Media secuencial

$$\hat{\mu}_{t+1} = \hat{\mu}_t + \frac{1}{N}(x_{t+1} - \hat{\mu}_t).$$

Media normal

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Varianza secuencial

$$\hat{\sigma}^2_{t+1} = \hat{\sigma}^2_t + \frac{1}{N}[(x_t - \hat{\mu})^2 - \hat{\sigma}^2_t].$$

Varianza normal

$$\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

Softmax:

$$\pi(a = j \mid s) = \frac{e^{\beta a_j}}{\sum_{i=1}^{k} e^{\beta a_i}}$$

Derivada softmax con respecto a la acción elegida

$$\frac{\partial S_j}{\partial a_j} = S_j(1 - S_j)$$

Derivada softmax con respecto a la acción no elegida

$$\frac{\partial S_j}{\partial a_i} = -S_j S_i$$

Actualización de parámetros para datos discretos

$$a_j \leftarrow a_j + \alpha\left(1 - s_j\right) R$$

$$a_i \leftarrow a_i + \alpha(0 - s_i) R$$

para $i \neq j$

- s_j es la probabilidad actual de la acción j.

Q learning

$$Q_{k+1}(s,a) = Q_k(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right]$$

- **γ**: descuento del future - Indica cuánto valen las recompensas futuras.
    - Si γ es **alto** (≈1):
      el agente es paciente → valora el futuro.
    - Si γ es **bajo**:
      el agente solo se preocupa por recompensas inmediatas.

Valor esperado

$$\mathbb{E}[f(X)] = \sum_i p(x_i) \cdot f(x_i)$$

Decaimiento exponencial de ε

```
eps_threshold = eps_end + (eps_start - eps_end) * exp(- steps / eps_decay)
```

- Con probabilidad **1 – ε**, el agente **explota**: elige la acción con el valor Q más alto.
- Con probabilidad **ε**, el agente **explora**: elige una acción aleatoria entre las disponibles.

PDF de una distribución normal

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$