A. Notation

Table 2. Notation

π^b, π^θ	Behavior policy, Evaluation policy
H	Horizon
T	Final optimization step
∇, ∇_θ	Derivative w.r.t θ
$\mathcal{H}_{a_t}$	History up to a_t, $(s_0, a_0, \cdot, s_t, a_t)$ (Rewards are excluded)
$\mathcal{H}_{s_t}$	History up to s_t, $(s_0, a_0, \cdot, s_t)$ (Rewards are excluded)
$\mathcal{T}_{a_t}$	History up to a_t, $(s_0, a_0, r_0, \cdot, a_t)$ (Rewards are included)
$\mathcal{T}_{s_t}$	History up to s_t, $(s_0, a_0, r_0, \cdot, s_t, a_t)$ (Rewards are included)
$\mathbb{E}_\pi[\cdot]$	Expectation is taken w.r.t policy π
$\mathbb{E}_n[\cdot]$	Empirical approximation
$J(\theta)$	Value of π
$Z(\theta)$	$\nabla J(\theta)$
MDP, NMDP	DAG MDP, Tree MDP
$q_t(s, a)$	State action value function at t
$v_t(s)$	State Value function at t
$\nu_{0:t}(\mathcal{H}_{a_t})$	$\prod_{0=i}^{H} \pi_i^\theta / \pi_i^b$
$\nu_{a:b}$	$\prod_{i=a}^{b} \pi_i^\theta / \pi_i^b$
$\tilde{\nu}_t$	π_t^θ / π_t^b
$\tilde{\mu}_t$	$p_{\pi^\theta}(s_t) / p_{\pi^b}(s_t)$
$\mu_t(s, a)$	Ratio of marginal distribution at t
$d_t^\nu(s, a), d_t^\mu(s, a)$	$\nabla \mu_t(s, a), \nabla \mu_t(s, a$
$d_t^q(s, a), d_t^v(s)$	$\nabla q_t(s, a), \nabla v_t(s, a)$
$\otimes g$	$g g^\top$
g_t	Score function of the policy: $\nabla \log \pi_t^\theta(a_t \mid s_t)$
C_1, C_2 C_1'	Distribution mismatch constants
$R_{\max}$	$0 \leq r_t \leq R_{\max}$
$G_{\max}$	$\|g_t\|_{\mathrm{op}} \leq G_{\max}$
$\mathcal{F}$	Function class
$\|\cdot\|_{\mathrm{op}}$	Operator norm
$\|\cdot\|_{L_b^2}$	L^2-integral norm with respect to the behavior policy
$\|\cdot\|_2$	Euclidean norm
Θ	Parameter space
$A \preceq B$	$B - A$ is a semi-positive definite matrix
$\mathcal{I}_{D \times D}$	Identity matrix
$\xi_{\mathrm{MDP}}, \xi_{\mathrm{NMDP}}$	Efficient influence functions (IFs) of $Z(\theta)$ under MDP and NMDP
$\mathcal{U}_k$	k-th partitioned data
$\mathcal{Z}_k$	Data except for $\mathcal{U}_k$
Proj	projection
Θ	Parameter space
Υ	Diameter of Θ
D	Dimension of θ
$\langle \cdot, \cdot \rangle$	Inner product $a^\top b$
$a \lnapprox b$	Inequality up to absolute constant
$\mathcal{I}$	Identiry matrix

B. Semiparametric Theory for Off-Policy RL: A General Conditioning Framework

Here, we summarize a general framework to obtain efficiency bounds and efficient influence functions for various quantities of interest under NMDP or MDP, which we then use in order to derive these for the policy gradient case. First, we present the framework in generality. Then, we show how to use this framework to re-derive the efficiency bounds and efficient influence functions for OPE of Kallus & Uehara (2019b), who derived it for the first time but from scratch. Our proofs for our policy gradient case in the subsequent sections use the observations from this section.

B.1. General Conditioning Framework

B.1.1. General Semiparametric Inference

Consider observing n iid observations $O^{(i)} \sim P$ from some distribution P. We are interested in the estimand $R(P)$ where the unknown P is assumed to live in some (nonparametric) model $P \in \mathcal{M}$ and $R : \mathcal{M} \to \mathbb{R}^D$. Estimators of this estimand are functions of the data, $\hat{R} = \hat{R}(O^{(1)}, \dots, O^{(n)})$. Regular estimators are, roughly speaking, those for which the distribution of $\sqrt{n}(\hat{R} - R(P))$ converges to a limiting distribution in a locally uniform sense in $\mathcal{M}$ (van der Vaart, 1998, Chapter 25). Under certain differentiability conditions on $R(\cdot)$, the efficiency bound is the smallest asymptotic MSE (the second moment of the distributional limit of $\sqrt{n}(\hat{R} - R(P))$) among all regular estimators $\hat{R}$ (van der Vaart, 1998, Theorem 25.20), which also lower bounds the limit infimum of $n\mathbb{E}[(\hat{R} - R(P))^2]$ via Fatou's lemma. The efficiency bound even lower bounds the limit infimum of the MSE of *any* estimator in a local asymptotic minimax sense (van der Vaart, 1998, Theorem 25.21). In particular, the efficiency bound is given by $\mathrm{var}_P[\phi^*(O)]$ for some function $\phi^*(O)$.

Asymptotically linear estimators $\hat{R}$ are ones for which there exists a function $\phi(O)$ such that $\hat{R} = \mathbb{E}_n\phi + \mathrm{o}_p(n^{-1/2})$, $\mathbb{E}\phi = R(P)$.[2] The function ϕ is known as the influence function of $\hat{R}$. Clearly, the asymptotic MSE of $\hat{R}$ is $\mathrm{var}_P[\phi(O)]$. Thus, an asymptotic linear estimator would be efficient if its influence function were ϕ^*, which is called the *efficient influence function*. In fact, under the same differentiability conditions on $R(\cdot)$, efficient (regular) estimators are exactly those with the influence function ϕ^* (van der Vaart, 1998, Theorem 25.23). Under certain regularity, the set of influence functions (minus $R(P)$) is equal to the set of pathwise derivatives of $R(\cdot)$ at P, and the function ϕ^* is exactly given by that with minimal L_2 norm among this set (Bickel et al., 1993; Klaassen, 1987). Thus, ϕ^* can be gotten by a projection of *any* influence function, which is a generic recipe for deriving the efficient influence function and the efficiency bound.

B.1.2. A Conditioning Framework for Nonparametric Factorable Models

We now summarize how additional graphical structure on the variable O can further simplify the above recipe for deriving the efficient influence function in a particular class of models, which includes the MDP and NMDP models. Suppose each observation O has J component variables, $O = (O_1, \dots, O_J)$. Suppose moreover that we have some tree on the nodes $[J] = \{1, \dots, J\}$ described by the parentage relationship $\mathrm{Pa} : [J] \to 2^{[J]}$ mapping a node to its parents and such that P satisfies the factorization

$$P(O) = \prod_{j=1}^{J} P_j(O_j \mid O_{\mathrm{Pa}(j)}). \tag{9}$$

Consider the nonparametric model of all distributions that satisfy this factorization

$$\mathcal{M} = \left\{ Q \;:\; Q(O) = \prod_{j=1}^{J} Q_j(O_j \mid O_{\mathrm{Pa}(j)}) \text{ for some conditional distributions } Q_j \right\}.$$

Then, a standard result (see van Der Laan & Robins, 2003, van der Laan & Rose, 2018, §A.7) is that, given any ϕ that is a valid influence function for $R(P)$ in $\mathcal{M}$, the efficient influence function for $R(P)$ is given by

$$\phi^*(O) - R(P) = \sum_{j=1}^{J} \left(\mathbb{E}[\phi(O) \mid O_j, O_{\mathrm{Pa}(j)}] - \mathbb{E}[\phi(O) \mid O_{\mathrm{Pa}(j)}] \right).$$

This arises due to the above-mentioned projection interpretation of the efficient influence function.

[2] Note that conventionally one restricts $\mathbb{E}\phi = 0$ and writes $\hat{R} - R(P) = \mathbb{E}_n\phi + \mathrm{o}_p(n^{-1/2})$, but we deviate slightly here for clearer and more succinct presentation in the main text.

Now, suppose that the estimand only depends on a particular part of the factorization:

$$R(Q) = R(Q') \;\text{ whenever } Q_j = Q'_j \text{ for all } j \in J_0, \tag{10}$$

for some index set $J_0 \subseteq [J]$. That is, $R(Q)$ is only a function of $Q_{J_0} = (Q_j)_{j \in J_0}$ and is independent of $Q_{J_0^C} = (Q_j)_{j \notin J_0}$. Consider the model where we assume that P_j is *known* for every $j \notin J_0$,

$$\mathcal{M}_0 = \left\{ Q \;:\; Q(O) = \prod_{j=1}^{J} Q_j(O_j \mid O_{\mathrm{Pa}(j)}) \;\text{ for some } Q_{J_0} \text{ and } Q_{J_0^C} = P_{J_0^C} \right\}.$$

Then, as long as $R(\cdot)$ satisfies Eq. (10), its efficient influence function under $\mathcal{M}$ and $\mathcal{M}_0$ must be the same (similarly for the efficiency bound).

Combining the above observations, we have that if (a) our model satisfies the nonparametric factorization as in Eq. (9) and (b) our estimand only depends on some subset J_0 of the factorization as in Eq. (10), then given any ϕ that is a valid influence function for $R(P)$ in $\mathcal{M}_0$, the efficient influence function for $R(P)$ under $\mathcal{M}$ is in fact also just given by

$$\phi^*(O) - R(P) = \sum_{j \in J_0^C} \left(\mathbb{E}[\phi(O) \mid O_j, O_{\mathrm{Pa}(j)}] - \mathbb{E}[\phi(O) \mid O_{\mathrm{Pa}(j)}] \right). \tag{11}$$

B.2. Application to Off-Policy RL

In off-policy RL, our data are observations of trajectories $\mathcal{T} = (s_0, a_0, r_0, \ldots, s_T, a_H, r_H, s_{H+1})$ generated by the behavior policy. Here $\mathcal{T}$ stands for a single observation (above O in the general case) and s_t, a_t, r_t are individual components (above O_j in the general case). Moreover, in the MDP model, the data-generating distribution satisfies a factorization like Eq. (9):

$$p_{\pi^b}(\mathcal{T}) = p_0(s_0) \prod_{t=0}^{H} \pi_t^b(a_t \mid s_t) p_t(r_t \mid s_t, a_t) p_t(s_{t+1} \mid s_t, a_t).$$

Finally, we have that off-policy quantities such as the policy value and policy gradient for π^θ are independent of the behavior policy, that is, satisfy Eq. (10) where J_0^C corresponds to the $\pi_t^b(a_t \mid s_t)$ part in the factorization above. Here, the model $\mathcal{M}_0$ would correspond to the model where the behavior policy is known (and indeed the efficiency bound is independent of whether it is known or not).

