Appendix for “Black-Box Importance Sampling”

Qiang Liu
Dartmouth College

Jason D. Lee
University of Southern California

1 Kernelized Stein Discrepancy and MMD

Given RKHS $\mathcal{H}$ with kernel $k(x, x')$, the maximum mean discrepancy (MMD) between two distributions with density $p(x)$ and $q(x)$ is defined as

$$\text{MMD}_{\mathcal{H}}(q, p) = \max_{f \in \mathcal{H}} \{ \mathbb{E}_q f - \mathbb{E}_p f \ s.t. \ ||f||_{\mathcal{H}} \leq 1 \},$$

which can be shown to be equivalent to

$$\text{MMD}_{\mathcal{H}}(q, p)^2 = \mathbb{E}_{x, x' \sim p}[k(x, x')] - 2\mathbb{E}_{x \sim p, y \sim q}[k(x, y)] + \mathbb{E}_{y, y' \sim q}[k(y, y')].$$

We show that kernelized discrepancy is equivalent to $\text{MMD}_{\mathcal{H}_p}(q, p)$, equipped with the $p$-Steinized kernel $k_p(x, x')$.

**Proposition 1.1.** Assume (3) is true, we have

$$S(q, p) = \text{MMD}_{\mathcal{H}_p}(q, p)^2.$$

**Proof.** Simply note that $\mathbb{E}_{x' \sim p}[k_p(x, x')] = 0$ for any $x$, we have

$$\text{MMD}_{\mathcal{H}_p}(q, p)^2 = \mathbb{E}_{x, x' \sim q}[k_p(x, x')] = S(q, p).$$

Similarly, we also have

$$\sqrt{S(\{w_i, x_i\}_{i=1}^n, p)} = \text{MMD}_{\mathcal{H}_p}(\{x_i, w_i\}, p) = \max_{f \in \mathcal{H}} \{ \sum_{i=1}^n w_i f(x_i) - \mathbb{E}_p f \ s.t. \ ||f||_{\mathcal{H}} \leq 1 \}.$$

**Proof of Proposition 3.1.** Let $\tilde{h}(x) = h(x) - \mathbb{E}_p h$, we have

$$|\sum_i w_i \tilde{h}(x_i)| = |\sum_i w_i \langle \tilde{h}, k_p(\cdot, x_i) \rangle_{\mathcal{H}_p}|$$

$$= |\langle \tilde{h}, \sum_i w_i k_p(\cdot, x_i) \rangle_{\mathcal{H}_p}|$$

$$\leq ||\tilde{h}||_{\mathcal{H}_p} \cdot ||\sum_i w_i k_p(\cdot, x_i)||_{\mathcal{H}_p}$$

$$= ||\tilde{h}||_{\mathcal{H}_p} \cdot \sqrt{S(\{w_i, x_i\}, p)}.$$

where we used Cauchy-Schwarz inequality and the fact that $||\sum_i w_i k_p(\cdot, x_i)||_{\mathcal{H}_p}^2 = \sum_{i,j} w_i w_j k_p(x_i, x_j) = S(\{w_i, x_i\}, p).$
2 Convergence Rate

We consider the error rate of our estimator \( \sum_i \hat{w}_i(x) h(x_i) \) with \( \{\hat{w}_i(x)\} \) given by the optimization in (6), under the assumption that \( x = \{x_i\}_{i=1}^n \) is i.i.d. drawn from an (unknown) distribution \( q(x) \). Based on the bound in Proposition 3.1, we can establish an error rate \( O(n^{-\delta}) \) by finding a set of oracle “reference weights” \( \{w_i(x)\} \), as a function of \( x \), such that \( S(x, w_i(x), p) = O(n^{-2\delta}) \), because

\[
\sum_i \hat{w}_i(x) h(x_i) - \mathbb{E}_p h \leq C_h \cdot \sqrt{S(x, w_i(x), p)} \leq C_h \cdot \sqrt{S(x, w_i(x), p)} = O(n^{-\delta}),
\]

where \( C_h = ||h - \mathbb{E}_p h||_{\mathcal{H}_w} \). This idea of using reference weights has been used in Briol et al. 2015b to study the convergence rate of Bayesian Monte Carlo.

Section 2.1 proves the \( O(n^{-1/2}) \) rate using the typical importance sampling weights as the reference weight. Section 2.2 proves a better \( o(n^{-1/2}) \) rate by using a reference weight based on a control variates method constructed with an orthogonal basis estimator.

