Policy Gradient Formulation

**Exact gradient formulation:** [Sutton et al., 2000]

\[
\nabla_\theta J_\theta (\theta) = \frac{1}{1 - \gamma} \int_S d^m \rho (s) \int_A \nabla_\theta \pi_a(s, \theta) Q^\pi(s, a) \, da \, ds.
\]

Policy parameter update rule: \(\theta' = \theta + \alpha \nabla_\theta J_\theta (\theta)\).

**Lower bound to policy performance:** [Pirotta et al., 2013]

\[
J_\theta (\theta' - J_\theta (\theta) \geq \frac{1}{1 - \gamma} \int_S d^m \rho (s) \int_A (\pi_a(s, \theta) - \pi_a(s, \theta')) Q^\pi(s, a) \, da - \frac{\gamma}{2(1 - \gamma)^2} \| \pi_a - \pi_a' \|_2 \| Q^\pi \|_\infty.
\]

**Optimal step size:** Exploiting a lower bound to Taylor's expansion it is possible to bound the policy step size, and obtain a new lower bound to the policy performance difference:

\[
J_\theta (\theta') - J_\theta (\theta) \geq \alpha \| \nabla_\theta J_\theta (\theta) \|^2_{\infty} + \frac{\alpha^2 s}{1 - \gamma} \int_S d^m \rho (s) \int_A \left( \sum_{i=1}^\infty \frac{\Delta \theta_i \Delta \theta_i}{2(1 - \gamma)} \right) Q^\pi(s, a) \, da
\]

The step size is polynomial in \(\alpha\) that admit an unique, positive root, that is the optimal step size.

Gaussian Policy Model

Given any pair of stationary Gaussian policies \(\pi_\theta \sim \mathcal{N}(\theta' \phi(s), \sigma^2)\) and \(\pi_\theta' \sim \mathcal{N}(\theta' \phi(s), \sigma^2)\), so that \(\theta' = \theta + \alpha \nabla_\theta J_\theta (\theta)\), then:

\[
J_\theta (\theta') - J_\theta (\theta) \geq \alpha \| \nabla_\theta J_\theta (\theta) \|^2_{\infty} + \frac{\alpha^2 s}{1 - \gamma} \int_S d^m \rho (s) \int_A \left( \sum_{i=1}^\infty \frac{\Delta \theta_i \Delta \theta_i}{2(1 - \gamma)} \right) Q^\pi(s, a) \, da
\]

Choosing the step size \(\alpha^*\) that maximizes the previous bound, we can guaranteed that:

\[
J_\theta (\theta') - J_\theta (\theta) \geq \frac{1}{2} \alpha^* \| \nabla_\theta J_\theta (\theta) \|^2_{\infty}.
\]

Approximate Framework

Given a policy gradient estimate \(\hat{\nabla}_\theta J_\theta (\theta)\), so that \(P \left( \| \nabla_\theta J_\theta (\theta) - \hat{\nabla}_\theta J_\theta (\theta) \| \geq \epsilon \right) \leq \delta\), the difference between the performance of \(\pi_\theta\) and \(\pi_\theta\) can be lower bounded at least with probability \((1 - \delta)\)\(^m\):

\[
J_\theta (\theta') - J_\theta (\theta) \geq \alpha \| \hat{\nabla}_\theta J_\theta (\theta) \|^2_{\infty} - \alpha^2 \left( \frac{1}{1 - \gamma} \sqrt{2\pi\sigma} \right)^2 \int_S d^m \rho (s) \int_A \left( \sum_{i=1}^\infty \frac{\Delta \theta_i \Delta \theta_i}{2(1 - \gamma)} \right) Q^\pi(s, a) \, da
\]

that is maximized by the following step size value:

\[
\alpha^* = \frac{1}{1 - \gamma} \sqrt{2\pi\sigma} \| \nabla_\theta J_\theta (\theta) \|_2 = \frac{1}{1 - \gamma} \sqrt{2\pi\sigma} \beta^* \| \nabla_\theta J_\theta (\theta) \|_2.
\]

\(\hat{\nabla}_\theta J_\theta (\theta)\) and \(\nabla_\theta J_\theta (\theta)\) are used to lower and upper bound the L2-norm and the L1-norm, respectively.