Similarly, in the NMDP model we have an alternative factorization, where each node's parent set is much larger:

$$p_{\pi^b}(\mathcal{T}) = p_0(s_0) \prod_{t=0}^{H} \pi_t^b(a_t \mid \mathcal{H}_{s_t}) p_t(r_t \mid \mathcal{H}_{a_t}) p_t(s_{t+1} \mid \mathcal{H}_{a_t}).$$

Again, off-policy quantities of interest are independent of of the behavior policy.

These observations imply that in order to derive the efficient influence function (and hence the efficiency bound) for any appropriate off-policy quantity, we simply need to identify one valid influence function in $\mathcal{M}_0$ and then compute Eq. (11). This is exactly the approach we take in our proofs for the policy gradient.

Before proceeding to our proofs, which for the first time derive the efficiency bounds for off-policy gradients, as an illustrative case we first show how we can use this framework to derive the efficient influence functions and efficiency bounds for $J(\theta)$ under MDP and NMDP, which was first derived by Kallus & Uehara (2019b).

Example 5 (Off-policy evaluation in MDP)**.** *First we derive the efficient influence function. Under the model $\mathcal{M}_0$ where the behavior policy is known we know that $J(\theta) = \mathbb{E}\left[\sum_{t=0}^{H} \nu_{0:t} r_t\right]$ and therefore $\hat{J}(\theta) = \mathbb{E}_n\left[\sum_{t=0}^{H} \nu_{0:t} r_t\right]$ is a consistent linear estimator for $J(\theta)$. Hence, $\phi(\mathcal{T}) = \left[\sum_{t=0}^{H} \nu_{0:t} r_t\right]$ must be a valid influence function. Plugging into the right-hand* *side of Eq.* (11)*, we obtain:*

$$\sum_{j=0}^{H}\left\{\mathbb{E}[\sum_{t=0}^{H}\nu_{0:t}r_t \mid r_j, s_j, a_j] - \mathbb{E}[\sum_{t=0}^{H}\nu_{0:t}r_t \mid s_j, a_j] + \mathbb{E}[\sum_{t=0}^{H}\nu_{0:t}r_t|s_j, a_{j-1}, s_{j-1}] - \mathbb{E}[\sum_{t=0}^{H}\nu_{0:t}r_t|a_{j-1}, s_{j-1}]\right\}$$

$$= \sum_{j=0}^{H}\{\mathbb{E}[\nu_{0:j}|s_j, a_j]r_j - \mathbb{E}[\nu_{0:j}r_j|s_j, a_j] + \mathbb{E}[\sum_{t=j}^{H}\nu_{0:t}r_t \mid s_j, a_j, s_{j-1}] - \mathbb{E}[\sum_{t=j}^{H}\nu_{0:t}r_t \mid s_{j-1}, a_{j-1}]\}$$

$$= \sum_{j=0}^{H}\{\mu_j r_j + \mathbb{E}[\sum_{t=j}^{H}\nu_{0:t}r_t \mid s_j, a_j, s_{j-1}] - \mathbb{E}[\sum_{t=j-1}^{H}\nu_{0:t}r_t \mid s_{j-1}, a_{j-1}] - \mathbb{E}[\nu_{0:H}r_H|s_T, a_H]\}$$

$$= \sum_{j=0}^{H}\{\mu_j r_j + \mathbb{E}[\sum_{t=j}^{H}\nu_{0:t}r_t \mid s_j, a_{j-1}, s_{j-1}] - \mathbb{E}[\sum_{t=j}^{H}\nu_{0:t}r_t \mid s_j, a_j]\}$$

$$= \sum_{j=0}^{H}\{\mu_j r_j + \mathbb{E}[\nu_{0:j-1}|s_j, a_{j-1}, s_{j-1}]\mathbb{E}[\sum_{t=j}^{H}\nu_{j:t}r_t \mid s_j] - \mathbb{E}[\nu_{0:j} \mid s_j, a_j]\mathbb{E}[\sum_{t=j}^{H}\nu_{j+1:t}r_t \mid s_j, a_j]\} - J(\theta)$$

$$= \sum_{j=0}^{H}\{\mu_j r_j + \mathbb{E}[\nu_{0:j-1}|s_j, a_{j-1}, s_{j-1}]\mathbb{E}[\sum_{t=j}^{H}\nu_{j:t}r_t \mid s_j] - \mathbb{E}[\nu_{0:j} \mid s_j, a_j]\mathbb{E}[\sum_{t=j}^{H}\nu_{j+1:t}r_t \mid s_j, a_j]\} - J(\theta)$$

$$= v_0(s_0) + \sum_{j=0}^{H}\mu_j(s_j, a_j)\{r_j + v_{j+1}(s_{j+1}) - q_j(s_j, a_j)\} - J(\theta).$$

And therefore the efficient influence function is

$$\phi^*(\mathcal{T}) = v_0(s_0) + \sum_{j=0}^{H}\mu_j(s_j, a_j)\{r_j + v_{j+1}(s_{j+1}) - q_j(s_j, a_j)\}.$$

The efficiency bound is given by its variance. This matches Kallus & Uehara (2019b).

Example 6 (Off-policy evaluation in NMDP)**.** *We repeat the above in the NMDP case. Again, we know that $\hat{J}(\theta) = \mathbb{E}_n\left[\sum_{t=0}^{H}\nu_{0:t}r_t\right]$ is still a consistent linear estimator for $J(\theta)$. Hence, $\phi(\mathcal{T}) = \left[\sum_{t=0}^{H}\nu_{0:t}r_t\right]$ must be a valid influence function. Plugging into the right-hand side of Eq.* (11)*, we obtain:*

$$\sum_{j=0}^{H}\left\{\mathbb{E}[\sum_{t=0}^{H}\nu_{0:t}r_t \mid r_j, h_{a_j}] - \mathbb{E}[\sum_{t=0}^{H}\nu_{0:t}r_t \mid \mathcal{H}_{a_j}] + \mathbb{E}[\sum_{t=0}^{H}\nu_{0:t}r_t|s_j, \mathcal{H}_{a_{j-1}}] - \mathbb{E}[\sum_{t=0}^{H}\nu_{0:t}r_t|\mathcal{H}_{a_{j-1}}]\right\}$$

$$= \sum_{j=0}^{H}\{\mathbb{E}[\nu_{0:j}|\mathcal{H}_{a_j}]r_j - \mathbb{E}[\nu_{0:j}r_j|\mathcal{H}_{a_j}] + \mathbb{E}[\sum_{t=j}^{H}\nu_{0:t}r_t \mid s_j, \mathcal{H}_{a_{j-1}}] - \mathbb{E}[\sum_{t=j}^{H}\nu_{0:t}r_t \mid \mathcal{H}_{a_{j-1}}]\}$$

$$= \sum_{j=0}^{H}\{\nu_{0:j}r_j + \mathbb{E}[\sum_{t=j}^{H}\nu_{0:t}r_t \mid \mathcal{H}_{s_j}] - \mathbb{E}[\sum_{t=j-1}^{H}\nu_{0:t}r_t \mid \mathcal{H}_{a_{j-1}}] - \mathbb{E}[\nu_{0:H}r_H|\mathcal{H}_{a_H}]\}$$

$$= \sum_{j=0}^{H}\{\nu_{0:j}r_j + \mathbb{E}[\sum_{t=j}^{H}\nu_{0:t}r_t \mid \mathcal{H}_{s_j}] - \mathbb{E}[\sum_{t=j}^{H}\nu_{0:t}r_t \mid \mathcal{H}_{a_j}]\} - J(\theta)$$

$$= \sum_{j=0}^{H}\{\nu_{0:j}r_j + \mathbb{E}[\nu_{0:j-1}|\mathcal{H}_{s_j}]\mathbb{E}[\sum_{t=j}^{H}\nu_{j:t}r_t \mid \mathcal{H}_{s_j}] - \mathbb{E}[\nu_{0:j} \mid \mathcal{H}_{a_j}]\mathbb{E}[\sum_{t=j}^{H}\nu_{j+1:t}r_t \mid \mathcal{H}_{a_j}]\} - J(\theta)$$

$$= v_0(s_0) + \sum_{j=0}^{H}\nu_{0:j}\{r_j + v_{j+1}(\mathcal{H}_{s_{j+1}}) - q_j(\mathcal{H}_{a_j})\} - J(\theta).$$

And therefore the efficient influence function is

$$\phi^*(\mathcal{T}) = v_0(s_0) + \sum_{j=0}^{H}\nu_{0:j}\{r_j + v_{j+1}(\mathcal{H}_{s_{j+1}}) - q_j(\mathcal{H}_{a_j})\}.$$

The efficiency bound is given by its variance. This matches Jiang & Li (2016); Kallus & Uehara (2019b); Thomas & Brunskill (2016).

