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 $\mathbb{X}$.

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 $D\mathcal{R}[h](f)$:

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

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

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

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, [Assumption] with Gateux derivative at h given by

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

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 [Assumption] Fréchet-differentiable, and both derivatives coincide. Therefore, under this assumption, which we henceforth make, $\nabla \mathcal{R}(h) = \mathcal{T}^*[\partial_2 \ell(r_0(\cdot), \mathcal{T}[h](\cdot))]$. [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 [Should we discuss this further?] stochastic version which will be easier to compute.

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.

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

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

46 Thus, we must take

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

47 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

49 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 $\{\widehat{r_0}(\boldsymbol{z}_m; \mathcal{D}_{r_0})\}_{m=1}^M$;

50 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 $\widehat{h} = \frac{1}{M} \sum_{m=1}^M \widehat{h}_m$;

---

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

> Should we do this?

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

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

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

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

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

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

3

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, the symbols $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$, when written without a subscript to specify which space they refer to, will act as the norm and inner product, respectively, of $L^2(X)$. 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

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

We then treat each of the three terms in the RHS of the inequality above separately:

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

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

4

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

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

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

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

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

76 Reversing back the expectations, we get

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

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

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

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

78 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
79 to estimate Φ, which we denote by $\mathcal{R}_\mathbb{W}$. This assumption implies

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

$$= \left\|\widehat{\Phi}\right\|_{\mathcal{R}_\mathbb{W}} \sqrt{k(\boldsymbol{w},\boldsymbol{w})} \leq \|\Phi\|_{\mathcal{R}_\mathbb{W}}$$

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

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

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

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

81 To bound the expectation, we assume the loss is quadratic and then [Assumption]

$$\mathbb{E}_{\boldsymbol{z}_{1:m}}\left[\left(\widehat{\mathcal{T}}[\widehat{h}_{m-1}](\boldsymbol{z}_m) - \widehat{r_0}(\boldsymbol{z}_m)\right)^2\right]$$

$$= \mathbb{E}_{\boldsymbol{z}_{1:m}}\left[\left(\left(\widehat{\mathcal{T}}[\widehat{h}_{m-1}](\boldsymbol{z}_m) - \mathcal{T}[\widehat{h}_{m-1}](\boldsymbol{z}_m)\right) + (r_0(\boldsymbol{z}_m) - \widehat{r_0}(\boldsymbol{z}_m))\right.\right.$$

$$\left.\left. + \left(\mathcal{T}[\widehat{h}_{m-1}](\boldsymbol{z}_m) - r_0(\boldsymbol{z}_m)\right)\right)^2\right]$$

$$\leq 3\mathbb{E}_{\boldsymbol{z}_{1:m}}\left[\left(\widehat{\mathcal{T}}[\widehat{h}_{m-1}](\boldsymbol{z}_m) - \mathcal{T}[\widehat{h}_{m-1}](\boldsymbol{z}_m)\right)^2 + (r_0(\boldsymbol{z}_m) - \widehat{r_0}(\boldsymbol{z}_m))^2\right.$$

$$\left. + \left(\mathcal{T}[\widehat{h}_{m-1}](\boldsymbol{z}_m) - r_0(\boldsymbol{z}_m)\right)^2\right]$$

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

$$\left. + \mathbb{E}_{\boldsymbol{z}_{1:m-1}}\left[\left\|\mathcal{T}[\widehat{h}_{m-1}] - r_0\right\|_{L^2(\mathbb{Z})}^2\right]\right\}.$$

5

82 We treat each part of this expression separately. Firstly,
$$\left\|(\widehat{\mathcal{T}} - \mathcal{T})[\widehat{h}_{m-1}]\right\|_{L^2(\mathbb{Z})}^2 \le \left\|\widehat{\mathcal{T}} - \mathcal{T}\right\|_{\mathrm{op}}^2 \left\|\widehat{h}_{m-1}\right\|_{L^2(\mathbb{X})}^2 \le M^2 \left\|\widehat{\mathcal{T}} - \mathcal{T}\right\|_{\mathrm{op}}^2.$$

