Appendix

A Hyperparameters in Algorithm 1

Hyper-parameters are presented below in the order of four main components— updating the critic, the density ratio, the emphatic weights, and the actor. $\alpha_\nu \in [0,1]$ is the stepsize in the critic update; $\alpha_\psi \in [0,1]$ is the stepsize in the density ratio update; $\lambda^{(1)} \in [0,1]$, $\lambda^{(2)} \in [0,1]$ and $\hat{\gamma} \in [0,1)$ can be found more details in Appendix B for the emphatic weights update; k, w, and β are inherited from STORM for the actor update. By default, w is set as 10 and $\beta = 100$.

B Emphatic weights update component of GeoffPAC [Zhang et al., 2019]

Figure 5 contains the updates for the emphatic weights in GeoffPAC. In this figure, $\lambda^{(1)}$ and $\lambda^{(2)}$ are parameters that are used for bias-variance tradeoff, $C(s) = \frac{d_{\hat{\gamma}}(s)}{d_\mu(s)}$ is the density ration function (Gelada and Bellemare 2019 call it covariate shift), and $i(s)$ is the intrinsic interest function that is defined from the extrinsic interest function $\hat{i}(s)$ as $i(s) = C(s)\hat{i}(s)$. In practice, $\hat{i}(s) = 1$. At time-step t, $F_t^{(1)}$ and $F_t^{(2)}$ are the follow-on traces, $M_t^{(1)}$ and $M_t^{(2)}$ are the emphatic weights, I_t is the gradient of the intrinsic interest, δ_t is the temporal-difference (TD) error, and finally Z_t is an unbiased sample of $\nabla J_{\hat{\gamma}}$. For more details about these parameters and their update formulas, we refer the reader to the GeoffPAC paper [Zhang et al., 2019].

HYPER-PARAMETER: $\lambda^{(1)}, \lambda^{(2)}$.
INPUT:$F_{t-1}^{(1)}, F_{t-1}^{(2)}, \rho_{t-1}, \rho_t, C(s_t;\psi_t), V(s_t;\nu_t), \delta_t, \hat{i}(s_t)$.
OUTPUT:$F_t^{(1)}, M_t^{(1)}, I_t, F_t^{(2)}, M_t^{(2)}, Z_t(a_t, s_t; \theta_t)$.
Compute $F_t^{(1)} = \gamma\rho_{t-1}F_{t-1}^{(1)} + \hat{i}(s_t)C(s_t;\psi_t)$.
Compute $M_t^{(1)} = (1-\lambda^{(1)})\hat{i}(s_t)C(s_t;\psi_t) + \lambda^{(1)}F_t^{(1)}$.
Compute $I_t = C(s_{t-1};\psi_{t-1})\rho_{t-1}\nabla_\theta \log\pi(a_{t-1}|s_{t-1};\theta_{t-1})$.
Compute $F_t^{(2)} = \hat{\gamma}\rho_{t-1}F_{t-1}^{(2)} + I_t$.
Compute $M_t^{(2)} = (1-\lambda^{(2)})I_t + \lambda^{(2)}F_t^{(2)}$.
Compute $Z_t(a_t, s_t; \theta_t) = \hat{\gamma}\hat{i}(s_t)V(s_t;\nu_t)M_t^{(2)} + \rho_t M_t^{(1)}\delta_t\nabla_\theta \log\pi(a_t|s_t;\theta_t)$.

Figure 5: Emphatic weights update component of GeoffPAC [Zhang et al., 2019]

C ACE-STORM Algorithm

The pseudo-code of ACE-STORM is shown in Algorithm 2.

D Comparison of Stochastic Variance Reduction Methods

This table is adapted from [Cutkosky and Orabona, 2019].

E Proof of Theorem 1

Before conducting the proof, we first denote ϵ_t: $\epsilon_t = g_t - \nabla J_{\hat{\gamma}}(\theta_t)$.

