Interactive Proofs
For Verifying ML

Jonathan Shafer

MIT

March 2026

Interactive Proofs

For Verifying ML

Example: Delegation of Machine Learning

Pay someone else to do the work!

Example: Delegation of Machine Learning

Seller

Example: Delegation of Machine Learning

Seller

Buyer

Example: Delegation of Machine Learning

AI Models for sale!

Buyer

Example: Delegation of Machine Learning

Fresh! Organic! Gluten-free!

Buyer

Example: Delegation of Machine Learning

Supposedly:

- Collects lots of good data
- Trains good ML model

Buyer

Example: Delegation of Machine Learning

Here's my model.
Want to try it out?

Buyer

Example: Delegation of Machine Learning

Here's my model.
Want to try it out?

80% accuracy on validation set

Example: Delegation of Machine Learning

Price: $1,000,000

80% accuracy on validation set

Example: Delegation of Machine Learning

Price: $1,000,000

80% accuracy on validation set
Should they accept?

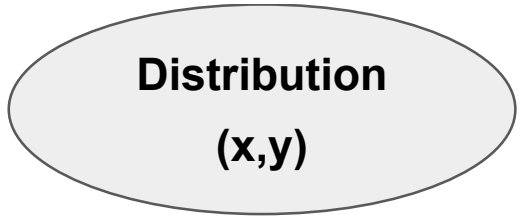
Example: Delegation of Machine Learning

Price: $1,000,000

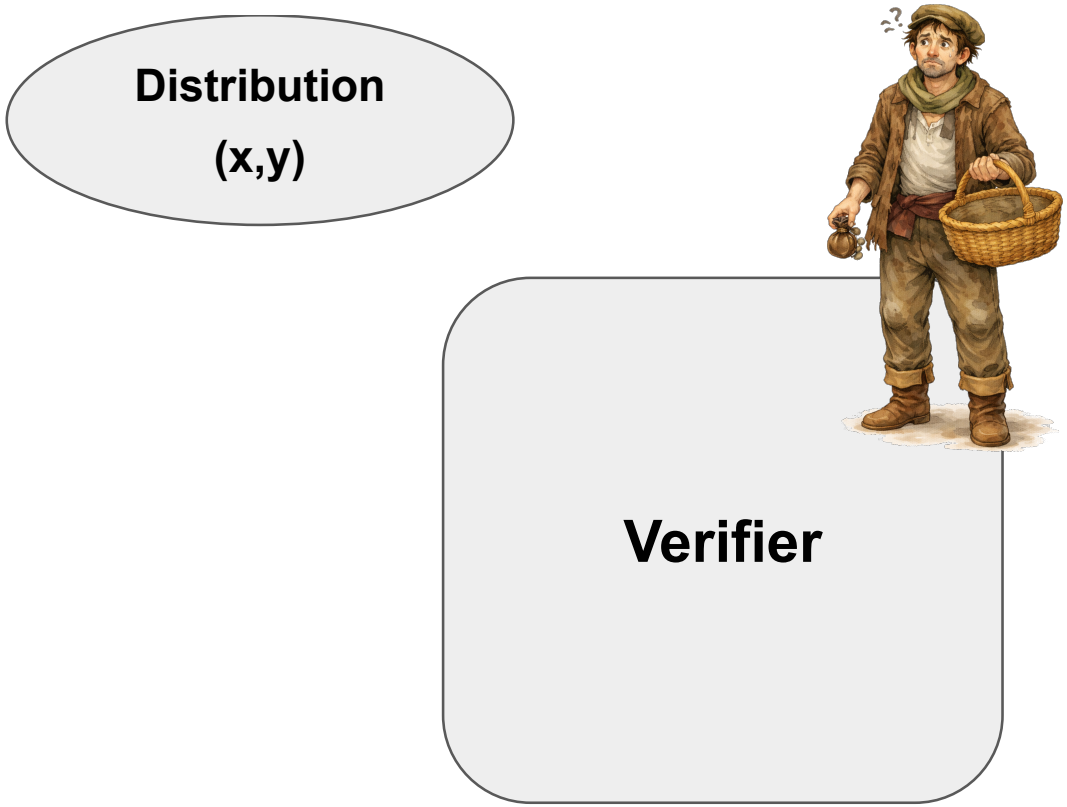
80% accuracy on validation set

Compare to benchmark!

