Ejercicio 1

Demostrar que, si conociéramos exactamente el valor de cada acción, es decir, si $Q_t(a) = E\left[R_t \middle| A_t = a\right]$, entonces la acción *greedy* $A_t = argmax_a Q_t(a)$ es la acción óptima en el sentido de que permite maximizar las recompensas totales.

Solución:

Sea,
a: Acción arbitraria
A_t: Acción tomada en el tiempo t
R_t: Recompensa obtenida en el tiempo t
$Q_t(a)$: Valor estimado de la acción a en el tiempo t
$q_*(a)$: Recompensa esperada dado que se toma la acción a
Acción *greedy*: acción cuyo valor estimado es el mayor de todos los valores estimados de las acciones posibles en el tiempo t

El valor $q_*(a)$ de una acción arbitraria a equivale al valor esperado de la recompensa si a es seleccionada.

$$q_*(a) \doteq \mathbb{E}\left[R_t \middle| A_t = a\right]$$

La aplicación del algoritmo *greedy* implica guardar los valores estimados de las acciones en cada iteración, en una estrategia no exploratoria. Se calcula $Q_t(a)$ como el valor estimado de la acción a en el tiempo t, o bien, promediando las recompensas obtenidas, es decir,

$$Q_t(a) \doteq \frac{\text{suma de las recompensas al tomar } \mathbf{a} \text{ antes de } \mathbf{t}}{\text{veces que se ha tomado } \mathbf{a} \text{ antes de } \mathbf{t}}$$

Donde $\mathbb{1}$ denota una variable aleatoria que toma el valor 1 si la acción a fue seleccionada en el tiempo t, y 0 en caso contrario[1]. La estrategia no exploratoria se interpreta matemáticamente como:

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

De la cual se toma el valor máximo del *Q-valor*, es decir

$$A_t = argmax_a Q_t(a)$$

Si se conociera el valor de $q_*(a)$, entonces la acción *greedy* será la acción que maximiza el valor esperado de la recompensa, es decir, la acción a tal que $Q_t(a) \approx q_*(a)$.

Ejercicio 2

En una selección de acciones tipo $\varepsilon - greedy$ con dos acciones posibles y $\varepsilon = 0{,}1$, ¿Cuál es la probabilidad de seleccionar la acción *greedy*?

Solución:

Ejercicio 3 Demostrar que el valor de una acción después de haber sido seleccionada $n-1$ veces, definido como

$$Q_n = \frac{R_1 + R_2 + \ldots + R_{n-1}}{n-1}$$

puede calcularse incrementalmente con la siguiente fórmula:

$$Q_{n+1} = Q_n + \frac{1}{n}\left[R_n - Q_n\right]$$

[1] Si el denominador es 0, se define $Q_t(a)$ como un valor por defecto, sea 0

Describa la ventaja de esta fórmula desde un punto de vista computacional.

Solución:

Sea,
Q_n: Valor estimado de la acción después de haber sido seleccionada $n-1$ veces
R_i: Recompensa obtenida después de la i-ésima selección de la acción
n: Número de veces que la acción ha sido seleccionada

Dado que Q_n es el promedio de las recompensas obtenidas después de haber seleccionado la acción $n-1$ veces, se tiene que

$$Q_n = \frac{R_1 + R_2 + \ldots + R_{n-1}}{n-1}$$

O bien puede escribirse como

$$Q_n = \frac{1}{n-1} \sum_{i=1}^{n-1} R_i$$

O su equivalente en términos de Q_{n+1}, al sustituir n por $n+1$ en la fórmula anterior

$$Q_{n+1} = \frac{1}{n} \sum_{i=1}^{n} R_i$$

Si se desea calcular el valor de la acción después de haber sido seleccionada n veces, se puede hacer de manera incremental, es decir, se puede calcular el valor de la acción después de haber sido seleccionada n veces a partir del valor de la acción después de haber sido seleccionada $n-1$ veces, de la siguiente manera:

$$\begin{aligned} Q_{n+1} &= \frac{R_1 + R_2 + \ldots + R_{n-1} + R_n}{n} \\ &= \frac{n-1}{n} \frac{R_1 + R_2 + \ldots + R_{n-1}}{n-1} + \frac{1}{n} R_n \end{aligned}$$