C. Proofs

Proof of Theorem 2. **Part 1: deriving the efficient influence function.** We use the general framework from Appendix B. Let $\overline{g}_t = \sum_{i=0}^{t} g_i$. Noting that $Z(\theta) = \mathbb{E}\left[\sum_{t=0}^{H} r_t \nu_{0:t} \overline{g}_t\right]$, we see that $\sum_{t=0}^{H} r_t \nu_{0:t} \overline{g}_t$ is an influence function for $Z(\theta)$ in the model where the behavior policy is known. Plugging it into the right-hand-side of Eq. (11), we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{H} r_t \nu_{0:t} \overline{g}_t\right] \\
&= \sum_{j=0}^{H} \{ \mathbb{E}\left[\sum_{t=0}^{H} r_t \nu_{0:t} \overline{g}_t \mid r_j, s_j, a_j\right] - \mathbb{E}\left[\sum_{t=0}^{H} r_t \nu_{0:t} \overline{g}_t \mid s_j, a_j\right] + \mathbb{E}\left[\sum_{t=0}^{H} r_t \nu_{0:t} \overline{g}_t \mid s_j, a_{j-1}, s_{j-1}\right] \\
&- \mathbb{E}\left[\sum_{t=0}^{H} r_t \nu_{0:t} \overline{g}_t \mid a_{j-1}, s_{j-1}\right] \} \\
&= \sum_{j=0}^{H} \{ \mathbb{E}\left[\nu_{0:j} \overline{g}_j \mid s_j, a_j\right] r_j - \mathbb{E}\left[\nu_{0:j} \overline{g}_j r_j \mid s_j, a_j\right] + \mathbb{E}\left[\sum_{t=j}^{H} r_t \nu_{0:t} \overline{g}_t \mid s_j, a_{j-1}, s_{j-1}\right] \\
&- \mathbb{E}\left[\sum_{t=j}^{H} r_t \nu_{0:t} \overline{g}_t \mid a_{j-1}, s_{j-1}\right] \} \\
&= \sum_{j=0}^{H} \left\{ \mathbb{E}\left[\nu_{0:j} \overline{g}_j \mid s_j, a_j\right] r_j - \mathbb{E}\left[\sum_{t=j}^{H} r_t \nu_{0:t} \overline{g}_t \mid s_j, a_{j-1}, s_{j-1}\right] + \mathbb{E}\left[\sum_{t=j}^{H} r_t \nu_{0:t} \overline{g}_t \mid a_j, s_j\right] \right\} - Z(\theta).
\end{aligned}
$$

Then, by substituting $\overline{g}_t = \sum_{i=0}^{t} g_i$, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=j}^{H} r_t \nu_{0:t} \left\{\sum_{i=0}^{t} g_i\right\} \mid a_j, s_j\right] \\
&= \mathbb{E}\left[\sum_{t=j}^{H} r_t \nu_{0:t} \left\{\sum_{i=j+1}^{t} g_i\right\} \mid a_j, s_j\right] + \mathbb{E}\left[\sum_{t=j}^{H} r_t \nu_{0:t} \left\{\sum_{i=0}^{j} g_i\right\} \mid a_j, s_j\right] \\
&= \mathbb{E}[\nu_{0:j} | a_j, s_j] \mathbb{E}\left[\sum_{t=j}^{H} r_t \nu_{j+1:t} \left\{\sum_{i=j+1}^{t} g_i\right\} \mid a_j, s_j\right] + \mathbb{E}\left[\nu_{0:j} \left\{\sum_{i=0}^{j} g_i\right\} \mid a_j, s_j\right] \mathbb{E}\left[\sum_{t=j}^{H} r_t \nu_{j+1:t} \mid a_j, s_j\right].
\end{aligned}
$$

In addition,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=j}^{H} r_t \nu_{0:t} \left\{\sum_{i=0}^{t} g_i\right\} \mid s_j, a_{j-1}, s_{j-1}\right] \\
&= \mathbb{E}\left[\sum_{t=j}^{H} r_t \nu_{0:t} \left\{\sum_{i=j+1}^{t} g_i\right\} \mid s_j, a_{j-1}, s_{j-1}\right] + \mathbb{E}\left[\sum_{t=j}^{H} r_t \nu_{0:t} \left\{\sum_{i=0}^{j-1} g_i\right\} \mid s_j, a_{j-1}, s_{j-1}\right] \\
&+ \mathbb{E}\left[\sum_{t=j}^{H} r_t \nu_{0:t} g_j \mid s_j, a_{j-1}, s_{j-1}\right] \\
&= \mathbb{E}[\nu_{0:j-1}|a_{j-1}, s_{j-1}]\mathbb{E}\left[\sum_{t=j}^{H} r_t \nu_{j:t} \left\{\sum_{i=j+1}^{t} g_i\right\} \mid s_j\right] + \mathbb{E}[\nu_{0:j-1}\left\{\sum_{i=0}^{j-1} g_i\right\}|a_{j-1}, s_{j-1}]\mathbb{E}\left[\sum_{t=j}^{H} r_t \nu_{j:t} \mid s_j\right] \\
&+ \mathbb{E}_{\pi^\theta}[Q_j g_j \mid s_j].
\end{aligned}
$$

To sum up, the efficient influence function of $Z(\theta)$ under MDP is

$$
\begin{aligned}
&\sum_{j=0}^{H}\{d^{\mu}_j(s_j, a_j) r_j - \mu_j(s_j, a_j) d^{q}_j(s_j, a_j) - d^{\mu}_j(s_j, a_j) q_j(s_j, a_j) \\
&\qquad + \mu_{j-1}(s_{j-1}, a_{j-1}) d^{v}_j(s_j) + d^{\mu}_{j-1}(s_{j-1}, a_{j-1}) v_j(s_j)\},
\end{aligned}
\tag{12}
$$

where

$$
\begin{aligned}
\mathbb{E}[\nu_{0:j} \mid s_j, a_j] &= \mu_j(s_j, a_j), \\
\mathbb{E}\left[\sum_{t=j+1}^{H} r_t \nu_{j+1:t} \left\{\sum_{i=j+1}^{t} g_i\right\} \mid s_j, a_j\right] &= d^{q}_j(s_j, a_j), \\
\mathbb{E}\left[\nu_{0:j} \left\{\sum_{i=0}^{j} g_i\right\} \mid s_j, a_j\right] &= d^{\mu}_j(s_j, a_j), \\
\mathbb{E}_{\pi^\theta}[d^{q}_j(s_j, a_j) + q_j g_j \mid s_j] &= d^{v}(s_j).
\end{aligned}
$$

Part 2: calculating the variance. Next, we calculate the variance of the efficient influence function using a law of total variance:

$$
\begin{aligned}
&\mathrm{var}\left[d^{\mu}_0(s_0, a_0) + \sum_{j=0}^{H} d^{\mu}_j(s_j, a_j)\{r_j - q_j(s_j, a_j) + v_{j+1}(s_{j+1})\} + \mu_j(s_j, a_j)\{d^{v}_{j+1}(s_{j+1}) - d^{q}_j(s_j, a_j)\}\right] \\
&= \sum_{k=0}^{H+1} \mathbb{E}\left[\mathrm{var}\left[\mathbb{E}[d^{\mu}_0(s_0, a_0) + \sum_{j=0}^{H} d^{\mu}_j(s_j, a_j)\{r_j - q_j(s_j, a_j) + v_{j+1}(s_{j+1})\} + \mu_j(s_j, a_j)\{d^{v}_{j+1}(s_{j+1}) - d^{q}_j(s_j, a_j)\}|\mathcal{T}_{a_k}]|\mathcal{T}_{a_{k-1}}\right]\right] \\
&= \sum_{k=0}^{H+1} \mathbb{E}\left[\mathrm{var}\left[\mathbb{E}[\sum_{j=k-1}^{H} d^{\mu}_j(s_j, a_j)\{r_j - q_j(s_j, a_j) + v_{j+1}(s_{j+1})\} + \mu_j(s_j, a_j)\{d^{v}_{j+1}(s_{j+1}) - d^{q}_j(s_j, a_j)\}|\mathcal{T}_{a_k}]|\mathcal{T}_{a_{k-1}}\right]\right] \\
&= \sum_{k=0}^{H+1} \mathbb{E}\left[\mathrm{var}\left[d^{\mu}_{k-1}(s_{k-1}, a_{k-1})\{r_{k-1} - q_{k-1}(s_{k-1}, a_{k-1}) + v_k(s_k)\} + \mu_{k-1}(s_{k-1}, a_{k-1})\{d^{v}_k(s_k) - d^{q}_{k-1}(s_{k-1}, a_{k-1})\}|\mathcal{T}_{a_{k-1}}\right]\right] \\
&= \sum_{k=0}^{H+1} \mathbb{E}\left[\mathrm{var}\left[d^{\mu}_{k-1}(s_{k-1}, a_{k-1})\{r_{k-1} - q_{k-1}(s_{k-1}, a_{k-1}) + v_k(s_k)\} + \mu_{k-1}(s_{k-1}, a_{k-1})\{d^{v}_k(s_k) - d^{q}_{k-1}(s_{k-1}, a_{k-1})\}|s_{k-1}, a_{k-1}\right]\right].
\end{aligned}
$$

From the third line to the fourth line, we have used the following Bellman equations:

$$0=\mathbb{E}[r_k-q_k+v_{k+1}\mid s_k,a_k],\quad 0=\mathbb{E}[-d^q_k+d^v_{k+1}\mid s_k,a_k].$$

Next, note that

$$d^{\mu}_j(s,a)=\mu_j(s,a)\nabla\log p^{\pi^\theta}_j(s,a).$$

Therefore, the variance is written as

$$\sum_{k=0}^{H}\mathbb{E}[\mu^2_{k-1}(s_{k-1},a_{k-1})\mathrm{var}[\nabla\log p^{\pi^\theta}_{k-1}(s_{k-1},a_{k-1})\{r_{k-1}-q_{k-1}(s_{k-1},a_{k-1})+v_k(s_k)\}$$
$$+\{d^v_k(s_k)-d^q_{k-1}(s_{k-1},a_{k-1})\}|s_{k-1},a_{k-1}]]. \tag{13}$$

Remark 6 (More specific presentation of the variance)**.** *Note that by covariance formula, the above efficiency bound is equal to*

$$\sum_{k=0}^{H+1}\mathbb{E}\left[\mu^2_{k-1}(s_{k-1},a_{k-1})\{\otimes\nabla\log p^{\pi^\theta}_{k-1}(s_{k-1},a_{k-1})\}\mathrm{var}\left[r_{k-1}|s_{k-1},a_{k-1}\right]\right]$$
$$+\sum_{k=0}^{H+1}\mathbb{E}\left[\mu^2_{k-1}(s_{k-1},a_{k-1})\nabla\log p^{\pi^\theta}_{k-1}(s_{k-1},a_{k-1})\mathbb{E}\left[\{r_{k-1}-q_{k-1}(s_{k-1},a_{k-1})+v_k(s_k)\}d^v_k(s_k)^\top|s_{k-1},a_{k-1}\right]\right]$$
$$+\sum_{k=0}^{H+1}\mathbb{E}\left[\mu^2_{k-1}(s_{k-1},a_{k-1})\mathrm{var}\left[d^v_k(s_k)|s_{k-1},a_{k-1}\right]\right].$$

Part 3: a simple bound for the variance.

