Optimal Categorical Instrumental Variables

Thomas Wiemann
University of Chicago

June 27, 2023

Introduction

Instrumental variables often results in high-variance estimators

- ▷ In practice: Researchers use multiple instruments (e.g., interactions)
- ▷ "Optimal" IVs (Amemiya, 1974; Chamberlain, 1987; Newey, 1990)

Problem when # instruments is large relative to sample size

- ▷ Overfit in the first stage $\Rightarrow \tau_{\text{TSLS}}$ biased
- ▷ LIML estimators consistent w/ many IVs (e.g., Bekker, 1994)
- ▷ But: Not weakly causal estimands w/ unobserved heterogeneity (Kolesár, 2013)

Belloni et al. (2012): Lasso-based nonparametric estimation of optimal IV

- ▷ But not a universal solution: Approximate sparsity
- ▷ Angrist and Frandsen (2022): Little benefit in calibrated simulations

Contribution

This paper: Semiparametric efficiency w/ *almost* many *categorical* IVs.

- ▷ Regime: # categories *almost* grows at the sample rate
- ▷ Key assumption: $\exists$ few *latent* categories w/ same first-stage fit
- ▷ Allows for mapping IVs to Optimal IVs at exponential rate

Key advantages:

- ▷ Regularization assumption is economically meaningful
- ▷ Robust to small categories & achieves efficiency bound (same as LIML)
- ▷ Admits weakly causal interpretation under misspecification (unlike LIML)

Literature

Literature:

1. Many instruments: Bekker (1994); Angrist and Krueger (1995); Chamberlain and Imbens (2004); Chao and Swanson (2005); Hausman et al. (2012); ...

2. Optimal instruments: Amemiya (1974); Chamberlain (1987); Newey (1990); Donald and Newey (2001); Belloni et al. (2012); Carrasco (2012); ...

3. Group-fixed effects: Hahn and Moon (2010); Bonhomme and Manresa (2015); Su et al. (2016); Bonhomme et al. (2022); ...

Outline

1. Setup
2. Estimation & Inference
3. Monte Carlo Simulation
4. Application: Returns to Schooling

Outline

1. **Setup**

2. Estimation & Inference

3. Monte Carlo Simulation

4. Application: Returns to Schooling

Setup

Data generating process: P_n

P_n is defined by the law of the random vector

$$W \equiv (Y, D, Z, U), \quad \operatorname{supp} W \subset R^4$$

- $Y \equiv$ outcome
- $D \equiv$ endogenous variable of interest
- $Z \equiv$ instrument
- $U \equiv$ structural residual

Allow P_n to change with the sample size n

- Asymptotics that better approximate finite sample behavior
- Importantly: Will allow $|\operatorname{supp} Z| \to \infty$ as $n \to \infty$

Subsequent assumptions characterize P_n uniformly over n

Identification

I consider linear IV under mean independence:

Assumption 1

$\exists \tau_0 \in \mathbb{R}: \; Y = D\tau_0 + U, \; E[U|Z] = 0.$

Assumption 1 implies

$$E[(Y - \tau_0 D)(m_0(Z) - E[m_0(Z)])] = 0, \quad \text{w/ } m_0(z) \equiv E[D|Z = z] \quad (1)$$

Assumption 2

$Var(m_0(Z))$ is bounded away from zero.

Assumptions 1-2 imply the moment solution:

$$\tau_0 = \frac{E[(Y - E[Y])(m_0(Z) - E[D])]}{E[(D - E[D])(m_0(Z) - E[D])]} \quad (2)$$

Infeasible Sample Analogue Estimator

Moment solution (2) holds for any $f : Cov(D, f(Z)) \neq 0$

- Why focus on $m_0(z) = E[D|Z = z]$?