2.1 \( O(n^{-1/2}) \) Rate

We use the typical importance sampling weight as a reference weight and establish \( O(n^{-1/2}) \) rate on the error of our estimator.

**Assumption 2.1.** Assume \( p(x)/q(x) > 0 \) for \( \forall x \in \mathcal{X} \) and \( \mathbb{E}_{x \sim q}[(p(x)/q(x))^2] < \infty \), \( \mathbb{E}_{x \sim q}[(p(x)/q(x))^2 k_p(x, x)] < \infty \), and \( \mathbb{E}_{x, x' \sim q}[(p(x)/q(x))^2 k_p(x, x')] < \infty \).

**Lemma 2.2.** Assume \( \{x_i\}_{i=1}^n \) is i.i.d. drawn from \( q(x) \)

\[
w_i^* = \frac{1}{Z} p(x_i)/q(x_i), \quad Z = \sum_i p(x_i)/q(x_i),
\]

then under Assumption 2.1, we have

\[
S(x, w_i^*, p) = O(n^{-1}).
\]

**Proof.** Define \( v_i^*(x_i) = \frac{1}{n} p(x_i)/q(x_i) \), and

\[
S(x, v_i^*, p) = \frac{1}{n^2} \sum_{ij} \frac{p(x_i) p(x_j)}{q(x_i) q(x_j)} k_p(x_i, x_j),
\]

then \( S(x, v_i^*, p) \) is a degenerate V-statistic since by (3) we have

\[
\mathbb{E}_{x' \sim q}[(p(x)/q(x))^2 k_p(x, x')] = \mathbb{E}_{x' \sim p} [k_p(x, x')] = 0, \quad \forall x \in \mathcal{X}
\]

then we have [see e.g., (7)]

\[
S(x, v_i^*, p) = O(n^{-1}).
\]

In addition, note that \( \sum_{i=1}^n v_i^* = 1 + O(n^{-1/2}) \), we have

\[
S(x, v_i^*, p) = \frac{S(x, v_i^*, p)}{(\sum_i v_i^*)^2} = O(n^{-1}).
\]

**Theorem 2.3.** Assume \( \{x_i\} \) is i.i.d. drawn from \( q(x) \), and \( \{\hat{w}_i(x)\} \) is given by (6), then under Assumption 2.1, we have

\[
\sum_{i=1}^n \hat{w}_i(x) h(x_i) - \mathbb{E}_p h = O(n^{-1/2}).
\]

**Proof.** Simply note that

\[
S(x, w_i^*, p) \leq S(x, w_i^*, p) = O(n^{-1}),
\]

and combining with Proposition 3.1 gives the result.

\[\Box\]
2.2 \( o(n^{-1/2}) \) Rate

We prove Theorem 3.3 that shows an \( o(n^{-1/2}) \) rate for our estimator. Our method is based on constructing a reference weight by using a two-fold control variate method based on the first \( L \) orthogonal eigenfunctions \( \{\phi_\ell\} \) of kernel \( k_p(x,x') \).

We first re-state the assumptions made in Theorem 3.3.

Assumption 2.4. 1. Assume \( k_p(x,x') \) has the following eigen-decomposition

\[
k_p(x,x') = \sum_\ell \lambda_\ell \phi_\ell(x)\phi_\ell(x'),
\]

where \( \lambda_\ell \) are the positive eigenvalues sorted in non-increasing order, and \( \phi_\ell \) are the eigenfunctions orthonormal w.r.t. distribution \( p(x) \), that is,

\[
E_p[\phi_\ell \phi_{\ell'}] \overset{\text{def}}{=} \int p(x)\phi_\ell(x)\phi_{\ell'}(x)dx = \mathbb{I}[\ell = \ell'].
\]

2. \( \text{trace}(k_p(x,x')) = \sum_{\ell=1}^\infty \lambda_\ell < \infty \).

3. \( \text{var}_{x\sim q}[w_*(x)^2 \phi_\ell(x)\phi_{\ell'}(x)] \leq M \) for all \( \ell \) and \( \ell' \), where \( w_*(x) = p(x)/q(x) \).

4. \( |\phi_\ell(x)|^2 \leq M_2 \), and \( w_*(x) \overset{\text{def}}{=} p(x)/q(x) \leq M_3 \) for any \( x \in \mathcal{X} \).

