From Weighting to Modeling: A Nonparametric Estimator for Off-Policy Evaluation

Anonymous authors
Paper under double-blind review

Abstract

We study off-policy evaluation in the setting of contextual bandits, where we aim to evaluate a new policy using historical data that consists of contexts, actions and received rewards. This historical data typically does not faithfully represent action distribution of the new policy accurately. A common approach, inverse probability weighting (IPW), adjusts for these discrepancies in action distributions. However, this method often suffers from high variance due to the probability being in the denominator. The doubly robust (DR) estimator reduces variance through modeling reward but does not directly address variance from IPW. In this work, we address the limitation of IPW by proposing a Nonparametric Weighting (NW) approach that constructs weights using a nonparametric model. Our NW approach achieves low bias like IPW but typically exhibits significantly lower variance. To further reduce variance, we incorporate reward predictions—similar to the DR technique— resulting in the Model-assisted Nonparametric Weighting (MNW) approach. We show that MNW yields accurate value estimates when either the reward model or the behavior policy model is well specified. Extensive empirical comparisons show that our approaches consistently outperform existing techniques, achieving lower variance in value estimation while maintaining low bias.

1 Introduction

We study the off-policy value evaluation problem for decision making in environments where feedback is available only for chosen actions. In this context, we aim to estimate the value of a new policy using data collected under the actual action-generating policy (Langford & Zhang, 2008; Strehl et al., 2010). This problem is critical in many real-world applications of reinforcement learning, especially when it is impractical to estimate a policy by direct implementation due to high costs, risks, or ethical and legal concerns (Li et al., 2011). For example, in health care, we collect a set of data according a policy that receive information only whether a patient received a treatment, but receive no information about if the patient were not receive the treatment, and we aim to evaluate a new treatment policy using the collected dataset.

The first approach address off-policy learning in contextual bandits is inverse probability weighting (IPW) (Horvitz & Thompson, 1952), uses importance weights to correct for action imbalances in the logged data, but often incurs high variance, especially when the data logging policy has a low probability of choosing some actions. The second approach, the Direct Method (DM), which estimates the reward function from data and substitutes it as a substitute for the true reward to evaluate policy value across contexts. DM heavily relies on correct model specification and can suffer high bias when the reward model is misspecified—an issue common in practice. The third approach, doubly robust (DR) estimator (Cassel et al., 1976; Robins & Rotnitzky, 1995; Lunceford & Davidian, 2004; Kang & Schafer, 2007), combines DM and IPS and retains unbiasedness if either component is correctly specified. Although DR reduces variance through reward modeling, it does not directly address the variance introduced by the weighting mechanism itself.

In this paper, we address the high variance of the IPW approach by introducing the nonparametric weighting (NW) method, which uses a nonparametric model to link target policy-weighted rewards to behavior policy probabilities, thereby reducing the instability inherent in IPS. By employing P-splines to estimate this flexible function, the NW approach constructs weights that substantially reduce variance while maintaining low

bias, similar to IPW. We establish convergence rates for both the bias and mean squared error of the NW estimator.

To further reduce variance, we extend our framework by incorporating reward predictions, akin to the DR technique, resulting in the Model-assisted Nonparametric Weighting (MNW) approach. We show that the MNW estimator corrects bias from the reward model and provide convergence rates for its bias and mean squared error. When applied to policy evaluation, both the NW and MNW estimators consistently outperform existing methods, achieving lower variance without sacrificing bias.

2 Off-policy Evaluation

Let $\mathcal{X}$ be an input space, $\mathcal{A}$ a finite action space with $K = |\mathcal{A}|$, and $\mathcal{R}$ a reward space. Consider a policy evaluation problem in contextual bandits that is specified by a distribution $\mathcal{D}$ over pairs (x, r), where $x \in \mathcal{X}$ is the context and $r \in \mathcal{R}^K$ is a vector of rewards for all actions. A dataset of size n is generated as follows: the environment draws a sample (x_i, r_i), and the policy chooses an action $a \in \mathcal{A}$ from an unknown distribution p. The reward r_{ia} corresponding to the context-action pair (x_i, a) is then revealed. We assume

$$r_{ia} = \mu_{ia} + \epsilon_{ia}, \tag{1}$$

where $\mu_{ia} = \mathbb{E}[r_a|x_i]$ is the expected reward given the context-action pair (x_i, a), and ϵ_{ia} is a zero-mean noise term with finite variance σ^2_{ia}.

We are interested in two tasks: policy evaluation and policy optimization. In policy evaluation, our goal is to evaluate the policy π. Given x_i, the policy π chooses an action: $\pi_{ia} = \pi(a|x_i)$, $a \in \mathcal{A}$. we aim to estimate the value of a stationary policy π,

$$V^\pi = \mathbb{E}_{(x,r)\sim\mathcal{D}}\left[r_{\pi(x)}\right]. \tag{2}$$

In policy optimization, the aim is to find an optimal policy with maximum value $\pi^* = \arg\max_\pi V^\pi$. In this paper, we focus on the problem of policy evaluation. It is expected that more accurate evaluation generally leads to better optimization Strehl et al. (2010). Further investigation is needed into variants of this approach.

The key challenge in estimating policy value, given the data as described in the previous section, is the fact that we only have partial information about the reward, hence we cannot directly simulate our proposed policy on the data set $\mathcal{S}$. Given a data set $\mathcal{S} = \{x_i, a_i, r_{ia_i}\}_{i=1}^n$ collected as above. Define the probability associated with x_i as

$$p_{ia} = \mathbb{P}(a|x_i).$$

In practice, this probability needs to be estimated, and we denote its estimator as $\hat{p}_{ia}$. Generally, there are three main approaches for overcoming this limitation.

