
Stochastic Gradient Descent in NPIV estimation

Anonymous Author(s)

Affiliation

Address

email

1 Binary response models

2 We want to be able to employ the same risk minimization procedure:

$$\arg \min_{h \in \mathcal{F}} \mathcal{R}(h) = \arg \min_{h \in \mathcal{F}} \mathbb{E}_Z [\ell(r_0(Z), \mathcal{P}[h](Z))]. \quad (1)$$

3 Let's see what data generating procedure makes this possible. Firstly, let

$$Y \mid X, \varepsilon \sim \text{Bernoulli}(\sigma(h^*(X) + \varepsilon)), \quad (2)$$

4 where σ is the logistic function, $\mathbb{E}[\varepsilon \mid X] \neq 0$ and $\mathbb{E}[\varepsilon \mid Z] = 0$. For (1) to make sense, we'd like
5 $r_0(Z) = \mathbb{E}[Y \mid Z]$ and $\mathcal{P}[h^*](Z) = \mathbb{E}[h^*(X) \mid Z]$ to be close according to a suitable loss function ℓ ,
6 at least close enough so that h^* is a solution to (1). Let's see if this is the case under (2):

$$\mathbb{E}[Y \mid Z] = \mathbb{P}[Y = 1 \mid Z],$$

7 Assuming (2), we may compute this conditioning on X and ε and then integrating them out:

$$\begin{aligned} \mathbb{P}[Y = 1 \mid Z = z] &= \int_{\mathcal{X} \times \mathbf{R}} \mathbb{P}[Y = 1 \mid Z = z, X = x, \varepsilon = e] p_{X, \varepsilon \mid Z}(x, e \mid z) \, dx d\varepsilon \\ &= \int_{\mathcal{X} \times \mathbf{R}} \sigma(h^*(x) + e) p_{X, \varepsilon \mid Z}(x, e \mid z) \, dx d\varepsilon \\ &= \mathbb{E}[\sigma(h^*(X) + \varepsilon) \mid Z = z]. \end{aligned}$$

8 There are a two main problems here. The first one is that ε appears inside σ and, hence, does not
9 vanish after conditioning on $Z = z$. I cannot think of a way to remove it without assuming known
10 the distribution of ε given X , which is prohibitive. The second problem is that, even if there was no
11 ε , the expectation is outside the function σ . In order for (1) to work under (2), we'd like set

$$\ell(y, y') = \text{BCE}(y, \sigma(y')),$$

12 where BCE is the binary cross entropy loss function:

$$\text{BCE}(y, p) = -[y \log p + (1 - y) \log(1 - p)].$$

13 That is, we'd like to have $\sigma(\mathbb{E}[h(X) \mid Z])$ inside $\mathcal{R}(h)$, instead of $\mathbb{E}[\sigma(h(X)) \mid Z]$.

14 The second option is to set

$$Y = \mathbf{1}[h^*(X) + \varepsilon > 0]. \quad (3)$$

15 Here, we have

$$\mathbb{E}[Y \mid Z = z] = \mathbb{P}[h^*(X) + \varepsilon > 0 \mid Z = z]. \quad (4)$$

16 To try to make this lead somewhere, let's define $\eta = h^*(X) - \mathbb{E}[h^*(X) \mid Z] + \varepsilon$, so that

$$Y = \mathbf{1}[\mathbb{E}[h^*(X) \mid Z] + \eta > 0]$$

17 and $\mathbb{E}[\eta \mid Z] = 0$. Let $t(Z) = \mathbb{E}[h^*(X) \mid Z]$. This implies

$$\begin{aligned}\mathbb{E}[Y \mid Z] &= \mathbb{P}[t(Z) + \eta > 0 \mid Z] \\ &= 1 - F_{\eta|Z}(-t(Z)).\end{aligned}$$

18 Hence, we have

$$t(Z) = -F_{\eta|Z}^{-1}(r_0(Z) - 1).$$

19 Or, equivalently:

$$\mathbb{E}[h^*(X) \mid Z] = -F_{\eta|Z}^{-1}(\mathbb{E}[Y \mid Z] - 1).$$

20 This looks promising: If we assume to know the conditional distribution of η given Z , we have a
21 couple of options. We can minimize

$$\text{BCE}(r_0(Z), 1 - F_{\eta|Z}(-\mathbb{E}[h(X) \mid Z])),$$

22 or

$$\left(\mathbb{E}[h(X) \mid Z] + F_{\eta|Z}^{-1}(r_0(Z) - 1)\right)^2.$$

23 This assumption was used on the paper “Nonparametric Instrumental Variable Estimation of Binary
24 Response Models”, by P. L. Florens, from where I took the ideas for these calculations.

25 In an unpublished version of that paper, they assume that $\eta = \frac{1}{\zeta(Z)}v$, where $v \mid Z \sim$
26 $\text{KnownDistribution}(0, \sigma_v^2)$. This implies

$$\begin{aligned}\mathbb{E}[Y \mid Z] &= \mathbb{P}[t(Z) + \eta > 0 \mid Z] \\ &= \mathbb{P}[t(Z) + \frac{v}{\zeta(Z)} > 0 \mid Z] \\ &= \mathbb{P}[v > -t(Z)\zeta(Z) \mid Z] \\ &= 1 - F_v(-t(Z)\zeta(Z)) \\ &\triangleq 1 - F_v(-\gamma(Z)).\end{aligned}$$

27 Equivalently, this means that

$$\gamma(Z) = -F_v^{-1}(1 - \mathbb{E}[Y \mid Z]),$$

28 where $\gamma(Z) = \mathbb{E}[h^*(X) \mid Z]\zeta(Z)$. They proceed to use γ to estimate r_0 (this involves splitting Z
29 into two parts and is the main contribution in their article) and then use this estimate of r_0 to estimate
30 h^* through Tikhonov regularization.

31 However, on the published version, the authors assume that η is *independent* of Z , which is good for
32 us.