Online Supplement to “Solving Bayesian Risk Optimization via Nested Stochastic Gradient Estimation”
by “Cakmak, Wu, and Zhou”

Contents

1 Overview 1

2 Discussion on the projection operator 2

3 Complete list of assumptions 3

4 Discussion on the assumptions 4
   4.1 A simple example . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4
   4.2 A general class of functions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
      4.2.1 Extensions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

5 Details of numerical experiments 7
   5.1 Details of the quadratic example . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7
   5.2 Convergence analysis of MCMC output . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

6 Results and proofs 9

1 Overview

Recall that the SA algorithm takes the form

\[ x_{t+1} = \Pi_X [x_t + \epsilon_t Y_t], \]

where we use \( Y_t = -\varphi_{n_t,m_t}(x_t) \) or \( Y_t = -\psi_{n_t,m_t}(x_t) \) depending on the choice of the risk measure. Our
estimators are given by
\[
\varphi_{\alpha}^{n,m}(x) := \partial_x \hat{H}^m(x; \theta) \big|_{H^m(x; \theta) = \hat{\psi}^n_{\alpha} m \theta} = \hat{D}^m(x; \hat{\theta}_b^{\alpha n})
\]
for VaR; and
\[
\psi_{\alpha}^{n,m}(x) := \frac{1}{n(1 - \alpha)} \sum_{i=1}^{n} \hat{D}^m(x; \theta_i) 1_{(H^m(x; \theta_i) \geq \hat{\psi}^n_{\alpha} m)}(x)
\]
for CVaR.

We start with a brief discussion on the projection operator \( \Pi_X \). For the readers’ convenience, we repeat the full set of assumptions here. We continue with a detailed discussion on the assumptions, verify them on a simple example, and show that Assumption 3.5 is satisfied for a general class of problems. We fill in the details of the numerical experiments, and conclude with the proofs of the results presented in the paper.

2 Discussion on the projection operator

The SA algorithm requires the use of a projection operator to ensure that the iterates remain within \( X \). The projection often requires solving the following optimization problem:

\[
\min_x \|x - y\|_2 \quad \text{s.t. } x \in X,
\]

where \( y = x_t + \epsilon_t Y_t \), and the solution \( x \) is the projection of \( y \) onto \( X \), i.e., the point in \( X \) with the shortest Euclidean distance to \( y \).

When \( X \) is a simple set, e.g., the hyper-rectangle \( X = \{x : a^i \leq x^i \leq b^i\} \), where \( x^i \) is the \( i^{th} \) coordinate of \( x \), the projection is very simple. If \( y^i < a^i \), we set \( x^i = a^i \); if \( y^i > b^i \), we set \( x^i = b^i \); and set \( x^i = y^i \) otherwise. Similarly, when \( X = \{x : \sum_i (x^i)^2 \leq S\} \) is a sphere, the projection reduces to \( x = S y/\|y\|_2^2 \). Other examples where the projection is available in a closed form can be constructed.

For general \( X \), one still needs to solve the optimization problem (1). When \( X \) is a polytope, (1) defines a quadratic program, which can be solved using commercially available optimization packages. For general convex \( X \), (1) is a convex optimization problem, which can again be efficiently optimized. The main challenge is when \( X \) is a non-convex set. In this case, we are not aware of general purpose projection algorithms, and the methods tend to be problem specific. Below, we describe an alternative approach that only requires a membership oracle of \( X \) and can be used to avoid projection when no efficient projection algorithm is available.

In the worst case, when no efficient projection algorithm is available, one can avoid projection altogether by shrinking the step-size. If \( y = x_t + \epsilon_t Y_t \) is not in \( X \), we can instead try a smaller step-
size, e.g., \( \epsilon_t/2 \). We can keep reducing the step-size until we find a feasible step, or we reach a tolerance threshold where we decide no meaningful step is possible. Note that in the multi-variate case, it is possible that while a step in the \( i^{th} \) coordinate is feasible, a step in the \( j^{th} \) \( (i \neq j) \) coordinate is infeasible. In such setting, we can treat each dimension independently, and choose to take a step in the direction that permits the largest step-size. Many other variations on this is possible, and the design of an ideal strategy is left to the reader. We conclude this discussion by noting that when no positive step size is feasible, the iterate has reached a stationary point of the solution set, and the algorithm can be terminated.

## 3 Complete list of assumptions

### Assumption 3.2. Zhu, Liu, and Zhou (2020)

1. For all \( x \in \mathcal{X} \), the response \( h(x, \xi(\theta)) \) has finite conditional second moment, i.e.,

\[
\tau_2^2 = \mathbb{E}_{\mathcal{F}}[h(x, \xi)^2] < \infty \quad \text{w.p. 1 (P}^N) \quad \text{and} \quad \tau^2 = \mathbb{E}_{\mathcal{F}}[h(x, \xi)^2] = \int \tau_2^2 dP^N < \infty.
\]

2. The joint density \( p_m(h, e) \) of \( H(x; \theta) \) and \( \xi^m(x; \theta) \), and its partial gradients \( \frac{\partial}{\partial h} p_m(h, e) \) and \( \frac{\partial^2}{\partial^2 h} p_m(h, e) \) exist for each \( m \), all pairs of \((h, e)\) and for all \( x \in \mathcal{X} \).

3. For all \( x \in \mathcal{X} \), there exists non-negative functions \( g_{0,m}(\cdot), g_{1,m}(\cdot) \) and \( g_{2,m}(\cdot) \) such that \( p_m(h, e) \leq g_{0,m}(e), |\frac{\partial}{\partial h} p_m(h, e)| \leq g_{1,m}(e), \ |\frac{\partial^2}{\partial^2 h} p_m(h, e)| \leq g_{2,m}(e) \) for all \((h, e)\). Furthermore, \( \sup_m \int |e|^r g_{i,m}(e) d\theta < \infty \) for \( i = 0, 1, 2 \), and \( 0 \leq r \leq 4 \).

### Assumption 3.3. There exists a random variable \( K(\xi(\theta)) \) such that \( \mathbb{E}_{\mathcal{F}}[K(\xi)] < \infty \), and the following holds in a probability 1 (\( P^N \)) subset of \( \Theta \).

1. \( |h(x_2, \xi(\theta)) - h(x_1, \xi(\theta))| \leq K(\xi(\theta))|x_2 - x_1| \) w.p.1 \( (P_\theta) \) for all \( x_1, x_2 \in \mathcal{X} \).

2. The sample path gradient \( d(x, \xi(\theta)) \) exists w.p.1 \( (P_\theta) \).

### Assumption 3.4. Hong (2009) For any \( x \in \mathcal{X} \), \( H(x, \theta) \) has a continuous density \( f(t; x) \) in a neighborhood of \( t = v_\alpha(x) \), and \( \partial_x F(t; x) \) exists and is continuous w.r.t. both \( x \) and \( t \) at \( t = v_\alpha(x) \).

### Assumption 3.5. Assume that there exists a family of measures \( G_m(\cdot) \) and a number \( \eta > 0 \) such that for all \( t \in B_\eta(v_\alpha(x)) \) and for all \( \Delta y \subset (-\infty, \infty) \),

\[
|\nu(\Delta y, t) - \nu^m(\Delta y, t)| \leq G_m(\Delta y) \quad \text{and} \quad \int \frac{|y|G_m(dy)}{G_m(dy)} \rightarrow 0 \text{ as } m \rightarrow \infty.
\]

### Assumption 3.6. \( \sup_y \mathbb{E}_{\mathcal{F}}[d(x; \xi)^2] < \infty \).

### Assumption 3.10. Hong and Liu (2009)

1. The VaR function \( v_\alpha(x) \) is differentiable for any \( x \in \mathcal{X} \).

2. For any \( x \in \mathcal{X} \), \( P[H(x; \theta) = v_\alpha(x)] = 0 \).
4 Discussion on the assumptions

In this paper, we present many technical assumptions that are needed for the results to hold. Here, we present a non-technical interpretation of each assumption to make it more intuitive to understand. Ignoring some pathological cases, we believe that most, if not all, assumptions are satisfied when \( h(\cdot, \xi) \) is Lipschitz continuous for almost all \( \xi \), the \( \xi \) and \( \theta \) are continuous random variables with light tails, and the variables \( x, \xi, \theta \) are confined to compact (bounded) spaces.

- **Assumption 3.2:** The random function \( h(x, \xi) \) is smooth and has light tails. For any given \( \theta \), the error function is a continuous random variable.

- **Assumption 3.3:** The function \( h(x, \xi) \) is Lipschitz continuous for a given \( \xi \), and is differentiable almost everywhere.

- **Assumption 3.4:** In a neighborhood of its \( \alpha \) quantile, the expected performance \( H(x; \theta) \) is a continuous random variable (induced by \( \theta \)), and its distribution function is smooth in \( x \).

- **Assumption 3.5:** This is a rather technical assumption that results from conditioning on measure zero events. It regulates the asymptotic behavior of the gradient observations, and requires the error of the gradient observations to converge to zero as the error of the function observation converges to zero. It is shown below that the assumption is satisfied for a general class of problems.

- **Assumption 3.6:** The gradient observations \( d(x, \xi) \) have a finite second moment.

