Policy gradient is a type of reinforcement learning that rely on optimizing parametrized policies with respect to the expected return by gradient descent. The policy is represented by a parametric probability distribution $\pi_\theta(a|s) = \mathcal{P}[a|s;\theta]$ that stochastically selects action a in state s with parameters θ. In continuous problem, stochastic policy gradient would result more computation because it integrates from both state and action space which is high dimension in continuous case. Silver et al. (2014) introduced deterministic policy gradient method which only integrates over state space $a = \mu_\theta(s)$.

Policy evaluation method estimates the action-value function $Q^\mu(s,a)$ which is used by an actor, and it selects actions in a greedy manner, $\mu^{k+1}(s) = argmax_a Q^{\mu^k}(s,a)$.
In general, behaving according to deterministic policy will not ensure adequate exploration and can lead to suboptimal solution, so an off-policy is used in this case $\beta(a,s) \neq \pi_\theta(a,s)$ to generate trajectories.
The performance objective of off-policy deterministic policy gradient will be set at:

$$J_\beta(\mu_\theta) = \int_S p^\beta(s) Q^\mu(s, \mu_\theta(s)) ds$$

.
Differentiating the performance objective and applying an approximation give the deterministic policy gradient:

$$\nabla_\theta J_\beta(\mu_\theta) = E_{s\sim p^\beta}[\nabla_\theta \mu_\theta(s) \nabla_a Q^\mu(s,a)|_{a=\mu_\theta(s)}].$$

A critic estimates action-value function $Q^w(s,a) \approx Q^\mu(s,a)$ with off-policy generated by $\beta(a,s)$. The authors have proven that there always exists a compatible function approximator of the form $Q^w(s,a) = A^w(s,a) + V^v(s)$ where $A^w(s,a) = (a - \mu_\theta(s))^\top \nabla_\theta \mu_\theta(s)^\top w$ is the advantage of take action a over action $\mu_\theta(s)$, and $V^v(s)$ may be any differentiable baseline function that is independent of action a with parameter v that $V^v(s) = v^\top \phi(s)$.
The parameters (θ for actor, w for critic, and v for baseline optimization) can be updated as follow with δ is the temporal difference error evaluated by the critic:

$$\delta_{t+1} = r_t + \gamma Q^w(s_{t+1}, a_{t+1}) - Q^w(s_t, a_t)$$
$$\theta_{t+1} = \theta_t + \alpha_\theta \nabla_\theta \mu_\theta(s_t)(\nabla_\theta \mu_\theta(s_t)^\top w_t)$$
$$w_{t+1} = w_t + \alpha_w \delta_t a_t^\top \nabla_\theta \mu_\theta(s_t)$$
$$v_{t+1} = v_t + \alpha_v \delta_t \phi(s_t)$$

The deterministic policy gradient algorithm has shown that it converges faster than stochastic policy gradient method. Later, the algorithm has been improved with Deep Neural Network in Deep Deterministic Policy Gradient (Lillicrap et al., 2015).

In DDPG, Lillicrap et al. added two target networks that the parameters are slowly updated by the two online networks which are optimized by the deterministic policy gradient algorithm. The target value for training is also evaluated from the target networks. A replay buffer was used for having a better training samples in each episode. OU method was also used for the exploration.