- **DM**. The direct method forms an estimate $\hat{\mu}_{ia} = \hat{\mu}_a(x_i)$ of the expected reward on the context and action (x_i, a). The policy value is then estimated by

$$\hat{V}^\pi_{\mathrm{dm}} = n^{-1}\sum\nolimits_{i=1}^n \sum\nolimits_{a\in\mathcal{A}} \pi_{ia}\hat{\mu}_{ia}.$$

- **IPW**. The inverse probability weighting (IPW) uses the inverse probability weighting to correct the gap between data-collection policy and the target policy:

$$\hat{V}^\pi_{\mathrm{ipw}} = n^{-1}\sum\nolimits_{i=1}^n p^{-1}_{ia_i}\pi_{ia_i} r_{ia_i}.$$

- **DR**. The doubly robust (DR) approach take advantage of both the estimate of the expected reward $\hat{\mu}_{ia} = \hat{\mu}_a(x_i)$ and the estimate of action probabilities $\hat{p}(a|x, \mathcal{H})$ and construct a DR estimator

$$\hat{V}^\pi_{\mathrm{dr}} = n^{-1}\sum\nolimits_{i=1}^n \left[p^{-1}_{ia_i}\pi_{ia_i}(r_{ia_i} - \hat{\mu}_{ia_i}) + \sum\nolimits_{a\in\mathcal{A}} \pi_{ia}\hat{\mu}_{ia}\right].$$

In practice, the estimated $\hat{p}_{ia}$ is used in place of the true p_{ia} in the IPW and DR approaches. The DM approach requires an accurate reward model of rewards; however, if the model is poorly specified, the DM method can incur high bias. Generally, accurately modeling rewards can be challenging, making this specification potentially restrictive. The IPW approach often suffers from large variance, particularly when the past policy differs substantially from the policy being evaluated. The DR approach improves the inference reliability by integrating the DM and IPW approaches. However, it primarily reduces variance through reward estimation, rather than specifically addressing the variance introduced by the IPW technique itself.

3 Nonparametric framework of policy evaluation

3.1 Nonparametric Model Framework

We now connect two weighing methods to two distinct models and demonstrate that the resulting estimators achieve optimal efficiency under their respective models.

Case 1: IPW. The IPW estimator can be viewed as a model-assisted estimator (Little, 2004). We construct the following working linear model to relate r_{ia} to the probability p_{ia}. For each (x_i, a), we model

$$\pi_{ia} r_{ia} = p_{ia}\beta + p_{ia}\epsilon_{ia}, \tag{3}$$

where $\{\epsilon_{ia}\}$ are assumed be independently distributed with mean 0 and variance σ^2. We estimate β from $\mathcal{S} = \{x_i, a_i, \pi_{ia_i} r_{ia_i}\}_{i=1}^n$. Given $\mathcal{S}$, the model follows the generalized least squares estimator $\hat{\beta}^\pi$ is given as

$$\hat{\beta}^\pi = n^{-1} \sum\nolimits_{i=1}^n p_{ia_i}^{-1} \pi_{ia_i} r_{ia_i}.$$

It follows that the estimator of $\mathbb{E}[\pi_{ia} r_{ia} | p_{ia}]$ is given by $\hat{\mu}_{ia} = p_{ia}\hat{\beta}^\pi$. The corresponding estimate of V^π is $\hat{V}^\pi = n^{-1} \sum_{i=1}^n \sum_{a \in \mathcal{A}} p_{ia} \hat{\beta}^\pi = \hat{\beta}^\pi$, where the last equality holds due to the fact $\sum_{a \in \mathcal{A}} p_{ia} = 1$ for each i. This demonstrates $\hat{V}^\pi = \hat{V}^\pi_{\text{ipw}}$, indicating that the IPW off-policy evaluation can be considered as a prediction under the model (3).

Case 2: Simple Weighting. Simple weighting could perform best under an alternative model. Assume that

$$\pi_{ia} r_{ia} = \mu + \epsilon_{ia}, \tag{4}$$

where $\epsilon_1, \cdots, \epsilon_N$ are assumed be independently distributed with mean 0 and variance σ^2. This model means that there is no any relation between r_{ia} and p_{ia}. Given the dataset $\mathcal{S}$, the model (4) leads to the following estimator for μ:

$$\hat{\mu}^\pi = n^{-1} \sum\nolimits_{i=1}^n \pi_{ia_i} r_{ia_i},$$

which minimizes the variance. The corresponding value estimator is $\hat{V}^\pi_{\text{sw}} = \hat{\mu}^\pi$. This demonstrates that the simple weighting estimator is the most efficient estimator, achieving the minimum variance.

The results from the two cases above indicate that the efficiency of the estimators depends on which model holds. This observation motivates us to consider a flexible model for $\pi_{ia} r_{ia}$ as a function of p_{ia}. We propose the following model to establish a connection between $\pi_{ia} r_{ia}$ with p_{ia}:

$$\pi_{ia} r_{ia} = f^\pi(p_{ia}) + \epsilon^\pi_{ia}, \tag{5}$$

where $f^\pi(p_{ia})$ is an unknown function of p_{ia} and ϵ^π_{ia} is an error term with mean zero and variance $\sigma^2_{\pi,ai}$.

Intuitively, (5) models the conditional expectation $\mathbb{E}[\pi_{ia} r_{ia} | p_{ia}]$, which implicitly defines $f^\pi(p) = \mathbb{E}[\pi_{ia} r_{ia} | p_{ia} = p]$. That is, $f^\pi(p)$ captures the systematic part of $\pi_{ia} r_{ia}$ that can be explained by p_{ia}, while the remainder part is absorbed by the model error term ϵ^π_{ia}.

In the model framework (5), we allow for a flexible structure of $f^\pi(p_{ia})$ without imposing a specific functional form, maintaining an agnostic stance toward model specification. By employing the nonparametric estimation procedure outlined in the next section, we facilitate efficient estimation of $f^\pi(p_{ia})$ within this framework. Importantly, the model framework (5) is intended as an approximation of the relationship between $\pi_{ia} r_{ia}$ and p_{ia}, rather than as a generative model for the reward process.