Lemma 1. *Suppose $\eta_t \leq \frac{1}{4L}$ for all t. Then*

$$\mathbb{E}\big[J_{\hat{\gamma}}(\theta_t) - J_{\hat{\gamma}}(\theta_{t+1})\big] \leq \mathbb{E}\big[-\eta_t/4\|\nabla J_{\hat{\gamma}}(\theta_t)\|^2 + 3\eta_t/4\|\epsilon_t\|^2\big]$$

Algorithms		Sample Complexity	Reference Sets Needed?
SVRG	[Reddi et al., 2016a] [Allen-Zhu and Hazan, 2016]	$O(n^{2/3}/\epsilon)$	$O(1/\epsilon)$
SARAH	[Nguyen et al., 2017a,b]	$O(n + 1/\epsilon^2)$	✓
SPIDER	[Fang et al., 2018]	$O(1/\epsilon^{3/2})$	✓
STORM	[Cutkosky and Orabona, 2019]	$O(1/\epsilon^{3/2})$	×

Table 2: Comparison of convergence rates to achieve $||\nabla J(x)||^2 \leq \epsilon$ for *nonconvex* objective functions.

Proof of Lemma 1. According to the smoothness of $J_{\hat{\gamma}}$,

$$\begin{aligned}
\big[-J_{\hat{\gamma}}(\theta_{t+1})\big] &\leq \mathbb{E}[-J_{\hat{\gamma}}(\theta_t) - \nabla J_{\hat{\gamma}}(\theta_t)\cdot\eta_t g_t + \frac{L\eta_t^2}{2}\|g_t\|^2] \\
&= \mathbb{E}[-J_{\hat{\gamma}}(\theta_t) - \eta_t\|\nabla J_{\hat{\gamma}}(\theta_t)\|^2 - \eta_t\nabla J_{\hat{\gamma}}(\theta_t)\cdot\epsilon_t + \frac{L\eta_t^2}{2}\|g_t\|^2] \\
&\leq \mathbb{E}[-J_{\hat{\gamma}}(\theta_t) - \frac{\eta_t}{2}\|\nabla J_{\hat{\gamma}}(\theta_t)\|^2 + \frac{\eta_t}{2}\|\epsilon_t\|^2 + \frac{L\eta_t^2}{2}\|g_t\|^2] \\
&\leq \mathbb{E}[-J_{\hat{\gamma}}(\theta_t) - \frac{\eta_t}{2}\|\nabla J_{\hat{\gamma}}(\theta_t)\|^2 + \frac{\eta_t}{2}\|\epsilon_t\|^2 + L\eta_t^2\|\epsilon_t\|^2 + L\eta_t^2\|\nabla J_{\hat{\gamma}}(\theta_t)\|^2] \\
&\leq \mathbb{E}[-J_{\hat{\gamma}}(\theta_t) - \frac{\eta_t}{2}\|\nabla J_{\hat{\gamma}}(\theta_t)\|^2 + \frac{3\eta_t}{4}\|\epsilon_t\|^2 + \frac{\eta_t}{4}\|J_{\hat{\gamma}}(\theta_t)\|^2
\end{aligned}$$

□

The following technical observation is key to our analysis: it provides a recurrence that enables us to bound the variance of the estimates g_t.