83 We leave the second part as $\|r_0 - \widehat{r_0}\|_{L^2(\mathbb{Z})}^2$. Finally, for the third part, we have
$$
\begin{aligned}
\left\|\mathcal{T}[\widehat{h}_{m-1}] - r_0\right\|_{L^2(\mathbb{Z})}^2 &= \mathbb{E}_Z\left[\left(\mathcal{T}[\widehat{h}_{m-1}](Z) - r_0(Z)\right)^2\right] \\
&= \mathbb{E}_Z\left[\left(\mathbb{E}\left[\widehat{h}_{m-1}(X) - Y \mid Z\right]\right)^2\right] \\
&\le \mathbb{E}_{(X,Y)}\left[\left(\widehat{h}_{m-1}(X) - Y\right)^2\right] \\
&\le 2\left(\mathbb{E}_X\left[\widehat{h}_{m-1}(X)^2\right] + \mathbb{E}\left[Y^2\right]\right) \\
&= 2\left(\left\|\widehat{h}_{m-1}\right\|_{L^2(\mathbb{X})}^2 + \mathbb{E}\left[Y^2\right]\right) \\
&\le 2\left(M^2 + \mathbb{E}\left[Y^2\right]\right).
\end{aligned}
$$

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

85 We still have to use convergence results for $\widehat{\mathcal{T}}$ and $\widehat{r_0}$ to finish this bound. It doesn't need to be good,
86 we only need to bound this by something which remains bounded as $|\mathcal{D}_{\Phi,\mathcal{T},r_0}|$ and the number of
87 iterations grow.

88 **Third term**

89 Our goal is to open up the inner product and make explicit the estimation errors of our model's
90 different components, like we did before. Here, we define $\Psi_m(Z) \triangleq \partial_2 \ell(r_0(Z), \mathcal{T}[\widehat{h}_{m-1}](Z))$. The
91 hat version $\widehat{\Psi}_m$ is defined accordingly, replacing r_0 and $\mathcal{T}$ by their estimators.

