which in turn defines a transformed MDP. This transformation leaves the discounted advantage function $A^{\pi,\gamma}$ unchanged for any policy $\pi$. To see this, consider the discounted sum of rewards of a trajectory starting with state $s_t$:

$$\sum_{l=0}^{\infty} \gamma^l \tilde{r}(s_{t+l}, a_t, s_{t+l+1}) = \sum_{l=0}^{\infty} \gamma^l r(s_{t+l}, a_{t+l}, s_{t+l+1}) - \Phi(s_t). \tag{21}$$

Letting $\tilde{Q}^{\pi,\gamma}, \tilde{V}^{\pi,\gamma}, \tilde{A}^{\pi,\gamma}$ be the value and advantage functions of the transformed MDP, one obtains from the definitions of these quantities that

$$\tilde{Q}^{\pi,\gamma}(s,a) = Q^{\pi,\gamma}(s,a) - \Phi(s) \tag{22}$$

$$\tilde{V}^{\pi,\gamma}(s,a) = V^{\pi,\gamma}(s) - \Phi(s) \tag{23}$$

$$\tilde{A}^{\pi,\gamma}(s,a) = (Q^{\pi,\gamma}(s,a) - \Phi(s)) - (V^{\pi,\gamma}(s) - \Phi(s)) = A^{\pi,\gamma}(s,a). \tag{24}$$

Note that if $\Phi$ happens to be the state-value function $V^{\pi,\gamma}$ from the original MDP, then the transformed MDP has the interesting property that $\tilde{V}^{\pi,\gamma}(s)$ is zero at every state.

Note that (Ng et al., 1999) showed that the reward shaping transformation leaves the policy gradient and optimal policy unchanged when our objective is to maximize the discounted sum of rewards $\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1})$. In contrast, this paper is concerned with maximizing the undiscounted sum of rewards, where the discount $\gamma$ is used as a variance-reduction parameter.

Having reviewed the idea of reward shaping, let us consider how we could use it to get a policy gradient estimate. The most natural approach is to construct policy gradient estimators that use discounted sums of shaped rewards $\tilde{r}$. However, Equation (21) shows that we obtain the discounted sum of the original MDP's rewards $r$ minus a baseline term. Next, let's consider using a "steeper" discount $\gamma\lambda$, where $0 \le \lambda \le 1$. It's easy to see that the shaped reward $\tilde{r}$ equals the Bellman residual term $\delta^V$, introduced in Section 3, where we set $\Phi = V$. Letting $\Phi = V$, we see that

$$\sum_{l=0}^{\infty} (\gamma\lambda)^l \tilde{r}(s_{t+l}, a_t, s_{t+l+1}) = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}^V = \hat{A}_t^{\mathrm{GAE}(\gamma,\lambda)}. \tag{25}$$

Hence, by considering the $\gamma\lambda$-discounted sum of shaped rewards, we exactly obtain the generalized advantage estimators from Section 3. As shown previously, $\lambda = 1$ gives an unbiased estimate of $g^\gamma$, whereas $\lambda < 1$ gives a biased estimate.

To further analyze the effect of this shaping transformation and parameters $\gamma$ and $\lambda$, it will be useful to introduce the notion of a response function $\chi$, which we define as follows:

$$\chi(l; s_t, a_t) = \mathbb{E}\left[r_{t+l} \mid s_t, a_t\right] - \mathbb{E}\left[r_{t+l} \mid s_t\right]. \tag{26}$$

Note that $A^{\pi,\gamma}(s,a) = \sum_{l=0}^{\infty} \gamma^l \chi(l; s, a)$, hence the response function decomposes the advantage function across timesteps. The response function lets us quantify the temporal credit assignment problem: long range dependencies between actions and rewards correspond to nonzero values of the response function for $l \gg 0$.

Next, let us revisit the discount factor $\gamma$ and the approximation we are making by using $A^{\pi,\gamma}$ rather than $A^{\pi,1}$. The discounted policy gradient estimator from Equation (6) has a sum of terms of the form

$$\nabla_\theta \log \pi_\theta(a_t \mid s_t) A^{\pi,\gamma}(s_t, a_t) = \nabla_\theta \log \pi_\theta(a_t \mid s_t) \sum_{l=0}^{\infty} \gamma^l \chi(l; s_t, a_t). \tag{27}$$

Using a discount $\gamma < 1$ corresponds to dropping the terms with $l \gg 1/(1-\gamma)$. Thus, the error introduced by this approximation will be small if $\chi$ rapidly decays as $l$ increases, i.e., if the effect of an action on rewards is "forgotten" after $\approx 1/(1-\gamma)$ timesteps.

If the reward function $\tilde{r}$ were obtained using $\Phi = V^{\pi,\gamma}$, we would have $\mathbb{E}\left[\tilde{r}_{t+l} \mid s_t, a_t\right] = \mathbb{E}\left[\tilde{r}_{t+l} \mid s_t\right] = 0$ for $l > 0$, i.e., the response function would only be nonzero at $l = 0$. Therefore, this shaping transformation would turn temporally extended response into an immediate response. Given that $V^{\pi,\gamma}$ completely reduces the temporal spread of the response function, we can hope that a good approximation $V \approx V^{\pi,\gamma}$ partially reduces it. This observation suggests an interpretation of Equation (16): reshape the rewards using $V$ to shrink the temporal extent of the response function, and then introduce a "steeper" discount $\gamma\lambda$ to cut off the noise arising from long delays, i.e., ignore terms $\nabla_\theta \log \pi_\theta(a_t \mid s_t) \delta_{t+l}^V$ where $l \gg 1/(1-\gamma\lambda)$.