- **Assumption 3.10:** When we are interested in CVaR rather than VaR, this assumption weakens Assumption 3.4. It requires the VaR function to be differentiable, and the expected performance \( H(x; \theta) \) to be a continuous random variable in a neighborhood of VaR.

4.1 A simple example

The following example is modified from Hong (2009). We proceed to show that the assumptions are satisfied, starting with the Assumption 3.5.

**Example 1.** Let \( \theta \sim N(0, 1) \), and \( h(x, \xi(\theta)) = x\theta_1 + \theta_2 + x\xi(\theta) \) where \( \xi(\theta) \sim N(0, \theta_1^2) \). Then, \( H(x; \theta) = x\theta_1 + \theta_2 \). It follows that \( D(x; \theta) = \theta_1, d(x, \xi(\theta)) = \theta_1 + \xi(\theta), H^m(x; \theta) = x\theta_1 + \theta_2 + x \frac{1}{m} \sum_{j=1}^m \xi_j(\theta) \) and \( \hat{D}^m(x; \theta) = \theta_1 + \frac{1}{m} \sum_{j=1}^m \xi_j(\theta) \).
To show that Assumption 3.5 holds, we have

\[
\nu((\infty, y]; t) = P(\theta_1 \leq y \mid x\theta_1 + \theta_2 = t) \\
= P\left(\theta_1 \leq y \mid \theta_1 = \frac{t - \theta_2}{x}\right) \\
= P(t - xy \leq \theta_2) \\
= 1 - \Phi(t - xy),
\]

where \(\Phi(\cdot)\) is the CDF of \(N(0, 1)\). For the noisy counterpart,

\[
\hat{\nu}^m((\infty, y]; t) = P\left(\theta_1 + \frac{1}{m}\sum_{j=1}^{m}\xi_j(\theta) \leq y \mid x\theta_1 + \theta_2 + x\frac{1}{m}\sum_{j=1}^{m}\xi_j(\theta) = t\right) \\
= P\left(\theta_1 + \frac{1}{m}\sum_{j=1}^{m}\xi_j(\theta) \leq y \mid \theta_1 + \frac{1}{m}\sum_{j=1}^{m}\xi_j(\theta) = \frac{t - \theta_2}{x}\right) \\
= P(t - xy \leq \theta_2) \\
= 1 - \Phi(t - xy).
\]

Thus, for this example, we have \(\nu((\infty, y]; t) = \hat{\nu}^m((\infty]; t)\), and Assumption 3.5 is satisfied with \(G_m(y) = 0\). We now verify the remaining assumptions.

- **Assumption 3.2**: For the first part, \(\tau^2 = x^2\theta_1^2 + (x\theta_1 + \theta_2)^2 < \infty\), and \(\tau^2 = 2x^2 + 1 < \infty\). For the remaining parts, see the discussion in Gordy and Juneja (2010) where it is implied that the assumption holds when the distribution of \(h(x, \xi)\) is Gaussian.

- **Assumption 3.3**: The first part holds with \(K(\xi(\theta)) = \theta_1 + \xi(\theta)\). For the second part, the derivative exists everywhere and is given by \(d(x, \xi(\theta)) = \theta_1 + \xi(\theta)\).

- **Assumption 3.4**: \(H(x; \theta)\) has a continuous density everywhere. Similarly \(\partial_x F(t; x)\) exists and is continuous everywhere.

- **Assumption 3.6**: \(\mathbb{E}_\theta[d(x, \xi)^2] = 2\theta_1^2\). The supremum here is not finite, so the assumption is technically violated. However, this violation is purely technical and happens only due to unbounded domain of the normal distribution. We can get around this by truncating the domain \(\Theta\) to some large interval, which is always done in practice due to limitations of machine precision.

- **Assumption 3.10**: The VaR function is given by \(v_\alpha(x) = z_\alpha\sqrt{x^2 + 1}\) where \(z_\alpha\) is the \(\alpha\) quantile of \(N(0, 1)\). It is differentiable for any \(x \in X\). The second part holds as \(H(x; \theta)\) is a continuous random variable.
4.2 A general class of functions

Here, we consider a class of functions of the form \( h(x; \xi(\theta)) = H(x; \theta) + \xi(\theta) \) with \( \xi(\theta) \) a mean zero finite variance random variable, and show that Assumption 3.5 holds for such problems. We will extend this to a more general class of functions below.

\[
\nu((\infty, y]; t) = P(D(x; \theta) \leq y \mid H(x; \theta) = t) \\
\nu^m((\infty, y]; t) = P(D(x; \theta) \leq y \mid H(x; \theta) + \xi(\theta) = t)
\]

If \( H(x; \theta) \) is invertible with a continuous inverse function and \( D(x; \theta) \) continuous in \( \theta \), then the result follows as \( \xi(\theta) \) converges to 0, and the convergence rate is \( O(m^{-1/2}) \).

For the case where the solution to \( H(x; \theta) = t \) is not unique, we define \( \vartheta(t) := \{ \theta : H(x; \theta) = t \} \) as the restricted random variable that satisfies \( H(x; \vartheta(t)) = t \). We require that \( \vartheta(t) \) is sample path continuous, which follows if \( H(x; \theta) \) is strictly monotone, i.e. has a non-zero gradient, in \( \theta \) in a neighborhood of \( t \). Then,

\[
P(D(x; \theta) \leq y \mid H(x; \theta) + \bar{\xi}(\theta) = t) = \mathbb{E}_\theta[P(D(x; \vartheta(t) - \bar{\xi}(\theta)) \leq y)] \\
\xrightarrow{m \to \infty} \mathbb{E}_\theta[P(D(x; \vartheta(t)) \leq y)] \\
= P(D(x; \theta) \leq y \mid H(x; \theta) = t).
\]

Note that \( \bar{\xi}(\theta) \xrightarrow{m \to \infty} 0 \) uniformly in \( \theta \) by Assumption 3.2. Thus, if \( D(x; \theta) \) is Lipschitz continuous in \( \theta \), the convergence is uniform in \( y \), and the assumption follows.

4.2.1 Extensions

A simple extension is to functions of the form \( h(x; \xi(\theta)) = H(x; \theta) + x\xi(\theta) \). The same analysis here holds, just with some extra terms which again disappear as \( m \to \infty \). In this case, we have

\[
P(D(x; \theta) + \bar{\xi}(\theta) \leq y \mid H(x; \theta) + x\bar{\xi}(\theta) = t) = \mathbb{E}_\theta[P(D(x; \vartheta(t) - x\bar{\xi}(\theta)) + \bar{\xi}(\theta) \leq y)] \\
\xrightarrow{m \to \infty} \mathbb{E}_\theta[P(D(x; \vartheta(t)) \leq y)] \\
= P(D(x; \theta) \leq y \mid H(x; \theta) = t).
\]

This again follows by uniform convergence of \( \bar{\xi}(\theta) \), and the Lipschitz continuity assumption on \( D(x; \theta) \).

This line of argument extends to the functions of the form \( h(x; \xi(\theta)) = H(x; \theta) + g(x)\xi(\theta) \) where \( g(\cdot) \) is a differentiable function of \( x \). Since the domain \( X \) is compact, the convergence is still uniform and the assumption holds.

We note that the analysis presented here implicitly assumes that \( \mathbb{E}_\theta[\xi] = 0 \), which is not restrictive.
as we can always replace $\xi$ with $\xi - \mathbb{E}_{\theta} [\xi]$ to obtain a mean zero random variable. The limiting factor in this analysis is that we only considered error terms in which the degree of $\xi$ is one. When the error term includes higher order terms of $\xi$, the uniform convergence is no longer implied by Assumption 3.2, and this line of argument fails to hold. One can still check for uniform convergence, and use the same argument if it holds, or verify the assumption in some other way. We conclude this analysis by noting that when the error term is a first order function of $\xi$, the integral in Assumption 3.5 is $O(m^{-1/2})$.

5 Details of numerical experiments

In this section, we fill out the details of the numerical experiments that were left out from the paper due to space constraints.

5.1 Details of the quadratic example

In this subsection, we fill in the details of the quadratic example presented in Section 4.1 in the paper. We start by verifying the assumptions, continue with problem setup, and obtain an analytical solution to the BRO optimization problem.

This example is modified from the simple example by Hong (2009) presented above. Assuming a posterior distribution of the form $\theta_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, the assumptions can be verified in a similar way. Here, we briefly highlight the differences.

- Assumption 3.2: $\tau_2^2 = x^2\theta_1^2 + (x\theta_1 + x^2\theta_2)^2 < \infty$ and $\tau^2 = 2x^2(\sigma_1^2 + \mu_1^2) + 2x^3\mu_1\mu_2 + x^4(\sigma_2^2 + \mu_2^2) < \infty$.

- Assumption 3.3: We have $K(\xi(\theta)) = \sup_{x \in X} \theta_1 + 2x\theta_2 + \xi(\theta)$ which is integrable since $X$ is compact. The derivative is given above and exists everywhere.

- Assumption 3.4: Same as before.