Tras simplificar la expresión anterior, se obtiene

$$\begin{aligned} Q_{n+1} &= \frac{n-1}{n} Q_n + \frac{1}{n} R_n \\ &= Q_n + \frac{1}{n} \left[R_n - Q_n \right] \end{aligned}$$

Ventajas de la fórmula incremental

- Memoria constante para estimar el valor promedio de la recompensa de una acción.
- Reducción de la complejidad computacional al no tener que almacenar todas las recompensas obtenidas, ni tener que llevar a cabo la suma del numerador.
- El cálculo del valor de Q_{n+1} únicamente requiere guardar el valor de Q_n y n para poder calcular el valor de la acción después de haber sido seleccionada $n+1$ veces.

Ejercicio 4

Considere un problema *k-armed bandit* con $k = 4$ acciones. Considere la aplicación de un algoritmo *bandit* usando selección de acciones ε - *greedy*, estimación incremental de los valores de cada acción y valores iniciales nulos $Q_1(a) = 0\ \forall a$. Suponga la siguiente secuencia de acciones y recompensas: $A_1 = 1,\ R_1 = 1,\ A_2 = 2,\ R_2 = 1,\ A_3 = 2,\ R_3 = -2,\ A_4 = 2,\ R_4 = 2,\ A_5 = 3,\ R_5 = 0$. En algunos de estos pasos se ha tomado una decisión aleatoria.

- ¿En qué pasos definitivamente se tomaron decisiones aleatorias?
- ¿En qué pasos es posible que la decisión haya sido aleatoria?

Solución:

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Ejercicio 5

[Programación] Aplique el algoritmo bandit $\varepsilon - greedy$ con

- $\varepsilon = 0$ (greedy)
- $\varepsilon = 0.01$
- $\varepsilon = 0.1$

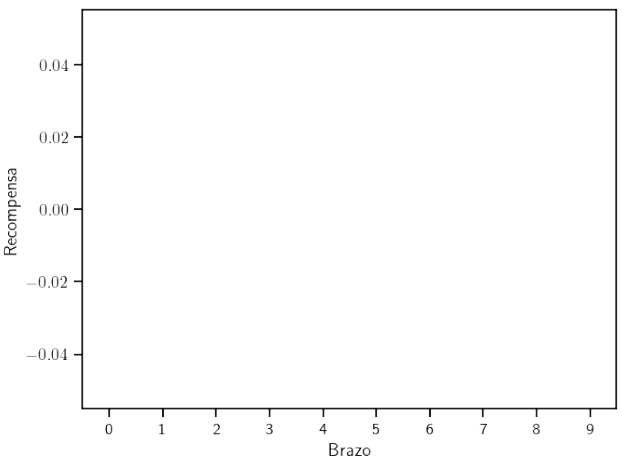
A un problema *k-armed bandit* con $k = 10$ acciones. Considere recompensas con medias aleatorias y desvío estándar constante σ. Analice experimentalmente el efecto del desvío estándar σ evaluando tres casos:

- $\sigma = 0$ (determinístico)
- $\sigma = 1$
- $\sigma = 10$

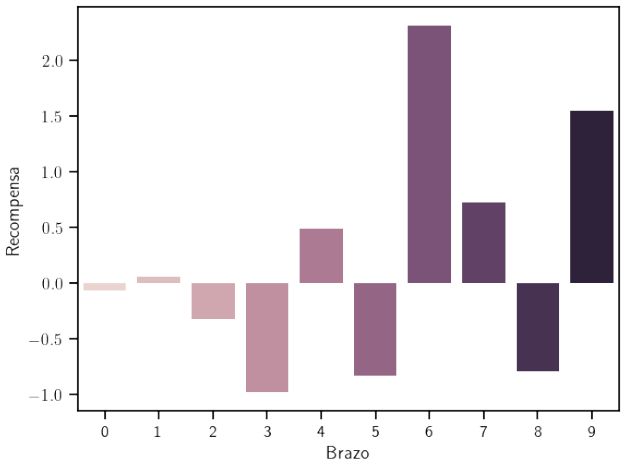
¿Qué conclusiones puede sacar?

