Stochastic Gradient Descent in NPIV estimation

Anonymous Author(s)
Affiliation
Address
email

1 Problem setup

1.1 Basic definitions

Fix a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Given $X \in L^2(\Omega; \mathbb{X} \subseteq \mathbf{R}^p)$, we define

$$L^2(X) \triangleq \left\{ h : \mathbb{X} \to \mathbf{R} \ : \ \mathbb{E}[h(X)^2] < \infty \right\},$$

that is, $L^2(X) = L^2(\mathbb{X}, \mathcal{B}(\mathbb{X}), \mathbb{P}_X)$[1], a Hilbert space equipped with the inner product $\langle h, g \rangle_{L^2(X)} = \mathbb{E}[h(X)g(X)]$. The regression problem we are interested in has the form

$$Y = h^\star(X) + \varepsilon, \tag{1}$$

where $h^\star \in L^2(X)$ and ε is an integrable r.v. such that $\mathbb{E}[\varepsilon \mid X] \neq 0$. We assume there exists $Z \in L^2(\Omega; \mathbb{Z} \subseteq \mathbf{R}^q)$ such that $Z \not\perp\!\!\!\perp X$, Z influences Y only through X and $\mathbb{E}[\varepsilon \mid Z] = 0$. This variable is called the instrumental variable. The problem consists of estimating $h^\star$ based on independent joint samples from X, Z and Y.

Conditioning (1) in Z, we find

$$\mathbb{E}[Y \mid Z] = \mathbb{E}[h^\star(X) \mid Z]. \tag{2}$$

This motivates us to introduce the operator $\mathcal{T} : L^2(X) \to L^2(Z)$ defined by

$$\mathcal{T}[h](z) \triangleq \mathbb{E}[h(X) \mid Z = z].$$

Clearly $\mathcal{T}$ is linear and, using Jensen's inequality, one may prove that it's bounded. It's also interesting to notice that its adjoint $\mathcal{T}^* : L^2(Z) \to L^2(X)$ satisfies

$$\mathcal{T}^*[g](x) = \mathbb{E}[g(Z) \mid X = x]. \tag{3}$$

Define $r_0 : \mathbb{Z} \to \mathbf{R}$ by $r_0(Z) = \mathbb{E}[Y \mid Z]$. Again by Jensen's inequality, we have $r_0 \in L^2(Z)$, and thus we can rewrite (2) as

$$\mathcal{T}[h^\star] = r_0. \tag{4}$$

Hence, (1) can be formulated as an inverse problem, where we wish to invert the operator $\mathcal{T}$.

Discuss the other implication, that if h satisfies $\mathcal{T}[h] = r_0$, then $h = h^\star$. This is false, but the reason can be connected to the strength of the instrument Z.

1.2 Risk measure

Let $\ell : \mathbf{R} \times \mathbf{R} \to \mathbf{R}_+$ be a pointwise loss function, which, with respect to its second argument, is convex and differentiable. We use the symbol ∂_2 to denote a derivative with respect to the second argument. The example to keep in mind is the quadratic loss function $\ell(y, y') = (y - y')^2$. Given $h \in L^2(X)$, we define the *populational risk* associated with it to be

$$\mathcal{R}(h) \triangleq \mathbb{E}[\ell(r_0(Z), \mathcal{T}[h](Z))].$$

We would like to solve

$$\inf_{h \in \mathcal{F}} \mathcal{R}(h),$$

where $\mathcal{F} \subseteq L^2(X)$ is a closed, convex set such that $h^\star \in \mathcal{F}$.

Assumption

[1] We denote by $\mathbb{P}_X$ the distribution of the r.v. X and by $\mathcal{B}(\mathbb{X})$ the Borel σ-algebra in $\mathbb{X}$.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

2 Gradient computation

We'd like to compute $\nabla\mathcal{R}(h)$ for $h \in L^2(X)$. We start by computing the directional derivative of $\mathcal{R}$ at h in the direction f, denoted by $D\mathcal{R}[h](f)$:

$$\begin{aligned}
D\mathcal{R}[h](f) &= \lim_{\delta\to 0}\frac{1}{\delta}\left[\mathcal{R}(h+\delta f) - \mathcal{R}(f)\right] \\
&= \lim_{\delta\to 0}\frac{1}{\delta}\mathbb{E}\left[\ell(r_0(Z), \mathcal{T}[h+\delta f](Z)) - \ell(r_0(Z), \mathcal{T}[h](Z))\right] \\
&= \lim_{\delta\to 0}\frac{1}{\delta}\mathbb{E}\left[\ell(r_0(Z), \mathcal{T}[h](Z) + \delta\mathcal{T}[f](Z)) - \ell(r_0(Z), \mathcal{T}[h](Z))\right] \\
&= \lim_{\delta\to 0}\frac{1}{\delta}\mathbb{E}\Bigg[\delta\partial_2\ell(r_0(Z), \mathcal{T}[h](Z))\mathcal{T}[f](Z) \\
&\qquad\qquad + \frac{\delta^2}{2}\partial_2^2\ell(r_0(Z), \mathcal{T}[h+\theta f](Z))\mathcal{T}[f](Z)^2\Bigg] \\
&= \mathbb{E}\left[\partial_2\ell(r_0(Z), \mathcal{T}[h](Z))\mathcal{T}[f](Z)\right] \\
&\qquad\qquad + \lim_{\delta\to 0}\mathbb{E}\left[\frac{\delta}{2}\partial_2^2\ell(r_0(Z), \mathcal{T}[h+\theta f](Z))\mathcal{T}[f](Z)^2\right] \\
&= \mathbb{E}\left[\partial_2\ell(r_0(Z), \mathcal{T}[h](Z))\mathcal{T}[f](Z)\right],
\end{aligned}$$