- Assumption 3.5: The assumption is satisfied with $G_m(\cdot) = 0$ as shown below.

\[
\nu((\infty, y]; t) = P(D(x; \theta) \leq y \mid H(x; \theta) = t) \\
= P(\theta_1 + 2x\theta_2 \leq y \mid x\theta_1 + x^2\theta_2 = t) \\
= P(\theta_1 + 2x\theta_2 \leq y \mid \theta_1 = \frac{t - x^2\theta_2}{x}) \\
= P(t + x^2\theta_2 \leq xy) \\
= P(\theta_2 \leq \frac{xy - t}{x^2})
\]
\[ \hat{\nu}^m((-\infty, y]; t) = P(\hat{D}^m(x; \theta) \leq y \mid \hat{H}^m(x; \theta) = t) \]
\[ = P(\theta_1 + 2x\theta_2 + \bar{\xi}(\theta) \leq y \mid x\theta_1 + x^2\theta_2 + x\bar{\xi}(\theta) = t) \]
\[ = P \left( \theta_1 + 2x\theta_2 + \bar{\xi}(\theta) \leq y \mid \theta_1 + \bar{\xi}(\theta) = \frac{t - x^2\theta_2}{x} \right) \]
\[ = P \left( \theta_2 \leq \frac{xy - t}{x^2} \right) \]

Thus, \[ |\nu((-\infty, y]; t) - \hat{\nu}^m((-\infty, y]; t)| = 0, \text{ and } G_m(\cdot) = 0 \] satisfies the assumption.

- Assumption 3.6: We again run into same technical violation, which is resolved by bounding the domain \( \Theta \).

- Assumption 3.10: The expression for the VaR is given below and is continuously differentiable almost everywhere. The second part holds as \( H(x; \theta) \) is a continuous random variable.

With the given posterior of the form \( \theta_1 \sim N(\mu_1, \sigma_1^2) \), \( H(x; \theta) \) is a normal random variable, and the VaR and CVaR objectives are given by

\[ v_\alpha(x) = \mu_H + z_\alpha \sigma_H \text{ and } c_\alpha(x) = \mu_H + \frac{\phi(z_\alpha)}{1 - \alpha} \sigma_H, \]

where \( z_\alpha \) and \( \phi(\cdot) \) are the \( \alpha \) quantile and the PDF of the standard normal distribution respectively, and \( \mu_H, \sigma_H \) are the mean and standard deviation of \( H(x; \theta) \) given by

\[ \mu_H = x\mu_1 + x^2\mu_2 \text{ and } \sigma_H = \sqrt{x^2\sigma_1^2 + x^4\sigma_2^2}. \]

The gradients of VaR and CVaR can be computed in a similar manner, and are given by \( v_\alpha'(x) = \mu_H' + z_\alpha \sigma_H' \) and \( c_\alpha'(x) = \mu_H' + \frac{\phi(z_\alpha)}{1 - \alpha} \sigma_H' \), where \( \mu_H' \) and \( \sigma_H' \) denote the gradients of \( \mu_H \) and \( \sigma_H \) respectively.

For a given \( \theta = (\theta_1, \theta_2) \) with \( \theta_2 > 0 \), the minimizer is given by \( x^*(\theta) = -\frac{\theta_1}{2\theta_2} \). Once \((\mu_1, \mu_2, \sigma_1, \sigma_2)\) are specified, the minimizers of BRO objectives can also be computed analytically.

We consider the case where \( \theta_1, \theta_2 \) are estimated from samples \( \{\xi_i^j\}_{i=1,...,N} \overset{iid}{\sim} N(\bar{\theta}_i, \tilde{\sigma}_i^2) \) where \( \tilde{\sigma}_i \) is the known variance. Then, using a degenerate normal prior on \( \theta \), the posterior distribution is given by \( \theta_i \sim N(\bar{\theta}_i, \tilde{\sigma}_i^2/\scriptsize N) \). In this numerical example, we suppose that the posterior is given by \( \theta_1 \sim N(-15, 16) \) and \( \theta_2 \sim N(10, 4) \), and consider the BRO objective with risk measure CVaR at risk level \( \alpha = 0.75 \).

With the given parameters, the optimal solution is found at \( x^* = 0.474775 \) with the corresponding BRO CVaR objective value of \(-2.38647 \).

To keep things simple, we use a fixed budget sequence of \( n_t = n = 100 \) and \( m_t = m = n/5 \). The gradient-based algorithms use the estimators developed in this paper, and the gradient-free alternatives use the nested estimators of Zhu et al. (2020) to estimate the objective value. The benchmark algorithms we consider are originally developed for deterministic optimization. Thus, in addition to stochastic eval-
uations of the objective, we also consider the Sample Average Approximation (SAA, Kim, Pasupathy, & Henderson, 2015) which converts the stochastic optimization problem into an approximate deterministic optimization problem. This is done by fixing a random draw of $\theta_i$ and $\xi_j(\theta_i)$ before the optimization starts, and using this fixed set of samples to calculate the value of the estimators. For comparison, the stochastic evaluations draw a new set of random variables $\theta_i$ and $\xi_j(\theta_i)$ for each evaluation of the estimators. A different set of $\theta_i$ and $\xi_j(\theta_i)$ is used for each replication of SAA.

5.2 Convergence analysis of MCMC output

It is known that MCMC methods converge to a steady state distribution, however, detecting when this convergence occurs, without access to the distribution, is an open question. In this paper, we use an improvised Wasserstein distance analysis to perform an empirical convergence analysis.

The idea behind this analysis is as follows. We treat subsets of the MCMC chain as empirical distributions drawn from the chain’s distribution at a given time. The subsets should be sufficiently large so that the correlation between the samples can be ignored. If the chain has not converged to the steady state distribution, the underlying distribution is actively changing and this should show up as a distance between the two empirical distributions. As the convergence occurs, we can expect the distance between subsequent empirical distributions to get smaller, and stabilize as we converge to the stationary distribution. Note that, even when the empirical distributions are drawn from the stationary distribution, the distance between the empirical distributions will be non-zero as these are essentially two random draws from the same underlying distribution.

However, once we have reached the stationary distribution, the distance between any two arbitrary subsets, not necessarily subsequent, should be roughly the same, apart from the random noise. We observe this behavior in the empirical posterior distributions used in the experiments. The Wasserstein distance between subsets of size $10^5$ are observed to be about $10^{-3}$ ($\pm$ noise) regardless of the ordering of the subsets. Thus, we conclude that the underlying distribution is not actively changing and the empirical distribution can be treated as coming from the steady state, i.e. the true posterior, distribution.

6 Results and proofs

Since the results are given for a fixed $x$, we drop the dependence on $x$ to simplify notation in the proofs. Moreover, $\hat{H}_i^m$ and $H_i$ are commonly used in place of $\hat{H}_i^m(x;\theta_i)$ and $H(x;\theta_i)$ respectively. We do the same with $\hat{D}_i^m$ and $D_i$ as well.

**Proposition 3.7.** Suppose that Assumptions 3.2, 3.3, 3.4, 3.5 and 3.6 hold. Then $E_{\theta,P_{\theta}}[\varphi_{\alpha}^{n,m}(x)] - v_{\alpha}'(x) \rightarrow 0$ as $n, m \rightarrow \infty$.

Moreover, if in addition the integral in Assumption 3.5 is $O(m^{-1/2})$, $g(t;x)$ is differentiable w.r.t.
Proof. We will show that \( \lim_{t \to \infty} \Theta(t) = \Theta(\alpha) \) for more details. where \( \Theta(\alpha) \), and let \( \hat{\theta}(\alpha) \). Next, we show that \( \hat{\theta}(\alpha) \) follows from the fact that \( \hat{\theta}(\alpha) \) is the same as conditioning on \( \hat{\theta}(\alpha) \) is the same as conditioning on

1. there exists some \( i^* \) such that \( \hat{\theta}(\alpha) = t \);
2. for \( i \neq i^* \), there are \( \alpha n - 1 \) values of \( \hat{\theta}(\alpha) \) which are \( < t \), and the rest are \( > t \).