3.2 Nonparametric Estimation

We provide a brief introduction to the P-spline approach (Eilers & Marx, 1996), which we adopt in this paper. It is worth noting that, under the general framework (5), various nonparametric methods can be employed. Here, we use the P-spline as a representative example due to its wide adoption in the nonparametric modeling community (Eilers & Marx, 1996). Other alternatives may also be promising, though their exploration is beyond the scope of this paper. Consider a univariate regression model

$$y_i = f(x_i) + \zeta_i, \quad i = 1, \cdots, n,$$

where, conditionally given x_i, ζ_i has mean zero and variance $\sigma^2(x_i)$. We assume $f(\cdot) \in W^q[a,b]$, where $W^q[a,b]$ denotes the Sobolev space of order q; that is, $f(\cdot)$ has $q-1$ absolutely continuous derivatives and satisfies $\int_a^b [f^{(q)}(x)]^2 dx < \infty$. For simplicity, we assume $x \in (0,1)$. We adopt the P-spline approach (Eilers & Marx, 1996), which represents the function as

$$f(x) = \sum\nolimits_{k=1}^{J(n)+d} \beta_k B_k(x),$$

where $\{B_k(x) : k = 1, \cdots, J(n)+d\}$ are d-th degree B-spline basis functions with knots $0 = \kappa_0 < \kappa_1 < \cdots < \kappa_{J(n)} = 1$. The number of knots $J(n)$ is chosen such that $J(n) = o(n)$.

The coefficient vector $\beta = (\beta_1, \cdots, \beta_{J(n)+d})^\top$ is estimated using a penalized least-squares approach. Specifically, the estimator $\hat{\beta}$ is obtained by minimizing the following objective function:

$$\sum\nolimits_{i=1}^{n} \left[y_i - \sum\nolimits_{k=1}^{J(n)+d} \beta_k B_k(x) \right]^2 + \lambda_n \sum\nolimits_{k=2}^{J(n)+d} (\Delta \beta_k)^2,$$

where Δ is the first-oder difference operator, i.e., $\Delta\beta_k = \beta_k - \beta_{k-1}$. Solving this optimization problem yields the estimator $\hat{\beta} = My$, where M is the smoothing (hat) matrix resulting from the penalized least-squares procedure and $y = (y_1, \cdots, y_n)^\top$. Denoting $B(x) = (B_1(x), \cdots, B_{J(n)+d}(x))^\top$, the resulting nonparametric regression estimator of $f(x)$ is given by

$$\hat{f}(x) = B(x)^\top \hat{\beta} = B(x)^\top My. \tag{6}$$

Under regularity conditions, the penalized spline estimator is consistent; see Eubank (1999); Claeskens et al. (2009); Xiao (2019). Specifically, for any $x \in (0,1)$, as long as $\lambda_n n^{2q-1} \to \infty$, the mean squared error satisfies $\mathbb{E}[(\hat{f}_n(x) - f(x))^2] = O(\lambda_n/n) + \sigma^2 O(n^{1/(2q-1)} \lambda_n^{-1/(2q)})$. In particular, when $\lambda_n = O(n^{-1/(1+2q)})$, we obtain the optimal convergence rate:

$$\mathbb{E}[(\hat{f}_n(x) - f(x))^2] = O(n^{-2q/(1+2q)}). \tag{7}$$

Throughout the theoretical analysis in the paper, we adopt this optimal rate.

3.3 Nonparametric Weighting

Under the model (5), we have $V^\pi = \mathbb{E}\left[\sum_{a \in \mathcal{A}} f^\pi(p_{ia})\right]$. To estimate $f^\pi(p_{ia})$ from the sample $\mathcal{S} = \{x_i, a_i, \pi_{ia_i} r_{ia_i}\}_{i=1}^n$, we apply the nonparametric estimation described in Section 3.2. Let $y = (\{\pi_{ia_i} r_{ia_i}\}_{i=1}^n)^\top$. Using the P-spline approach in (6), we obtain the estimator $\hat{f}^\pi(p_{ia}) = y^\top w_{\mathrm{nw}}(p_{ia})$, which is linear in $\{\pi_{ia_i} r_{ia_i}\}_{i=1}^n$. Here, the weight vector $w_{\mathrm{nw}}(p_{ia}) = B(p_{ia})^\top M$ is derived from the penalized least-squares estimator presented in (6). We then estimate V^π by

$$\hat{V}^\pi_{\mathrm{nw}} = n^{-1} \sum\nolimits_{i=1}^{n} \sum\nolimits_{a \in \mathcal{A}} \hat{f}^\pi(p_{ia}). \tag{8}$$

This reveals that $\hat{V}^\pi_{\mathrm{nw}}$ is a weighted average of the elements in $\{\pi_{ia_i} r_{ia_i}\}_{i=1}^n$, with the overall weight vector $n^{-1} \sum_{i=1}^n \sum_{a \in \mathcal{A}} w_{\mathrm{nw}}(p_{ia})$. We refer to this approach as *Nonparametric Weighting* (NW).

The procedure is summarized in Algorithm 1 below.

Algorithm 1 Policy Evaluation using the Nonparametric Weighting approach

Given the policy π to evaluate and the values of p_{ia_i}, or estimates thereof obtained from the collected data.
Step 1: Use the P-spline approach in (6) on $\{\pi_{ia_i} r_{ia_i}\}_{i=1}^n$ over $\{p_{ia_i}\}_{i=1}^n$, and obtain the fitted function $\hat{f}^\pi(\cdot)$;
Step 2: Obtain the estimate $\hat{V}^\pi_{\text{nw}}$ in (8).

Robustness to the estimation of p_{ia}. In practice, the true values of p_{ia} are typically unknown and must be estimated from the data. To address this, the function $f(\cdot)$ is subsequently fitted using the estimates $\hat{p}_{ia}$. We now examine the implications of using $\hat{p}_{ia}$, accounting for both the estimation error and bias arising from potential misspecification in the estimation of p_{ia}.

Given the estimates $\hat{p}_{ia}$, we write $\hat{p}_{ia} = p^*_{ia} + o_p(1)$, indicating that $\hat{p}_{ia} \to p^*_{ia}$ in probability. Here, the difference $p^*_{ia} - p_{ia}$ reflects the bias due to potential model misspecification, while the term $o_p(1)$ captures estimation error. Suppose that $f(\cdot)$ is Lipschitz continuous, and that p^*_{ia} is a transformation of p_{ia} through a mapping $h(\cdot)$, i.e., $p^*_{ia} = h(p_{ia})$. Then we can express $f(\hat{p}_{ia})$ as

$$f(\hat{p}_{ia}) = f(h(p_{ia})) + o_p(1) = f^{(h)}(p_{ia}) + o_p(1),$$

where $f^{(h)}(p_{ia})$ denotes the composition of $f(h(p_{ia}))$ with $h(\cdot)$. This shows that the function evaluated at the estimated probabilities corresponds to a transformed version of the function evaluated at the true probabilities, up to a negligible error term. Due to the flexibility of the unknown function, the impact of this transformation is minimal. As a result, the procedure remains robust to the errors in estimating p_{ia}—that is, using $\hat{p}_{ia}$ in place of p_{ia} does not substantially affect the estimation accuracy.

