Variance in Policy Gradients

$$\nabla_\theta J(\theta) = \frac{1}{N} \sum_{\tau_s} \nabla_\theta \log \pi_\theta(\tau) \times r(\tau)$$

Now, if you add a constant to all rewards, say +100 to all rewards, then, the reward fn is still practically the same!

But then, the gradients for policy significantly changes!

For example, say there are 2 trajectories

A	B	
500	-500	← rewards

Add 1 million to both

1,000,500	999,500

In this case, one τ looks BAD & one τ looks GOOD

In this case, BOTH τs are looking good!

Also, mathematically since the gradients are in "product" of rewards! if every $r(\tau)$ gets added a constant b!

Then every gradient is multiplied by a constant b!

Say, there are 'n' nos $a_1, a_2, \ldots a_n$ whose mean is μ, & variance is σ, Now, scale all elements by b → $a_1 b, a_2 b, \ldots a_n b$

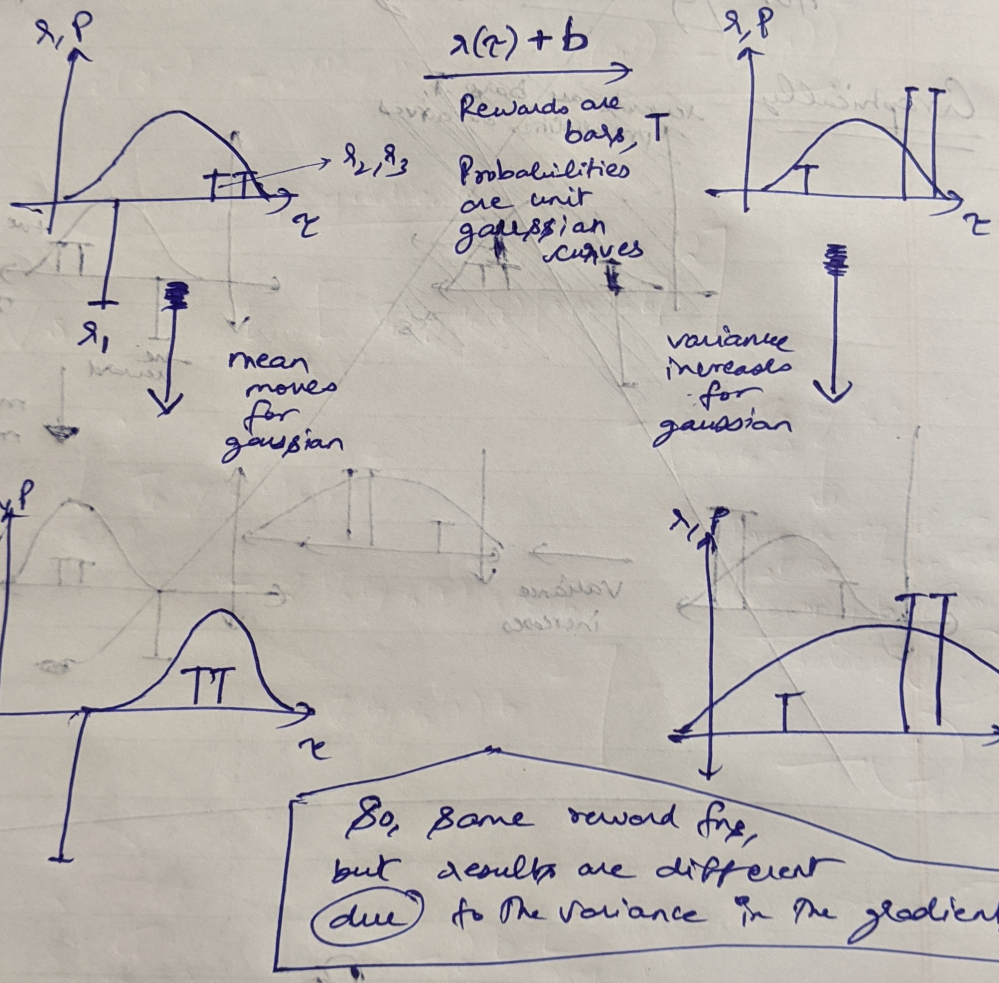
$$\text{mean} = \frac{a_1 b + a_2 b \ldots a_n b}{N} = b\mu$$

$$\text{Variance} = \frac{(a_1 b - b\mu)^2 + (a_2 b - b\mu)^2 \ldots (a_n b - b\mu)^2}{N} = b^2 \sigma$$

The variance of those 'n' numbers increases!

So, if the reward fn is modified [equivalently] by adding a constant, Then the gradients are multiplied by a constant factor increasing it's variance [QUADRATICALLY]

Graphically, Let's say these are 2 reward fns off by a constant factor



So, same reward fns, but results are different (due) to the variance in the gradients!

So, if we take the [mean] of rewards & subtract it to all rewards, this would make both reward fns in this example [equivalent]!

This mean subtraction would (NULLIFY) any constants that may (have) been added to the rewards!