Verification



Verification



Verification

Distribution

(x,y)

Learner / Prover

Verifier

Verification

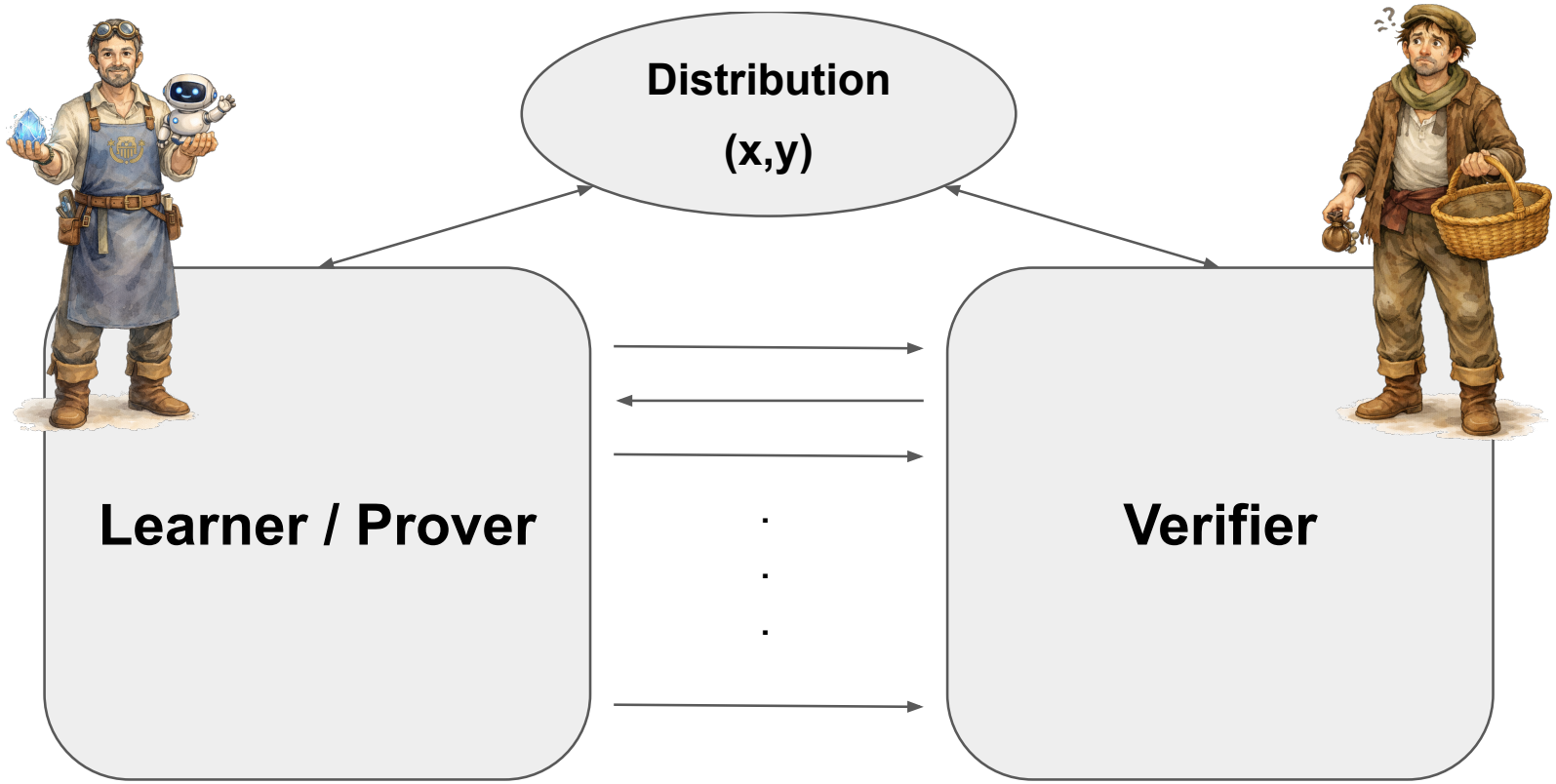
Distribution

(x,y)