3.4 Error Analysis

We now analyze the error bounds of the NW estimator $\hat{V}^\pi_{\text{nw}}$ relative to V^π to address the question: How well does this estimator perform? Let $\Delta^\pi_f(p) = \hat{f}^\pi(p) - f^\pi(p)$, and define $\bar{V}^\pi = n^{-1}\sum_{i=1}^n \sum_{a\in\mathcal{A}} f^\pi(p_{ia})$ as the sample average of $f^\pi(p_{ia})$. Then, the error of the NW estimator can be decomposed as

$$\begin{aligned}\hat{V}^\pi_{\text{nw}} - V^\pi =& \hat{V}^\pi_{\text{nw}} - \bar{V}^\pi + \bar{V}^\pi - V^\pi \\ =& n^{-1}\sum\nolimits_{i=1}^n \sum\nolimits_{a\in\mathcal{A}} \Delta^\pi_f(p_{ia}) + n^{-1}\sum\nolimits_{i=1}^n \sum\nolimits_{a\in\mathcal{A}} [f^\pi(p_{ia})] - V^\pi. \end{aligned} \tag{9}$$

Noting $\mathbb{E}[(\Delta^\pi_f(p_{ia}))^2] = O(n^{-2q/(1+2q)})$, as established in Eqn. (7), and $\mathbb{E}[(\bar{V}^\pi - V^\pi)^2] = O(n^{-1})$, we see from Eqn. (9) that the uncertainty in $\hat{V}^\pi_{\text{nw}}$ relative to V^π mainly stems from $\Delta^\pi_f(p_{ia})$, which captures the estimation error in the nonparametric procedure. The following proposition establishes the convergence rates for the bias and mean squared error of $\hat{V}^\pi_{\text{nw}}$.

Proposition 3.1. *(1) The bias of the NW estimator is given by*

$$\mathbb{E}(\hat{V}^\pi_{nw}) - V^\pi = O(Kn^{-q/(1+2q)}).$$

(2) The MSE of the NW estimator is given by

$$\mathbb{E}[(\hat{V}^\pi_{nw} - V^\pi)^2] = O(K^2 n^{-2q/(1+2q)}).$$

Proposition 3.1 shows that the convergence rates depend on both K and n, under the condition that $Kn^{-q/(1+2q)} \to 0$ as $n \to \infty$. As a result, convergence is guaranteed even in the presence of a large action space, provided that $K = o(n^{q/(1+2q)})$.

3.5 An Illustration Example

We present an illustrative example to demonstrates how our proposed nonparametric weighting method can improve efficiency. We simulate a K-armed bandit setting, where each arm $k = 1, \cdots, K$ has rewards y^2_{ik}, with $y_{ik} \sim N(0,1)$, for each context $i = 1, \cdots, 300$, using three different ways:

- Sorting $\{y_{i1}^2, \cdots, y_{iK}^2\}$ in increasing order;
- Sorting $\{y_{i1}^2, \cdots, y_{iK}^2\}$ in decreasing order;
- Leaving $\{y_{i1}^2, \cdots, y_{iK}^2\}$ unsorted.

We perform sampling without replacement, proportional to a set of K values that are drawn from a uniform distribution over $(0, 1)$ and sorted in increasing order.

Sorting the rewards in increasing order induces a strong positive correlation between the rewards and the sampling probabilities, while sorting them in decreasing order induces a strong negative correlation. In such cases, an approximate working model can exploit this explanatory power through nonparametric modeling of the probabilities. In contrast, when the rewards remain unsorted, this correlation disappears, leaving the rewards and sampling probabilities uncorrelated. In this case, the probabilities offer no additional explanatory power, and the simple weighting (SW) estimator provides the most efficient evaluation.

We set $K = 20$ and obtain a sample of size $n = 300$. Our goal is to evaluate the uniform target policy $\pi_k = 1/K$, leading to the objective of estimating $V = (nK)^{-1} \sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik}^2$. We estimate V using simple weighting (SW), inverse probability weighting (IPW), and the proposed nonparametric weighting (NW). With $B = 2000$ iterations, we calculate the bias, standard deviation (s.d.), and root mean square error (RMSE) for each estimator. The results are presented in Table 1. From the table, the variance of the NW

Table 1: Performance of Example 1

Case	method	Bias	s.d.	RMSE
decreasing	SW	-0.5858	**0.0401**	0.5871
	IPW	**0.0165**	0.4913	0.4916
	NW	-0.0701	0.1481	**0.1639**
increasing	SW	0.5766	0.0904	0.5836
	IPW	**-0.0018**	0.0447	0.0448
	NW	0.0253	**0.0310**	**0.0400**
unsorted	SW	0.0181	**0.0814**	**0.0833**
	IPW	**0.0037**	0.2577	0.2577
	NW	0.0068	0.1095	0.1097

estimator is significantly smaller than that of the IPW estimator across all three cases, resulting in much higher efficiency in terms of MSE, while the NW estimator exhibits slightly larger absolute bias compared to the IPW estimator. Compared to the SW estimator, which has the lowest variance in the unsorted and decreasing cases but the largest absolute bias, the NW estimator performs much better in the decreasing and increasing cases while being slightly worse in the unsorted case, as expected.

4 Model-assisted Nonparametric Weighting

In this section, we extend the nonparametric weighting approach by integrating the DM estimator and propose a model-assisted nonparametric weighting (MNW) estimator. Given an estimate $\hat{\mu}_{ia}$ of the expected reward for the context-action pair (x_i, a), we compute the residual as $\pi_{ia}(r_{ia} - \hat{\mu}_{ia})$. We model the relationship between $\pi_{ia}(r_{ia} - \hat{\mu}_{ia})$ and the probability p_{ia} using the following nonparametric model for each (x_i, a):

$$\pi_{ia}(r_{ia} - \hat{\mu}_{ia}) = g^{\pi}(p_{ia}) + \xi_{ia}^{\pi}, \tag{10}$$

where $\{\xi_{ia}^{\pi}\}$ are independently distributed with mean 0 and variance $\nu_{\pi,ai}^2$.