Algorithm 2 ACE-STORM

V: value function parameterized by ν
π: policy function parameterized by θ
Input: Initial parameters ν_0 and θ_0. Initialize $F_{-1}^{(1)} = 0$, $\rho_{-1} = 1$, $i(\cdot) = 1$, and hyper-parameters $\lambda^{(1)}$, k, w, β and α_ν.
for timestep $t = 0$ to T **do**
  Sample a transition S_t, A_t, R_t, S_{t+1} according to behavior policy μ.
  Compute $\delta_t = R_t + \gamma V(S_{t+1};\nu_t) - V(S_t;\nu_t)$
  Update the parameter for value function: $\nu_{t+1} = \nu_t + \alpha_\nu\delta_t\nabla_\nu V(S_t;\nu_t)$
  Compute $F_t^{(1)} = \gamma\rho_{t-1}F_{t-1}^{(1)} + i(S_t)$
  Compute $M_t^{(1)} = (1-\lambda^{(1)})i(S_t) + \lambda^{(1)}F_t^{(1)}$
  Compute $Z_t^{(1)}(A_t, S_t;\theta_t) = \rho_t M_t^{(1)}\delta_t\nabla_\theta\log\pi(A_t|S_t;\theta_t)$.
  Compute $G_t = ||Z_t^{(1)}(A_t, S_t;\theta_t)||$.
  Compute $\alpha_t = \beta\eta_{t-1}^2$
  Compute $Z_t^{(1)}(A_t, S_t;\theta_{t-1}) = \rho_t M_t^{(1)}\delta_t\nabla_\theta\log\pi(A_t|S_t;\theta_{t-1})$.
  Compute $g_t = Z_t^{(1)}(A_t, S_t;\theta_t) + (1-\alpha_t)\big(g_{t-1} - Z_t^{(1)}(A_t, S_t;\theta_{t-1})\big)$.
  Compute $\eta_t = \frac{k}{(w+\sum_{i=1}^t G_t^2)^{\frac{1}{3}}}$.
  Update the parameter for the actor: $\theta_{t+1} = \theta_t + \eta_t g_t$
end for
Output I: Parameters ν_{T+1}, θ_{T+1}.
Output II:Parameters ν_{T+1}, θ_τ, where τ is sampled with a probability of $p(\tau = t) \propto \frac{1}{\eta_t^2}$.

Lemma 2. *With the notation in Algorithm, we have*

$$\begin{aligned}&\mathbb{E}\big[\|\epsilon_t\|^2/\eta_{t-1}\big]\\ \leq&\mathbb{E}\big[2\beta^2\eta_{t-1}^3\sigma^2+(1-\alpha_t)^2(1+4L^2\eta_{t-1}^2)\|\epsilon_{t-1}\|^2/\eta_{t-1}+4(1-\alpha_t)^2L^2\eta_{t-1}\|\nabla J_{\hat{\gamma}}(\theta_{t-1})\|^2\big].\end{aligned}$$

The proof of Lemma 2 is motivated by the proof of Lemma 2 in [Cutkosky and Orabona, 2019].

Proof of Theorem 1. We first construct a Lyapunov function of $\Phi_t = J_{\hat{\gamma}}(\theta_t) + \frac{1}{32L^2\eta_{t-1}}\|\epsilon_t\|^2$. We will upper bound $\Phi_{t+1} - \Phi_t$ for each t, which will allow us to bound Φ_T in terms of Φ_1 by summing over t. First, observe that since $w \geq (4Lk)^3$, we have $\eta_t \leq \frac{1}{4L}$. Further, since $\alpha_{t+1} = \beta\eta_t^2$, we have $\alpha_{t+1} \leq \frac{\beta k}{4Lw^{1/3}} \leq 1$ for all t. Then, we first consider $\eta_t^{-1}\|\epsilon_{t+1}\|^2 - \eta_{t-1}^{-1}\|\epsilon_t\|^2$. Using Lemma 2, we obtain

$$\begin{aligned}&\mathbb{E}\big[\eta_t^{-1}\|\epsilon_{t+1}\|^2-\eta_{t-1}^{-1}\|\epsilon_t\|^2\big]\\ \leq&\mathbb{E}\big[2c^2\eta_t^3G^2+\frac{(1-\alpha_{t+1})^2(1+4L^2\eta_t^2)\|\epsilon_t\|^2}{\eta_t}+4(1-\alpha_{t+1})^2L^2\eta_t\|\nabla J_{\hat{\gamma}}(\theta_t)\|^2-\frac{\|\epsilon_t\|^2}{\eta_{t-1}}\big]\\ \leq&\mathbb{E}\Big[\underbrace{2c^2\eta_t^3G^2}_{A_t}+\underbrace{\big(\eta_t^{-1}(1-\alpha_{t+1})(1+4L^2\eta_t^2)-\eta_{t-1}^{-1}\big)\|\epsilon_t\|^2}_{B_t}+\underbrace{4L^2\eta_t\|\nabla J_{\hat{\gamma}}(\theta_t)\|^2}_{C_t}\Big].\end{aligned}$$