The following is an expended version of Theorem 3.3.

Theorem 2.5. Assume \( \{x_i\}_{i=1}^n \) is i.i.d. drawn from \( q(x) \), and \( \hat{w}_i \) is calculated by

\[
\hat{w} = \arg\min_w wK_p w, \quad \text{s.t.} \sum_i w_i = 1, \quad w_i \geq 0,
\]

and \( h - E_p h \in \mathcal{H}_p \). Under Assumption 2.4 we have

\[
E_{x\sim q}(\|\sum_i \hat{w}_i h(x_i) - E_p h\|^2) = O\left(\frac{1}{\sqrt{n}}\gamma(n)\right),
\]

where \( \gamma(n) = \min_{L \in \mathbb{N}} \left\{ \frac{M_3}{2} \mathbb{R}(L) + \frac{M_4 L}{2n} + M_f n(n+2) \exp\left(-\frac{n}{L^2M_0}\right) \right\} \),

where \( \mathbb{N}^+ \) is the set of positive integers, and \( \mathbb{R}(L) = \sum_{\ell>L} \lambda_\ell \) is the residual of the spectrum, and \( M_3 = 2M_3M \text{trace}(k_p) \), and \( M_f = \text{trace}(k_p(x,x')) M_2 \) and \( M_0 = \max(M_2^2 M_3, M_3^2 (M_2 M_3 + \sqrt{2})^2) \).

Remark To see how Theorem 2.5 implies Theorem 3.3 we just need to observe that we obviously have \( \gamma(n) \geq 2M_3^{1/2}, \) and \( \gamma(n) = o(1) \) by taking \( L = n^{1/4} \).

Based on Proposition 3.1 to prove Theorem 2.5 we just need to show that for any \( x = \{x_i\}_{i=1}^n \), there exists a set of positive and normalized weights \( \{w_i^+(x)\} \), as a function of \( x \), such that

\[
E_{x\sim q}[S(\{w_i^+(x), x_i\}, p)] = O\left(\frac{\gamma(n)}{n}\right).
\]

In the sequel, we construct such a weight based on a control variates method which uses the top eigenfunctions \( \phi_\ell \) as the control variates. Our proof includes the following steps:

1. Step 1: Construct a control variate estimator based on the orthogonal eigenfunction basis, and obtain the corresponding weights \( \{w_i(x)\} \).

2. Step 2: Bound \( E_{x\sim q}[S(\{w_i(x), x_i\}, p)] \).

3. Step 3. Construct a set of positive and normalized weights by \( w_i^+(x) = \frac{\max(0,w_i(x))}{\sum_i \max(0,w_i(x))} \), and establish the corresponding bound.
Appendix for “Black-Box Importance Sampling”

Proof of Theorem 2.5. Combine the bound in Lemma 2.7 and Lemma 2.9 below.

We note that the idea of using reference weights was used in Briol et al. [2015b] to establish the convergence rate of Bayesian Monte Carlo. Related results is also presented in Bach [2015]. The main additional challenge in our case is to meet the non-negative and normalization constraint (Step 3); this is achieved by showing that the \{w_i(x)\} constructed in Step 2 is non-negative with high probability, and their sum approaches to one when \( n \) is large, and hence \{w_i^+(x)\} is not significantly different from \{w_i(x)\}.

Note that if we discard the non-negative and normalization constraint (Step 3), the error bound would be \( \mathcal{O}(\gamma_0(n)n^{-1}) \), where

\[
\gamma_0(n) = \min_{L \in \mathbb{N}^+} \{2M_3\mathbb{R}(L) + 2M_4\frac{L}{n}\},
\]

as implied by Lemma 2.7. Therefore, the third term in \( \gamma(n) \) is the cost to pay for enforcing the constraints. However, this additional term does not influence the rate significantly once \( \mathbb{R}(L) = \sum_{\ell>L} \lambda_\ell \) decays sufficiently fast. For example, when \( \mathbb{R}(L) = \mathcal{O}(L^{-\alpha}) \) where \( \alpha > 1 \), both \( \gamma(n) \) and \( \gamma_0(n) \) equal \( \mathcal{O}(n^{-1+1/(\alpha+1)}) \); when \( \mathbb{R}(L) = \mathcal{O}(\exp(-\alpha L)) \) with \( \alpha > 0 \), both \( \gamma(n) \) and \( \gamma_0(n) \) equal \( \mathcal{O}(\log n) \). An open question is to derive upper bounds for the decay of eigenvalues \( \mathbb{R}(L) \) for given \( p \) and \( k(x,x') \), so that actual rates can be determined.