Consider the on-policy case when $\mu_t=1$. Then, from (13), the efficiency bound of $Z(\theta)$ under MDP is

$$\textstyle\sum_{k=0}^{H+1}\mathbb{E}_{p^{\pi^\theta}}\left[\mathrm{var}\left[\nabla\log p^{\pi^\theta}_{k-1}(s_{k-1},a_{k-1})\{r_{k-1}-q_{k-1}(s_{k-1},a_{k-1})+v_k(s_k)\}+\{d^v_k(s_k)-d^q_{k-1}(s_{k-1},a_{k-1})\}|s_{k-1},a_{k-1}\right]\right]. \tag{14}$$

Since this is the lower bound regarding asymptotic MSE among regular estimators of $Z(\theta)$, it is smaller than the variance of

$$\textstyle\sum_{t=0}^{H}r_t\sum_{k=0}^{t}g_k(s_k\mid a_k),$$

noting $\mathbb{E}_n[\sum_{t=0}^{H}r_t\sum_{k=0}^{t}g_k(s_k\mid a_k)]$ is an asymptotic linear estimator. The variance of this estimator is bounded by

$$\begin{aligned}\mathrm{var}_{p^{\pi^\theta}}\left[\textstyle\sum_{t=0}^{H}r_t\sum_{k=0}^{t}g_k(s_k\mid a_k)\right]&\preceq R^2_{\max}\mathrm{var}_{p^{\pi^\theta}}\left[\sum_{t=0}^{H}\sum_{k=0}^{t}g_k(s_k\mid a_k)\right]\\&=R^2_{\max}\sum_{t=0}^{H}\sum_{k=0}^{t}\mathrm{var}_{p^{\pi^\theta}}\left[g_k(s_k\mid a_k)\right]\\&\preceq R^2_{\max}G_{\max}\left\{\frac{(H+1)(H+2)}{2}\right\}^2\mathcal{I}_{D\times D}.\end{aligned}\tag{15}$$

Here, from the first line to the second line, we use the fact that the covariance across the time is zero:

$$\mathrm{cov}_{p^{\pi^\theta}}[g_k(s_k\mid a_k),g_j(s_j\mid a_j)]=0,\ (k\neq j),$$

since when $k<j$

$$\mathrm{cov}_{p^{\pi^\theta}}[g_k(s_k\mid a_k),g_j(s_j\mid a_j)]=\mathbb{E}_{p^{\pi^\theta}}[g_kg_j]-\mathbb{E}_{p^{\pi^\theta}}[g_k]\mathbb{E}_{p^{\pi^\theta}}[g_j]=\mathbb{E}_{p^{\pi^\theta}}[g_k\mathbb{E}_{p^{\pi^\theta}}[g_j\mid s_j,a_j]]=0.$$

Therefore, the quantity (14) is also bounded by RHS of (15).

Let us go back to the general off–policy case. For any functions $k(s_t,a_t)$ taking a real number, by importance sampling, we have

$$\mathbb{E}_{p^{\pi^b}}[\mu_t^2(s_t,a_t)k(s_t,a_t)] = \mathbb{E}_{p^{\pi^\theta}}[\mu_t(s_t,a_t)k(s_t,a_t)] \leq \mathbb{E}_{p^{\pi^\theta}}[k(s_t,a_t)]C_2,$$

since μ_t is upper bounded by C_2. Therefore, noting the difference of (13) and (14), the quantity $\|\mathrm{var}[\xi_{\mathrm{MDP}}]\|_{\mathrm{op}}$ is upper-bounded by

$$C_2R_{\max}^2G_{\max}\left\{\tfrac{(H+1)(H+2)}{2}\right\}^2.$$

□

Proof of Theorem 3. We omit the proof of the first and second parts since it is almost the same as Theorem 3, where we simply replace $\mu_t(s_t,a_t), q_t(s_t,a_t), d_t^q(s_t,a_t), d_t^\mu(s_t,a_t)$ with $\nu_{0:t}(\mathcal{H}_{a_t}), q_t(\mathcal{H}_{a_t}), d_t^q(\mathcal{H}_{a_t}), d_t^\nu(\mathcal{H}_{a_t})$. Then, based on (12), the efficient influence function of $Z(\theta)$ under NMDP is

$$\begin{aligned}\xi_{\mathrm{NMDP}} = \sum_{j=0}^{H}\{&d_j^\nu(\mathcal{H}_{a_j})r_j - \nu_{0:j}(\mathcal{H}_{a_j})d_j^q(\mathcal{H}_{a_j}) - d_j^\nu(\mathcal{H}_{a_j})q_j(\mathcal{H}_{a_j})\\ &+\nu_{0:j-1}(\mathcal{H}_{a_{j-1}})d_j^v(\mathcal{H}_{s_j}) + d_{j-1}^\nu(\mathcal{H}_{a_{j-1}})v(\mathcal{H}_{s_j})\}.\end{aligned}$$

The efficiency bound of $Z(\theta)$ under NMDP is

$$\textstyle\sum_{k=0}^{H+1}\mathbb{E}[\nu_{k-1}^2(\mathcal{H}_{a_{k-1}})\mathrm{var}[\nabla\log p_{k-1}^{\pi^\theta}(\mathcal{H}_{a_{k-1}})\{r_{k-1}-q_{k-1}(\mathcal{H}_{a_{k-1}})+v_k(\mathcal{H}_{s_k})\}+\{d_k^v(\mathcal{H}_{s_k})-d_{k-1}^q(\mathcal{H}_{a_{k-1}})\}|\mathcal{H}_{a_{k-1}}]], \tag{16}$$

where $\nabla\log p_k^{\pi^\theta}(\mathcal{H}_{a_k}) = \sum_{j=0}^k g_j(\mathcal{H}_{a_j})$. Again, consider the on-policy case where $\nu_{0:t}=1$. Then, the above is equal to

$$\textstyle\sum_{k=0}^{H+1}\mathbb{E}_{p^{\pi^\theta}}[\mathrm{var}[\nabla\log p_{k-1}^{\pi^\theta}(\mathcal{H}_{a_{k-1}})\{r_{k-1}-q_{k-1}(\mathcal{H}_{a_{k-1}})+v_k(\mathcal{H}_{s_k})\}+\{d_k^v(\mathcal{H}_{s_k})-d_{k-1}^q(\mathcal{H}_{a_{k-1}})\}|\mathcal{H}_{a_{k-1}}]]. \tag{17}$$

Again, this quantity is bounded by RHS of (15). Go back to the general off-policy case. For any functions $k(s_t,a_t)$ taking a real number, by importance sampling, we have

$$\mathbb{E}_{p^{\pi^b}}[\nu_t^2(s_t,a_t)k(s_t,a_t)] = \mathbb{E}_{p^{\pi^\theta}}[\nu_{0:t}(s_t,a_t)k(s_t,a_t)] \leq \mathbb{E}_{p^{\pi^\theta}}[k(s_t,a_t)]C_1^t,$$

noting $\nu_{0:t}\leq C_1^t$. Therefore, noting the difference of (13) and (14), the term $\|\mathrm{var}[\xi_{\mathrm{MDP}}]\|_{\mathrm{op}}$ is upper-bounded by

$$C_1^HR_{\max}^2G_{\max}\left\{\tfrac{(H+1)(H+2)}{2}\right\}^2.$$

□

Proof of Theorem 4. The efficiency bound of $Z(\theta)$ under NMDP is written as $\sum_{k=0}^{H+1}\mathbb{E}\left[\nu_{k-1}^2\alpha_{k-1}(\mathcal{H}_{k-1})\right]$, where

$$\alpha_{k-1}(\mathcal{H}_{k-1}) = \mathrm{var}[\nabla\log p_{k-1}^{\pi^\theta}(\mathcal{H}_{a_{k-1}})\{r_{k-1}-q_{k-1}(\mathcal{H}_{a_{k-1}})+v_k(\mathcal{H}_{s_k})\}+\{d_k^v(\mathcal{H}_{s_k})-d_{k-1}^q(\mathcal{H}_{a_{k-1}})\}|\mathcal{H}_{a_{k-1}}].$$

From the assumption, this efficiency bound is lower bounded:

$$\textstyle\sum_{k=0}^{H+1}\mathbb{E}_{p^{\pi^b}}\left[\nu_{k-1}^2\alpha_{k-1}(\mathcal{H}_{a_{k-1}})\right] \succeq \mathbb{E}_{p^{\pi^b}}\left[\nu_H^2\alpha_H(\mathcal{H}_{a_H})\right] \succeq C_3^{2H}\mathbb{E}_{p^{\pi^b}}\left[\alpha_H(\mathcal{H}_{a_H})\right] \succeq C_3^{2H}c.$$

Here, we also have used $\alpha_k(\mathcal{H}_{a_k})$ for each $-1\leq k\leq H-1$ is a semi-positive definite matrix, and

$$\alpha_H(\mathcal{H}_{a_H}) = \mathrm{var}[\nabla\log p_H^{\pi^\theta}(\mathcal{H}_{a_H})\{r_H-q_H\}\mid\mathcal{H}_{a_H}].$$

□

Proof of Theorem 5. We have

$$\begin{aligned}
\mathrm{var}[\textstyle\sum_{t=0}^{H+1}\nu_{0:t}r_t\sum_{s=0}^{t}g_s] &= \sum_{k=0}^{H+1}\mathbb{E}[\mathrm{var}[\mathbb{E}[\sum_{t=0}^{H}\nu_{0:t}r_t\sum_{s=0}^{t}g_s \mid \mathcal{T}_{a_k}] \mid \mathcal{T}_{a_{k-1}}]]\\
&= \sum_{k=0}^{H+1}\mathbb{E}[\mathrm{var}[\mathbb{E}[\sum_{t=k-1}^{H}\nu_{0:t}r_t\sum_{s=k-1}^{t}g_s \mid \mathcal{T}_{a_k}] \mid \mathcal{T}_{a_{k-1}}]]\\
&= \sum_{k=0}^{H+1}\mathbb{E}[\nu_{k-1}^2\mathrm{var}[\mathbb{E}[\sum_{t=k-1}^{H}\nu_{k:t}r_t\sum_{s=k-1}^{t}g_s \mid \mathcal{T}_{a_k}] \mid \mathcal{T}_{a_{k-1}}]].\\
&= \sum_{k=0}^{H+1}\mathbb{E}[\nu_{k-1}^2\mathrm{var}[\mathbb{E}[\sum_{t=k-1}^{H}\nu_{k:t}r_t\sum_{s=k-1}^{t}g_s \mid \mathcal{H}_{a_k}] \mid \mathcal{H}_{a_{k-1}}]].
\end{aligned}$$