Moreover, due to independence and symmetry, \( (\dagger) \) is the same as (without loss of generality) \( \hat{\theta}(\alpha) \) is the same as conditioning on \( \{\hat{\theta}(\alpha) = t \} \) is the same as conditioning on

\[
\mathbb{E}_{\theta, P_\alpha}[\varphi_{\alpha n}^{n,m}] = \mathbb{E}_{\theta, P_\alpha}[\hat{D}^m(\hat{\varphi}_{\alpha n}^{n,m})] \\
= \int \mathbb{E}_{\theta, P_\alpha}[\hat{D}^m(\hat{\varphi}_{\alpha n}^{n,m}) \mid \hat{H}^m(\hat{\varphi}_{\alpha n}^{n,m}) = t] dF_\nu(t) \\
= \int \mathbb{E}_{\theta, P_\alpha}[\hat{D}^m(\theta) \mid \hat{H}^m(\theta) = t] dF_\nu(t) \\
= \int \hat{g}^m(t; x) dF_\nu(t) \\
= \mathbb{E}_{\theta, P_\alpha}[\hat{g}^m(\hat{\varphi}_{\alpha n}^{n,m}; x)],
\]

where \( (\dagger\dagger) \) follows from the fact that \( \theta_1, \theta_2, \ldots, \theta_n \) are i.i.d., and thus \( \{\hat{H}^m(\theta_i)\}_{i=1}^n \) are i.i.d.; conditioning on \( \{\hat{H}^m(\theta_i)\} = t \) is the same as conditioning on

\[
\frac{g(t; x)}{g(t; x)} \to g(x; x) \text{ uniformly on } B_\eta(v_n) \text{ as } m \to \infty. \text{ Note that we can write,}
\]

\[
g(t; x) = \mathbb{E}_\theta[\partial_x H(x; \theta) \mid H(x; \theta) = t] = \int_R y \nu(dy, t);
\]

and similarly,

\[
\hat{g}^m(t; x) = \int_R y \hat{\nu}^m(dy, t).
\]

Then, for all \( t \in B_\eta(v_n) \),

\[
|\hat{g}^m(t; x) - g(t; x)| = \left| \int_R y (\hat{g}^m(dy, t) - \nu(dy, t)) \right| \\
\leq \int_R |y| |\hat{g}^m(dy, t) - \nu(dy, t)| \\
\leq \int_R |y| G_m(dy) \to 0 \quad \text{as } m \to \infty,
\]

where the first inequality follows from the triangle inequality and the definition of Lebesgue integral, and the last inequality follows from Assumption 3.5. Since \( G_m(\cdot) \) does not depend on \( t \), the convergence is
uniform in \( t \). We now claim that \( \hat{g}^m(\hat{v}^{n,m}_\alpha; x) \to g(v_\alpha; x) \) a.s. as \( n, m \to \infty \). To see this, take a sample path on which \( \hat{v}_\alpha^{n,m} \to v_\alpha \) as \( n, m \to \infty \), which occurs a.s. due to Assumption 3.2. Therefore, it suffices to show convergence on this sample path. Take any \( \epsilon > 0 \), and notice the following.

- Since \( g(\cdot; x) \) is continuous by Assumption 3.4, there exists \( \delta_1 > 0 \) such that
  \[
  |g(t; x) - g(v_\alpha; x)| < \epsilon/2, \quad \forall t \in B_{\delta_1}(v_\alpha).
  \]

- Take \( \delta_2 := \min\{\delta, \eta\} \). Then, there exists \( N_1, M_1 \in \mathbb{Z}^+ \) such that
  \[
  |\hat{v}_\alpha^{n,m} - v_\alpha| < \delta_2, \quad \forall n \geq N_1, m \geq M_1.
  \]

- Furthermore, due to uniform convergence, there exists \( M_2 \in \mathbb{Z}^+ \) such that \( \forall m \geq M_2, \)
  \[
  |\hat{g}^m(t; x) - g(t; x)| < \epsilon/2, \quad \forall t \in B_{\delta_2}(v_\alpha).
  \]

Combining the above, we get

\[
|\hat{g}^m(\hat{v}_\alpha^{n,m}; x) - g(v_\alpha; x)|
\leq |\hat{g}^m(\hat{v}_\alpha^{n,m}; x) - g(\hat{v}_\alpha^{n,m}; x)| + |g(\hat{v}_\alpha^{n,m}; x) - g(v_\alpha; x)|
\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,
\]

for all \( n \geq N_1, m \geq \max\{M_1, M_2\} \). Thus,

\[
\hat{g}^m(\hat{v}_\alpha^{n,m}; x) \to g(v_\alpha; x) \quad \text{a.s. as } n, m \to \infty.
\] (2)

Finally, since

\[
\mathbb{E}_{\theta,\mathbb{P}}[\hat{g}^m(\hat{v}_\alpha^{n,m}; x)^2] = \int \mathbb{E}_{\theta,\mathbb{P}}[\partial_x \hat{H}^m(x; \theta) | \hat{H}^m(x; \theta) = t]dF_x(t)
\leq \int \mathbb{E}_{\theta,\mathbb{P}} \left\{ \left[ \partial_x \hat{H}^m(x; \theta) \right]^2 | \hat{H}^m(x; \theta) = t \right\} dF_x(t) \quad \text{(Jensen's inequality)}
= \mathbb{E}_{\theta,\mathbb{P}} \left\{ \left[ \partial_x \hat{H}^m(x; \theta_{\lfloor [\alpha n] \rfloor}) \right]^2 \right\}
\leq \sup_{\theta} \mathbb{E}_{\theta,\mathbb{P}}[d(x; \xi)^2] < \infty, \quad \text{(Similar to (††))}
\]

we get that \( \hat{g}^m(\hat{v}_\alpha^{n,m}; x) \) is uniformly integrable. This together with (2) yields that

\[
\lim_{n,m \to \infty} \mathbb{E}_{\theta,\mathbb{P}}[\varphi^{n,m}_\alpha] = g(v_\alpha; x) = v'_\alpha.
\]
This completes the first part of the proof. For the second part of the proposition, given that the integral is $O(m^{-1/2})$, we have that $\hat{g}^m(t; x) - g(t; x) = O(m^{-1/2})$ uniformly for all $t \in B_\eta(v_\alpha)$. Then, for $n, m$ large enough (so that $\hat{v}^{n,m}_\alpha \in B_\eta(v_\alpha)$),

$$\left| \hat{g}^m(\hat{v}^{n,m}_\alpha; x) - g(v_\alpha; x) \right| \leq \left| g(\hat{v}^{n,m}_\alpha; x) - g(v_\alpha; x) \right| + O(m^{-1/2}).$$

For any fixed $\hat{v}^{n,m}_\alpha$, using Taylor’s theorem,

$$\left| g(\hat{v}^{n,m}_\alpha; x) - g(v_\alpha; x) \right| = \left| \partial_t g(v_\alpha; x)(\hat{v}^{n,m}_\alpha - v_\alpha) + o(\hat{v}^{n,m}_\alpha - v_\alpha) \right|$$

Putting it all together, we have

$$\mathbb{E}_{\theta, \phi_x}[\varphi^{n,m}_\alpha] - v'_\alpha = \mathbb{E}_{\theta, \phi_x}[\hat{g}^m(\hat{v}^{n,m}_\alpha; x)] - g(v_\alpha; x)$$

$$\leq \mathbb{E}_{\theta, \phi_x}[|g(\hat{v}^{n,m}_\alpha; x) - g(v_\alpha; x)||] + O(m^{-1/2})$$

$$= \mathbb{E}_{\hat{v}^{n,m}_\alpha}[|\partial_t g(v_\alpha; x)(\hat{v}^{n,m}_\alpha - v_\alpha) + o(\hat{v}^{n,m}_\alpha - v_\alpha)||] + O(m^{-1/2}).$$

Theorem 3.6 of Zhu et al. (2020) shows that under Assumption 3.2, $n = o(m^2)$ is a sufficient and necessary condition for

$$\sqrt{n}(\hat{v}^{n,m}_\alpha - v_\alpha) \Rightarrow \mathcal{N}(0, \sigma_v)$$

where $\sigma_v = \frac{\alpha(1-\alpha)}{f_\eta}$. Therefore, the term inside the absolute value (when scaled by $\sqrt{n}$) converges to a mean zero normal random variable, and the expectation is $O(n^{-1/2})$. Putting it together with $n = \Theta(m)$, we get that the bias is $O(n^{-1/2})$.

**Theorem 3.8.** Suppose that Assumptions 3.2, 3.3, 3.4, 3.5 and 3.6 hold, then

$$\varphi^{n,m,k}_\alpha(x) \xrightarrow{P} v'_\alpha(x) \text{ as } n, m, k \to \infty,$$

where $\xrightarrow{P}$ denotes convergence in probability.

**Proof.** By Proposition 3.7, we have $\mathbb{E}_{\theta, \phi_x}[\varphi^{n,m}_\alpha] \to v'_\alpha$ as $n, m \to \infty$. For any $\epsilon > 0$, by Chebyshev’s inequality,

$$P \left( |\varphi^{n,m,k}_\alpha - \mathbb{E}_{\theta, \phi_x}[\varphi^{n,m}_\alpha]| > \epsilon \right) \leq \frac{\text{Var}(\varphi^{n,m}_\alpha)}{k\epsilon^2} \leq \frac{\text{Var}_{\phi_x}[\varphi^{n,m}_\alpha]}{k\epsilon^2} \leq \frac{\sup_{\theta} \mathbb{E}_{\phi_x}[|d(\xi)|^2]}{k\epsilon^2}.$$ 

Therefore, $\varphi^{n,m,k}_\alpha \xrightarrow{P} \mathbb{E}[\varphi^{n,m}_\alpha]$ as $k \to \infty$ uniformly for each $n, m$. Combining with the result of Proposition 3.7, $\varphi^{n,m,k}_\alpha \xrightarrow{P} v'_\alpha$ as $n, m, k \to \infty$. 