Consider an i.i.d. sample $\{(Y_i, D_i, Z_i)\}_{i=1}^n$ from P_n

Moment solution suggests the estimator

$$\hat{\tau}_n^* = \frac{\frac{1}{n}\sum_{i=1}^n \left(Y_i - \bar{Y}_n\right)\left(m_0(Z_i) - \bar{D}_n\right)}{\frac{1}{n}\sum_{i=1}^n \left(D_i - \bar{D}_n\right)\left(m_0(Z_i) - \bar{D}_n\right)}$$

$m_0(Z_i)$ is the "optimal" instrument (Amemiya, 1974):

- $\hat{\tau}_n^*$ achieves efficiency bound (under homoskedasticity)

Categorical Instrumental Variables

m_0 not (generally) known:

- $\hat{\tau}_n^*$ is generally infeasible
- Need to estimate optimal instruments

This paper focuses on *categorical* instruments Z:

- $\forall z \in \operatorname{supp} Z, \Pr(Z = z) > 0$
- Estimator for $m_0(z)$ simply $\frac{1}{N_z} \sum_{i:Z_i=z} D_i$

To approximate settings with few observations per category:

- $\Pr(Z = z) \to 0$ as $n \to \infty$.

(Almost) Many Categorical Instruments

When $\Pr(Z = z) = o(n^{-0.5})$

- TSLS estimator not $\sqrt{n}$ normal details

When $\Pr(Z = z) = o(n^{-1})$

- LIML is $\sqrt{n}$ normal (e.g., Bekker and Van der Ploeg, 2005)

I consider the slightly less demanding setting to prove optimality:

Assumption 3

$\forall z \in \operatorname{supp} Z, \exists \lambda_z \in (0, 1] : \Pr(Z = z) n^{1-\lambda_z} \to a_z > 0.$

- LIML is semiparametrically efficient
  (Donald and Newey, 2001; Bekker and Van der Ploeg, 2005)
- CIV benefit: Admits weakly causal interpretation details

Optimal Instrument with Fixed Support

Key regularization assumption:

Assumption 4

$\exists K_0 \in \mathbb{N} : |\operatorname{supp} E[D|Z]| = K_0.$

Implies existence of latent categorical variable with fixed support

For every $n \in \mathbb{N}$, exists partition $(\mathcal{Z}_g)_{g=1}^{K_0}$ of $\operatorname{supp} Z$ such that

$$\forall g \in \{1, \ldots, K_0\}, \quad m_0(z') = m_0(z), \quad \forall z', z \in \mathcal{Z}_g$$

Estimation assumes K_0 is known

- Can be estimated under additional assumptions details

Example: Returns to Education

Angrist and Krueger (1991):

- Returns to schooling for male Americans born 30-40s
- IV: Quarter-of-birth $\times$ Year-of-birth $\times$ Place-of-birth
- 1530 indicator instruments in the first stage
- Key motivation for weak & many IV literature (e.g., Bound et al., 1995; Angrist and Krueger, 1995; Angrist et al., 1999; Hansen et al., 2008; Angrist and Frandsen, 2022; Mikusheva and Sun, 2022)

Instrument idea:

- QOB affects schooling due to mandatory attendance laws
- Interaction w/ YOB $\times$ POB b/c laws change across time & space

Is a student born in a particular quarter constraint / not constraint?

- $K = 1530$ but $K_0 = 2$!

Outline

1. Setup
2. **Estimation & Inference**
3. Monte Carlo Simulation
4. Application: Returns to Schooling

Categorical Instrumental Variable Estimator

Finite support assumption motivates the Categorical IV estimator (CIV):

$$\hat{\tau}_n = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}_n\right)\left(\hat{m}_n(Z_i) - \bar{D}_n\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_n(Z_i) - \bar{D}_n\right)^2}, \tag{3}$$

where $\hat{m}_n(Z_i)$ is an estimator for $m_0(Z_i)$ defined by

$$\hat{m}_n = \underset{\substack{m\colon \operatorname{supp} Z \to \mathcal{M} \\ |m(\operatorname{supp} Z)| = K_0}}{\arg\min} \sum_{i=1}^{n} \left(D_i - m(Z_i)\right)^2 \tag{4}$$

Assumption 5

$\operatorname{supp} E[D|Z] \subset \mathcal{M}$, and $\mathcal{M} \subset \mathbb{R}$ is compact.

