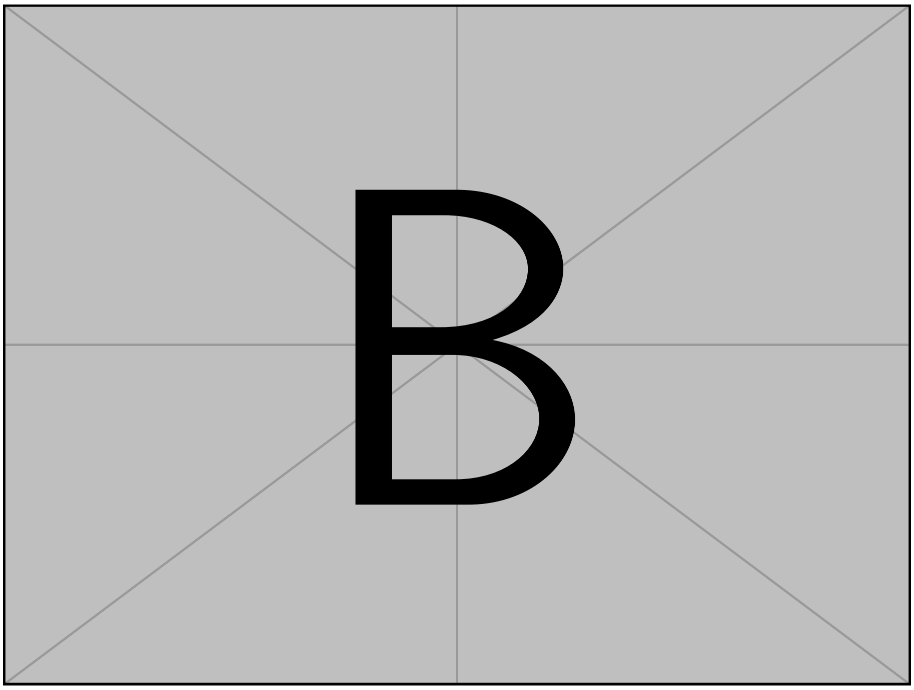

Figure 2 (b)

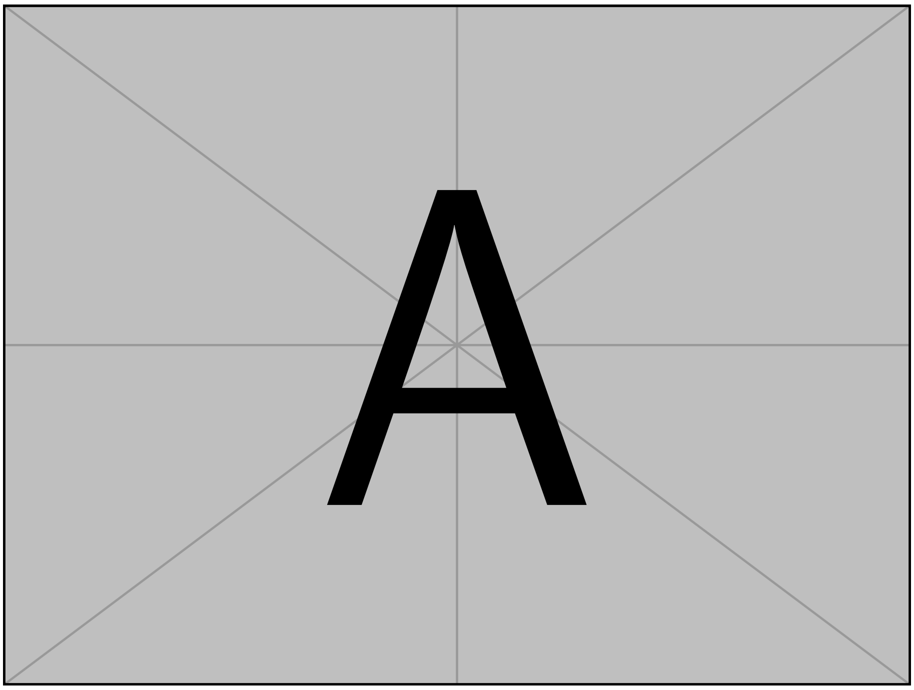

Figure 3 (a)