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}_{1:m}}\left[\langle\nabla\mathcal{R}(\widehat{h}_{m-1}) - u_m, \widehat{h}_{m-1} - h^\star\rangle \mid \mathcal{D}_{\Phi,\mathcal{T},r_0}\right] \\
&= \mathbb{E}_{\boldsymbol{z}_{1:m}}\left[\langle\nabla\mathcal{R}(\widehat{h}_{m-1}) - u_m, \widehat{h}_{m-1} - h^\star\rangle\right] && (\boldsymbol{z}_{1:m} \perp\!\!\!\perp \mathcal{D}_{\Phi,\mathcal{T},r_0}) \\
&= \mathbb{E}_{\boldsymbol{z}_{1:m-1}}\left[\mathbb{E}_{\boldsymbol{z}_m}\left[\langle\nabla\mathcal{R}(\widehat{h}_{m-1}) - u_m, \widehat{h}_{m-1} - h^\star\rangle\right]\right] && (\boldsymbol{z}_m \perp\!\!\!\perp \boldsymbol{z}_{1:m-1}) \\
&= \mathbb{E}_{\boldsymbol{z}_{1:m-1}}\left[\langle\nabla\mathcal{R}(\widehat{h}_{m-1}) - \mathbb{E}_{\boldsymbol{z}_m}[u_m], \widehat{h}_{m-1} - h^\star\rangle\right] \\
&\le \mathbb{E}_{\boldsymbol{z}_{1:m-1}}\left[\left\|\nabla\mathcal{R}(\widehat{h}_{m-1}) - \mathbb{E}_{\boldsymbol{z}_m}[u_m]\right\|\left\|\widehat{h}_{m-1} - h^\star\right\|\right] \\
&\le D\mathbb{E}_{\boldsymbol{z}_{1:m-1}}\left[\left\|\nabla\mathcal{R}(\widehat{h}_{m-1}) - \mathbb{E}_{\boldsymbol{z}_m}[u_m]\right\|\right] && (\operatorname{diam}\mathcal{F} < D) \\
&\le D\mathbb{E}_{\boldsymbol{z}_{1:m-1}}\left[\mathbb{E}_X\left[\left(\nabla\mathcal{R}(\widehat{h}_{m-1})(X) - \mathbb{E}_{\boldsymbol{z}_m}[u_m]\right)^2\right]\right]^{\frac{1}{2}} && (\text{Jensen}) \\
&= D\mathbb{E}_{\boldsymbol{z}_{1:m-1}}\left[\mathbb{E}_X\left[\left(\mathbb{E}_Z[\Phi(X,Z)\Psi_m(Z)]\right.\right.\right. \\
&\qquad\qquad\qquad\qquad \left.\left.\left. - \mathbb{E}_{\boldsymbol{z}_m}\left[\widehat{\Phi}(X,\boldsymbol{z}_m)\widehat{\Psi}_m(\boldsymbol{z}_m)\right]\right)^2\right]\right]^{\frac{1}{2}} \\
&= D\mathbb{E}_{\boldsymbol{z}_{1:m-1}}\left[\mathbb{E}_X\left[\mathbb{E}_Z\left[\Phi(X,Z)\Psi_m(Z) - \widehat{\Phi}(X,Z)\widehat{\Psi}_m(Z)\right]^2\right]\right]^{\frac{1}{2}} && (Z \overset{\text{iid}}{\sim} \boldsymbol{z}_m) \\
&= D\mathbb{E}_{\boldsymbol{z}_{1:m-1}}\left[\mathbb{E}_X\left[\mathbb{E}_Z\left[\Psi_m(Z)\left(\Phi(X,Z) - \widehat{\Phi}(X,Z)\right)\right.\right.\right. \\
&\qquad\qquad\qquad\qquad \left.\left.\left. + \widehat{\Phi}(X,Z)\left(\Psi_m(Z) - \widehat{\Psi}_m(Z)\right)\right]^2\right]\right]^{\frac{1}{2}}
\end{aligned}
$$

$$\leq D\mathbb{E}_{\mathbf{z}_{1:m-1}}\left[\mathbb{E}_X\left[\left(\|\Psi_m\|_{L^2(Z)}\left\|\Phi(X,\cdot)-\widehat{\Phi}(X,\cdot)\right\|_{L^2(Z)}\right.\right.\right.$$

$$\left.\left.\left.+\left\|\widehat{\Phi}(X,\cdot)\right\|_{L^2(Z)}\left\|\Psi_m-\widehat{\Psi}_m\right\|_{L^2(Z)}\right)^2\right]\right]^{\frac{1}{2}}$$

$$\leq\sqrt{2}D\mathbb{E}_{\mathbf{z}_{1:m-1}}\left[\mathbb{E}_X\left[\|\Psi_m\|^2_{L^2(Z)}\left\|\Phi(X,\cdot)-\widehat{\Phi}(X,\cdot)\right\|^2_{L^2(Z)}\right.\right.$$

$$\left.\left.+\left\|\widehat{\Phi}(X,\cdot)\right\|^2_{L^2(Z)}\left\|\Psi_m-\widehat{\Psi}_m\right\|^2_{L^2(Z)}\right]\right]^{\frac{1}{2}}$$

$$=\sqrt{2}D\left(\mathbb{E}_{\mathbf{z}_{1:m-1}}\left[\|\Psi_m\|^2_{L^2(Z)}\mathbb{E}_X\left[\left\|\Phi(X,\cdot)-\widehat{\Phi}(X,\cdot)\right\|^2_{L^2(Z)}\right]\right]\right.$$

