Stochastic Gradient Descent in NPIV estimation

Anonymous Author(s)
Affiliation
Address
email

Abstract

TODO

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 $Z \not\perp\!\!\!\perp X$ 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

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

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

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

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

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

[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 $\mathbb{X}$.

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 subspace such that $h^\star \in \mathcal{F}$.

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

Assumption

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, with Gateux derivative at h given by

Assumption

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

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

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 $L^2(X)$.

Do this.

[2] It is if ℓ is quadratic.

After doing this, the first steps in the proof are the same as in the previous paper. We show that $\mathcal{R}$ is convex in $\mathcal{F}$ and then simple algebraic manipulation allows us to write

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

We then treat each term 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 $\mathbb{E}\left[\|u_m\|^2_{L^2(X)}\right]$ by a constant.
- 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.