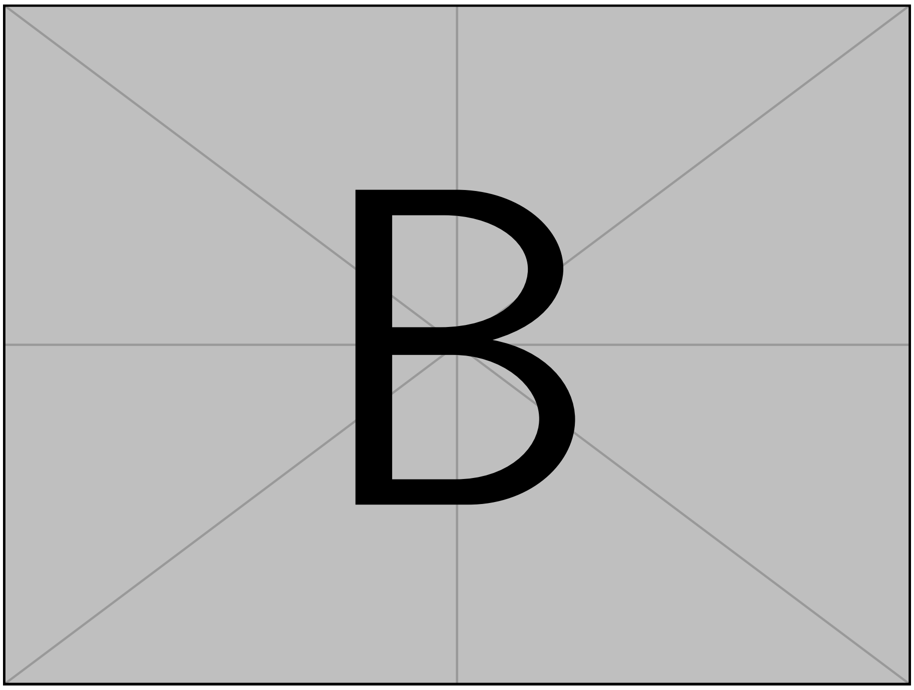

Figure 3 (b)