Proposition 3.11. Suppose Assumption 3.2 holds and \( P(H(x; \theta) = v_\alpha(x)) = 0 \). Then

\[
\frac{1}{n} \sum_{i=1}^{n} \left| I_{(H_i \geq \hat{v}_\alpha^m)} - I_{(H_i \geq v_\alpha)} \right| \to 0 \text{ w.p.1 as } n, m \to \infty.
\]

Proof.

\[
\frac{1}{n} \sum_{i=1}^{n} \left| I_{(H_i \geq \hat{v}_\alpha^m)} - I_{(H_i \geq v_\alpha)} \right| = \frac{1}{n} \sum_{i=1}^{n} \left| I_{(H_i \geq \hat{v}_\alpha^m)} - I_{(H_i \geq v_\alpha)} \right| - \frac{1}{n} \sum_{i=1}^{n} \left| I_{(H_i \geq v_\alpha)} - I_{(H_i \geq \hat{v}_\alpha^m)} \right|
\]

\[
\leq \frac{1}{n} \sum_{i=1}^{n} \left| I_{(H_i \geq \hat{v}_\alpha^m)} - I_{(H_i \geq \hat{v}_\alpha^m)} \right| + \frac{1}{n} \sum_{i=1}^{n} \left| I_{(H_i \geq \hat{v}_\alpha^m)} - I_{(H_i \geq v_\alpha)} \right|
\]

We will now proceed to bound each term individually and show that the sum is bounded above by zero in the limit.

\[
(P2) = \left( I_{(\hat{v}_\alpha^m \leq v_\alpha)} - I_{(\hat{v}_\alpha^m > v_\alpha)} \right) \frac{1}{n} \sum_{i=1}^{n} \left( I_{(H_i \geq \hat{v}_\alpha^m)} - I_{(H_i \geq v_\alpha)} \right)
\]

\[
= I_{(\hat{v}_\alpha^m \leq v_\alpha)} \frac{1}{n} \sum_{i=1}^{n} \left( I_{(H_i \geq \hat{v}_\alpha^m)} - I_{(H_i \geq \hat{v}_\alpha^m)} \right) + I_{(\hat{v}_\alpha^m > v_\alpha)} \frac{1}{n} \sum_{i=1}^{n} \left( I_{(H_i \geq \hat{v}_\alpha^m)} - I_{(H_i \geq v_\alpha)} \right)
\]

Note that both \((P3)\) and \((P4)\) are non-negative as long as the accompanying indicator is one.

\[
(P3) = I_{(\hat{v}_\alpha^m \leq v_\alpha - \epsilon)} (P3) + I_{(\hat{v}_\alpha^m \geq v_\alpha - \epsilon)} (P3)
\]

\[
\leq I_{(\hat{v}_\alpha^m \leq v_\alpha - \epsilon)} + I_{(\hat{v}_\alpha^m \geq v_\alpha - \epsilon)} \frac{1}{n} \sum_{i=1}^{n} \left( I_{(H_i \geq v_\alpha - \epsilon)} - I_{(H_i \geq v_\alpha)} \right)
\]

\[
\leq I_{(\hat{v}_\alpha^m \leq v_\alpha - \epsilon)} + \frac{1}{n} \sum_{i=1}^{n} I_{(v_\alpha > H_i \geq v_\alpha - \epsilon)}
\]

Similarly,

\[
(P4) = I_{(\hat{v}_\alpha^m \leq v_\alpha + \epsilon)} (P4) + I_{(\hat{v}_\alpha^m > v_\alpha + \epsilon)} (P4)
\]

\[
\leq I_{(\hat{v}_\alpha^m > v_\alpha + \epsilon)} + I_{(\hat{v}_\alpha^m \leq v_\alpha + \epsilon)} \frac{1}{n} \sum_{i=1}^{n} \left( I_{(H_i \geq v_\alpha + \epsilon)} - I_{(H_i \geq v_\alpha)} \right)
\]

\[
\leq I_{(\hat{v}_\alpha^m > v_\alpha + \epsilon)} + \frac{1}{n} \sum_{i=1}^{n} I_{(v_\alpha + \epsilon > H_i \geq v_\alpha)}
\]
Putting them together, we get

\[(P2) \leq \mathbb{I}(\hat{c}_{\alpha}^{n,m} < v_0 - \epsilon) + \mathbb{I}(\hat{c}_{\alpha}^{n,m} > v_0 + \epsilon) + \sum_{i=1}^{n} \left( \mathbb{I}(v_0 > H_i \geq v_0 - \epsilon) + \mathbb{I}(v_0 + \epsilon > H_i \geq v_0) \right) \]

\[= (1 - \mathbb{I}(v_0 + \epsilon \geq \hat{c}_{\alpha}^{n,m} \geq v_0 - \epsilon)) + \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(v_0 + \epsilon < H_i \geq v_0 - \epsilon) \]

We can now look at the other term.

\[(P1) = \frac{1}{n} \sum_{i=1}^{n} \left| \mathbb{I}(\hat{H}_{\alpha}^{n,m} \geq \hat{c}_{\alpha}^{n,m}) - \mathbb{I}(H_i \geq \hat{c}_{\alpha}^{n,m}) \right| \]

\[= \frac{1}{n} \sum_{i=1}^{n} \left( \mathbb{I}(\hat{H}_{\alpha}^{n,m} \geq \hat{c}_{\alpha}^{n,m} > H_i) + \mathbb{I}(H_i \geq \hat{c}_{\alpha}^{n,m} > \hat{H}_{\alpha}^{n,m}) \right) \]

\[= \frac{1}{n} \sum_{i=1}^{n} \left[ \mathbb{I}(\hat{H}_{\alpha}^{n,m} \leq H_i + \epsilon) \mathbb{I}(\hat{H}_{\alpha}^{n,m} \geq \hat{c}_{\alpha}^{n,m} > H_i) + \mathbb{I}(\hat{H}_{\alpha}^{n,m} > H_i + \epsilon) \mathbb{I}(\hat{H}_{\alpha}^{n,m} \geq \hat{c}_{\alpha}^{n,m} > H_i) \right. \]

\[+ \left. \mathbb{I}(\hat{H}_{\alpha}^{n,m} \geq H_i - \epsilon) \mathbb{I}(H_i \geq \hat{c}_{\alpha}^{n,m} > \hat{H}_{\alpha}^{n,m}) + \mathbb{I}(\hat{H}_{\alpha}^{n,m} < H_i - \epsilon) \mathbb{I}(H_i \geq \hat{c}_{\alpha}^{n,m} > \hat{H}_{\alpha}^{n,m}) \right] \]

\[\leq \frac{1}{n} \sum_{i=1}^{n} \left[ \mathbb{I}(\hat{H}_{\alpha}^{n,m} \leq H_i + \epsilon) \mathbb{I}(H_i \geq \hat{c}_{\alpha}^{n,m} > H_i) + \mathbb{I}(\hat{H}_{\alpha}^{n,m} > H_i - \epsilon) \mathbb{I}(H_i \geq \hat{c}_{\alpha}^{n,m} > H_i - \epsilon) \right] \]

\[+ \frac{1}{n} \sum_{i=1}^{n} \left( \mathbb{I}(\hat{H}_{\alpha}^{n,m} > H_i + \epsilon) + \mathbb{I}(\hat{H}_{\alpha}^{n,m} < H_i - \epsilon) \right) \]

\[(P5) \leq \frac{1}{n} \sum_{i=1}^{n} \left( \mathbb{I}(H_i + \epsilon \geq \hat{c}_{\alpha}^{n,m} > H_i) + \mathbb{I}(H_i \geq \hat{c}_{\alpha}^{n,m} > H_i - \epsilon) \right) \]

\[= \frac{1}{n} \sum_{i=1}^{n} \left( \mathbb{I}(H_i + \epsilon \geq \hat{c}_{\alpha}^{n,m} > H_i - \epsilon) \right) = \frac{1}{n} \sum_{i=1}^{n} \left( \mathbb{I}(H_i + \epsilon \geq \hat{c}_{\alpha}^{n,m} > H_i) \mathbb{I}(\hat{c}_{\alpha}^{n,m} > H_i - \epsilon) \right) \]

\[= \frac{1}{n} \sum_{i=1}^{n} \left[ \mathbb{I}(v_0 + \epsilon \geq \hat{c}_{\alpha}^{n,m} > v_0 - \epsilon) \mathbb{I}(H_i + \epsilon \geq \hat{c}_{\alpha}^{n,m} > H_i) \mathbb{I}(\hat{c}_{\alpha}^{n,m} > H_i - \epsilon) \right. \]

\[+ \left. (1 - \mathbb{I}(v_0 + \epsilon \geq \hat{c}_{\alpha}^{n,m} > v_0 - \epsilon)) \mathbb{I}(H_i + \epsilon \geq \hat{c}_{\alpha}^{n,m} > H_i) \mathbb{I}(\hat{c}_{\alpha}^{n,m} > H_i - \epsilon) \right] \]

\[\leq \frac{1}{n} \sum_{i=1}^{n} \left( \mathbb{I}(v_0 + \epsilon \geq \hat{c}_{\alpha}^{n,m} > v_0 - \epsilon) \mathbb{I}(H_i + \epsilon \geq \hat{c}_{\alpha}^{n,m} > H_i - 2\epsilon) + (1 - \mathbb{I}(v_0 + \epsilon \geq \hat{c}_{\alpha}^{n,m} > v_0 - \epsilon)) \right) \]

\[\leq \frac{1}{n} \sum_{i=1}^{n} \left( \mathbb{I}(H_i + 2\epsilon \geq v_0 > H_i - 2\epsilon) + (1 - \mathbb{I}(v_0 + \epsilon \geq \hat{c}_{\alpha}^{n,m} > v_0 - \epsilon)) \right) \]