Estimator (4) implemented using K_0-Means

- ▷ Adapted from Bonhomme and Manresa (2015)

Additional Assumptions

Define the CEF residual:

$$V \equiv D - E[D|Z]$$

Assumptions 6-7 place tail restrictions on first and second stage errors

Assumption 6

$\exists L < \infty$ such that $E\left[U^4\right] \leq L$ and $E\left[V^4\right] \leq L$.

Assumption 7

$\exists b_1, b_2 : \Pr(|V| > v) \leq \exp\left\{1 - \left(\frac{v}{b_1}\right)^{b_2}\right\}, \forall v > 0.$

Additional Assumptions (Contd.)

Assumptions 8-9 ensure the optimal instrument is well-separated

Assumption 8

$\exists c > 0 : (d_z - \tilde{d}_z)^2 \geq c, \forall d_z \neq \tilde{d}_z \in \operatorname{supp} E[D|Z]$.

Assumption 9

$\exists \xi > 0 : \Pr(E[D|Z] = d_z) > \xi, \forall d_z \in \operatorname{supp} E[D|Z]$.

Assumption 10 is the standard i.i.d. sampling assumption

Assumption 10

The data is an i.i.d. sample $\{(Y_i, D_i, Z_i)\}_{i=1}^n$ from P_n.

Main Theorem

Theorem 1

Let assumptions 1-10 hold. Then, as $n \to \infty$,

$$\sqrt{n}\left(\hat{\tau}_n - \tau_0\right)/\sigma \xrightarrow{d} N\left(0, 1\right),$$

where $\sigma = \sqrt{Var(m_0(Z)U)}/Var(m_0(Z))$. If in addition, U is homoskedastic, then $\hat{\tau}_n$ is semiparametrically efficient for estimating τ_0.

Device: Exponential misclassification probabilities in first stage Proof sketch

The result continues to hold when σ is consistently estimated:

$$\hat{\sigma}_n \equiv \sqrt{\frac{1}{n}\sum_{i=1}^{n} \hat{m}_n(Z_i)^2 \left(Y_i - D_i\hat{\tau}_n\right)^2} \Bigg/ \left(\frac{1}{n}\sum_{i=1}^{n} \hat{m}_n(Z_i)^2\right)$$

Outline

1. Setup
2. Estimation & Inference
3. **Monte Carlo Simulation**
4. Application: Returns to Schooling

Monte Carlo Simulation

Simple DGP:

$$
\begin{aligned}
Y_i &= D_i \tau_0 + U_i \\
D_i &= m_0(Z_i) + V_i
\end{aligned}
$$

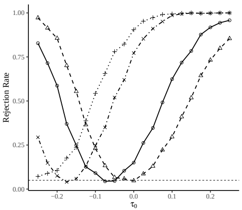
where

- Z_i takes values in $\{1, \ldots, 50\}$ and $E[V_i|Z_i] = 0$
- Each category in the sample has equal observations n_z
- $m_(z) = \frac{-p}{2}$ for $z \leq 25$
- $m(z) = \frac{p}{2}$ for $z > 25$

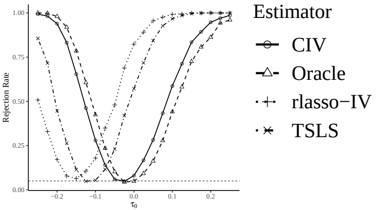
$$
Cov(U_i, V_i | Z_i = z) = \begin{bmatrix} \sigma_U^2(z) & \frac{1}{2}\sigma_U(z)\sigma_V(z) \\ \frac{1}{2}\sigma_U(z)\sigma_V(z) & \sigma_V^2(z) \end{bmatrix}
$$

where $\sigma_U(z)$ and $\sigma_V(z)$ are independent draws from a uniform $U(\frac{1}{2}, \frac{3}{2})$.