where $\theta \in \mathbf{R}$ is due to Taylor's formula and can be assumed to be inside a fixed interval $(-\theta_0, \theta_0)$, with θ_0 arbitrarily small. We have assumed at the last step that there exists $\theta_0 > 0$ such that

$$\sup_{|\theta|<\theta_0}\mathbb{E}\left[\partial_2^2\ell(r_0(Z), \mathcal{T}[h+\theta f](Z))\mathcal{T}[f](Z)^2\right] < \infty.$$

Assumption

Assumption

This is a mild integrability condition which can be shown to hold in the quadratic case.

We can in fact expand the calculation a bit more, as follows:

$$\begin{aligned}
D\mathcal{R}[h](f) &= \mathbb{E}\left[\partial_2\ell(r_0(Z), \mathcal{T}[h](Z))\mathcal{T}[f](Z)\right] \\
&= \langle \partial_2\ell(r_0(\cdot), \mathcal{T}[h](\cdot)), \mathcal{T}[f]\rangle_{L^2(Z)} \\
&= \langle \mathcal{T}^*[\partial_2\ell(r_0(Z), \mathcal{T}[h](\cdot))], f\rangle_{L^2(X)},
\end{aligned}$$

where we are assuming that $\partial_2\ell(r_0(\cdot), \mathcal{T}[h](\cdot)) \in L^2(Z)$. This shows that $\mathcal{R}$ is Gateux-differentiable, with Gateux derivative at h given by

$$D\mathcal{R}[h] = \mathcal{T}^*[\partial_2\ell(r_0(\cdot), \mathcal{T}[h](\cdot))].$$

Assumption

If we assume[2] that $h \mapsto D\mathcal{R}[h]$ is a continuous mapping from $L^2(Z)$ to $L^2(Z)$, then $\mathcal{R}$ is also Fréchet-differentiable, and both derivatives coincide. Therefore, under this assumption, which we henceforth make, $\nabla\mathcal{R}(h) = \mathcal{T}^*[\partial_2\ell(r_0(\cdot), \mathcal{T}[h](\cdot))]$.

Assumption

Talk about which conditions ℓ can satisfy so that this is continuous.

3 Unbiased estimator of the gradient

We have found that

$$\nabla\mathcal{R}(h)(x) = \mathcal{T}^*[\partial_2\ell(r_0(\cdot), \mathcal{T}[h](\cdot))](x) = \mathbb{E}[\partial_2\ell(r_0(Z), \mathcal{T}[h](Z)) \mid X = x].$$

This turns out to be hard to estimate in practice, as we have two nested conditional expectation operators. Our objective in this section is to find a random element $u_h \in L^2(X)$ such that $\mathbb{E}[u_h(x)] = \nabla\mathcal{R}(h)(x)$, so we can replace $\nabla\mathcal{R}(h)(x)$ by $u_h(x)$ in a gradient descent algorithm, obtaining a stochastic version which will be easier to compute.

Should we discuss this further?

Our strategy to obtain u_h will be to write $\nabla\mathcal{R}(h)(x) = \mathbb{E}[\Phi(x, Z)\partial_2\ell(r_0(Z), \mathcal{T}[h](Z))]$, for some suitable kernel Φ. To ease the notation, define $\xi_h(z) \triangleq \partial_2\ell(r_0(z), \mathcal{T}[h](z))$. Assuming that X and

[2] It is if ℓ is quadratic.

Z have a joint distribution which is absolutely continuous with respect to Lebesgue measure in $\mathbf{R}^{p+q}$, we can write

$$\begin{aligned}
\nabla\mathcal{R}(h)(x) &= \mathbb{E}[\xi_h(Z) \mid X = x] \\
&= \int_{\mathbb{Z}} p(z \mid x)\xi_h(z)\,\mathrm{d}z \\
&= \int_{\mathbb{Z}} p(z)\frac{p(z \mid x)}{p(z)}\xi_h(z)\,\mathrm{d}z \\
&= \mathbb{E}\left[\frac{p(Z \mid x)}{p(Z)}\xi_h(Z)\right].
\end{aligned}$$

Thus, we must take

$$\Phi(x, z) = \frac{p(z \mid x)}{p(z)} = \frac{p(x \mid z)}{p(x)} = \frac{p(x, z)}{p(x)p(z)}.$$

With this choice, setting $u_h(x) = \Phi(x, Z)\xi_h(Z)$ we clearly have $\mathbb{E}[u_h(x)] = \nabla\mathcal{R}(h)(x)$.

