
Stochastic Gradient Descent in NPIV estimation

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 TODO

2 1 Problem setup

3 1.1 Basic definitions

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

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

5 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 $\langle h, g \rangle_{L^2(X)} =$
6 $\mathbb{E}[h(X)g(X)]$. The regression problem we are interested in has the form

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

7 where $h^* \in L^2(X)$ and ε is an integrable r.v. such that $\mathbb{E}[\varepsilon \mid X] \neq 0$. We assume there exists
8 $Z \in L^2(\Omega; \mathbb{Z} \subseteq \mathbf{R}^q)$ such that $Z \not\perp X$ and $\mathbb{E}[\varepsilon \mid Z] = 0$. This variable is called the instrumental
9 variable. The problem consists of estimating h^* based on independent joint samples from X, Z and
10 Y .

11 Conditioning (1) in Z , we find

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

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

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

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

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

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

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

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

18 1.2 Risk measure

19 Let $\ell : \mathbf{R} \times \mathbf{R} \rightarrow \mathbf{R}_+$ be a pointwise loss function, which, with respect to its second argument, is
20 convex and differentiable. We use the symbol ∂_2 to denote a derivative with respect to the second

¹We 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} .

21 argument. The example to keep in mind is the quadratic loss function $\ell(y, y') = (y - y')^2$. Given
 22 $h \in L^2(X)$, we define the *populational risk* associated with it to be

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

23 We would like to solve

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

24 where $\mathcal{F} \subseteq L^2(X)$ is a subspace such that $h^* \in \mathcal{F}$.

25 2 Gradient computation

26 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}
 27 at h in the direction f , denoted by $D\mathcal{R}[h](f)$:

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

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

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

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

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

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

32 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,
 33 with Gateux derivative at h given by

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

34 If we assume² that $h \mapsto D\mathcal{R}[h]$ is a continuous mapping from $L^2(Z)$ to $L^2(Z)$, then \mathcal{R} is also
 35 Fréchet-differentiable, and both derivatives coincide. Therefore, under this assumption, which we
 36 henceforth make, $\nabla \mathcal{R}(h) = \mathcal{T}^*[\partial_2 \ell(r_0(\cdot), \mathcal{T}[h](\cdot))]$.

37 3 Unbiased estimator of the gradient

38 We have found that

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

Assumption

Assumption

Assumption

Assumption

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

39 This turns out to be hard to estimate in practice, as we have two nested conditional expectation
 40 operators. Our objective in this section is to find a random element $u_h \in L^2(X)$ such that $\mathbb{E}[u_h(x)] =$
 41 $\nabla \mathcal{R}(h)(x)$, so we can replace $\nabla \mathcal{R}(h)(x)$ by $u_h(x)$ in a gradient descent algorithm, obtaining a
 42 stochastic version which will be easier to compute.

Should we discuss this further?

43 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
 44 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
 45 Z have a joint distribution which is absolutely continuous with respect to Lebesgue measure in \mathbb{R}^{p+q} ,
 46 we can write

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

47 Thus, we must take

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

48 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 $u_h \in L^2(X)$.

49 4 Algorithm

50 Having an unbiased estimator of the gradient, we can construct an SGD algorithm for estimating h^* .

Algorithm 1: SGD-NPIV

input : Datasets $\mathcal{D}_{r_0} = \{(y_i, z_i)\} \stackrel{\text{iid}}{\sim} \mathbb{P}_{YZ}$, $\mathcal{D}_{\Phi} = \{(\mathbf{x}_i, z_i)\} \stackrel{\text{iid}}{\sim} \mathbb{P}_{XZ}$,
 $\mathcal{D}_{\mathcal{T}} = \{(\mathbf{x}_i, z_i)\} \stackrel{\text{iid}}{\sim} \mathbb{P}_{XZ}$, discretization $\{\mathbf{x}_k\}_{k=1}^K$ of \mathbb{X} which contains the observed
 values of X , sequence of learning rates $(\alpha_m)_{m=1}^M$.

output : $\{\hat{h}(\mathbf{x}_k)\}_{k=1}^K$

Compute $\{\hat{r}_0(z_m; \mathcal{D}_{r_0})\}_{m=1}^M$;