□

Proof of Theorem 6. Based on Theorem 5, as in the proof of Theorem 4, when $c\mathcal{I} \preceq \mathrm{var}[r_H g_H \mid \mathcal{H}_{a_H}]$, this variance is lower bounded by $C_3^{2H}c$. □

Proof of Theorem 7. For the simplicity of the notation, we prove the case where $K = 2$. Recall that the influence function of ξ_{MDP} is

$$\xi_{\mathrm{MDP}}(\mathcal{T};q,\mu,d^q,d^\mu) = \sum_{j=0}^{H}\{d_j^\mu(s_j,a_j)r_j - \mu_j(s_j,a_j)d_j^q(s_j,a_j) - d_j^\mu(s_j,a_j)q_j(s_j,a_j) \tag{18}$$

$$+ \mu_{j-1}(s_{j-1},a_{j-1})d_j^v(s_j) + d_{j-1}^\mu(s_{j-1},a_{j-1})v_j(s_j)\}. \tag{19}$$

Here, $\mu = \{\mu_j\}, q = \{q_j\}$, $d^q = \{d_j^q\}$, $d^\mu = \{d_j^\mu\}$. Then, the estimator $\hat{Z}^{\mathrm{EOPPG}}(\theta)$ is

$$0.5\mathbb{E}_{\mathcal{U}_1}[\xi_{\mathrm{MDP}}(\mathcal{T};\hat{q}^{(1)},\hat{\mu}^{(1)},\hat{d}^{q(1)},\hat{d}^{\mu(1)})] + 0.5\mathbb{E}_{\mathcal{U}_2}[\xi_{\mathrm{MDP}}(\mathcal{T};\hat{q}^{(2)},\hat{\mu}^{(2)},\hat{d}^{q(2)},\hat{d}^{\mu(2)})].$$

Then, we have

$$\sqrt{n}\{\mathbb{E}_{\mathcal{U}_2}[\xi_{\mathrm{MDP}}(\mathcal{T};\hat{q}^{(2)},\hat{\mu}^{(2)},\hat{d}^{q(2)},\hat{d}^{\mu(2)})] - Z(\theta)\}$$

$$= \sqrt{n}\{\mathbb{G}_{\mathcal{U}_2}[\xi_{\mathrm{MDP}}(\mathcal{T};\hat{q}^{(2)},\hat{\mu}^{(2)},\hat{d}^{q(2)},\hat{d}^{\mu(2)}) - \xi_{\mathrm{MDP}}(\mathcal{T};q,\mu,d^q,d^\mu)] \tag{20}$$

$$+ \sqrt{n}\{\mathbb{E}[\xi_{\mathrm{MDP}}(\mathcal{T};\hat{q}^{(2)},\hat{\mu}^{(2)},\hat{d}^{q(2)},\hat{d}^{\mu(2)}) \mid \hat{q}^{(2)},\hat{\mu}^{(2)},\hat{d}^{q(2)},\hat{d}^{\mu(2)}] - \mathbb{E}[\xi_{\mathrm{MDP}}(\mathcal{T};q,\mu,d^q,d^\mu)]\} \tag{21}$$

$$+ \sqrt{n}\{\mathbb{E}_{\mathcal{U}_2}[\xi_{\mathrm{MDP}}(\mathcal{T};q,\mu,d^q,d^\mu)] - Z(\theta)\}.$$

The first term (20) is $\mathrm{o}_p(1)$ following the proof of Theorem 5 (Kallus & Uehara, 2019b) (Also from doubly robust struture of EIF from the following lemma). The second term (21) is following Lemma 14.

Lemma 14. *The term* (21) *is* $\mathrm{o}_p(1)$.

Proof.

$$
\begin{aligned}
&\mathbb{E}[\xi_{\mathrm{MDP}}(\mathcal{T};\hat{q}^{(2)},\hat{\mu}^{(2)},\hat{d}^{q(2)},\hat{d}^{\mu(2)})\mid \hat{q}^{(2)},\hat{\mu}^{(2)},\hat{d}^{q(2)},\hat{d}^{\mu(2)}]-\mathbb{E}[\xi_{\mathrm{MDP}}(\mathcal{T};q,\mu,d^{q},d^{\mu})]\\
&=\mathbb{E}[\sum_{k=0}^{H}(\hat{\mu}^{(2)}_{k}-\mu_{k})(-\hat{d}^{q(2)}_{k}+d^{q}_{k})+(\hat{d}^{\mu(2)}_{k}-d^{\mu}_{k})(-\hat{q}^{(2)}_{k}+q_{k})\mid \mathcal{L}_2]\\
&+\mathbb{E}[\sum_{k=0}^{H}(\hat{\mu}^{(2)}_{k-1}-\mu_{k-1})(\hat{d}^{v(2)}_{k}-d^{v}_{k})+(\hat{d}^{\mu(2)}_{k-1}-d^{\mu}_{k-1})(\hat{v}^{(2)}_{k}-v_{k})\mid \mathcal{L}_2]\\
&+\mathbb{E}[\sum_{k=0}^{H}d^{\mu}_{k}(q_{k}-\hat{q}^{(2)}_{k})+\mu_{k}(d^{q}_{k}-\hat{d}^{q(2)}_{k})\mid \mathcal{L}_2]\\
&+\mathbb{E}[\sum_{k=0}^{H}d^{\mu}_{k-1}(\hat{v}^{(2)}_{k}-v_{k})+\mu_{k-1}(\mathbb{E}_{\pi^{e}}[\hat{d}^{q(2)}_{k}+\hat{q}^{(2)}_{k}g_{k}\mid s_{k}]-\mathbb{E}_{\pi^{e}}[d^{q}_{k}+q_{k}g_{k}\mid s_{k}])\mid \mathcal{L}_2]\\
&+\mathbb{E}[\sum_{k=0}^{H}(\hat{\mu}^{(2)}_{k}-\mu_{k})(-d^{q}_{k}+d^{v}_{k+1})+(\hat{d}^{\mu(2)}_{k}-d^{\mu}_{k})(r_{k}-q_{k}+v_{k+1})\mid \mathcal{L}_2]\\
&=\mathbb{E}[\sum_{k=0}^{H}(\hat{\mu}^{(2)}_{k}-\mu_{k})(-\hat{d}^{q(2)}_{k}+d^{q}_{k})+(\hat{d}^{\mu(2)}_{k}-d^{\mu}_{k})(-\hat{q}^{(2)}_{k}+q_{k})\mid \mathcal{L}_2]\\
&+\mathbb{E}[\sum_{k=0}^{H}(\hat{\mu}^{(2)}_{k-1}-\mu_{k-1})(\hat{d}^{v(2)}_{k}-d^{v}_{k})+(\hat{d}^{\mu(2)}_{k-1}-d^{\mu}_{k-1})(\hat{v}^{(2)}_{k}-v_{k})\mid \mathcal{L}_2].
\end{aligned}
$$

Here, we use

$$
\begin{aligned}
0&=\mathbb{E}[\mu_{k}f(s_k,a_k)-\mu_{k-1}\mathbb{E}_{\pi^{\theta}}[f(s_k,a_k)\mid s_k]],\\
0&=\mathbb{E}[d^{\mu}_{k}f(s_k,a_k)-d^{\mu}_{k-1}\mathbb{E}_{\pi^{\theta}}[f(s_k,a_k)\mid s_k]-\mu^{k-1}\mathbb{E}_{\pi^{\theta}}[f(s_k,a_k)g_k\mid s_k]],\\
0&=\mathbb{E}[f(s_k,a_k)(r_k-q_k+v_{k+1})],\\
0&=\mathbb{E}[f(s_k,a_k)(-d^{q}_{k}+d^{v}_{k+1})].
\end{aligned}
$$

Then, from Hölder's inequality, the Euclidean norm of the above is upper bounded by up to some absolute constant:

$$
\begin{aligned}
&\sum_{k=0}^{H}\|\hat{\mu}^{(2)}_{k}-\mu_{k}\|_{L^2_b}\|\|\hat{d}^{q(2)}_{k}-d^{q}_{k}\|_2\|_{L^2_b}+\|\|\hat{d}^{\mu(2)}_{k}-d^{\mu}_{k}\|_2\|_{L^2_b}\|\hat{q}^{(2)}_{k}-q_{k}\|_{L^2_b}\\
&+\sum_{k=0}^{H}\|\hat{\mu}^{(2)}_{k-1}-\mu_{k-1}\|_{L^2_b}\|\|\hat{d}^{v(2)}_{k}-d^{v}_{k}\|_2\|_{L^2_b}+\|\|\hat{d}^{\mu(2)}_{k-1}-d^{\mu}_{k-1}\|_2\|_{L^2_b}\|\hat{v}^{(2)}_{k}-v_{k}\|_{L^2_b}\\
&=\mathrm{o}_p(n^{-\alpha_1})\mathrm{o}_p(n^{-\alpha_4})+\mathrm{o}_p(n^{-\alpha_2})\mathrm{o}_p(n^{-\alpha_3})+\mathrm{o}_p(n^{-\alpha_1})\mathrm{o}_p(n^{-\min(\alpha_3,\alpha_4)})+\mathrm{o}_p(n^{-\alpha_2})\mathrm{o}_p(n^{-\alpha_3})\\
&=\mathrm{o}_p(n^{-\min\{\alpha_1,\alpha_2\}})\mathrm{o}_p(n^{-\min\{\alpha_3,\alpha_4\}})=\mathrm{o}_p(n^{-1/2}).
\end{aligned}
$$

Here, the convergence rates of $\hat{v}$, $\hat{d}^{v}$ are proved as follows:

$$
\begin{aligned}
\|\hat{v}_k-v_k\|^2_{L^2_b}&=\mathbb{E}_{\pi^{b}}[\{\mathbb{E}_{\pi^{\theta}}[\hat{q}_k(s_k,a_k)\mid s_k]-\mathbb{E}_{\pi^{\theta}}[q_k(s_k,a_k)\mid s_k]\}^2\mid \hat{q}_k]\\
&\leq \mathbb{E}_{\pi^{b}}[\mathbb{E}_{\pi^{\theta}}[\{\hat{q}_k(s_k,a_k)-q_k(s_k,a_k)\}^2\mid s_k]\mid \hat{q}_k]\\
&\leq C_1\mathbb{E}_{\pi^{b}}[\mathbb{E}_{\pi^{b}}[\{\hat{q}_k(s_k,a_k)-q_k(s_k,a_k)\}^2\mid s_k]\mid \hat{q}_k]\\
&\leq C_1\mathbb{E}_{\pi^{b}}[\{\hat{q}_k(s_k,a_k)-q_k(s_k,a_k)\}^2\mid \hat{q}_k]=\mathrm{o}_p(n^{-\alpha_3})
\end{aligned}
$$