Must discuss why $u_h \in L^2(X)$.

4 Algorithm

Having an unbiased estimator of the gradient, we can construct an SGD algorithm for estimating $h^\star$.

Discuss everything we don't know and must estimate.

Comment on exactly what is needed to estimate each unknown (samples from which r.v.'s).

Discuss necessity of discretizing $\mathbb{X}$.

```
Algorithm 1: SGD-NPIV
input : Datasets $\mathcal{D}_{r_0} = \{(y_i, \boldsymbol{z}_i)\} \overset{\text{iid}}{\sim} \mathbb{P}_{YZ}$, $\mathcal{D}_\Phi = \{(\boldsymbol{x}_i, \boldsymbol{z}_i)\} \overset{\text{iid}}{\sim} \mathbb{P}_{XZ}$,
        $\mathcal{D}_{\mathcal{T}} = \{(\boldsymbol{x}_i, \boldsymbol{z}_i)\} \overset{\text{iid}}{\sim} \mathbb{P}_{XZ}$, discretization $\{\boldsymbol{x}_k\}_{k=1}^K$ of $\mathbb{X}$ which contains the observed
        values of $X$, sequence of learning rates $(\alpha_m)_{m=1}^M$.
output: $\left\{\widehat{h}(\boldsymbol{x}_k)\right\}_{k=1}^K$
Compute $\{\widehat{r_0}(\boldsymbol{z}_m; \mathcal{D}_{r_0})\}_{m=1}^M$ ;
Compute $\widehat{\Phi}(\boldsymbol{x}, \boldsymbol{z}; \mathcal{D}_\Phi)$ ;
for $1 \leq m \leq M$ do
    Compute $\widehat{\mathcal{T}[\widehat{h}_{m-1}]}(\boldsymbol{z}_m; \mathcal{D}_{\mathcal{T}})$ ;
    Set $u_m(\boldsymbol{x}_k) = \widehat{\Phi}(\boldsymbol{x}_k, \boldsymbol{z}_m)\partial_2\ell\left(\widehat{r_0}(\boldsymbol{z}_m, \mathcal{D}_{r_0}), \widehat{\mathcal{T}[\widehat{h}_{m-1}]}(\boldsymbol{z}_m; \mathcal{D}_{\mathcal{T}})\right)$   for $1 \leq k \leq K$ ;
    Set $\widehat{h}_m(\boldsymbol{x}_k) = \widehat{h}_{m-1}(\boldsymbol{x}_k) - \alpha_m u_m(\boldsymbol{x}_k)$   for $1 \leq k \leq K$ ;
end
Set $\widehat{h} = \frac{1}{M}\sum_{m=1}^M \widehat{h}_m$ ;
```

An option we have is to project onto the closed, convex, bounded set $\mathcal{F}$ after applying the stochastic gradient, that is, constructing the new estimate as

Should we do this?

$$\widehat{h}_m = P_{\mathcal{F}}\left[\widehat{h}_{m-1} - \alpha_m u_m\right].$$

From what I can see, this would require minor changes to the proof and would justify the assumption that $\widehat{h}_m \in \mathcal{F}$ for all m.

A possible choice for the set $\mathcal{F}$ is

$$\mathcal{F} \triangleq \left\{h \in L^2(X) : \|h\|_\infty \leq M\right\},$$

where $M > 0$ is a constant chosen *a priori*. This set is obviously closed, convex and bounded in the $L^2(X)$ norm. Furthermore, the operator $P_{\mathcal{F}}$ is very easy to compute, as $P_{\mathcal{F}}[h]$ is obtained by cropping h inside $[-M, M]$. More formally,

$$P_{\mathcal{F}}[h] = h^+ \wedge M - h^- \wedge M.$$

5 Proof of convergence

The first problem is proving our sequence of estimates is, in fact, contained in $L^2(X)$. This amounts to proving $u_m \in L^2(X)$ for every m. It's not even immediate why $u_h(x) = \Phi(x, Z)\xi_h(Z)$ (the unbiased gradient when we know r_0, Φ and $\mathcal{T}$) belongs to $L^2(X)$

We'll need to bound the norm of u_m by a constant later in the proof.