$$\left.+\mathbb{E}_{\mathbf{z}_{1:m-1}}\left[\left\|\Psi_m-\widehat{\Psi}_m\right\|^2_{L^2(Z)}\mathbb{E}_X\left[\left\|\widehat{\Phi}(X,\cdot)\right\|^2_{L^2(Z)}\right]\right]\right)^{\frac{1}{2}}$$

$$=\sqrt{2}D\left(\mathbb{E}_X\left[\left\|\Phi(X,\cdot)-\widehat{\Phi}(X,\cdot)\right\|^2_{L^2(Z)}\right]\mathbb{E}_{\mathbf{z}_{1:m-1}}\left[\|\Psi_m\|^2_{L^2(Z)}\right]\right.$$

$$\left.+\mathbb{E}_X\left[\left\|\widehat{\Phi}(X,\cdot)\right\|^2_{L^2(Z)}\right]\mathbb{E}_{\mathbf{z}_{1:m-1}}\left[\left\|\Psi_m-\widehat{\Psi}_m\right\|^2_{L^2(Z)}\right]\right)^{\frac{1}{2}}$$

$$=:\sqrt{2}D(A+B)^{\frac{1}{2}}.$$

We proceed to analyze each term separately:

- To bound A, first notice that

$$\mathbb{E}_X\left[\left\|\Phi(X,\cdot)-\widehat{\Phi}(X,\cdot)\right\|^2\right]=\mathbb{E}_X\left[\mathbb{E}_Z\left[\left(\Phi(X,Z)-\widehat{\Phi}(X,Z)\right)^2\right]\right]=\left\|\Phi-\widehat{\Phi}\right\|^2_{L^2(X\otimes Z)},$$

  where $L^2(X\otimes Z)$ is the space of square integrable functions with respect to the measure induced by independent copies of X and Z. If we estimate $\widehat{\Phi}$ using the uLSIF algorithm described in [1], under some regularity conditions, and decreasing the regularization parameter according to a specific rate, we have the following estimate:

<!-- margin note: Create section describing how we are estimating each term. -->

$$\left\|\Phi-\widehat{\Phi}\right\|^2_{L^2(X\otimes Z)}=\mathcal{O}_p\left(\left(\frac{\log|\mathcal{D}_\Phi|}{|\mathcal{D}_\Phi|}\right)^{\frac{2}{2+\gamma}}\right).$$

  Furthermore, we can bound $\|\Psi_m\|^2_{L^2(Z)}$ as follows:

$$\|\Phi_m\|^2_{L^2(Z)}=\left\|r_0-\mathcal{T}[\widehat{h}_{m-1}]\right\|^2_{L^2(Z)}$$

$$\leq 2\left(\|r_0\|^2_{L^2(Z)}+\left\|\mathcal{T}[\widehat{h}_{m-1}]\right\|^2_{L^2(Z)}\right)$$

$$\leq 2\left(\mathbb{E}[Y^2]+\|\mathcal{T}\|^2_{\mathrm{op}}\left\|\widehat{h}_{m-1}\right\|^2_{L^2(Z)}\right)$$

$$\leq 2\left(\mathbb{E}[Y^2]+M^2\right)\qquad\qquad(\|\mathcal{T}\|_{\mathrm{op}}\leq 1).$$

  In total, what we have is

$$A=\mathbb{E}_X\left[\left\|\Phi(X,\cdot)-\widehat{\Phi}(X,\cdot)\right\|^2_{L^2(Z)}\right]\mathbb{E}_{\mathbf{z}_{1:m-1}}\left[\|\Psi_m\|^2_{L^2(Z)}\right]$$

$$\leq\left\|\Phi-\widehat{\Phi}\right\|^2_{L^2(Z)}\cdot 2(\mathbb{E}[Y^2]+M^2)$$

$$=\mathcal{O}_p\left(\left(\frac{\log|\mathcal{D}_\Phi|}{|\mathcal{D}_\Phi|}\right)^{\frac{2}{2+\gamma}}\right).$$

- To bound B, notice that, by Assumption 14.15 of [1], we have

$$\mathbb{E}_X\left[\left\|\widehat{\Phi}(X,\cdot)\right\|_{L^2(Z)}^2\right] = \mathbb{E}_X\left[\mathbb{E}_Z\left[\widehat{\Phi}(X,Z)^2\right]\right] \leq \left\|\widehat{\Phi}\right\|_{\mathcal{R}_\mathbb{W}}^2.$$

We still need to bound this norm somehow.