Learner / Prover

Verifier

Verification



Verification

Distribution

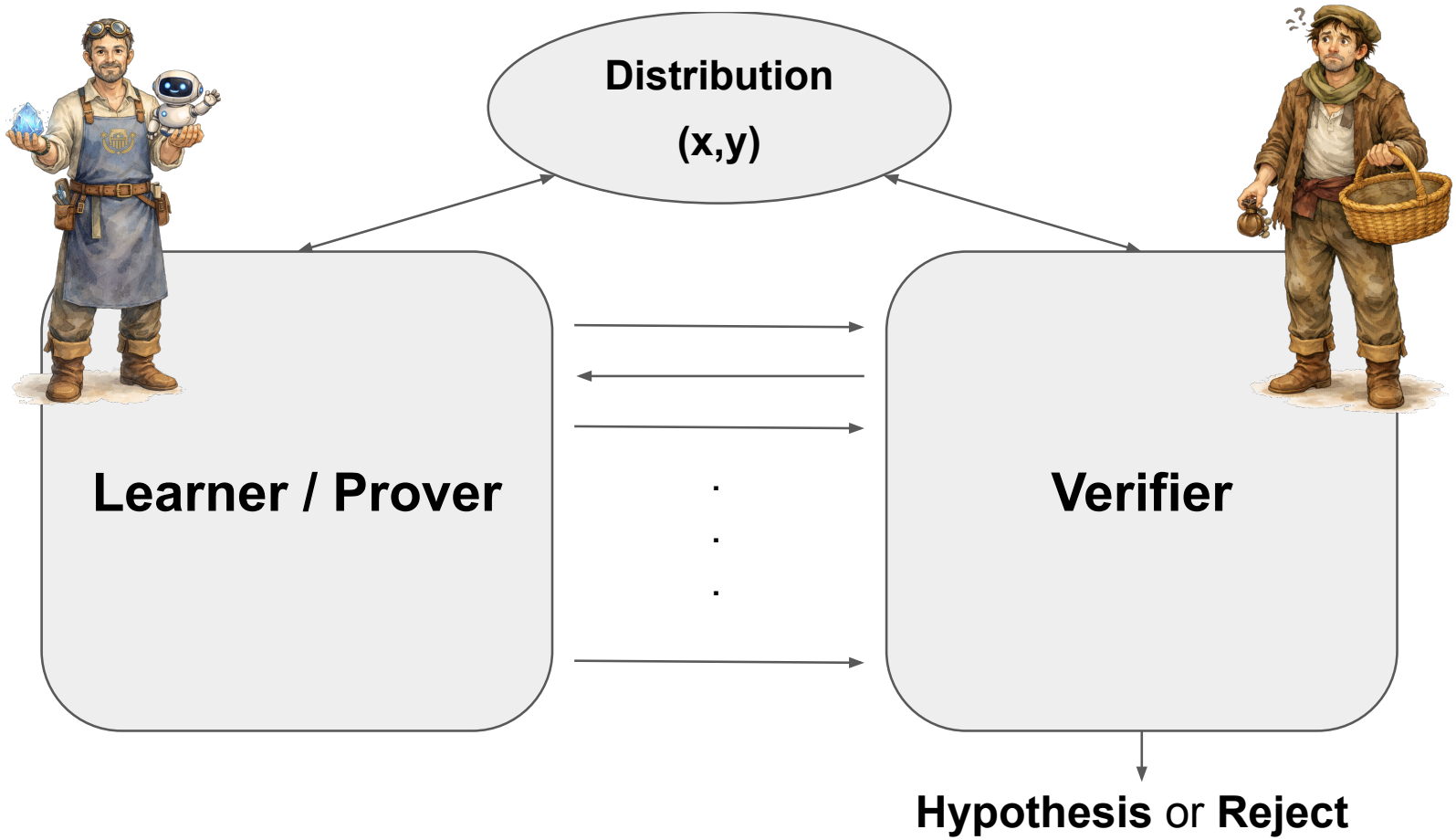
(x,y)

Learner / Prover

Verifier

Hypothesis or **Reject**

Verification



Verification must be cheaper than learning

Definition (GRSY21).