After doing this, we check that $\mathcal{R}$ is convex in $\mathcal{F}$: if $h, g \in \mathcal{F}$ and $\lambda \in [0, 1]$, then

$$\begin{aligned}\mathcal{R}(\lambda h + (1-\lambda)g) &= \mathbb{E}[\ell(r_0(Z), \mathcal{T}[\lambda h + (1-\lambda)g](Z))] \\ &= \mathbb{E}[\ell(r_0(Z), \lambda\mathcal{T}[h](Z) + (1-\lambda)\mathcal{T}[g](Z))] \\ &\leq \lambda\mathbb{E}[\ell(r_0(Z), \mathcal{T}[h](Z))] + (1-\lambda)\mathbb{E}[\ell(r_0(Z), \mathcal{T}[g](Z))] \\ &= \lambda\mathcal{R}(h) + (1-\lambda)\mathcal{R}(g).\end{aligned}$$

To lighten the notation, we denote the norm and inner product in $L^2(X)$ by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$, respectively. By the Algorithm 1 procedure, we have

$$\begin{aligned}\frac{1}{2}\left\|\widehat{h}_m - h^\star\right\|^2 &= \frac{1}{2}\left\|\widehat{h}_{m-1} - \alpha_m u_m - h^\star\right\|^2 \\ &= \frac{1}{2}\left\|\widehat{h}_{m-1} - h^\star\right\|^2 - \alpha_m\langle u_m, \widehat{h}_{m-1} - h^\star\rangle + \frac{\alpha_m^2}{2}\|u_m\|^2.\end{aligned}$$

After adding and subtracting $\alpha_m\langle\nabla\mathcal{R}(\widehat{h}_{m-1}), \widehat{h}_{m-1} - h^\star\rangle$, we are left with

$$\frac{1}{2}\left\|\widehat{h}_{m-1} - h^\star\right\|^2 - \alpha_m\langle u_m - \nabla\mathcal{R}(\widehat{h}_{m-1}), \widehat{h}_{m-1} - h^\star\rangle + \frac{\alpha_m^2}{2}\|u_m\|^2 - \alpha_m\langle\nabla\mathcal{R}(\widehat{h}_{m-1}), \widehat{h}_{m-1} - h^\star\rangle.$$

Applying the basic convexity inequality on the last term give us, in total,

$$\begin{aligned}\frac{1}{2}\left\|\widehat{h}_m - h^\star\right\|^2 \leq \frac{1}{2}\left\|\widehat{h}_{m-1} - h^\star\right\|^2 &- \alpha_m\langle u_m - \nabla\mathcal{R}(\widehat{h}_{m-1}), \widehat{h}_{m-1} - h^\star\rangle \\ &+ \frac{\alpha_m^2}{2}\|u_m\|^2 - \alpha_m(\mathcal{R}(\widehat{h}_{m-1}) - \mathcal{R}(h^\star)).\end{aligned}$$

Rearranging terms, we get

$$\begin{aligned}\mathcal{R}(\widehat{h}_{m-1}) - \mathcal{R}(h^\star) \leq \frac{1}{2\alpha_m}&\left(\left\|\widehat{h}_{m-1} - h^\star\right\|^2 - \left\|\widehat{h}_m - h^\star\right\|^2\right) \\ &+ \frac{\alpha_m}{2}\|u_m\|^2 - \langle u_m - \nabla\mathcal{R}(\widehat{h}_{m-1}), \widehat{h}_{m-1} - h^\star\rangle.\end{aligned}$$

Finally, summing over $1 \leq m \leq M$ leads to

$$\begin{aligned}\sum_{n=1}^{M}\left[\mathcal{R}(\widehat{h}_{m-1}) - \mathcal{R}(h^\star)\right] \leq \sum_{m=1}^{M}&\frac{1}{2\alpha_m}\left(\left\|\widehat{h}_{m-1} - h^\star\right\|^2 - \left\|\widehat{h}_m - h^\star\right\|^2\right) \\ &+ \sum_{m=1}^{M}\frac{\alpha_m}{2}\|u_m\|^2 \\ &- \sum_{m=1}^{M}\langle u_m - \nabla\mathcal{R}(\widehat{h}_{m-1}), \widehat{h}_{m-1} - h^\star\rangle.\end{aligned}$$

5.1 Option 1: averaging with respect to everything

We then treat each term (summation) separately:

- The first term is bounded using the assumption that $\operatorname{diam}\mathcal{F} = D < \infty$.

Assumption

- The bound on the second term depends on bounding $\mathbb{E}\left[\|u_m\|_{L^2(X)}^2\right]$ by a constant independent of m or M.
- The third term must vanish because of the unbiasedness of u_m, but we don't know that our u_m is unbiased, and it may very well not be.

First term By assumption, we have $\operatorname{diam}\mathcal{F} = D < \infty$. Hence

$$\begin{aligned}\sum_{m=1}^{M}\frac{1}{2\alpha_m}\left(\left\|\widehat{h}_{m-1}-h^\star\right\|^2-\left\|\widehat{h}_m-h^\star\right\|^2\right) &= \sum_{m=2}^{M}\left(\frac{1}{2\alpha_m}-\frac{1}{2\alpha_{m-1}}\right)\left\|\widehat{h}_{m-1}-h^\star\right\|^2 \\ &\quad + \frac{1}{2\alpha_1}\left\|\widehat{h}_0-h^\star\right\|^2-\frac{1}{2\alpha_M}\left\|\widehat{h}_M-h^\star\right\|^2 \\ &\le \sum_{m=2}^{M}\left(\frac{1}{2\alpha_m}-\frac{1}{2\alpha_{m-1}}\right)D^2+\frac{1}{2\alpha_1}D^2=\frac{D^2}{2\alpha_M}.\end{aligned}$$