So, Expectat gradie mean

This is same! r₂ J(θ) differe

Also, By adding a constant to all rewards, we **only** reduce the variance of gradients, keeping its mean **same**

Proof

$$\nabla_\theta J(\theta) = \mathop{E}_{\tau \sim \pi_\theta(\tau)} \nabla_\theta \log \pi_\theta(\tau) [R(\tau) - b]$$

$$= \mathop{E}_{\tau \sim \pi_\theta(\tau)} \nabla_\theta \log \pi_\theta(\tau) R(\tau) - \mathop{E}_{\tau \sim \pi_\theta(\tau)} \nabla_\theta \log \pi_\theta(\tau) \times b$$

Unlike $R(\tau)$, b doesn't change for **each** τ, b is constant for all τs, so it can be taken **COMMON**

$$= -b \mathop{E}_{\tau \sim \pi_\theta(\tau)} \nabla_\theta \log \pi_\theta(\tau)$$

$$= -b \sum_{\text{all } \tau\text{s}} \pi_\theta(\tau) \nabla_\theta \log \pi_\theta(\tau)$$

$$= -b \sum_{\text{all } \tau\text{s}} \nabla_\theta \pi_\theta(\tau)$$

$$= -b \nabla_\theta \left(\sum_{\text{all } \tau} \pi_\theta(\tau) \right) \quad \left(\sum_{\text{all } \tau} \pi_\theta(\tau) = 1\right)$$

$$= -b \nabla_\theta (1) \quad (\nabla_\theta (1) = 0)$$

$$= 0$$

So, Expectation of gradients, the mean is **SAME**

This is because, adding same / constant to all τs **just** increases $J(\theta)$ by a constant, so, differentiation of $J(\theta)$ would differentiate the constant & make it 0!

But the variance of it is not **zero**! → **Proof PTO**

Variance of 'n' nos whose mean is μ!

$$\frac{(a_1-\mu)^2 + (a_2-\mu)^2 \cdots (a_n-\mu)^2}{N}$$

$$= \frac{a_1^2 + 2a_1\mu + \mu^2 \cdots}{N}$$

$$= \frac{a_1^2 + a_2^2 \cdots + a_n^2}{N} - 2\mu\left(\frac{a_1 + a_2 \cdots a_n}{N}\right) + \frac{N\mu^2}{N}$$

$\left(\frac{a_1 + a_2 \cdots a_n}{N}\right) \to \mu$

$$= E(x^2) - 2\mu^2 + \mu^2 = \underline{E(x^2) - (E(x))^2}$$

So, variance of gradients

$$\text{Var} = \underset{\tau \sim \pi_\theta(\tau)}{E}\left[\left(\nabla_\theta \log \pi_\theta(\tau)\left[r(\tau) - b\right]\right)^2\right]$$

$$= \underset{\tau \sim \pi_\theta(\tau)}{E}\left[\left(\nabla_\theta \log \pi_\theta(\tau) r(\tau)\right)^2 - 2\left(\nabla_\theta \log \pi_\theta(\tau)\right)^2 r(\tau) b + \left(\nabla_\theta \log \pi_\theta(\tau)\, b\right)^2\right]$$

$$- \left(\underset{\tau \sim \pi_\theta(\tau)}{E} \nabla_\theta \log \pi_\theta(\tau)\left[r(\tau) - b\right]\right)^2$$

↓ This is equal to just the expectation, ignoring b, (proved previously!)

— (1)

$$= \text{old } \sigma \text{ (if b was never added)} - 2b \underset{\tau \sim \pi_\theta(\tau)}{E}\left(\nabla_\theta \log \pi_\theta(\tau)\right)^2 r(\tau) + b^2 \underset{\tau \sim \pi_\theta(\tau)}{E}\left(\nabla_\theta \log \pi_\theta(\tau)\right)^2$$

$$E\left[\left(\nabla_\theta \log \pi_\theta(\tau) r(\tau)\right)^2\right] - \left(E\left[\nabla_\theta \log \pi_\theta(\tau) r(\tau)\right]\right)^2$$

So, old σ + non-zero terms!

Hence Proved, the variance (has CHANGED)

So, what should b, be, so Variance is @ minimum?

$\frac{\partial Var}{\partial b} = 0$ So, from (1)

$$Var = E_{\tau \sim \pi_\theta(\tau)}\left[\left(\nabla_\theta \log \pi_\theta(\tau)\, x(\tau)\right)^2\right] - \left(E\left[\nabla_\theta \log \pi_\theta(\tau)\, x(\tau)\right]\right)^2$$

$$\frac{\partial Var}{\partial b} = \underbrace{-2b\, E\left(\nabla_\theta \log \pi_\theta(\tau)\right)^2 x(\tau) + b^2 E\left(\nabla_\theta \log \pi_\theta(\tau)\right)^2}_{0\ (\text{no } b\text{'s})} - 2E\left(\nabla_\theta \log \pi_\theta(\tau)\right)^2 x(\tau) + 2b\, E\left(\nabla_\theta \log \pi_\theta(\tau)\right)^2$$

$$0 = -2b\, E\left(\nabla_\theta \log \pi_\theta(\tau)\right)^2 - 2E\left(\nabla_\theta \log \pi_\theta(\tau)\right)^2 x(\tau)$$

$$\frac{E\left(\left(\nabla_\theta \log \pi_\theta(\tau)\right)^2 x(\tau)\right)}{E\left(\left(\nabla_\theta \log \pi_\theta(\tau)\right)^2\right)} = b$$

So, expectation of rewards weighted by $\left(\nabla_\theta \log \pi_\theta(\tau)\right)^2$

Since the mean of this individual [illegible]

So, the approximation, it's just