Putting them together,

\[(P1) \leq (1 - \mathbb{I}(v_0 + \epsilon \geq \hat{c}_{\alpha}^{n,m} > v_0 - \epsilon)) + \frac{1}{n} \sum_{i=1}^{n} \left( \mathbb{I}(H_i + 2\epsilon \geq v_0 > H_i - 2\epsilon) + \mathbb{I}(\hat{H}_{\alpha}^{n,m} > H_i + \epsilon) + \mathbb{I}(\hat{H}_{\alpha}^{n,m} < H_i - \epsilon) \right) \]

\[= (1 - \mathbb{I}(v_0 + \epsilon \geq \hat{c}_{\alpha}^{n,m} > v_0 - \epsilon)) + \frac{1}{n} \sum_{i=1}^{n} \left( \mathbb{I}(H_i + 2\epsilon \geq v_0 > H_i - 2\epsilon) + (1 - \mathbb{I}(H_i + \epsilon \geq \hat{H}_{\alpha}^{n,m} > H_i - \epsilon)) \right) \]

\[+ \frac{1}{n} \sum_{i=1}^{n} \left( \mathbb{I}(\hat{H}_{\alpha}^{n,m} > H_i + \epsilon) + \mathbb{I}(\hat{H}_{\alpha}^{n,m} < H_i - \epsilon) \right) \]
Thus, we have
\[
\frac{1}{n} \sum_{i=1}^{n} |\mathbb{I}(\hat{\epsilon}_{\alpha,m}^n \geq v_\alpha) - \mathbb{I}(H_i \geq v_\alpha)| \leq 2(1 - \mathbb{I}(v_\alpha + \epsilon \geq \hat{\epsilon}_{\alpha,m}^n \geq v_\alpha - \epsilon)) + \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(H_i + 2\epsilon \geq v_\alpha \geq H_i - 2\epsilon) + \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(H_i + \epsilon \geq H_i^m \geq H_i - \epsilon)
\]
By strong consistency of \(\hat{\epsilon}_{\alpha,m}^n\), the first term goes to 0 w.p.1. The second term is equivalent to \(P(H + 2\epsilon \geq v_\alpha > H - 2\epsilon)\) (w.p.1) in the limit which goes to 0 as \(\epsilon \to 0\) since \(P(H = v_\alpha) = 0\). The last term goes to 0 w.p.1 by a similar argument. Let’s focus on the third term. We can rewrite it as
\[
(*) = \frac{1}{n} \sum_{i=1}^{n} (1 - \mathbb{I}(H_i + \epsilon \geq H_i^m \geq H_i - \epsilon)) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}((\mathcal{E}(\theta_i)/\sqrt{m}) > \epsilon).
\]
Pick \(\delta > 0\).
\[
P\left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}((\mathcal{E}(\theta_i)/\sqrt{m}) > \epsilon) > \delta\right) \leq \frac{E_{\theta, P_\delta}[\mathbb{I}((\mathcal{E}(\theta_i)/\sqrt{m}) > \epsilon)]}{\delta}
\]
\[
= \frac{P((\mathcal{E}(\theta_i)/\sqrt{m}) > \epsilon)}{\delta}
\]
\[
\leq \frac{P((\mathcal{E}(\theta_i)/\sqrt{m}) > \epsilon)}{\delta}
\]
\[
\leq E_{\theta, P_\delta}[(\mathcal{E}(\theta_i)/\sqrt{m})^4]/\delta \epsilon^4 m^4
\]
By assumption 3.2, \(E_{\theta, P_\delta}[(\mathcal{E}(\theta_i)/\sqrt{m})^4] < \infty\). Then, by Borel-Cantelli lemma, \((*) \to 0\) w.p.1 as \(n, m \to \infty\).
We have that
\[
\limsup_{n, m \to \infty} \frac{1}{n} \sum_{i=1}^{n} |\mathbb{I}(\hat{\epsilon}_{\alpha,m}^n \geq v_\alpha) - \mathbb{I}(H_i \geq v_\alpha)| \leq 0.
\]
Noting that \(\frac{1}{n} \sum_{i=1}^{n} |\mathbb{I}(\hat{\epsilon}_{\alpha,m}^n \geq v_\alpha) - \mathbb{I}(H_i \geq v_\alpha)| \geq 0\) due to absolute value, the proof is complete. \(\square\)

**Proposition 3.12.** Under Assumptions 3.2, 3.3, 3.6, 3.10, the bias \(E_{\theta, P_\delta}[(\hat{\epsilon}_{\alpha,m}^n(x) - c'_\alpha(x)) \to 0\) as \(n, m \to \infty\). Moreover, if in addition \(n = \Theta(m)\), then the bias is of the order \(O(n^{-1/2})\).

**Proof.** Note that \(E_{\theta} \left[\frac{1}{(1-\alpha)n} \sum_{i=1}^{n} D_i \mathbb{I}(H_i \geq v_\alpha)\right] = c'_\alpha\). Therefore we will work with \(E_{\theta} \left[\frac{1}{(1-\alpha)n} \sum_{i=1}^{n} D_i \mathbb{I}(H_i \geq v_\alpha)\right]\) instead of \(c'_\alpha\).
\[
\mathbb{E}_{\theta, P_n}[\psi_{\alpha}^{n,m}] = \mathbb{E}_{\theta} \left[ \frac{1}{(1-\alpha)n} \sum_{i=1}^{n} D_i 1_{(H_i \geq \psi_{\alpha}^m)} \right] = \mathbb{E}_{\theta, P_n} \left[ \frac{1}{1-\alpha} \frac{1}{n} \sum_{i=1}^{n} D_i 1_{(H_i^m \geq \psi_{\alpha}^m)} \right]
\]

\[
= \frac{1}{1-\alpha} \mathbb{E}_{\theta, P_n} \left[ \frac{1}{n} \sum_{i=1}^{n} D_i 1_{(H_i^m \geq \psi_{\alpha}^m)} \right] = \mathbb{E}_{\theta, P_n} \left[ \frac{1}{n} \sum_{i=1}^{n} D_i 1_{(H_i^m \geq \psi_{\alpha}^m)} \right] + \mathbb{E}_{\theta, P_n} \left[ \frac{1}{n} \sum_{i=1}^{n} D_i (1_{(H_i^m \geq \psi_{\alpha}^m)} - 1_{(H_i \geq \psi_{\alpha}^m)}) \right]
\]

We will show that both parts go to 0 w.p.1.

\[
|P1| \leq \mathbb{E}_{\theta, P_n} \left[ \frac{1}{n} \sum_{i=1}^{n} |\hat{D}_i^m - D_i| 1_{(H_i^m \geq \psi_{\alpha}^m)} \right] \leq \mathbb{E}_{\theta, P_n} \left[ \frac{1}{n} \sum_{i=1}^{n} |\hat{D}_i^m - D_i| \right] = \mathbb{E}_{\theta, P_n} [|\hat{D}^m(\theta) - D(\theta)|]
\]

We have that for any \( \theta \), \( \mathbb{E}_{P_n}[|\hat{D}^m(\theta) - D(\theta)|] \to 0 \) w.p.1 by the strong consistency of \( \hat{D}^m \). We will show how this carries on to \( \mathbb{E}_{\theta, P_n}[|\hat{D}^m(\theta) - D(\theta)|] \) using the assumption \( \sup_{\theta} \mathbb{E}_{P_n}[d(x, \xi)^2] < \infty \).

\[
\mathbb{E}_{\theta, P_n}[|\hat{D}^m(\theta) - D(\theta)|] \leq \sup_{\theta} \mathbb{E}_{P_n}[|\hat{D}^m(\theta) - D(\theta)|] = \sup_{\theta} \mathbb{E}_{P_n} \left[ \frac{1}{m} \sum_{j=1}^{m} d(\xi_j) - D(\theta) \right] \leq \sup_{\theta} \sqrt{\mathbb{E}_{P_n} \left[ \left( \frac{1}{m} \sum_{j=1}^{m} d(\xi_j) - D(\theta) \right)^2 \right]} = \sup_{\theta} \sqrt{\frac{\text{Var}_{P_n}(d(\xi))}{m}}
\]

\[
\leq \frac{1}{\sqrt{m}} \sup_{\theta} \sqrt{\mathbb{E}_{P_n}[d(\xi)^2]} \to 0
\]

as \( m \to \infty \) since \( \sup_{\theta} \mathbb{E}_{P_n}[d(x, \xi)^2] < \infty \). Therefore \( P1 \) \( \to 0 \) as \( n, m \to \infty \).
\[ |(P2)| \leq \mathbb{E}_{\theta, P_\theta} \left[ \frac{1}{n} \sum_{i=1}^{n} |D_i| \left| \mathbb{1}_{(\hat{H}_m \geq \hat{v}_\alpha^m)} - \mathbb{1}_{(H_i \geq v_{\alpha})} \right| \right] \\
= \sup_{\theta} |D(\theta)| \mathbb{E}_{\theta, P_\theta} \left[ \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(\hat{H}_m \geq \hat{v}_\alpha^m)} - \mathbb{1}_{(H_i \geq v_{\alpha})} \right] \\
\quad \text{(P3)}
\]

Note that the term inside the expectation is bounded by 1. Then, by Proposition 3.11 and Dominated Convergence Theorem, \( (P3) \to 0 \) as \( \sup_{\theta} |D(\theta)| < \infty \) follows from the assumption \( \sup_{\theta} \mathbb{E}_{P_\theta}[d(x, \xi)^2] < \infty \). Thus, we get \( (P2) \to 0 \) as \( n, m \to \infty \). Therefore the bias converges to 0 as \( n, m \to \infty \).