Second term Define $\mathcal{D}$ to be the set of all observed data, that is, all of the variables in $\mathcal{D}_\Phi, \mathcal{D}_{r_0}, \mathcal{D}_\mathcal{T}$ and $\{\boldsymbol{z}_i\}_{i=1}^M$. Let's evaluate $\mathbb{E}\left[\|u_m\|^2\right]$:

$$\mathbb{E}\left[\|u_m\|^2\right] = \mathbb{E}_\mathcal{D}\left[\mathbb{E}_X\left[\widehat{\Phi}(X,\boldsymbol{z}_m;\mathcal{D}_\Phi)^2\partial_2\ell\left(\widehat{r_0}(\boldsymbol{z}_m;\mathcal{D}_{r_0}),\widehat{\mathcal{T}[\widehat{h}_{m-1}]}(\boldsymbol{z}_m;\mathcal{D}_\mathcal{T})\right)^2\right]\right],$$

where the second expectation is with respect to a copy of X which is independent of $\mathcal{D}$. Continuing:

$$\begin{aligned}&\mathbb{E}_\mathcal{D}\left[\mathbb{E}_X\left[\widehat{\Phi}(X,\boldsymbol{z}_m;\mathcal{D}_\Phi)^2\partial_2\ell\left(\widehat{r_0}(\boldsymbol{z}_m;\mathcal{D}_{r_0}),\widehat{\mathcal{T}[\widehat{h}_{m-1}]}(\boldsymbol{z}_m;\mathcal{D}_\mathcal{T})\right)^2\right]\right] \\ &\qquad = \mathbb{E}_\mathcal{D}\left[\partial_2\ell\left(\widehat{r_0}(\boldsymbol{z}_m;\mathcal{D}_{r_0}),\widehat{\mathcal{T}[\widehat{h}_{m-1}]}(\boldsymbol{z}_m;\mathcal{D}_\mathcal{T})\right)^2\mathbb{E}_X\left[\widehat{\Phi}(X,\boldsymbol{z}_m;\mathcal{D}_\Phi)^2\right]\right].\end{aligned}$$

If ℓ is quadratic, we have

$$\begin{aligned}&\mathbb{E}_\mathcal{D}\left[\partial_2\ell\left(\widehat{r_0}(\boldsymbol{z}_m;\mathcal{D}_{r_0}),\widehat{\mathcal{T}[\widehat{h}_{m-1}]}(\boldsymbol{z}_m;\mathcal{D}_\mathcal{T})\right)^2\mathbb{E}_X\left[\widehat{\Phi}(X,\boldsymbol{z}_m;\mathcal{D}_\Phi)^2\right]\right] \\ &\qquad = \mathbb{E}_\mathcal{D}\left[\left(\widehat{\mathcal{T}[\widehat{h}_{m-1}]}(\boldsymbol{z}_m;\mathcal{D}_\mathcal{T})-\widehat{r_0}(\boldsymbol{z}_m;\mathcal{D}_{r_0})\right)^2\mathbb{E}_X\left[\widehat{\Phi}(X,\boldsymbol{z}_m;\mathcal{D}_\Phi)^2\right]\right].\end{aligned}$$

Find a way to finish this bound. Maybe switch the order of expectations? First in X and then in $\mathcal{D}$.

Third term We have to work with

$$\mathbb{E}\left[\langle u_m-\nabla\mathcal{R}(\widehat{h}_{m-1}),\widehat{h}_{m-1}-h^\star\rangle\right].$$

The strategy employed on the SIP paper was to condition on the σ-algebra generated by the training samples observed up until iteration iteration $m-1$. In our case, that would be $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_{m-1}$. The problem which arises is that we no longer have measurability of $\widehat{h}_{m-1}$ with respect to this σ-algebra, as it depends on the datasets $\mathcal{D}_\Phi,\mathcal{D}_\mathcal{T},\mathcal{D}_{r_0}$, used to estimate $\widehat{\Phi},\widehat{\mathcal{T}}$ and $\widehat{r}_0$ in an offline manner. The other option would be to condition on more things, namely the σ-algebra generated by $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_{m-1},\mathcal{D}_\Phi,\mathcal{D}_\mathcal{T},\mathcal{D}_{r_0}$. We gain measurability of $\widehat{h}_{m-1}$, but we are no longer integrating out $\mathcal{D}_\Phi,\mathcal{D}_\mathcal{T},\mathcal{D}_{r_0}$, which is needed to use some sort of unbiasedness of the estimators $\widehat{\Phi},\widehat{\mathcal{T}},\widehat{r}_0$. Let's try the latter and see what we end up with. Define $\mathcal{D}_{m-1}\triangleq\mathcal{D}_\Phi\cup\mathcal{D}_\mathcal{T}\cup\mathcal{D}_{r_0}\cup\{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_{m-1}\}$. Then,