Step 1: Constructing the weights

We first construct a set of unnormalized, potentially negative reference weights, by using a two-fold control variates method based on the orthogonal eigenfunctions \( \{\phi_\ell\} \) of kernel \( k_p(x,x') \). Assume \( n \) is an even number, and we partition the data \( \{x_i\}_{i=1}^n \) into two parts \( D_0 = \{1, \ldots, \frac{n}{2}\} \) and \( D_1 = \{\frac{n}{2} + 1, \ldots \} \). For any \( h \in \mathcal{H}_p \), we have \( \mathbb{E}_p h = 0 \) by (3), and

\[
h(x) = \sum_{\ell=1}^\infty \beta_\ell \phi_\ell(x), \quad \beta_\ell = \mathbb{E}_{x \sim p}[h(x)\phi_\ell(x)].
\]

We now construct an orthogonal series estimator \( \hat{h}(x) \) for \( h(x) \) based on \( x_{D_0} \),

\[
\hat{h}_{D_0}(x) = \sum_{\ell=1}^L \hat{\beta}_{\ell,0} \phi_\ell(x), \quad \text{where} \quad \hat{\beta}_{\ell,0} = \frac{2}{n} \sum_{i \in D_0} h(x_i)\phi_\ell(x_i)\frac{p(x_i)}{q(x_i)},
\]

where we approximate \( \beta_\ell \) with an unbiased estimator \( \hat{\beta}_{\ell,0} \) since

\[
\mathbb{E}_{x \sim q}[\hat{\beta}_{\ell,0}] = \mathbb{E}_{x \sim q}[h(x)\phi_\ell(x)\frac{p(x)}{q(x)}] = \int p(x)h(x)\phi_\ell(x)dx = \beta_\ell.
\]

We also truncate at the \( L \)th basis functions to keep \( \hat{h}_{D_0}(x) \) a smooth function, as what is typically done in orthogonal basis estimators. We will discuss the choice of \( L \) later. Based on this we define a control variates estimator:

\[
\hat{Z}_0[h] = \frac{2}{n} \sum_{i \in D_1} [w_i(x_i)(h(x_i) - \hat{h}_{D_0}(x_i))],
\]

which gives an unbiased estimator for \( \mathbb{E}_p h = 0 \) because

\[
\mathbb{E}_{x \sim q}[^{\hat{Z}_0[h]}] = \int q(x)\frac{p(x)}{q(x)} (h(x) - \hat{h}_{D_0}(x_i))dx = \mathbb{E}_{x \sim p} h - \mathbb{E}_{x \sim q} [\mathbb{E}_{x \sim p} [\hat{h}_{D_0}(x) | x_{D_0}]] = 0,
\]

where the last step is because \( \mathbb{E}_{x \sim p} [\hat{h}_{D_0}(x) | x_{D_0}] = \sum_{\ell=1}^L \hat{\beta}_{\ell,0} \mathbb{E}_{x \sim p} [\phi_\ell(x)] = 0 \). Switching \( D_0 \) and \( D_1 \), we get another estimator

\[
\hat{Z}_1[h] = \frac{2}{n} \sum_{i \in D_0} [w_i(x_i)(h(x_i) - \hat{h}_{D_1}(x_i))],
\]

Averaging them gives

\[
\hat{Z}[h] = \frac{\hat{Z}_0[h] + \hat{Z}_1[h]}{2}.
\]
Lemma 2.6. Given $\hat{Z}[h]$ defined as above, for any $h \in \mathcal{H}_p$, we have

$$\hat{Z}[h] = \sum_{i=1}^{n} w_i(x) h(x_i), \quad \text{with} \quad w_i(x) = \begin{cases} \frac{1}{n} w_s(x_i) - \frac{2}{n^2} \sum_{j \in D_1} w_s(x_i) w_s(x_j) k_L(x_j, x_i), & \forall i \in D_0 \\ \frac{1}{n} w_s(x_i) - \frac{2}{n^2} \sum_{j \in D_0} w_s(x_i) w_s(x_j) k_L(x_j, x_i), & \forall i \in D_1 \end{cases}$$

where $w_s(x) = p(x)/q(x)$ and $k_L(x, x') = \sum_{\ell=1}^{L} \phi_{\ell}(x) \phi_{\ell}(x')$.