The first line to the second line is proved by conditional Jensen's inequality. In the same way, by defining $q_{k,i}$ as a i-th component of q_k,

$$
\begin{aligned}
\|\hat{d}^v_{k,i} - d^v_{k,i}\|^2_{L^2_b} &< \mathbb{E}_{\pi^b}[\{\mathbb{E}_{\pi^\theta}[(\hat{d}^q_{k,i} - d^q_{k,i}) + (\hat{q}_k - q_k)g_{k,i} \mid s_k]\}^2 \mid \hat{q}_k, \hat{d}^q_k] \\
&\leq \mathbb{E}_{\pi^b}[\mathbb{E}_{\pi^\theta}[\{(\hat{d}^q_{k,i} - d^q_{k,i}) + (\hat{q}_k - q_k)g_{k,i}\}^2 \mid s_k] \mid \hat{q}_k, \hat{d}^q_k] \\
&\leq C_1\mathbb{E}_{\pi^b}[\mathbb{E}_{\pi^b}[\{(\hat{d}^q_{k,i} - d^q_{k,i}) + (\hat{q}_k - q_k)g_{k,i}\}^2 \mid s_k] \mid \hat{q}_k, \hat{d}^q_k] \\
&\leq C_1\mathbb{E}_{\pi^b}[\{(\hat{d}^q_{k,i} - d^q_{k,i}) + (\hat{q}_k - q_k)g_{k,i}\}^2 \mid \hat{q}_k, \hat{d}^q_k] = \mathrm{o}_p(n^{-\min(\alpha_3,\alpha_4)}).
\end{aligned}
$$

□

Finally, combining everything, we have

$$
\begin{aligned}
&0.5\mathbb{E}_{\mathcal{U}_1}[\xi_{\mathrm{MDP}}(\mathcal{T}; \hat{q}^{(1)}, \hat{\mu}^{(1)}, \hat{d}^{q(1)}, \hat{d}^{\mu(1)})] + 0.5\mathbb{E}_{\mathcal{U}_2}[\xi_{\mathrm{MDP}}(\mathcal{T}; \hat{q}^{(2)}, \hat{\mu}^{(2)}, \hat{d}^{q(2)}, \hat{d}^{\mu(2)})] \\
&= 0.5\mathbb{E}_{\mathcal{U}_1}[\xi_{\mathrm{MDP}}(\mathcal{T}; q, \mu, d^q, d^\mu)] + 0.5\mathbb{E}_{\mathcal{U}_2}[\xi_{\mathrm{MDP}}(\mathcal{T}; q, \mu, d^q, d^\mu)] + \mathrm{o}_p(n^{-1/2}) \\
&= \mathbb{E}_n[\xi_{\mathrm{MDP}}(\mathcal{T}; q, \mu, d^q, d^\mu)] + \mathrm{o}_p(n^{-1/2}).
\end{aligned}
$$

Finally, CLT concludes the proof. □

Proof of Theorem 8. For the simplicity of the notation, we prove the case where $K = 2$. Recall that the influence function of ξ_{MDP} is

$$
\xi_{\mathrm{MDP}}(\mathcal{T}; q, \mu, d^q, d^\mu) = \sum_{j=0}^{H}\{d^\mu_j(s_j, a_j)r_j - \mu_j(s_j, a_j)d^q_j(s_j, a_j) - d^\mu_j(s_j, a_j)q_j(s_j, a_j) \tag{22}
$$

$$
+ \mu_{j-1}(s_{j-1}, a_{j-1})\mathbb{E}[d^q_j(s_j, a_j)|s_j] + d^\mu_{j-1}(s_{j-1}, a_{j-1})\mathbb{E}[q_j(s_j, a_j)|s_j]\}. \tag{23}
$$

Here, $\mu = \{\mu_j\}, q = \{q_j\}$, $d^q = \{d^q_j\}$, $d^\mu = \{d^\mu_j\}$. Then, the estimator $\hat{Z}^{\mathrm{EOPPG}}(\theta)$ is

$$
0.5\mathbb{E}_{\mathcal{U}_1}[\xi_{\mathrm{MDP}}(\mathcal{T}; \hat{q}^{(1)}, \hat{\mu}^{(1)}, \hat{d}^{q(1)}, \hat{d}^{\mu(1)})] + 0.5\mathbb{E}_{\mathcal{U}_2}[\xi_{\mathrm{MDP}}(\mathcal{T}; \hat{q}^{(2)}, \hat{\mu}^{(2)}, \hat{d}^{q(2)}, \hat{d}^{\mu(2)})].
$$

Then, we have

$$
\{\mathbb{E}_{\mathcal{U}_2}[\xi_{\mathrm{MDP}}(\mathcal{T}; \hat{q}^{(2)}, \hat{\mu}^{(2)}, \hat{d}^{q(2)}, \hat{d}^{\mu(2)})] - Z(\theta)\}
$$

$$
\{\mathbb{G}_{\mathcal{U}_2}[\xi_{\mathrm{MDP}}(\mathcal{T}; \hat{q}^{(2)}, \hat{\mu}^{(2)}, \hat{d}^{q(2)}, \hat{d}^{\mu(2)}) - \xi_{\mathrm{MDP}}(\mathcal{T}; q^\dagger, \mu^\dagger, d^{\mu\dagger}, d^{q\dagger})] \tag{24}
$$

$$
+ \{\mathbb{E}[\xi_{\mathrm{MDP}}(\mathcal{T}; \hat{q}^{(2)}, \hat{\mu}^{(2)}, \hat{d}^{q(2)}, \hat{d}^{\mu(2)}) \mid \hat{q}^{(2)}, \hat{\mu}^{(2)}, \hat{d}^{q(2)}, \hat{d}^{\mu(2)}] - \mathbb{E}[\xi_{\mathrm{MDP}}(\mathcal{T}; q^\dagger, \mu^\dagger, d^{\mu\dagger}, d^{q\dagger})]\} \tag{25}
$$

$$
+ \{\mathbb{E}_{\mathcal{U}_2}[\xi_{\mathrm{MDP}}(\mathcal{T}; q^\dagger, \mu^\dagger, d^{\mu\dagger}, d^{q\dagger})] - Z(\theta)\}.
$$

The first term (24) is $\mathrm{o}_p(1/\sqrt{n})$ following the proof of Theorem 5 (Kallus & Uehara, 2019b). The second term (25) is 0 following Lemma 14.

Finally,

$$
\begin{aligned}
&0.5\mathbb{E}_{\mathcal{U}_1}[\xi_{\mathrm{MDP}}(\mathcal{T}; \hat{q}^{(1)}, \hat{\mu}^{(1)}, \hat{d}^{q(1)}, \hat{d}^{\mu(1)})] + 0.5\mathbb{E}_{\mathcal{U}_2}[\xi_{\mathrm{MDP}}(\mathcal{T}; \hat{q}^{(2)}, \hat{\mu}^{(2)}, \hat{d}^{q(2)}, \hat{d}^{\mu(2)})] \\
&= 0.5\mathbb{E}_{\mathcal{U}_1}[\xi_{\mathrm{MDP}}(\mathcal{T}; q^\dagger, \mu^\dagger, d^{\mu\dagger}, d^{q\dagger})] + 0.5\mathbb{E}_{\mathcal{U}_2}[\xi_{\mathrm{MDP}}(\mathcal{T}; q^\dagger, \mu^\dagger, d^{\mu\dagger}, d^{q\dagger})] + \mathrm{o}_p(1) \\
&= \mathbb{E}_n[\xi_{\mathrm{MDP}}(\mathcal{T}; q^\dagger, \mu^\dagger, d^{\mu\dagger}, d^{q\dagger})] + \mathrm{o}_p(1).
\end{aligned}
$$

Finally, the law of large number concludes the proof since the mean is $Z(\theta)$ under the condition in the theorem. We use $\mathbb{E}[\xi_{\mathrm{MDP}}(\mathcal{T}; q^\dagger, \mu^\dagger, d^{\mu\dagger}, d^{q\dagger})] = Z(\theta)$.

Lemma 15. $\mathbb{E}[\xi_{\mathrm{MDP}}(\mathcal{T}; q^\dagger, \mu^\dagger, d^{\mu\dagger}, d^{q\dagger})] = Z(\theta)$.