$$\begin{aligned}&\mathbb{E}_\mathcal{D}\left[\langle u_m-\nabla\mathcal{R}(\widehat{h}_{m-1}),\widehat{h}_{m-1}-h^\star\rangle\right] \\ &\qquad = \mathbb{E}_{\mathcal{D}_{m-1}}\left[\mathbb{E}\left[\langle u_m-\nabla\mathcal{R}(\widehat{h}_{m-1}),\widehat{h}_{m-1}-h^\star\rangle \mid \mathcal{D}_{m-1}\right]\right] \\ &\qquad = \mathbb{E}_{\mathcal{D}_{m-1}}\left[\langle \mathbb{E}\left[u_m\mid\mathcal{D}_{m-1}\right]-\nabla\mathcal{R}(\widehat{h}_{m-1}),\widehat{h}_{m-1}-h^\star\rangle\right] \qquad \text{(Justify this properly.).}\end{aligned}$$

Now we must evaluate the expression inside the inner product. We restrict ourselves to the quadratic case. Define $\psi(z)=\mathcal{T}[\widehat{h}_{m-1}](z)-r_0(z)$ and $\widehat{\psi}(z;\mathcal{D}_\text{proj})=\widehat{\mathcal{T}[\widehat{h}_{m-1}]}(z;\mathcal{D}_\mathcal{T})-\widehat{r}_0(z;\mathcal{D}_{r_0})$, where

$\mathcal{D}_{\text{proj}} = \mathcal{D}_{\mathcal{T}} \cup \mathcal{D}_{r_0}$. Then:

$$\begin{aligned}
&\mathbb{E}\left[u_m(x) \mid \mathcal{D}_{m-1}\right] - \nabla\mathcal{R}(\widehat{h}_{m-1})(x) \\
&\qquad = \mathbb{E}\left[\widehat{\Phi}(x, Z; \mathcal{D}_\Phi)\widehat{\psi}(Z; \mathcal{D}_{\text{proj}}) \mid \mathcal{D}_{m-1}\right] - \mathbb{E}_Z\left[\Phi(x, Z)\psi(Z)\right] \\
&\qquad = \mathbb{E}_Z\left[\widehat{\Phi}(x, Z; \mathcal{D}_\Phi)\widehat{\psi}(Z; \mathcal{D}_{\text{proj}})\right] - \mathbb{E}_Z\left[\Phi(x, Z)\psi(Z)\right] \\
&\qquad = \mathbb{E}_Z\left[\left(\widehat{\Phi}(x, Z; \mathcal{D}_\Phi) - \Phi(x, Z)\right)\widehat{\psi}(Z; \mathcal{D}_{\text{proj}}) + \left(\widehat{\psi}(Z; \mathcal{D}_{\text{proj}}) - \psi(Z)\right)\Phi(x, Z)\right] \\
&\qquad = \left\langle \widehat{\Phi}(x, \cdot; \mathcal{D}_\Phi) - \Phi(x, \cdot), \widehat{\psi}(\cdot; \mathcal{D}_{\text{proj}})\right\rangle_{L^2(Z)} + \left\langle \widehat{\psi}(\cdot; \mathcal{D}_{\text{proj}}) - \psi(\cdot), \Phi(x, \cdot)\right\rangle_{L^2(Z)} \\
&\qquad \leq \left\|\widehat{\Phi}(x, \cdot; \mathcal{D}_\Phi) - \Phi(x, \cdot)\right\|_{L^2(Z)} \left\|\widehat{\psi}(\cdot; \mathcal{D}_{\text{proj}})\right\|_{L^2(Z)} \\
&\qquad\qquad + \left\|\widehat{\psi}(\cdot; \mathcal{D}_{\text{proj}}) - \psi(\cdot)\right\|_{L^2(Z)} \left\|\Phi(x, \cdot)\right\|_{L^2(Z)}.
\end{aligned}$$

5.2 Option 2: conditioning on $\mathcal{D}_{\Phi,\mathcal{T},r_0}$ and averaging with respect to $z_1, \ldots, z_M$

With this strategy, our aim isn't to get a convergence guarantee analogous to that of the SIP paper, but some sort of "given $\mathcal{D}_{\Phi,\mathcal{T},r_0}$, this inequality is true with high probability" guarantee.

We similarly treat each of the three terms separately.

First term It is bounded in the exact same way as before, since this bound is deterministic and does not involve expectations.