Similar to Section 3.3, we estimate the function $g^{\pi}(\cdot)$ from the sample $\mathcal{S}$ and obtain the nonparametric estimate $\hat{g}^{\pi}(\cdot)$ under the model (10). Given $\hat{g}^{\pi}(\cdot)$, we construct the MNW estimator as

$$\hat{V}_{\text{mnw}}^{\pi} = n^{-1} \sum\nolimits_{i=1}^{n} \sum\nolimits_{a \in \mathcal{A}} (\hat{g}_{ia}^{\pi} + \pi_{ia} \hat{\mu}_{ia}). \tag{11}$$

Like the the doubly robust (DR) estimator, the MNW estimator incorporates $\hat{\mu}_{ia}$ as a baseline when a reward model is available, reducing variance by nonparametric modeling the residuals $r_{ia} - \hat{\mu}_{ia}-$. At the same time, this incorporation does not introduce bias. In other words, the MNW estimator remains robust to misspecification in $\hat{\mu}_{ia}$: when $\hat{\mu}_{ia}$ is biased, $\hat{V}^{\pi}_{\text{mnw}}$ compensates through the nonparametric adjustment provided by $\hat{g}^{\pi}(\cdot)$, which captures and corrects the residual bias. As a result, the MNW estimator effectively adjusts for misspecification in the reward model $\mu_a(\cdot)$, achieving higher efficiency when $\hat{\mu}_a(\cdot)$ is accurate and maintaining robustness when it is not. The procedure is summarized in Algorithm 2 below.

Algorithm 2 Policy Evaluation using the Model-assisted Nonparametric Weighting approach

Given the policy π to evaluate and the values of p_{ia_i}, or estimates thereof obtained from the collected data.
Step 1: Estimate μ_{ia} and obtain the fitted $\hat{\mu}_{ia}$;
Step 2: Use the P-spline approach on $\{\pi_{ia_i}(r_{ia_i} - \hat{\mu}_{ia_i})\}_{i=1}^{n}$ over $\{p_{ia_i}\}_{i=1}^{n}$, and obtain the fitted function $\hat{g}^{\pi}_{ia}(\cdot)$;
Step 3: Obtain the estimate $\hat{V}^{\pi}_{\text{mnw}}$ according to (11).

4.1 Error Analysis

We now analyze the bias and MSE of the MNW estimator. Denoting $\mu^*_{ia} = \mathbb{E}(\hat{\mu}_{ia})$, the bias introduced by the reward model is given by $\mu_{ia} - \mu^*_{ia}$. Under the nonparametric model (10), we have

$$\mathbb{E}[\pi_{ia}(r_{ia} - \hat{\mu}_{ia})] = \pi_{ia}(\mu_{ia} - \mu^*_{ia}) = g^{\pi}(p_{ia}). \tag{12}$$

Define $\bar{V}^{\pi}_{\text{mnw}} = n^{-1}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}}\pi_{ia}\mu_{ia}$. Using (12), we obtain

$$\bar{V}^{\pi}_{\text{mnw}} = n^{-1}\sum\nolimits_{i=1}^{n}\sum\nolimits_{a\in\mathcal{A}}\left(\pi_{ia}\mu^*_{ia} + g^{\pi}(p_{ia})\right). \tag{13}$$

Let $\Delta^{\pi}_{g}(p) = \hat{g}^{\pi}(p) - g^{\pi}(p)$. From (13), we have $\hat{V}^{\pi}_{\text{mnw}} - \bar{V}^{\pi}_{\text{mnw}} = \Delta^{\pi}_{g}(p_{ia}) + (\hat{\mu}_{ia} - \mu^*_{ia})$. This leads to the following decomposition:

$$\begin{aligned}\hat{V}^{\pi}_{\text{mnw}} - V^{\pi} =& \hat{V}^{\pi}_{\text{mnw}} - \bar{V}^{\pi}_{\text{mnw}} + \bar{V}^{\pi}_{\text{mnw}} - V^{\pi} \\ =& n^{-1}\sum\nolimits_{i=1}^{n}\sum\nolimits_{a\in\mathcal{A}}\left[\Delta^{\pi}_{g}(p_{ia}) + (\hat{\mu}_{ia} - \mu^*_{ia})\right] + n^{-1}\sum\nolimits_{i=1}^{n}\left[V^{\pi}_{\text{mnw}}(i) - V^{\pi}\right].\end{aligned} \tag{14}$$

where $V^{\pi}_{\text{mnw}}(i) = \sum_{a\in\mathcal{A}}\left(\pi_{ia}\mu^*_{ia} + g^{\pi}(p_{ia})\right)$. Based on this decomposition, we derive the convergence rates for the bias and mean squared error of $\hat{V}^{\pi}_{\text{mnw}}$, as stated in the following proposition.

Proposition 4.1. *Assume $\mathbb{E}[(\hat{\mu}_{ia} - \mu^*_{ia})^2] = O(n^{-1})$. Then, we have the following results: (1) The bias of the MNW estimator is given by*

$$\mathbb{E}(\hat{V}^{\pi}_{mnw}) - V^{\pi} = O(Kn^{-q/(2q+1)}).$$

(2) The MSE of the MNW estimator is given by

$$\mathbb{E}[(\hat{V}^{\pi}_{mnw} - V^{\pi})^2] = O(K^2 n^{-2q/(2q+1)}).$$

Proposition 4.1 demonstrates that the MNW estimator remains consistent despite discrepancies in reward modeling (i.e., $\mu_{ia} - \mu^*_{ia}$), as these discrepancies are corrected by $g^{\pi}(\cdot)$.