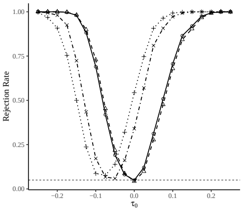
Power Curves ($K_0 = 2$, $p = 1$)



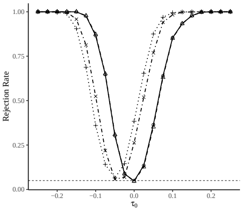
(a) $n_z = 30$



(b) $n_z = 50$

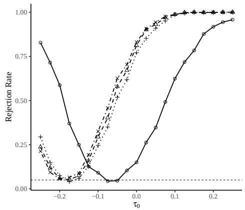


(c) $n_z = 100$

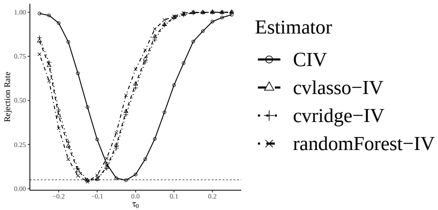


(d) $n_z = 150$

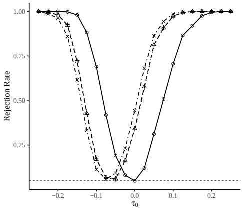
Additional Power Curves ($K_0 = 2$, $p = 1$)



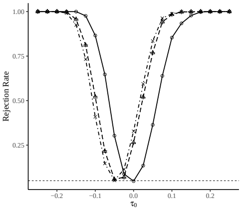
(a) $n_z = 30$



(b) $n_z = 50$

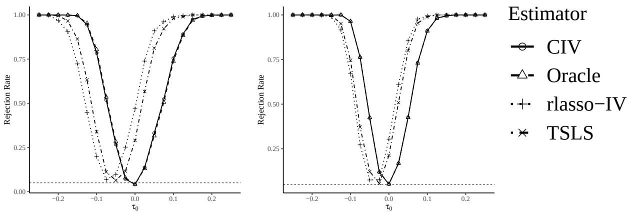


(c) $n_z = 100$

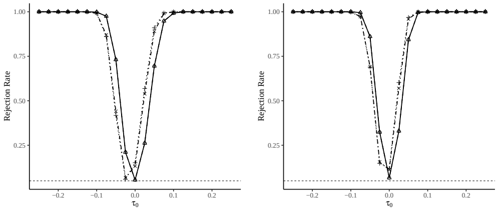


(d) $n_z = 150$

Power Curves ($K_0 = 2$, $p = 2$)



(a) $n_z = 30$ (b) $n_z = 50$



(c) $n_z = 100$ (d) $n_z = 150$

Outline

1. Setup
2. Estimation & Inference
3. Monte Carlo Simulation
4. **Application: Returns to Schooling**

Estimating Returns to Schooling: Revisited

Table 1: Results on Returns to Schooling

$n =$		32,950	98,852	167,754	296,558	329,509
CIV ($K_0 = 2$)	Mean $\hat{\tau}_n$	0.070	0.072	0.074	0.078	0.078
	Mean $se(\hat{\tau}_n)$	0.010	0.009	0.009	0.008	0.008
	Std. Dev. $\hat{\tau}_n$	0.008	0.008	0.006	0.004	-
CIV ($K_0 = 3$)	Mean $\hat{\tau}_n$	0.069	0.069	0.074	0.074	0.074
	Mean $se(\hat{\tau}_n)$	0.035	0.368	0.018	0.060	0.060
	Std. Dev. $\hat{\tau}_n$	0.037	0.137	0.024	0.087	-
TSLS	Mean $\hat{\tau}_n$	0.067	0.068	0.069	0.071	0.071
	Mean $se(\hat{\tau}_n)$	0.005	0.005	0.005	0.005	0.005
	Std. Dev. $\hat{\tau}_n$	0.005	0.005	0.004	0.002	-
OLS	Mean $\hat{\tau}_n$	0.067	0.067	0.067	0.067	0.067
	Mean $se(\hat{\tau}_n)$	0.001	0.001	0.001	0.000	0.000
	Std. Dev. $\hat{\tau}_n$	0.001	0.001	0.000	0.000	-

Estimating Returns to Schooling: Revisited (Contd.)