Definition (GRSY21). $\mathcal{H}$ is verifiable if

Definition (GRSY21). $\mathcal{H}$ is verifiable if

$$\exists\ V, P$$

Definition (GRSY21). $\mathcal{H}$ is verifiable if

$\exists\ V, P$

$\forall\ \varepsilon, \delta \geq 0\ \forall$ distribution $\mathcal{D}$:

Definition (GRSY21). $\mathcal{H}$ is verifiable if

$\exists\ V, P$

$\forall\ \varepsilon, \delta \geq 0\ \forall$ distribution $\mathcal{D}$:

Completeness.

Definition (GRSY21). $\mathcal{H}$ is verifiable if

$\exists\ V, P$

$\forall\ \varepsilon, \delta \geq 0\ \forall$ distribution $\mathcal{D}$:

Completeness. Honest P convinces V to output ε-optimal classifier

Definition (GRSY21). $\mathcal{H}$ is verifiable if

$\exists\ V, P$

$\forall\ \varepsilon, \delta \geq 0\ \forall$ distribution $\mathcal{D}$:

Completeness.

Definition (GRSY21). $\mathcal{H}$ is verifiable if

$\exists\ V, P$

$\forall\ \varepsilon, \delta \geq 0\ \forall$ distribution $\mathcal{D}$:

Completeness. $h = [V, P]$ satisfies

Definition (GRSY21). $\mathcal{H}$ is verifiable if

$\exists\ V, P$

$\forall\ \varepsilon, \delta \geq 0\ \forall$ distribution $\mathcal{D}$:

Completeness. $h = [V, P]$ satisfies

$$\mathbb{P}[\qquad\qquad\qquad\qquad] \geq 1 - \delta$$

Definition (GRSY21). $\mathcal{H}$ is verifiable if

$\exists\ V, P$

$\forall\ \varepsilon, \delta \geq 0\ \forall$ distribution $\mathcal{D}$:

Completeness. $h = [V, P]$ satisfies

$$\mathbb{P}[h \neq \text{reject} \qquad\qquad] \geq 1 - \delta$$

Definition (GRSY21). $\mathcal{H}$ is verifiable if

$\exists\ V, P$

$\forall\ \varepsilon, \delta \geq 0\ \forall$ distribution $\mathcal{D}$:

Completeness. $h = [V, P]$ satisfies

$$\mathbb{P}[h \neq \text{reject} \ \wedge \ L_{\mathcal{D}}(h) \leq \text{Opt}_{\mathcal{H}} + \varepsilon] \geq 1 - \delta$$

Definition (GRSY21). $\mathcal{H}$ is verifiable if

$\exists\ V, P$

$\forall\ \varepsilon, \delta \geq 0\ \forall$ distribution $\mathcal{D}$:

Completeness. $h = [V, P]$ satisfies

$$\mathbb{P}[h \neq \text{reject} \ \wedge\ L_{\mathcal{D}}(h) \leq \text{Opt}_{\mathcal{H}} + \varepsilon] \geq 1 - \delta$$

Soundness. No P' convinces the V to output bad classifier

Definition (GRSY21). $\mathcal{H}$ is verifiable if

$\exists\ V, P$

$\forall\ \varepsilon, \delta \geq 0\ \forall$ distribution $\mathcal{D}$:

Completeness. $h = [V, P]$ satisfies

$$\mathbb{P}[h \neq \text{reject} \ \wedge\ L_{\mathcal{D}}(h) \leq \text{Opt}_{\mathcal{H}} + \varepsilon] \geq 1 - \delta$$

Soundness.