Comparison with the NW estimator. Eqn. (14) shows that $\hat{V}^{\pi}_{\text{mnw}} - \bar{V}^{\pi}_{\text{mnw}}$ can be decomposed into two components: $\Delta^{\pi}_{g}(p_{ia})$, which reflects the error in estimating $g^{\pi}(\cdot)$, and $\hat{\mu}_{ia} - \mu^*_{ia}$, the deviation of the estimated reward from its expectation. When the variability of $\hat{\mu}_{ia} - \mu^*_{ia}$ is smaller than that of $\Delta^{\pi}_{g}(p_{ia})$, the dominant source of error arises from $\Delta^{\pi}_{g}(p_{ia})$. Comparing $\Delta^{\pi}_{g}(p_{ia})$ in the MNW estimator with $\Delta^{\pi}_{f}(p_{ia})$ in the NW estimator, it becomes evident that the MNW estimator can achieve higher efficiency when $\pi_{ia}\hat{\mu}_{ia}$ effectively captures context-dependent reward information.

5 Experiments

This section provides empirical evidence supporting the effectiveness of the NW estimator over IPW and the MNW estimator over DR. Although more recent methods for off-policy evaluation have been developed (see Section 6 below), we restrict our comparison to the IPW, and DR estimators. The rationale is that these estimators are conceptually representative and sufficient to highlight the essential differences between weighting-based and nonparametric approaches. Including a larger set of methods would not necessarily provide additional insights, as our primary aim is to examine the relative performance and robustness of the NW-type estimators in comparison with the standard IPW framework.

Following Dudík et al. (2011), we consider a multi-class classification problem with bandit feedback, using public benchmark datasets. We use the same datasets as in Dudík et al. (2011). All experiments were conducted on a MacBook Air equipped with M3.

In a classification task, we assume that the data are drawn independently and identically distributed from a fixed distribution: $(x, c) \sim P$, where $x \in \mathcal{X}$ is the feature vector and $c \in \{1, \cdots, K\}$ is the class label. The goal is to find a classifier $\pi : \mathcal{X} \to \{1, \cdots, K\}$ that minimizes the classification error $e^\pi = \mathbb{E}_P[I(\pi(x) \neq c)]$. We turn the data point (x, c) into a cost-sensitive classification example $(x, l_1, \cdots, l_K)$, where l_a denotes the loss for predicting a. We consider a noisy loss where l_a is drawn from a Gaussian distribution with mean $I(a \neq c)$ and $\sigma = 0.2$. Under this formulation, the classifier π can be viewed as an action-selection policy, and its classification error is exactly the policy's expected loss. The target policy π is defined as the deterministic decision of a logistic regression classifier learned on the multi-class data.

The logging policy b is generated by sampling probability proportional to values drawn from a uniform distribution; specifically, we draw $p_a \sim \text{unif}(0, 1)$ and assign action a according to $b_a \propto p_a$. To predict the value for action a for unit i in the testing dataset, we first fit an l_2-regularized least-squares estimates using the training data, and then generate predictions for each data point in the test set.

We randomly split data into training and test sets of (roughly) the same size. On the training set, we train a multinomial classifier to define the policy π, and compute the classification error on the test data, treating it as the ground truth for comparing various estimates. We perform Monte Carlo simulation with 500 iterations to generate a partially labeled set from the test data using the logging policy b. We compute policy evaluation estimates using various methods and calculate the resulting bias and root mean squared error (RMSE). We repeat this entire process 20 times, and report the average bias and RMSE values across these 20 runs in Table 2.

Table 2: Bias and RMSE of various estimators for classification error with the true logging policy.

Data	Bias					RMSE				
	DM	IPW	DR	NW	MNW	DM	IPW	DR	NW	MNW
letter	0.507	0.001	0.009	**0.000**	0.009	0.507	0.070	0.075	**0.036**	0.045
glass	0.218	-0.003	0.034	**0.000**	0.036	0.218	0.233	0.238	0.238	**0.193**
ecoli	0.215	0.002	0.037	**0.000**	0.034	0.215	0.270	0.229	0.243	**0.208**
opt	0.292	0.001	-0.001	**0.000**	-0.001	0.292	0.047	0.060	**0.027**	0.037
page	0.086	**0.000**	0.012	-0.001	0.012	0.086	0.021	0.028	**0.014**	0.020
pen	0.400	0.000	0.004	**0.000**	0.004	0.400	0.033	0.051	**0.022**	0.032
sat	0.192	0.000	0.020	**0.000**	0.021	0.192	0.041	0.043	**0.024**	0.031
vehicle	0.239	0.001	0.017	**0.000**	0.016	0.239	0.079	0.078	0.058	**0.057**
yeast	0.215	**0.000**	0.070	-0.002	0.070	0.215	0.139	0.140	**0.098**	0.106

From the table, the RMSE of the NW estimator is consistently lower than that of the IPW estimator across all datasets, while its bias is remains negligible and comparable to that of the IPW estimator. The DM estimator performs the worst, reflecting the poor fit of the linear model used to predict rewards. In comparison, the MNW estimator achieves substantially lower RMSE than the DR estimator, while maintaining a similar level of bias.

Table 3: Bias and RMSE of various estimators for classification error under estimated logging policy.

Data	Bias					RMSE				
	DM	IPW	DR	NW	MNW	DM	IPW	DR	NW	MNW
letter	0.505	0.066	-0.085	**0.000**	0.009	0.505	0.319	0.337	**0.034**	0.041
glass	0.215	0.044	0.015	**-0.002**	0.039	0.215	0.302	0.282	0.226	**0.188**
ecoli	0.206	0.060	0.012	**0.002**	0.039	0.206	0.334	0.272	0.221	**0.190**
opt	0.291	0.006	-0.052	**0.000**	-0.001	0.291	0.057	0.115	**0.025**	0.035
page	0.088	0.004	-0.001	**0.000**	0.012	0.088	0.043	0.048	**0.013**	0.019
pen	0.398	0.020	-0.095	**0.000**	0.004	0.398	0.081	0.216	**0.020**	0.030
sat	0.192	0.028	-0.007	**-0.002**	0.021	0.192	0.074	0.085	**0.023**	0.030
vehicle	0.236	0.021	-0.027	**-0.001**	0.015	0.236	0.126	0.164	0.054	**0.053**
yeast	0.218	0.044	-0.043	**-0.003**	0.069	0.218	0.230	0.470	**0.089**	0.101

How much does estimating the logging policy affect the results? To assess the impact of estimating the logging policy b, we conduct another experiment using a perturbed logging policy defined as $\tilde{b} \propto p_a * \delta$, where $\delta \sim \mathcal{N}(1, 0.09)$. Here, δ introduces Gaussian noise into the probability estimates, simulating errors in estimating the logging policy that generated the data. We calculate the bias and RMSE values of each estimator using $\tilde{b}$ and report the results in Table 3. As shown in the table, the IPW and DR estimators exhibit significantly larger RMSE, highlighting their sensitivity to errors in probability estimation. Notably, the IPW estimator shows a marked increase in bias, suggesting that it may become biased under noisy probability estimates. In contrast, the NW and MNW estimators yield RMSE values similar to those obtained when the logging policy is known, demonstrating that our proposed methods are robust to inaccuracies in the estimated probabilities.