Let start with upper bounding the second term B_t we have

$$B_t \leq (\eta_t^{-1}-\eta_{t-1}^{-1}+\eta_t^{-1}(4L^2\eta_t^2-\alpha_{t+1}))\|\epsilon_t\|^2 = \big(\eta_t^{-1}-\eta_{t-1}^{-1}+\eta_t(4L^2-\beta)\big)\|\epsilon_t\|^2\,.$$

Let us focus on $\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}$ for a minute. Using the concavity of $x^{1/3}$, we have $(x+y)^{1/3} \leq x^{1/3} + yx^{-2/3}/3$. Therefore:

$$\begin{aligned}\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}} &= \frac{1}{k}\Big(w+\sum_{i=1}^t G_i^2\Big)^{1/3}-\frac{1}{k}\Big(w+\sum_{i=1}^{t-1}G_i^2\Big)^{1/3} \leq \frac{G_t^2}{3k(w+\sum_{i=1}^{t-1}G_i^2)^{2/3}}\\ &\leq \frac{G_t^2}{3k(w-G^2+\sum_{i=1}^t G_i^2)^{2/3}} \leq \frac{G_t^2}{3k(w/2+\sum_{i=1}^t G_i^2)^{2/3}}\\ &\leq \frac{2^{2/3}G_t^2}{3k(w+\sum_{i=1}^t G_i^2)^{2/3}} \leq \frac{2^{2/3}G_t^2}{3k^3}\eta_t^2 \leq \frac{2^{2/3}G^2}{12Lk^3}\eta_t \leq \frac{G^2}{7Lk^3}\eta_t\end{aligned}$$

where we have used that that $w \geq (4Lk)^3$ to have $\eta_t \leq \frac{1}{4L}$.

Further, since $\beta = 28L^2 + G^2/(7Lk^3)$, we have

$$\eta_t(4L^2-\beta) \leq -24L^2\eta_t - G^2\eta_t/(7Lk^3).$$

Thus, we obtain

$$B_t \leq -24L^2\eta_t\|\epsilon_t\|^2$$

Now, we are ready to analyze the potential Φ_t. Since $\eta_t \leq \frac{1}{4L}$, we can use Lemma 1 to obtain

$$\mathbb{E}[\Phi_t - \Phi_{t+1}] \leq \mathbb{E}\left[-\frac{\eta_t}{4}\|\nabla J_{\hat{\gamma}}(\theta_t)\|^2+\frac{3\eta_t}{4}\|\epsilon_t\|^2+\frac{1}{32L^2\eta_t}\|\epsilon_{t+1}\|^2-\frac{1}{32L^2\eta_{t-1}}\|\epsilon_t\|^2\right].$$

Summing over t, we obtain Rearranging terms we get,

$$\begin{aligned}&\mathbb{E}[\frac{\eta_t}{8}\|\nabla J_{\hat{\gamma}}(\theta_t)\|^2] \leq \mathbb{E}[\Phi_{t+1}-\Phi_t]+\mathbb{E}[\frac{\beta^2\eta_t^3G^2}{16L^2}]\\ \iff&\mathbb{E}[\frac{1}{8\eta_t^2}\|\nabla J_{\hat{\gamma}}(\theta_t)\|^2] \leq \mathbb{E}[\frac{1}{8\eta_t^3}[\Phi_{t+1}-\Phi_t]]+\frac{\beta^2G^2}{16L^2}\end{aligned}$$