Proof. We have

$$\hat{Z}_0[h] = \frac{2}{n} \left[ \sum_{i \in D_1} w_s(x_i) \left( h(x_i) - \hat{h}_{D_0}(x_i) \right) \right]$$

$$= \frac{2}{n} \left[ \sum_{i \in D_1} w_s(x_i) \left( h(x_i) - \sum_{\ell=1}^{L} \hat{\beta}_{x_0, \ell} \phi_{\ell}(x) \right) \right]$$

$$= \frac{2}{n} \left[ \sum_{i \in D_1} w_s(x_i) \left( h(x_i) - \frac{2}{n} \sum_{\ell=1}^{L} \sum_{j \in D_0} h(x_j) w_s(x_j) \phi_{\ell}(x_j) \phi_{\ell}(x_i) \right) \right]$$

$$= \frac{2}{n} \sum_{i \in D_1} w_s(x_i) h(x_i) - \frac{4}{n^2} \sum_{j \in D_0} \sum_{i \in D_1} h(x_j) w_s(x_j) w_s(x_i) \sum_{\ell=1}^{L} \phi_{\ell}(x_j) \phi_{\ell}(x_i)$$

$$= \frac{2}{n} \sum_{i \in D_1} w_s(x_i) h(x_i) - \frac{4}{n^2} \sum_{j \in D_0} \sum_{i \in D_1} h(x_j) w_s(x_j) w_s(x_i) k_L(x_i, x_j)$$

$$\overset{\text{def}}{=} \sum_{i=1}^{n} w_{i,0} h(x_i),$$

where

$$w_{i,0} = \begin{cases} -\frac{4}{n^2} \sum_{j \in D_1} w_s(x_i) w_s(x_j) k_L(x_j, x_i), & \forall i \in D_0 \\ \frac{2}{n} w_s(x_i), & \forall i \in D_1 \end{cases} \quad (2)$$

We can derive the same result for $\hat{Z}_1[h]$ and averaging them would gives the result. \qed

Step 2: Calculating $\mathbb{E}_{x \sim q}(\mathbb{S}(\{x, w_i(x)\}, \ p))$

Lemma 2.7. Under Assumption 2.4 for the weights $\{w_i(x)\}$ defined in Lemma 2.6 we have

$$\mathbb{E}_{x \sim q}[\mathbb{S}(\{x_i, w_i(x)\}, \ p)] \leq \frac{2}{n} [M_3 \mathbb{R}(L) + M_4 \frac{L}{n}]$$

where $M_3$ is the upper bound of $p(x)/q(x)$, $\forall x \in \mathcal{X}$ and $\mathbb{R}(L) = \sum_{\ell > L} \lambda_{\ell}$ and $M_4 = 2 M_3 \max_{\mathcal{P}} \{\sum_{\ell} \lambda_{\ell} \rho_{\ell'}\} \leq 2 M_3 \text{trace}(k_p)$. 
Proof. First, for any $h \in \mathcal{H}_p$ (such that $\mathbb{E}_p[h] = 0$), we have