Solución:

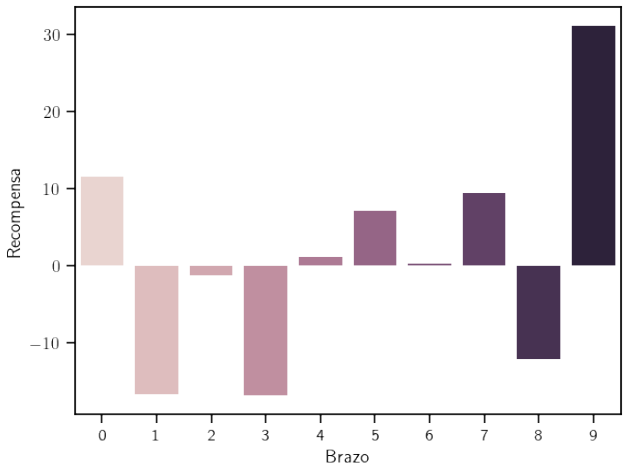
Las recompensas con medias aleatorias y desvío estándar constante σ se muestran en la Figura 1.
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.
Los resultados de los valores estimados de las acciones para $\varepsilon = 0$ (greedy), $\varepsilon = 0.01$ y $\varepsilon = 0.1$ se muestran en la Figura 2.
Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis.



(a) $\sigma = 0$



(b) $\sigma = 1$



(c) $\sigma = 10$

Figura 1: Recompensas con medias aleatorias y desvío estándar constante σ

Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Ejercicio 6

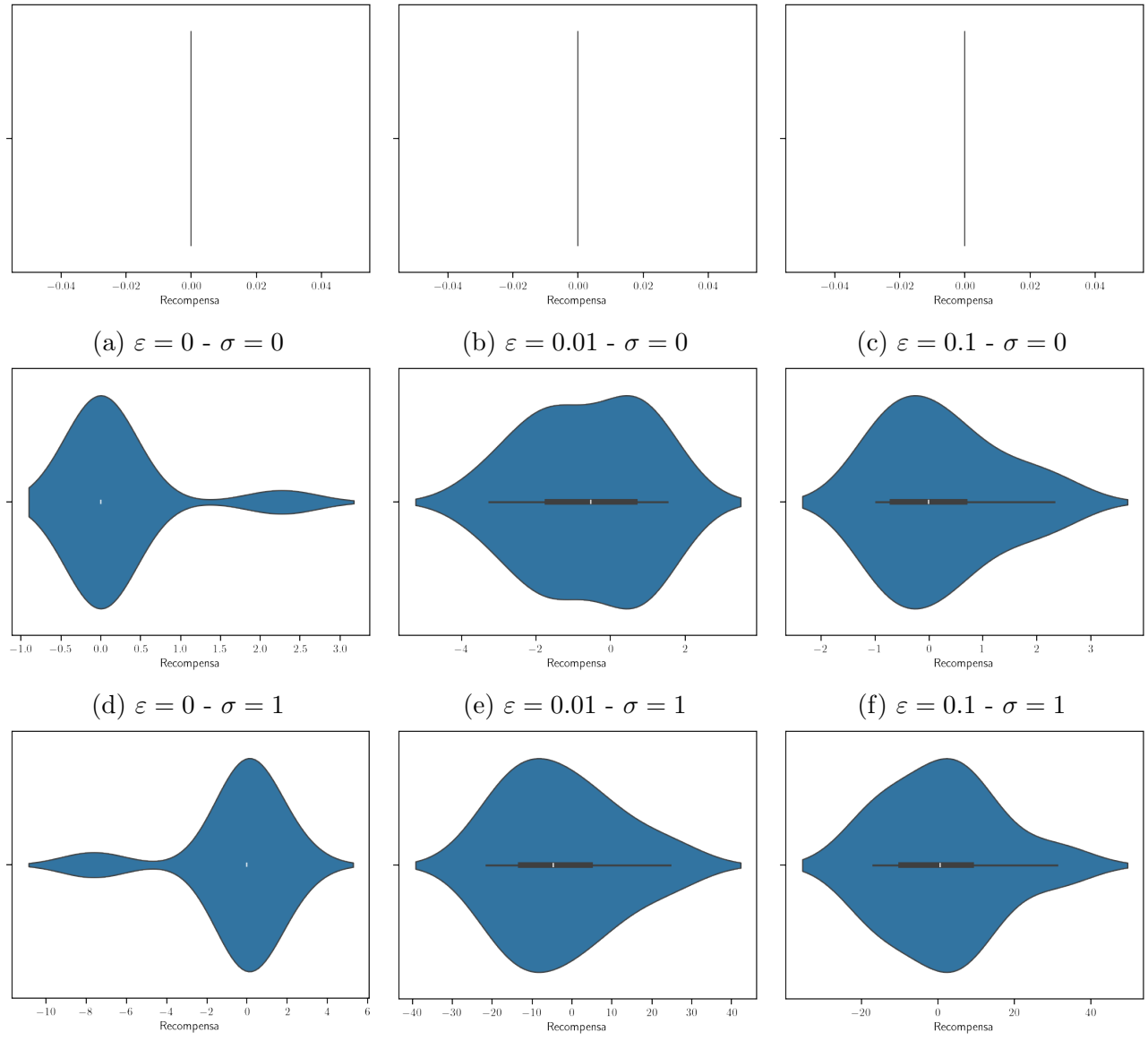
Dada la fórmula adaptativa del valor $Q_{n+1} = Q_n + \alpha\,[R_n - Q_n]$ con $\alpha \in (0, 1]$, demostrar que