51 Compute $\hat{\Phi}(\mathbf{x}, z; \mathcal{D}_{\Phi})$;

for $1 \leq m \leq M$ **do**

 Compute $\mathcal{T}[\hat{h}_{m-1}](z_m; \mathcal{D}_{\mathcal{T}})$;

 Set $u_m(\mathbf{x}_k) = \hat{\Phi}(\mathbf{x}_k, z_m) \partial_2 \ell \left(\hat{r}_0(z_m, \mathcal{D}_{r_0}), \mathcal{T}[\hat{h}_{m-1}](z_m; \mathcal{D}_{\mathcal{T}}) \right)$ for $1 \leq k \leq K$;

 Set $\hat{h}_m(\mathbf{x}_k) = \hat{h}_{m-1}(\mathbf{x}_k) - \alpha_m u_m(\mathbf{x}_k)$ for $1 \leq k \leq K$;

end

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

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} .

52 5 Proof of convergence

53 The first problem is proving our sequence of estimates is, in fact, contained in $L^2(X)$. This amounts
 54 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
 55 unbiased gradient when we know r_0 , Φ and \mathcal{T}) belongs to $L^2(X)$.

Do this.

²It is if ℓ is quadratic.

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

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

58 We then treat each term separately:

- 59 • The first term is bounded using the assumption that $\text{diam } \mathcal{F} = D < \infty$. Assumption
- 60 • The bound on the second term depends on bounding $\mathbb{E} [\|u_m\|_{L^2(X)}^2]$ by a constant indepen-
- 61 dent of m .
- 62 • The third term must vanish because of the unbiasedness of u_m , but we don't know that our
- 63 u_m is unbiased, and it may very well not be.

64 **Second term** Define \mathcal{D} to be the set of all observed data, that is, all of the variables in $\mathcal{D}_\Phi, \mathcal{D}_{r_0}, \mathcal{D}_\mathcal{T}$
 65 and $\{z_i\}_{i=1}^M$. Let's evaluate $\mathbb{E} [\|u_m\|_{L^2(X)}^2]$:

$$\mathbb{E} [\|u_m\|_{L^2(X)}^2] = \mathbb{E}_\mathcal{D} \left[\mathbb{E}_X \left[\hat{\Phi}(X, z_m; \mathcal{D}_\Phi)^2 \partial_2 \left(\hat{r}_0(z_m; \mathcal{D}_{r_0}), \mathcal{T}[\widehat{h_{m-1}}](z_m; \mathcal{D}_\mathcal{T}) \right)^2 \right] \right],$$

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

$$\begin{aligned} \mathbb{E}_\mathcal{D} \left[\mathbb{E}_X \left[\hat{\Phi}(X, z_m; \mathcal{D}_\Phi)^2 \partial_2 \left(\hat{r}_0(z_m; \mathcal{D}_{r_0}), \mathcal{T}[\widehat{h_{m-1}}](z_m; \mathcal{D}_\mathcal{T}) \right)^2 \right] \right] \\ = \mathbb{E}_\mathcal{D} \left[\partial_2 \left(\hat{r}_0(z_m; \mathcal{D}_{r_0}), \mathcal{T}[\widehat{h_{m-1}}](z_m; \mathcal{D}_\mathcal{T}) \right)^2 \mathbb{E}_X [\hat{\Phi}(X, z_m; \mathcal{D}_\Phi)^2] \right]. \end{aligned}$$

67 If ℓ is quadratic, we have

$$\begin{aligned} \mathbb{E}_\mathcal{D} \left[\partial_2 \left(\hat{r}_0(z_m; \mathcal{D}_{r_0}), \mathcal{T}[\widehat{h_{m-1}}](z_m; \mathcal{D}_\mathcal{T}) \right)^2 \mathbb{E}_X [\hat{\Phi}(X, z_m; \mathcal{D}_\Phi)^2] \right] \\ = \mathbb{E}_\mathcal{D} \left[\left(\mathcal{T}[\widehat{h_{m-1}}](z_m; \mathcal{D}_\mathcal{T}) - \hat{r}_0(z_m; \mathcal{D}_{r_0}) \right)^2 \mathbb{E}_X [\hat{\Phi}(X, z_m; \mathcal{D}_\Phi)^2] \right]. \end{aligned}$$