\[
\mathbb{E}_{x \sim q}\left[ \hat{Z}_0[h]^2 \right] = \mathbb{E}_{x \sim q}\left[ \left( \frac{2}{n} \sum_{i \in \mathcal{D}_1} w_s(x_i)(h(x_i) - \hat{h}_{\mathcal{D}_0}(x_i)) \right)^2 \right] = \frac{4}{n^2} \mathbb{E}_{x \sim q}\left\{ \sum_{i \in \mathcal{D}_1} \mathbb{E}_{x_i \sim q}\left[ w_s(x_i)^2(h(x_i) - \hat{h}_{\mathcal{D}_0}(x_i))^2 \right] + \sum_{i \neq j, i, j \in \mathcal{D}_1} \mathbb{E}_{x_i, x_j \sim q}\left[ w_s(x_i)(h(x_i) - \hat{h}_{\mathcal{D}_0}(x_i)) w_s(x_j)(h(x_j) - \hat{h}_{\mathcal{D}_0}(x_j)) \right] \right\} \\
= \frac{4}{n^2} \mathbb{E}_{x \sim q}\left\{ \sum_{i \in \mathcal{D}_1} \mathbb{E}_r\left[ (h(x_i) - \hat{h}_{\mathcal{D}_0}(x_i))^2 \right] + \sum_{i \neq j, i, j \in \mathcal{D}_1} \mathbb{E}_p\left[ (h(x_i) - \hat{h}_{\mathcal{D}_0}(x_i))(h(x_j) - \hat{h}_{\mathcal{D}_0}(x_j)) \right] \right\} \\
= \frac{2}{n} \mathbb{E}_{x \sim q}\left\{ \int \frac{p(x)^2}{q(x)} (h(x) - \hat{h}_0(x))^2 dx \right\} \quad \text{(because $\mathbb{E}_p h = \mathbb{E}_p \hat{h} = 0$)} \\
\leq \frac{2 M_3}{n} \mathbb{E}_{x \sim q}\left\{ \mathbb{E}_p[(h(x) - \hat{h}_0(x))^2] \right\} \quad \text{(because $p(x)/q(x) \leq M_3$ by assumption)} \\
= \frac{2 M_3}{n} \mathbb{E}_{x \sim q}\left\{ \sum_{\ell > L} \beta^2_\ell + \sum_{\ell \leq L} (\beta_\ell - \hat{\beta}_\ell, 0)^2 \right\} \\
= \frac{2 M_3}{n} \mathbb{E}_{x \sim q}\left\{ \sum_{\ell > L} \beta^2_\ell + \sum_{\ell < L} \text{var}_{x \sim q}(\hat{\beta}_\ell, 0) \right\} \quad \text{(because $\mathbb{E}_{x \sim q}[\hat{\beta}_\ell, 0] = \beta_\ell$)} \\
= \frac{2 M_3}{n} \left[ \sum_{\ell > L} \beta^2_\ell + \frac{2}{n} \sum_{\ell < L} \text{var}_{x \sim q}[w_s(x)\phi_\ell(x)h(x)] \right].
\]

We can derive the same result for $\hat{Z}_1[h]$ and hence

\[
\mathbb{E}_{x \sim q}\left[ \hat{Z}[h]^2 \right] \leq \frac{1}{2} (\mathbb{E}_{x \sim q}\left[ \hat{Z}_0[h]^2 \right] + \mathbb{E}_{x \sim q}\left[ \hat{Z}_1[h]^2 \right]) = \frac{2 M_3}{n} \left[ \sum_{\ell > L} \beta^2_\ell + \frac{2}{n} \sum_{\ell < L} \text{var}_{x \sim q}[w_s(x)\phi_\ell(x)h(x)] \right].
\]

Taking $h(x) = \phi_{\ell'}(x)$ for which we have $\beta_\ell = 1[\ell = \ell']$, we get

\[
\mathbb{E}_q[\hat{Z}[\phi_{\ell'}]^2] \leq \begin{cases} 
\frac{4 M_3}{n^2} \sum_{\ell \leq L} \text{var}_{x \sim q}[w_s(x)\phi_\ell(x)\phi_{\ell'}(x)] & \text{if } \ell' \leq L \\
\frac{2 M_3}{n} + \frac{4 M_3}{n^2} \sum_{\ell < L} \text{var}_{x \sim q}[w_s(x)\phi_\ell(x)\phi_{\ell'}(x)] & \text{if } \ell' > L.
\end{cases}
\]
Define \( \rho_{\ell \ell} = \text{var}_{x \sim q}[w_s(x) \phi_\ell(x) \phi_\ell(x)] \) and we have \( \rho_{\ell \ell} \leq M \) by Assumption 2.4. We have

\[
\mathbb{E}_{x \sim q}[S(\{x_i, w_i(x)\}, p)] = \mathbb{E}_{x \sim q}[\sum_{i,j=1}^{n} w_i(x)w_j(x)k_p(x_i, x_j)] \\
= \mathbb{E}_{x \sim q}[\sum_{i,j=1}^{n} w_i(x)w_j(x) \sum_{\ell=1}^{n} \lambda_\ell \phi_\ell(x_i) \phi_\ell(x_j)] \\
= \sum_{\ell}^{n} \lambda_\ell \mathbb{E}_{x \sim q}[\sum_{i=1}^{n} w_i(x) \phi_\ell(x_i)]^2 \\
= \sum_{\ell}^{n} \lambda_\ell \mathbb{E}_{x \sim q}[\hat{Z}[\phi_\ell]^2] \\
\leq \frac{2M_3}{n} \sum_{\ell>0} \lambda_\ell + \frac{2}{n} \sum_{\ell=1}^{\infty} \lambda_\ell \sum_{\ell' < \ell} \rho_{\ell \ell'} \\
\leq \frac{2}{n} [M_3 \sum_{\ell>0} \lambda_\ell + M_4 L],
\]

where \( M_4 = 2M_3 \max_{\ell} \{ \sum_{\ell} \lambda_\ell \rho_{\ell \ell'} \} \leq 2M_3 \text{trace}(k_p). \)