Definition (GRSY21). $\mathcal{H}$ is verifiable if

$\exists\ V, P$

$\forall\ \varepsilon, \delta \geq 0\ \forall$ distribution $\mathcal{D}$:

Completeness. $h = [V, P]$ satisfies

$$\mathbb{P}[h \neq \text{reject} \ \wedge\ L_{\mathcal{D}}(h) \leq \text{Opt}_{\mathcal{H}} + \varepsilon] \geq 1 - \delta$$

Soundness. $\forall P'$: $h = [V, P']$ satisfies

Definition (GRSY21). $\mathcal{H}$ is verifiable if

$\exists\ V, P$

$\forall\ \varepsilon, \delta \geq 0\ \forall$ distribution $\mathcal{D}$:

Completeness. $h = [V, P]$ satisfies

$$\mathbb{P}[h \neq \text{reject} \ \wedge \ L_{\mathcal{D}}(h) \leq \text{Opt}_{\mathcal{H}} + \varepsilon] \geq 1 - \delta$$

Soundness. $\forall P'$: $h = [V, P']$ satisfies

$$\mathbb{P}[\qquad\qquad\qquad] \geq 1 - \delta$$

Definition (GRSY21). $\mathcal{H}$ is verifiable if

$\exists\ V, P$

$\forall\ \varepsilon, \delta \geq 0\ \forall$ distribution $\mathcal{D}$:

Completeness. $h = [V, P]$ satisfies

$$\mathbb{P}[h \neq \text{reject} \ \wedge \ L_{\mathcal{D}}(h) \leq \text{Opt}_{\mathcal{H}} + \varepsilon] \geq 1 - \delta$$

Soundness. $\forall P'$: $h = [V, P']$ satisfies

$$\mathbb{P}[h = \text{reject} \qquad\qquad\qquad\qquad] \geq 1 - \delta$$

Definition (GRSY21). $\mathcal{H}$ is verifiable if

$\exists\ V, P$

$\forall\ \varepsilon, \delta \geq 0\ \forall$ distribution $\mathcal{D}$:

Completeness. $h = [V, P]$ satisfies

$$\mathbb{P}[h \neq \text{reject} \ \wedge\ L_{\mathcal{D}}(h) \leq \text{Opt}_{\mathcal{H}} + \varepsilon] \geq 1 - \delta$$

Soundness. $\forall P'$: $h = [V, P']$ satisfies

$$\mathbb{P}[h = \text{reject} \ \vee\ L_{\mathcal{D}}(h) \leq \text{Opt}_{\mathcal{H}} + \varepsilon] \geq 1 - \delta$$

Definition (GRSY21). $\mathcal{H}$ is verifiable if

$\exists\ V, P$

$\forall\ \varepsilon, \delta \geq 0\ \forall$ distribution $\mathcal{D}$:

Completeness. $h = [V, P]$ satisfies

$$\mathbb{P}[h \neq \text{reject} \ \wedge\ L_{\mathcal{D}}(h) \leq \text{Opt}_{\mathcal{H}} + \varepsilon] \geq 1 - \delta$$

Soundness. $\forall P'$: $h = [V, P']$ satisfies

$$\mathbb{P}[h = \text{reject} \ \vee\ L_{\mathcal{D}}(h) \leq \text{Opt}_{\mathcal{H}} + \varepsilon] \geq 1 - \delta$$

Realizable case is easy

Result: Upper Bound

Theorem (MS23, GRSY21).

Result: Upper Bound

Theorem (MS23, GRSY21). $\exists$ class $\mathcal{H}_d$

Result: Upper Bound

Theorem (MS23, GRSY21). $\exists$ class $\mathcal{H}_d$ with $\mathrm{VC}(\mathcal{H}_d) = d$

Result: Upper Bound

Theorem (MS23, GRSY21). $\exists$ class $\mathcal{H}_d$ with $\mathrm{VC}(\mathcal{H}_d) = d$ that is verifiable

Result: Upper Bound

Theorem (MS23, GRSY21). $\exists$ class $\mathcal{H}_d$ with $\mathrm{VC}(\mathcal{H}_d) = d$ that is verifiable with V using

$$O\left(\frac{\sqrt{d}}{\varepsilon^{2.5}}\right)$$

i.i.d. samples.