To obtain the convergence rate, we use the same decomposition. Define \( \mathcal{E}'(\xi(\theta)) = d(\xi(\theta)) - D(\theta) \) as the zero mean error term. Note that this term has a bounded variance by Assumption 3.6. For \( (P1) \), it was shown that

\[ |(P1)| \leq \mathbb{E}_{\theta, P_\theta} \left[ \left| \hat{D}^m(\theta) - D(\theta) \right| \right] = \mathbb{E}_{\theta, P_\theta} \left[ \frac{1}{m} \sum_{j=1}^{m} \mathcal{E}'(\xi_j(\theta)) \right] . \]

Scaling with \( \sqrt{m} \), we get

\[ \sqrt{m}\mathbb{E}_{\theta, P_\theta} \left[ \left| \frac{1}{m} \sum_{j=1}^{m} \mathcal{E}'(\xi_j(\theta)) \right| \right] = \mathbb{E}_{\theta, P_\theta} \left[ \sqrt{m} \frac{1}{m} \sum_{j=1}^{m} \mathcal{E}'(\xi_j(\theta)) \right] \]

where the term inside the absolute value converges to a mean zero normal random variable by CLT. It follows that \( \sqrt{m}(P1) = O(1) \) and \( (P1) = O(m^{-1/2}) \).

We have the following lemma by Zhu et al. (2020) that is useful when working with \( (P3) \). In the following \( \hat{v}_\alpha^m \) is the \( \alpha \) quantile of \( \hat{H}^m \), i.e. \( \hat{F}_m(\hat{v}_\alpha^m) = \alpha \) where \( \hat{F}_m(\cdot) \) is the CDF of the noised response function and \( \hat{f}_m(\cdot) \) is the PDF of it.

**Lemma A.1.** Lemma B.4, Zhu et al. (2020) - Under Assumption 3.2,

\[ \hat{v}_\alpha^n - \hat{v}_\alpha^m = \frac{1}{\hat{f}_m(\hat{v}_\alpha^m)} \left( \alpha - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(\hat{H}_m(\theta_i) \leq \hat{v}_\alpha^m)} \right) + A_n, \]

where \( A_n = O_{a.s.}(n^{-3/4} \log n)^{3/4} \) and holds uniformly for all \( m \).

Recall that

\[ (P2) = \mathbb{E}_{\theta, P_\theta} \left[ \frac{1}{n} \sum_{i=1}^{n} D_i \left( \mathbb{1}_{(\hat{H}_m \geq \hat{v}_\alpha^m)} - \mathbb{1}_{(H_i \geq v_{\alpha})} \right) \right] = \mathbb{E}_{\theta, P_\theta} \left[ D_i \left( \mathbb{1}_{(\hat{H}_m \geq \hat{v}_\alpha^m)} - \mathbb{1}_{(H_i \geq v_{\alpha})} \right) \right] . \]

Since the difference of the indicators is \( \in \{-1, 0, 1\} \) and the gradient is bounded by Assumption 3.6, we
can write

\[-\sup_{\theta} |D(\theta)| \mathbb{E}_{\theta, \bar{P}_n} \left[ 1(H_n > \hat{v}^{n,m}_\alpha) - 1(H_i > v_o) \right] \leq \mathbb{E}_{\theta, \bar{P}_n} \left[ D_i(1(H_n > \hat{v}^{n,m}_\alpha) - 1(H_i > v_o)) \right] \leq \sup_{\theta} |D(\theta)| \mathbb{E}_{\theta, \bar{P}_n} \left[ 1(H_n > \hat{v}^{n,m}_\alpha) - 1(H_i > v_o) \right].\]

It follows that \((P2) = O((P4))\). In the following, we use a trick from Section 4.2 of Hong and Liu (2009) to mitigate the dependency of \(\hat{H}^m(\theta)\) and \(\hat{v}^{n,m}_\alpha\). After applying the trick, we have the following where \(\hat{v}^{(n-1),m}\) is calculated using \(\theta_2, \theta_3, \ldots, \theta_n\) and is independent of \(\hat{H}^m(\theta_1)\). Note that by definition

\[
\mathbb{E}_{\theta, \bar{P}_n}[1(H_i > v_o)] = \alpha = \mathbb{E}_{\theta, \bar{P}_n}[1(\hat{H}^m(\theta_1) > v_o)].
\]

\[(P4) = \mathbb{E}_{\theta, \bar{P}_n} \left[ 1(H^n(\theta_1) > \hat{v}^{(n-1),m}) \right] - \alpha
= \mathbb{E}_{\theta, \bar{P}_n} \left[ 1(H^n(\theta_1) > \hat{v}^{(n-1),m}) - 1(H^n(\theta_1) > \hat{v}^m) \right]
= \mathbb{E}_{\hat{v}^{(n-1),m}} \mathbb{E}_{\theta, \bar{P}_n} \left[ 1(H^n(\theta_1) > \hat{v}^{(n-1),m}) - 1(H^n(\theta_1) > \hat{v}^m) \bigg| \hat{v}^{(n-1),m} \right]
= \mathbb{E}_{\hat{v}^{(n-1),m}} \left[ \hat{F}^m(\hat{v}^{(n-1),m}) - \hat{F}^m(\hat{v}^m) \right]
= \mathbb{E}_{\hat{v}^{(n-1),m}} \left[ \hat{f}^m(\hat{v}^m)(\hat{v}^{(n-1),m} - \hat{v}^m) + o(\hat{v}^{(n-1),m} - \hat{v}^m) \right].
\]

If we ignore the constant and the \(o(\cdot)\) terms, we need to show that

\[
\sqrt{n} \mathbb{E}_{\hat{v}^{(n-1),m}} \left[ (\hat{v}^{(n-1),m} - \hat{v}^m) \right] \to 0. \tag{3}
\]

It is seen from the Proof of Theorem 3.6 of Zhu et al. (2020) that under Assumption 3.2 \(n = o(m^2)\) is a sufficient and necessary condition for

\[
\lim_{n,m \to \infty} \sqrt{n} \mathbb{E}_{\hat{v}^{(n-1),m}} \left[ (\hat{v}^{(n-1),m} - \hat{v}^m) \right] \Rightarrow \mathcal{N}(0, \sigma_v) \tag{4}
\]

where \(\sigma_v = \frac{n(1-\alpha)}{f^2(v_o)}\). All that is left is to justify the interchange of the limit and expectation. For this, it suffices to show that \(\sup_{n,m} \mathbb{E}_{\theta, \bar{P}_n}[n(\hat{v}^{(n-1),m} - \hat{v}^m)^2] = \sup_{n,m} \mathbb{E}_{\theta, \bar{P}_n}[(n+1)(\hat{v}^{n,m} - \hat{v}^m)^2] < \infty\), i.e. the
sequence is uniformly integrable.

$$E_{\theta, P}(n + 1)(\hat{v}_{\alpha}^m - \bar{v}_\alpha^m)^2 = (n + 1)E_{\theta, P}[\left(\hat{m}(\bar{v}_\alpha^m) \left( \alpha - \frac{1}{n} \sum_{i=1}^{n} 1_{(\hat{m}(\theta_i) \leq \bar{v}_\alpha^m)} \right) + A_n \right)^2]$$

$$= (n + 1)(\hat{m}(\bar{v}_\alpha^m))^2E_{\theta, P}\left[ (\alpha - \frac{1}{n} \sum_{i=1}^{n} 1_{(\hat{m}(\theta_i) \leq \bar{v}_\alpha^m)})^2 \right]$$

$$+ 2(n + 1)\hat{m}(\bar{v}_\alpha^m)E_{\theta, P}\left[ \alpha - \frac{1}{n} \sum_{i=1}^{n} 1_{(\hat{m}(\theta_i) \leq \bar{v}_\alpha^m)} \right] A_n$$

$$+ (n + 1)A_n^2.$$  

Here, the expectation in the second term equals zero and the third term is bounded by definition. So, ignoring the constant terms, we only need to show that the following is bounded.

$$(n + 1)E_{\theta, P}\left[ (\alpha - \frac{1}{n} \sum_{i=1}^{n} 1_{(\hat{m}(\theta_i) \leq \bar{v}_\alpha^m)})^2 \right] = (n + 1)E_{\theta, P}\left[ (\alpha - \frac{1}{n} \sum_{i=1}^{n} 1_{(\hat{m}(\theta_i) \leq \bar{v}_\alpha^m)})^2 \right]$$

$$= (n + 1)E_{\theta, P}\left[ \alpha^2 - 2\alpha \frac{1}{n} \sum_{i=1}^{n} 1_{(\hat{m}(\theta_i) \leq \bar{v}_\alpha^m)} + \left( \frac{1}{n} \sum_{i=1}^{n} 1_{(\hat{m}(\theta_i) \leq \bar{v}_\alpha^m)} \right)^2 \right]$$

$$= (n + 1)\left( \alpha^2 - 2\alpha^2 + \frac{1}{n^2}(n\alpha + n(n - 1)\alpha^2) \right)$$

$$= \frac{n + 1}{n}\alpha - \alpha^2$$

where the results holds uniformly for all $m$ and $\sup_{\alpha} \frac{n + 1}{n}\alpha - \alpha^2 < \infty$. It follows that

$$\lim_{n,m \to \infty} \sqrt{n}E_{\theta, P}[\hat{v}_{\alpha}^{m-1} - \bar{v}_\alpha^m] \to 0 \text{ as } n \to \infty. \quad (5)$$

Therefore, we get that $(P2) = o(n^{-1/2})$. Recalling the relation $n = \Theta(m)$ and putting the two together, we conclude that the bias is $O(n^{-1/2})$.

\[\square\]

**Theorem 3.13.** Under Assumptions 3.2, 3.3, 3.6, 3.10, $\psi_{\alpha}^{n,m}(x)$ is a strongly consistent estimator of $c'_\alpha$.