Furthermore, we also have

$$\begin{aligned}
\left\|\Phi_m - \widehat{\Phi}_m\right\|_{L^2(Z)}^2 &= \left\|\left(\mathcal{T}[\widehat{h}_{m-1}] - r_0\right) - \left(\widehat{\mathcal{T}}[\widehat{h}_{m-1}] - \widehat{r_0}\right)\right\|_{L^2(Z)}^2 \\
&= \left\|\left(\mathcal{T}[\widehat{h}_{m-1}] - \widehat{\mathcal{T}}[\widehat{h}_{m-1}]\right) - (r_0 - \widehat{r_0})\right\|_{L^2(Z)}^2 \\
&\leq 2\left(\left\|\mathcal{T}[\widehat{h}_{m-1}] - \widehat{\mathcal{T}}[\widehat{h}_{m-1}]\right\|_{L^2(Z)}^2 + \|r_0 - \widehat{r_0}\|_{L^2(Z)}^2\right) \\
&\leq 2\left(\left\|\mathcal{T} - \widehat{\mathcal{T}}\right\|_{op}^2\left\|\widehat{h}_{m-1}\right\|_{L^2(Z)}^2 + \|r_0 - \widehat{r_0}\|_{L^2(Z)}^2\right) \\
&\leq 2\left(M^2\left\|\mathcal{T} - \widehat{\mathcal{T}}\right\|_{op}^2 + \|r_0 - \widehat{r_0}\|_{L^2(Z)}^2\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
B &= \mathbb{E}_X\left[\left\|\widehat{\Phi}(X,\cdot)\right\|_{L^2(Z)}^2\right]\mathbb{E}_{\boldsymbol{z}_{1:m-1}}\left[\left\|\Psi_m - \widehat{\Psi}_m\right\|_{L^2(Z)}^2\right] \\
&\leq \left\|\widehat{\Phi}\right\|_{\mathcal{R}_\mathbb{W}}^2 \mathbb{E}_{\boldsymbol{z}_{1:m-1}}\left[2\left(M^2\left\|\mathcal{T} - \widehat{\mathcal{T}}\right\|_{op}^2 + \|r_0 - \widehat{r_0}\|_{L^2(Z)}^2\right)\right] \\
&= 2\left\|\widehat{\Phi}\right\|_{\mathcal{R}_\mathbb{W}}^2\left(M^2\left\|\mathcal{T} - \widehat{\mathcal{T}}\right\|_{op}^2 + \|r_0 - \widehat{r_0}\|_{L^2(Z)}^2\right).
\end{aligned}$$

What's left to do:

- Bound $\left\|\widehat{\Phi}\right\|_{\mathcal{R}_\mathbb{W}}$

- Use some estimate on $\left\|\mathcal{T} - \widehat{\mathcal{T}}\right\|_{op}$ (Adapt notation and setup in the KIV paper)

- Find way to estimate r_0 which gives estimate on $\|r_0 - \widehat{r_0}\|_{L^2(Z)}$.

For the rest of the paper:

- Create section which describes, in detail, how we are estimating Φ, $\mathcal{T}$ and r_0, lists all the references, states the main convergence theorems and lists all of the assumptions that are being made.

- Adapt the algorithm section to use the KIV first stage, which directly estimates $\mathcal{T}$.

- Find better letter for either the number of iterations or the upper bound for the set $\mathcal{F}$. Right now, both are being denoted by the letter M.

References

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