Second term We are fixing the offline data $\mathcal{D}_{\Phi,\mathcal{T},r_0}$ and averaging with respect to the other samples of the instrumental variable. Therefore, what we wish to compute is

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{z}_{1:M}}\left[\|u_m\|^2 \mid \mathcal{D}_{\Phi,\mathcal{T},r_0}\right] &= \mathbb{E}_{\boldsymbol{z}_{1:m}}\left[\mathbb{E}_X\left[\widehat{\Phi}(X, \boldsymbol{z}_m)^2 \partial_2\ell\left(\widehat{r_0}(\boldsymbol{z}_m), \widehat{\mathcal{T}}[\widehat{h}_{m-1}](\boldsymbol{z}_m)\right)^2\right] \;\middle|\; \mathcal{D}_{\Phi,\mathcal{T},r_0}\right] \\
&= \mathbb{E}_X\left[\mathbb{E}_{\boldsymbol{z}_{1:m}}\left[\widehat{\Phi}(X, \boldsymbol{z}_m)^2 \partial_2\ell\left(\widehat{r_0}(\boldsymbol{z}_m), \widehat{\mathcal{T}}[\widehat{h}_{m-1}](\boldsymbol{z}_m)\right)^2 \;\middle|\; \mathcal{D}_{\Phi,\mathcal{T},r_0}\right]\right].
\end{aligned}$$

Since $\boldsymbol{z}_{1:m}$ is independent from $\mathcal{D}_{\Phi,\mathcal{T},r_0}$, this is equal to

$$\mathbb{E}_X\left[\mathbb{E}_{\boldsymbol{z}_{1:m}}\left[\widehat{\Phi}(X, \boldsymbol{z}_m)^2 \partial_2\ell\left(\widehat{r_0}(\boldsymbol{z}_m), \widehat{\mathcal{T}}[\widehat{h}_{m-1}](\boldsymbol{z}_m)\right)^2\right]\right].$$

Reversing back the expectations, we get

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}_{1:m}}\left[\mathbb{E}_X\left[\widehat{\Phi}(X, \boldsymbol{z}_m)^2 \partial_2\ell\left(\widehat{r_0}(\boldsymbol{z}_m), \widehat{\mathcal{T}}[\widehat{h}_{m-1}](\boldsymbol{z}_m)\right)^2\right]\right] \\
&\qquad = \mathbb{E}_{\boldsymbol{z}_{1:m}}\left[\mathbb{E}_X\left[\widehat{\Phi}(X, \boldsymbol{z}_m)^2\right] \partial_2\ell\left(\widehat{r_0}(\boldsymbol{z}_m), \widehat{\mathcal{T}}[\widehat{h}_{m-1}](\boldsymbol{z}_m)\right)^2\right].
\end{aligned}$$

Now we use Assumption 14.5.1 in [1], which states that

$$\sup_{w \in \mathbb{W}} k(\boldsymbol{w}, \boldsymbol{w}) \leq 1,$$

where $\mathbb{W} = \mathbb{X} \times \mathbb{Z}$, $\boldsymbol{w} = (\boldsymbol{x}, \boldsymbol{z})$ and $k : \mathbb{W} \times \mathbb{W} \to \mathbf{R}$ is the kernel corresponding to the RKHS used to estimate Φ, which we denote by $\mathcal{R}_{\mathbb{W}}$. This assumption implies

$$\begin{aligned}
\widehat{\Phi}(\boldsymbol{w}) = \langle \widehat{\Phi}, k(\boldsymbol{w}, \cdot)\rangle_{\mathcal{R}_{\mathbb{W}}} &\leq \left\|\widehat{\Phi}\right\|_{\mathcal{R}_{\mathbb{W}}} \|k(\boldsymbol{w}, \cdot)\|_{\mathcal{R}_{\mathbb{W}}} = \left\|\widehat{\Phi}\right\|_{\mathcal{R}_{\mathbb{W}}} \sqrt{\langle k(\boldsymbol{w}, \cdot), k(\boldsymbol{w}, \cdot)\rangle} = \\
&= \left\|\widehat{\Phi}\right\|_{\mathcal{R}_{\mathbb{W}}} \sqrt{k(\boldsymbol{w}, \boldsymbol{w})} \leq \|\Phi\|_{\mathcal{R}_{\mathbb{W}}}
\end{aligned}$$

for all $\boldsymbol{w} \in \mathbb{W}$. Therefore,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}_{1:m}}\left[\mathbb{E}_X\left[\widehat{\Phi}(X,\boldsymbol{z}_m)^2\right]\partial_2\ell\left(\widehat{r_0}(\boldsymbol{z}_m),\widehat{\mathcal{T}}[\widehat{h}_{m-1}](\boldsymbol{z}_m)\right)^2\right]\\
&\qquad\le \mathbb{E}_{\boldsymbol{z}_{1:m}}\left[\mathbb{E}_X\left[\left\|\widehat{\Phi}\right\|^2_{\mathcal{R}_{\mathbb{W}}}\right]\partial_2\ell\left(\widehat{r_0}(\boldsymbol{z}_m),\widehat{\mathcal{T}}[\widehat{h}_{m-1}](\boldsymbol{z}_m)\right)^2\right]\\
&\qquad= \left\|\widehat{\Phi}\right\|^2_{\mathcal{R}_{\mathbb{W}}}\mathbb{E}_{\boldsymbol{z}_{1:m}}\left[\partial_2\ell\left(\widehat{r_0}(\boldsymbol{z}_m),\widehat{\mathcal{T}}[\widehat{h}_{m-1}](\boldsymbol{z}_m)\right)^2\right].
\end{aligned}$$