**Proof.** Recall that $\psi_{\alpha}^{n,m} = \frac{1}{(1-\alpha)n} \sum_{i=1}^{n} \hat{D}_i\sum_{i=1}^{n} 1_{(\hat{m}(\theta_i) \geq \bar{v}_\alpha^m)}$. We need to show that $\psi_{\alpha}^{n,m} - c'_\alpha \to 0 \text{ w.p.1 as}$
\[ n, m \to \infty. \]

\[
\psi_{\alpha, m}^{\theta} - \psi_{\alpha, m}^0 = \frac{1}{(1 - \alpha)} \left[ \frac{1}{n} \sum_{i=1}^{n} \hat{D}_{i}^{m} I_{(H_{i} \geq \psi_{\alpha, m}^0)} - \mathbb{E}_{\theta}[D(\theta) I_{(H(\theta) \geq \psi_{\alpha})}] \right]
= \frac{1}{(1 - \alpha)} \left[ \frac{1}{n} \sum_{i=1}^{n} \hat{D}_{i}^{m} I_{(H_{i} \geq \psi_{\alpha, m}^0)} - \frac{1}{n} \sum_{i=1}^{n} \hat{D}_{i}^{m} I_{(H_{i} \geq v_{\cdot})} \right] \\
\begin{align*}
&\quad + \frac{1}{n} \sum_{i=1}^{n} \hat{D}_{i}^{m} I_{(H_{i} \geq v_{\cdot})} - \mathbb{E}_{\theta}[D(\theta) I_{(H(\theta) \geq v_{\cdot})}] \\
&\quad \text{(P1)}
\end{align*}
\]

We will show that each part individually goes to zero w.p.1.

\[
(P2) = \frac{1}{n} \sum_{i=1}^{n} \hat{D}_{i}^{m} I_{(H_{i} \geq v_{\cdot})} - \mathbb{E}_{\theta}[D(\theta) I_{(H(\theta) \geq v_{\cdot})}] = \frac{1}{n} \sum_{i=1}^{n} \hat{D}_{i}^{m} I_{(H_{i} \geq v_{\cdot})} - \frac{1}{n} \sum_{i=1}^{n} D_{i} I_{(H_{i} \geq v_{\cdot})} + \frac{1}{n} \sum_{i=1}^{n} D_{i} I_{(H_{i} \geq v_{\cdot})} - \mathbb{E}_{\theta}[D(\theta) I_{(H(\theta) \geq v_{\cdot})}]
\]

Here, the second part goes to zero with probability one by Strong Law of Large Numbers. For the first part,

\[
\left| \frac{1}{n} \sum_{i=1}^{n} \hat{D}_{i}^{m} I_{(H_{i} \geq v_{\cdot})} - \frac{1}{n} \sum_{i=1}^{n} D_{i} I_{(H_{i} \geq v_{\cdot})} \right| \leq \frac{1}{n} \sum_{i=1}^{n} |\hat{D}_{i}^{m} - D_{i}| I_{(H_{i} \geq v_{\cdot})}
\leq \sup_{\theta} |\hat{D}_{i}^{m}(\theta) - D(\theta)|
= \sup_{\theta} \left| \frac{1}{m} \sum_{j=1}^{m} d(\xi(\theta)) - D(\theta) \right|
\]

which is the sample average of a mean 0 random variable with finite variance (by the assumption \( \sup_{\theta} \mathbb{E}_{\theta}[d(x, \xi)^2] < \infty \)). Therefore, by SLLN, \((P2) \to 0 \) w.p.1 as \( n, m \to \infty \). We need to show that \((P1) = \frac{1}{n} \sum_{i=1}^{n} \hat{D}_{i}^{m}(1_{(H_{i} \geq \psi_{\alpha, m}^0)} - 1_{(H_{i} \geq v_{\cdot})}) \to 0 \) w.p.1 as \( n, m \to \infty \). By Cauchy-Schwarz inequality, we have

\[
\left| \frac{1}{n} \sum_{i=1}^{n} \hat{D}_{i}^{m}(1_{(H_{i} \geq \psi_{\alpha, m}^0)} - 1_{(H_{i} \geq v_{\cdot})}) \right| \leq \left[ \frac{1}{n} \sum_{i=1}^{n} |\hat{D}_{i}^{m}|^2 \right]^{1/2} \left[ \frac{1}{n} \sum_{i=1}^{n} |1_{(H_{i} \geq \psi_{\alpha, m}^0)} - 1_{(H_{i} \geq v_{\cdot})}|^2 \right]^{1/2}
\]

In order to show that the first term is finite, we will show that \( \frac{1}{n} \sum_{i=1}^{n} (\hat{D}_{i}^{m})^2 \to \mathbb{E}_{\theta}[D(\theta)^2] \) w.p.1 which is bounded by the assumption \( \sup_{\theta} \mathbb{E}_{\theta}[d(x, \xi)^2] < \infty \).

\[
\frac{1}{n} \sum_{i=1}^{n} (\hat{D}_{i}^{m})^2 - \mathbb{E}_{\theta}[D(\theta)^2] = \frac{1}{n} \sum_{i=1}^{n} (\hat{D}_{i}^{m})^2 - \frac{1}{n} \sum_{i=1}^{n} (D_{i})^2 + \frac{1}{n} \sum_{i=1}^{n} (D_{i})^2 - \mathbb{E}_{\theta}[D(\theta)^2]
\]
The second part goes to zero w.p.1 by SLLN. We need to show that \( \frac{1}{n} \sum_{i=1}^{n} ((\hat{D}^m_i)^2 - (D_i)^2) \to 0 \) w.p.1.

Define \( E'((\xi(\theta)) := d(\xi(\theta)) - D(\theta) \).

\[
\frac{1}{n} \sum_{i=1}^{n} ((\hat{D}^m_i)^2 - (D_i)^2) \leq \sup_{\theta} \left( \hat{D}^m(\theta)^2 - D(\theta)^2 \right) \]

\[
= \sup_{\theta} \left[ \left( \frac{1}{m} \sum_{j=1}^{m} E'((\xi_j(\theta))) \right)^2 - 2D(\theta) \frac{1}{m} \sum_{j=1}^{m} E'((\xi_j(\theta))) \right]
\]

Note that \( \sup_{\theta} D(\theta) < \infty \) and \( E'((\xi_j(\theta))) \) is a mean zero random variable with finite variance by the assumption \( \sup_{\theta} E \mathbb{P}_{\theta} [d(x, \xi)^2] < \infty \). Therefore, both terms converge to zero w.p.1 and we get \( \left[ \frac{1}{n} \sum_{i=1}^{n} |\hat{D}^m_i|^2 \right]^{1/2} < \infty \).

To complete the proof, we need to show that \( \left[ \frac{1}{n} \sum_{i=1}^{n} |1(\hat{H}_m^i \geq \hat{v}_{n,m}^i) - 1(H_i \geq v_{\alpha})|^2 \right]^{1/2} \to 0 \) w.p.1 as \( n, m \to \infty \). This follows from Continuous Mapping Theorem and Proposition 3.11. Therefore, \( \psi_{n,m}^{\alpha} \to c_{\alpha}' \) w.p.1 as \( n, m \to \infty \).

References

Gordy, M. B., & Juneja, S. (2010). Nested simulation in portfolio risk measurement. Management Science, 56(10).

Hong, L. J. (2009). Estimating quantile sensitivities. Operations Research, 57(1).

Hong, L. J., & Liu, G. (2009). Simulating sensitivities of conditional value at risk. Management Science, 55(2).

Kim, S., Pasupathy, R., & Henderson, S. G. (2015). A guide to sample average approximation. In M. C. Fu (Ed.), Handbook of simulation optimization. Springer New York.

Zhu, H., Liu, T., & Zhou, E. (2020). Risk quantification in stochastic simulation under input uncertainty. ACM Trans. Model. Comput. Simul., 30(1).