Proof.

$$
\begin{aligned}
&\mathbb{E}[\xi_{\mathrm{MDP}}(\mathcal{T}; q^\dagger, \mu^\dagger, d^{\mu\dagger}, d^{\mu\dagger}) \mid q^\dagger, \mu^\dagger, d^{\mu\dagger}, d^{\mu\dagger}] - \mathbb{E}[\xi_{\mathrm{MDP}}(\mathcal{T}; q, \mu, d^q, d^\mu)] \\
&= \mathbb{E}[\sum_{k=0}^{H}(\mu_k^\dagger - \mu_k)(-d_k^{q\dagger} + d_k^q) + (d_k^{\mu\dagger} - d_k^\mu)(-q_k^\dagger + q_k)] \\
&+ \mathbb{E}[\sum_{k=0}^{H}(\mu_{k-1}^\dagger - \mu_{k-1})(d_k^{v\dagger} - d_k^v) + (d_{k-1}^{\mu\dagger} - d_{k-1}^\mu)(v_k^\dagger - v_k)] \\
&+ \mathbb{E}[\sum_{k=0}^{H} d_k^\mu(q_k - q_k^\dagger) + \mu_k(d_k^q - d_k^{q\dagger})] \\
&+ \mathbb{E}[\sum_{k=0}^{H} d_{k-1}^\mu(v_k^\dagger - v_k) + \mu_{k-1}(\mathbb{E}_{\pi^\theta}[(q_k^\dagger - q_k)g_k \mid s_k]) + \mu_{k-1}(\mathbb{E}_{\pi^\theta}[d_k^{q\dagger} \mid s_k] - \mathbb{E}_{\pi^\theta}[d_k^q \mid s_k])] \\
&+ \mathbb{E}[\sum_{k=0}^{H}(\mu_k^\dagger - \mu_k)(-d_k^q + d_k^v) + (d_k^{\mu\dagger} - d_k^\mu)(r_k - q_k + v_{k+1})] \\
&= \mathbb{E}[\sum_{k=0}^{H}(\mu_k^\dagger - \mu_k)(-d_k^{q\dagger} + d_k^q) + (d_k^{\mu\dagger} - d_k^\mu)(q_k^\dagger - q_k)] \\
&+ \mathbb{E}[\sum_{k=0}^{H}(\mu_{k-1}^\dagger - \mu_{k-1})(-d_k^{v\dagger} + d_{k+1}^v) + (d_{k-1}^{\mu\dagger} - d_{k-1}^\mu)(v_k^\dagger - v_k)].
\end{aligned}
$$

Here, we use the relations for $\forall f(s,a)$:

$$
\begin{aligned}
0 &= \mathbb{E}[\mu_k f(s_k, a_k) - \mu_{k-1}\mathbb{E}_{\pi^\theta}[f(s_k, a_k) \mid s_k]], \\
0 &= \mathbb{E}[d_k^\mu f(s_k, a_k) - d_{k-1}^\mu \mathbb{E}_{\pi^\theta}[f(s_k, a_k) \mid s_k] - \mu^{k-1}\mathbb{E}_{\pi^\theta}[f(s_k, a_k)g_k \mid s_k]], \\
0 &= \mathbb{E}[f(s_k, a_k)(r_k - q_k + v_{k+1})], \\
0 &= \mathbb{E}[f(s_k, a_k)(-d_k^q + d_{k+1}^v)].
\end{aligned}
$$

Then, when $\mu = \mu^\dagger, d^\mu = d^{\mu\dagger}$ or $q = q^\dagger$, $d^q = d^{q\dagger}$ or $\mu = \mu^\dagger$, $q = q^\dagger$, we have

$$
\begin{aligned}
&\mathbb{E}[\sum_{k=0}^{H}(\mu_k^\dagger - \mu_k)(d_k^q - d_k^{q\dagger}) + (d_k^{\mu\dagger} - d_k^\mu)(q_k - q_k^\dagger)] + \mathbb{E}[\sum_{k=0}^{H}(\mu_{k-1}^\dagger - \mu_{k-1})(d_k^{v\dagger} - d_k^v) + (d_{k-1}^{\mu\dagger} - d_{k-1}^\mu)(v_k^\dagger - v_k)] \\
&= 0 + 0 + 0 + 0 = 0.
\end{aligned}
$$

This concludes the proof.

Remark 7. *The above is not equal to* 0 *when* $d^\mu = d^{\mu\dagger}$, $d^q = d^{q\dagger}$*. The reason is in that case:*

$$\mathbb{E}[\textstyle\sum_{k=0}^{H}(\mu_{k-1}^\dagger - \mu_{k-1})(d_k^{v\dagger} - d_k^v)] \neq 0.$$

since $d_k^{v\dagger} \neq d_k^v$.

□

□

Proof of Theorem 9. First, we have

$$q_j(s_j, s_j) = \mathbb{E}[\textstyle\sum_{t=j}^{H} r_t \nu_{j+1:t} \mid a_j, s_j].$$

By differentiating w.r.t θ, we have

$$d^q_j(s_j,a_j)=\mathbb{E}\left[\sum_{t=j}^{H} r_t\nu_{j+1:t}\left\{\sum_{i=j+1}^{t} g_i(a_i|s_i)\right\}\mid a_j,s_j\right]=\mathbb{E}\left[\sum_{t=j+1}^{H} r_t\nu_{j+1:t}\left\{\sum_{i=j+1}^{t} g_i(a_i|s_i)\right\}\mid a_j,s_j\right],$$

noting

$$\nabla\nu_{j+1:t}=\nu_{j+1:t}\nabla\log\nu_{j+1:t}=\nu_{j+1:t}\{\textstyle\sum_{i=j+1}^{t} g_i(a_i|s_i)\}.$$

Second, we have

$$\mu_j(s_j,a_j)=\mathbb{E}\left[\nu_{0:j}\mid a_j,s_j\right].$$

By differentiating w.r.t θ, we have

$$d^{\mu}_j(a_j,s_j)=\mathbb{E}\left[\nu_{0:j}\left\{\textstyle\sum_{i=0}^{j} g_i(a_i|s_i)\right\}\mid a_j,s_j\right],$$

noting

$$\nabla\nu_{0:j}=\nu_{0:j}\nabla\log\nu_{0:j}=\nu_{0:j}\{\textstyle\sum_{i=0}^{j} g_i(a_i|s_i)\}.$$

□

Proof of Theorem 10. The following recursive equations (Bellman equations) hold:

$$\begin{aligned}q_j(s_j,a_j)&=\mathbb{E}[r+q_{j+1}(s_{j+1},\pi^{\theta})\mid s_j,a_j],\\ \mu_j(s_j,a_j)&=\mathbb{E}[\mu_{j-1}(s_{j-1},a_{j-1})\tilde{\nu}_j|s_j,a_j].\end{aligned}$$

Then, by differentiating w.r.t θ, we have

$$\begin{aligned}d^q_j(s_j,a_j)&=\mathbb{E}[\mathbb{E}_{\pi^{\theta}}[d^q_{j+1}(s_{j+1},a_{j+1})+g_{j+1}(s_{j+1},a_{j+1})q_{j+1}(s_{j+1},a_{j+1})\mid s_{j+1}]|s_j,a_j],\\ d^{\mu}_j(s_j,a_j)&=\mathbb{E}[d^{\mu}_{j-1}(s_{j-1},a_{j-1})|s_j,a_j]\tilde{\nu}_j+\mathbb{E}[\mu_{j-1}(s_{j-1},a_{j-1})|s_j,a_j]g_j(a_j,s_j)\tilde{\nu}_j\\ &=\mathbb{E}[d^{\mu}_{j-1}(s_{j-1},a_{j-1})|s_j,a_j]\tilde{\nu}_j+\mu_j(s_j,a_j)g_j(a_j,s_j).\end{aligned}$$

□

Proof of Theorem 11 . We modify the proof of Theorem 1 (Khamaru & Wainwright, 2018) so that we can deal with the noise gradient. In this proof, define $f(\theta)=-J(\theta)$. Then, by M-smoothness,

$$f(\theta)\leq f(\theta_k)+\langle\nabla f(\theta_k),x-\theta_k\rangle+\tfrac{M}{2}\|x-\theta_k\|_2.$$

Then, by replacing θ with $\theta_{k+1}=\theta_k-\alpha_k\nabla f(\theta_k)-\alpha_k B_k$,

$$f(\theta_{k+1})\leq f(\theta_k)+\langle\nabla f(\theta_k),\theta_{k+1}-\theta_k\rangle+\tfrac{M}{2}\|\theta_{k+1}-\theta_k\|_2.$$

Thus,

$$\begin{aligned}f(\theta_k)-f(\theta_{k+1})&\geq-\langle\nabla f(\theta_k),\theta_{k+1}-\theta_k\rangle-\frac{M}{2}\|\theta_{k+1}-\theta_k\|_2^2\\ &=\alpha_k\langle\nabla f(\theta_k),\nabla f(\theta_k)+B_k\rangle-\frac{M\alpha_k^2}{2}\|\nabla f(\theta_k)+B_k\|_2^2\\ &=\alpha_k\|\nabla f(\theta_k)\|_2^2+\alpha_k\langle\nabla f(\theta_k),B_k\rangle-\frac{M\alpha_k^2}{2}\|\nabla f(\theta_k)+B_k\|_2^2\\ &=\alpha_k\|\nabla f(\theta_k)\|_2^2-\alpha_k|\langle\nabla f(\theta_k),B_k\rangle|-\frac{M\alpha_k^2}{2}\|\nabla f(\theta_k)+B_k\|_2^2\\ &\geq\alpha_k\|\nabla f(\theta_k)\|_2^2-0.5\alpha_k(\|\nabla f(\theta_k)\|^2+\|B_k\|_2^2)-M\alpha_k^2(\|\nabla f(\theta_k)\|_2^2+\|B_k\|_2^2)\\ &\geq 0.25\alpha_k\|\nabla f(\theta_k)\|_2^2-0.5\alpha_k\|B_k\|_2^2-0.25\alpha_k\|B_k\|_2^2.\end{aligned}$$

Here, from the fourth line to the fifth line, we use inequalities parallelogram law:

$$2|\langle a,b\rangle| < \|a\|_2^2+\|b\|_2^2,\ \|a+b\|_2^2 \leq 2\|a\|_2^2+2\|b\|_2^2.$$

From the fifth line to the sixth line, we use a condition regarding M. This yields,

$$f(\theta_k)-f(\theta_{k+1})+0.75\alpha_k\|B_k\|_2^2 \geq 0.25\alpha_k\|\nabla f(\theta_k)\|_2^2.$$

Then, by telescoping sum,

$$\tfrac{1}{T}\{f(\theta_1)-f^*\}+\tfrac{1}{T}\textstyle\sum_t 0.75\alpha_t\|B_t\|_2^2 \geq \tfrac{1}{T}\sum_t 0.25\alpha_t\|\nabla f(x_t)\|_2^2.$$

Noting $f(\theta)=-J(\theta)$,

$$\tfrac{1}{T}\{J^*-J(\theta_1)\}+\tfrac{1}{T}\textstyle\sum_t 0.75\alpha_t\|B_t\|_2^2 \geq \tfrac{1}{T}\sum_t 0.25\alpha_t\|\nabla J(x_t)\|_2^2.$$