Summing over $1, \cdots, t$, we have

$$\sum_{t=1}^{T} \mathbb{E}[\frac{1}{\eta_t^2}\|\nabla J_{\hat{\gamma}}(\theta_t)\|^2] \leq \sum_{t=1}^{T} \mathbb{E}[\frac{8}{\eta_t^3}[\Phi_{t+1} - \Phi_t]] + \frac{G^2 T}{2L^2}$$
$$\iff \sum_{t=1}^{T} \mathbb{E}[\frac{1}{\eta_t^2}\|\nabla J_{\hat{\gamma}}(\theta_t)\|^2] \leq \sum_{t=1}^{T} \mathbb{E}[\frac{8}{\eta_t^3}[\Phi_{t+1} - \Phi_t]] + \frac{\beta^2 G^2 T}{2L^2}$$
$$\iff \sum_{t=1}^{T} \mathcal{W}_{1t}\mathbb{E}[\|\nabla J_{\hat{\gamma}}(\theta_t)\|^2] \leq \sum_{t=1}^{T} 8\mathcal{W}_{2t}\mathbb{E}[\Phi_{t+1} - \Phi_t] + \frac{\beta^2 G^2 T}{2L^2}$$

As $G_{t+1}^2 \leq G^2$, therefore $\eta_t \sim \Omega((\frac{k}{w+tG^2})^{1/3})$. As a result, $\mathcal{W}_{1t} = \frac{1}{\eta_t^2} = \frac{(w+tG^2)^{2/3}}{k^2} \sim O(t^{2/3})$, $\mathcal{W}_{2t} = \frac{1}{\eta_t^3} = \frac{(w+tG^2)}{k^3} \sim O(t)$.

$$\sum_{t=1}^{T} t\mathbb{E}[\Phi_{t+1} - \Phi_t] = \sum_{t=1}^{T} \mathbb{E}[(t+1)\Phi_{t+1} - (t)\Phi_t] - \sum_{t=1}^{T} \Phi_{t+1}$$
$$= (T+1)\Phi_{T+1} - \Phi_1 - \sum_{t=1}^{T} \Phi_{t+1} = \sum_{t=1}^{T+1} (\Phi_{T+1} - \Phi_t) \leq (T+1)\Delta_\Phi$$

where $\Delta_\Phi \leq \Delta_{J_{\hat{\gamma}}} + \frac{\|\epsilon_0\|^2}{32\eta_0 L^2}$, $\Delta_{J_{\hat{\gamma}}} = J_{\hat{\gamma}}(\theta^*) - J_{\hat{\gamma}}(\theta), \forall \theta \in R^d$, and $\theta^\star$ is the maximizer of $J_{\hat{\gamma}}$.

$$\sum_{t=1}^{T} \mathcal{W}_{1t} = \sum_{t=1}^{T} t^{2/3} \geq \int_{t=1}^{T} t^{2/3} dt = \frac{3}{5}(T^{5/3} - 1) \geq \frac{2}{5}T^{5/3}.$$

Then we have

$$\frac{\sum_{t=1}^{T} \mathcal{W}_{1t}\mathbb{E}[\|\nabla J_{\hat{\gamma}}(\theta_t)\|^2}{\sum_{t=1}^{T} \mathcal{W}_{1t}} \leq \frac{\sum_{t=1}^{T} 8\mathcal{W}_{2t}\mathbb{E}[\Phi_t - \Phi_{t+1}]}{\sum_{t=1}^{T} \mathcal{W}_{1t}} + \frac{\beta^2 G^2 T}{2L^2 \sum_{t=1}^{T} \mathcal{W}_{1t}}$$
$$\leq \frac{8(T+1)\Delta_\Phi}{\frac{2}{5}(T^{5/3})} + \frac{\eta^2 G^2 T}{2L^2(\frac{2}{5}T^{5/3})}$$
$$\leq \frac{40\Delta_\Phi}{T^{2/3}} + \frac{2\beta^2 G^2}{L^2 T^{2/3}}$$

where $\beta = 28L^2 + \sigma^2/(7Lk^3)$.