- $Q_{n+1} = (1-\alpha)^n Q_n + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i$
- $(1-\alpha) + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1$, es decir, Q_{n+1} es un promedio pesado de $Q_n, R_1, R_2, \ldots, R_n$.

Solución:

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Ejercicio 7



(a) $\varepsilon = 0$ - $\sigma = 0$ (b) $\varepsilon = 0.01$ - $\sigma = 0$ (c) $\varepsilon = 0.1$ - $\sigma = 0$

(d) $\varepsilon = 0$ - $\sigma = 1$ (e) $\varepsilon = 0.01$ - $\sigma = 1$ (f) $\varepsilon = 0.1$ - $\sigma = 1$

(g) $\varepsilon = 0$ - $\sigma = 10$ (h) $\varepsilon = 0.01$ - $\sigma = 10$ (i) $\varepsilon = 0.1$ - $\sigma = 10$

Figura 2: Valores estimados de las acciones para ε σ

Demostrar que fórmula adaptativa para calcular el valor $Q_{n+1} = Q_n + \alpha\,[R_n - Q_n]$ con *step-size* $\alpha \in (0, 1]$ constante no verifica las hipótesis del teorema de convergencia y, por lo tanto, no está garantizada su convergencia.

Solución:

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Ejercicio 8

En la *Figura 2.3* del libro *Sutton&Barto (2018)*, se observa un *spike* en el paso número 11 cuando se utiliza inicialización optimista. De una explicación de este fenómeno.

Solución:

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Ejercicio 9

Demuestre que la función SOFTMAX: $p(a) = \frac{e^{H(a)}}{\sum_{a'=1}^{K} e^{H(a')}}$, define una distribución de probabilidades discreta válida.

Solución:

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Ejercicio 10

Demostrar que las derivadas de la función SOFTMAX $p(x)$ respecto de sus parámetros $H(a)$, $a = 1, 2, \ldots, K$, son iguales a:

$$\frac{\partial p(x)}{\partial H(a)} = \begin{cases} p(x)(1 - p(x)) & \text{si } x = a \\ -p(x)p(a) & \text{si } x \neq a \end{cases}$$

Solución:

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Ejercicio 11

Demostrar que la regla de actualización por gradiente ascendente estocástico:

$$H_{t+1}(a) = H_t(a) + \alpha \frac{\partial E[R_t]}{\partial H_t(a)}$$

con $E[R_t] = \sum_{x=1}^{K} p_t(x) q*(x)$, puede escribirse de la siguiente manera:

$$H_{t+1}(a) = \begin{cases} H_t(a) + \alpha(R_t - C)(1 - p_t(a)) & \text{si } a = A_t \\ H_t(a) - \alpha(R_t - C)p_t(a) & \text{si } a \neq A_t \end{cases}$$

Solución:

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Referencias Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Apéndice Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.