Result: Upper Bound

Theorem (MS23, GRSY21). $\exists$ class $\mathcal{H}_d$ with $\mathrm{VC}(\mathcal{H}_d) = d$ that is verifiable with V using

$$O\left(\frac{\sqrt{d}}{\varepsilon^{2.5}}\right)$$

i.i.d. samples.

Proof Idea

1. P sends discretized distribution with support of size $k = O(d/\varepsilon)$

Result: Upper Bound

Theorem (MS23, GRSY21). $\exists$ class $\mathcal{H}_d$ with $\mathrm{VC}(\mathcal{H}_d) = d$ that is verifiable with V using

$$O\left(\frac{\sqrt{d}}{\varepsilon^{2.5}}\right)$$

i.i.d. samples.

Proof Idea

1. P sends discretized distribution with support of size $k = O(d/\varepsilon)$
2. V performs distribution identity testing using $O\left(\sqrt{k}/\varepsilon^2\right)$ samples

Result: Lower Bound

Theorem (MS23). $\forall$ class $\mathcal{H}$:

Result: Lower Bound

Theorem (MS23). $\forall$ class $\mathcal{H}$:

if $\mathrm{VC}(\mathcal{H}) = d$

Result: Lower Bound

Theorem (MS23). $\forall$ class $\mathcal{H}$:

if $\mathrm{VC}(\mathcal{H}) = d$

and (V, P) verifies $\mathcal{H}$

Result: Lower Bound

Theorem (MS23). $\forall$ class $\mathcal{H}$:

if $\mathrm{VC}(\mathcal{H}) = d$

and (V, P) verifies $\mathcal{H}$ then:

V uses $\Omega\left(\sqrt{d}/\varepsilon^2\right)$ i.i.d. samples.

Result: Lower Bound

Theorem (MS23). $\forall$ class $\mathcal{H}$:

if $\mathrm{VC}(\mathcal{H}) = d$

and (V, P) verifies $\mathcal{H}$ then:

V uses $\Omega\left(\sqrt{d}/\varepsilon^2\right)$ i.i.d. samples.

Proof Idea. Reduction from distribution testing lower bound of [Pan08].

More Verification:

- Qualitative separation [GRSY21]
- General statistical algorithms [MS23]
- Quantum [CHI+24]
- AC^0, juntas, improved GL [GJK+24]
- Equilibria in games [CRS25]
- ...

Thank You!

References

[CHI+24] Matthias C. Caro, Marcel Hinsche, Marios Ioannou, Alexander Nietner, and Ryan Sweke. Classical verification of quantum learning. In Venkatesan Guruswami, editor, *15th Innovations in Theoretical Computer Science Conference, ITCS 2024, Berkeley, CA, USA, January 30 - February 2, 2024*, volume 287 of *LIPIcs*, pages 24:1–24:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2024.

[CRS25] Miranda Christ, Daniel Reichman, and Jonathan Shafer. Protocols for verifying smooth strategies in bandits and games. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[GJK+24] Tom Gur, Mohammad Mahdi Jahanara, Mohammad Mahdi Khodabandeh, Ninad Rajgopal, Bahar Salamatian, and Igor Shinkar. On the power of interactive proofs for learning. In Bojan Mohar, Igor Shinkar, and Ryan O'Donnell, editors, *Proceedings of the 56th Annual ACM Symposium on Theory of Computing, STOC 2024, Vancouver, BC, Canada, June 24-28, 2024*, pages 1063–1070. ACM, 2024.

[GRSY21] Shafi Goldwasser, Guy N. Rothblum, Jonathan Shafer, and Amir Yehudayoff. Interactive proofs for verifying machine learning. In James R. Lee, editor, *12th Innovations in Theoretical Computer Science Conference, ITCS 2021, January 6-8, 2021, Virtual Conference*, volume 185 of *LIPIcs*, pages 41:1–41:19. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021.

[MS23] Saachi Mutreja and Jonathan Shafer. PAC verification of statistical algorithms. In Gergely Neu and Lorenzo Rosasco, editors, *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*, volume 195 of *Proceedings of Machine Learning Research*, pages 5021–5043. PMLR, 2023.

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.