6 Related Work

Off-policy evaluation for bandits has been extensively studied. Below, we provide a concise review of the most relevant work to the best of our knowledge.

To address the high variance of IPW-based methods, several simple techniques have been proposed. One common approach is weight clipping, which truncates large inverse probability weights (Ionides, 2008; Swaminathan & Joachims, 2015a; Su et al., 2019). Another is normalization, which rescales the weights to stabilize estimates (Swaminathan & Joachims, 2015b). In contrast, our paper takes a fundamentally different approach by modeling the probabilities directly, thereby avoiding the issue of large weights in a data-driven way.

Doubly robust estimation (Robins et al., 1994; Lunceford & Davidian, 2004; Kang & Schafer, 2007) is widely used for parameter estimation in statistical inference. Dudík et al. (2011) first applied this framework to policy evaluation and optimization, with later extensions to the reinforcement learning setting by Jiang & Li (2016) and Thomas & Brunskill (2016). More recent developments include the work of Wang et al. (2017), Farajtabar et al. (2018), and Su et al. (2020). While the goal of our MNW approach is not to guarantee the standard doubly robust property, it explicitly models and mitigates the bias introduced by reward modeling. Doubly robust estimation has also been explored in observational settings for estimating average treatment effects using asymptotically optimal estimators (Hirano et al., 2003; Imbens et al., 2007).

Several data-driven approaches have been proposed for off-policy evaluation. Saito et al. (2021) selected one of multiple logging policies as a pseudo-target policy, directly estimated its value from the dataset, and used it to identify the off-policy estimator with the best performance. Udagawa et al. (2023) introduced two surrogate policies constructed from the logged data. Cief et al. (2024) proposed a cross-validated off-policy evaluation framework. In contrast to these methods, our approach leverages flexible modeling, offering both ease of implementation and reliable performance.

Finally, it is worth noting that nonparametric modeling of sampling probabilities is not new. In survey sampling, such techniques have been employed to correct for sample selection bias and enable finite population inference; see the review article (Little, 2004). Motivated by this line of work, we propose a nonparametric wighting approach for policy evaluation, and further enhance it by incorporating the regression-based adjustment to improve accuracy.

7 Conclusion

In this paper, we have proposed nonparametric weighting estimators for off-policy evaluation. The weights are constructed from the P-spline approach. Furthermore, we incorporate the DM estimator to further reduce the variance in the rewards and develop a model-assisted nonparametric weighting estimator. Extensive experiments on a multi-class classification with bandit feedback, using public benchmark datasets, demonstrate the efficacy of these estimators and emphasize the role of nonparametric weighting in achieving superior performance. As a result, we anticipate that the NW approach will become standard alternatives to the IPW approach.

This study has some limitations. First, our method does not rely on selection bias adjustment but instead depends heavily on the general model framework (5). When the rewards are only weakly correlated with the selection probabilities, the nonparametric model offers little improvement, as there is no strong relationship to capture; in such cases, the NW estimator essentially reduces to simple averaging. Second, the NW-type estimators depend on the choice of nonparametric method. Although the P-spline approach demonstrated robust performance in our experiments, implementing the framework with alternative nonparametric methods could further enhance its flexibility and predictive accuracy. This direction warrants further investigation.

Several promising directions for further research arise from this work. Given the widespread use of IPW in policy evaluation across various domains, we anticipate that our NW approach will have competitive applications in these areas. For instance, extending it to combinatorial contextual bandits (Swaminathan et al., 2017) could offer valuable insights, particularly in decision-making scenarios with complex action spaces. Furthermore, integrating this methodology within the broader reinforcement learning framework presents a promising opportunity to improve policy evaluation and learning in sequential decision-making problems. Finally, our paper primarily focuses on settings with small action spaces. Extending our approach to large action spaces is a promising direction, as importance weighting can break down in such cases due to extreme variance from large importance weights (Saito & Joachims, 2022).

References

C. M. Cassel, C. E. Särndal, and J. H. Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63:615–620, 1976.

M. Cief, B. Kveton, and M. Kompan. Cross-validated off-policy evaluation. In *AAAI*, 2024.

G. Claeskens, T. Krivobokova, and J. D. Opsomer. Asymptotic properties of penalized spline estimators. *Biometrika*, 96(3):529–544, 2009.

M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

P.H.C. Eilers and B.D. Marx. Flexible smoothing with b-splines and penalties (with discussion). *Statistical Science*, 11:89–121, 1996.

R.L. Eubank. *Nonparametric Regression and Spline Smoothing*. New York: Marcel Dekker., 1999.

M. Farajtabar, Y. Chow, and M. Ghavamzadeh. More robust doubly robust off-policy evaluation. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

K. Hirano, G. W Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71:1161–1189, 2003.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952.

G. Imbens, W. Newey, and G. Ridder. Mean-squared-error calculations for average treatment effects. Available at SSRN: http://dx.doi.org/10.2139/ssrn.954748, 2007.

E. L Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2): 295–311, 2008.

N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2016.

J. D. Y. Kang and J. L. Schafer. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22:523–539, 2007.

J. Langford and T Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*, 2008.

L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *International Conference on Web Search and Data Mining*, 2011.

Roderick J Little. To model or not to model? competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99(466):546–556, 2004.

J. K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23:2937–2960, 2004.

J. Robins and A Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90:122–129, 1995.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866, 1994.

Y. Saito and T. Joachims. Off-policy evaluation for large action spaces via embeddings. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Y. Saito, T. Udagawa, H. Kiyohara, K. Mogi, Y. Narita, and K. Tateno. Evaluating the robustness of off-policy evaluation. In *Proceedings of the 15th ACM Conference on Recommender Systems*, 2021.