**Step 3: Meeting the Non-negative and Normalization Constraint**

The weights defined in (2.6) is not normalized to sum to one, and may also have negative values. To complete the proof, we define a set of new weights,

\[
w_i^+(x) = \frac{\max(0, w_i(x))}{\sum_{i} \max(0, w_i(x))}.
\]

We need to give the bound for \( S(\{x_i, w_i^+(x)\}, p) \) based on the bound of \( O(S(\{x_i, w_i(x)\}, p)) \). The key observation is that we have \( \sum_{i=1}^{n} w_i(x) \xrightarrow{p} 1 \) and \( w_i(x) \geq 0 \) with high probability for the weights given by in Lemma 2.6.

**Lemma 2.8.** For the weights \( \{w_i(x)\} \) defined in Lemma 2.6 under Assumption 2.4, we have

i). When \( x = \{x_i\}_{i=1}^{q} \sim q \), we have

\[
\Pr[w_i(x) < 0] \leq \exp(-\frac{n}{LM_2^2 M_3^2}), \quad \text{for } \forall i \leq n. \tag{3}
\]

ii). We have \( \mathbb{E}_{x \sim q}[\sum_i w_i(x)] = 1 \). Assume \( L \geq 1 \), we have

\[
\Pr(S < 1 - t) \leq 2 \exp(-\frac{n}{L^2 M_s^2}) \quad \text{ where } \quad M_s = M_2^2 (M_2 M_3 + \sqrt{2})^2/4, \tag{4}
\]

**Proof.** i). Recall that

\[
w_i(x) = \begin{cases} 
\frac{1}{n} w_s(x_i) - \frac{2}{n^2} \sum_{j \in D_1} w_s(x_i)w_s(x_j)k_L(x_j, x_i), & \forall i \in D_0 \\
\frac{1}{n} w_s(x_i) - \frac{2}{n^2} \sum_{j \in D_0} w_s(x_i)w_s(x_j)k_L(x_j, x_i), & \forall i \in D_1.
\end{cases}
\]

We just need to prove (3) for \( i \in D_0 \). Note that

\[
w_i(x) = \frac{1}{n} w_s(x_i) [1 - T], \quad \text{ where } \quad T = \frac{2}{n} \sum_{j \in D_1} w_s(x_j)k_L(x_j, x_i).
\]

Because \( \mathbb{E}(T \mid x_i) = \mathbb{E}_{x' \sim q}[w_s(x')k_L(x', x_i)] = 0 \) for \( \forall x \) and \( |w(x')k_L(x, x')| \leq LM_2 M_3, \forall x, x' \in X \), using Hoeffding’s inequality, we have

\[
\Pr(w_i(x) < 0) = \Pr(T > 1) \leq \exp(-\frac{n}{L^2 M_2^2 M_3^2}).
\]
where
\[ S_1 = \frac{1}{n} \sum_{i=1}^{n} w_i(x_i), \quad S_2 = -\frac{2}{nt^2} \sum_{i \in B_0} \sum_{j \in B_1} w_i(x_i)w_j(x_j)k_L(x_i, x_j), \]

where the first term is the standard importance sampling weights and the second term comes from the control variate. It is easy to show that \( \mathbb{E}[S_1] = 1 \) and \( \mathbb{E}[S_2] = 0 \), and hence \( \mathbb{E}[S] = 1 \). To prove the tail bound, note that for any \( t_1 + t_2 = t, t_1, t_2 > 0 \), we have
\[
\Pr(S < 1 - t) \leq \Pr(S_1 < 1 - t_1) + \Pr(S_2 \leq t_2) \leq \exp(-\frac{2nt^2}{M^2}) + \exp(-\frac{4nt^2}{L^2 M^2 M_3^2}),
\]
where the bound for \( S_2 \) uses the Hoeffding’s inequality for two-sample U statistics [?; Section 5b]. We take \( t_1 = \sqrt{2t}/(LM_2 M_3 + \sqrt{2}) \), we have
\[
\Pr(S < 1 - t) \leq 2 \exp(-\frac{4nt^2}{L^2 M_2^2 M_3^2 + \sqrt{2}/L}) \leq 2 \exp(-\frac{nt^2}{L^2 M_s}),
\]
where \( M_s = M_2^2 (M_2 M_3 + \sqrt{2})/4 \) (we assume \( \Lambda \geq 1 \)).