Table 2: Additional Results on Returns to Schooling

$n =$		32,950	98,852	167,754	296,558	329,509
CIV ($K_0 = 2$)	Mean $\hat{\tau}_n$	0.070	0.072	0.074	0.078	0.078
	Mean $se(\hat{\tau}_n)$	0.010	0.009	0.009	0.008	0.008
	Std. Dev. $\hat{\tau}_n$	0.008	0.008	0.006	0.004	-
rlasso-IV-1	Mean $\hat{\tau}_n$	0.128	0.085	0.086	0.086	0.086
	Mean $se(\hat{\tau}_n)$	0.019	0.037	0.035	0.027	0.025
	Std. Dev. $\hat{\tau}_n$	0.037	0.032	0.025	0.009	-
rlasso-IV-2	Mean $\hat{\tau}_n$	0.098	0.046	-	-	-
	Mean $se(\hat{\tau}_n)$	0.043	0.035	-	-	-
	Std. Dev. $\hat{\tau}_n$	0.077	NA	-	-	-
LIML	Mean $\hat{\tau}_n$	0.127	0.128	0.080	0.102	0.102
	Mean $se(\hat{\tau}_n)$	0.067	0.033	0.024	0.016	0.014
	Std. Dev. $\hat{\tau}_n$	1.886	0.676	0.710	0.020	-

Conclusion

This paper:

- Categorical IV w/ few observations per category
- Propose new CIV estimator
- Conditions for first-order oracle equivalence of CIV
- Application to returns to schooling

References I

Amemiya, T. (1974). Multivariate regression and simultaneous equation models when the dependent variables are truncated normal. *Econometrica*, pages 999–1012.

Angrist, J. D. and Frandsen, B. (2022). Machine labor. *Journal of Labor Economics*, 40(S1):S97–S140.

Angrist, J. D., Imbens, G. W., and Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67.

Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4):979–1014.

Angrist, J. D. and Krueger, A. B. (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, 13(2):225–235.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica: Journal of the Econometric Society*, pages 657–681.

Bekker, P. A. and Van der Ploeg, J. (2005). Instrumental variable estimation based on grouped data. *Statistica Neerlandica*, 59(3):239–267.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

References II

Bonhomme, S., Lamadon, T., and Manresa, E. (2022). Discretizing unobserved heterogeneity. *Econometrica*, 90(2):625–643.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450.

Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics*, 170(2):383–398.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334.

Chamberlain, G. and Imbens, G. (2004). Random effects estimators with many instrumental variables. *Econometrica*, 72(1):295–306.

Chao, J. C. and Swanson, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica*, 73(5):1673–1692.

Donald, S. G. and Newey, W. K. (2001). Choosing the number of instruments. *Econometrica*, 69(5):1161–1191.

Hahn, J. and Moon, H. R. (2010). Panel data models with finite number of multiple equilibria. *Econometric Theory*, 26(3):863–881.

References III

Hansen, C., Hausman, J., and Newey, W. (2008). Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, 26(4):398–422.

Hausman, J. A., Newey, W. K., Woutersen, T., Chao, J. C., and Swanson, N. R. (2012). Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics*, 3(2):211–255.

Kolesár, M. (2013). Estimation in an instrumental variables model with treatment effect heterogeneity. Working Paper.

Mikusheva, A. and Sun, L. (2022). Inference with many weak instruments. *Review of Economic Studies*, 89(5):2663–2686.

Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica*, pages 809–837.

Su, L., Shi, Z., and Phillips, P. C. (2016). Identifying latent structures in panel data. *Econometrica*, 84(6):2215–2264.