A. Strehl, J. Langford, L. Li, and S. M Kakade. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems*, 2010.

Y. Su, L. Wang, M. Santacatterina, and T. Joachims. CAB: Continuous adaptive blending estimator for policy evaluation and learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Y. Su, M. Dimakopoulou, A. Krishnamurthy, and M. Dudík. Doubly robust off-policy evaluation with shrinkage. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems*, 2015a.

A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems*, 2015b.

A. Swaminathan, A. Krishnamurthy, A. Agarwal, M. Dudik, J. Langford, D. Jose, and I. Zitouni. Off-policy evaluation for slate recommendation. In *Advances in Neural Information Processing Systems*, 2017.

P. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2016.

T. Udagawa, H. Kiyohara, Y. Narita, Y. Saito, and K. Tateno. Policy-adaptive estimator selection for off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Y.X. Wang, A. Agarwal, and M. Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

L. Xiao. Asymptotic theory of penalized splines. *Electronic Journal of Statistics*, 13:747–794, 2019.

Appendix

The Proof of Proposition 3.1

Eqn. (9) directly follows

$$\mathbb{E}[\hat{f}^{\pi}(p_{ia})] - \pi_{ia}\mu_{ia} = \mathbb{E}[\Delta_f^{\pi}(p_{ia})] = O(n^{-q/(1+2q)}).$$

Thus, we have the bias

$$\mathbb{E}[\hat{V}_{\mathrm{nw}}^{\pi}] - V^{\pi} = n^{-1}\sum\nolimits_{i=1}^{n}\sum\nolimits_{a\in\mathcal{A}}\mathbb{E}\left[\Delta_f^{\pi}(p_{ia})\right] = O(Kn^{-q/(1+2q)}).$$

Noting $\mathbb{E}[(\hat{V}_{\mathrm{nw}}^{\pi} - \bar{V}_{\mathrm{nw}}^{\pi})(\bar{V}_{\mathrm{nw}}^{\pi} - V^{\pi})] = 0$ from (9), we have the mean squared error of $\hat{V}_{\mathrm{nw}}^{\pi}$:

$$\mathbb{E}[(\hat{V}_{\mathrm{nw}}^{\pi} - V^{\pi})^2] = \mathbb{E}[(\hat{V}_{\mathrm{nw}}^{\pi} - \bar{V}_{\mathrm{nw}}^{\pi})^2] + \mathbb{E}[(\bar{V}_{\mathrm{nw}}^{\pi} - V^{\pi})^2]. \tag{15}$$

For the first term of the righ-hand side of (15), we have

$$\begin{aligned}\mathbb{E}[(\hat{V}_{\mathrm{nw}}^{\pi} - \bar{V}_{\mathrm{nw}}^{\pi})^2] &\leq \left(n^{-1}\sum\nolimits_{i=1}^{n}\mathbb{E}\left[\sum\nolimits_{a\in\mathcal{A}}\Delta_f^{\pi}(p_{ia})\right)^2\right] \\ &\leq \left(n^{-1}\sum\nolimits_{i=1}^{n}|\mathcal{A}|\sum\nolimits_{a\in\mathcal{A}}\mathbb{E}\left[\Delta_f^{\pi}(p_{ia})\right]^2\right) \\ &= O(K^2 n^{-2q/(1+2q)}).\end{aligned} \tag{16}$$

For the second term of the righ-hand side of (15), $\bar{V}^{\pi}$ is a sample simple average of a set of sample with expectation V^{π}, following we have

$$\mathbb{E}[(\bar{V}_{\mathrm{nw}}^{\pi} - V^{\pi})^2] = O(n^{-1}). \tag{17}$$

Inserting (16) and (17) to (15), the bound on the MSE in Proposition 3.1 is proved.

The Proof of Proposition 4.1

Similar to the proof of Proposition 3.1, we have

$$\mathbb{E}[\hat{V}_{\mathrm{mnw}}^{\pi} - V^{\pi}] = O(Kn^{-q/(2q+1)}) + O(Kn^{-1}).$$

For the MSE term, similar to the proof of Proposition 3.1, we have

$$\begin{aligned}\mathbb{E}[(\hat{V}_{\mathrm{mnw}}^{\pi} - V^{\pi})^2] &= \mathbb{E}[(\hat{V}_{\mathrm{mnw}}^{\pi} - \bar{V}_{\mathrm{mnw}}^{\pi})^2] + \mathbb{E}[(\bar{V}_{\mathrm{mnw}}^{\pi} - V^{\pi})^2]; \\ \mathbb{E}[(\bar{V}_{\mathrm{mnw}}^{\pi} - V^{\pi})^2] &= O(n^{-1})\end{aligned}$$

Now we investigate the term $\mathbb{E}[(\hat{V}_{\mathrm{mnw}}^{\pi} - \bar{V}_{\mathrm{mnw}}^{\pi})^2]$, and have

$$\begin{aligned}\mathbb{E}[(\hat{V}_{\mathrm{mnw}}^{\pi} - \bar{V}_{\mathrm{mnw}}^{\pi})^2] &= \mathbb{E}\left(n^{-1}\sum\nolimits_{i=1}^{n}\sum\nolimits_{a\in\mathcal{A}}\left[\Delta_g^{\pi}(p_{ia}) + \pi_{ia}(\hat{\mu}_{ia} - \mu_{ia}^{*})\right]^2\right) \\ &\leq n^{-1}|\mathcal{A}|\sum\nolimits_{i=1}^{n}\sum\nolimits_{a\in\mathcal{A}}\mathbb{E}\left(\left[\Delta_g^{\pi}(p_{ia}) + \pi_{ia}(\hat{\mu}_{ia} - \mu_{ia}^{*})\right]^2\right) \\ &\leq 2n^{-1}|\mathcal{A}|\sum\nolimits_{i=1}^{n}\sum\nolimits_{a\in\mathcal{A}}\left(\mathbb{E}\left[\Delta_g^{\pi}(p_{ia})\right]^2 + \mathbb{E}\left(\pi_{ia}(\hat{\mu}_{ia} - \mu_{ia}^{*})\right)^2\right) \\ &= O(K^2 n^{-q/(1+2q)}) + O(K^2 n^{-1}).\end{aligned}$$

Thus, the MSE expression in the theorem follows.