\[ \square \]

**Lemma 2.9.** Under Assumption 2.4 we have
\[
\mathbb{E}[S(\{x_i, w_i \dagger(x_i)\}, p)] \leq \frac{1}{4} \mathbb{E}[S(\{x_i, w_i(x_i)\}, p)] + M_f(n + 2) \exp(-\frac{n}{L^2 M_0}),
\]
where \( M_f = \text{trace}(k_p(x, x')) M_2 \) and \( M_0 = \text{max}(M_2^2 M_3, M_3^2 (M_2 M_3 + \sqrt{2})^2) \).

**Proof.** We use short notation \( f(w^+) = S(\{x_i, w_i \dagger(x_i)\}, p) \) for convenience. We have
\[
|f(w^+)| = \sum_{\ell} \lambda_{\ell}(\sum_i w_i^+ \phi_{\ell}(x_i))^2 \leq \text{trace}(k_p(x, x')) M_2 \overset{\text{def}}{=} M_f.
\]

Define \( \mathcal{E}_n \) to be the event that all \( w_i > 0 \) and \( \sum_i w_i \geq 1/2 \), that is, \( \mathcal{E}_n = \{\sum_i w_i \geq 1/2, \ w_i \geq 0, \ \forall i \in [n]\}. \) We have from Lemma 2.8 that
\[
\Pr(\mathcal{E}_n) \leq n \exp(-\frac{n}{L^2 M_2^2 M_3}) + 2 \exp(-\frac{n}{4L^2 M_s}).
\]

Note that under event \( \mathcal{E}_n \), we have \( w = w^+ \). Therefore,
\[
\mathbb{E}[f(w^+)] = \mathbb{E}[f(w^+) \mid \mathcal{E}_n] \cdot \Pr(\mathcal{E}_n) + \mathbb{E}[f(w^+) \mid \mathcal{E}_n^c] \cdot \Pr(\mathcal{E}_n^c) - \frac{1}{4} \mathbb{E}[f(w) \mid \mathcal{E}_n] \cdot \Pr(\mathcal{E}_n) + M_f \cdot \Pr(\mathcal{E}_n^c)
\]
\[
\leq \mathbb{E}[f(w^+) \mid \mathcal{E}_n] \cdot \Pr(\mathcal{E}_n) + M_f \cdot (1 - \Pr(\mathcal{E}_n))
\]
\[
\leq \frac{1}{4} \mathbb{E}[f(w) \mid \mathcal{E}_n] + M_f \cdot \Pr(\mathcal{E}_n^c)
\]
\[
\leq \frac{1}{4} \mathbb{E}[f(w)] + M_f \cdot \left[n \exp(-\frac{n}{L^2 M_2^2 M_3}) + 2 \exp(-\frac{n}{4L^2 M_s})\right]
\]
\[
\leq \frac{1}{4} \mathbb{E}[f(w)] + M_f(n + 2) \exp(-\frac{n}{L^2 M_0}).
\]

\[ \square \]

### 3 Additional Empirical Results

Here we show in Figure 1 an additional empirical result when \( p(x) \) is a Gaussian mixture model shown in Figure 1(a) and \( \{x_i\}_{i=1}^n \) is generated by running \( n \) independent chains of MALA for 10 steps.
Figure 1: Gaussian Mixture Example. (a) The contour of the distribution $p(x)$ that we use, and $\{x_i\}_{i=1}^n$ is generated by running $n$ independent MALA for 10 steps. (b) - (c) The MSE of the different weighting schemes for estimating $E(h(x))$, where $h(x)$ equals $x$, $x^2$, and $\cos(\omega x + b)$, respectively. For $h = \cos(\omega x + b)$ in (c), we draw $\omega \sim \mathcal{N}(0, 1)$ and $b \sim \text{Uniform}([0, 2\pi])$ and average the MSE over 20 random trials.