Properties of the Naive CIV Estimator

back

Numerator of $\sqrt{Kn_Z}(\hat{\tau}_n - \tau_0)$ written as $O_p(1)$-term plus

$$A_n \equiv \frac{1}{\sqrt{Kn_Z}} \sum_{k=1}^{K} \sum_{i=1}^{n_Z} U_{ki} \left(\hat{m}_n(k) - m_0(k) \right).$$

Naive estimator uses $\hat{m}_n(k) = \frac{1}{n_Z} \sum_{i=1}^{n_Z} D_{ki}$ so that

$$\begin{aligned} A_n &= \frac{1}{\sqrt{Kn_Z}} \sum_{k=1}^{K} \sum_{i=1}^{n_Z} U_{ki} \left(\frac{1}{n_Z} \sum_{i=1}^{n_Z} V_{ki} \right) \\ &= \frac{\sqrt{n_Z}}{\sqrt{K}} \sum_{k=1}^{K} \left(\frac{1}{n_Z} \sum_{i=1}^{n_Z} U_{ki} \right) \left(\frac{1}{n_Z} \sum_{i=1}^{n_Z} V_{ki} \right) \end{aligned}$$

In expectation, $E[A_n] \approx \sqrt{K/n_Z} Cov(U_{ki}, V_{ki})$.

- Diverges unless $K/n_Z = K^2/n \to c < \infty$

Weakly Causal Interpretation

Under the LATE assumptions, we have

$$\tau_0 = \sum_{m=1}^{K} \lambda_m \text{LATE}(z_m, z_{m-1})$$

where

$$\text{LATE}(z_m, z_{m-1}) = E\left[Y(1) - Y(0) | D(z_m) > D(z_{m-1})\right]$$

and

$$\lambda_m \equiv \frac{(m_0(z_m) - m_0(z_{m-1}))\left(\sum_{l=m}^{K}(m_0(z_l) - E[D])\, m_0(z_l)\right)}{\sum_{j=1}^{K}(m_0(z_j) - m_0(z_{j-1}))\left(\sum_{l=j}^{K}(m_0(z_l) - E[D])\, m_0(z_l)\right)}$$

Importantly: $\lambda_m \geq 0, \forall m$ and $\sum_{m=1}^{K} \lambda_m = 1$

Estimating the Number of K_0

back

Connection to factor model literature. Following Bai and Ng (2002)

$$I(M) = \frac{1}{Kn_Z} \sum_{k=1}^{K} \sum_{i=1}^{n_Z} \left(D_{ki} - \hat{m}^{(K)}(k) \right)^2 + M \times h(K, n_Z),$$

where $\hat{m}^{(M)}$ is the estimator w/ M support points, and h is such that

- ▷ $\lim_{K, n_Z \to \infty} h(K, n_Z) = 0$,
- ▷ $\lim_{K, n_Z \to \infty} min(K, n_Z) h(K, n_Z) = \infty$.

Then take

$$\hat{K} = \operatorname*{arg\,min}_{M \in \{1, \ldots, K_{max}\}} I(M).$$

Known K_{max} crucial for consistency of $\hat{K}$ *and* semiparametric efficiency.

Proof Sketch

back

Proof in three steps:

1. Show that $\forall \delta > 0 : \hat{m}_n = \tilde{m}_n + o_p(n^{-\delta})$
2. Show that $\hat{\tau}_n = \tilde{\tau}_n + o_p(n^{-\delta})$
3. Show that
$$\sqrt{n}(\tilde{\tau}_n - \tau_0) \xrightarrow{d} N(0, \sigma^2),$$
where $\sigma^2 = Var(m_0(Z)U)/Var(m_0(Z))^2$.

Proof Sketch (Contd.)

back

Step 1. heavily leverages arguments of Bonhomme and Manresa (2015)

Most importantly:

Lemma 1 (Lemma B.5 in Bonhomme and Manresa (2015))

Let z_t be a strongly mixing process with zero mean, with strong mixing coefficients $\alpha[t] \leq \exp\left(-at^{d_1}\right)$, and with tail probabilities $P(|z_t| > z) \leq \exp\left(1 - \left(\frac{z}{b}\right)^{d_2}\right)$, where a, b, d_1, and d_2 are positive constants. Then, $\forall z \geq 0$, we have, $\forall \delta > 0$,

$$T^{\delta} P\left(\left|\frac{1}{T}\sum_{t=1}^{T} z_t\right| \geq z\right) \overset{T \to \infty}{\to} 0. \tag{5}$$

Application:

- "Missclassification" probability vanishes exponentially
- Can learn partition $(\mathcal{Z}_g)_{g=1}^{K_0}$ of supp Z *very quickly*