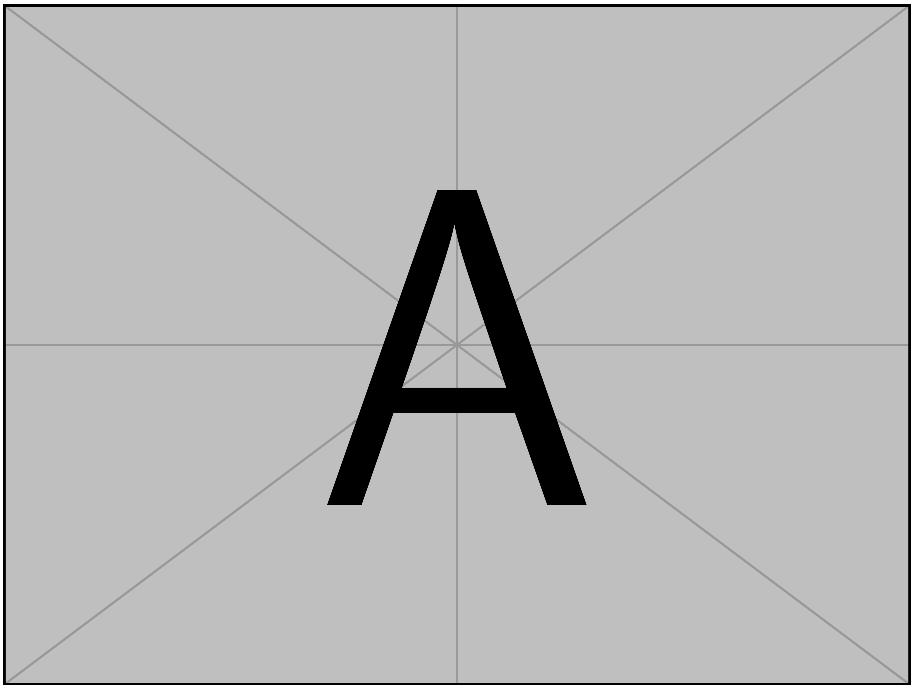

Figure 4

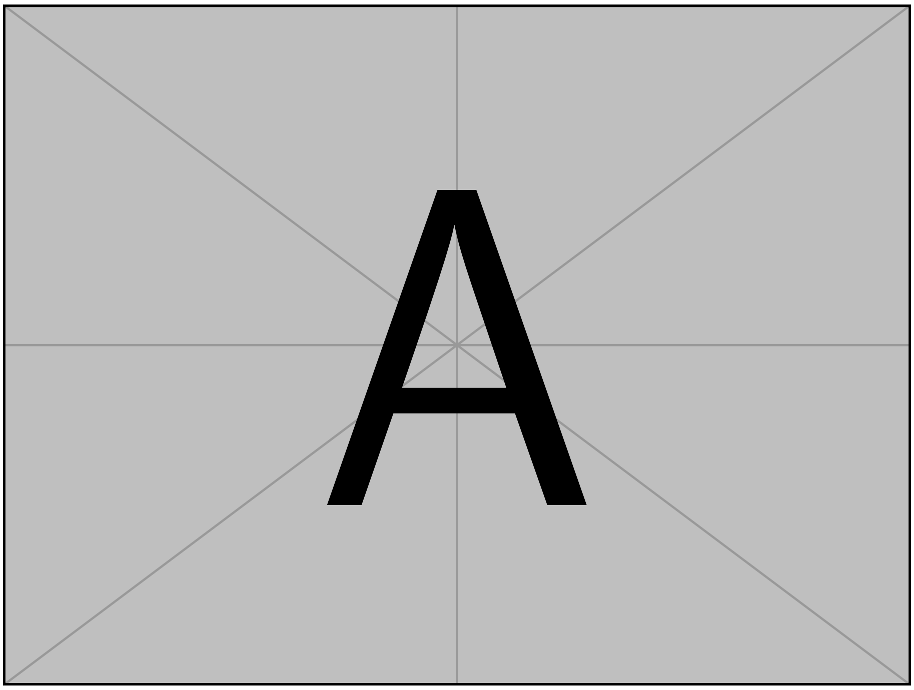

Figure 5 (a)

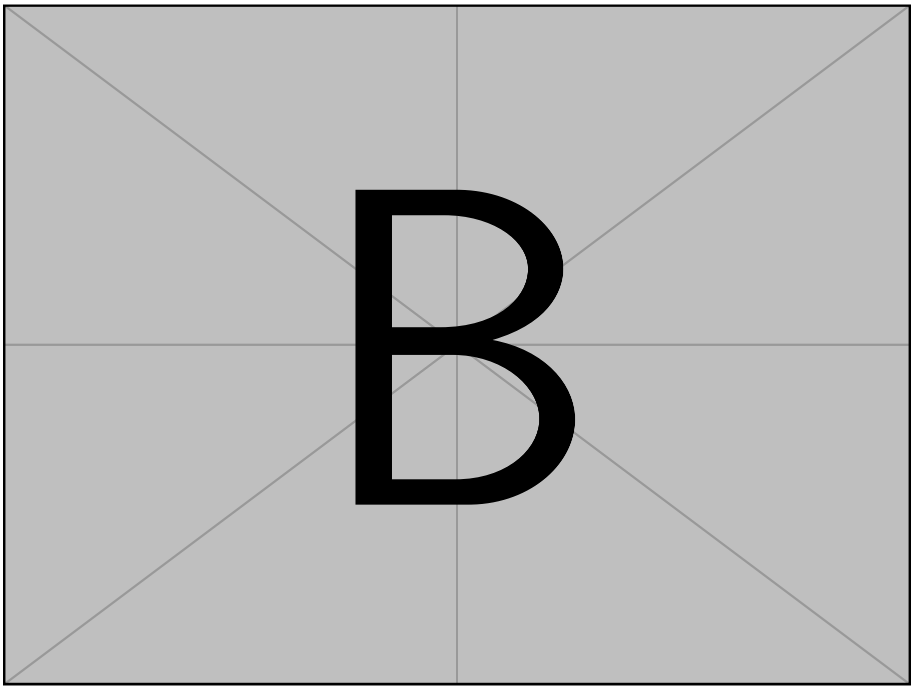

Figure 5 (b)