To bound the expectation, we assume the loss is quadratic and then

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}_{1:m}}\left[\left(\widehat{\mathcal{T}}[\widehat{h}_{m-1}](\boldsymbol{z}_m)-\widehat{r_0}(\boldsymbol{z}_m)\right)^2\right]\\
&\qquad= \mathbb{E}_{\boldsymbol{z}_{1:m}}\Bigg[\Bigg(\left(\widehat{\mathcal{T}}[\widehat{h}_{m-1}](\boldsymbol{z}_m)-\mathcal{T}[\widehat{h}_{m-1}](\boldsymbol{z}_m)\right)+\left(r_0(\boldsymbol{z}_m)-\widehat{r_0}(\boldsymbol{z}_m)\right)\\
&\qquad\qquad\qquad+\left(\mathcal{T}[\widehat{h}_{m-1}](\boldsymbol{z}_m)-r_0(\boldsymbol{z}_m)\right)\Bigg)^2\Bigg]\\
&\qquad\le 3\mathbb{E}_{\boldsymbol{z}_{1:m}}\Bigg[\left(\widehat{\mathcal{T}}[\widehat{h}_{m-1}](\boldsymbol{z}_m)-\mathcal{T}[\widehat{h}_{m-1}](\boldsymbol{z}_m)\right)^2+\left(r_0(\boldsymbol{z}_m)-\widehat{r_0}(\boldsymbol{z}_m)\right)^2\\
&\qquad\qquad\qquad+\left(\mathcal{T}[\widehat{h}_{m-1}](\boldsymbol{z}_m)-r_0(\boldsymbol{z}_m)\right)^2\Bigg]\\
&\qquad= 3\Bigg\{\mathbb{E}_{\boldsymbol{z}_{1:m-1}}\left[\left\|(\widehat{\mathcal{T}}-\mathcal{T})[\widehat{h}_{m-1}]\right\|^2_{L^2(\mathbb{Z})}\right]+\mathbb{E}_{\boldsymbol{z}_{1:m}}\left[\|r_0-\widehat{r_0}\|^2_{L^2(\mathbb{Z})}\right]\\
&\qquad\qquad\qquad+\mathbb{E}_{\boldsymbol{z}_{1:m-1}}\left[\left\|\mathcal{T}[\widehat{h}_{m-1}]-r_0\right\|^2_{L^2(\mathbb{Z})}\right]\Bigg\}.
\end{aligned}$$

We treat each part of this expression separately. Firstly,

$$\left\|(\widehat{\mathcal{T}}-\mathcal{T})[\widehat{h}_{m-1}]\right\|^2_{L^2(\mathbb{Z})}\le\left\|\widehat{\mathcal{T}}-\mathcal{T}\right\|^2_{\mathrm{op}}\left\|\widehat{h}_{m-1}\right\|^2_{L^2(\mathbb{X})}\le M^2\left\|\widehat{\mathcal{T}}-\mathcal{T}\right\|^2_{\mathrm{op}}.$$

We leave the second term as $\|r_0-\widehat{r_0}\|^2_{L^2(\mathbb{Z})}$. Finally, for the third term, we have

$$\begin{aligned}
\left\|\mathcal{T}[\widehat{h}_{m-1}]-r_0\right\|^2_{L^2(\mathbb{Z})} &= \mathbb{E}_Z\left[\left(\mathcal{T}[\widehat{h}_{m-1}](Z)-r_0(Z)\right)^2\right]\\
&= \mathbb{E}_Z\left[\left(\mathbb{E}\left[\widehat{h}_{m-1}(X)-Y\mid Z\right]\right)^2\right]\\
&\le \mathbb{E}_{(X,Y)}\left[\left(\widehat{h}_{m-1}(X)-Y\right)^2\right]\\
&\le 2\left(\mathbb{E}_X\left[\widehat{h}_{m-1}(X)^2\right]+\mathbb{E}\left[Y^2\right]\right)\\
&= 2\left(\left\|\widehat{h}_{m-1}\right\|^2_{L^2(\mathbb{X})}+\mathbb{E}\left[Y^2\right]\right)\\
&\le 2\left(M^2+\mathbb{E}\left[Y^2\right]\right).
\end{aligned}$$

Putting everything together, what we conclude is

$$\mathbb{E}_{\boldsymbol{z}_{1:m}}\left[\|u_m\|^2_{L^2(\mathbb{X})}\mid\mathcal{D}_{\Phi,\mathcal{T},r_0}\right]\le 3\left\|\widehat{\Phi}\right\|^2_{\mathcal{R}_{\mathbb{W}}}\left(M^2\left\|\widehat{\mathcal{T}}-\mathcal{T}\right\|^2_{\mathrm{op}}+\|r_0-\widehat{r_0}\|^2_{L^2(\mathbb{Z})}+2\left(M^2+\mathbb{E}[Y^2]\right)\right).$$

Third term

References

[1] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012. DOI: 10.1017/CBO9781139035613.