Expanding by 4 yields the result. □

Proof of Theorem 12. Here,

$$B_t=\mathbb{E}_n[\xi_{\mathrm{MDP}}(\theta_t)-Z(\theta_t)]+\mathrm{o}_p(n^{-1/2}).$$

from the proof of Theorem 7. Then, the j^{th} component of B_t^2 is

$$B_{t,j}^2=(\mathbb{E}_n[\xi_{\mathrm{MDP},j}(\theta_t)-Z_j(\theta_t)]+\mathrm{o}_p(n^{-1/2}))^2=\mathbb{E}_n[\xi_{\mathrm{MDP},j}(\theta_t)-Z_j(\theta_t)]^2+\mathrm{o}_p(n^{-1}), \qquad (26)$$

where $\xi_{\mathrm{MDP},j}$ is a j^{th} component of IF and Z_j is a j-th term of $Z(\theta)$, $\xi_{\mathrm{MDP}}(\theta)$ is ξ_{MDP} at θ. Here, we use $O_p(n^{-1/2})\mathrm{o}_p(n^{-1/2})=\mathrm{o}_p(n^{-1})$, $\mathrm{o}_p(n^{-1/2})\mathrm{o}_p(n^{-1/2})=\mathrm{o}_p(n^{-1})$. Here, noting θ_t is a random variable, we have to bound the main term uniformly as

$$\mathbb{E}_n[\xi_{\mathrm{MDP},j}(\theta_t)-Z_j(\theta_t)]^2 \leq (\textstyle\sup_{\theta\in\Theta}\mathbb{E}_n[\xi_{\mathrm{MDP},j}(\theta)-Z_j(\theta)])^2.$$

By following Theorem 8.5 (Sen, 2018) based on a standard empirical process theory combining Rademacher complexity and Talagland inequality, with probability $1-\delta$,

$$\begin{aligned}
&\sup_{\theta\in\Theta}\mathbb{E}_n[\xi_{\mathrm{MDP},j}(\theta)-Z_j(\theta)]\\
&\lessapprox \mathbb{E}[\sup_{\theta\in\Theta}|\frac{1}{n}\sum_{i=1}^{n}\epsilon^{(i)}\xi^{(i)}_{\mathrm{MDP},j}(\theta)|]+\sqrt{\frac{\sup_{\theta\in\Theta}\mathrm{var}[\xi_{\mathrm{MDP},j}]\log(1/\delta)}{n}}+\frac{\log(1/\delta)}{n}\\
&\lessapprox L\sqrt{D/n}+\sqrt{\frac{C_2G_{\max}R_{\max}^2(H+1)^2(H+2)^2\log(1/\delta)}{n}}+\frac{\log(1/\delta)}{n}.
\end{aligned}$$

Then, with probability $1-\delta$,

$$\begin{aligned}
&\{\sup_{\theta\in\Theta}\mathbb{E}_n[\xi_{\mathrm{MDP},j}(\theta)-Z_j(\theta)]\}^2\\
&\lessapprox \left\{L\sqrt{D/n}+\sqrt{\frac{C_2G_{\max}R_{\max}^2(H+1)^2(H+2)^2\log(1/\delta)}{n}}+\frac{\log(1/\delta)}{n}\right\}^2.
\end{aligned}$$

Here, we also use the Rademacher complexity of the Lipschitz class on Euclidean ball is bounded by $L\sqrt{D/n}$ based on the assumption Θ is a compact space (Example 4.6 (Sen, 2018)).

Considering an error term $\mathrm{o}_p(1/n)$ in (26) and taking an union bound over $t\in[1,\cdots,T], j\in[1,\cdots,D]$, we conclude that there exists N_δ such that with probability at least $1-\delta, \forall n\geq N_\delta$

$$\begin{aligned}
\tfrac{1}{T}\textstyle\sum_t\|B_t\|_2^2 &= \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{D}B_{t,j}^2\\
&\lessapprox D\left\{L\sqrt{\frac{D}{n}}+\sqrt{\frac{C_2G_{\max}R_{\max}^2(H+1)^2(H+2)^2\log(TD/\delta)}{n}}+\frac{\log(TD/\delta)}{n}\right\}^2\\
&\lessapprox D\frac{L^2D+C_2G_{\max}R_{\max}^2(H+1)^2(H+2)^2\log(TD/\delta)}{n}.
\end{aligned}$$

□

Proof of Theorem 13. We modify the proof of Theorem 3.1 (Hazan, 2015) so that we can deal with a noise gradient. Define $-J(\theta)$ as $f(\theta)$ and redefine $Z(\theta), \hat{Z}(\theta)$ as $-Z(\theta), -\hat{Z}^{\mathrm{EOPPG}}(\theta)$. Then, the algorithm is redefined as

- $\tilde{\theta}_t = \theta_t - \alpha_t \hat{Z}(\theta_t)$
- $\theta_t = \mathrm{Proj}_{\Theta}\, \tilde{\theta}_t$

Then, from convexity assumption for $-J(\theta)$,

$$f(\theta_t) - f(\theta^*) \leq Z(\theta_t)(\theta_t - \theta^*).$$

In addition, from Theorem 2.1 (Hazan, 2015),

$$\|\theta_{t+1} - \theta^*\|_2 = \| \mathrm{Proj}_{\Theta}(\theta_t - \alpha_t \hat{Z}(\theta_t)) - \theta^*\|_2 \leq \|\theta_t - \alpha_t \hat{Z}(\theta_t) - \theta^*\|_2.$$

Hence,

$$2\langle \hat{Z}(\theta_t), (\theta_t - \theta^*) \rangle \leq \tfrac{\|\theta_t - \theta^*\|_2^2 - \|\theta_{t+1} - \theta^*\|_2^2}{\alpha_t} + \alpha_t \|\hat{Z}(\theta_t)\|_2^2. \tag{27}$$

Noting

$$\|\hat{Z}(\theta_t)\|_2^2 = \|B_t + Z(\theta_t)\|_2^2 \leq 2\|B_t\|_2^2 + 2\|Z(\theta_t)\|_2^2,$$

as in the proof of Theorem 12, there exists N_δ such that $n \geq N_\delta$ with probability at least $1 - \delta$,

$$2\langle Z(\theta_t), (\theta_t - \theta^*) \rangle \lessapprox \tfrac{\|\theta_t - \theta^*\|_2^2 - \|\theta_{t+1} - \theta^*\|_2^2}{\alpha_t} + (\sup_{\theta \in \Theta} \|Z(\theta)\|^2 + \tilde{U})\alpha_t,$$

where $\tilde{U} = D\frac{L^2 D + C_2 G_{\max} R_{\max}^2 (H+1)^2 (H+2)^2 \log(D/\delta)}{n}$. Then, based on

$$\begin{aligned} \textstyle\sum_{t=1}^T f(\theta_t) - f(\theta^*) &\leq \sum_{t=1}^T \langle Z(\theta), (\theta_t - \theta^*) \rangle \\ &\leq \sum_{t=1}^T \langle \hat{Z}^{\mathrm{EOPPG}}(\theta), (\theta_t - \theta^*) \rangle + \sum_{t=1}^T \langle Z(\theta) - \hat{Z}^{\mathrm{EOPPG}}(\theta), (\theta_t - \theta^*) \rangle, \end{aligned} \tag{28}$$

we analyze the first term and second term of (28).

First term of (28)

From (27), there exists N_δ such that $n > N_\delta$ with at least $1 - \delta$,

$$\textstyle\sum_{t=1}^T \langle \hat{Z}^{\mathrm{EOPPG}}(\theta), (\theta_t - \theta^*) \rangle \lessapprox \sum_{t=1}^T \frac{\|\theta_t - \theta^*\|_2^2 - \|\theta_{t+1} - \theta^*\|_2^2}{\alpha_t} + (U + \sup_{\theta \in \Theta} \|Z(\theta)\|_2)\alpha_t.$$

where $U = D\frac{L^2 D + C_2 G_{\max} R_{\max}^2 (H+1)^2 (H+2)^2 \log(TD/\delta)}{n}$. (We also take an union bound over t). Then, under this event,

$$\begin{aligned} &\sum_t \frac{\|\theta_t - \theta^*\|_2^2 - \|\theta_{t+1} - \theta^*\|_2^2}{\alpha_t} + \sum_t \{\sup \|Z(\theta)\|_2^2 + U\}\alpha_t \\ &\leq \sum_t \|\theta_t - \theta^*\|_2^2 (\frac{1}{\alpha_t} - \frac{1}{\alpha_{t+1}}) + \sum_t \{\sup \|Z(\theta)\|_2^2 + U\}\alpha_t \\ &\leq \sum_{t=1}^T \Upsilon^2 (\frac{1}{\alpha_t} - \frac{1}{\alpha_{t+1}}) + \sum_t \{\| \sup Z(\theta)\|_2^2 + U\}\alpha_t \\ &\leq \Upsilon^2 \frac{1}{\alpha_T} + \{\sup Z(\theta)^2 + U\} \sum_t \alpha_t \lessapprox \Upsilon \sqrt{\{\sup \|Z(\theta)\|_2^2 + U\} T}, \end{aligned}$$

Here, we take $\alpha_t = \Upsilon/\sqrt{t\{\sup_{\theta\in\Theta}\|Z(\theta)\|^2+U\}}$. The last inequality follows since $\sum 1/\sqrt{t} \leq \sqrt{T}$.

Second term of (28)

We have

$$\textstyle\sum_{t=1}^{T}\{Z(\theta)-\hat{Z}^{\mathrm{EOPPG}}(\theta)\}^{\top}(\theta_t-\theta^*) \leq \sum_{t=1}^{T}\{Z(\theta)-\hat{Z}^{\mathrm{EOPPG}}(\theta)\}^{\top}(\theta_t-\theta^*) \leq \sum_{t=1}^{T}\|B_t\|_2\times\Upsilon.$$

Then, as in the proof of Theorem 12, with probability $1-\delta$, this is bounded by

$$\textstyle\sum_{t=1}^{T}\|B_t\|_2\times\Upsilon \lessapprox T\times\sqrt{U}\times\Upsilon.$$

Combining the first term and second term of (28)

We combine the first term and second term of (28). Then, we have

$$\begin{aligned} f\left(\frac{1}{T}\sum_{t=1}^{T}\theta_t\right)-f(\theta^*) &< \frac{1}{T}\left(\sum_{t=1}^{T}f(\theta_t)-f(\theta^*)\right) \\ &\lessapprox \sqrt{U}\Upsilon+\Upsilon\sqrt{\frac{\sup\|Z(\theta)\|_2^2+U}{T}} \\ &\lessapprox \Upsilon\left\{\sqrt{\frac{\sup\|Z(\theta)\|_2^2}{T}}+\sqrt{U}\left(1+\frac{1}{\sqrt{T}}\right)\right\}. \end{aligned}$$

Here, we use an inequality $\sqrt{x+y}\leq\sqrt{x}+\sqrt{y}$ for $x>0, y>0$. □