□

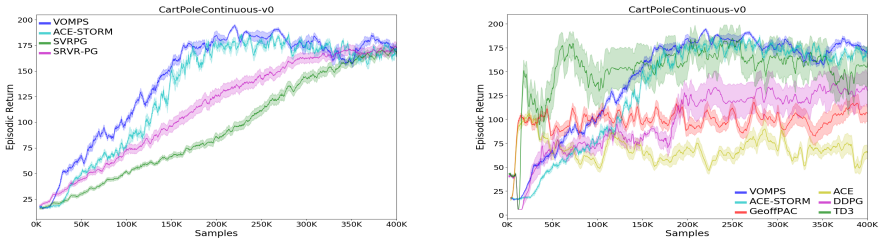
F Details of Experiments

For VOMPS and ACE-STORM, the policy function π is parameterized as a diagonal Gaussian distribution where the mean is the output of a two-hidden-layer network (64 hidden units with ReLU) and the standard deviation is fixed. For GeoffPAC, ACE, SVRPG, SRVR-PG, DDPG and TD3, we use the same parameterization as Zhang et al. [2019], Papini et al. [2018], Xu et al. [2019b], Lillicrap et al. [2015] and Fujimoto et al. [2018] respectively.

Cartpole `CartPoleContinuous-v0` has 4 dimensions for a state and 1 dimension for an action. The only difference between `CartPoleContinuous-v0` and `CartPole-v0` (provided by OpenAI Gym) is that `CartPoleContinuous-v0` has a continuous value range of $[-1, 1]$ for action space. The episodic return for the comparison with on-policy and off-policy methods is shown in Fig. 6(a), 6(b). The relative performance matches with that of the Monte Carlo return.

Hopper `Hopper-v2` attempts to make a 2D robot hop that has 11 dimensions for a state and 3 dimensions for an action. The episodic return for the comparison with on-policy and off-policy methods is shown in Fig. 7(a), 8(a).

HalfCheetah `HalfCheetah-v2` attempts to make a 2D cheetah robot run that has 17 dimensions for a state and 6 dimensions for an action. The episodic return for the comparison with on-policy and off-policy methods is shown in Fig. 7(b), 8(b).

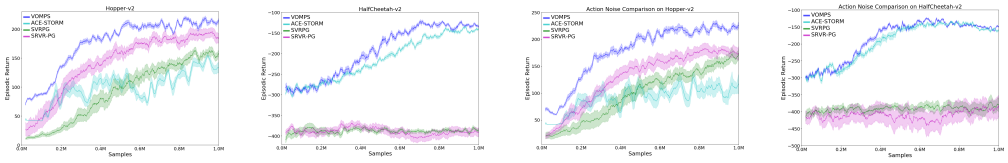


(a) Comparison with on-policy methods (b) Comparison with off-policy methods

Figure 6: Episodic Return on `CartPoleContinuous-v0`

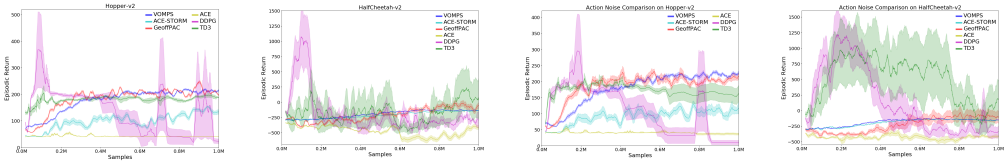
Besides, the episodic return for the 20% action noise comparison on Mujoco (including `Hopper-v2` and `HalfCheetah-v2`) is shown in Fig. 7(c), 8(c), 7(d), 8(d) respectively.

The parameter settings for GeoffPAC and ACE are insensitive on `CartPoleContinuous-v0`. Therefore, we keep the setting of $\lambda^{(1)} = 0.7$, $\lambda^{(2)} = 0.6$, $\hat{\gamma} = 0.2$ for GeoffPAC, and $\lambda^{(1)} = 0$ for ACE in all of the experiments. For DDPG and TD3, we use the same parameter settings as Lillicrap et al. [2015] and Fujimoto et al. [2018] respectively.



(a) Hopper (b) HalfCheetah (c) Hopper (action noise) (d) HC (action noise)

Figure 7: Comparison with on-policy PG methods (Mujoco), "HC" is short for HalfCheetah.



(a) Hopper (b) HalfCheetah (c) Hopper (action noise) (d) HC (action noise)

Figure 8: Comparison with off-policy PG methods (Mujoco), "HC" is short for HalfCheetah.